INTRODUCCIÓN AL

BIG DATA

José Francisco Aldana Montes
Alejandro Baldominos Gómez
José Manuel García Nieto
Juan Carlos González Cabañas
Francisco Mochón Morcillo
Ismael Navas Delgado

Introducción al Big Data

José Francisco Aldana Montes
Alejandro Baldominos Gómez
José Manuel García Nieto
Juan Carlos Gonzálvez Cabañas
Francisco Mochón Morcillo
Ismael Navas Delgado

MADRID

Introducción al Big Data

ISBN: 978-84-15793-94-6
Depósito legal: M-8052-2016

Preimpresión: FER Fotocomposición
Impreso en: FER Impresión digital. c/ Alfonso Gómez, 38, 3.° C - 28037 Madrid

IMPRESO EN ESPAÑA - PRINTED IN SPAIN

Dedicado a todos aquellos con ganas de empezar a trabajar con cantidades casi inabarcables de datos. ¡Ánimo!

José Francisco Aldana Montes
Catedrático de Universidad
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga

Alejandro Baldominos Gómez
Estudiante de Doctorado
Departamento de Informática
Universidad Carlos III de Madrid

José Manuel García Nieto
Investigador Doctor
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga

Juan Carlos Gonzálvez Cabañas
Director de Gestión de Innovación y Nuevos Productos
ZED Worldwide

Francisco Mochón Morcillo
Catedrático de Universidad
Departamento de Análisis Económico II
Universidad Nacional de Educación a Distancia

Ismael Navas Delgado
Profesor Contratado Doctor
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga

Tabla de Contenidos

Introducción al Big Data

Prólogo

La importancia del Big Data descansa en la inmensa cantidad de datos que se generan cada día, especialmente a raíz de la eclosión de las redes sociales online (Facebook, Twitter, Google Plus. etc.) y del crecimiento exponencial de dispositivos, tales como los *smartphones*, *smartwatches*, *wearables*, etc. o las redes de sensores; así como en la posibilidad de tener en cuenta información cada vez más actualizada y variada para la toma de decisiones.

Esta proliferación de datos generados proporciona información que, debidamente utilizada, permite que el proceso de toma de decisiones sea más objetivo y se base menos en la intuición.

El Big Data permite detectar tendencias, obtener modelos con información histórica para realizar predicción de sucesos futuros, o extraer patrones del comportamiento de los usuarios, adaptando mejor los servicios a sus necesidades.

En este libro se exponen de forma sencilla las distintas fases del proceso de gestión de los datos: recolección, almacenamiento, procesamiento, análisis, visualización, toma de decisiones y estrategias a seguir. Se presentan los conocimientos necesarios para poder identificar las técnicas y herramientas requeridas para gestionar de forma eficiente grandes cantidades de datos.

El libro puede ser de interés tanto para lectores que no tengan una formación técnica, como para aquellos con formación o amplia experiencia en el mundo de las TIC. Pensando en los primeros, se introducen con claridad los aspectos básicos de uso de las tecnologías de gestión de la información, facilitando su comprensión para un amplio espectro

de lectores, independientemente de cuáles sean sus conocimientos o su experiencia. Para los lectores con un nivel técnico elevado este libro aporta, además de conocimientos específicos en herramientas para la integración de Big Data, una formación muy orientada a la utilización de las citadas herramientas para resolver problemas reales de las organizaciones.

Para tratar de alcanzar estos objetivos, los temas se introducen por un lado a un nivel eminentemente aplicado, de forma que los lectores sin conocimientos técnicos previos puedan visualizar y comprobar, vía ejemplos y ejercicios, las posibilidades que ofrece el Big Data. Por otro lado, pensando en aquellos alumnos que tengan una formación técnica suficiente, los principios teóricos esbozados y el uso de las herramientas propuestas en el curso, constituyen una rigurosa introducción al manejo de los datos.

El contenido del libro se ha estructurado de forma que se ofrezca una visión global de todos los temas que forman parte de un análisis de Big Data y está integrado por once capítulos.

En el primer capítulo se realiza una presentación de lo que supone la gestión de los datos. El capítulo segundo se dedica a introducir el concepto de Big Data y esbozar el tipo de casos que se pueden abordar con esta herramienta de análisis. El capítulo tercero presenta la forma en que se debe abordar el Big Data en las organizaciones. Con el capítulo cuarto se inicia un bloque de capítulos de contenido más aplicado; en concreto se realiza una introducción a lo que supone el trabajo con datos. El capítulo quinto se dedica a estudiar cómo se realiza la recolección de datos y en el sexto se presentan técnicas y herramientas para llevar a cabo el almacenamiento de estos datos. En el capítulo séptimo se indica cómo realizar un procesamiento y transformación eficaz de los datos, mientras que el octavo presenta con cierto detalle cómo realizar un análisis de Big Data para extraer valor de negocio. El capítulo noveno trata la visualización y consumo de datos, con el fin de producir informes y visualizaciones que permitan transmitir la esencia de los datos de forma eficaz. Finalmente, el capítulo décimo se ocupa de presentar conceptos de seguridad y gobernanza, y en último capítulo se presenta un caso de estudio que aplica todos los conceptos y procedimientos explicados anteriormente.

Si se trabajan de forma suficiente los temas señalados, los lectores serán capaces de:

- Saber qué es Big Data (Volumen, Velocidad y Variedad en la gestión de los datos), cuáles son sus orígenes, sus principales fuentes, los retos a los que se enfrenta y los principales tipos de aplicaciones que se pueden llevar a cabo.
- Determinar cómo y cuándo surgen estos retos en diferentes campos de actividad.
- Identificar las consecuencias que la generalización del uso del Big Data está teniendo en las organizaciones.

- Conocer las diferentes etapas del trabajo con datos.
- Valorar las ventajas e inconvenientes del uso de la computación en la nube en el despliegue de plataformas de Big Data.
- Explorar el modelo relacional SQL y su aplicación al Big Data
- Entender las características de los sistemas NoSQL así como sus ventajas e inconvenientes.
- Identificar MapReduce como paradigma de procesamiento de Big Data, así como reconocer las ventajas, beneficios y novedades que supone su uso en cuanto al rendimiento y la robustez.
- Conocer en qué consisten las técnicas de análisis predictivo y análisis de patrones, y entender el funcionamiento de los principales algoritmos de aprendizaje automático (tanto supervisado como no supervisado).
- Valorar la importancia que la visualización de los datos tiene en entornos Big Data.
- Entender la relevancia y complejidad que la seguridad y la gobernanza de los datos tienen cuando trabajamos en entornos Big Data.
- Reconocer de forma práctica cómo se desarrollan casos de uso reales de análisis del Big Data, desarrollando y discutiendo versiones simplificadas pero completas.

El libro se ha estructurado con un carácter eminentemente aplicado y a lo largo de los distintos capítulos se proponen ejercicios y actividades, siguiendo el principio: «la mejor forma de aprender es haciendo». Además, en el último capítulo se presenta un caso de estudio completo.

Por ultimo, cabe señalar que este libro se ha concebido como un texto de introducción al Big Data y se ha redactado procurando que pueda utilizarse como manual de referencia para realizar una primera toma de contacto con el análisis de los datos. Por ello, los conceptos se abordan de forma relativamente sencilla pero rigurosa. En cualquier caso, debe tenerse en cuenta que los temas objeto de estudio, si bien son sumamente interesantes, presentan un cierto grado de complejidad; y por este motivo se ha optado por mostrar cómo se deben abordar, pero sin entrar en los fundamentos teóricos y matemáticos de los mismos. En definitiva, este libro debe considerarse como lo que es; un manual de introducción al Big Data, con un marcado enfoque aplicado; por lo que puede ser especialmente idóneo para utilizarse en un curso de introducción a la materia.

1 La Gestión de los Datos

1.1 Evolución en los Sistemas de Gestión y Análisis de Datos

1.1.1 Introducción

Desde que, allá por los años 60 del siglo XX, la computación irrumpiese en el mundo empresarial, ha habido grandes cambios en los sistemas orientados a la gestión y el análisis de los datos. La evolución en estos sistemas habitualmente ha surgido como consecuencia de una nueva necesidad de resolver un problema específico asociado a los datos, ya sea de capacidad de almacenamiento, de procesamiento o de gestión y análisis de la información.

Cada cambio producido, independientemente de la trascendencia final que haya tenido, se ha fundamentado en las bases de las técnicas previas.

Habitualmente se asocian estos cambios a la evolución del software: la aparición de nuevos lenguajes de programación, de nuevos procedimientos o de nuevos paradigmas, sin embargo, si analizamos el contexto en el que se han producido, nos daremos cuenta de que en realidad han sido consecuencia de dos factores:

- **La aparición y consolidación de nuevos avances tecnológicos**. Estos han tenido lugar, no sólo en el terreno del software, sino también en campos como el hardware, el almacenamiento, las redes o la aparición de nuevos modelos computacionales (como por ejemplo en la actualidad el Cloud Computing).
- **El continuo abaratamiento de los costes asociados a estas tecnologías**. Procesadores más rápidos y baratos, discos con mayores capacidades de almace-

namiento y más eficientes a costes similares que sus predecesores, redes con velocidades de transferencia superiores sin incrementos de costes o infraestructuras de alta calidad sin necesidad de realizar grandes inversiones iniciales y con capacidad para adaptarse a las necesidades de crecimiento de forma dinámica, que son consumidos y pagados como servicio (Virtualización de servidores, IaaS - *Infrastructure as a Service*).

Cuando hacemos referencia a sistemas de gestión y análisis de datos, podemos distinguir dos tipos de sistemas:

- **Sistemas operacionales**. Entendidos como aquellos que dan soporte a las herramientas operativas de la empresa. Son sistemas que permiten la realización de las transacciones que las empresas necesitan realizar para el desarrollo de su actividad. Tradicionalmente se utilizan bases de datos relacionales como sistemas de gestión que dan soporte al almacenamiento y procesamiento.

  Algunos ejemplos de este tipo de sistemas son los siguientes: los sistemas sobre los que se realizan las operaciones que se llevan a cabo en un cajero automático, las operaciones que se realizan en un terminal punto de venta de una caja de un supermercado, las bases de datos en las que se sustenta un ERP (*Enterprise Resource Planning*) de una compañía o las operaciones que debe procesar un AdServer para realizar las impresiones publicitarias programadas; son algunos ejemplos de este tipo de sistemas.

  A los sistemas operacionales se les denomina sistemas transaccionales. Un sistema transaccional debe controlar las transacciones para mantener la seguridad y consistencia de los datos involucrados. A este tipo de sistemas se les conoce como sistemas OLTP (*OnLine Transaction Processing*, Procesamiento de Transacciones en Línea), por su capacidad para facilitar y administrar aplicaciones transaccionales, ya que normalmente son utilizados para entrada de datos, recuperación y procesamiento de transacciones.

- **Sistemas de apoyo a la toma de decisiones**. El concepto de sistema de soporte a la toma de decisiones (DSS, *Decision Support System*), puede definirse como un conjunto de herramientas que permiten realizar el análisis de las diferentes variables de negocio para apoyar el proceso de toma de decisiones de los directivos. La información en la que se basan estos sistemas procede fundamentalmente de los sistemas operacionales de la compañía, aunque cada vez más, esta información se combina, complementa y enriquece con información procedente de fuentes externas de información.

  A este tipo de sistemas se los denomina OLAP (*OnLine Analytical Processing*, procesamiento analítico en línea) y su objetivo es agilizar el análisis de grandes cantidades de datos.

La finalidad de los sistemas operacionales es controlar las transacciones que se realizan en las organizaciones, de forma que se mantenga la seguridad y consistencia de los datos. Así, por ejemplo, cuando un cliente transfiere dinero de una cuenta a otra cuenta dentro de una misma entidad financiera, la cantidad de dinero que se descuenta de la cuenta emisora debe ser igual a la que se suma en la cuenta receptora. De no ser así, la acción (transacción) no se realiza. Por esta razón los sistemas operacionales, también se denominan como sistemas transaccionales.

Los sistemas de apoyo a la toma de decisiones, se denominan de forma genérica como sistemas informacionales. Un sistema informacional es un tipo de sistema diseñado para recolectar, almacenar, modificar y recuperar todo tipo de información que es generada por las transacciones en una organización, no para gestionar las transacciones en sí mismas, sino para poder obtener información relevante. Por ejemplo, tener información de todas las transacciones que se producen a lo largo del año puede ayudarnos a hacer previsiones a futuro o identificar tendencias de mercado.

1.1.2 Historia de los Sistemas de Apoyo a la Toma de Decisiones

En las primeras etapas del desarrollo de los sistemas de apoyo a la toma de decisiones, las necesidades de información por parte de las empresas eran muy simples, se limitaban a la mera descripción de determinados parámetros, como por ejemplo identificar cuantos artículos se han vendido en un establecimiento, o la contabilización de las existencias de un almacén. En estos casos, la aproximación generalmente adoptada consistía en realizar herramientas ad-hoc para la generación de informes. El proceso en estos casos se basaba en realizar peticiones de tipos de informes, que el departamento TI (Tecnologías de la Información) se encargaba de desarrollar, a través de herramientas orientadas a generar estos informes a partir de los datos operacionales de la empresa. Sin embargo, en muchas ocasiones los departamentos TI no llegaban a consolidar las herramientas en un sistema completo, y se limitaban a realizar software sencillo capaz de generar el informe que se pedía en cada momento.

El procedimiento era un proceso completamente reactivo a la demanda de los diferentes departamentos de la compañía y se utilizaba normalmente lo que se denomina «fuerza bruta», es decir desarrollos a medida que debían repetirse o crearse cada vez que se realizaba una petición. Estos desarrollos estaban normalmente creados en lenguajes de programación de alto nivel que sólo el personal especializado del departamento TIC era capaz de ejecutar.

Un poco más adelante, cuando los sistemas tratan de anticipar la respuesta a las posibles peticiones, el software comienza a organizarse para reducir el tiempo entre la solicitud de un nuevo tipo de informe y la capacidad de comenzar a generarlos cada vez que fueran necesarios.

Posteriormente, se identifica un nuevo problema relevante que es el proceso de extracción de los datos a partir de los sistemas operacionales. En una primera aproximación, los datos extraídos para hacer un informe se conservan fuera del sistema transaccional por si podían ser reutilizados en otros tipos de informes sin necesidad de volver a extraer datos de estos sistemas vitales para el funcionamiento diario de la empresa. Basándose en esta idea, se comienzan a desarrollar pequeñas aplicaciones de extracción de datos que son capaces de repetir un proceso de extracción de datos del sistema transaccional. Estos datos, fuera del sistema transaccional, sirven para la generación de aplicaciones que permitan crear informes a medida en base a parámetros que pudieran definir los usuarios para filtrar los datos. Estos informes podían ser generados como documentos que posteriormente eran usados en la toma de decisiones o en los sistemas más avanzados, además de mostrarse en pantalla para su uso por parte de los directivos encargados de esta toma de decisiones.

A comienzos de los años 70 las grandes empresas comienzan a crear lo que se denominaban centros de información, que básicamente eran centros donde los usuarios podían solicitar informes específicos o incluso visualizaciones de los mismos en pantalla. En estos casos se trataba normalmente de informes predeterminados o volcados en pantalla. En estos centros de información se disponía de personal TIC capaz de dar soporte a los usuarios en la generación de estos informes.

1.1.2.1 El modelo relacional

Más tarde, durante la década de 1970, las cosas cambiaron con la invención del modelo de datos relacional y el sistema de gestión de bases de datos relacionales (*Relational Database Management System*, RDBMS) que impuso una estructura y un método para mejorar el rendimiento.

El modelo relacional es una forma de representar la información, basado en el concepto matemático de relación. En este modelo la información se representa en forma de "tablas" o relaciones, donde cada fila de la tabla se interpreta como una relación ordenada de valores (un conjunto de valores relacionados entre sí).

Una de las cosas más relevantes que introdujo el modelo relacional fue la capacidad de añadir un nuevo nivel de abstracción (el lenguaje de consultas estructurado –SQL–, generadores de informes y herramientas de gestión de datos) que permitía a los programadores extraer los datos de los sistemas operacionales de una forma más sencilla, de manera que fueran capaces de satisfacer las crecientes demandas del negocio para extraer valor de los datos.

El modelo relacional ofrecía un conjunto de herramientas que permitían ayudar a las empresas en su creciente necesidad de organizar sus datos y realizar comparaciones y correlaciones entre los mismos. Los gerentes podían realizar comparaciones que les facilitaba

la toma de decisiones. Por ejemplo se podían comparar las existencias de mercancías en el almacén con los pedidos realizados por los clientes.

Sin embargo, estas nuevas funcionalidades no estuvieron exentas de problemas. La creciente demanda de información útil disparaba el número de peticiones, lo que hacía que las necesidades de almacenamiento de información creciesen y almacenar este creciente volumen de datos era caro. Además, cuanto mayor era el volumen de información almacenado más lento era el acceso. Para empeorar las cosas, se producía un tremendo volumen de datos duplicados, haciendo más difícil la aportación de valor al negocio.

En aquel momento, la tecnología existente tenía la necesidad urgente de encontrar un nuevo conjunto de tecnologías para apoyar el modelo relacional. Surgió entonces el modelo Entidad-Relación (ER) que añadía un nivel adicional de abstracción permitiendo optimizar los datos y aumentar la capacidad ser utilizados. Este modelo permite a los desarrolladores crear nuevas relaciones entre las fuentes de datos sin necesidad de tener que recurrir a rutinas de programación complejas. El modelo ER supuso un gran avance permitiendo a los desarrolladores crear modelos más complejos. El modelo ER permitió que el mercado de las bases de datos relacionales creciese de forma exponencial y todavía hoy en día sigue vigente y en pleno uso, siendo especialmente importante para la gestión de datos de sistemas transaccionales.

Sin embargo, esta explosión en el uso de los datos, hizo que el volumen que era necesario gestionar creciese también exponencialmente. Aunque hubo propuestas tecnológicas basadas en sacar provecho de los datos o información transaccional que manejaban las aplicaciones de la empresa, pronto se comprendió que estos sistemas no serían capaces de ofrecer soluciones orientadas al tipo de usuario que tenía que realizar procesos de toma de decisiones y que en el caso de grandes empresas, estas aproximaciones podían llegar a interferir con las operaciones diarias de la empresa.

1.1.2.2 Almacenes de datos

Como resultado de este proceso se propuso un nuevo paradigma para permitir un análisis y exploración de datos orientado a la toma de decisiones: los almacenes de datos, o «*Data Warehouse*», el término anglosajón ampliamente extendido en el mundo empresarial. El uso de almacenes de datos proporcionó a las empresas que lo adoptaron en esos comienzos una gran ventaja competitiva.

Los almacenes de datos permitieron a los responsables de tecnología de las empresas seleccionar aquel subconjunto de datos que proporcionaría una mayor información para la toma de decisiones estratégicas de las empresas. Además, seleccionando sólo los datos útiles para la toma de decisiones, se obtienen dos beneficios, por un lado la reducción del volumen de información a almacenar y por otro, la mejora en el rendimiento en el acceso a la misma. Los almacenes de datos tienen además otras ventajas:

- Permiten separar la información operativa de la información utilizada para la toma de decisiones, permitiendo una mejora del rendimiento, ya que ambos sistemas disponen de recursos técnicos independientes.
- Permiten gestionar una mayor cantidad de información de forma eficiente.
- Facilitan el uso de información procedente de distintos orígenes.
- La integración de datos de distinta procedencia permite realizar correlaciones y comparaciones de datos, que hasta la fecha resultaban imposibles de realizar.
- Facilitan el almacenamiento de información de diferentes periodos, lo que permite realizar análisis que hagan aflorar tendencias y patrones de comportamiento.

1.1.2.3 Data marts

A medida que la información crecía y puesto que los almacenes de datos son el repositorio de toda la información transaccional, especialmente para las empresas grandes, los almacenes de datos empezaron a tener que gestionar enormes volúmenes de datos, lo que sin duda afectaba al rendimiento. Surgen entonces los Data Marts, que en lenguaje coloquial, podríamos decir que son almacenes de datos, que contienen información de una sola área o temática. Mientras que el almacén de datos, continúa siendo el lugar donde se almacenan todos los datos de toda la compañía, los *Data Marts* se alimentan de la información que reside en el almacén de datos, pero sólo contendrán un subconjunto de la información de la compañía, aumentando la eficiencia en la gestión de los datos.

Los almacenes de datos y los *Data Marts* resolvieron muchos de los problemas a los que tenían que enfrentarse las empresas en la gestión de los datos, sin embargo estas soluciones sólo eran apropiadas para la gestión de datos estructurados (ver sección 1.2.1). Pero a medida que la tecnología era capaz de almacenar más información, esta se volvía más compleja y los directivos demandaban gestionar información que no sólo procedía de los sistemas operacionales sino también de otras fuentes. En muchos casos estos datos carecían de estructura y procedían de sistemas que no siempre eran bases de datos, por ejemplo ficheros de texto o imágenes. La tecnología también encontró una solución adecuada para gestionar estos datos, surgieron entonces los BLOBs (*Binary Large Objects* que permitían almacenar y gestionar este tipo de datos, y aparecieron las bases de datos orientadas a objetos: ODBMS(*Object Database Management System*).

Las ventajas que para las empresas supone el acceso y utilización de los datos como fuente de ventaja competitiva, ha popularizado el uso de los almacenes de datos en todas las empresas. Sin embargo, las empresas cada vez demandan más información, de manera más rápida y utilizando fuentes más variadas y complejas. Ante estas necesidades, los almacenes de datos no resultaban adecuados, pues se diseñaron para ser alimentados mediante procesamiento por lotes, a intervalos, normalmente diarios. Esta forma de

proceder funciona bien para la planificación financiera, para la generación de informes o incluso para la programación de campañas publicitarias tradicionales, pero resulta demasiado lenta para entornos empresariales en los que se requieren respuestas en tiempo real o entornos en los que se requiere interacción con el consumidor.

El nuevo entorno competitivo en el que las empresas deben desarrollar su actividad comercial ha evolucionado desde un entorno aislado y desconectado, a uno continuamente conectado, en el que Internet se ha convertido en una especie de plataforma incorporada a los procesos de negocio, en la que la información se encuentra en un volumen y una variedad de formatos inusual hasta la fecha. Las organizaciones están empezando a comprender que necesitan gestionar una nueva generación de fuentes de datos, con una cantidad y variedad de datos sin precedentes, que necesita ser procesada una velocidad inaudita.

1.1.3 Orígenes de Big Data

Son muchas las fuentes que están produciendo de forma continuada grandes volúmenes de datos. En esta sección ilustramos algunas de estas fuentes, para que el lector pueda hacerse a la idea de la inmensidad de datos que se producen en el mundo digital. No obstante, el lector podrá imaginar muchas otras fuentes de Big Data que no aparezcan aquí enumeradas.

1.1.3.1 Páginas web

En septiembre de 2014 se superó por primera vez en la historia la cifra de 1000 millones de sitios web activos en la *World Wide Web*[1], mostrando un crecimiento anual que se asemeja a una curva exponencial. Evidentemente, puesto que cada sitio web suele contener varias páginas, el número de páginas web indexadas[2] es mayor. A fecha de mayo de 2015 existen al menos 4,52 mil millones de páginas web indexadas y la cifra es mucho mayor si se consideran aquellas páginas que no están recogidas por los buscadores.

Pensemos por un momento en el trabajo que realiza un motor de búsqueda como Google o Bing. Estas herramientas deben examinar todas las páginas web existentes, construir (o actualizar) un índice con las palabras que aparecen en cada una de estas páginas y además asignar un valor de relevancia a cada una de ellas, para decidir así el orden en el que se mostrarán a los usuarios que realicen una búsqueda. Esta tarea la realiza un componente que se llama *web crawler* o *web spider*, cuyo funcionamiento básico consiste en comenzar con un conjunto pequeño de páginas web y a continuación ir siguiendo enlaces (hipervínculos) a otras páginas, de tal manera que las nuevas páginas se añaden

[1] Internet Live Stats. *Total Number of Websites*. http://www.internetlivestats.com/total-number-of-websites/. [Online; consultado el 8 de mayo de 2015]

[2] WorldWideWebSize.com. *The Size of the World Wide Web*. http://www.worldwidewebsize.com. [Online; consultado el 8 de mayo de 2015]

Figura 1 – Ejemplo de un log de acceso HTTP

```
127.0.0.1 - - [08/May/2015:11:33:21 +0100] "GET /index HTTP/1.1" 200 1453
10.0.0.1 - john [08/May/2015:11:35:12 +0100] "GET /private HTTP/1.1" 200 326
```

al conjunto inicial, haciéndolo crecer, al tiempo que evita seguir enlaces a contenido duplicado. A continuación, se construye un índice que relaciona cada página web con los términos que en ella aparecen, de tal modo que se puedan realizar las búsquedas de un modo mucho más eficiente. Por último, los buscadores suelen calcular para cada página un valor de «importancia» (en el caso de Google se llama PageRank[3]), que permite ordenar los resultados de búsqueda en función de la relevancia de un sitio con respecto a otros. Normalmente, la relevancia de un sitio web viene definido por la cantidad de enlaces de otros sitios que apuntan a él, pero además, también depende de la propia relevancia de estos.

Si la cantidad de datos que implica la existencia de estos sitios web es ya de por sí impresionante, no debemos olvidar que cada vez que un usuario de Internet visita una página web se generan más datos. En la mayoría de los casos, los sitios web almacenan al menos dos registros o *logs*: uno de accesos y otro de errores. El primero registrará información relativa al usuario que está visitando el sitio web, así como al recurso web (página, imagen, etc.) que ha solicitado. El segundo registrará entradas cuando se hayan producido errores o advertencias: páginas que no pueden cargar, recursos que no existen, etc.

La Figura 1 muestra un par de entradas de ejemplo en un registro de acceso HTTP (*HyperText Transfer Protocol*), para el servidor web Apache. Los datos contenidos en estos registros están separados por espacios y contienen respectivamente la IP (dirección de Internet) del usuario, el nombre de usuario que visita la web (en caso de que esté acreditado mediante contraseña, en caso contrario no se muestra), la fecha y la hora a la que se produce la petición web, una serie de valores que indican la petición realizada, el recurso accedido y la versión del protocolo HTTP empleado y por último, un código de respuesta (que es 200 si la petición es correcta, 404 si no se encuentra el recurso, etc.) y el tamaño del fichero devuelto.

Estos datos se generan en tiempo real, a medida que los usuarios acceden a un sitio web. Aunque los registros se pueden eliminar, es más frecuente comprimir los ficheros de *logs* más antiguos para poder almacenarlos. El hecho de almacenar esta información permite analizarla con posterioridad, detectando patrones de acceso al servidor web (ubicación de los usuarios, periodos de tiempo con más accesos, cantidad de errores registrados, etc.), si bien supone un coste de almacenamiento importante.

[3]Sergey Brin y Lawrence Page. «The Anatomy of a Large-Scale Hypertextual Web Search Engine». En: *Computer Networks and ISDN Systems*. 1998, págs. 107-117

1.1.3.2 Comercio electrónico

La posibilidad de comprar a través de Internet resulta muy atractiva para muchas personas, puesto que permite comparar productos y sus precios de forma sencilla, y adquirirlos de forma cómoda desde cualquier lugar en el que estemos, a través de nuestro ordenador o dispositivo móvil.

Tan interesante se muestra esta posibilidad que las transacciones de compra/venta online se remontan a antes incluso del nacimiento de Internet: a principio de los 70, estudiantes de Stanford y del MIT se pusieron en contacto a través de ARPANET (la red de ordenadores que nació en 1969 y posteriormente acabaría evolucionando en Internet) para concertar una venta de marihuana[4].

No obstante, es a mediados de los años 90 cuando los sitios web para la compra/venta de productos comienzan a afianzarse. En 1995 se fundan dos compañías que aún hoy mantienen su liderazgo mundial: eBay y Amazon. El número de empresas dedicadas al comercio electrónico continúa creciendo a finales de los años 90, debido a la mentalidad existente sobre la «nueva economía» y dando pie a la aparición de la burbuja *puntocom*, que alcanza su pico máximo en el año 2000.

En la actualidad, muchas empresas de comercio electrónico ven crecer su facturación año tras año, lo que muestra un interés creciente por parte de la población en utilizar este tipo de servicios. Por ejemplo, los ingresos brutos de Amazon.com han pasado de 7,64 mil millones de dólares en 2010 hasta 25,11 mil millones de dólares en 2014, mostrando una tendencia de crecimiento anual lineal [5]; mientras que los de eBay han pasado de 6,39 mil millones a 11,94 mil millones en el mismo periodo, mostrando una tendencia similar pero con una pendiente menor[6].

No solo se sigue extendiendo la práctica de comprar a través de Internet, sino que cada vez es más frecuente hacerlo a través de dispositivos móviles. Como muestra de ello, en la campaña de Navidad de 2014 el 60 % de las ventas de Amazon se realizaron a través de dispositivos móviles[7], como *smartphones* y *tablets*.

Pero, ¿qué tiene que ver el afianzamiento de esta práctica con la generación de Big Data? Debemos tener en cuenta que cada vez que un usuario realiza una compra a través de

[4]Mike Power. *Online Highs Are Old as the Net: the First e-Commerce Was a Drugs Deal*. Ed. por The Guardian. http://www.theguardian.com/science/2013/apr/19/online-high-net-drugs-deal. [Online; publicado el 19 de abril de 2013]

[5]MarketWatch. *AMZN Annual Income Statement*. http://www.marketwatch.com/investing/stock/amzn/financials. [Online; consultado el 9 de mayo de 2015]

[6]MarketWatch. *EBAY Annual Income Statement*. http://www.marketwatch.com/investing/stock/ebay/financials. [Online; consultado el 9 de mayo de 2015]

[7]Darrell Etherington. *Amazon's 2014 Holiday Sees Mobile Shopping Approach 60 % Of Total Volume*. Ed. por TechCrunch. http://techcrunch.com/2014/12/26/amazon-2014-holiday-sales/. [Online; publicado el 26 de diciembre de 2014]

Internet, se produce una transacción que se registra, indicando los datos del usuario, los productos adquiridos, su coste, la forma de pago, la dirección de envío, etc. Evidentemente, esto da lugar a un importante volumen de datos, pero la cosa no queda ahí.

Muchas empresas almacenan datos no solo cuando los clientes adquieren un producto, sino cuando visitan su página, registrando así toda la lista de productos visitados por un usuario. De esta forma, compañías como Amazon son capaces de recomendar productos a sus usuarios basándose en su afinidad con otros productos visitados recientemente, lo que podría indicar un interés por parte del usuario en adquirir próximamente algún producto de un cierto tipo.

Además, la aparición de herramientas para analíticas de comportamiento, como Google Analytics, permiten registrar casi todas las interacciones que realiza un usuario con un sitio web: qué páginas sigue, cuánto tiempo invierte en cada una de ellas, cuándo efectúa una compra o una suscripción a un boletín, etc. Estos sistemas están generando grandes cantidades de datos a una gran velocidad, a medida que los usuarios navegan por los sitios web e interactúan con ellos.

1.1.3.3 Redes sociales

A mediados de los años 2000 se comienza a producir un cambio de paradigma en lo que se refiere al uso de Internet por parte de sus usuarios. Hasta el momento, había existido una marcada diferenciación entre los productores de contenidos (desarrolladores de sitios web que introducían información, noticias, productos, etc.) y los consumidores que accedían a esta información. Este cambio de paradigma supone la evolución de la Web 1.0 a la Web 2.0, donde los usuarios también toman el rol de creadores de contenidos, lo que hace que la diferencia entre productores y consumidores deje de estar tan clara.

Un factor clave en este cambio de paradigma es la aparición de las redes sociales online, que comienzan a surgir a mediados de la década de los 2000, siendo Facebook (nacido en 2004) y Twitter (en 2006) algunos de los ejemplos más destacables por su gran volumen de usuarios, que constituyen un porcentaje importante de la población mundial.

En esencia, una red social hace referencia a un grupo de usuarios que se relacionan entre sí. Las redes sociales suelen permitir a sus usuarios establecer vínculos de amistad entre ellos, compartir información, fotografías, vídeos, etc.

La Figura 2 muestra el primer tuit (micropublicación en Twitter) escrito por uno de los fundadores de esta red social en 2006. Desde entonces, Twitter ha alcanzado la cifra de 320 millones de usuarios activos, que publican aproximadamente unos 500 millones de tuits al día[8]. Estas publicaciones pueden contener imágenes, vídeos, menciones a otros

[8]Twitter. *Sobre Twitter, Inc.* https://about.twitter.com/es/company. [Online; consultado el 10 de mayo de 2015]

Figura 2 – Primer tuit de la historia, escrito por un fundador de Twitter en 2006

usuarios o *hashtags*, que son como palabras clave dentro del texto, reconocibles porque comienzan con el símbolo '#'.

No obstante, la red social más popular es Facebook, que además de contar en la actualidad con algo más de mil millones de usuarios activos diarios de media[9], es la segunda página más visitada en Internet, solo después de Google[10]. Otra red social a destacar por su volumen de tráfico es YouTube, en la que se suben más de 300 horas de vídeo por minuto y se visualizan cientos de millones de horas de vídeo diariamente, lo que corresponde a miles de millones de reproducciones[11].

Estas estadísticas evidencian que las redes sociales constituyen uno de los principales orígenes de Big Data en la actualidad. No solo es relevante el volumen de los datos generados, sino la velocidad a la que se generan y su variedad, ya que estos contenidos pueden ser textos, imágenes, vídeos, etc.

1.1.3.4 Internet de las cosas

Una fuente de datos más reciente y con más posibilidades aún que las anteriores, es aquella que surge a partir del Internet de las cosas (IoT, *Internet of Things*). Lo que persigue esta nueva filosofía es que la mayoría de los objetos de uso cotidiano estén conectados a Internet. Según Gartner[12], en 2020 habrá 26 mil millones de objetos conectados a Internet, por lo que la cantidad de datos que pueden originar es gigantesca.

La implantación del Internet de las cosas supondría un paso importante para lograr el desarrollo de los entornos inteligentes y alcanzar el concepto de computación ubicua

[9]Facebook. *Company Info | Facebook Newsroom*. http://newsroom.fb.com/company-info/. [Online; consultado el 10 de mayo de 2015]

[10]Alexa. *Alexa Top 500 Global Sites*. http://www.alexa.com/topsites. [Online; consultado el 10 de mayo de 2015]

[11]Youtube. *Estadísticas - YouTube*. https://www.youtube.com/yt/press/es/statistics.html. [Online; consultado el 10 de mayo de 2015]

[12]Janessa Rivera y Rob van der Meulen. *Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020*. Ed. por Gartner. http://www.gartner.com/newsroom/id/2636073. [Online; publicado el 12 de diciembre de 2013]

que describió Mark Weiser a principio de los años noventa[13]. Según este paradigma, los objetos del mundo físico estarían conectados, existiendo multitud de sensores capaces de obtener información del entorno y de actuadores capaces de modificar el propio entorno.

Un ejemplo emergente de este tipo de entornos es el de las ciudades inteligentes, que buscan realizar un desarrollo de la ciudad sostenible y mejorar la calidad de vida en una ciudad por medio de la tecnología. De este modo, se podrían incluir sensores en la ciudad que permitieran detectar necesidades específicas o tomar medidas para mejorar el estado de la ciudad. En este sentido se puede pensar en sensores de temperatura o de emisiones de CO2, sensores de aparcamiento, etc. En algunos casos, si los valores detectados no son los adecuados, se podrían tomar medidas automáticas o alertar de forma autónoma a las autoridades pertinentes. Así, si la contaminación por emisiones de dióxido de carbono es muy elevada en una zona, se podría restringir el tráfico de automóviles en esa zona hasta que la situación se revirtiese.

Acercándonos a la realidad

John Cohn, uno de los principales impulsores del Internet de las cosas ha afirmado: «En los próximos años los objetos serán inteligentes por estar conectados»[14].

- ***¿Cree justificada su afirmación? ¿En qué sentido Internet de las cosas puede contribuir a mejorar temas como la salud o el tráfico?***

1.2 Fundamentos del Trabajo con Datos

A lo largo de esta sección, se esbozarán los conocimientos y conceptos fundamentales del trabajo con datos, para a continuación, avanzar en la obtención de conocimiento y valor a partir de los mismos.

1.2.1 ¿Qué es un Dato?

Si vamos a trabajar con datos, parece lógico pensar que lo primero que deberíamos saber es a qué nos referimos cuando hablamos de datos.

Si echamos un vistazo al diccionario, nos dice que dato es el:

> *Antecedente necesario para llegar al conocimiento exacto de algo o para deducir las consecuencias legítimas de un hecho.*

[13] Mark Weiser. «The Computer for the 21st Century». En: *Scientific American* 265 (1991), págs. 94-104

[14] Joel Dalmau. *John Cohn, 'Yoda' de Internet de las cosas: "En los próximos años los objetos serán inteligentes por estar conectados"*. Ed. por El País. http://one.elpais.com/john-cohn-yoda-de-internet-de-las-cosas-en-los-proximos-anos-los-objetos-seran-inteligentes-por-estar-conectados/. [Online; publicado el 30 de septiembre de 2015]

Ya en la propia definición de la palabra, el diccionario nos indica que el dato, no es un valor en sí mismo, sino el camino o el medio, bien para llegar a un conocimiento o bien para iniciar un proceso deductivo.

El dato es una representación simbólica (numérica, alfabética, algorítmica, espacial, etc.) de un atributo o variable cuantitativa o cualitativa.

La primera propiedad de los datos es que se utilizan para describir, aquellos hechos, sucesos o entidades que están relacionados con el propósito de nuestro estudio. A este conjunto de datos objeto de nuestro estudio es a lo que en estadística se le denomina la población o el corpus de datos.

Algunos ejemplos ilustrativos, podrían ser: datos acerca de la actividad que tienen los usuarios de una página web, de las transacciones de compra de los clientes de un sitio de comercio electrónico, los historiales médicos de los pacientes de un hospital, las imágenes tomadas en tiempo real de los pasajeros que pasan el control de seguridad de un aeropuerto, los datos proporcionados por la web de un ayuntamiento local acerca de los niveles de polen en el ambiente o el resultado que nos proporciona un secuenciador de ADN tras el análisis de una muestra.

Lo siguiente que nos dice la definición de dato recogida en el diccionario, es que un dato es la representación simbólica, (es decir el valor que toma) de un atributo o variable. Las variables o los atributos, son las características que describen los hechos, sucesos o entidades.

Esas características pueden ser de dos tipos:

- Cualitativas: Aquellas propiedades que no son cuantificables. Siguiendo con algunos de los ejemplos citados anteriormente, podría ser la temática del sitio web del que procedía un visitante de nuestro sitio, las preferencias de compra de nuestro cliente, el sexo de los pacientes del hospital, la complexión de los pasajeros del aeropuerto, los diferentes barrios de la ciudad o el tipo de nucleótido en el ejemplo de la secuenciación de ADN.

- Cuantitativas: Son aquellas que si son cuantificables. Algunos ejemplos pueden ser: el número de veces que visita nuestra web un usuario al mes, el importe de la última compra que realizó un cliente, la presión sanguínea de los pacientes del hospital, la altura de los pasajeros, la concentración en suspensión de un determinado alérgeno o el número de apariciones de un nucleótido.

Cuando trabajamos con datos, podemos distinguir entre dos tipos de datos dependiendo del tratamiento al que hayan sido sometidos:

- **Datos en bruto (*raw data*)**, también conocidos como datos primarios, son aquellos datos recogidos directamente de una fuente de datos, independientemente de que esta sea manual o electrónica.
    - Estos datos no han sido objeto de ningún tipo de procesamiento o manipulación.
    - Generalmente no pueden ser utilizados directamente para su análisis.
    - Estos datos pueden contener errores, no están validados, presentar diferentes formatos o no tener formato. Un ejemplo muy típico de esto son las fechas que pueden venir en diferentes formatos: «31 de enero 1999», «31/01/1999», «31/1/99», «31 de enero», o «hoy».
    - Aunque los datos en bruto tiene el potencial de convertirse en «información», para que esto ocurra, es necesario realizar procesos de extracción, organización, y en ocasiones, análisis y formateo de la información para su presentación.
- **Datos procesados**. Son el resultado de uno o varios procesos de transformación de los datos en bruto. Respecto a los datos procesados cabe señalar que:
    - Los procesos de manipulación a los que son sometidos pueden incluir entre otros, el fusionado de varios datos, el agrupamiento, la transformación, la normalización, etc.
    - Los procesos de transformación suelen estar estandarizados.
    - Están listos para ser utilizados en los procesos de análisis.
    - Todos los pasos seguidos para la transformación de los datos deben ser registrados.

Un terminal de punto de venta (TPV) en un supermercado con cierta actividad, recoge grandes volúmenes de datos en bruto de cada día, sin embargo, estos datos tal y como son proporcionados por la máquina no nos darán mucha información hasta que los procesemos. Una vez procesados, los datos pueden indicar los elementos particulares que cada cliente compra, cuando los compran y a qué precio. Tal información podría convertirse en datos para el procesamiento de por ejemplo, campañas de marketing predictivo.

Como resultado del procesamiento, los datos a veces terminan en una base de datos, que permite que estos datos en bruto sean accesibles para su posterior procesamiento y análisis de diferentes maneras.

Pero entonces ¿qué aspecto tienen los datos?

En el trabajo con datos se podría pensar que estos van a venir empaquetados en formato tabular, como por ejemplo en grandes ficheros del estilo de las hojas de cálculo, en el que los datos están clasificados por columnas y filas, cada columna se correspondería

con un atributo o variable y cada fila con un valor para la misma a modo de registro. Sin embargo, la realidad es completamente distinta y lo extraño es recibir los datos en este formato. Lo normal en el trabajo diario es que los datos se nos entreguen en un formato en bruto, en muchos casos puede ser la salida directa de una máquina, en otros un *log* (registro) de una aplicación, ...

En este sentido, podemos categorizar los datos en dos grandes grupos, dependiendo de la estructura (o la falta de estructura) que estos tengan:

- **Datos estructurados**: Son aquellos datos que tienen una estructura o esquema definido y fijo, es decir se repiten las mismas estructuras o reglas en todo el conjunto de datos. Precisamente por esta razón, los datos estructurados son fácilmente «entendibles» por las máquinas, lo que facilita su procesamiento, tratamiento, almacenamiento y análisis, por parte de los ordenadores. Estos datos son almacenados en bases de datos relacionales, de forma que se pueden establecer interconexiones o relaciones entre los datos (que están guardados en tablas), y a través de dichas conexiones relacionar los datos de ambas tablas.

  Los datos estructurados residen en campos fijos como parte de un registro o fichero. Esto incluye los datos almacenados en bases de datos y hojas de cálculo. Los datos estructurados dependen de la creación de un modelo de datos, es decir, un modelo de los tipos de datos de los diferentes campos y cómo éstos son almacenados, procesados y accedidos. Esto incluye cómo estos datos son almacenados: tipos de datos (numérico, moneda, alfabético, nombre, fecha, dirección, etc.) y cualquier restricción sobre los datos que se pueden almacenar (número de caracteres, patrón del texto almacenados, rango de valores numéricos permitidos, etc.). La ventaja de los datos estructurados es que son fácilmente introducidos, almacenados, consultados y analizados. Debido a las limitaciones que existía en el hardware que procesaba los datos las únicas alternativas viables eran las bases de datos y las hojas de cálculos. Sin embargo con el abaratamiento de los costes de almacenamiento en disco y de la memoria para el procesamiento de los datos se están dando alternativas a los datos estructurados cuando estos modelos no se ajustan al problema que necesitamos resolver.

- **Datos no estructurados**: En contraposición a los datos estructurados, los datos no estructurados son aquellos que carecen de una estructura definida. Los datos no estructurados son fácilmente entendibles o interpretables por el ser humano, sin embargo su procesamiento para su posterior análisis por parte de las máquinas resulta muy complicado y requiere del uso de algoritmos complejos diseñados para tal fin. La visualización de una fotografía, la audición de un fichero de audio o la lectura de un texto por parte de una persona son suficientes para extraer los datos que contienen y comprender la información que

proporcionan, sin embargo procesar los datos contenidos en estos archivos por parte de una máquina, es una tarea muy compleja. El almacenamiento de este tipo de datos, requiere de bases de datos más complejas que las tradicionales, esto es de bases de datos relacionales. Ejemplos de datos no estructurados son: archivos en formato binario como fotografías, archivos de audio o ficheros de texto como los archivos en formato PDF.

- **Datos semiestructurados**. En una posición intermedia entre las dos grandes categorías de datos que se acaban de presentar, los datos estructurados y los no estructurados, están los datos semiestructurados. Los datos semiestructurados son aquellos que a pesar de no tener estructuras fijas, contienen ciertos elementos (etiquetas o marcadores) estructurados. Los datos semiestructurados ofrecen una alternativa para aquellos datos que no conforman una estructura que encaje en un modelo de datos estructurado. Se basan normalmente en el uso de etiquetas para identificar estos elementos, por lo que también se conocen como estructuras auto-descriptivas. Un ejemplo típico de este tipo de datos son los datos procedentes de la web, ya que contienen las típicas etiquetas HTML, XML o JSON.

Veamos unos ejemplos:

En la Figura 3 se muestra un extracto del fichero de salida de un secuenciador de proteínas o uno de ADN. Como podemos ver, la salida es un fichero de texto en el que, además de otra información, encontramos una secuencia de letras (la parte marcada en gris) que se corresponde con los nucleótidos que lo forman.

En otras ocasiones, los datos no proceden de máquinas, sino que son el resultado del consumo de la información suministrada a través de una API proporcionada por un tercero. En la Figura 4 se recoge el resultado de una petición a la API de Twitter. Como vemos, la información está estructurada, pero desde luego no está en forma de filas y columnas.

En otras ocasiones podemos encontrarnos con ficheros de texto (Figura 5), por ejemplo la ficha del historial médico de un paciente. Los datos no tienen por qué estar siempre en formato de texto o numérico.

Figura 3 – Extracto del fichero de salida de un secuenciador de proteínas

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAG
ACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDD
```

Figura 4 – Resultado de una petición a la API de Twitter

```
GET  https://api.twitter.com/1.1/blocks/list.json?skip_status=true&cursor=-1

{
  "previous_cursor": 0,
  "previous_cursor_str": "0",
  "next_cursor": 0,
  "users": [
    {
      "profile_sidebar_fill_color": "DDEEF6",
      "profile_background_tile": false,
      "profile_sidebar_border_color": "C0DEED",
      "name": "Javier Heady \r",
      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
      "profile_image_url":
"http://a0.twimg.com/sticky/default_profile_images/default_profile_4_normal.pr
      "location": "",
      "is_translator": false,
      "follow_request_sent": false,
      "profile_link_color": "0084B4",
      "id_str": "509466276",
      "entities": {
        "description": {
          "urls": [
...
```

Figura 5 – Ejemplo de datos en formato de texto

```
********** ALLERGIES **********

Last Updated: 01 Dec 2011 @ 0851


Allergy Name:         TRIMETHOPRIM
Location:             DAYT29
Date Entered:         09 Mar 2011
Reaction:
Allergy Type:         DRUG
VA Drug Class:        ANTI-INFECTIVES,OTHER
Observed/Historical:  HISTORICAL
Comments:             The reaction to this allergy was MILD (NO
SQUELAE)

Allergy Name:         TRAMADOL
Location:             DAYT29
Date Entered:         09 Mar 2011
Reaction:             URINARY RETENTION
Allergy Type:         DRUG
VA Drug Class:        NON-OPIOID ANALGESICS
Observed/Historical:  HISTORICAL
Comments:             gradually worsening difficulty emptying bladder
try tramadol again cautiously because pt. reported pain relief

********** MEDICATION HISTORY **********

Last Updated: 11 Apr 2011 @ 1737


Medication: AMLODIPINE BESYLATE 10MG TAB
Instructions: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR 1
GRAPEFRUIT JUICE--
Status: Active
Refills Remaining: 3
Last Filled On: 20 Aug 2010
Initially Ordered On: 13 Aug 2010
Quantity: 45
Days Supply: 90
Pharmacy: DAYTON
Prescription Number: 2718953

Medication: IBUPROFEN 600MG TAB
Instructions: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOOD
Status: Active
Refills Remaining: 3
Last Filled On: 20 Aug 2010
Initially Ordered On: 01 Jul 2010
Quantity: 360
```

A veces los datos están en formato binario. Si recordamos el ejemplo que comentamos del control de seguridad de un aeropuerto, tendremos que procesar imágenes de las caras de los pasajeros tomadas por las cámaras de seguridad.

Resulta, por lo tanto, que el formato de los datos con los que vamos a tener que trabajar es muy variado. Viendo estos pocos ejemplos, nos podemos hacer una idea de la variedad y complejidad con la que nos encontramos en el mundo real y de lo laboriosos que serán los procesos de preparación de los datos, es decir los procesos de convertir los datos en bruto en datos procesados listos para el análisis.

A menudo se dice que el 80 % del tiempo que se dedica al análisis de los datos se consume en el proceso de limpieza y preparación de los mismos[15]. Además, el proceso de preparación de datos no es un proceso que se haga una vez, deberemos repetirlo en diferentes ocasiones, a medida que vayan saliendo a la luz nuevos problemas o se recopilen nuevos datos.

Y, ¿cómo son los datos procesados?

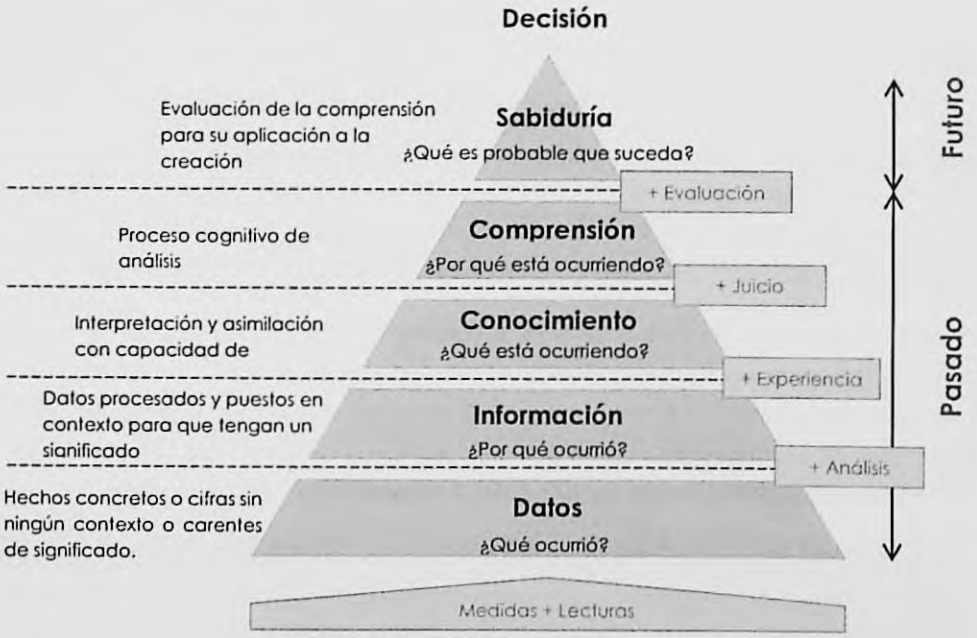
Una vez que realizamos el proceso de limpieza y preparación de los datos en bruto, estos son normalmente almacenados en bases de datos. Esto significa que estarán en un formato tabular, similar a una hoja de cálculo de forma que cada columna será una variable y cada fila contendrá los valores que toma cada variable.

1.2.2 Descubrimiento de Conocimiento en Bases de Datos (KDD)

El verdadero valor del trabajo con datos, independientemente del tipo y volumen de datos con los que trabajemos, no está en la capacidad de procesarlos y almacenarlos, sino en la habilidad de obtener conocimiento a partir de los mismos. Los datos no dejan de ser la materia prima que utilizaremos para convertir estos en valor.

Se suele representar este hecho en lo que se ha dado en llamar la «Jerarquía del Conocimiento», también conocida como «Jerarquía DIKW» (*Data Information Knowledge Wisdom Chain*), o «Pirámide del Conocimiento» (Figura 6).

Figura 6 – Pirámide del conocimiento



[15] Tamraparni Dasu y Theodore Johnson. *Exploratory Data Mining and Data Cleaning*. Wiley, 2003

De las diferentes fuentes de información, recolectamos y almacenamos **datos**. Estos datos, son una representación simbólica de los hechos acontecidos en el pasado. Nos dan una medida de los mismos, nos dicen qué ocurrió, pero de forma descontextualizada, de manera que no tienen mucho valor en sí mismos.

En esos datos, existe una gran cantidad de información oculta a la que no se puede acceder a través de la simple observación de los mismos o mediante las técnicas clásicas de recuperación de la información. El descubrimiento de esta información oculta, sólo es posible tras aplicar mecanismos y procesos de análisis de los datos. Esta tarea se conoce como Minería de Datos (*Data Mining*).

Tras el análisis de los datos, transformamos estos en **información**. La información nos explica por qué ocurrieron los hechos que representaban los datos. El procesamiento de los datos permite contextualizar su significado.

La experiencia en una determinada materia facilita la interpretación de la información su conversión en **conocimiento**, permitiéndonos averiguar qué está ocurriendo.

La valoración y reflexión sobre el conocimiento, nos permite averiguar por qué está ocurriendo lo que está ocurriendo. Hablamos entonces de la **comprensión** del conocimiento.

Hasta este momento, hemos estado trabajando e interpretando hechos pasados. La evaluación de la comprensión y su aplicación, nos permite identificar o inferir, qué es probable que ocurra en el futuro próximo, lo que dentro de la pirámide del conocimiento se conoce como **sabiduría**. La última fase de este proceso y el fin último del mismo, es la toma de una **decisión**.

El proceso que permite transformar el dato en valor es, como se ha señalado, el Descubrimiento del Conocimiento en Bases de Datos o KDD (*Knowledge Discovery in Databases*).

El término *Knowledge Discovery in Databases*, fue acuñado por primera vez en 1989, para referirse al proceso general de búsqueda de conocimiento en los datos, enfatizando la aplicación de métodos de minería de datos de «alto nivel».

El KDD se define como la extracción, no trivial, de información previamente desconocida y potencialmente útil, en grandes colecciones de datos[16]. Puede considerarse como una búsqueda de reglas interesantes, patrones o excepciones en grandes colecciones de datos. Es un área interdisciplinaria en la que hay que utilizar conocimientos de: estadística, bases de datos, aprendizaje automático, inteligencia artificial, teoría de la información, computación paralela y distribuida y visualización, entre otros.

Dentro del KDD, las técnicas más frecuentemente utilizadas pueden catalogarse en:

[16]Usama Fayyad, Gregory Piatetsky-Shapiro y Padhraic Smyth. «From Data Mining to Knowledge Discovery in Databases». En: *AI Magazine* 17.3 (1996), págs. 37-54

- **Descriptivas**: El objetivo de estos procedimientos es la búsqueda de la caracterización o discriminación de un conjunto de datos. Las técnicas más conocidas son: agrupamiento o *clustering*, reglas de asociación, análisis de patrones secuenciales, análisis de componentes principales y detección de desviación.

- **Predictivas**: El propósito de estos métodos es aprender una hipótesis que pueda clasificar a nuevos individuos. Los algoritmos principales son los regresión y los de clasificación (árboles de decisión, clasificación bayesiana, redes neuronales, algoritmos genéticos, conjuntos y lógica difusa).

1.2.2.1 Data mining (minería de datos)

Aunque el término *Data Mining* o DM, es una etapa dentro del proceso de descubrimiento de conocimiento en bases de datos; en el mundo empresarial y en la literatura científica, es común que ambos términos se utilicen indistintamente.

En la bibliografía nos encontramos con diferentes definiciones; una definición generalmente aceptada podría ser:

> *Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos.*[16]

En el mundo empresarial nos encontramos con definiciones como:

> *La integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión.*[17]

Aunque desde fechas recientes es un término muy en boga, por su asociación al concepto de Big Data, la idea del *data mining* no es nada nueva. En los años sesenta los profesionales de la estadística manejaban términos como «*Data Fishing*», «*Data Mining*» o «*Data Archaeology*» para referirse a la idea de encontrar correlaciones, sin partir de hipótesis previas, en bases de datos que pueden contener datos que no han sido tratados y por tanto pueden contener «ruido».

El término comenzó a consolidarse en la década de los ochenta cuando empezaron a publicarse estudios acerca de KDD y *Data Mining*[18].

---

[17] Luis Carlos Molina y S. Ribeiro. «Descubrimiento Conocimiento para el Mejoramiento Bovino Usando Técnicas de Data Mining». En: *Actas del IV Congreso Catalán de Inteligencia Artificial*. 2001, págs. 123-130

[18] Usama Fayyad y col., eds. *Advances in Knowledge and Data Mining*. MIT Press, 1996

1.2.3 ¿Qué es la Inteligencia de Negocio (Business Intelligence)?

Desde la creación allá por los años 80 de los *Data Warehouse* o Almacenes de Datos (véase la sección 1.2.6), el volumen de datos procesados y el nivel de detalle almacenado de los mismos, han ido creciendo a pasos agigantados. A partir de ese momento, surge la necesidad de controlar la información de cualquier cambio en las organizaciones evitando los inconvenientes de los sistemas de gestión de datos tradicionales y obsoletos, ya que día a día, aumentan los datos operacionales asociados a dichos sistemas y consecuentemente los datos que se pueden analizar y almacenar de forma organizada en estas inmensas bases de datos.

Hoy vivimos en lo que se ha dado en llamar «La Sociedad de la Información», donde cada vez tenemos acceso a una mayor cantidad de información, gracias a Internet y al desarrollo creciente de la tecnología y de los sistemas de información. En este contexto, las organizaciones de cualquier índole, las empresas y sus directivos pueden acceder a un número cada vez mayor de fuentes de datos que proveen de más información, que además cada vez es de mejor calidad y a la que se puede acceder de forma cada vez más rápida.

El potencial que todo esto ofrece para mejorar el conocimiento en general y la toma de decisiones, en particular, guiando a las empresas hacia la consecución de sus objetivos es enorme. Sin embargo, toda esta información se convierte en conocimiento útil, sólo cuando se analiza y se estructura de forma inteligente. Esta tarea, no exenta de complejidades técnicas, requiere que la información sea procesada y analizada en tiempos cada vez más cortos. Tener cada vez más datos y no ser capaz de analizarlos a tiempo, convierte estos en información inútil. En la actualidad, la globalización, la creciente internacionalización de los mercados, y la consiguiente intensificación del entorno competitivo, convierten este hecho en un auténtico reto de gestión.

La capacidad para tomar decisiones con rapidez, basadas en un adecuado conocimiento de la realidad de la empresa así como del mercado y sus tendencias, ha pasado a convertirse en una nueva fuente de ventaja competitiva. Poseer un conocimiento proveniente de información comprensible, detallada, relevante y útil es vital para lograr y sostener una ventaja competitiva en el entorno empresarial. Para sacar partido a los datos y convertirlos en información, y ésta a su vez en conocimiento, se necesitan distintas técnicas y procesos. A todos estos procesos de tratamiento de datos se les atribuye el término de *Business Intelligence* (BI).

A lo largo del capítulo utilizaremos el término anglosajón *Business Intelligence* por considerar que es un término ampliamente extendido y utilizado en el mundo empresarial. Aunque hay diferentes traducciones para este término, la más común es la de «Inteligencia de Negocio». El objetivo básico de la *Business Intelligence* es apoyar de forma sostenible y continuada a las organizaciones para mejorar su competitividad, facilitando la información necesaria para la toma de decisiones. El primero que acuñó el término fue

Howard Dresner que, cuando era consultor de Gartner, popularizó el término *Business Intelligence* o BI como un término paraguas para describir un conjunto de conceptos y métodos que mejoran la toma de decisiones, utilizando información sobre qué hechos habían sucedido. Mediante el uso de tecnologías y las metodologías de *Business Intelligence* pretendemos convertir datos en información y a partir de la información ser capaces de descubrir conocimiento.

El origen de la *Business Intelligence* va ligado a proveer acceso directo a la información a los usuarios de negocio para ayudarles en la toma de decisiones, sin intervención de los departamentos de Sistemas de Información.

Una amplia definición es la que proponen en The Datawarehouse Institute[19]:

> *Business Intelligence (BI) es un término paraguas que abarca los procesos, las herramientas, y las tecnologías para convertir datos en información, información en conocimiento y planes para conducir de forma eficaz las actividades de los negocios. BI abarca las tecnologías de datawarehousing los procesos en el 'back end', consultas, informes, análisis y las herramientas para mostrar información (estas son las herramientas de BI) y los procesos en el 'front end'.*

1.2.4 Necesidad Creciente de Información Estratégica

La información estratégica es necesaria para los ejecutivos y directivos de las empresas cuya responsabilidad es mantener o aumentar la competitividad de la empresa. Esta información les servirá para tomar las decisiones necesarias para este objetivo. Necesitarán por tanto información para plantear la estrategia de negocio, establecer objetivos y hacer un seguimiento de los resultados de las decisiones tomadas. Veamos algunos ejemplos de **objetivos empresariales**.

- Incrementar el número de clientes en un porcentaje de al menos un 10 % en los próximos 3 años.
- Incrementar las ventas en el próximo trimestre.
- Mejorar el servicio de atención al cliente para reducir la pérdida de clientes en un 10 % en los próximos 2 años.
- Comenzar nuevas líneas de productos en el próximo año.
- Corregir la pérdida de ventas de las sedes europeas de la empresa.

[19] TWDI. *TDWI | Advancing all things data. | Business Intelligence, Data Warehousing, Analytics | Education & Research*. http://www.tdwi.org. [Online; consultado el 2 de diciembre de 2015]

La **toma de decisiones** para alcanzar estos objetivos requiere poner a disposición de los ejecutivos mecanismos para conocer el funcionamiento de la empresa, revisar y monitorizar indicadores de rendimiento (KPI, *Key Performance Indicators*) y como unos modifican a otros, trazar los cambios en los aspectos clave del negocio a lo largo del tiempo, comparar el rendimiento de la empresa respecto a la competencia, etc. Estos ejecutivos deberán centrarse en el análisis de las necesidades de los clientes y sus preferencias, los resultados de ventas y de campañas publicitarias, la calidad de los servicios y productos, etc. Esta información debe además ser ofrecida de forma amplia abarcando la empresa al completo sin limitarse a información orientada a un departamento específico. De esta forma podrán tomarse decisiones estratégicas para llegar a los objetivos empresariales de la empresa. Esta información no tiene que dar soporte a las operaciones diarias. Es decir, esta información no debe generar recibos, envíos, reclamaciones o servir para realizar operaciones bancarias. Esta información estratégica debe ser de utilidad para el funcionamiento continuo y la supervivencia de la empresa a medio-largo plazo.

Esta información debe tener las siguientes características:

- **Integrada**. La información estratégica debe ofrecer una perspectiva global del negocio.
- **Correcta**. Esta información debe ser precisa y descrita conforme a las reglas del negocio.
- **Accesible**. La accesibilidad de esta información y la facilidad de uso y generación de nuevos análisis es de vital importancia para su utilidad estratégica.
- **Fiable**. Esta información debe estar libre de errores que puedan llevar a la toma de decisiones incorrectas.
- **Proporcionada a tiempo**. Debe proporcionarse en el momento justo para su uso en la toma de decisiones.

1.2.4.1 La crisis de la información

En las empresas de tamaño medio-grande se dispone normalmente de estructuras para la gestión de la información que generan. Generalmente, además se cuenta con grandes cantidades de datos acumulados a lo largo de los años de funcionamiento de la empresa y que frecuentemente no están centralizados en un único almacén de datos, lo que hace que su uso efectivo para la toma de decisiones sea complejo. Este problema se incrementa constantemente por el hecho de que las empresas que tienen actividad no paran de generar información. Se estima que la cantidad de información que manejan las empresas se duplica cada 18 meses.

¿Dónde está por tanto la crisis de información? Precisamente la crisis viene del problema de tener grandes cantidades de datos, pero su naturaleza y forma de gestión tradicional

en el contexto de las operaciones diarias de la empresa hacen que **extraer información de estos datos sea un proceso costoso** que no todas las empresas pueden abordar apropiadamente.

Los datos necesarios para la toma de decisiones estratégicas deben estar en formatos que puedan ser usados para su apropiado análisis destacando los elementos de interés a tener en cuenta en los diferentes aspectos del negocio. Los ejecutivos de la empresa necesitan ver estos aspectos del negocio a lo largo del tiempo para detectar anomalías, tendencias u oportunidades. Las **toneladas de datos operacionales** de la empresa no pueden ser usadas directamente para este tipo de análisis. Los datos operacionales se suelen derivar de eventos que producen cambios puntuales en los datos del sistema transaccional.

Para la toma de decisiones estratégicas los ejecutivos necesitan tener acceso a los datos desde distintos puntos de vista, pero siempre con una visión global de los diferentes aspectos del negocio. Por ejemplo, deben poder acceder a relaciones entre ventas y productos, almacenes y ventas, vendedores y su efectividad, etc.

1.2.5 Sistemas Operacionales frente a la Toma de Decisiones

Uno de los principales motivos del fracaso de las aproximaciones comentadas anteriormente lo podemos encontrar en un hecho común a todas ellas: el intento de obtener información estratégica de los sistemas operacionales de la empresa. Estos sistemas operacionales no han sido diseñados para este tipo de tareas y por ello no son ideales para producir sistemas de apoyo a la toma de decisiones. Por lo tanto, la información estratégica debe obtenerse de diversos sistemas de información de la empresa para ofrecer soporte a la toma de decisiones. Los sistemas operacionales son sistemas transaccionales (OLTP). Estos sistemas son los que usan las aplicaciones que dan soporte al funcionamiento diario de la empresa, son los que mantienen en movimiento los mecanismos empresariales y por tanto dan soporte a los distintos procesos empresariales. Estos sistemas almacenan los datos normalmente en bases de datos, de forma que los datos de cada transacción se almacenan en tablas de estas bases de datos como parte de transacciones SQL.

Por otro lado, los sistemas de apoyo a la toma de decisiones no están pensados para formar parte del proceso diario del negocio ni de los procesos de negocio. Estos sistemas deben estar diseñados para observar el negocio en funcionamiento, sin interferir de forma directa en los sistemas transaccionales que le dan soporte. La forma de que estos sistemas tengan un efecto en los procesos de negocio es a través de las decisiones que tomen los directivos gracias a la información proporcionada por los sistemas de apoyo a la toma de decisiones.

Los sistemas de apoyo a la toma de decisiones deben ser capaces de extraer información de las bases de datos, al contrario que los sistemas transaccionales que están más enfocados en insertar datos en las mismas.

En resumen, **para proporcionar información estratégica es necesario desarrollar sistemas informacionales** en contraposición a los sistemas operacionales que dan soporte al funcionamiento de los procesos de negocio. Las principales características que deben reunir los sistemas informacionales son las siguientes:

- Servir a diferentes propósitos.
- Permitir abordar diferentes aspectos del negocio.
- Incluir datos de naturaleza distinta a la de los sistemas transaccionales.
- Ofrecer mecanismos de acceso diferentes basados en el uso que se hará de los mismos (no orientado a transacciones puntuales).

1.2.6 Almacenes de Datos

Llegados a este punto, queda claro que necesitamos un sistema capaz de llevar a cabo análisis sobre los datos (no transaccionales) para descubrir tendencias, hacer un seguimiento del rendimiento de los procesos de negocio, etc. Este tipo de sistema, y las bases de datos que lo integran, debe tener ciertas características esenciales para ayudar a la toma de decisiones que se pueden sintetizar como sigue:

- Bases de datos orientadas a tareas analíticas.
- Datos provenientes de distintos sistemas o aplicaciones de la empresa (o externas a la misma).
- Fácil de usar y orientado a sesiones interactivas de análisis de datos.
- Intensivas en operaciones de lectura de datos.
- Interfaces de usuario orientadas a usuarios involucrados en la toma de decisiones sin la necesidad de interacción con personal TIC.
- Contenidos actualizados periódicamente y estables a lo largo del tiempo.
- Datos históricos y actuales.
- Capacidad de resolver consultas sin necesidad de desarrollo software ad-hoc adicional.
- Funciones que permitan a los usuarios iniciar el desarrollo de informes de forma autónoma.

La mayor parte de los procesos en este entorno serán de naturaleza analítica, lo que implica algunos requisitos funcionales:

- Capacidad de ejecutar consultas sobre datos actuales e históricos.
- Capacidad de ofrecer análisis de tipo «Y-SI».
- Capacidad de consultar, volver atrás, analizar y continuar procesando tantas veces como sea necesario en un proceso cíclico.
- Capacidad de mostrar tendencias históricas y aplicarlas en procesos de negocio posteriores.

Como ya introdujimos anteriormente, **este entorno analítico centrado en los datos que tanto necesitaban las empresas se conoce como los almacenes de datos**. Este nuevo entorno analítico se mantiene separado de los sistemas transaccionales para evitar su interferencia en los procesos de negocio.

En este capítulo introductorio podemos decir que un almacén de datos contiene métricas críticas para los procesos de negocio almacenados en términos de las diferentes dimensiones del negocio. Por ejemplo, un almacén de datos contendrá unidades como ventas, productos, días, clientes, distritos, regiones, promociones, etc. En este contexto las dimensiones de negocio podrían ser los productos, días, clientes, distritos, regiones y promociones. La unidad de ventas serán las métricas que podemos tomar para cada producto; día, cliente, distrito, región y promoción.

Los datos procedentes de los sistemas operacionales deberán ser la fuente de datos de la que se completen estas dimensiones y métricas para conformar el almacén de datos sobre el que se desarrollarán los sistemas de apoyo a la toma de decisiones.

Entre estos dos sistemas (el operacional y el almacén de datos) debería haber un área de almacenamiento intermedia en la que, de forma desacoplada, se extraigan los datos del sistema operacional, se realicen operaciones de filtrado y limpieza, se calculen valores agregados y finalmente se realice la carga de los datos en el almacén de datos.

Llegados a este punto podemos tratar de definir funcionalmente un almacén de datos como sigue:

- Proporcionan una visión completa del negocio.
- Proporcionan la información histórica y actual de forma sencilla.
- Proporcionan operaciones de apoyo a la toma de decisiones sin interferir en los sistemas operacionales.
- La información que representan es consistente.
- Presentan una fuente de información interactiva y flexible.

En definitiva el concepto de almacén de datos es un concepto simple, pero que ofrece una solución eficaz para un problema existente. Básicamente se basa en un cambio en el paradigma de almacenamiento de los datos para ofrecer una funcionalidad con unas características bien definidas. Estos sistemas dan soporte a los procesos decisionales en vez de a los procesos transaccionales de la empresa. Es por tanto un concepto simple: toma los datos de la empresa, los limpia y los transforma para ofrecer una visión de los mismos orientada a las necesidades de negocio. Para alcanzar este objetivo en un almacén de datos se realizan las siguientes tareas:

- Se toman todos los datos de los sistemas operacionales, incluyendo los datos históricos que quizás están almacenados en copias de seguridad.
- Cuando sea necesario se toman datos externos, como indicadores de mercado.
- Se integran los datos de las distintas fuentes (departamentales, externos, etc.).
- Se eliminan las inconsistencias y se transforman los datos para su gestión orientada al análisis de los mismos.
- Se almacenan los datos de forma que puedan usarse en las herramientas de apoyo a la toma de decisiones.

Los almacenes de datos al igual que los almacenes de mercancías deben estar orientados a ofrecer soporte a un amplio abanico de usos analíticos de los datos. Cuando sea necesario, estos datos pueden pasarse a sistemas especializados para ciertos tipos de usuario (*data marts*).

Acercándonos a la realidad

La idea defendida en esta sección, de que el análisis de datos puede ayudar a los gerentes de organizaciones a tomar decisiones que mejoren su rendimiento e impulsen su ventaja competitiva en el mercado es algo que ha sido ratificado por los resultados de una encuesta realizada por KPMG[20]. Según los resultados de la citada encuesta el 69 % de la alta dirección en el mundo considera que el concepto de procesamiento analítico de datos es importante para efectos estratégicos en sus planes de crecimiento.

- ***¿Cómo se explica que, sin embargo, en el mismo trabajo de KPMG se señale que hoy en día un 96 % cree que su compañía no está utilizando con eficiencia la gestión analítica de los datos?***

[20]KPMG. *Generan Ventaja Competitiva a través de Business Intelligence*. http://www.kpmg.com/mx/es/issuesandinsights/articlespublications/paginas/cp-generan-ventaja-competitiva-business-intelligence.aspx. [Online; publicado el 6 de enero de 2014]

1.3 El Ciclo de Vida de los Datos

1.3.1 Introducción

En la sección 1.2.1 ya se introdujo el concepto de dato como la representación simbólica (numérica, alfabética, algorítmica, espacial, etc.) de un atributo o variable cuantitativa o cualitativa, cuya principal característica es la de ser capaz de describir, aquellos hechos, sucesos o entidades que están relacionados con el propósito de nuestro estudio.

También se hizo una primera clasificación de los datos atendiendo al tratamiento al que estos hubiesen sido sometidos. Así distinguíamos entre datos sin ningún tipo de tratamiento, o datos en bruto (*raw data*), entendidos como aquellos que son almacenados tal y como vienen de la fuente de la que proceden; y datos procesados, que son aquellos que han sido sometidos a algún tratamiento previo antes de su almacenamiento. En esta última categoría, podríamos hacer una segunda clasificación, distinguiendo entre datos intermedios, aquellos que han sido previamente tratados y que posteriormente serán utilizados en otros procesos o que tras posteriores tratamientos serán modificados; y datos finales, entendidos como aquellos que son el resultado final de un análisis o estudio.

Sin embargo, es conveniente introducir una nueva categoría o estado de los datos: los datos obsoletos. Estos hacen referencia a aquellos datos que ya no tienen validez, algo que puede suceder en un determinado momento por múltiples razones: obsolescencia temporal o de otra índole, cambio de condiciones, cambios de criterios, etc.

En términos generales, el concepto «ciclo de vida» precisamente hace referencia a este hecho y se suele representar como una curva que marca un punto de inicio y un punto final con un lapso de existencia que sube y baja dentro de unos criterios de valoración. Dentro de cualquier ciclo de vida, hay distintas fases: crecimiento, madurez y declive. El ejemplo más conocido es probablemente el del ciclo de vida de un producto.

Obviamente este concepto es aplicable a los datos, hay un punto de inicio y, a continuación, un período de utilización, que normalmente finaliza en una fase de desuso. Una vez que los datos han quedado obsoletos, estos pueden ser destruidos, archivados, para un uso potencial posterior o simplemente olvidados.

Podríamos entonces definir el ciclo de vida de los datos en cualquier sistema de información, como «el lapso de tiempo en el que los datos existen, desde el momento de su creación, hasta su transformación: evolución, modificación, reutilización o destrucción».

El concepto, a pesar de ser simple, es de vital importancia, ya que como apuntamos, los datos no son un valor en sí mismos, sino el medio para llegar a un conocimiento, y alcanzar este conocimiento, no resulta viable sin un plan asociado para su gestión (y desarrollo).

Los riesgos de la falta de gestión de datos son la creación de datos de forma indiscriminada y sin normalización ni control institucional, lo que conlleva un altísimo riesgo de pérdida de información a largo plazo y una escasa reutilización e interoperabilidad. Estos

inconvenientes están directamente relacionados con el aumento del coste de la obtención de datos y con una escasa transparencia de los resultados.

1.3.2 Gestión del Ciclo de Vida de los Datos

Para que los datos se conviertan en un activo importante dentro de una organización (hecho que cada vez es más común), es imprescindible dedicar los recursos y los esfuerzos necesarios para su correcta gestión, que nos permita, no sólo su correcto almacenamiento y custodia, sino también su correcta reutilización.

La relevancia de la gestión del ciclo de vida de los datos se pone de manifiesto, entre otras razones, por el hecho de que todos los grandes desarrolladores de software y los grandes actores dentro de la industria relacionada con los datos, disponen de paquetes especializados para tal fin. Sin embargo, es importante remarcar que la gestión del ciclo de vida de los datos (DLM, *Data Life Cycle Management*) no es un paquete de software, sino un enfoque basado en políticas para gestionar el flujo de datos de un sistema de información durante todo su ciclo de vida. En la literatura se puede encontrar a veces el término ILM (*Information Life-Cycle Management*) para hacer referencia también a la gestión del ciclo de vida de los datos. Ambos términos son utilizados indistintamente para hacer referencia a este concepto, el lector encontrará que hay autores que distinguen ambos términos, aunque si bien es verdad, los matices son muy sutiles.

1.3.2.1 ¿Qué es la gestión del ciclo de vida de los datos?

La gestión del ciclo de vida de los datos incluye procedimientos y procesos (Figura 7), así como aplicaciones o paquetes de software específicos.

Los objetivos principales de la gestión del ciclo de vida de los datos son asegurar que los datos:

- Pueden ser utilizados y son fácilmente accesibles (en el sentido de que pueden ser descubiertos o encontrados, por alguien que busca información).

Figura 7 – Ciclo de vida de los datos[21]

[21] Sanda Ionescu. *Introduction to DDI 3.0.* Inf. téc. CESSDA Expert Seminar

- Están bien descritos. Es decir, existe información adicional que permite identificar a que hace referencia o que describe un dato.
- El acceso está garantizado y protegido. Es decir, existen políticas que garantizan que o bien, quien hace las consultas tiene permiso para acceder a los datos, o bien se han aplicado los procesos necesarios a los mismos para garantizar la confidencialidad (por ejemplo, anonimización de la información confidencial).
- Puedan ser utilizados en un futuro, más allá del objetivo inicial para el que fueron pensados.

Recientemente, la explosión de los datos, especialmente en el entorno académico y en formato digital, ha dado lugar a la acuñación de un término que engloba el concepto de la gestión del ciclo de vida de los datos: *Data Curation*. El término no tiene un único equivalente en castellano, pudiendo encontrarse en la literatura denominaciones como «curación de datos» o «preservación de datos».

También podemos definir el ciclo de vida de los datos en función de todas aquellas tareas que es necesario realizar para poder hacer uso de los mismos, es decir desde que se identifica dónde están localizados los datos hasta su descatalogación o borrado, si así fuese considerado. Podríamos entonces definir el ciclo de vida de los datos como el esquema que se representa en la Figura 8, es decir:

1. **Captura**. Se inicia con un proceso de captura de los datos, que incluye la identificación de las fuentes donde encontraremos los datos que nos interesan.
2. **Organización**. Procesos que incluyen la exploración, limpieza, transformación, así como el aseguramiento de la calidad de los datos. Estos procesos también incluyen la descripción de los datos que permitirán una posterior mejor utilización (meta información asociada a los datos). Esta fase concluye con la definición del modelo de datos y su preparación para la carga.

Figura 8 – Ciclo de vida de los datos

3. **Integración**. Procesos que permitan la carga y el modelo de datos definido, así como el modo en el que se persistirán los mismos (almacenarán para su posterior reutilización). También incluye la forma en la que se procesarán los datos.

4. **Análisis**. Mecanismos y algoritmos aplicados a los datos para la obtención de valor.

5. **Actuación**. Decisiones tomadas como consecuencia del análisis de los datos. Estas decisiones incluyen tanto el consumo de los datos (por ejemplo mediante *dashboards* o informes de visualización de la información), como la aplicación del conocimiento al negocio o al área de conocimiento correspondiente, o el reprocesamiento de los datos, como el enriquecimiento de los mismos con otros datos adicionales o mediante la aplicación de nuevos algoritmos que permiten obtener un mayor valor o su archivado como datos históricos para un uso futuro, o por qué no, su eliminación permanente del sistema.

Normalmente estos procesos se representan como un ciclo y no como un proceso lineal, en primer lugar porque la gestión de los datos es un proceso dinámico y en segundo lugar, porque el comportamiento natural del tratamiento de los datos es cíclico, es decir tras la captura, procesamiento y análisis de los datos suele llegar un proceso de toma de decisión que casi siempre desemboca en una nueva demanda de información que requerirá de nuevos datos, cuya fuente debe ser identificada, los datos capturados, procesados integrados, almacenados, analizados, etc. de manera que el ciclo vuelve a empezar.

Hasta este punto hemos tratado aspectos genéricos sobre los datos y algunos conceptos relacionados con ellos, que son de aplicación general, tanto a los métodos tradicionales de gestión de los mismos como a los métodos más novedosos. En el próximo capítulo entraremos de lleno en el concepto de Big Data.

Acercándonos a la realidad

En el artículo «Gestión del ciclo de vida de datos documentos: acercando posiciones»[22] publicado el 11 enero de 2013 por Elisa García-Morales en Grupo ThinkEPI, se señala que hay los dos tipos de información en las organizaciones: la estructurada o datos en bases de datos relacionales, y la no estructurada o documental. La autora señala la necesidad de que los programas informáticos puedan gestionar ambos tipos, yendo hacia una convergencia de los modelos *Information Lifecycle Management / Data Lifecycle Management* (ILM/DLM) y *Records Management* (RM).

- ***¿En qué sentido cree que los desarrollos recientes en materia de tratamiento de datos han alcanzado esta convergencia?***

[22] Elisa García-Morales. *Gestión del Ciclo de Vida de Datos Documentos: Acercando Posiciones*. Inf. téc. [Online; publicado el 11 de enero de 2013]. Grupo ThinkEPI

2 Introducción al Big Data

Las organizaciones han construido sus almacenes de datos para poder analizar la actividad de su negocio y proporcionar información de valor a aquellos que tienen que tomar decisiones con el fin de, por ejemplo, mejorar el desempeño empresarial o la efectividad operativa. Si bien es cierto que las tecnologías asociadas a los almacenes de datos y la inteligencia de negocio, así como las bases de datos en las que se sustentan, son muy maduras y robustas; el entorno en el que las empresas deben competir ha cambiado sustancialmente desde que estas tecnologías aparecieron.

Una sociedad hiperconectada como la actual, en la que Internet se incorpora de forma creciente a los procesos de negocio y en la que la necesidad de información también aumenta; es capaz de generar datos en un volumen, a una velocidad y de una variedad inconcebibles hasta la fecha.

Estas características y los requisitos que conllevan, comenzaron a poner de manifiesto que las tecnologías tradicionalmente utilizadas para el tratamiento, almacenamiento, gestión y análisis de los datos, no eran adecuadas.

Si bien es cierto que las tecnologías de *Business Intelligence* y almacenes de datos no han quedado obsoletas, todavía tienen cabida en el entorno empresarial y son soluciones adecuadas para ciertos entornos; estas tecnologías no son capaces de gestionar las características que los datos y la información tienen en Internet (ver sección 1.1.3 – Orígenes de Big Data del capítulo 1).

La necesidad de gestionar mayores volúmenes de datos, a una velocidad y de una variedad creciente, ha dado lugar al desarrollo de nuevas tecnologías, que siguen evolucionando, y que son capaces de hacerlo a unos costes razonables.

2.1 Introducción al Big Data

2.1.1 ¿Qué es Big Data?

El término «Big Data» hace referencia a una acumulación masiva de datos tal, que supera la capacidad de las herramientas tradicionales para que sean capturados, gestionados y procesados en un tiempo razonable. En este sentido, la idea predominante es que un conjunto de datos entra dentro de la categoría de Big Data si es demasiado grande como para manejarlo de forma apropiada con los programas convencionales de software ampliamente disponibles, y que por lo tanto requiere analistas especializados[23].

Una forma intuitiva de explicar en lo que consiste Big Data es diciendo que su esencia radica en utilizar los datos para resolver problemas. Estos problemas pueden ser de índole empresarial, personal o pública. Así, se puede utilizar Big Data para resolver un problema de atención al cliente en una empresa de telefonía, para darle consejos nutricionales a una persona a partir de los datos facilitados por su nevera que registra los alimentos que utiliza diariamente o para abordar un problema relacionado con los turnos de los enfermeros en un hospital público.

¿Qué es Big Data para los gerentes?

Un sondeo reciente realizado por IBM entre un grupo de gerentes muestra cómo los encuestados tienen puntos de vista divididos sobre si Big Data se define mejor en relación al gran volumen de datos de hoy en día, a los nuevos tipos de datos y análisis, o a las nuevas necesidades de análisis de información en tiempo real. Así, a la pregunta ¿qué entienden los directivos por "Big Data"? las respuestas más frecuentes son las siguientes[24,25]:

- Un mayor abanico de información – 18 %.
- Nuevos tipos de datos y análisis – 16 %.
- Información en tiempo real – 15 %.
- Flujo de datos de nuevas tecnologías medios no tradicionales – 13 %.

En cualquier caso, aunque es cierto que en esencia Big Data es una herramienta para analizar grandes cantidades de datos, resulta más interesante identificar Big Data con

[23] Anita McGahan. «Unlocking the Big Promise of Big Data». En: *Rotman Magazine* (2013)

[24] Michael Schroeck y col. *Analytics: el Uso de Big Data en el Mundo Real*. Inf. téc. [Online; consultado el 8 de diciembre de 2015]. IBM Institute for Business Value, 2012

[25] BBVA Innovation Edge. *Big Data: Es Hora de Generar Valor de Negocio con los Datos*. http://es.slideshare.net/cibbva/big-data-castellano. [Online; publicado el 27 de junio de 2013]

el tipo de información que nos pueden aportar los datos y en concreto con las cuatro preguntas siguientes[26]:

1. La primera es de carácter descriptivo: ¿qué ha ocurrido?
2. La segunda está relacionada con el diagnóstico: no me digas únicamente qué ocurrió, ¿por qué ocurrió?
3. La tercera es predictiva: ¿qué va a ocurrir?
4. La cuarta pregunta es prescriptiva: ¿qué puedo hacer para que eso ocurra?

2.1.1.1 Big Data como un cambio de paradigma

Es frecuente malinterpretar lo que realmente supone la revolución del Big Data. Algunos creen que lo esencial y novedoso del Big Data es aplicar las nuevas tecnologías al tratamiento de los datos, mientras que otros, excesivamente condicionados por el uso intenso que los expertos en marketing digital hacen de la analítica de datos, piensan que fundamentalmente el Big Data son unas nuevas herramientas enfocadas a tratar los datos y obtener el máximo de información de los usuarios de las redes sociales. Pero desde una perspectiva diferente puede decirse que lo esencial del Big Bata es que supone un profundo cambio de paradigma. Hasta ahora la investigación científica funcionaba, esquemáticamente como sigue. En primer lugar, un investigador tenía una intuición, seguidamente formulaba una hipótesis y por último, buscaba datos para corroborarla o refutarla.

Esta forma de llevar a cabo la investigación es una aproximación dirigida por las hipótesis. El Big Data es todo lo contrario: primero se buscan patrones en los datos y luego se formula una hipótesis sobre lo que se observa. Es una aproximación dirigida por los datos. Dando un paso más, algunos incluso dicen que es el fin de la teoría, pues desde ésta perspectiva cuantos más datos se tengan más descubrimientos se podrán hacer.

Una afirmación más ponderada sería decir que la clave más que en los datos en sí está en la capacidad de utilizar de forma inteligente los datos: aquellas organizaciones que sepan hacer las preguntas adecuadas en el momento preciso serán las que incrementen su ventaja competitiva. Ello se debe a que en cierto sentido el Big Data puede tener las respuestas a todas nuestras preguntas; lo importante es saber qué preguntar.

Acercándonos a la realidad

Como se señala en el artículo «*Why Big Data is the new competitive advantage*» escrito por Tim McGuire, James Manyika y Michael Chui en julio-agosto de 2012[27], la mayor

[26] David Simchi-Levi. *Overheard at MIT - Four Questions to Answer*. http://sloanreview.mit.edu/article/overheard-at-mit. [Online; publicado el 19 de diciembre de 2013]

[27] Tim McGuire, James Manyika y Michael Chui. *Why Big Data is the New Competitive Advantage*. http://iveybusinessjournal.com/publication/why-big-data-is-the-new-competitive-advantage. [Online; publicado en agosto de 2012]

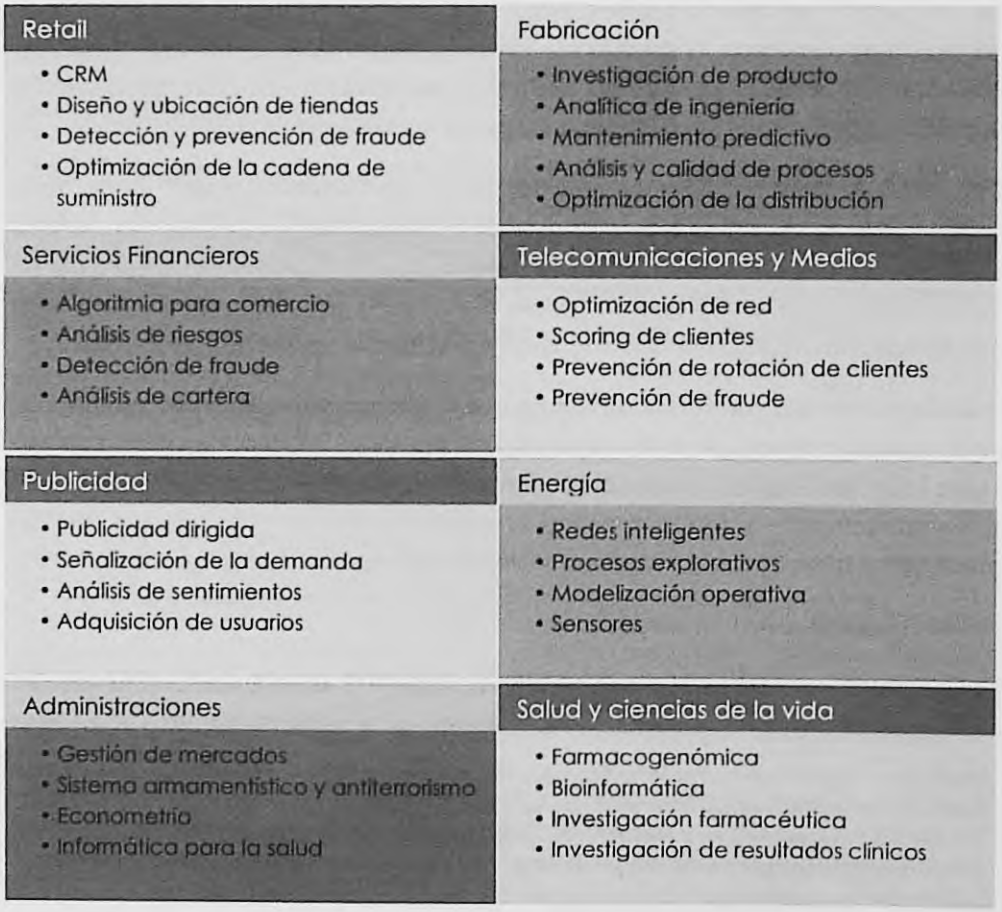
parte de los sectores reconocen ya que Big Data y el análisis de datos pueden disparar la productividad, hacer que los procesos sean más visibles y mejorar las predicciones sobre el comportamiento.

- ***¿En qué sentido el Big Data puede marcar la diferencia entre los perdedores y los vencedores en el futuro?***

2.1.1.2 ¿Quién utiliza Big Data?

El uso de Big Data se está generalizando entre organizaciones de todo tipo (Figura 9). Es un hecho que las grandes empresas, las startups, las agencias gubernamentales y las organizaciones no gubernamentales, de forma progresiva se están viendo animadas a utilizar los microdatos generados por los dispositivos digitales para poder actuar de forma más eficiente. Estos microdatos, cuando se agregan se convierten en cantidades masivas de datos cuya gestión requiere herramientas y capacidades especializadas[23].

Figura 9 – Utilización de Big Data en diferentes industrias



No hay duda de que las organizaciones están nadando en un mar cada vez mayor de datos que son demasiado voluminosos o demasiado poco estructurados para ser gestionados y analizados a través de medios tradicionales. Entre las fuentes de estos datos están las derivadas de la secuencia de clics desde la Web, el contenido de los social media; tuits, blogs, publicaciones del muro de Facebook (sólo Facebook cuenta con más de 1.000 millones de usuarios activos generando datos de interacción social) o los sistemas de identificación por radiofrecuencia, que generan hasta 1.000 veces más datos que los sistemas convencionales de códigos de barras, (sólo Walmart gestiona más de 1 millón de transacciones con clientes por hora).

En el mundo se registran cada segundo 10.000 transacciones de pagos con tarjetas. Más de 5.000 millones de personas telefonean, mandan mensajes de texto y navegan por internet con teléfonos móviles. Cada día se envían 340 millones de tuits, unos 4.000 por segundo[25]. Al día se generan 2,5 trillones de bytes de datos. Sin embargo, muy poca de esa información tiene el formato de filas y columnas de las bases de datos tradicionales.

Respondiendo de forma específica a la pregunta de quién utiliza Big Data, podríamos decir, siguiendo a Davenport[28], que si bien la gestión de la información y de los datos es algo que se utiliza en prácticamente todas las áreas, hay algunas áreas en las que Big Data se emplea con especial intensidad: Marketing analítico, cadena de suministros, recursos humanos, finanzas, análisis de los clientes, *web/mobile/social analytics*, efectividad de las operaciones (mejora de procesos: productividad) y fraude y análisis del riesgo.

Acercándonos a la realidad

En la publicación del BBVA Innovation Edge Big Data, de junio de 2013[25], se presenta la siguiente relación de sectores que están utilizando Big Data:

> *Sectores que están utilizando Big Data para transformar los modelos de negocio y mejorar el rendimiento en muchas áreas: -Manufacturas -Investigación de productos -Análisis de ingeniería -Mantenimiento predictivo -Análisis de procesos y calidad -Optimización de la distribución -Optimización de redes -Valoración de los clientes -Evitar pérdida de clientes -Prevención del fraude -Medios y telecomunicaciones -Energía -Redes inteligentes -Exploración -Modelos operacionales -Sensores de tendido eléctrico -Salud y ciencias de la vida -Farmacogenómica -Bioinformática -Investigación farmacéutica -Investigación de resultados clínicos Servicios financieros -«Trading» automatizado -Análisis de riesgos -Detección del fraude -Análisis de carteras Venta minorista -Gestión de relaciones con el cliente -Ubicación y distribución de tiendas -Detección y prevención del fraude -Optimizar la cadena de suministros -Precios dinámicos Gobierno -Gobernanza del mercado -Sistemas de armas y contraterrorismo -Econometría -Informática aplicada a la salud -Publicidad y*

[28]Thomas H. Davenport. *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Harvard Business Review, 2014

relaciones públicas -Gestión de señales de demanda -Publicidad personalizada -Análisis de sentimiento del mercado -Adquisición de clientes

- ***Reflexionando sobre esta relación de sectores y actividades que ya emplean Big Data ¿Se te ocurren algunas otras posibles actividades en las que se podría emplear Big Data?***

2.1.1.3 ¿Qué tipo de situaciones se pueden analizar mediante Big Data?

Para ilustrar sobre el tipo de situaciones que pueden generar datos susceptibles de ser analizados mediante técnicas de Big Data y la utilidad derivada de su uso, cabe citar los siguientes ejemplos:

- «Información sobre todas las direcciones a las que los miembros de una cooperativa de taxistas, de una gran ciudad viajaron a lo largo del último año». De esta forma se podrían inferir las rutas, los horarios y las paradas más eficientes de los cooperativistas.
- «El tipo de información facilitada y número de minutos que los clientes individuales de una empresa financiera pasaron consultando con sus agentes antes de decidirse a hacer una compra efectiva de un producto financiero». El objetivo de la empresa financiera consistiría en evaluar la productividad de los agentes y qué tipo de información se debe facilitar a los clientes.
- «Las características, en términos de bacterias, del agua de una playa de una ciudad del litoral Mediterráneo de acuerdo con pruebas diarias». El objetivo sería evaluar la efectividad de las depuradoras en las distintas épocas del año.
- «Los datos climatológicos tomados cada día en diversos puntos de España a lo largo de los últimos años». La razón de ser de esta información sería anticipar la posibilidad de inundaciones y otras catástrofes climatológicas.
- «Datos diarios sobre los trayectos de los viajeros de los autobuses municipales en una gran ciudad». Con esta información se podrían optimizar las rutas y las frecuencias de los autobuses.
- «Las conversaciones de los clientes de una empresa con su centro de atención al cliente». El objetivo de la empresa puede ser inferir los sentimientos de los clientes sobre un determinado servicio o producto.

La relación de posibles casos presentada es una mera ilustración de las enormes posibilidades que ofrece Big Data para predecir multitud de acontecimiento y para ayudar a tomar decisiones. La combinación de las ingentes cantidades de datos, muchos en tiempo real, que ofrece la economía digital y las modernas capacidades de computación abre

las posibilidades para abordar el análisis de casi cualquier tipo de problema, tanto en el sector privado, como en el público.

Big Data y las administraciones públicas

Vamos a empezar analizando las posibilidades, muchas de ellas ya hechas realidad, en la gestión de lo que se conoce como las *smart cities*. La gestión inteligente de las ciudades es consecuencia de la hiperconectividad, y eso es algo en lo que se continuará avanzando. Así, por ejemplo, las rutas de los autobuses se adaptarán automáticamente a la demanda, el tráfico y los aparcamientos se gestionarán de forma cada vez más automatizada, los servicios sanitarios de urgencias serán cada vez más eficientes, etc. Todo ello, porque cuando se analizan las ciudades se observa que las infraestructuras y la población se comportan de acuerdo a patrones muy específicos. Aunque cada individuo puede actuar al azar, cuando se mira todo desde una perspectiva poblacional, se ve cómo emergen patrones. Si se cruzan diferentes fuentes de datos demográficos, climáticos o de tráfico, se puede empezar a predecir todo lo que ocurre, desde el flujo de pasajeros en el transporte público a accidentes de coche y crímenes. Esto permite crear soluciones a problemas incluso antes de que ocurran, y hacer que nuestras ciudades sean más eficientes y seguras.

En cualquier caso, muchas oportunidades para el empleo de Big Data están relacionadas con los grandes problemas de la sociedad actual como el cambio climático, la fragilidad de nuestro sistema financiero, las enfermedades epidémicas, la corrupción generalizada, la privación de derechos de los pobres, o el agotamiento de minerales que son extraíbles a un bajo coste, por nombrar sólo algunos de estos[23]. En pocas palabras: la «gran promesa» de Big Data radica en contribuir a avanzar en la solución de la mayoría de los problemas más importantes de gestión de nuestro siglo.

Para ilustrar lo señalado vamos a presentar unos ejemplos en los que se ha utilizado Big Data para abordar problemas relacionados con la administración pública y con la política.

El fracaso escolar en Gwinnett y el Álgebra I

En 2002 el sistema público de enseñanza del condado de Gwinnett de Greater Atlanta, en Estados Unidos, tenía un problema: el rendimiento de los estudiantes estaba descendiendo y la tasa de abandono escolar estaba aumentando[29]. Por ello, los responsables del sistema público de enseñanza decidieron investigar las causas de aumento de la tasa de abandono escolar para tratar de revertir la situación y optaron por utilizar Big Data.

Un análisis riguroso de los datos les permitió comprobar que la mejor variable explicativa del abandono escolar era haber suspendido la asignatura Álgebra I. En otras palabras, los datos evidenciaron que ningún predictor era más potente de finalizar los estudios que haber aprobado la asignatura Álgebra I, que se impartía en noveno o décimo grado.

[29]Steve LaValle y col. «Big Data, Analytics and the Path from Insights to Value». En: *MIT Sloan Management Review* 52.2 (2011), págs. 21-31

Seguidamente los investigadores se formularon la siguiente pregunta: ¿qué se podría hacer para que los niños que habían sufrido históricamente con las matemáticas– que habían fracasado en los cursos anteriores –mejorasen lo suficiente como para que pudieran aprobar Álgebra I? Un análisis cuidadoso de los datos mostró una respuesta un tanto asombrosa: el mejor predictor de éxito en la asignatura Álgebra I, para estudiantes que habían fracasado los cursos anteriores de matemáticas, era haber superado con éxito la asignatura de Escritura creativa de octavo grado. Con ese descubrimiento en la mano, los responsables del sistema educativo le dedicaron atención y recursos a ayudar a los estudiantes a tener éxito en las clases de Escritura creativa.

El resultado de estas iniciativas fue aleccionador. A medida que mejoraron las tasas de éxito de Escritura creativa, también lo hicieron las tasas de éxito de la asignatura Álgebra I y, con el tiempo, también lo hizo la tasa de graduación de la escuela secundaria; al reducirse la tasa de abandono escolar. La tendencia aparentemente irreversible se invirtió. Y en el otoño de 2010, Gwinnett ganó el prestigioso Premio Broad, que honra a los distritos escolares en los que los estudiantes han alcanzado los mejores logros y han presentado la tasa de mejora más elevada.

Este es un ejemplo de cómo una aplicación correcta de Big Data permite convertir la información en conocimiento y el conocimiento en acción. Y la acción ha producido un resultado histórico.

El caso de Gwinnett ilustra como un análisis de los datos bien realizado puede contribuir a cambiar la vida de los jóvenes a través de la mejora del funcionamiento de las instituciones. En el caso de Gwinnett también hubo descubrimientos incómodos. Los datos de Gwinnett revelaron, por ejemplo, que algunos enfoques consagrados en las políticas de educación no sólo eran ineficaces sino contraproducentes. Lo positivo de este caso es que los responsables de Gwinnett decidieron aceptar las consecuencias del análisis impulsado por los datos y llevar a cabo las reformas pertinentes. Reconocieron que algunas creencias arraigadas estaban equivocadas e impulsaron los cambios oportunos. La moraleja de este caso es que cuando las organizaciones gestionan de forma rigurosa basándose en los datos, efectivamente pueden llegar a liderar los resultados en sus respectivos sectores.

La campaña de Barack Obama de 2012 y el Big Data

La campaña del presidente estadounidense Barack Obama de 2012 le debe mucho de su éxito al Big Data. El análisis realizado permitió identificar, por ejemplo, qué personas probablemente se inclinarían a votar por él después de recibir un *flyer*, una llamada telefónica o una visita a su domicilio. Este tipo de factores contribuyeron a inclinar la balanza a su favor en los estados cruciales.

Así pues, el Big Data ha sido en buena medida el secreto que empujó la campaña de Barack Obama al éxito. Como señaló Jim Messina, responsable de la campaña, «Vamos a

medir cualquier cosa en esta campaña». Entre las iniciativas que se tomaron cabe destacar las siguientes[30]:

- El equipo de analítica de datos multiplicó el número de empleados por cinco, en relación a los contratados en la campaña anterior (2008).
- El equipo de analistas estaba dirigido por un "científico de datos" de prestigio.
- El trabajo de los analistas se mantuvo en paralelo y sin mucho contacto con el resto de miembros de la campaña, y para mantener la confidencialidad emplearon nombres en código para los temas en los que estaban trabajando.
- Integraron las diversas bases de datos en una única que sumaba la información que día a día recopilaban los voluntarios con las bases de datos históricas que se habían conseguido gracias al registro en la web de anteriores campañas de Obama.
- Manejando la información los analistas descubrieron cosas curiosas, como por ejemplo, que una cena de Obama con George Clooney sería uno de procedimientos más eficientes para recaudar fondos entre las mujeres de 40 a 49 años.
- La analítica de datos también permitió descubrir que Michelle Obama era un gran reclamo para conseguir financiación.
- El Big Data también les ayudó a abordar los estados complicados. Estudiaron durante meses los posibles votantes de los estados que creían que iban a ser claves. De esta forma pudieron conocer el estado real de la intención de voto y por tanto actuar de forma realista. En este sentido, en Ohio, donde Obama terminó ganando con un 50,1 % de los votos totales, se llegaron a identificar a 29.000 votantes indecisos, a los que se dirigieron con una campaña específica para ellos.
- El análisis de los datos también fue clave en la compra de publicidad o en la elección de soportes para lanzar los mensajes. Así, en Florida detectaron que el grupo clave para ganar eran las mujeres menores de 35 años. Tras profundizar en el análisis descubrieron que a la mayoría de este colectivo les gustaban las mismas series de televisión, así que invirtieron en anuncios que se mostraron durante la emisión de dichas series.
- Barack Obama fue el primer presidente en conceder una entrevista con los internautas en directo en el agregador Reddit, una de las páginas más populares en todo el mundo. ¿El resultado? En cifras tangibles, más de 24.000 comentarios y más de 5 millones de páginas vistas. La repercusión online fue mucho

[30] TICbeat. *La Campaña de Big Data que Dio la Victoria a Obama*. http://www.ticbeat.com/bigdata/campana-big-data-dio-victoria-obama/. [Online; publicado el 8 de noviembre de 2012]

más allá, con multitud de medios recogiendo el hecho de que el presidente se sentara a responder, él mismo, las preguntas de los usuarios[31].

Así, pues, al igual que puede decirse que las elecciones de 2008 fueron las elecciones de las redes sociales, en 2012 puede decirse que se celebraron las elecciones del Big Data.

Acercándonos a la realidad

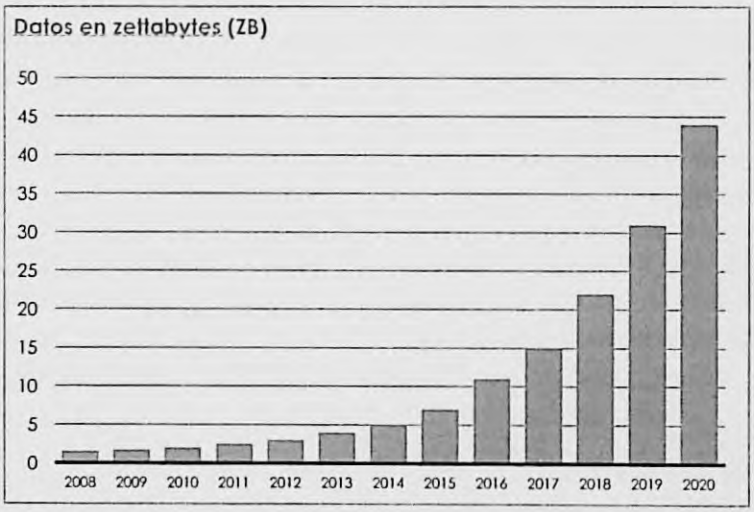
Según se desprende de la información aparecida en un artículo de Constance L. Hays, publicado el 14 de noviembre de 2014 en New York Times («What Wal-Mart Knows About Customers' Habits»)[32], Wal-Mart ha aprendido gracias a Big Data, que la amenaza inminente de que un huracán va a azotar un área, no sólo hace que la demanda de linternas aumente sino también la de Pop-Tarts (tipo de cereales de desayuno) de fresa.

- ***¿Se le ocurre otros casos en los que las empresas podrían utilizar Big Data para mejorar sus resultados?***

2.1.1.4 La «Dataficación» de la sociedad y sus implicaciones

Según las estimaciones de Eric Siegel, estamos agregando 2,5 trillones de bytes de datos cada día[33]. La Figura 10 muestra la previsión de crecimiento anual del volumen de datos, que muestra una tendencia de crecimiento exponencial. Esto se debe en buena

Figura 10 – Crecimiento anual del volumen de datos

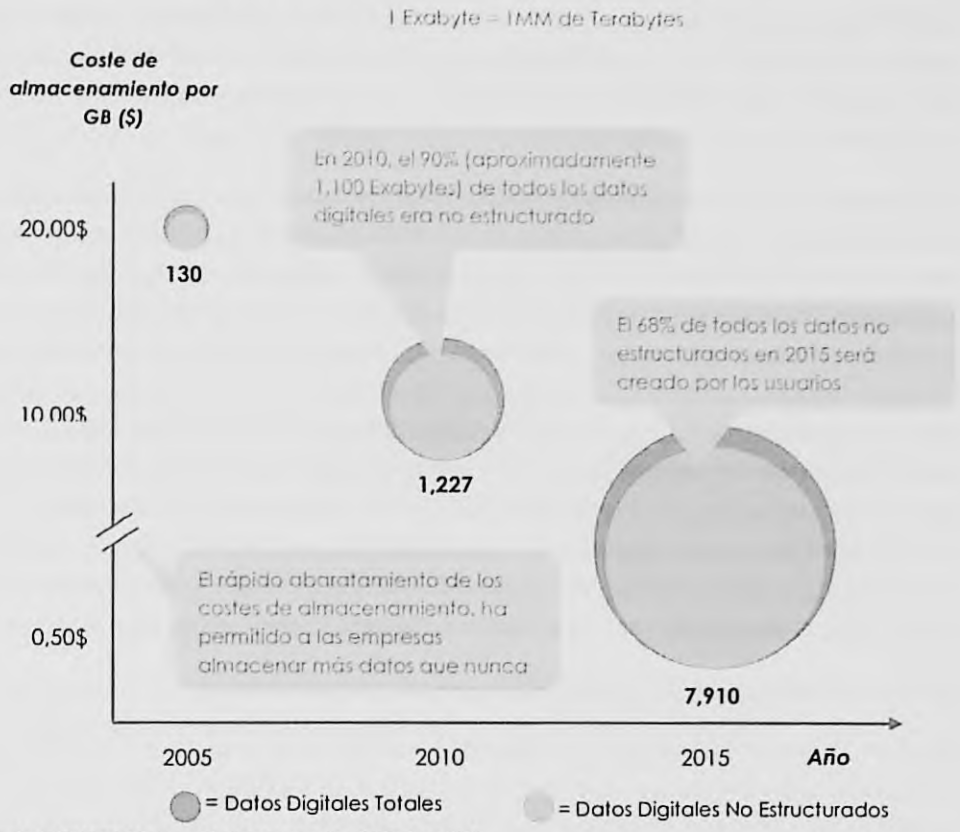


[31]Zenith. *El Triunfo de Obama en Internet: Caso de Estudio de las Campañas de 2008 y 2012 (II)*. http://blogginzenith.zenithmedia.es/el-triunfo-de-obama-en-internet-caso-de-estudio-de-las-campanas-de-2008-y-2012-ii/. [Online; publicado el 10 de enero de 2013]

[32]Constance L. Hays. *What Wal-Mart Knows About Customers' Habits*. The New York Times. [Publicado el 14 de noviembre de 2004]

[33]Eric Siegel. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Wiley, 2013

Figura 11 – Almacenamiento de datos en Exabytes



Fuente: IDC's Digital Universe Study, sponsored by EMC, June 2011

medida a que las palabras se han convertido en datos; los estados físicos de nuestra maquinaria se han convertido en datos; nuestras ubicaciones físicas se han convertido en datos; e incluso nuestras interacciones con los demás se han convertido en datos. Los datos con frecuencia pueden ser recopilados pasivamente, sin mucho esfuerzo o incluso sin conciencia por parte de los que están siendo grabados. Y debido a que el coste de almacenamiento ha caído drásticamente (Figura 11), es más fácil justificar mantener los datos que desecharlos, tal como señalan Viktor Mayer-Schönberger y Kenneth Cukier[34]. Los investigadores se refieren a este nuevo fenómeno como la «dataficación» de todas las cosas.

De hecho, estamos inundados de información, pero ¿qué significa todo esto? ¿Qué nos dicen los datos? Ciertamente algunas empresas se han convertido en expertas en la manipulación selectiva de los datos y han descubierto todo tipo de correlaciones valiosas. Algunas no son del todo sorprendentes. Por ejemplo, según los informes de Siegel, las personas que compran pequeñas almohadillas de fieltro que se adhieren a la parte inferior de las patas de la silla (para proteger el suelo) suelen presentar un menor riesgo

[34]Viktor Mayer-Schönberger y Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt, 2013

de crédito que la media[35]. Otros resultados son, sin embargo, bastante inesperados. Los fumadores en algunos lugares de trabajo tienden a sufrir menos del síndrome del túnel carpiano (tal vez debido a que generalmente toman más pausas en el trabajo), y los vegetarianos tienden a perder menos vuelos (tal vez porque encargan una comida especial y están por lo tanto más comprometidos a hacer el viaje).

La realidad es que existen grandes cantidades de datos digitales sobre prácticamente cualquier tema de interés de todo negocio. Los teléfonos móviles, las compras online, las redes sociales, la comunicación electrónica, los GPS y la maquinaria instrumental producen torrentes de datos como un subproducto de sus operaciones ordinarias. Cada uno de nosotros es ahora un generador andante de datos. Los datos disponibles están generalmente en forma no estructurada, no están organizados en una base de datos y resultan difíciles de manejar; pero hay una enorme cantidad de señal en el ruido. La señal, esto es, la información útil, simplemente está esperando a ser liberada. En este sentido, la clave no radica en contar con mejores algoritmos sino en tener más datos y en saber cómo gestionarlos[36]. Para tratar esta ingente cantidad de información, si bien las herramientas de análisis de la información anteriores a Big Data aportaron técnicas rigurosas que ayudaban en el proceso de toma de decisiones; Big Data es a la vez más simple y potente.

Acercándonos a la realidad

Por lo señalado hasta resulta fácil aceptar el enorme potencial de Big Data, pero algo que puede resultar sorprendente es que Big Data pueda incidir positivamente en la curación de las personas tal como se apunta en el artículo de James Kobielus «Big Data and the power of positive curation», publicado el 11 de diciembre de 2014[37].

- ***Reflexione sobre el contenido del artículo y analice como Big Data puede ayudar a las personas a confiar en el uso de los datos como la «única versión de la realidad».***

2.1.2 Factores Diferenciadores y Rasgos Característicos del Big Data

En esta sección vamos a analizar, en primer lugar, en qué sentido las organizaciones que optan por Big Data se diferencian de las que siguen con los métodos tradicionales de análisis de datos y en segundo lugar vamos a revisar los rasgos característico del Big Data.

2.1.2.1 Big Data vs. análisis tradicional

Muchas personas en el mundo de la tecnología de la información y en las salas de juntas de muchas organizaciones de está hablando de Big Data. Pero, ¿en qué medida los

[35]Alden M. Hayashi. «Thriving in a Big Data World». En: *MIT Sloan Management Review* 55.2 (2013), págs. 35-39

[36]Andrew McAfee y Erik Brynjolfsson. *Big Data: The Management Revolution.* Harvard Business Review, 2012

[37]James Kobielus. *Big Data and the Power of Positive Curation.* http://www.ibmbigdatahub.com/blog/big-data-and-power-positive-curation. [Online; publicado el 11 de diciembre de 2014]

potenciales puntos de vista de Big Data difieren de los que se pueden deducir a partir de análisis tradicionales de datos? Muchos creen que, para las empresas que lo hacen bien, Big Data podrá abrir la ventana a nuevas capacidades organizativas y a crear valor. Pero, ¿en qué sentido su potencial difiere de lo que los administradores pueden generar a partir de los análisis tradicionales?

Muchos proveedores de tecnologías de la información y de soluciones informáticas utilizan el término Big Data como una palabra de moda para referirse a los análisis más inteligentes de datos. Pero Big Data es realmente mucho más que eso. De hecho, las empresas que aprenden a aprovechar Big Data utilizarán información en tiempo real de los sensores, la identificación por radiofrecuencia y la información generada por las redes sociales para conocer sus entornos de negocio a un nivel más granular, para crear nuevos productos y servicios, y para responder a los cambios en los patrones de uso a medida que estos tienen lugar. Así, por ejemplo en las ciencias de la salud, esta inmediatez en el análisis, su realización por expertos en Big Data y su conexión con el núcleo o core donde se toman las decisiones en las organizaciones permitirá que la información obtenida pueda allanar el camino a los tratamientos y curas para enfermedades mortales, suponiendo un avance significativo en pro de la vida.

En este sentido, puede afirmarse que las organizaciones que optan por Big Data se distinguen del análisis tradicional de datos por los tres hechos siguientes[38]:

1. Por prestarle atención prioritaria al flujo de datos en lugar de a un *stock* fijo de los datos.
2. Por basarse en expertos en Big Data, «*Data Scientists*» (científicos de datos), en lugar de en analistas tradicionales de datos.
3. Por integrar el Big Data en el núcleo del negocio y en las áreas operacionales y de producción.

Prestarle atención prioritaria a los flujos de datos en lugar de a los *stocks* fijos de datos

Hay varios tipos de aplicaciones de Big Data que evidencian las ventajas de prestarle atención a los flujos de datos en vez de a los *stocks* de datos:

- El primer tipo de ejemplos de análisis centrados en los flujos de información son ciertos procesos de cara al cliente que permiten hacer cosas como identificar el fraude en tiempo real o, en el área de la salud, evaluar aquellos pacientes que son de alto riesgo.

[38]Thomas H. Davenport, Paul Barth y Randy Bean. «How 'Big Data' Is Different». En: *MIT Sloan Management Review* 54.1 (2012), págs. 22-24

- Un segundo tipo de ejemplos que también implican un proceso continuo de monitorización serían los realizados para detectar temas tales como los cambios en los sentimientos del consumidor o la necesidad de un servicio mantenimiento de un motor de un avión.

- Un tercer tipo de ejemplos serían aquellos que utilizan grandes volúmenes de datos para explorar las relaciones en la red como amigos sugeridos en LinkedIn y Facebook.

En todas estas aplicaciones, los datos a analizar no son el *stock* en un almacén de datos, sino un flujo continuo. Esto representa un cambio sustancial respecto al pasado, cuando los analistas de datos llevaban a cabo múltiples análisis para encontrarle sentido a una oferta fija de datos. Hoy en día, en lugar de mirar los datos para evaluar lo que ocurrió en el pasado, las organizaciones tienen que pensar en términos de flujos y procesos continuos. «*Streaming analytics* permite procesar los datos durante un evento para mejorar los resultados», señala Tom Deutsch, director del programa de Big Data y tecnologías de analítica aplicada en IBM[39]. Esta capacidad es cada vez más importante en campos como la sanidad. Así, por ejemplo, en el Hospital para Niños Enfermos de Toronto, algoritmos de *machine learning* –aprendizaje máquina– son capaces de descubrir patrones que anticipan infecciones en los bebés prematuros antes de que ocurran.

Desde una perspectiva general, hay que admitir que el aumento del volumen y la velocidad de datos en entornos relacionados con la producción, significa que las organizaciones tendrán que desarrollar procesos continuos para recoger, analizar e interpretar los datos. Las ideas obtenidas de estos esfuerzos pueden vincularse con las aplicaciones y los procesos productivos para permitir el procesamiento continuo. Aunque pequeños *stocks* de datos ubicados en almacenes de datos pueden seguir siendo útiles para el desarrollo y el perfeccionamiento de los modelos de análisis utilizados en Big Data, pero una vez que se han desarrollado los modelos, se necesita procesar flujos de datos de forma rápida, precisa y continua.

Algunos ejemplos de análisis basados en flujos de datos

El comportamiento de las compañías de tarjetas de crédito ofrece un buen ejemplo de esta dinámica. En el pasado, los grupos de marketing directo de estas compañías creaban modelos para estimar el comportamiento de los clientes. De esta manera podían determinar la probabilidad de que un cliente adquiriera un determinado producto, información que se obtenía de un almacén de datos de gran tamaño. El proceso de extracción de datos, preparación y análisis requería semanas para prepararse – y varias semanas más para ejecutarse. Sin embargo, las compañías de tarjetas de crédito, frustradas por su incapacidad para actuar con rapidez, determinaron que había una manera mucho más

[39]Thomas H. Davenport y Jinho Kim. *Keeping Up with the Quants: Your Guide to Understanding and Using Analytics*. Harvard Business School Publishing, 2013

rápida para satisfacer la mayoría de sus necesidades. De hecho, ellas fueron capaces de crear una base de datos «lista para salir al mercado» y un sistema que permite que un vendedor pueda analizar, seleccionar y enviar ofertas en un solo día. A través de iteraciones frecuentes y la monitorización de la página web y actividades del *call-center*, las empresas pueden realizar ofertas personalizadas en milisegundos, y luego optimizar las ofertas a lo largo del tiempo.

Algunos entornos Big Data, como el análisis de los sentimientos del consumidor no están diseñados para la automatización de decisiones, sino que son más adecuados para la monitorización en tiempo real del entorno. Dado el volumen y la velocidad de Big Data, los enfoques convencionales de elevada certeza en la toma de decisiones no suelen ser apropiados en esos ámbitos; pues para el momento en que la organización tiene la información que necesita para tomar una decisión, ya hay disponibles nuevos datos que hacen que la decisión quede obsoleta. En contextos de monitorización en tiempo real, las organizaciones deben adoptar un enfoque más continuo para el análisis y la toma de decisiones sobre la base de una serie de los flujos de información y pequeños cambios de comportamientos. Así, por ejemplo, en la analítica de las redes sociales lo importante es capturar las rápidas rupturas de las tendencias de los sentimientos de los clientes sobre productos, marcas y empresas. Aunque las empresas pueden estar interesadas en saber si cambios de una hora o un día en los sentimientos online se correlacionan con cambios de las ventas, para el momento en que un análisis tradicional se hubiera completado, se dispondría de una serie de nuevos datos que habría que analizar. Por lo tanto, en entornos de Big Data es importante analizar, decidir y actuar rápidamente, y hacerlo con frecuencia.

Pero no basta con ser capaz de controlar un flujo continuo de información. La organización que decida utilizar Big Data también tiene que estar preparada para tomar decisiones y tomar medidas. Las organizaciones necesitan establecer procesos para determinar cuándo las decisiones y acciones específicas son necesarias, como por ejemplo, cuando los valores de algunos datos caen fuera de ciertos límites. Establecer este tipo de procedimientos ayuda a determinar las posibles decisiones que los responsables debería tomar, a que se definan los procesos de toma de decisiones y a que se establezcan los criterios y plazos en los que las decisiones se deben tomar.

Basarse en científicos de datos en lugar de analistas tradicionales de datos

Los profesionales clave son los científicos de datos y los desarrolladores de productos y procesos, quedando muy en segundo plano los analistas tradicionales de datos. Aunque siempre ha existido la necesidad de profesionales de análisis de la información para apoyar la capacidad de análisis de los datos, los requisitos para el personal de apoyo son diferentes con Big Data (véase capítulo 3). Debido a que la interacción con los datos en sí –esto es, la obtención, extracción, manipulación y estructuración– es fundamental para cualquier análisis, las personas que trabajan con Big Data necesitan ciertos conocimientos informáticos y actuar de forma creativa. También tienen que estar cerca de

los productos y procesos dentro de las organizaciones. Esto implica que el trabajo de estos profesionales se debe organizar de manera diferente a como el personal de análisis estaba organizado en el pasado.

Los «científicos de datos», como se conoce a estos profesionales, deben entender la analítica, y también deben tener conocimientos de computación, de ciencias sociales y deben conocer el funcionamiento de las redes sociales. Sus capacidades incluyen, por lo tanto, la gestión de datos, un cierto conocimiento de los negocios y la capacidad de comunicarse de manera efectiva con los que toman las decisiones. Por lo tanto el perfil profesional, de un especialista en Big Data, va mucho más allá de lo que se requería para ser un analista de datos en el pasado. Reunir este tipo de capacidades no es fácil, por lo que la demanda de estos profesionales es muy considerable. Por ello, muchas organizaciones que emplean Big Data han optado por trabajar para desarrollar su propio talento en colaboración con las universidades y centros especializados.

En este sentido, EMC Corporation, por ejemplo, tradicionalmente un proveedor de tecnologías de almacenamiento de datos, se ha especializado en Big Data y ha comenzado a ofrecer servicios integrales incluyendo una oferta educativa para los científicos de datos. Su objetivo es convertir a Big Data en una base para que las organizaciones puedan crear nuevos niveles de valor ofertando productos y servicios personalizados y produciendo mayor satisfacción para los clientes. La idea es que los negocios que explotan Big Data para mejorar la estrategia y la ejecución se diferencien de los competidores. La solución de Big Data de EMC ofrece capacidad de almacenamiento, una plataforma de análisis unificada y herramientas de desarrollo de aplicaciones y procesos de negocios. El objetivo es lograr que las organizaciones extraigan información más detallada y se vuelvan más predictivas.

Por las razones señaladas, los usuarios de Big Data también están rediseñando sus estructuras organizativas para encajar mejor a sus expertos en Big Data en el seno de la estructura de la organización. Tradicionalmente, los analistas de datos formaban parte del departamento de TI y actuaban a modo de consultores internos que asesoraban a los administradores o directivos en la toma de decisiones. Sin embargo, en los sectores que adoptan Big Data, como por ejemplo, las redes sociales online, las empresas que se dedican a la gamificación y alguna empresas representativas de la industria farmacéutica, los profesionales que trabajan en Big Data son parte del núcleo de la organización o de la áreas responsables del desarrollo de nuevos productos y de la definición de las características de los producto. En Merck & Co. Inc., por ejemplo, los expertos en Big Data, esto es los científicos en genética estadística trabajan en las áreas de descubrimiento de nuevos fármacos y organización del desarrollo.

Integrar Big Data en el negocio principal y en las áreas operacionales y de producción

Al integrar Big Data en el negocio principal y en las líneas de negocio lo que se ha pretendido es ganar cercanía con los problemas que se deben abordar. La creciente cantidad

de datos requiere mejoras en las bases de datos y en las técnicas de análisis. La captura, filtrado, almacenamiento y análisis de los flujos de datos pueden inundar las redes tradicionales, los sitios de almacenamiento y las plataformas de bases de datos relacionales. Los intentos de reproducir y ampliar las tecnologías tradicionales no permitirán atender las demandas de Big Data. Big Data está cambiando la tecnología, las capacidades y los procesos del área de tecnologías de la información con el objetivo de integrar el análisis con el negocio.

El mercado ha respondido con una amplia gama de nuevos productos diseñados para atender a los requerimientos de Big Data. Incluyen plataformas de código abierto como Hadoop, inventadas por los pioneros de Internet para apoyar la escala masiva de datos que generaban y gestionaban. Hadoop permite cargar, almacenar y consultar conjuntos masivos a bajo costo, así como ejecutar análisis avanzados en paralelo. Las bases de datos relacionales también se han transformado: los nuevos productos han aumentado el rendimiento de las consultas en un factor de 1000 y son capaces de manejar la gran variedad de fuentes de Big Data. Los paquetes de análisis estadísticos también han evolucionado para trabajar con estas nuevas plataformas de datos, con los nuevos tipos de datos y con los algoritmos.

Un hecho que está contribuyendo a revolucionar la gestión de los datos y que requiere revisar los procesos, es la prestación de servicios de Big Data través de «la nube». A pesar de que aún no se ha adoptado ampliamente en las grandes corporaciones, la informática basada en la nube se adapta muy bien a Big Data. Los grandes volúmenes de datos que se manejan, la variedad de datos que se puede utilizar y la velocidad a la que hay que hacerlo cuando hablamos de Big Data, requieren de infraestructuras flexibles que permitan adaptarse a requisitos dinámicos de recursos, en muchas ocasiones hay que hacer crecer la capacidad de almacenamiento, la de procesamiento (CPU) o la de memoria, en algunas ocasiones este crecimiento será permanente, como suele ocurrir con el almacenamiento, y en otras es temporal, sólo hay que hacer crecer la infraestructura para atender picos de trabajo. Los únicos que pueden proporcionar un servicio de estas características, a unos precios razonables, son los proveedores de servicios basados en la nube.

Otro enfoque para la gestión de Big Data consiste en trabajar con los datos en la propia fuente, sin necesidad de procesamientos adicionales. Los llamados *data marts* virtuales (pequeños almacenes de datos) permiten a los científicos de datos compartir datos existentes sin replicarlos. EBay Inc., por ejemplo, solía tener un enorme problema de replicación de datos, con entre 20 y 50 versiones de los mismos datos esparcidos en sus diferentes *data marts*. Ahora, gracias a sus *data marts* virtuales, el problema de replicación de la compañía se ha reducido drásticamente. EBay también ha establecido un «centro de datos» – un sitio web interno para que sea más fácil para los administradores y analistas servirse a sí mismos, compartir datos y realizar análisis de toda la organización. De esta forma, eBay ha creado una red social en torno a la analítica y los datos.

Big Data está motivando a las organizaciones a repensar sus supuestos básicos sobre la relación entre el negocio y las tecnologías de la información y sus respectivos roles. El papel tradicional de las TI, en buena medida consistente en automatizar los procesos de negocio impone requisitos precisos, la adhesión a las normas estándar y el control de los cambios que se lleven a cabo. Big Data sigue un enfoque muy distinto. Un principio clave de Big Data es que el mundo y los datos que lo describen están cambiando constantemente, y las organizaciones pueden reconocer los cambios y reaccionar con rapidez. Mientras que en el enfoque tradicional las características más valoradas de los negocios y de TI solían ser la estabilidad y la escala, las nuevas ventajas se basan en el descubrimiento y la agilidad. Lo que más se valora ahora es la capacidad de explotar las fuentes de datos existentes y las nuevas fuentes; buscando continuamente patrones, eventos y oportunidades.

Todo lo señalado requiere un cambio radical en la actividad de las tecnologías de la información dentro de las organizaciones. A medida que el volumen de datos se dispara, las organizaciones necesitarán herramientas analíticas que sean fiables, robustas y capaces de ser automatizadas y que funcionen de forma integrada con el negocio. Al mismo tiempo, la analítica, los algoritmos y los interfaces de usuario que se empleen tendrán que facilitar la interacción de la gente que trabaja con las herramientas. Las organizaciones para alcanzar el éxito deberán formar y contratar a profesionales con un nuevo conjunto de habilidades y competencias de forma que puedan integrar las nuevas capacidades analíticas en sus entornos operacionales.

Otra forma en que Big Data altera los roles tradicionales del negocio y las tecnologías de la información es que pone el descubrimiento y el análisis, como la primera prioridad de la organización. Los procesos y sistemas en las tecnologías de la información propios del Big Data deben diseñarse para propiciar ideas innovadoras, no simplemente para lograr la automatización. En la arquitectura tradicional de las tecnologías de la información es frecuente que las aplicaciones o los servicios actúen como «cajas negras» que realizan tareas sin hacer públicos los datos y procedimientos internos. Pero en entornos Big Data los nuevos datos deben tener sentido y no basta con generar informes resumen. Esto significa que las aplicaciones de las tecnologías de la información tienen que medir e informar de forma transparente sobre una amplia variedad de aspectos, incluyendo interacciones de los clientes, uso de productos, acciones encaminadas a la prestación de servicios (*service actions*) y otras medidas dinámicas. A medida que Big Data evoluciona, la arquitectura deberá evolucionar para convertirse en un ecosistema de información: una red de servicios internos y externos que comparten continuamente la información, la optimización de decisiones, la comunicación de los resultados y la generación de nuevos conocimientos para las organizaciones[38].

2.1.2.2 Rasgos característicos del Big Data

Una vez presentado en qué sentido las organizaciones que optan por el Big Data se diferencian del análisis tradicional de datos podemos estudiar los tres rasgos característicos y diferenciadores del Big Data frente a las técnicas de análisis anteriores. Estos rasgos son el volumen, la velocidad y la variedad (las tres uves, Figura 12)[36].

Volumen

Como ya avanzamos en la sección 2.1.1.4 – La «Dataficación» de la sociedad, estamos generando y gestionando más cantidad de información en los últimos años que toda la existencia de la humanidad. Esto ofrece a las empresas una oportunidad de trabajar con muchos petabytes de datos en un único conjunto de datos, y no sólo a través de Internet.

Lo bueno de contar con tales cantidades de información es que las decisiones basadas en datos son mejores decisiones. El uso de grandes volúmenes de datos permite a los administradores decidir en base a hechos en lugar de tener que recurrir exclusivamente a la intuición. Por esta razón, contar con grandes volúmenes de datos tiene el potencial de revolucionar la gestión.

Velocidad

En muchas ocasiones para poder desarrollar ciertos tipos de aplicaciones, la velocidad de creación de datos es incluso más importante que el volumen. En el mundo Big Data, es

Figura 12 – Rasgos característicos de Big Data

VOLUMEN	VARIEDAD	VELOCIDAD
• El volumen de datos almacenados en los repositorios empresariales han crecido de los megabytes y gigabytes a los petabytes • El volumen de datos procesados por las empresas ha crecido significativamente: Google procesa 20 petabytes / día • En 2020, se espera que se generen 420 mil millones de pagos electrónicos. • El New York Stock Exchange genera 1 terabyte de datos al día frente a los feeds de Twitter que generan 8 terabytes de datos por día (o 80 MB por segundo)	• La variedad de datos con las que trabajar ya no es sólo la información estructurada que se almacena en los repositorios empresariales, sino que se ha extendido a la información no estructurada o la semiestructurada, como pueden ser: archivos de audio, video, XML, etc. • El streaming de datos, las cotizaciones de bolsa, las redes sociales, las comunicaciones entre máquinas (M2M – Machine-to-Machine), los datos procedentes de sensores, todos ellos han contribuido al aumento de la variedad que es necesario procesar y convertir en información	• La velocidad con la que se capturan, procesan o mueven datos, interna o externamente, se ha incrementado significativamente • Los modelos basados en tecnologías de Business Intelligence normalmente tardan días en procesar la información para su procesamiento - en comparación con los requisitos actuales de análisis "casi" en tiempo real utilizando flujo de entrada de datos de alta velocidad • eBay está abordando el fraude de uso de PayPal, mediante el análisis en tiempo real de 5 millones de transacciones cada día.

un hecho que la velocidad del movimiento de datos, del procesamiento y de la captura en y fuera de la empresa ha aumentado significativamente. Los modelos de inteligencia de negocios tradicionales requerían normalmente días para su procesamiento, mientras que en la actualidad es casi un requisito que el análisis sea en tiempo real, utilizando la corriente de entrada de datos de alta velocidad. Contar con información en tiempo real o casi en tiempo real hace posible que una empresa sea mucho más ágil que sus competidores. Así por ejemplo, eBay está abordando el fraude en el uso de PayPal mediante el análisis en tiempo real de 5 millones de transacciones cada día[40]. Otro ejemplo de la importancia de la velocidad es el aportado por Andrew McAfee y Erik Brynjolsson[36].

Un grupo de investigadores del MIT Media Lab utilizaron datos de localización de los teléfonos móviles para inferir cuántas personas estaban en los estacionamientos de Macy's el Black Friday, que supone el inicio de la temporada de compras navideñas en Estados Unidos. Con esta información pudieron estimar las ventas de los minoristas en ese día crítico incluso antes que el mismo Macy's hubiera registrado esas ventas. La posibilidad de poder realizar rápidamente inferencias de este tipo puede proporcionar una ventaja competitiva. Poder anticiparse en la toma de decisiones, puede ser un factor claro de éxito en determinadas actividades, como por ejemplo para los analistas de Wall Street.

Variedad

La variedad en el tipo de información utilizada es otro factor característico de Big Data. La información se recibe en forma de mensajes, actualizaciones e imágenes publicadas en las redes sociales, las lecturas de los sensores, las señales GPS de los teléfonos móviles, y otras más.

Las enormes cantidades de información de las redes sociales, por ejemplo, son sólo tan antiguas como las propias redes; Facebook fue lanzada en 2004 y Twitter en 2006. Las redes sociales con sus millones de usuarios y el uso que hacen de las mismas intercambiando mensajes, escribiendo post en blog, publicando fotos o vídeos, o simplemente comentando cosas de su interés son una continua fuente de variada información.

Lo mismo ocurre con los teléfonos inteligentes y otros dispositivos móviles que ahora proporcionan enormes flujos de datos vinculados a las personas, actividades y lugares. Debido a que estos dispositivos están en todas partes, es fácil olvidar que el iPhone salió al mercado en 2007 y el iPad en 2010. Ante esta variedad de fuentes y tipos de información, las bases de datos estructurados que almacenaban la mayoría de información corporativa hasta hace poco tiempo resultan poco adecuadas para almacenar y procesar los grandes y variados volúmenes de datos actuales.

[40]Ramesh Nair y Andy Narayanan. *Getting Results from Big Data: A Capabilities-Driven Approach to the Strategic Use of Unstructured Information*. Inf. téc. [Online; publicado el 25 de septiembre de 2012]. PwC Strategy&

A estas fuentes hay que añadir la eclosión de las señales capturadas por diferentes sensores, cuya última expresión es el concepto «Internet en las cosas» (véase el capítulo 1), donde se presupone que cada dispositivo será un captador o sensor y lo transmitirá de forma no estructurada ni lineal al mecanismo de registro. Todos estos datos tienen que procesarse y convertirse en información.

Por otro lado, a la esbozada variedad de fuentes de datos hay que añadir la evolución de los costes. La disminución constante del coste de todos los elementos de la informática de almacenamiento, memoria, procesamiento, ancho de banda, y demás implica que si bien hace unos años una gestión basada en un uso intensivo de los datos era costosa, ahora se está convirtiendo rápidamente en económica. Recientemente se está extendiendo la costumbre de añadir a la lista anterior otras dos uves más: es lo que se conoce como las cinco uves (5 Vs), que también asocian al Big Data las siguientes propiedades:

- **Veracidad**: los datos deben ser fieles a la realidad, no estar manipulados y ser fiables. Datos erróneos o mal interpretados pueden conducir a un análisis pobre de los mismos y a obtener conclusiones distorsionadas.
- **Valor**: esta propiedad hace referencia al hecho de que al disponer de una mayor cantidad de datos, estos se pueden cruzar y analizar para obtener un valor de negocio que los datos que se almacenaban tradicionalmente no eran capaces de revelar.

A medida que una mayor cantidad de información empresarial es digitalizada, nuevas fuentes de información y equipamientos cada vez más baratos se combinan para llevarnos a una nueva era; una era en la que compañías que nacieron como digitales, tales como Google y Amazon, ya se han convertido en auténticas expertas en gestión de datos. Pero lo importante es que el potencial para ganar ventaja competitiva puede ser incluso mayor para otras compañías. Interpretar correctamente y entender ese volumen de información requiere toda una nueva generación de tecnologías, de técnicas y una actitud innovadora ante la información. Las ventajas competitivas serán aprovechadas por los que comprendan mejor que sus competidores lo que está sucediendo y pongan en práctica un aprovechamiento ágil de las conclusiones que alcancen.

Acercándonos a la realidad

En el artículo de Guillermo Altares de 13 de septiembre de 2014 en El País – Comprar en la era de Big Data[41], se señala lo siguiente:

> *La capacidad para procesar cantidades ingentes de datos, lo que se conoce como Big Data, sumada a la información que ofrecemos voluntariamente y*

[41] Guillermo Altares. *Comprar en la Era del Big Data*. http://sociedad.elpais.com/sociedad/2014/09/13/actualidad/1410618299_290408.html. [Online; publicado el 13 de septiembre de 2014]

a las huellas que vamos dejando en Internet sin ser conscientes de ello, está revolucionando el consumo. El momento cumbre de este profundo cambio se produjo cuando un supermercado estadounidense de la cadena Target fue capaz de detectar que una adolescente estaba embarazada antes que sus padres, con un algoritmo que estudiaba sus hábitos de compra.

- ***Reflexione sobre las posibilidades que Big Data ofrece para estudiar los movimientos dentro de un centro comercial y aplicar a las ventas las conclusiones estadísticas. ¿En qué sentido nuestro ADN digital es una mina de información a la que nadie quiere renunciar?***

2.1.2.3 La interacción social en Big Data y la empresa

La importancia creciente de las redes sociales y el software empresarial de corte social (correo, chat,...) se explican por dos cosas: las conexiones entre la gente que las usa y la información que comparten e intercambian[42]. Igual que Facebook usa los datos extraídos por medio de sus herramientas analíticas sobre cómo se comportan los usuarios para facilitar la personalización y una mejor experiencia de usuario, el mismo fenómeno ha ocurrido en lo que podemos denominar empresas digitales. Contextualizar la interacción en entornos sociales y el reto mismo de hacerlo han ido empujando las fronteras de la tecnología disponible en los últimos años. Diferentes organizaciones en todo el espectro del negocio social están empezando a comprender la vasta cantidad de información que se puede deducir mediante el análisis de los millones de conversaciones que tienen lugar, la mayor parte de las veces de manera pública por los usuarios de medios sociales. Y si bien es cierto que hay problemas de privacidad, legalidad y regulación, incluso en redes sociales de uso interno, que tendrán que abordarse, no es probable que se demore mucho la adopción generalizada por parte de las organizaciones de dichas capacidades dado su valor potencial.

En el corto plazo la capacidad de analizar y de hacer minería de datos a la escala necesaria en las redes sociales está sacando a la luz a un conjunto de aplicaciones útiles que pueden conectarse en redes de medios sociales y utilizar el conocimiento derivado. Hacerlo bien, sin embargo, ha resultado ser poco trivial, como, por ejemplo, encontrar sentido a ciertos tipos de contenido que son muy opacos, como es el caso del vídeo de alta definición, o sensorizar las conexiones entre miles de mensajes en lenguaje natural con estructura indefinida. Todo esto requiere unas tecnologías nuevas que manejen estas enormes escalas de información, que hagan frente a las necesidades de computación de un modo eficiente en cuanto a costes y que cumplan unas exigencias crecientes de velocidad de respuesta.

La clave de las interacciones entre la gente en medios sociales es que deja un cierto conocimiento para que otros lo encuentren y lo vuelvan a utilizar. Esto puede ser el contenido

[42] Teófilo Redondo. *Big Data y la Interacción Social*. Economía Digital. Curso MOOC (F. Mochón, J.C. Gonzálvez y J. Calderón). Alfaomega. 2014

original que empezó la conversación o los comentarios posteriores así como la discusión, la valoración o los reenvíos (los conocidos «retuits» en Twitter). Estas conversaciones permanecerán en la red mucho tiempo después, normalmente para una eternidad digital, y conforman una historia valiosísima y un repositorio de conocimiento para la sociedad, la cultura y los negocios que podrán ser descubiertos, resucitados, compartidos o vueltos a utilizar para aprender. Por supuesto, parte de esta información no tiene un valor inherente en sí mismo. Encontrar lo que uno busca en el vasto océano de conversaciones humanas es una tarea complicada, pero este es el terreno de juego donde Big Data saca a relucir su potencial.

Pero el tema más profundo no es simplemente encontrar valor en la información enterrada en los medios sociales, sino más bien llegar a entender lo que es posible conocer a medida que los medios sociales se convierten en la forma predominante de comunicación. Uno de los objetivos de Big Data es ayudarnos a encontrar sentido en la inmensidad del flujo inacabable de nuestra actividad social.

En este sentido, las redes sociales de consumidores llevan mucho tiempo usando técnicas de Big Data en las plataformas construidas al efecto por los propios gestores de dichas redes sociales, siendo el caso de Google un ejemplo bien representativo de las posibilidades que abre este tipo de análisis.

Capitalizando las redes sociales: algunos retos a superar

Si se admite que las redes sociales contienen la suma visible de la comunicación humana y su interacción, el reto radica en encontrar y rastrear todo el conocimiento allí depositado, esto es, la inteligencia social del negocio. En este sentido, Big Data aporta dos aspectos novedosos.

En primer lugar, Big Data se aparta de métodos anteriores porque aplica nuevas formas de pensar sobre la captura, el almacenamiento y el procesado de cantidades ingentes de datos, precisamente el tipo de información que surge de los actuales ecosistemas de medios sociales. Esto también incluye la tecnología en la que se basa, por ejemplo redes para minería de datos o infraestructuras para MapReduce (modelo de programación utilizado por Google para dar soporte a grandes colecciones de datos), así como la arquitectura de software que es a menudo poco determinista y poco lineal en cuanto a su diseño. En la práctica esto significa que hay una clara división generacional y técnica entre el modo en que la mayoría de las organizaciones abordan el manejo de datos hoy en día y las cosas bien diferentes que se necesitan afrontar en el mundo de Big Data.

El segundo aspecto novedoso se relaciona con la idea de que «se gestiona solo lo que se puede medir» o lo que es lo mismo, lo que es medible es gestionable. El sentido de esta frase explica, tanto la importancia de la reciente explosión de datos digitales, como la relevancia creciente de Big Data[36]. Gracias a la enorme cantidad de datos disponibles y a las técnicas de Big Data, los directivos pueden conocer mucho mejor sus negocios y

traducir directamente ese conocimiento en una toma de decisiones más ajustada. Este tipo de análisis es lo que permite "detectar tendencias", "encontrar patrones" "saber lo que va a pasar antes de que ocurra", "adelantarse a la conversación y ver hacia dónde va". El análisis de sentimientos, la minería de conocimientos, la agregación de conversaciones hasta descubrir tendencias, todo esto es posible cuando se cuenta con información suficiente, se sabe lo que se hace y se dispone de la tecnología apropiada. Big Data introduce nuevas suposiciones para manejar la información y comprender el conocimiento encerrado en enormes cantidades de datos.

Acercándonos a la realidad

En el artículo de José Julio López «Las principales Técnicas Big Data y sus Aplicaciones», publicado el 11 de julio de 2014[43] se ofrece una introducción al dinamismo de las técnicas utilizadas en Big Data.

- ***Reflexione sobre la importancia de estar al día con la evolución de la tecnología y la importancia de la especialización.***

2.1.3 Big Data y el Rendimiento de las Empresas: Algunos Casos

La explotación de nuevos y enormes flujos de información puede mejorar radicalmente el rendimiento de las organizaciones. En este sentido, Big Data puede ser una relevante fuente de generación de valor, cuya importancia crecerá a medida que el mundo se va haciendo más y más social. Pero el hecho a destacar es que las ventajas competitivas serán una realidad para los que comprendan lo que está sucediendo mejor que sus homólogos y lo logren conectar directamente a los resultados de su negocio y a otros propósitos igualmente útiles. Paralelamente aquellos que no logren incorporar a sus procesos de toma de decisiones estás nuevas tecnologías competirán en inferioridad de condiciones y su supervivencia puede que se vea amenazada.

2.1.3.1 El caso de la venta de libros al por menor

Consideremos la venta al por menor de libros. Los libreros tradicionales, en sus tiendas físicas siempre podían rastrear qué libros habían vendido y los que no lograban vender. Si tenían un programa de fidelización y seguimiento de los clientes, podían ligar algunas de esas compras a clientes individuales. Y eso era todo. Una vez que las compras se han desplazado a las tiendas online, sin embargo, el conocimiento de los clientes ha aumentado radicalmente.

Los minoristas que venden libros online pueden rastrear no sólo lo que compran los clientes, sino también qué más han mirado; cómo navegan a través del sitio; en que medida

[43] José Julio López. *Las principales Técnicas Big Data y sus Aplicaciones*. https://josejuliolopezsantos.wordpress.com/2014/07/11/las-principales-tecnicas-big-data-y-sus-aplicaciones/. [Online; publicado el 11 de julio de 2014]

se ven influenciados por las promociones, las opiniones de otros clientes o por los diseños de página; así mismo, pueden analizar las similitudes entre los clientes individuales y los grupos de clientes. Cuando cuentan con una cierta cantidad de información y experiencia, desarrollan algoritmos para predecir qué libros les gustaría leer próximamente a los clientes individuales. Estos algoritmos van evolucionando conforme se añade información adicional, de forma que funcionan mejor cada vez que el cliente responde a una recomendación o la ignora. Los minoristas tradicionales simplemente no podían acceder a este tipo de información, y mucho menos actuar en base a ella de una manera apropiada. Por ello, no es de extrañar que Amazon haya hecho que cierren tantas librerías tradicionales.

Un hecho a destacar es que si bien se tiende a pensar que las empresas que son nativas digitales, como es el caso de Amazon, sean capaces de lograr cosas que los directivos hace unos años no podían ni soñar, la realidad es que Big Data también tiene la capacidad de transformar a las empresas tradicionales. Cualquier empresa puede obtener ventajas competitivas usando Big Data y puede medir y gestionar con mayor precisión que nunca. Puede hacer mejores predicciones y tomar decisiones más inteligentes. Puede orientar sus intervenciones de forma más eficaz, y puede hacerlo en temas que hasta ahora han estado dominados por el instinto y la intuición, más que por los datos y el rigor. De hecho, Big Data está contribuyendo a cambiar el valor de la experiencia y la práctica de la gestión organizacional. Por ello, los líderes empresariales están empezando a considerar Big Data como una revolución de la gestión, y como toda revolución no está exenta de retos[36].

La pregunta que nos formulamos es si hay evidencia de que el uso de Big Data de forma inteligente mejorará el rendimiento del negocio. En este sentido, un estudio dirigido por el MIT Center for Digital Business ha investigado en qué medida las empresas basadas en datos obtienen mejores resultados que las empresas que no utilizan Big Data[36]. Los resultados de este trabajo se pueden sintetizar como sigue: las empresas que ocupan posiciones en el tercio superior de su industria en el ranking según el uso de la toma de decisiones basada en datos fueron, en promedio, son un 5 % más productivas y un 6 % más rentables que sus competidores.

Para ilustrar lo señalado vamos a analizar los casos de dos empresas que han empleado Big Data con éxito. Una utiliza grandes volúmenes de datos para crear nuevos negocios y la otra para generar más ventas[36].

2.1.3.2 El caso de PASSUR: mejorando la estimación del tiempo de llegada

El dicho «El tiempo es oro», es completamente cierto para las aerolíneas. Las aerolíneas le prestan suma atención a las estimaciones de los tiempos de llegada de los vuelos a los aeropuertos. Si un avión aterriza antes que el personal de tierra esté preparado para ello, los pasajeros y la tripulación se verán de hecho atrapados en el avión. Por otro lado, si un avión aterriza más tarde de lo esperado, el personal del aeropuerto se encontrará

inactivo, lo que elevará los costes, y además los viajeros se sentirán molestos por la falta de puntualidad. Por eso, cuando una de las principales aerolíneas estadounidenses tuvo conocimiento, gracias a un estudio interno que alrededor del 10 % de los vuelos con destino a su principal centro (*hub*) habían tenido al menos una diferencia de 10 minutos entre la hora estimada de llegada y el tiempo real de llegada y el 30 % de los vuelos habían tenido una diferencia de por lo menos cinco minutos, se decidieron a tomar medidas.

Hasta ese momento, la aerolínea se basaba en la práctica tradicional de la industria de la aviación, consistente en la utilización de los tiempos estimados de llegada, esto es, los ETA (*Estimated Time of Arrivals*) proporcionados por los pilotos. Los pilotos realizaban estas estimaciones durante la fase de aproximación final al aeropuerto, cuando tenían muchas otras tareas que realizar que requerían su tiempo y atención. Para tratar de encontrar una solución mejor que los ETA facilitados por los pilotos, la aerolínea decidió contar con la colaboración de PASSUR Aerospace, proveedor de tecnologías de soporte para la toma de decisiones en la industria de la aviación. En 2001 PASSUR comenzó a ofrecer sus propias estimaciones de la llegada de los vuelos.

PASSUR realiza sus estimaciones de los tiempos de llegada combinando una amplia base de datos públicos sobre el clima, horarios de vuelo, y otros factores, con los datos de la propia empresa y los datos suministrados por una red de estaciones de radar pasivos que habían sido instalados cerca de los aeropuertos para recopilar datos sobre cada aeroplano en el cielo local. PASSUR comenzó con sólo unas pocas de estas instalaciones, pero antes de 2012 tenía más de 155. Cada 4,6 segundos recoge una amplia gama de información sobre todos los aeroplanos que «ve». Esto produce un enorme y constante flujo de datos digitales. La compañía conserva todos los datos que ha recogido a lo largo de los años, por lo que tiene una inmensa cantidad de información multidimensional que abarca más de una década. Estos datos permiten llevar a cabo sofisticados análisis y comparar los patrones. Las estimaciones del tiempo de llegada, comercializadas con el nombre de Right ETA esencialmente funcionan formulándose dos preguntas: «¿Qué ha pasado todas las veces anteriores que un avión se aproximaba este aeropuerto en estas condiciones? ¿Cuándo aterrizó en realidad?» Las estimaciones se derivan de algoritmos que son alimentados por múltiples fuentes de datos en tiempo real. El Right ETA incluye la hora estimada de llegada, la identificación del vuelo y la hora de llegada efectiva.

Después de cambiar a Right ETA, la aerolínea prácticamente eliminó las brechas entre los tiempos estimados y reales de llegada. PASSUR cree que la información facilitada por el Right ETA, al permitir a una aerolínea saber cuándo sus aviones van aterrizar y planificar en consecuencia, tiene un valor de varios millones de dólares al año en cada aeropuerto. La moraleja en este caso es simple: la utilización de Big Data permite mejores predicciones, y mejores predicciones generan mejores decisiones que permiten ahorrar costes y molestias a los pasajeros.

2.1.3.3 Sears Holdings: Promociones más rápidas y más personalizadas

Hace unos años, Sears Holdings llegó a la conclusión de que era necesario generar un mayor valor de las enormes cantidades de datos sobre los clientes, productos y promociones que recoge de sus marcas Sears, Craftsman y Lands' End. Los directivos de Sears Holdings pensaron que sería útil combinar y hacer uso de todos estos datos para adaptar promociones y otras ofertas a los clientes, y para personalizar las ofertas aprovechando las condiciones locales. El proyecto parecía interesante, pero difícil de llevar a cabo. Los procesos de la compañía no eran muy ágiles y la información, además de ser muy voluminosa, estaba muy dispersa y circulaba con bastante lentitud entre las distintas marcas y entre los diferentes departamentos. Baste señalar que Sears necesitaba aproximadamente ocho semanas para generar promociones personalizadas. El resultado era que cuando las promociones estaban listas para ofertarse a los clientes, en muchos casos ya no eran lo más adecuado para la empresa.

Ante esta situación los responsables de Sears Holdings, en un principio buscando simplemente llevar a cabo promociones y personalizar sus ofertas de forma rápida y eficiente, decidieron acudir al Big Data. Uno de sus primeros pasos consistió en crear un *cluster* de Hadoop, esto es, un grupo de servidores genéricos de bajo costo, cuyas actividades son coordinadas por un marco de software novedoso llamado Hadoop.

Además decidieron dejar de tener los datos en silos en muchos lugares, e iniciaron un proceso tendente a agrupar todos los datos en un solo lugar (el cluster) con el fin de tener en un solo punto toda la información sobre el cliente. Sears comenzó a usar el clúster para almacenar los datos entrantes de todas sus marcas y para mantener los datos de los almacenes de datos existentes. A continuación, realizaron análisis en el clúster directamente, evitando las complejidades que consumen mucho tiempo debido a la extracción de datos de varias fuentes y la necesidad de integrarlos, para que puedan ser analizados. Esta implantación gradual, al final ha logrado unos excelentes resultados pues el tiempo necesario para generar un conjunto integral de promociones se ha reducido de ocho semanas a una semana, y sigue reduciéndose. Y estas promociones son de mayor calidad, debido a que son más oportunas y más granulares, y más personalizadas.

Una vez implantado Big Data, Sears Holdings combina y mezcla gran cantidad de datos para establecer precios personalizados en tiempo casi real. Los datos de información del producto, las condiciones económicas locales, los precios de la competencia, etc. se combinan y se analizan mediante un algoritmo que utiliza la elasticidad de los precios. De esta forma Sears Holdings puede fijar el mejor precio para el producto adecuado, en el momento y lugar precisos a través de cupones personalizados que se dan a los compradores leales y también se utilizan para «actualizar» el inventario si se considera necesario.

Paralelamente, en los últimos años, Sears Holdings ha pasado de llevar a cabo estrategias de precios a nivel nacional, a hacerlas a nivel regional y también personal. Los cupones que reciben los clientes se basan en su lugar de residencia, la cantidad de productos

que están disponibles, así como los productos que se quieren quitar del inventario y qué productos los técnicos de Sears creen que les van a gustar a los clientes y en consecuencia van a comprar.

Además de para realizar promociones y personalizar sus oferta, Sears Holdings también aplica Big Data para combatir el fraude, realizar el seguimiento de la eficacia de las campañas de marketing, optimizar la fijación de precios y controlar mejor la cadena de suministro[44].

Un último hecho a señalar es lo fácil que fue para Sears la transición de lo viejo a los nuevos enfoques de gestión de datos y análisis de alto rendimiento. En cualquier caso, dado que las habilidades y conocimientos relacionados con las nuevas tecnologías de datos eran tan raras en 2010, cuando Sears comenzó la transición, se contrataron algunos de los trabajos a una empresa especializada en servicios de software basados en Hadoop. Pero con el tiempo la vieja guardia de TI y los profesionales de la analítica han empezado a sentirse cómodos con las nuevas herramientas y enfoques y a asumir progresivamente las nuevas actividades.

2.1.3.4 Otros ejemplos de los logros de Big Data

Los ejemplos PASSUR y Sears Holding ilustran el poder del Big Data, pues permite predicciones más precisas, mejores decisiones e intervenciones selectivas, y estos logros se pueden alcanzar a escala aparentemente ilimitada. Para justificar estas afirmaciones, seguidamente vamos enumerar algunos ejemplos de los logros alcanzados mediante la utilización de Big Data en una amplia y variada gama de actividades:

- En la gestión de la cadena de suministro, al lograr que las tasas de productos defectuosos de un fabricante de automóviles se reduzcan radicalmente.
- En el servicio al paciente, al permitir analizar e intervenir en las prácticas de cuidado de la salud de millones de personas.
- En la planificación organizacional, al poder, por ejemplo, anticipar mejor las ventas online sobre la base de un conjunto de características del producto.
- En las finanzas, al permitir una gestión en tiempo real y con plena información de las carteras de inversión.
- En el marketing, al hacer posible un seguimiento exhaustivo del comportamiento de los clientes.
- En cadenas hoteleras, al facilitar la gestión de los juegos de azar.

[44] Mark van Rijmenam. *Sears Became a Real-Time Digital Enterprise Due to Big Data*. https://datafloq.com/read/sears-real-time-digital-enterprise-big-data/265. [Online; publicado el 7 de julio de 2012]

- En la gestión de los recursos humanos, al lograr disponer de forma inmediata de una muy amplia información de todos los empleados que, por ejemplo permite una eficaz gestión de los turnos.
- En la reparación de máquinas, al facilitar ajustar eficientemente los tiempos de mantenimiento.
- En la agricultura, al hacer previsiones anuales muy precisas sobre el número de litros de agua por metro cuadrado en cada una de las parcelas de un viticultor, ayudándole a tomar decisiones menos arriesgadas.

Estos no son más que unos simples ejemplos que permiten ilustrar que casi ningún ámbito de la actividad empresarial se mantendrá al margen del Big Data.

Acercándonos a la realidad

En el artículo «Las cinco principales aplicaciones de Big Data»[45] se señala la importancia de conocer cuáles son los problemas principales de las organizaciones a los que Big Data puede aportar una solución.

❖ ***Reflexione sobre las cinco principales aplicaciones de Big Data presentadas en el artículo y piense en otros posibles problemas a cuya solución Big Bata podría contribuir.***

2.1.3.5 Triunfos de Big Data ante análisis más elaborados

En ocasiones, alguien que viene de fuera de una industria puede detectar una mejor manera de utilizar Big Data que alguien de dentro de la industria pues el análisis de tantas fuentes nuevas e inesperadas de datos puede que se lleve a cabo de forma más incisiva y creativa, con una mentalidad nueva y abierta. Esto, sin embargo, no obvia la conveniencia de dotarse de un conocimiento profundo del sector antes de aplicar Big Data.

Predicciones del mercado de la vivienda

Un caso que puede ilustrar esta posibilidad es el trabajo realizado por Lynn Wu y Erik Brynjolsson[46]. Estos investigadores no tenían ningún conocimiento especial del mercado de la vivienda y decidieron utilizar datos de acceso público de web para predecir las alteraciones en los precios de la vivienda en las áreas metropolitanas en los Estados Unidos. Su hipótesis de partida era que datos en tiempo real permitirían realizar buenas previsiones a corto plazo sobre el mercado de la vivienda, y los hechos vinieron a darle la razón.

[45] Lantares. *Las Cinco Principales Aplicaciones de Big Data*. http://www.lantares.com/blog/las-cinco-principales-aplicaciones-de-big-data. [Online; consultado el 8 de diciembre de 2015]

[46] Lynn Wu y Erik Brynjolfsson. «The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales». En: *Economic Analysis of the Digital Economy*. National Bureau of Economic Research, Inc, 2014, págs. 89-118

De hecho, sus predicciones resultaron ser más precisa que las de la National Asociation of Realtors (NAR), que había desarrollado un modelo mucho más complejo pero que se basaba en datos históricos que reflejan cambios relativamente lentos. Wu y Brynjolfsson tomaron el mercado de la vivienda simplemente como un ejemplo para evidenciar cómo la búsqueda online puede utilizarse para analizar el presente de determinadas actividades económicas y predecir las tendencias económicas futuras.

La mayoría de las fuentes de datos utilizadas en economía, ya sea del gobierno o de las empresas, están típicamente disponibles con un retraso sustancial respecto a la fecha en que se recoge la información, con un alto nivel de agregación, y sólo para aquellas variables que se han especificado y recogido previamente. Esto reduce la eficacia de las predicciones en tiempo real. La investigación realizada por Lynn Wu y Erik Brynjolfsson evidencia como los datos procedentes de los motores de búsqueda, por ejemplo de Google, pueden proporcionar una manera muy precisa, y simple para predecir las actividades futuras de los negocios.

En el caso concreto del mercado de la vivienda llevado a cabo por Lynn Wu y Erik Brynjolfsson, la aplicación de Big Data para predecir las tendencias del mercado inmobiliario permitió elaborar un índice de búsqueda de vivienda que es altamente predictivo de las futuras ventas del mercado inmobiliario y de los precios. En concreto, cada punto porcentual de aumento en el índice de búsqueda de vivienda se correlaciona con ventas adicionales de 67.220 casas en el próximo trimestre. El uso de datos provenientes de los motores de búsqueda produce predicciones con un error medio absoluto de sólo 0,102, lo que supone una mejora sustancial sobre el 0,441 de error medio absoluto del modelo de referencia que utiliza datos convencionales pero no incluye ningún dato de búsqueda.

Las implicaciones de este ejemplo se pueden sintetizarse como sigue; los economistas, los políticos y los inversores están esperando los datos gubernamentales publicados cada mes para evaluar el mercado de la vivienda actual y predecir su recuperación. Sin embargo, los datos del gobierno se publican a menudo con un retraso de meses, lo que conlleva un retraso en la evaluación de las condiciones económicas actuales. En la investigación de Lynn Wu y Erik Brynjolsson se propone una forma diferente de predecir el futuro precio de la vivienda a través de la frecuencia de las búsquedas online. Analizando las búsquedas de los consumidores según lo revelado por sus comportamientos, es posible descubrir las tendencias de ventas antes de que se publiquen los datos.

Desde una perspectiva más general puede afirmarse que Big Data, con este tipo de microdatos (búsquedas individuales online) pueden contribuir a transformar la predicción en numerosos mercados, y por lo tanto los negocios y la toma de decisiones del consumidor.

Predicción de brotes de gripe con Google Flu Trends

Otro ejemplo, en un principio sorprendente de la bondad de las predicciones basadas en modelos simples de Big Data frente a otras basadas en modelos más elaborados, realiza-

das por centros con mucha experiencia y prestigio es el caso siguiente. Investigadores de la Escuela de Medicina Johns Hopkins descubrieron que podían usar los datos de Google Flu Trends (un agregador gratuito de temas de investigación relevantes, a disposición del público) para predecir los aumentos repentinos en las visitas a las urgencias relacionadas con la gripe una semana antes que el Center for Disease Control (CDC) lo advirtiera[47].

La primera pregunta que se formularon los investigadores de la Escuela de Medicina Johns Hopkins fue si era conveniente realizar un seguimiento de las tendencias de la gripe en su comunidad. Para llevar a cabo su estudio, los investigadores analizaron conjuntamente los datos específicos de Baltimore de la página web de Google Flu Trends, que estima los brotes de gripe en base a búsquedas online de información sobre la gripe, datos del departamento de emergencias y ciertas estadísticas del laboratorio del Hospital Johns Hopkins.

Usando Google Flu Trends, los investigadores encontraron que el número de búsquedas online de información sobre la gripe aumentó, al mismo tiempo que la unidad de emergencias pediátricas del hospital experimentó un notable crecimiento de los casos de niños con síntomas de gripe. Los datos de Google Flu Trends presentaron una correlación moderada con el volumen de pacientes en la unidad de emergencias de adultos. Google Flu Trends mostró un aumento en los casos de gripe de siete a 10 días antes que el CDC.

En base en los resultados, los investigadores sugirieron que plataformas como Google Flu Trends podrían ayudar a los administradores del hospital a anticipar los brotes de gripe y tomar decisiones de personal y planificar la capacidad de forma apropiada.

Acercándonos a la realidad

Como se señala en el texto, en ocasiones las predicciones obtenidas con modelos simples de Big Data resultan ser más precisas que las derivadas de informes elaborados por entidades oficiales de prestigio que emplean modelos complejos. Analiza el caso de la actividad en Twitter en comparación con los informes oficiales en el seguimiento de la propagación del cólera en Haití tras el terremoto de enero de 2010 contenido en el artículo «Daily-Briefing. Google beats the CDC: Web tool predicts flu-related ED surge» publicado en enero de 2012[47].

- ***¿Se le ocurren ideas para poder llevar a cabo investigaciones similares?***

Acercándonos a la realidad

Suponga que trabaja en el departamento de tecnologías de la información en un gran almacén que se caracteriza por llevar una gestión de la información basada en los métodos tradicionales ¿Qué razonamientos emplearía y que tipo de aplicaciones les propondría

[47] Advisory. *Google Beats the CDC: Web Tool Predicts Flu-Related ED Surge.* https://www.advisory.com/Daily-Briefing/2012/01/13/Google-flu. [Online; publicado el 13 de enero de 2012]

realizar a sus jefes para convencerlos para que se decidiesen a implantar Big Data en su empresa?

2.2 Big Data: Retos y Oportunidades

Una vez presentadas las limitaciones de Big Data y las cautelas con que se debe utilizar esta potentísima herramienta, esta sección se centra en el análisis de los retos que plantea y en esbozar las oportunidades que ofrece. El mensaje que se trasmite es que los beneficios que se pueden derivar de emplear Big Data sobrepasan ampliamente cualquier problema que se pueda plantear.

2.2.1 Retos a Superar

La gestión eficiente de la cantidad de datos disponibles y los que cada día se generan supone un gran reto y una enorme oportunidad. Los datos son más importantes que nunca, pero su crecimiento exponencial, la complejidad de los formatos y la velocidad de entrega puede desbordar la capacidad de las empresas para gestionarlos de forma tradicional y obtener beneficios a partir de ellos[48].

2.2.1.1 Retos tecnológico-analíticos

Como reflexión general sobre los desafíos tecnológicos que conlleva la implementación de Big Data a gran escala cabe apuntar, en primer lugar, un problema de madurez de las soluciones. Aún existe un número relativamente limitado de grandes desarrollos de soluciones Big Data en la empresa. Muchos de los grandes desarrollos de empresa se encuentran en etapas piloto[40]. En segundo lugar, hay que admitir que el gran reto tecnológico se deriva de que la velocidad, el volumen y la variedad de los datos no dan muestras de reducir su ritmo, lo que plantea la necesidad de que las soluciones tecnológicas estén en un proceso de continua mejora.

Desde una perspectiva estrictamente tecnológica los tipos de retos a superar son de recolección de datos, almacenamiento, procesamiento y obtención de valor o análisis.

- **Recolección**: Los retos que se plantean en la fase de recolección son de distinta naturaleza según se trate de fuentes de tipo online u offline. En cualquiera de los casos, habrá que desarrollar unos sistemas de extracción de datos que resulten coherentes con los sistemas de almacenamiento (véase capítulo 5). Otro hecho a destacar es que es necesario conocer el ciclo de vida del tipo de dato (véase el capítulo 1) que recogemos hasta que lo tenemos que introducir en la base de datos. Esto no es una obviedad, una las causas más extendidas de

[48] Jordi Torres. *Retos del Big Data*. http://es.slideshare.net/jorditorres/retos-del-big-data. [Online; publicado el 2 de mayo de 2012]

pérdida de tiempo dentro del equipo de trabajo suele tener su origen en no haber tratado los datos bien desde el principio[49].

Por ello debemos asegurarnos que el *dataset* (fuente de origen completa de datos) que vamos a tratar esté consolidado correctamente. Una de las primeras tareas del científico de datos debe consistir es «limpiar» todos los datos y dejar los campos bien representados. La base de datos son los cimientos sobre los que se debe construir el análisis: si estos tienen deficiencias el edificio acabará teniendo grietas.

- **Almacenamiento**: hacen falta nuevas tecnologías de almacenamiento más baratas (ver capítulo 6). Además, las bases de datos relacionales no pueden recoger toda la información.

- **Procesamiento**: se requieren nuevos modelos de programación (ver capítulo 7). Para conseguir procesar grandes conjuntos de datos Google creó el modelo de programación MapReduce. Pero fue el desarrollo de Hadoop MapReduce, por parte de Yahoo, el que ha propiciado un ecosistema de herramientas *open source*.

- **Análisis**: los datos necesitan ser analizados para sacarle valor; la información no es conocimiento «accionable». Para ello se cuenta con técnicas de Data Mining. Pero muchos de los algoritmos de conocimiento ejecutan bien en miles de registros, pero son difícilmente aplicables cuando se tienen miles de millones. (véase capítulo 8).

2.2.1.2 ¿Datos excesivos o falta de ingenio?

Pero la velocidad, el volumen y la variedad de los datos no sólo plantean los retos tecnológicos citados también otros relacionados con la propia abundancia de datos ¿Hay demasiados datos? Los datos son algo muy positivo pero sólo si se cuenta con los medios para convertir estos datos en información útil y pertinente. Un problema típico a abordar mediante Big Data contiene una gran cantidad de variables, un montón de relaciones entre ellas, y por lo tanto se necesita una gran cantidad de operaciones para llegar a una solución. Podemos medir el tamaño del problema que genera Big Data por el número de operaciones necesarias para convertir Big Data en información útil.

Las empresas tienen que responder a dos tipos de preguntas:

- ¿Qué sucederá si hacemos esto?

- ¿Cuál es la mejor manera de conseguir que esto suceda?

[49] Francisco Javier Pulido. *La Etapa de Obtención de Datos en BigData (II) - Retos a Superar*. http://www.franciscojavierpulido.com/2014/03/la-etapa-de-obtencion-de-datos-en_10.html. [Online; publicado el 16 de diciembre de 2014]

Estos problemas toman Big Data como una entrada y generan información útil –planes, estrategias, modelos– como salidas. Pero conforme el tamaño del input se incrementa, el tamaño del problema no sólo va aumentar, sino que va a mega-aumentar[35]. La relación entre el número de variables de entrada y el número de operaciones necesarias es altamente no lineal. La mayoría de los problemas a los que enfrentan las empresas son técnicamente difíciles de tratar, porque el tamaño del problema aumenta exponencialmente con el de las entradas. Piense en el análisis del patrón de relaciones generadas por las redes personales en LinkedIn con el fin de predecir quién va a introducir a quién; o, el patrón de las decisiones de compra en Amazon que la gente en la misma red de Facebook hace, para poder predecir quienes son los influyentes clave.

Para resolver las ecuaciones que resultarán del modelado de estos patrones, además de necesitar una enorme cantidad de memoria para que sean «visibles» los datos, son muchas las operaciones que se tienen que realizar. Para atender a estas necesidades la gran la potencia de cálculo que tenemos en nuestro planeta sólo es suficiente "en parte". Si sólo se requiriese cálculo bruto el problema Big Data no sería nada más que un problema de diseño de chips y de computación: darles más FLOPS (operaciones de coma flotante por segundo o *floating point operations per second*). Pero, a pesar de que la inteligencia computacional colectiva está creciendo rápidamente, el problema Big Data no es nada fácil. De hecho, la capacidad de análisis de datos se presenta como el mayor factor diferenciador para las empresas en todo el mundo.

Big Data es un problema difícil de resolver aunque se supere la Ley de Moore y se consiga un número cada vez mayor de FLOPS, porque para afrontar los problemas a los que se enfrenta Big Data la inteligencia no es suficiente. El tamaño de los problemas tipo de Big Data no crece linealmente con el número de variables –o puntos de datos– que estamos teniendo en cuenta. Crece de manera exponencial, y por lo tanto más rápidamente de lo que la Ley de Moore sugiere. Para explicitar intuitivamente la naturaleza del problema acudamos a un caso.

El caso de la ruta óptima de una vendedora en Canadá

Considere el problema al que se enfrenta una vendedora que está tratando de cubrir en coche las 4.663 ciudades de Canadá en el menor tiempo posible. ¿Es un problema muy difícil? Bueno, es increíblemente difícil de resolver por métodos de fuerza bruta con los que trabajan los equipos tradicionalmente disponibles. Tendría que considerar unas 4.663 permutaciones dadas las ciudades que componen los caminos alternativos, que con los dispositivos computacionales actuales tendrían que estar funcionando a 10^{12} operaciones por segundo, lo que supondría alrededor de $1,6 \times 10^{1.383}$ años[35].

Lo que se necesita aquí no es inteligencia –la capacidad de hacer muchas operaciones en un corto período de tiempo– sino ingenio –la capacidad de encontrar y aplicar el mejor método de búsqueda de solución. Y, resulta que hay un método por el cual se puede encontrar la ruta más corta en unos seis minutos de tiempo de CPU en su PC con tecno-

logía Intel de escritorio, mediante el uso de un método de búsqueda inventado por Lin y Kernighan hace unos 40 años[50]. El algoritmo de Lin-Kernighan se basa en hacer una primera aproximación a un camino creíble, y luego hacer pequeños cambios en este camino, evaluando el resultado.

Como se indica en el ejemplo anterior, el ingenio es diferente de la inteligencia. No es la capacidad de realizar operaciones de forma rápida, sino la capacidad de diseñar la forma en que se secuencian las operaciones –el diseño de los algoritmos y heurísticas que utilizamos para resolver los grandes problemas que surgen de Big Data.

En una era en la que la inteligencia computacional en bruto ha crecido a pasos agigantados y amenaza con afectar a nuestra capacidad de crear dispositivos cada vez más pequeños, el ingenio se plantea como la barrera para el análisis de Big Data. Es el ingenio el que nos permite optimizar grandes plantas de fabricación, calcular interacciones críticas gen-gen en el genoma de individuos de alto riesgo, calcular el valor en riesgo (*value at risk*) durante la noche de grandes carteras diversificadas con varios niveles de liquidez, madurez y volatilidad o de diseñar *matching markets* para los médicos residentes e internos en el sistema médico estadounidense.

Acercándonos a la realidad

En ActiBva.com *Loslunesalsol* publicó el 24 de julio de 2013 el Artículo «Big Data, claves de negocio y retos oportunidades»[51]. En concreto se presentan algunos de los retos tecnológicos de debe superar Big Data.

- ***Reflexione sobre los retos tecnológicos y cuáles cree que son los temas más difíciles de superar.***

2.2.1.3 Los usuarios y la propiedad de los datos

Un tema controvertido, un reto por resolver, es la relación entre los usuarios y la propiedad de los datos personales. En general, nadie debería trabajar con datos personales sin pensar en las implicaciones sobre la privacidad, porque los usuarios son personas. Por otro lado, la cantidad de datos está aumentando a un ritmo más rápido del que las organizaciones pueden gestionar la seguridad de los datos de forma correcta (véase capítulo 10).

Conforme hay más datos disponibles, el uso aceptable de los datos personales se convierte en una gran preocupación para los usuarios. Por un lado, el enorme rastro que diariamente dejan los usuarios desde sus ordenadores y dispositivos móviles plantea du-

[50] Shen Lin y Brian W. Kernighan. «An Effective Heuristic Algorithm for the Traveling-Salesman Problem». En: *Operations Research* 21.2 (1973), págs. 498-516

[51] Loslunesalsol. *Big Data, Claves de Negocio y Retos Tecnológicos*. http://www.actibva.com/magazine/mas-que-economia/big-data-claves-de-negocio-y-retos-tecnologicos. [Online; publicado el 24 de julio de 2013]

das sobre la propiedad de la información y los datos propios. Una corriente de opinión, defendida entre otros por el inventor de la Web, Tim Berners-Lee, es que la propiedad sea de los usuarios[52].

Así pues, ante la pregunta ¿qué hacer con los enormes volúmenes de datos que genera la economía digital? La respuesta de Tim Berners-Lee es que no deben entregarse al dominio de terceros para recibir publicidad personalizada. Las personas deberían tener la propiedad sobre sus datos y la libertad de poder usarla cruzándola con toda la información de actividad personal que se puede recoger. Estos datos, que pueden combinar registros de desplazamientos, con otros de salud, o sobre las preferencias de consumo, pueden componer un conjunto más completo y relevante para el usuario que tener bloques de datos disgregados.

Tim Berners-Lee también defiende la posibilidad de que cualquier persona permita voluntariamente a otros individuos u organizaciones que accedan a sus registros de actividad recolectados online. Bajo esta perspectiva, en lugar de buscar más alternativas para encriptar y bloquear el acceso a la información personal, se deberían buscar herramientas para saber los datos que produce un individuo online y cómo combinarlos con otros de forma útil[53]. Este sistema integrado de información cruzada está ligado al principio de web semántica promovido desde el Consorcio W3C, fundado por Berners Lee. Este consorcio defiende la apertura y colaboración colectiva en la red, a la vez que los derechos individuales. En el mundo virtual se deben construir sistemas que den cabida a la privacidad, pues la gente tiene el derecho a saber cómo están usando sus datos[52].

En cualquier caso, las cosas no son tan simples como parece, pues cuando se regula tratando de proteger los derechos de los usuarios puede que estemos limitando iniciativas empresariales innovadoras, sobre todo cuando en unos países se implanta una regulación más proteccionista que en otros. Así, consideremos el caso de una empresa impulsada por un joven emprendedor español, Socialtech[54]. Esta empresa ha creado Wordfeeling, una plataforma web que permite a las empresas gestionar y monitorizar su presencia en las redes sociales (Facebook, Twitter, Instagram y Foursquare), así como en Internet o en medios de comunicación digitales. Esta plataforma utiliza, entre otros mecanismos, el análisis del lenguaje natural de los comentarios. La herramienta puede aplicar la analítica predictiva en relación, por ejemplo, a un cambio de comportamiento del consumidor. A través del estudio de actos anteriores de ese usuario (cómo ha reaccionado ante casos similares) se puede hacer una predicción probabilística de los siguientes pasos que va a realizar.

---

[52]Lina María Aguirre. *Tim Berners-Lee: Algo Mejor que 'Big Data'*. http://blogs.lavanguardia.com/tecladomovil/tim-berners-lee-algo-mejor-que-big-data-49131. [Online; publicado el 10 de octubre de 2014]

[53]Alex Hern. *Sir Tim Berners-Lee Speaks Out on Data Ownership*. The Guardian. [Publicado el 8 de octubre de 2014]

[54]Gina Tosas. *El Big Data es un Problema Más que una Tecnología*. La Vanguardia. [Publicado el 4 de junio de 2014]

Pues bien, esta empresa que está en las fronteras de la investigación, se ha encontrado en España con un obstáculo al que no ha podido hacer frente y que le supone tener que emigrar a los Estados Unidos. El obstáculo insalvable es la regulación española (y europea) en relación a la protección de datos. Según los responsables de la empresa, las certificaciones que se necesitan para operar en España no las podrán obtener porque se considera ilegal su modelo de negocio basado en la descarga de información de redes sociales. El tema de fondo es que la legislación española (y europea) no considera que Internet, Twitter o Facebook sean fuentes de datos abiertas al público. Así pues el debate sobre la propiedad de los datos es un reto que no está plenamente resuelto.

Acercándonos a la realidad

El enorme rastro que diariamente dejan los usuarios desde sus ordenadores y dispositivos móviles plantea una gran preocupación por el hecho de que emerja un escenario propio del Gran Hermano, de vigilancia del comportamiento y del conocimiento de la gente, y utilizarlo con fines para los que no fue concebido al principio. Esto ya es un tema de interés en los medios sociales, y la automatización y madurez crecientes de Big Data, medios sociales y herramientas analíticas lo harán más acusado. Los que se dediquen a navegar y rastrear, analizar y compartir datos extraídos de medios sociales tendrán que considerar cuidadosamente todas las implicaciones.

- ***¿En qué sentido las empresas deberán jugar el papel de guardianes en tanto que todos los temas sobre propiedad intelectual o propiedad de los datos sean resueltos por los reguladores y la propia industria?***

2.2.1.4 Retos de cara a la gestión

Como se ha señalado, el coste de la captura y almacenamiento de datos nunca ha sido más bajo. Las nuevas plataformas analíticas llave en mano, basadas en la nube hacen que poner en marcha una plataforma y lograr su rentabilidad sea más viable que nunca. A esto se le une que el potencial de las nuevas herramientas de cálculo propicia que a las organizaciones cada vez les resulte más fácil optar por utilizar todos los datos, no sólo una muestra. Pero esta profusión de datos plantea nuevos retos de gestión[23]:

- **Idoneidad de los micro-datos.** Hay que determinar si los datos micro-conductuales disponibles representan con precisión el marco más amplio de decisiones implicadas en la situación. El paso de contar con datos de carácter micro económico, basados en observaciones individuales, a una realidad macroeconómica, puede plantear problemas de agregación.

- **Escasez de expertos.** Se deben encontrar expertos en Big Data, con conocimientos de estadística e informática y con la capacidad de entender la naturaleza del negocio. Las estrategias tecnológicas y los procesos de negocio para Big Data

son muy diferentes. Las lagunas en el almacenamiento de datos y las estrategias de procesamiento, así como la falta de conocimiento o dirección por parte de los gerentes de sistemas pueden generar dudas en algunas organizaciones. Además, los profesionales de muchas organizaciones todavía carecen de conocimiento sobre las herramientas de gestión de Big Data. La formación técnica y de usuario final también pueden dificultar que algunas organizaciones adopten Big Data.

- **Seleccionar la información relevante**. Hay que decidir cómo gestionar toda la información que tenemos disponible y qué preguntas son las que se pueden abordar con rigor a partir de los datos. Hay que pensar cuidadosamente sobre qué temas se va a centrar el análisis y cómo se van a enmarcar los problemas. Desde este punto de vista, la abundancia de datos plantea un reto de gestión. Es un problema que tienen las empresas y los usuarios, pues se genera más información de la que somos capaces de tratar en tiempo real y en tiempo de gestión de negocio.

- **Alinear Big Data y el negocio**. Es necesario alinear el Big Data con el negocio. Es necesario establecer pasarelas de comunicación entre los científicos de datos y los gestores de la organización. Hay que ser conscientes que pueden surgir dificultades derivadas de la no siempre fácil alineación de la estrategia de implantación del Big Data con el negocio. Los principales directivos de las organizaciones suelen tener unos objetivos empresariales muy claros, pero puede que estos objetivos no estén alineados con las ideas relativas al Big Data, lo que puede ser una fuente de problemas. En este sentido, también hay que superar problemas de cultura organizacional, pues en algunos casos los responsables no han asimilado plenamente las implicaciones del Big Data en la gestión de los negocios; tanto en lo referente a la arquitectura de TI como a la forma en que se debe ejecutar (véase capítulo 3).

- **Seleccionar las correlaciones relevantes**. Hay que elegir entre todas las correlaciones aquellas que interesan al negocio. Cuando se cuenta con una ingente cantidad de datos se pueden encontrar correlaciones espurias que no se corresponden con patrones de comportamiento que permitan inferir relaciones de causalidad, sino que más bien pueden ser fruto de la casualidad. En este sentido conviene recordar que los datos son algo fenomenal para la interpolación, pero no tan bueno en la extrapolación[35].

- **Comunicar de forma efectiva**. Hay que saber comunicar para poder ejecutar. Una vez realizado el análisis, hay que decidir cómo se comunican los resultados, de forma de se propicie la toma de decisiones y se facilite la respuesta por parte de la organización. Los datos deben llevar a la acción. El valor inherente a los datos sólo puede asumirse cuando las organizaciones pueden actuar de forma consistente con las oportunidades que les suscitan interés.

Acercándonos a la realidad

Como se ha evidenciado en párrafos anteriores, Big Data es una herramienta muy potente pero su empleo plantea importantes retos. Una visión positiva de tales retos se recoge en el vídeo «¿Cómo superar el reto del análisis de Big Data?» publicado en Youtube el 19 de noviembre de 2014[55].

- ***Reflexione sobre como el uso de la tecnología puede facilitar las operaciones de proveedores.***

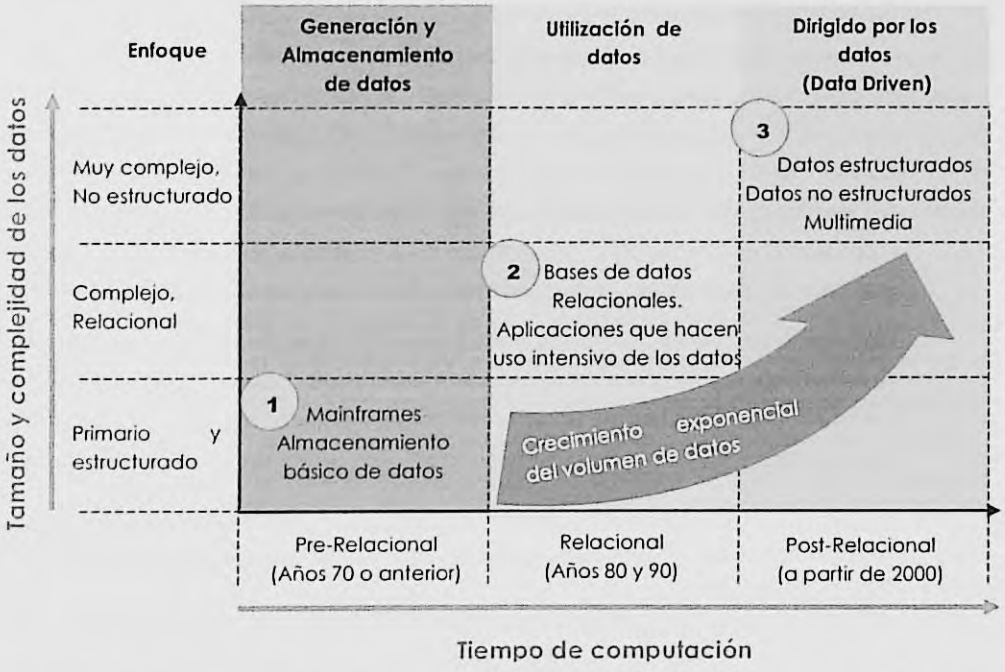
2.2.2 Oportunidades

En los párrafos siguientes vamos a enumerar una serie de razones que explicitan las oportunidades derivadas del uso de Big Data. La Figura 13 muestra la evolución del volumen de datos a lo largo de las últimas décadas.

2.2.2.1 La abundancia de datos como una gran oportunidad

Existen grandes cantidades de información digital sobre prácticamente cualquier tema de interés para un negocio. Big Data cuenta con técnicas rigurosas, pero a la vez más simples y más potentes que las anteriores, lo que unido a la enorme cantidad de datos disponible supone una importante fuente de oportunidades para las organizaciones. En

Figura 13 – Evolución de Big Data



[55] OSS Solutions. *¿Cómo Superar el Reto del Análisis de Big Data?* https://www.youtube.com/watch?v=rqYDc43Amq0. [Online; publicado el 19 de noviembre de 2014]

este sentido, el director de investigación de Google, Peter Norvig, es bastante explícito al destacar la importancia de contar con tantos datos: «Nosotros no tenemos mejores algoritmos. Sólo tenemos más datos»[56].

2.2.2.2 La interacción Big Data - medios sociales como fuente de oportunidades

Aunque la aplicación de Big Data a las redes sociales necesariamente se debe realizar con cautelas para que los resultados puedan ser generalmente aceptados, la superación de estas dificultades es lo que abre importantes oportunidades, algunas de las cuales se sintetizan en los puntos siguientes[42]:

- **Big Data, redes sociales de consumidores finales y redes empresariales.** Las aplicaciones Big Data que están integradas en las propias redes sociales permiten examinar los datos, realizar consultas, revisar las tendencias más destacadas y mostrar visualmente la información. Big Data y las técnicas analíticas asociadas hacen posible convertir todos esos datos en algo comprensible y de aplicación útil. Estas posibilidades se están convirtiendo en una oportunidad de negocios en sí y no sólo para las grandes empresas. Hay un número creciente de datos abiertos, así como mercados de datos a los que las compañías pequeñas pueden acudir.

  Un caso que ilustra como una empresa pequeña puede beneficiarse de Big Data es la empresa de juegos digitales Mediatonic[57]. Para esta empresa el análisis de datos es algo fundamental para su negocio. Los juegos de Mediatonic son exclusivamente digitales. Los usuarios generan enormes cantidades de datos como cuándo, dónde y por cuánto tiempo juegan, y qué partes del juego resultan fáciles o difíciles, y para Mediatonic entender el comportamiento del jugador es crucial. Dada la información facilitada por Big Data, se pueden probar diferentes versiones de un juego en diversos grupos demográficos al mismo tiempo y ajustarlos en respuesta a datos recibidos en tiempo real. En otras palabras, gracias a Big Data Mediatonic puede entender a cada jugador individualmente.

  Otro ejemplo de cómo se puede capitalizar la información facilitada por Big Data es la empresa francesa de anuncios online Criteo. Esta empresa al seguir la navegación de sus clientes en internet consiguió hacer anuncios personalizados ajustados a los intereses del usuario[58]. Los anuncios se pueden cambiar

[56] María José López. *De la Web 2.0 a la Inteligencia Colectiva Aumentada – Tim O'Reilly en FICOD*. http://mjlopezz.com/2011/11/de-la-web-2-0-a-la-inteligencia-colectiva-aumentada-tim-oreilly-en-ficod/. [Online; publicado el 28 de noviembre de 2011]

[57] Matthew Wall. *¿Puede una Empresa Pequeña Beneficiarse del Big Data?* http://www.bbc.com/mundo/noticias/2014/05/140426_economia_grandes_datos_pymes_finde_az. [Online; publicado el 4 de mayo de 2014]

[58] Criteo. *Maximizing Display Advertising's Potential Requires a Data-Centric Approach*. http://www.criteo.com/media/1308/maximizing-display-advertising.pdf. [Online; publicado en marzo de 2013]

en tiempo real según quién visita el lugar, así que dos personas distintas no se encuentran con la misma publicidad. Este tipo de servicios le ha permitido experimentar un fuerte crecimiento en un corto periodo de tiempo.

Acercándonos a la realidad

Muchos pequeños negocios no están en un sector que genera grandes cantidades de datos, así que cabría preguntarse si el Big Data tiene sentido. Así, como se señala en el artículo de Matthew Wall, «¿puede una empresa pequeña beneficiarse del Big Data?»[57] para un comerciante no es realista esperar que el Big Data tengan una aplicación directa, pero pueden acceder a servicios que dependen de grandes análisis de datos.

> ❖ ***¿En qué sentido cabe pensar que en un mundo poblado de aplicaciones para celulares inteligentes, las empresas podrán conectar con el análisis de Big Data a través de sus teléfonos?***

- **La integración en Big Data de los datos internos y externos.** El avance hacia una sociedad hiperconectada, en la que Big Data desempeña un papel relevante, incluye una mezcla creciente entre el mundo empresarial y el de los consumidores finales. Si antes solo se contaba con las técnicas tradicionales de análisis de datos internos de la empresa, ahora con el Big Data se amplía el ciclo del conocimiento, al posibilitar el análisis de los datos externos o de fuentes abiertas. Big Data facilita conocer qué ocurre fuera de la empresa y sacar provecho de ello, lo que requiere saber cómo capturar y sacar valor de la información externa de la empresa, e integrarla con la información interna.

  Big Data facilita extraer el verdadero valor de la información para los responsables de las organizaciones. Esta capacidad para extraer valor en muchas ocasiones va a marcar la diferencia entre el éxito y el fracaso. El éxito de las organizaciones se verá condicionado por la capacidad de montar un sistema que facilite el análisis de información externa no estructurada por parte de las personas clave de la organización.

- **Herramientas analíticas que te «encuentran».** Saber lo que hay que buscar es una virtud esencial a la hora de encontrar valor en el escenario actual de las redes sociales. Pero su confluencia con Big Data permitirá que datos vitales te «encuentren», incluso antes de que sepas que lo necesitas. Por ejemplo, el campo de analítica predictiva requiere software sofisticado combinado con la capacidad de ejecutar una enorme cantidad de consultas basadas en diversas hipótesis para ser corroboradas en un determinado plazo de tiempo.

- **Herramientas analíticas de Big Data en la Nube.** El Big Data masivo requiere una infraestructura flexible y adaptable. Los volúmenes de computación necesaria para Big Data, junto con la naturaleza descentralizada de las redes sociales, resultan ideales para un enfoque basado en la computación en la nube. Los

que sepan combinar satisfactoriamente Big Data, herramientas analíticas y redes sociales para construir soluciones bajo demanda a precios razonables serán los que triunfen.

2.2.2.3 Algunos ejemplos de aplicaciones de Big Data

Las aplicaciones de Big Data son muy variadas y se deben de ajustar a las capacidades de las organizaciones para encontrar valor en el uso de los datos. Algunos ejemplos de las aplicaciones que pueden llevarse a cabo son los siguientes:

- **Sistemas de recomendación**. Utilizan la información sobre el comportamiento de cada usuario para predecir sus intenciones e intereses, y presentarles («recomendarles») contenidos adecuados. Son muy utilizados en comercio electrónico.

- **Análisis de opiniones y sentimientos**. Basándose en conversaciones públicas (por ejemplo, Twitter, foros, ...) se detecta y clasifica la opinión sobre las preferencias o gustos de un usuario y se analiza su comportamiento con finalidades de diferente tipo como, por ejemplo, recomendaciones o inferir el estado de ánimo de los usuarios.

  Un ejemplo de este tipo de análisis se ha aplicado para medir la felicidad en América Latina a través del uso de las redes sociales[59]. En concreto, se ha utilizado la red social Twitter debido a sus características, tales como la facilidad de uso, disponibilidad y popularidad, datos geográficos, etc. Este trabajo evidencia que la medición de la felicidad a través del uso de las redes sociales es viable, y es tremendamente simple en comparación con los métodos tradicionales, como por ejemplo, las encuestas. El método utilizado en este trabajo consiste en inferir los sentimientos de los usuarios de redes sociales sobre la base de un análisis.

- **Predicción y gestión del impacto de catástrofes**. Las grandes cantidades de datos disponibles se utilizan en la detección de eventos como incendios o terremotos, de tal manera que se pueda predecir su impacto y generar una reacción temprana que minimice los efectos destructivos de tales eventos.

- **Juegos**. La gestión de grandes cantidades de información se ha aplicado a juegos concretos, como por ejemplo el ajedrez (Deep Blue) o de preguntas y respuestas (Watson). Estos dos casos son ejemplos de programas que analizan grandes cantidades de datos de partidas o conocimiento genérico y de sentido común para jugar con contrincantes humanos.

[59] Francisco Mochón y Óscar Sanjuán. «A First Approach to the Implicit Measurement of Happiness in Latin America Through the Use of Social Nework». En: *International Journal of Interactive Multimedia and Artificial Intelligence* 2.5 (2014), págs. 16-22

- **Reconocimiento y categorización**. Hay aplicaciones que permiten recorrer lugares, imágenes, caras o personas mediante el análisis del gran volumen de datos de este tipo disponibles online. Este tipo de aplicaciones se utilizan para llevar a cabo actividades de realidad aumentada.

- **Medicina genómica**. La medicina genómica personalizada (aún en el fase de investigación y prueba) analiza e integra datos genómicos y clínicos para el diagnóstico precoz y una aplicación más adecuada de las terapias. Así, la secuenciación masiva de genes parte del hecho que cada paciente es diferente. Por ello la mejor forma de tratar ciertas enfermedades, como el cáncer, es analizando en profundidad las alteraciones moleculares únicas y específicas de cada tumor y convertirlas en potenciales dianas terapéuticas. El análisis de datos aporta información relevante para predecir la respuesta a determinados fármacos y, por tanto, adecuar el tratamiento al perfil molecular de cada tumor.

- **Comportamiento inteligente de servicios públicos**. Utilizando la información proveniente de datos recopilados por sensores inteligentes puede mejorarse la distribución y consumo de recursos fundamentales como el agua o la energía eléctrica. En los programas de Smart cities se recurre a este tipo de aplicaciones.

- **Modelado de riesgos**. Algunas entidades bancarias y firmas de inversión punteras utilizan tecnologías de análisis de grandes cantidades de datos para determinar el riesgo de operaciones, evaluando un gran número de escenarios financieros hipotéticos.

- **Detección de fraude**. Utilizando técnicas para combinar bases de datos con el comportamiento de usuarios y datos transaccionales online pueden detectarse actividades fraudulentas, como por ejemplo el uso de una tarjeta de crédito robada.

- **Monitorización de redes**. Las redes de servidores producen una gran cantidad de datos que pueden ser analizados para identificar cuellos de botella o ataques a redes. Este tipo de análisis puede aplicarse también a otros tipos de redes, tales como redes de transportes, con el fin, por ejemplo, de optimizar el consumo de combustible.

- **Investigación y desarrollo**. Algunas empresas con fuerte componente investigadora, como las farmacéuticas, realizan análisis de grandes volúmenes de documentación (por ejemplo artículos científicos) y otro tipo de datos históricos para mejorar el desarrollo de sus productos.

La relación de ejemplos de aplicaciones de Big Data presentados no es una relación exhaustiva. Su utilidad estriba exclusivamente en ilustrar la diversidad de usos que se le puede dar a Big Data, debiéndose señalar, además que en muchos de los casos su implantación está en las primeras fases.

Acercándonos a la realidad

Julián García Barbosa, el 9 de octubre de 2014, publicó el artículo «La medicina del futuro pasa por Big Data»[60]. En el artículo se afirma que en la actualidad los sistemas sanitarios están inmersos en un mar de datos y gran parte de estos datos son desestructurados (radiografías, resonancias magnéticas, mensajes de Twitter...) y no pueden gestionarse con bases de datos tradicionales. Además, son generados, analizados y explotados a una gran velocidad, como los datos que envían en tiempo real los sensores que recogen las constantes vitales de un paciente.

- ***¿En qué sentido comparte la opinión del autor en el sentido de que el futuro de la medicina (especialmente de la medicina de las 4P: personalizada, predictiva, preventiva y participativa) pasa por el Big Data?***

2.3 Componentes de un Sistema Big Data

2.3.1 Introducción

Como se ha podido comprobar, las aplicaciones de la tecnología Big Data son múltiples y las áreas en las que se pueden aplicar, también son variadas. Además, muchas de las tecnologías en las que se sustentan los sistemas de Big Data, son proyectos muy jóvenes, en muchos casos sólo llevan aplicándose unos pocos años, lo que significa que son proyectos en continua evolución.

Además, los sistemas de Big Data deben convivir e integrarse con las tecnologías existentes en las empresas, fundamentalmente con los sistemas de almacenamiento tradicionales, como las bases de datos y los sistemas informacionales basados en soluciones de soporte a la toma de decisiones; para complementarlas o para mejorarlas.

Todos estos hechos, hacen que sea prácticamente imposible definir una arquitectura tecnológica estandarizada para el despliegue de sistemas de Big Data en el mundo real. Sin embargo, sí que podemos dividir, desde un punto de vista funcional, las diferentes tareas que deben realizarse por un sistema Big Data, lo que nos ayudará a identificar los componentes básicos con los que debe contar este tipo de sistemas.

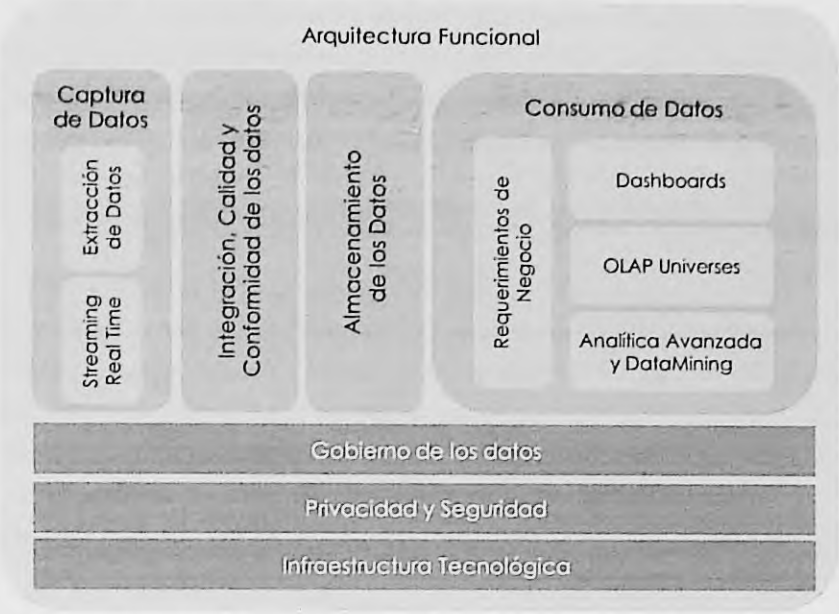
2.3.2 Arquitectura Funcional

A continuación trataremos de identificar las diferentes tareas que debe realizar un sistema de Big Data.

En la Figura 14 se muestra un diagrama de la arquitectura funcional de un sistema de Big Data, pudiéndose identificar cuatro grandes áreas funcionales de abajo a arriba (a menudo se utiliza el término «capa» para indicar cada una de las diferentes áreas).

[60] Julián García Barbosa. *La Medicina del Futuro Pasa por Big Data.* http://www.aunclicdelastic.com/la-medicina-del-futuro-pasa-por-big-data/. [Online; publicado el 9 de octubre de 2014]

Figura 14 – Arquitectura funcional de un sistema Big Data



- En la capa inferior nos encontramos la **infraestructura tecnológica** en la que se sustentan los sistemas. Se trata de los sistemas físicos (hardware, redes, etc.). La infraestructura física es fundamental para el funcionamiento y la escalabilidad de una arquitectura para Big Data. Para soportar un volumen inesperado o impredecible de datos, la infraestructura física para Big Data tiene que ser diferente a la de los sistemas tradicionales de gestión de datos. Big Data está basado en un modelo de computación distribuida. Es decir, los datos pueden estar físicamente almacenados en muchos sitios distintos y son conectados a través de la red, mediante el uso de sistemas de archivos distribuidos o mediante el uso de diferentes herramientas y aplicaciones de analítica de Big Data.

  La irrupción de la computación en la nube, ha cambiado la forma en la que entendemos el concepto de infraestructura. Esta ha pasado de ser un elemento físico a un servicio recibido. Proveedores como Amazon Web Services o Microsoft Azure, ofrecen a sus clientes robustas infraestructuras distribuidas, capaces de crecer dinámicamente en función de las necesidades de computación y almacenamiento, sin que los clientes tengan que adquirir una sola máquina. La nube ha hecho surgir nuevos modelos de negocio para consumir infraestructuras tecnológicas, como la Infraestructura como servicio (IaaS – *Infraestructure as a Services*), la Plataforma como Servicio (PaaS – *Platform as a Service*) o el Software como Servicio (SaaS – *Software as a Service*), véase la sección 2.4.1.4.

- A continuación se encuentra la capa de **privacidad y seguridad**. A medida que los datos se convierten en un activo más importante para una compañía, más relevancia toman la seguridad y la privacidad de los mismos (véase capítulo 10). La seguridad de los datos se fundamenta en garantizar que los datos serán accedidos por quien debe y sólo tendrá acceso a los datos estrictamente

necesarios para la actividad que necesita realizar. Tan importante es asegurar los datos como las operaciones que se pueden realizar con ellos. Los algoritmos desarrollados pueden inferir información que atente contra la privacidad de las personas. Imaginemos el caso de un hospital que, en sus sistemas, tiene información de sus pacientes. Será necesario determinar quién debe acceder a qué información y en qué circunstancias pueden hacerlo. Será necesario identificar a los usuarios que acceden a la información y proteger la identidad de los pacientes.

- Posteriormente tenemos la capa de **gobierno de los datos**. La gobernanza de los datos es una actividad importante especialmente cuando trabajamos con grandes volúmenes de datos (véase capítulo 10). El gobierno de los datos implica tener documentados los datos, es decir descritos en todas sus dimensiones; qué significan, donde están (hecho muy relevante en sistemas distribuidos), fuentes de las que proceden, transformaciones a las que han sido sometidos, qué validez tienen, etc.

- Finalmente nos encontramos con una capa que está subdividida en otras, se trataría de la capa de **gestión de datos**. En esta capa encontraremos todos los elementos necesarios para gestionar el ciclo de vida de los datos, empezando de izquierda a derecha nos encontramos con diferentes componentes que nos permitirán:

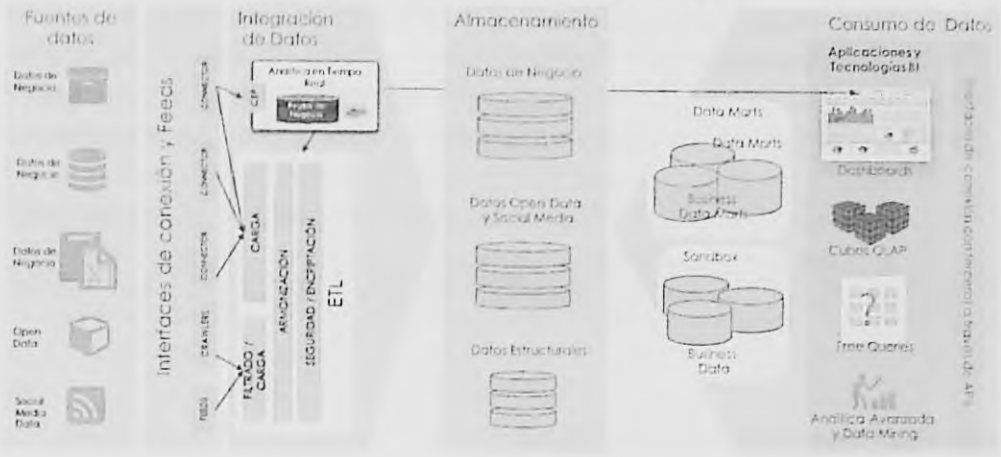
    - Realizar captura de datos, de diferentes fuentes (internas o externas) y en diferentes formas y formatos.
    - Proceder a la integración de los datos en nuestro sistema, incluyendo por su puesto aquellos que nos permitan garantizar la calidad de nuestros datos.
    - Almacenar de forma persistente y óptima nuestros datos.
    - Facilitar el consumo de los datos por terceros, ya sean estas personas (informes, cuadros de mandos, etc.), u otros sistemas automáticos (a través de interfaces de comunicación y entrega de datos, como las APIs).

Para describir con un poco más detalle esta última capa, podemos ayudarnos del esquema de la Figura 15 en el que podemos identificar los siguientes componentes:

- **Fuentes de datos**. En realidad no se trata de un componente, sino más bien de un requisito. Se trata de la materia prima con la que trabajará nuestro sistema.

- **Interfaces de conexión y *feeds***. Una característica importante de Big Data es la capacidad de gestionar grandes cantidades de datos, por lo que es necesario

Figura 15 – Componentes de la capa de datos



dotar al sistema de la capacidad de alimentarse de múltiples fuentes de datos. Hay que tener en cuenta que el sistema debe estar dotado de este tipo de componentes tanto en la parte de carga de los datos, como en la parte de consumo de los mismos.

- **Procesos de integración de los datos.** Se trata de los procesos de extracción, transformación y carga (ETL) de los datos en nuestro sistema. Una vez identificadas las fuentes de datos de donde obtendremos la información y antes de proceder a almacenarla, los datos deben ser transformados, limpiados, filtrados y redefinidos. Normalmente, la información que obtenemos de las distintas fuentes, ya sean los sistemas transaccionales de la compañía, o los datos obtenidos de las redes sociales, no está preparada para el uso que daremos posteriormente. Cuando trabajamos con Big Data en muchas ocasiones es necesario realizar analíticas en tiempo real lo que supone implementar procesos complejos de gestión de los datos (CEP – *Complex Event Procesing*).

- **Almacenamiento de datos.** Tradicionalmente los datos con los que las empresas trabajaban procedían de los sistemas transaccionales de la compañía, lo que significa que se trataba de datos altamente estructurados que eran almacenados en bases de datos relacionales (bases de datos SQL). La realidad actual implica que los datos proceden de múltiples fuentes y que no se trabaja sólo con datos estructurados. Esto hace que los sistemas deban estar preparados para gestionar este tipo de información y por tanto incluya, no sólo bases de datos relacionales, sino también sistemas NoSQL.

- **Herramientas que permitan el acceso a los datos.** Que nos proporcionen capacidad de cálculo, consultas, funciones de planificación, pronóstico y análisis de escenarios sobre grandes volúmenes de datos. La capacidad de almacenar grandes volúmenes de datos, implica la aparición de un nuevo reto, ¿cómo ser

capaz de gestionarlos? La respuesta viene de la mano de nuevos paradigmas de computación e innovaciones en este campo tales como MapReduce o el procesamiento de datos en memoria.

- **Herramientas de visualización y consumo de datos**, que nos permitirán el análisis y la navegación a través de los mismos. El sistema debe contemplar componentes que permitan tanto la interacción con humanos (*dashboards*, informes, consultas personalizadas, etc.), como la interacción con otros componentes propios o de terceros (mediante interfaces como las APIs o los servicios Web).

2.4 Big Data y Cloud Computing

2.4.1 Introducción a la Computación en la Nube

2.4.1.1 Breve reseña histórica

La computación en la nube (*cloud computing*) se puede definir de forma genérica como cualquier aplicación que haga uso de servicios proporcionados a través Internet. El término se acuñó hace pocos años, y se ha convertido en una tecnología consolidada de uso generalizado.

Los orígenes de la computación en la nube los podemos situar a mediados de los años 80, cuando la popularización de los ordenadores personales y las redes de comunicación hizo que fuera factible disponer de redes de ordenadores que permitan que éstos se puedan comunicar entre sí usando protocolos normalizados. De este modo se generaliza el uso de redes de área local (LAN – *Local Area Network*), como las redes Ethernet, y las redes de área extensa (WAN – *Wide Area Network*), siendo Internet el más claro exponente.

Una ventaja de las redes de ordenadores es que permiten el procesamiento distribuido, es decir, usar todos los ordenadores de la red conjuntamente para realizar cómputos y procesamiento de datos en paralelo.

Justo en esa época empiezan a surgir los supercomputadores paralelos, de muy elevado coste, y a finales de los 80 aparecen sistemas más asequibles, como los basados en procesadores Transputer, que posibilitan disponer de sistemas informáticos con decenas de procesadores a un coste razonable. Sin embargo, estos sistemas se componían de procesadores y hardware específico, y tenían problemas de escalabilidad, por lo que en general sólo se podía disponer de varias decenas de procesadores.

En los años 90 se produce un auge de los denominados «*clusters*», que consisten en sistemas compuestos de ordenadores convencionales interconectados a través de una red de área local. Estos sistemas eran, y son aún hoy día, fáciles de construir y de mantener, y es factible disponer de *clusters* con cientos de procesadores.

A finales de los 90 se acuña el término «*grid computing*», que hace referencia a la posibilidad de tener un sistema informático compuesto de miles de procesadores como resultado de interconectar vía Internet *clusters* de diferentes organizaciones. De este modo surgen proyectos como Seti@HOME (análisis de señales del espacio para detectar inteligencia extraterrestre), que mediante el uso de tiempo de CPU donados por voluntarios consigue usar varios cientos de miles de ordenadores. Sistemas como Boing, Globus o Condor son ejemplos de plataformas de computación *grid*.

En esa misma época, con la Web ya firmemente asentada sobra una Internet ya omnipresente, se desarrollan los denominados *web services*, que posibilitan la realización de aplicaciones que dan servicios a través de la Web. De esta forma, muchas empresas empiezan a ofrecer aplicaciones que se ejecutan directamente en el navegador Web, como es el caso de los documentos de Google. Es en este contexto en el que aparece la computación en la nube.

El término «*cloud*» (nube) parece tener su origen en algunos diagramas que representaban a Internet como una nube que da servicios (un ejemplo se observa en la Figura 16). Desde entonces, la computación en la nube, lejos de ser una tecnología de moda con una duración temporal limitada, se ha convertido en una infraestructura que tiene una gran influencia en muchos sectores, tanto de usuarios finales como de empresas, dadas las ventajas que ofrece.

Hay que resaltar que cuando la computación en la nube comenzó a ser conocida, se generaron discusiones acerca de si el concepto era una evolución de *grid computing*. Hoy día la diferencia está clara. La computación en la nube no es un paso más de la computación *grid* sino que son conceptos diferentes aunque en algunos casos pueden ser complemen-

Figura 16 – Computación en la nube

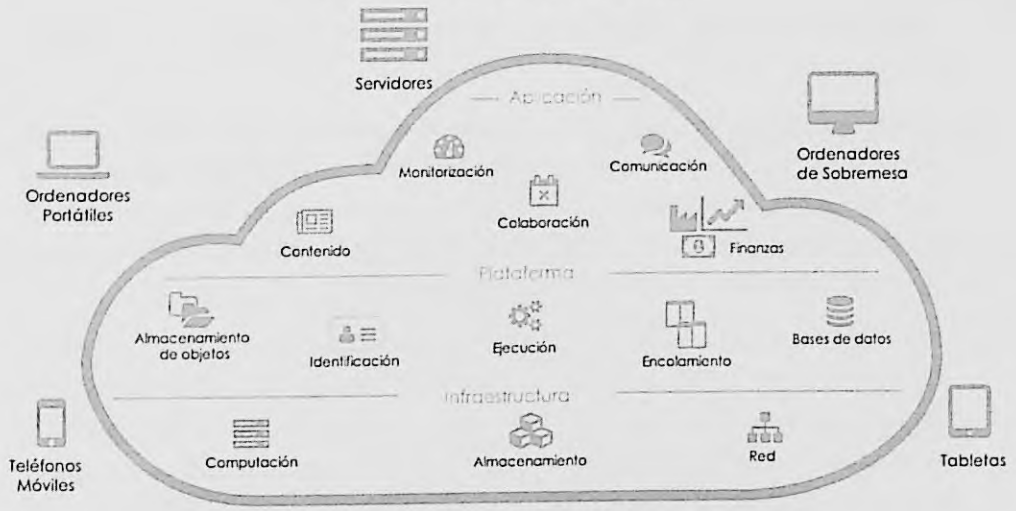


Figura 17 – De los ordenadores paralelos al *cloud computing*

tarios; por ejemplo, un servicio de la nube puede dar acceso a un sistema de computación *grid*. La Figura 17 muestra la aparición en el tiempo de las tecnologías mencionadas.

2.4.1.2 ¿Qué es la computación en la nube?

Existen muchas definiciones acerca de lo que es la computación en la nube.

Una definición que reúne los aspectos más destacados de la computación en la nube es la que señala que esta tiene que ver con servicios que se ofrecen a través de Internet, existiendo un proveedor de dichos servicios pudiéndose contratar estos de forma flexible, en función de las necesidades y abonándose según lo que se solicita y se consume.

2.4.1.3 Características de la computación en la nube

La computación en la nube tiene varias características, entre las que se pueden reseñar las que se mencionan a continuación:

1. **Accesibilidad**: se puede acceder a los recursos de la nube desde cualquier sitio, siempre que se disponga de conexión a Internet. Para ello se pueden usar PCs, portátiles, tabletas, *smartphones*, etc. A esto contribuye que el acceso a la nube se suele hacer a través de protocolos de comunicaciones normalizados (por ejemplo, los basados en HTTP).
2. **Mantenimiento**: el proveedor de servicios de la nube es el responsable de la puesta en marcha y mantenimiento de dichos servicios. Los usuarios no tienen que preocuparse de adquirir equipos, de actualizar los sistemas operativos o las bases de datos, ni de adquirir las últimas versiones de componentes software. Estas tareas son responsabilidad del proveedor, no del cliente.
3. **Elasticidad**: permite proporcionar servicios basados en demandas que se ajustan a las necesidades actuales, manteniendo siempre altos niveles de seguridad y fiabilidad. De esta forma, los recursos solicitados (memoria, procesador, disco, ancho de banda de comunicaciones) se pueden ajustar de forma dinámica, ya sea incrementándolos o decrementándolos. Esta flexibilidad es una de las características clave de los sistemas de computación en la nube.

4. **Pago por uso**: el carácter elástico de la computación en la nube permite pagar sólo por lo que se necesita, lo que presenta grandes ventajas a los clientes, que pueden en todo momento decidir qué cantidades desean dedicar a este tipo de servicios.

2.4.1.4 Modelos de servicios en la nube

Los modelos de servicio asociados a la computación en la nube más conocidos son, como se ilustra en las Figuras 18 y 19: infraestructura como servicio (*Intrastructure as a Service* o IaaS), plataforma como servicio (*Platform as a Service* o PaaS) y software como servicio (*Software as a Service* o SaaS). A continuación se describen estos tres modelos:

- ***Infrastructure as a Service*** **(IaaS)**: los consumidores controlan y gestionan los sistemas operativos, aplicaciones, almacenamiento y conexiones de red, pero no controlan la infraestructura de la nube. Los proveedores dan la infraestructura típicamente a través de una plataforma de virtualización. Los ejemplos más conocidos son Amazon EC2 y Microsoft Azure.

- ***Platform as a Service*** **(PaaS)**: los consumidores adquieren el acceso a las plataformas, pudiendo desplegar sus propias aplicaciones (bases de datos, servidores Web, etc.) Los sistemas operativos y el acceso a la red no es controlado por los clientes, y normalmente existen restricciones en cuanto a qué aplicaciones se pueden desplegar. Dos ejemplos son Microsoft Azure y Google App Engine.

- ***Software as a Service*** **(SaaS)**: los consumidores compran la posibilidad de acceder y usar aplicaciones o servicios que se alojan en la nube. Normalmente se usa la Web para proporcionar las aplicaciones de los proveedores, de forma que en la mayor parte de los casos las aplicaciones se ejecutan directamente en navegadores Web, sin necesidad de descargar o instalar ningún software.

Figura 18 – Modelos de *cloud computing*

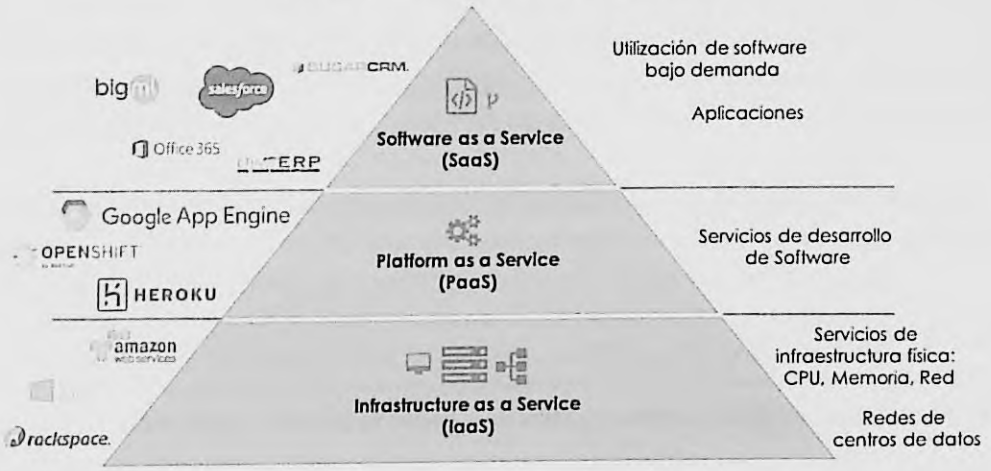
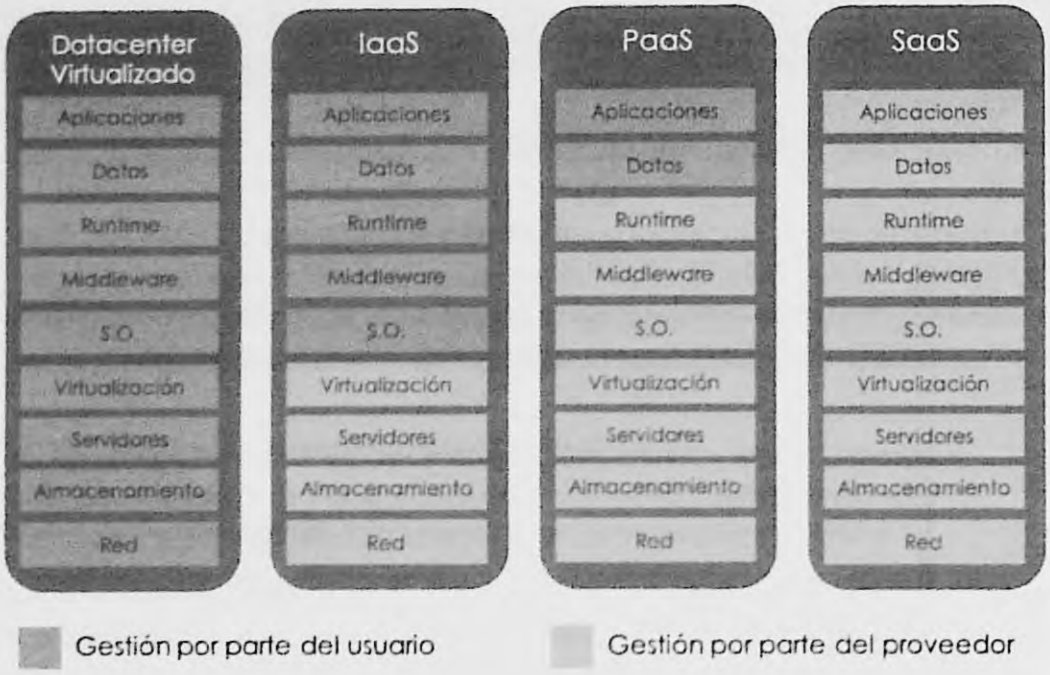


Figura 19 – Diferencias entre los distintos modelos de *cloud computing*



Un ejemplo de aplicaciones SaaS es Google Drive, que ofrece almacenamiento y herramientas ofimáticas para procesamiento de textos, hojas de cálculo y presentaciones; Microsoft ofrece algo parecido con Microsoft 365.

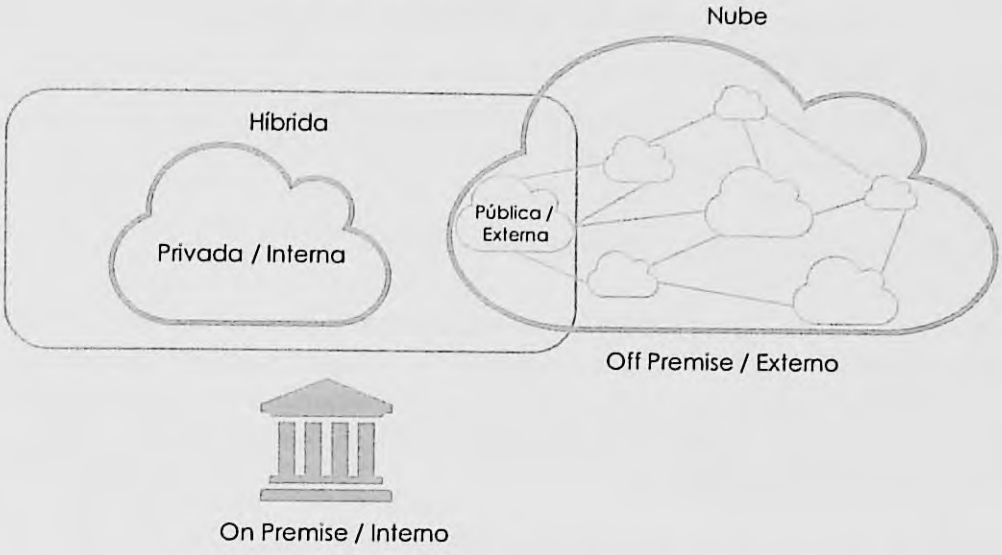
Se puede observar como la responsabilidad de los componentes del sistema varía del control propio total en el caso del *datacenter*, al control total por el proveedor de servicios en la nube en un modelo SaaS. Los modelos IaaS y PaaS representan diferentes compromisos entre lo que es gestión propia y gestión de terceros.

2.4.1.5 Modelos de despliegue

El despliegue de aplicaciones de computación en la nube puede variar dependiendo de las necesidades de las mismas. Se han identificado tres modelos básicos de despliegue, cada uno con unas características específicas para dar soporte a las necesidades de los de los servicios y de los usuarios (ver Figura 20):

- ***Cloud* privado**: la infraestructura de la nube es desplegada, mantenida y gestionada por una organización concreta. La gestión puede ser interna o llevada a cabo por un tercero. Este tipo de nube ofrece un elevado grado de seguridad y de control, porque requiere que la organización adquiera y mantenga toda la infraestructura, lo que conlleva una inversión que puede ser considerable.
- ***Cloud* público**: la infraestructura de la nube es proporcionada por un suministrador de servicios que los ofrece públicamente en base a un esquema comer-

Figura 20 – *Cloud* privado, externo e híbrido



cial (que puede incluir la posibilidad de ser gratuito). Esto permite a los usuarios desarrollar y desplegar sus propias aplicaciones en la nube con un coste financiero menor que si se optara por tener que mantener un *datacenter* o una nube privada.

- ***Cloud* híbrido**: la infraestructura de la nube híbrida consiste en tener varias nubes de distinto tipo, de forma que a través de sus interfaces es posible mover los datos y las aplicaciones de unas nubes a otras. Se suele usar como una combinación de nubes privadas y públicas, de forma que la privada contiene datos dentro de la propia organización y la pública se usa para otros servicios. De esta forma cada aspecto de la organización se localiza en el entorno más eficiente posible.

2.4.1.6 Beneficios de usar computación en la nube

Una vez que se conocen las características de la computación la nube, los modelos de servicio que existen y los distintos modelos de despliegue se pueden determinar los beneficios que pueden obtener los clientes y usuarios:

- **Menor coste**: las organizaciones pueden reducir su inversión en infraestructura hardware y software, ahorrando costes de mantenimiento y de personal técnico.
- **Escalabilidad y flexibilidad**: una empresa puede comenzar a usar la nube con un despliegue mínimo y posteriormente crecer o decrecer según sus necesidades. La flexibilidad permite usar recursos extra en situaciones de picos de trabajo, permitiendo así satisfacer sus necesidades.

- **Fiabilidad**: los servicios en la nube usan múltiples localizaciones redundantes, lo que permite mantener la continuidad de sus servicios y la recuperación en casos de fallos.

- **Movilidad**: la nube es accesible desde cualquier sitio y desde cualquier dispositivo con acceso a Internet.

2.4.1.7 Desventajas de la Computación en la Nube

Usar computación en la nube también tiene algunas desventajas, de las cuales se pueden destacar dos. La primera es que es necesario estar conectado a Internet para acceder a los servicios de la nube. Un corte en las comunicaciones en un momento crítico puede ser fatal para una organización que basa todo su negocio en aplicaciones y servicios desplegados en una nube pública.

El segundo mayor problema es la seguridad. Hay que tener en cuenta que los datos de la empresa están alojados en servidores externos, por lo que se pierde el control sobre ellos y no hay más remedio que confiar en el proveedor de servicios. Por otro lado, se es vulnerable a ataques a la seguridad del tipo «denegación de servicio» (DoS, «*Denial of Service*»), que pueden impedir el acceso a la nube. Estos aspectos se analizan con más detalle a continuación.

Privacidad y protección de datos en entornos de Cloud Computing

La pérdida de control sobre los datos almacenados se puede producir en varios contextos, algunos de los cuales se comentan a continuación y se plantea con más detalle en el capítulo 10:

1. Existe una falta de transparencia para los propietarios de los datos sobre cómo, cuándo, por qué y dónde éstos son procesados por parte del proveedor del servicio en la nube.

2. Los proveedores pueden analizar los datos, por ejemplo usando técnicas de minería de datos. Esto puede ser problemático en el caso de información privada o confidencial. Un ejemplo son las redes sociales, en las que todo lo que incluyen los usuarios es analizado.

3. Los dispositivos móviles, debido a sus limitadas capacidades de almacenamiento, usan almacenamiento en la nube en lugar de la memoria del dispositivo. Muchos usuarios no son conscientes de este hecho, lo que puede entrañar riesgos si piensan que los datos no salen de su teléfono móvil o tableta.

4. Los servicios en la nube siempre son accedidos de forma remota, por lo que existen riesgos si las conexiones no están debidamente protegidas, como *eavesdropping*

(leer conversaciones privadas sin autorización), DNS *spoofing* (suplantación de direcciones IP) o ataques de denegación de servicio.

5. Las leyes y regulaciones sobre privacidad de la información varían de unos países a otros, lo que es un problema cuando los centros de datos de los proveedores en la nube están repartidos en distintos puntos geográficos. Así, por ejemplo, las leyes de protección de datos vigentes en Europa puede que no se vean garantizadas si las plataformas en la nube se almacenan en países no europeos.

6. Un aspecto no muy conocido por los usuarios concierne al borrado de datos. Es complejo a veces borrar las copias de los datos por la sencilla razón de que puede ser difícil encontrar todas las copias existentes.

7. La seguridad de las comunicaciones de los servicios en la nube se suelen basar en la seguridad y fiabilidad de las redes de comunicación propietarias de empresas de telecomunicación. Esto introduce un tercer componente de riesgo que escapa del control de la empresa de computación en la nube.

8. Si la empresa de servicios en la nube quiebra, los servicios que ofrece se interrumpirán; lo que puede dar lugar a problemas para acceder a los datos, con los consecuentes efectos negativos para el cliente.

Como consecuencia de la gran variedad de aspectos que afectan a la privacidad y protección de los datos en entornos *cloud*, los usuarios de este tipo de servicios deben tomar las precauciones necesarias para evitar situaciones no deseadas. Las más importante es ser consciente de las situaciones que se pueden producir para de este modo disminuir los riesgos, lo que implica invertir en la formación del personal involucrado en las aplicaciones en la nube.

En cualquier caso debe señalarse que muchos de los problemas que afectan a la privacidad y a protección de datos no son exclusivos de la computación en la nube, sino de cualquier aplicación que se ejecute sobre la infraestructura de Internet.

A modo de balance final puede afirmarse que la computación en la nube es un paradigma que ya está firmemente implantado y que presenta una serie de ventajas que hacen que cada vez sea más usado. Ofrece diversos modelos de servicios y de despliegue, por lo que se puede adecuar a las necesidades particulares de cada cliente.

Hay que tener en cuenta, no obstante, los riesgos que conlleva, ya sea por la dificultad para acceder a los recursos remotos contratados o por la pérdida de control sobre los datos propios que se almacenan en los proveedores de los servicios en la nube.

2.4.2 Computación en la Nube y Big Data

En la actualidad Big Data y la computación en la nube están muy relacionados, ya que los requisitos de las aplicaciones de Big Data en cuanto necesidades de procesamiento,

memoria, almacenamiento y ancho de banda pueden ser ofrecidos por un proveedor de la nube.

El principal desafío del Big Data es permitir analizar la gran cantidad de datos disponibles con el fin de proporcionar al usuario final información lo más precisa, útil y valiosa posible. En este contexto se pueden concretar tres aspectos en los que usar computación en la nube para aplicaciones de análisis del Big Data tiene claras ventajas:

- La **inversión en analizar los resultados del Big Data puede ser considerable**. Los recursos disponibles en las empresas consisten generalmente en data centers. Usar una nube privada puede ofrecer una mejor relación coste/rendimiento, sobre todo si se configura con la posibilidad de acceder a nubes públicas en caso de necesidad. De esta forma, mediante el modelo de nube híbrida se pueden usar los servicios de cómputo y almacenamiento de la nube pública para ciertos tipos de análisis que pueden ser intensivos en recursos aunque de corta duración.

- **Las aplicaciones de Big Data pueden mezclar fuentes de datos internas y externas**. Aunque las organizaciones generalmente mantienen los datos más sensibles en sus servidores internos, volúmenes ingentes de datos, que pueden provenir de terceras partes, estarán localizados en sitios externos (en algunos casos, estarán en una nube). De este modo, mover datos internos fuera de la organización puede permitir analizar el conjunto global de datos en el sitio en que éstos están almacenados, lo que puede tener más sentido (en términos de eficiencia) que tener que traerlos a la organización.

- **Existen servicios de datos para extraer valor del Big Data**. Se está acuñando el término *Analytics as a Service* o AaaS y *Big Data as a Service* o BDaaS para hacer referencia a servicios de análisis del Big Data que pueden ser ofrecidos por nubes privadas, públicas o híbridas. En concreto, BDaaS es un concepto reciente que hace referencia a subcontratar varias de las funcionalidades del Big Data en la nube. No se trata únicamente de adquirir capacidad de almacenamiento y de procesamiento de datos, sino de proporcionar servicios para analizar dichos datos. De esta forma, el usuario pagaría por el tiempo que usa las herramientas de análisis o por la cantidad de datos procesados.

Otra vertiente, que es muy utilizada en la actualidad, es usar el modelo IaaS de proveedores en la nube para desplegar infraestructuras basadas en Hadoop en la nube. Tanto Amazon como Microsoft ofrecen la facilidad de contratar *clusters* Hadoop ya configurados, con lo que la puesta en marcha de las aplicaciones de los clientes se ve muy simplificada. Hay que tener en cuenta que disponer de un *cluster* Hadoop propio, al margen de adquirir el hardware necesario, conlleva un coste considerable para su instalación, gestión y mantenimiento.

A continuación se comentan las soluciones ofrecidas por ambas empresas.

2.4.2.1 Amazon EMR

Amazon Elastic Map Reduce (EMR) es un servicio más del abanico que ofrece Amazon AWS (Amazon Web Services) que permite ejecutar Hadoop en *clusters* de tamaño variable de instancias de Amazon EC2 (máquinas virtuales de capacidad variable según demanda). De acuerdo con la información proporcionada por este servicio, las ventajas que ofrece son:

- **Facilidad de uso**: Se puede lanzar un *cluster* en pocos minutos, sin tener que preocuparse por la provisión de nodos, instalación y configuración del mismo.
- **Elasticidad**: Se puede aumentar o reducir el tamaño de las instancias de forma sencilla, pagando sólo por lo que use.
- **Bajo coste**: Un *cluster* Hadoop de 10 nodos se puede ejecutar desde pocos céntimos de dólar / hora.
- **Fiabilidad**: El servicio supervisa el funcionamiento del *cluster*, reiniciando tareas que fallen y sustituye automáticamente instancias de pobre rendimiento por otras.
- **Seguridad**: Existe un cortafuegos que controla el acceso a las instancias y se ofrecen redes privadas virtuales.
- **Flexibilidad**: El cliente tiene pleno control del *cluster*.

Existe una enorme gama de posibilidades para poder configurar un *cluster* Hadoop usando Amazon EMR (véase capítulo 7). Al margen de los tipos de instancias Amazon EC2 (pequeñas, medianas, grandes, optimizadas para memoria, optimizadas para almacenamiento, etc.), los precios también pueden variar dependiendo de la zona o región que se seleccione (EEUU, UE, Asia Pacífico, o América del Sur).

Como es habitual en este tipo de servicios, se ofrece una calculadora para poder conocer el precio de los servicios contratados según se va ajustando lo que se pide. Así se facilita al cliente que pueda estudiar distintas configuraciones y pueda elegir la que más se adecúe a sus necesidades.

2.4.2.2 Microsoft Azure

Azure es la plataforma de Microsoft para proporcionar servicios en la nube. Comparte muchas de las ventajas de Amazon EMR, pero tiene algunas diferencias:

- Ofrece servicios de tipo IaaS y PaaS.
- Permite configuraciones según modelos híbridos de despliegue.
- Proporciona un servicio de aprendizaje máquina (*machine learning*)
- Ofrece un servicio de eventos (*event hub*), que permite analizar y procesar millones de eventos que se registran a través de dispositivos y aplicaciones registrados.

Como es lógico, debajo de Azure se encuentran plataformas de Microsoft como Windows o .NET, lo que en los casos en los que se tenga experiencia con esos sistemas puede hacer más interesante optar por Azure que por Amazon EMR.

Al igual que Amazon EMR, existe un amplio abanico de opciones a elegir en Azure, incluyendo configuraciones para sitios Web, análisis, integración híbrida, multimedia, etc., que pueden ajustar a través de la calculadora de precios que proporciona (véase capítulo 7).

2.4.2.3 Otros proveedores

Al margen de Amazon y Microsoft, existen otras muchas empresas que ofrecen servicios en la nube sobre los que ejecutar aplicaciones de Big Data. La siguiente lista incluye algunos de los más reseñables:

- Google Cloud Infrastructure
- IBM Cloud
- CloudSigma
- Rackspace Cloud Big Data
- HP Cloud

Las características de los servicios de computación en la nube, permiten en la actualidad el acceso al Big Data a usuarios para los que sería muy difícil montar su propio sistema de computación de Big Data.

No es sólo una cuestión de ahorro en equipamiento informático, sino también de coste de personal especializado y de tiempo, dado que los distintos proveedores permiten disponer de un sistema completo, configurado con las necesidades que se hayan contratado, en un breve período de tiempo.

2.4.3 Costes de Almacenamiento de Datos en la Nube

Elegir una configuración determinada para almacenamiento de datos en la nube puede ser una tarea muy compleja dado el número de proveedores disponibles y la gran cantidad de posibles configuraciones que cada uno ofrece. Una posibilidad es usar herramientas de gestión de costes de la nube como CloudVertical, que ofrece un comparador de precios de cinco proveedores (Amazon AWS, Microsoft Azure, HP Cloud, Google Compute and Rackspace). De esta forma, usando una determinada plataforma como referencia se puede obtener una estimación del precio de alternativas de otras empresas y así tener una visión de conjunto de forma fácil.

Otros servicios similares a CloudVertical son Cloudyn, RightScale y Teevity. La mayoría de estos servicios son de pago aunque suelen ofrecer un período de pruebas gratuito.

Para dar una idea de los rangos de precios que nos podemos encontrar hemos usado CloudVertical (con datos actualizados a febrero de 2015) para configurar varios tipos de sistemas. Para ello se han tomado los siguientes parámetros como base:

- **Región**: Estados Unidos
- **Moneda**: dólar americano
- **Precio**: coste por semana (168 horas)
- **Plataforma**: Linux
- **Proveedor de referencia**: Amazon AWS

Dentro del gran abanico de configuraciones que ofrece Amazon AWS (uso general, optimizadas para cómputo, optimizadas para memoria y optimizadas para almacenamiento) hemos considerado las dos que hacen énfasis en almacenamiento, que se denominan instancias I2 y HS1. Los nombres concretos y las prestaciones de las instancias de estos tipos se resumen en la Figura 21.

Figura 21 – Tabla de características técnicas de diferentes opciones de configuración

Modelo	vCPU	Memoria (GB)	Almacenamiento (GB)
i2.xlarge	4	30,5	1 x 800 SSD
i2.2xlarge	8	61	2 x 800 SSD
i2.4xlarge	16	122	4 x 800 SSD
i2.8xlarge	32	244	8 x 800 SSD
hs1.8xlarge	16	117	24 x 2000

Figura 22 – Tabla comparativa de costes de servicios de *cloud computing*

Instancia	Amazon AWS	Google	Windows Azure	HP Cloud	Rackspace
i2.xlarge	143$	**85$**	**84$**	*151$*	*228$*
i2.2xlarge	267$	**169$**	**168$**	*302$*	*457$*
i2.4xlarge	573$	**99$**	**168$**	**302$**	**457$**
i2.8xlarge	1,146$	**199$**	**750$**	**544$**	**685$**
hs1.8xlarge	773$	**199$**	**750$**	**544$**	*913$*

Como se puede observar, la tendencia actual en las instancias I2 es usar unidades de estado sólido SSD con el fin de primar el rendimiento sobre la capacidad de almacenamiento.

Tomando estas instancias como referencia, el coste por semana de sistemas equivalentes proporcionados por el resto de proveedores se muestra en la tabla de la Figura 22.

Se han resaltado en negrita los precios más bajos respecto a los de referencia y en cursiva los más altos. De acuerdo con estos datos se aprecia que tanto Google como Azure ofrecen precios más competitivos que Amazon, mientras que HP Cloud y Rackspace requieren más inversión para las dos instancias más pequeñas.

Esta información concuerda con algunos estudios realizados. Por ejemplo, la revista Computerworld, en su edición australiana, hizo una comparativa Amazon vs. Google vs. Azure en cuanto a rendimiento en marzo de 2014[61], en el que concluye que Google es la opción de menor coste, seguida por Azure, siendo en general Amazon el proveedor más caro. Este estudio realiza además varias pruebas de rendimiento en las que muestra que Google es el que proporciona en general las mejores prestaciones.

Configuración de servicios en la nube

La gran cantidad de configuraciones posibles que ofrecen los distintos proveedores hace que buscar la opción más adecuada no sea una tarea trivial. La presencia de herramientas de gestión de gastos en la nube, puede ayudar a tener una visión de conjunto de las distintas opciones.

En cualquier caso, es fundamental tener muy claros los requisitos que requieren los sistemas o aplicaciones a desplegar en la nube para hacer una elección correcta.

[61] Peter Wayner. *Amazon vs. Google vs. Windows Azure: Cloud Computing Speed Showdown*. http://www.computerworld.com.au/article/539633/amazon_vs_google_vs_windows_azure_cloud_computing_speed_showdown/. [Online; publicado el 4 de marzo de 2014]

3 Big Data en las Organizaciones

3.1 ¿Es Siempre Adecuado el Uso de Big Data?

En unos años Big Data se ha hecho omnipresente. De forma generalizada se habla de sus bondades, pues al combinar el poder de la computación moderna con la abundancia de datos digitales, parece como si cualquier tipo de problema pudiera resolverse simplemente analizando en profundidad los datos. En este sentido uno de los más optimistas defensores de Big Data, Patrick Tucker[62] nos dice que en dos décadas seremos capaces de predecir grandes áreas del futuro con mucha más precisión que nunca antes en la historia, incluyendo acontecimientos que se ha creído que están más allá del reino de la inferencia humana.

Pero, ¿realmente Big Data es algo tan bueno como parece? No hay duda de que Big Data es una valiosa herramienta, que ya ha tenido un impacto fundamental en ciertas áreas y que en un futuro inmediato su campo de aplicación será aún mayor. Pero precisamente por ello conviene preguntarse si siempre es adecuado el uso de Big Data y cuáles son sus limitaciones[63]. En este sentido el análisis se va a desarrollar en dos fases, en primer lugar, se presenta una serie de limitaciones o cautelas generales relacionadas con el uso de Big Data, y en segundo lugar se realiza un estudio específico de las cautelas que se deben tener cuando se trabaja con las redes sociales, fundamentalmente desde la perspectiva del marketing.

---

[62]Patrick Tucker. *The Naked Future: What Happens in a World that Anticipates Your Every Move*. Amazon, 2014

[63]Xavier Guiteras Vila. *5 Críticas al Empleo del Big Data*. http://www.investigacionmercados.es/5-criticas-al-empleo-del-big-data/. [Online; publicado el 30 de junio de 2014]

3.1.1 Cautelas al Utilizar Big Data

Para presentar las cautelas que se deben tener al utilizar Big Data los hechos que se van a considerar son los siguientes: la necesaria diferenciación entre causalidad y correlación, reconocer Big Data como una mera herramienta, la imposibilidad de sustituir el trabajo humano, los errores en las predicciones mecánicas, las limitaciones derivadas de utilizar datos autogenerados y los problemas que surgen cuando nos enfrentamos a acontecimientos poco frecuentes.

3.1.1.1 Correlación no equivale a causalidad

Las grandes cantidades de datos utilizadas en Big Data son especialmente efectivas para encontrar correlaciones entre acontecimientos. Pero todas estas correlaciones no tienen por qué estar fundamentadas en una relación causa-efecto o en una asociación de variables mediada por una tercera variable. Pueden deberse, simplemente, a la casualidad. Para evidenciar que la correlación entre dos variables no significa necesariamente que una causa a la otra, Tyler Vigen ha elaborado una serie de gráficos que muestran una amplia batería «correlaciones espurias». Una correlación espuria se produce cuando dos cosas –como la tasa de divorcios en Maine y el desplome del consumo de margarina de dicho estado– parecen estar relacionados, pero en realidad no hay una relación causa efecto. En este sentido, uno de los riesgos en el análisis de datos es que a partir de correlaciones espurias se hagan inferencias falsas. Un ejemplo de ello sería cuando al analizar las calificaciones de lectura en un colegio, se encuentra que el tamaño del zapato de un niño aparece como un buen predictor de su puntuación; lo que parece no tener mucho sentido. Una explicación podría ser que se agruparon los datos de todos los niños, desde los de primer grado hasta los alumnos de sexto grado. Lógicamente los niños mayores podían leer mejor, y lógicamente también tenían zapatos más grandes.

Así, pues, Big Data es muy bueno en la detección de correlaciones, correlaciones especialmente útiles cuando se tratan grande cantidades de datos y que un análisis de conjuntos más reducidos de datos podría no detectar, pero nunca nos dirá cuáles son las correlaciones que realmente son significativas. Si alguien analizara bases de Big Data sin aplicar el sentido común o sus conocimientos científicos, podría llegar a la conclusión, por ejemplo, de que existe una clara relación positiva entre el gasto público en I+D y la cantidad de suicidios llevados a cabo por ahorcamiento, tal como se desprendería de las correlaciones encontradas por el citado Tyler Vigen. Cuando se analiza la correlación entre muchas variables siempre nos arriesgamos a encontrar, por pura casualidad, correlaciones falsas que aparecen como estadísticamente significativas, a pesar de que no hay ninguna conexión real entre las variables. Por ello, la falta de una supervisión cuidadosa de las variables analizadas por Big Data puede ser una importante fuente de errores.

En cualquier caso, no es un problema de Big Data, esto es algo que se aplica al análisis estadístico entre las variables: «correlación no equivale a causalidad».

3.1.1.2 Big Data es una herramienta muy importante, pero no es la investigación en sí misma

En segundo lugar, Big Data puede funcionar bien como un complemento a la investigación científica, pero difícilmente puede actuar como un sustituto de la misma. Como señalan Marcus y Davis[64], a los biólogos moleculares, por ejemplo, les gustaría mucho ser capaces de deducir la estructura tridimensional de las proteínas simplemente a partir de su secuencia de ADN subyacente. Los científicos que trabajan en el problema utilizan grandes volúmenes de datos como una de sus muchas herramientas. Pero ningún científico piensa que puede resolver este problema mediante el simple análisis de los datos. Por muy poderosas que sean las herramientas de análisis estadístico; la base del análisis debe radicar en una sólida comprensión de la física y la bioquímica.

Desde una perspectiva general cabe afirmar que la investigación no consiste únicamente en manejar los datos; los investigadores tienen que aportar luz a los datos gracias a la experiencia acumulada y al conocimiento sobre la materia investigada, las personas y su comportamiento.

3.1.1.3 Big Data no sustituye al trabajo humano: a las máquinas se les puede hacer trampas

En tercer lugar, muchos algoritmos que se utilizan en Big Data pueden ser fácilmente burlados. Por ejemplo, al analizar cómo se evalúan en Estados Unidos los trabajos (*essays*) de los estudiantes a través del Big Data se llegó a la conclusión de que las variables que estaban más correlacionadas con una buena nota son la longitud de la oración y lo sofisticadas que fueran las palabras utilizadas. Cuando los estudiantes descubrieron cómo funciona el algoritmo, comenzaron a escribir frases largas y a usar palabras sofisticadas, en lugar de aprender cómo estructurar correctamente las frases y escribir textos claros y coherentes.

Otro ejemplo de este tipo de situaciones es que incluso el prestigiado motor de búsqueda de Google, con razón considerado como un gran caso de éxito de Big Data, no es inmune a los «bombardeos de Google» y «*spamdexing*», técnicas astutas para elevar artificialmente el posicionamiento web o SEO (*Search Engine Optimization*).

Así, pues no podemos dejar en manos del Big Data aquellas tareas que necesitan de la interpretación humana. El funcionamiento a través de patrones lógicos siempre deja una vía a aquellos que buscan alcanzar un objetivo de forma no meritoria u honesta; a las máquinas se le puede hacer trampas y corresponde a las personas detectarlas. En cualquier caso, Big Data es una herramienta que puede ser un excelente apoyo y facilitar el trabajo humano del analista.

---

[64] Gary Marcus y Ernest Davis. *Eight (No, Nine!) Problems with Big Data*. http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html. [Online; publicado el 6 de abril de 2014]

3.1.1.4 Las predicciones de Big Data no son infalibles

En cuarto lugar, incluso cuando los resultados de un análisis de Big Data no estén falseados intencionalmente, a menudo resultan ser menos robustos de lo que inicialmente parecen. Recuérdese el caso Google Flu Trends (véase capítulo 2), inicialmente considerado como un sorprendente caso de éxito de Big Data. En 2009, Google informó que mediante el análisis de las consultas de búsqueda relacionadas con la gripe, se había podido detectar la propagación de la gripe con mayor precisión y más rápidamente que los Centros para el Control y la Prevención de Enfermedades. Unos años más tarde, sin embargo, Google Flu Trends comenzó a tambalearse; y durante algunos años ha hecho más predicciones malas que buenas. Las estimaciones de Google Flu Trends han sobrestimado el número de pacientes que sufrirían esta enfermedad.

Como en un reciente artículo en la revista Science se ha explicado[65], una de las causas principales de los fracasos de Google Flu Trends puede haber sido que el motor de búsqueda de Google en sí cambia constantemente, de tal manera que los patrones observados en los datos recogidos en un momento determinado, no se aplican necesariamente a los datos recogidos en otro momento. Además, hay evidencia de que Google Flu Trends no está usando toda la información de que dispone para hacer mediciones precisas de la incidencia de la gripe. En este sentido, en el caso que estamos considerando quizás el problema real no sea el empleo del Big Data, sino el uso que de él se hace. Si se hubiera construido un algoritmo más preciso y eficiente, fundamentado en una teoría robusta, se hubieran podido minimizar las desviaciones.

Como el estadístico Kaiser Fung[66] ha señalado, los datos utilizados en Big Data que se obtienen de la web, en ocasiones mezclan datos recogidos de diferentes maneras y con diferentes propósitos, de forma que a veces el resultado puede que no sea muy adecuado. Por ello, puede ser arriesgado extraer conclusiones a partir de conjuntos de datos si estos no se han sometido a un análisis y «limpiado» apropiado.

Desde una perspectiva más general Kaiser Fung señala que Big Data es:

- Observacional: gran parte de los nuevos datos provienen de los sensores o dispositivos de seguimiento que se supervisan continuamente y de manera indiscriminada y sin diseño, en lugar de tener su origen en cuestionarios, entrevistas, o experimentos con diseño específico
- Está falto de controles: no hay controles propiamente dichos, lo que hace difícil realizar comparaciones y análisis válidos.

[65] David Lazer y col. «Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season». En: *SSRN* 2408560 (2014)

[66] Kaiser Fung. *Google Flu Trends' Failure Shows Good Data >Big Data*. Ed. por Harvard Business Review. https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data/. [Online; publicado el 25 de marzo de 2014]

- Aparentemente completo: la disponibilidad de datos para la mayoría de las unidades mensurables y el enorme volumen de datos generado no tiene precedentes, pero contar con muchos datos puede crear pistas falsas y callejones sin salida, lo que puede complicar la búsqueda de estructuras con sentido y predecibles.
- Adaptado: son terceras personas las que recopilan los datos, a menudo con unos fines no relacionados con los científicos de datos, lo que puede generar dificultades para su interpretación.
- Fusionado: diferentes conjuntos de datos se combinan, lo que puede agravar los problemas relacionados con la falta de definición de objetivos y acoplamiento entre Big data y el negocio.

Esta visión de Big Data puede parecer que no es muy optimista pero en realidad el sentido de las cautelas de Kaise Fung es que siempre debemos estar atentos a los retos que plantea el análisis de datos y aplicar mucha prudencia en la interpretación de los resultados. Y esto es aplicable a cualquier trabajo que tenga su soporte en el análisis de los datos; en otras palabras estas cautelas son también aplicables a la estadística y a la econometría.

3.1.1.5 Un círculo vicioso: la fuente de información de Big Data es un producto de Big Data

Una quinta preocupación podría ser el llamado efecto de eco-cámara, una especie de círculo vicioso en el que se reproducen, en cada etapa con más vigor, los errores del pasado. El origen de este problema proviene de que Big Data analiza la gran mayoría de las veces la información disponible en la red que es, al mismo tiempo, una fuente *per se* de Big Data. Dado que la fuente de información para el análisis de Big Data es en sí misma un producto de Big Data, las posibilidades de que aparezcan círculos viciosos son bastante frecuentes. Considérese el caso de los programas de traducción como Google Translate, que se basan en muchos pares de textos paralelos de diferentes idiomas –por ejemplo, la misma entrada de la Wikipedia en dos idiomas diferentes– para descubrir los patrones de traducción entre los idiomas. Esta es una estrategia perfectamente razonable, excepto por el hecho de que con algunos de los idiomas menos comunes, muchos de los propios artículos de Wikipedia pueden haber sido escritos utilizando Google Translate. En esos casos, los errores iniciales en el traductor Google infectan a Wikipedia,... mientras que Google Translate, posteriormente, habiendo analizado el contenido de Wikipedia, refuerza el error y se orienta hacia una traducción siempre equivocada del texto en cuestión.

3.1.1.6 Problemas de análisis ante acontecimientos poco frecuentes

Finalmente, Big Data está en un entorno apropiado cuando se analizan patrones de comportamiento que son comunes, pero a menudo no da la talla cuando se trata de analizar

las cosas que son menos frecuentes y lo que se pretende es descubrir patrones poco comunes o que están empezando a generar una corriente de viralidad. Por ejemplo, los programas que utilizan Big Data para analizar textos, como los motores de búsqueda y los programas de traducción, a menudo se basan en buena medida en algo que se llama trigramas: secuencias de tres palabras en una fila (como «en una fila»), que son la base a partir de la cual los software de traducción realizan su tarea. Información estadística fiable puede compilarse sobre trigramas comunes, precisamente porque aparecen con frecuencia. Pero ningún cuerpo existente de datos nunca será lo suficientemente grande como para incluir a todos los trigramas que la gente podría utilizar, debido a la continua inventiva del lenguaje.

3.1.1.7 Balance final de las limitaciones de Big Data

Algunos de los problemas relacionados con el empleo del Big Data son superables. De la propia utilización de Big Data y de sus fracasos se puede aprender cómo mejorar los algoritmos o cómo evitar que los usuarios «manipulen» las reglas de los robots para conseguir las salidas deseadas, por ejemplo combatiendo el *spamdexing* o el *Google Bombing*. En cualquier caso, para que Big Data logre superar todos los posibles problemas y pueda alcanzar los resultados deseados la clave está en la labor del científico de los datos. A éste le corresponde una adecuada tarea de recolección y limpieza de datos, un correcto almacenamiento, un riguroso procesamiento y análisis y, por último, una esmerada labor orientada a facilitar la visualización e interpretación de los resultados por parte del consumidor de los datos. Sin el científico que gestione e interprete adecuadamente la información, el Big Data no sirve de mucho. En cualquier caso, no debe olvidarse que al Big Data no se le debe pedir más de lo que puede ofrecer; es un importante apoyo para el investigador o analista, pero no el agente en sí mismo.

Acercándonos a la realidad

Tradicionalmente según se señala en «El BIG DATA o el verdadero valor de la información»[67] los tres motivos aducidos por las empresas para no implantar Big Data son:

- La falta de una infraestructura adecuada que presume de gran complejidad.
- La falta de habilidades para exportar un sistema de información para obtener información válida de la red.
- La barrera de entrada típica: el coste económico.

❖ ***¿En qué sentido estos motivos actualmente han perdido buena parte de su vigencia?***

[67] Papeles de Inteligencia. *EL BIG DATA o el Verdadero Valor de la Información.* http://papelesdeinteligencia.com/big-data/. [Online; consultado el 10 de diciembre de 2015]

3.1.2 Limitaciones al Empleo de Big Data en las Redes Sociales

En esta sección se presentan una serie de consideraciones sobre las limitaciones del empleo de Big Data como herramienta de análisis de la redes sociales, especialmente de cara a su utilización en el marketing[68].

3.1.2.1 Los perfiles de las redes sociales

La utilización de los datos de las redes sociales para identificar los perfiles de los clientes objetivo se ve limitada en la práctica por la información que los usuarios de las redes sociales opten por compartir sobre sí mismos y los títulos que ellos elijan para autoproclamarse (*self-proclaimed titles*).

Por otro lado, hay que tener en cuenta que Twitter está infestado con cuentas spam «robot» y que Facebook estima que tiene 83 millones de cuentas falsas. Dada esta situación de hecho, resulta altamente probable que cualquier conjunto de datos de las redes sociales contenga cuentas falsas.

En cualquier caso, y al margen de la calidad de los datos de partida, las empresas quieren tener identificado el perfil de sus clientes, desean poder segmentarlos, y están interesadas en «escuchar» lo que desea su base de clientes. De esta forma, podrán saber quiénes son, qué les gusta y a quién están conectados. El resultado será contar con una segmentación más rica de su cartera de clientes y poder planificar su estrategia comercial de forma más incisiva.

Para el logro de estos objetivos, el buen analista deberá identificar las limitaciones de los datos que utiliza y ser cauto al elaborar sus predicciones.

3.1.2.2 La cartografía de la conversación social por zonas geográficas: información estadísticamente no significaiva

Según Pulsar[68], en una plataforma de inteligencia de datos sociales, menos del 5 % de los usuarios de medios sociales ponen a disposición del público su ubicación en sus perfiles. Esto significa que, si bien existe la tecnología para permitir a los medios de comunicación social hacer el «mapeo» de las conversaciones, los escasos datos que se tiene de los usuarios tienden a hacer que los resultados sean estadísticamente poco significativos.

A pesar de estas limitaciones, las organizaciones quieren poder señalar donde está situado su público objetivo, pues con esta información resulta más fácil asignar el gasto de marketing y el patrocinio de eventos y será más probable obtener un mayor retorno a la inversión (ROI). Si la información sobre la ubicación es estadísticamente relevante,

[68]Anna Lawlor. *Five Inconvenient Truths about Social Data*. Ed. por The Guardian. http://www.theguardian.com/media/2014/aug/04/five-inconvenient-truths-social-data-marketers. [Online; publicado el 4 de agosto de 2014]

este dato podría tener enormes consecuencias para el presupuesto de la organización y la asignación de recursos.

3.1.2.3 Dificultades para analizar de forma automática los sentimientos

Los proveedores de datos sociales de forma generalizada señalan que el análisis automatizado de sentimientos tiene entre el 70 % y el 80 % de precisión. Sin embargo, una investigación de FreshMinds, una consultora de prospectiva e innovación, encontró que estos porcentajes puede enmascarar lo que realmente suele ocurrir[69]. En un caso piloto realizado en Starbucks, se observó que aproximadamente el 80 % de todos los comentarios encontrados eran de naturaleza neutral. Eran meras declaraciones de hechos o informaciones sin manifestar sentimientos positivos o negativos. Este porcentaje es consistente con lo manifestado por muchas otras empresas y en términos generales puede decirse que la mayoría de las conversaciones online son neutrales. Este tipo de conversaciones tienen menor interés para una empresa que quiere tomar una decisión o realizar una acción en base a lo que se dice online. Para las empresas las conversaciones que pueden calificarse como positivas o negativas tienen mucha mayor importancia y es aquí donde el análisis automatizado de los sentimientos realmente puede fallar.

Es difícil distinguir automáticamente entre conversaciones positivas y negativas

Las dificultades surgen cuando se observa que al eliminar las frases neutrales, las herramientas automatizadas por lo general analizan los sentimientos incorrectamente. Cuando se comparan con un analista humano, las herramientas presentan unos resultados que reflejan, como media un 30 % de aciertos al decidir si una declaración era positiva o negativa. Para cualquier empresa que pretenda utilizar la monitorización de las redes sociales para ayudarle a interactuar y responder a los comentarios positivos o negativos estos resultados no son aceptables. De hecho, dado el bajo porcentaje de aciertos, muy a menudo, un comentario positivo puede ser clasificado como negativo o a la inversa.

¿Por qué este fallo preocupa a las empresas?

Estas limitaciones en el análisis automatizado de los sentimientos pueden causar problemas reales para las empresas, sobre todo si basan algunos de sus procesos internos en la monitorización de las redes sociales. Por ejemplo, imagine que usted envía todas sus conversaciones negativas a que las analice su equipo de atención al cliente para contestarles a los que formularon los comentarios negativos. Pero si resulta que dos tercios (o quizás más) de las conversaciones «negativas» que son contestadas, en realidad eran positivas, entonces este proceso comienza a no tener sentido. Además, y quizás esto sea lo más importante, resultará que una gran cantidad de las conversaciones que efectivamen-

[69] Matt Rhodes. *The Problem with Automated Sentiment Analysis*. Ed. por FreshMinds. http://www.freshminds.net/2010/05/the-problem-with-automated-sentiment-analysis/. [Online; publicado el 28 de mayo de 2010]

te eran negativas nunca llegarán al equipo de atención al cliente, pues incorrectamente habían sido clasificadas como positivas. Lógicamente, a estos clientes insatisfechos no se les canalizará hacia las personas adecuadas y no se conseguirá que sus problemas sean abordados. Este tipo de fallos del análisis automatizado de los sentimientos es importante porque las organizaciones utilizan los datos sociales como indicador adelantado y como indicador retrasado.

- Como un indicador adelantado; el análisis de los sentimientos puede ayudar a predecir los resultados (desde acontecimientos políticos hasta lanzamientos de productos al mercado) y proporcionar información en tiempo real de cuáles son los sentimientos de la gente en relación con una marca y sus productos – y cómo estos sentimientos se comparan con los experimentados ante los productos de los competidores de la marca.

- Como un indicador retrasado; el análisis de los sentimientos puede ayudar en una amplia variedad de temas, desde medir el ROI de una campaña de marketing a poner de relieve la insatisfacción con «puntos de contacto» con los clientes.

En última instancia, una marca puede hacer bien en utilizar un proveedor de datos para tratar los mensajes «neutrales» y emplear a un ser humano para tamizar los *posts* restantes y así obtener una visión más ajustada a la realidad.

Entonces, ¿qué podemos hacer?

Ante la imprecisión mostrada por las herramientas de monitorización de las redes sociales investigadas puesta de manifiesto por los investigadores de FreshMinds[69], pero dado el interés de las empresas en tomar decisiones o realizar acciones sobre la base de que una conversación sea positiva o negativa, la pregunta que cabe formular es ¿qué se puede hacer?

Por supuesto, hay muchas cosas que se pueden hacer y con el tiempo las herramientas pueden prepararse para aprender y para mejorar la forma en que evalúan las conversaciones referidas a una empresa determinada. En cualquier caso, la tarea no es sencilla pues la lingüística de los humanos es compleja y los comentarios pueden interpretarse de forma abierta, sobre todo cuando no se conoce ni el contexto y ni el posible sarcasmo utilizado. En este sentido, debe destacarse la importancia de las investigaciones sobre el Procesamiento del Lenguaje Natural (PLN), cuyo objetivo es enseñar a un ordenador a entender cómo se comunican las personas y para ello se crean tecnologías específicas. El PLN trata de diseñar mecanismos para comunicarse que se puedan realizar por medio de programas que ejecuten o simulen la comunicación.

3.1.2.4 Las palabras clave y búsquedas temáticas pueden resultar decepcionantes

La investigación del contenido de los mensajes de las redes sociales debe dar una idea de la naturaleza de las conversaciones online pertinentes. Sin embargo, posiblemente debido a la gran cantidad de mensajes «neutrales» y retuits, las palabras clave y los temas de búsqueda pueden resultar decepcionantes y proporcionar una gran cantidad de palabras no relacionadas, como «tiempo», «ayuda» y «2015», que sin contexto no tienen sentido.

Por otro lado, la mayoría de los comentarios de los medios sociales no son acerca de las marcas o servicios, y cuando se refieren a ellos tienden a representar visiones extremas (la muy buena o muy mala). Esto deja grandes agujeros en la información recopilada.

Sin embargo, las organizaciones quieren saber lo que su público objetivo realmente están diciendo, la terminología que utilizan y lo que les importa. En muchos casos, la palabra clave y los temas de búsqueda prometen el mundo pero poco entregan.

Podría decirse que la recopilación de datos global puede proporcionar un contexto muy necesario pero los datos por sí solos no pueden sustituir una inversión significativa de horas de análisis humanos; en parte dedicado a depurar y limpiar los datos.

3.1.2.5 Cambios continuos en la dinámica de las redes sociales

Es un hecho que la realidad social está experimentando un cambio continuo. En primer lugar, los datos demográficos de los usuarios que entran en las redes sociales cambian. En segundo lugar, las propias redes sociales cambian; nuevas redes sociales surgen y otras desaparecen. Y en tercer lugar, el comportamiento de los usuarios evoluciona. El efecto global de estos cambios es que resulta difícil hacer comparaciones significativas entre los conjuntos de datos sociales. Los medios de comunicación social están en constante cambio; y no de una forma consistente, de forma que la precisión de las mediciones de las redes sociales es limitada, algo que los científicos de datos deben tener en cuenta cuando realizan sus análisis[68].

Acercándonos a la realidad

En el artículo «La nueva ciencia de las ciudades y las limitaciones del Big Data» de Manu Fernández publicado el 5 de diciembre de 2014, se abordan temas como las *smart cities*, la ciencia de las ciudades y Big Data[70].

- ***¿En qué sentido el movimiento del Big Data y su capacidad de análisis de sentimientos puede ofrecer nuevas oportunidades para el estudio de las ciudades?***

[70] Manu Fernández. *La Nueva Ciencia de las Ciudades y las Limitaciones del Big Data.* http://www.ciudadesaescalahumana.org/2014/12/la-nueva-ciencia-de-las-ciudades-y-las.html. [Online; publicado el 5 de diciembre de 2014]

3.2 Cultura Analítica en la Organización: Data Driven Business

3.2.1 Introducción

Uno de los principales problemas que se deben resolver al implantar una estrategia de Big Data no es ni de datos, ni tecnológico; las barreras de adopción más difíciles de superar son de gestión y culturales. Los directivos deben convencer a toda la organización de las ventajas que se derivan de basar la gestión de la empresa en el análisis de los datos. Es muy importante crear una cultura analítica, en la que se admita que la toma de decisiones debe basarse en el análisis de los datos. En este contexto, Big Data es un instrumento fundamental para dar paso a un nuevo tipo de empresa, la empresa guiada (o dirigida) por los datos (*data-driven business*).

3.2.2 La Empresa Guiada por los Datos

Muchos estudios sobre la importancia del Big Data se han centrado en los exitosos modelos de negocio basados en grandes empresas estadounidenses como Facebook, Google o Twitter. Pero la revolución del Big Data no vendrá sólo por el uso de grandes cantidades de datos por parte de las grandes empresas, sino cuando muchas pequeñas y medianas empresas se decidan a utilizar las grandes cantidades de datos disponibles en la actualidad. De hecho, de forma progresiva, cada vez más compañías están adoptando Big Data, y está apareciendo toda una nueva generación de *startups* caracterizadas por el uso intensivo de los datos. Estas empresas, a su vez están actuando como pioneras de nuevos modelos de negocio. El análisis de estas nuevas empresas ofrece lecciones importantes sobre el papel que el Big Data está jugando y jugará en la economía del mañana.

Recientemente se ha puesto de moda el concepto de toma de decisiones de gestión basada en los datos (DDDM – *Data-Driven Decision Management*) como un nuevo enfoque de la gestión empresarial que valora aquellas decisiones respaldadas con datos que pueden ser verificados. El enfoque basado en datos está ganando adeptos dentro de la empresa, por el aumento de la disponibilidad de una mayor cantidad de datos, junto con la cada vez mayor presión y la creciente competencia que caracteriza a los mercados.

La toma de decisiones basada en datos se ha convertido en una fuente de ventaja competitiva para las empresas. Un estudio del MIT[71] para Digital Business reveló que aquellas organizaciones que basan la toma de decisiones en los datos, tenían unos mejores índices de productividad, un 4 % superiores a la media y unos beneficios superiores, un 6 % por encima de la media.

Data-driven disaster

Sin embargo, el éxito del enfoque basado en los datos depende de dos factores: de la calidad de los datos recogidos y de la eficacia de su análisis e interpretación. Los errores

[71] Andrew McAfee y Erik Brynjolfsson. *Big Data: The Management Revolution*. Harvard Business Review, 2012

pueden introducirse en los procesos de análisis de datos en cualquier etapa y ser arrastrados a lo largo de todo el proceso, lo que puede tener graves consecuencias. Es lo que se conoce como *data-driven disaster*, o lo que podríamos traducir como desastre generado por los datos.

Según el Instituto de Data Warehousing, los problemas de calidad de datos cuestan a las empresas en los Estados Unidos más de 600 mil millones de dólares al año. Además del impacto económico, los problemas con la calidad y el análisis de los datos pueden tener un serio impacto en otras áreas como la seguridad, el cumplimiento de las normas, la gestión de proyectos y la gestión de recursos humanos, entre otras.

Los errores pueden introducirse en el análisis de datos en cualquier momento. En primer lugar, la calidad de los datos puede ser insuficiente,, la información puede ser incompleta, inexacta, no actualizada, o puede que no represente de forma fiable lo que buscamos. En las fases de análisis de datos e interpretación también pueden surgir dificultades: factores que introducen ruido y pueden provocar confusión, modelos matemáticos defectuosos o inapropiados, correlaciones que erróneamente sugieren causalidad, atribución errónea de significación estadística, cuando en realidad los datos no lo soportan.

Hay muchos ejemplos, tanto en el mundo empresarial, como en otros ámbitos, de errores cometidos como consecuencia de toma de decisiones basadas en procesos guiados por datos. Dos ejemplos trágicos, reconocidos incluso por sus responsables, son: la explosión del transbordador espacial Challenger en 1986 y el derribo de un Airbus iraní por el USS Vincennes en 1988.

En la mayoría de los casos conocidos de decisiones erróneas basadas en datos, se repiten dos situaciones: insuficiencia de recursos dedicados a los procesos de datos y un exceso de confianza en la validez de los mismos. Por tanto, para evitar o minimizar los errores, es importante poner en práctica procesos continuos que permitan examinar y evaluar tanto la calidad de los datos, como los procesos analíticos, y no olvidar que hay que seguir prestando atención al sentido común. Cuando los datos parecen estar indicando algo que no tiene sentido lógico o simplemente parece estar mal, es el momento de cuestionar los resultados y volver a examinar los datos y los métodos de análisis utilizados.

Transformar una organización siempre es un reto importante y no exento de problemas, realizar esta transformación en la forma en la que se toman las decisiones, por supuesto no es una excepción. Sin embargo la mayoría de las organizaciones pretenden realizar esta transformación mediante la tecnología, adquiriendo más y mejores recursos tecnológicos: nuevos y mejores paquetes de software o mejores equipamientos de hardware. Sin embargo, como hemos repetido en más de una ocasión, el problema es mucho más complejo que el mero hecho de invertir en tecnología; orientar a la compañía para que las decisiones sean tomadas teniendo en cuenta los datos objetivos, requiere de un cambio cultural profundo.

Convertirse en una organización dirigida o guiada por los datos requiere no sólo una inversión en tecnología sino también mucho aprendizaje y entrenamiento. Las empresas necesitan adoptar un enfoque más científico en la toma de decisiones. Un enfoque en el que los datos son considerados como un activo importante de la empresa, facilita y anima a poner en práctica la teoría, la investigación e incluso la experimentación, fomentando su uso a lo largo de todos los estamentos de la empresa.

Tener acceso a grandes volúmenes de datos ha abierto las puertas a formas de medir que nunca habían sido posibles antes. Sin embargo, tener datos no es suficiente para que una empresa se convierta en «*data-driven*». Esta transformación, no sucede por el toque de la varita mágica de la tecnología. La toma de decisiones basada en el análisis de los datos debe convertirse en una filosofía que forma parte de la naturaleza de la compañía, un hábito integrado en la forma de:

- pensar,
- tomar decisiones,
- discutir, y
- en las actividades cotidianas.

Por lo tanto, es razonable pensar que aquellas empresas que, de forma satisfactoria, toman sus decisiones basándose en los datos, tienen un comportamiento determinado o ponen en práctica acciones que les distinguen de aquellas que no lo hacen. O visto de otro modo, si queremos transformar nuestra empresa, orientándola hacia la toma de decisiones basada en datos, debemos identificar aquellos comportamientos y prácticas de empresas que ya tienen implementada una cultura analítica.

Comportamientos que facilitan la toma de decisiones basada en los datos

Pero ¿existen realmente esos comportamientos? ¿Hay alguna forma de construir una cultura analítica en la empresa? No existe ninguna ciencia exacta que podamos exponer, sin embargo sí que hay unos comportamientos que podemos identificar y si se ponen en práctica ayudan a conseguir que una organización construya una cultura analítica:

- **Alguien es responsable de la calidad de sus datos.** Hay alguien en la empresa que tiene la responsabilidad de garantizar que los datos recogidos y comunicados tienen la calidad adecuada, lo que genera confianza en los datos en el resto de la organización.
- **En cualquier nuevo proyecto, la obtención de datos es una de las primeras prioridades.** Antes incluso de escribir una sola línea de código, se hace el planteamiento acerca de las preguntas necesarias que permitan establecer las normas de etiquetado que permitan la recolección de datos, de manera que una

vez que el proyecto se pone en producción, se dispone de datos que permiten el análisis del rendimiento.

- **Cuando alguien da una opinión, la acompaña de números y datos.** Si alguien da una opinión que no está fundamentada en datos, lo adecuado es preguntar en que datos fundamenta su opinión. Las decisiones no se deben tomar basándose en prejuicios, ideas preconcebidas o creencias, sino en hechos reales; lo que reduce el riesgo.

- **Los datos se comunican siempre, aunque muestren malos resultados.** Cuando el rendimiento es inferior a la media, nadie trata de ocultarlo, de forma que, se sabe en qué áreas hay que mejorar.

- **Todo el mundo puede acceder a los datos que tienen relación con su trabajo.** Todos los empleados tienen acceso a los datos que tienen relación con las actividades de su grupo de trabajo, de manera que todos los miembros del equipo puedan entender las fortalezas y debilidades del mismo.

- **Cada objetivo tiene un indicador y un valor a alcanzar.** Cuando se asignan objetivos a un equipo o persona, se conocen los indicadores que permiten juzgar si se ha alcanzado el objetivo. De este modo, los empleados saben si se logran los objetivos y tienen una mejor visión de su contribución a la consecución de los objetivos.

- **Los equipos reciben formación en análisis de datos.** Cada empleado ha sido entrenado en el análisis e interpretación de los datos. De esta manera, se consigue que los empleados confíen en su capacidad en las tareas de interpretación de datos, conocen los métodos que deben emplear para analizarlos y saben actuar en consecuencia.

- **Los proyectos que implican recolección de datos no tienen problemas para recibir financiación.** Cuando la compañía se plantea proyectos en los que se ve la posibilidad de obtener nuevos datos, el presupuesto necesario es fácil de conseguir. Lo que demuestra que la dirección valora y apoya la obtención de datos para su posterior uso en la toma de decisiones.

- **Los datos nunca se utilizan para señalar un culpable en caso de fallo.** Cuando algo falla, los datos también se comparten, no se trata de identificar al responsable ni de generar sentimiento de culpa, sino de compartir el conocimiento que permita mejorar. Los empleados tienen menos miedo a asumir riesgos, y son menos reticentes a utilizar los datos, aunque estos puedan revelar un desempeño potencialmente menor que el de la media.

- **La realización de pruebas es algo natural.** Antes de que se adopten medidas correctivas, se llevan a cabo pruebas con una parte de la población para comprobar los efectos que estas pueden tener en la mejora del rendimiento. Los

datos justifican y validan los proyectos de mejora que se llevan a cabo y permiten adquirir una mejor comprensión del comportamiento de los usuarios y su motivación.

No se debe confundir o malinterpretar el concepto «*data-driven*». Las empresas deben utilizar los datos para comprobar si están yendo en la dirección adecuada o si están alcanzando los objetivos previamente establecidos y si están muy cerca o muy lejos de alcanzarlos. En ningún caso los datos deben dirigir o establecer los objetivos de la compañía. Las compañías se dirigen de acuerdo con las estrategias planificadas o la visión establecida por sus dirigentes y ejecutadas por sus empleados. En este sentido, podríamos decir que las empresas deben estar dirigidas o guiadas por la estrategia o por la visión e informadas por los datos.

3.2.3 Big Data y la Analítica de Datos

Big Data se utiliza, por un lado, para lograr mejoras incrementales en las prácticas actuales y, por otro, para tratar de optimizar la prestación de servicios. Ejemplos de este tipo de aplicaciones de Big Data podrían ser procurar la optimización de la relación con el cliente, la detección de fraude o la reducción de la tasa de abandono (*churn rate*). Por otro lado, se puede innovar sobre nuevos productos y modelos de negocio o mejorar los ya existentes, basándonos en la información generada mediante el uso de Big Data.

Los datos pueden actuar como un nuevo factor de producción al impulsar un amplio abanico de actividades innovadoras. La analítica de datos permite pasar de la mera descripción, a la predicción y de esta, al análisis prescriptivo, lo que representa un nuevo y potente instrumento para facilitar la toma de decisiones. En este sentido, recuérdese que el análisis descriptivo incluye la presentación de informes de negocios y las respuestas a preguntas del tipo; ¿Qué pasó? y / o ¿Qué está pasando? El análisis predictivo se refiere a la utilización de técnicas de aprendizaje de máquina (*machine learning*) y modelos matemáticos para predecir el resultado futuro de ciertos acontecimientos, dados los datos existentes. El análisis preceptivo, por su parte, busca determinar la toma de decisiones óptimas, dado un complejo conjunto de objetivos, requisitos y restricciones.

Muchas *startups* están recurriendo al análisis de datos y están experimentando con nuevos modelos de negocio impulsados por los datos. Cada vez es más común encontrar nuevas aplicaciones y servicios basados en los datos, que contribuyen a hacer la vida más fácil y a costes más bajos. De esta forma el análisis de los datos está contribuyendo a cambiar nuestros hábitos de conducta y la forma en que nos relacionamos.

Acercándonos a la realidad

En el artículo: «Big data: una 'revolución industrial' en la gestión de los datos digitales»[72] se señalan algunos ejemplos de Big Data: las consultas y resultados de los motores de

[72] Fidelity. *Big Data: una 'Revolución Industrial' en la Gestión de los Datos Digitales*. 2012

búsqueda, los datos de las redes sociales (como los tuits), los datos meteorológicos, los datos astronómicos, la vigilancia militar, los datos económicos y bursátiles, los historiales médicos, los experimentos físicos (Gran Colisionador de Hadrones), los archivos fotográficos, la radio y la televisión, los vídeos (CCTV y YouTube) y los datos sobre transacciones.

❖ ***Reflexione sobre otros posibles campos para utilizar Big Data.***

3.2.4 Modelos de Negocio y las Fuentes de Datos Clave

El análisis de datos será particularmente transformacional en una serie actividades y sectores vitales para el funcionamiento de la economía actual, entre los que cabe destacar los siguientes: las finanzas, la logística, la fabricación, el desarrollo de nuevos productos, el comercio minorista, los medios de comunicación, los servicios online, el marketing, la salud, la energía, los servicios públicos, el transporte o el turismo.

La cadena de valor de los datos permite a nuevas empresas, ágiles y disruptivas, poder adoptar diferentes modelos de negocio, desde la recolección de datos y su transformación en información útil (análisis de datos), el uso de esta información para la optimización de procesos de negocio (*Business Intelligence*) y la provisión de servicios relacionados con todos los aspectos de consultoría de Big Data. Así mismo, están prosperando empresas de desarrollo de software en temas relacionados con los datos.

Analizando las diferentes fuentes de datos y las actividades que se pueden realizar sobre ellos, podemos identificar distintos modelos de negocio potenciales en torno a los datos como activo principal.

3.2.4.1 Actividades a realizar sobre los datos

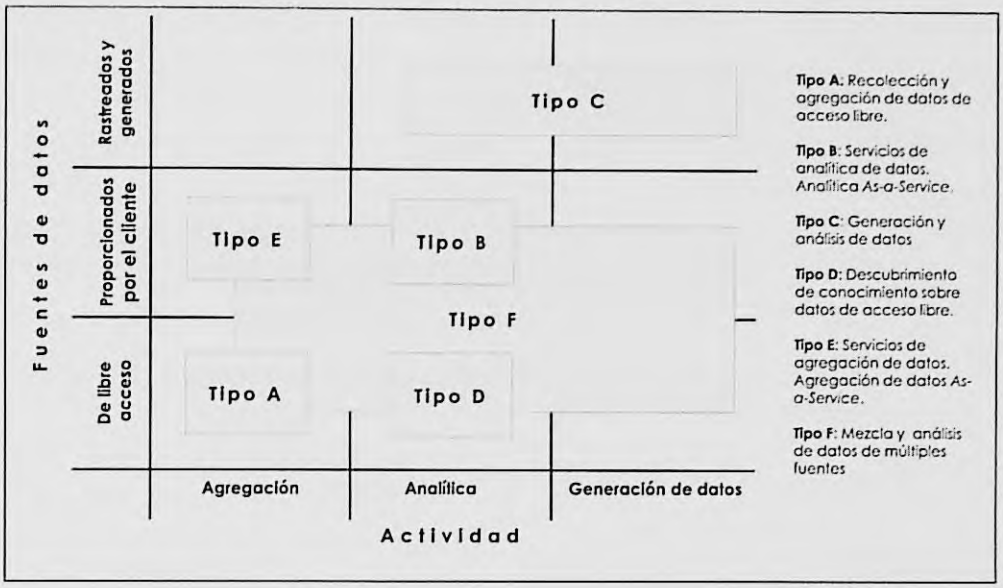
Por lo que respecta a las actividades a desarrollar con los datos, podemos identificar tres actividades clave:

- Agregación: Recolección de datos de diferentes fuentes y compilación de los mismos en una única fuente sin tratamiento adicional sobre los datos.
- Analítica de datos.
- Generación de datos. Creación de nuevos datos a partir de los obtenidos.

Teniendo en cuenta los tres tipos de datos (abiertos, del cliente y generados o rastreados en la web) y las tres actividades clave citadas, surgen seis tipos de modelos de negocio impulsados por los datos empleados por las *startups* que se sintetizan en la Figura 23.

En función de los datos utilizados, los seis tipos de modelos de modelos de negocio basados en los datos son los siguientes:

Figura 23 – Matriz de modelos de negocios impulsados por los datos



- Los modelos de negocio tipos «A» y «D» se basan en datos abiertos o datos de libre disposición.
- Los tipos «B» y «E» se basan en los datos proporcionados por los clientes y / o socios.
- El tipo «C» se basa en los datos rastreados en la web o generados.
- El modelo de negocio tipo «F» combina fuentes de datos proporcionadas por el cliente con otros de libre disposición.

En cuanto a las principales actividades realizadas por los diferentes tipos de empresas también se observan patrones distintos. Los grupos de empresas «A» y «E» se basan en la agregación de datos de diferentes fuentes. Los tipos de empresas «B», «C» y «D» sólo realizan análisis sin ocuparse de la agregación. El tipo de empresas «F» lleva a cabo tanto la agregación como el análisis.

En fechas recientes, han surgido dos iniciativas que se han centrado en analizar las oportunidades y los retos que el desarrollo de nuevos negocios impulsados por los datos; la liderada por el Foro Digital Europeo (European Digital Forum) y la llevada a cabo en el seno de la Universidad de Cambridge[73].

La diferencia principal entre ambos trabajos estriba en cómo conciben los datos obtenidos mediante el seguimiento de las personas: en el trabajo de la Universidad de Cambridge se habla de datos rastreados y generados, mientras que en el trabajo del Foro Digital

[73]Philipp Max Hartmann y col. «Big Data for Big Business? A Taxonomy of Data-Driven Business Models Used by Start-Up Firms». En: *SSRN* 2416869 (2014)

Europeo se habla de datos personales. Como se puede observar en la Figura 23, se ha seguido el criterio de la Universidad de Cambridge.

Acercándonos a la realidad

En el artículo «Siete modelos de negocio rentables para Big Data» de Enrique Checa y Ravi Chabaldas[74], publicado el 18 de agosto de 2014, se comentan diversos modelos de negocio basados en los datos. Al comentar la modalidad de *pay per use*, se señala que esta opción ofrece a los clientes una amplia selección de ofertas para que elijan solo la que realmente utilizan.

- ***Reflexione sobre las razones que explican que la modalidad pay per use es la más frecuentemente utilizada.***

3.3 El Dilema del Directivo: Intuición vs. Datos

Aunque la intuición y la experiencia siguen y seguirán desempeñando un papel importante en la toma de decisiones, desde hace unos años, hay dos factores que están contribuyendo a aconsejar que las decisiones se deberían fundamentar mucho más en los datos y menos en la intuición. Por un lado, los progresos alcanzados en el aprendizaje de máquina y, por otro, la creciente abundancia de datos de muy diversa naturaleza y el desarrollo de Big Data.

3.3.1 Machine Learning (Aprendizaje Máquina)

La ventaja comparativa de los seres humanos sobre el software se ha ido erosionando en los últimos años, pues las máquinas y sus algoritmos basados en el aprendizaje de máquina han avanzado en la capacidad de reconocimiento de patrones y en la interpretación y comunicación matizada de información compleja. Meramente como ejemplo, cabe señalar que el ganador de una prueba de reconocimiento de señales de tráfico borrosas con un 99,4 % de aciertos fue un algoritmo, mientras que el porcentaje de las personas que las identificaron correctamente fue un 98,5[75].

En cualquier caso, lo relevante es que en fechas recientes ha cobrado un renovado interés el debate sobre el papel que han de desarrollar las personas y sobre si las computadoras son complementarias o sustitutivas del trabajo humano y más concretamente en la toma de decisiones.

Este debate ha venido impulsado por el notable desarrollo experimentado en las últimas décadas por los algoritmos que permiten a las computadoras aprender. En efecto,

[74] Enrique Checas y Ravi Chabaldas. *Siete Modelos de Negocio Rentables para 'Big Data'*. Ed. por CincoDías. http://cincodias.com/cincodias/2014/08/14/tecnologia/1408040000_071970.html. [Online; publicado el 18 de agosto de 2014]

[75] Rik Kirkland. *Artificial Intelligence Meets the C-Suite*. Inf. téc. [Online; publicado en septiembre de 2014]. McKinsey & Company

el «aprendizaje de máquina» (*machine learning*) es una rama de la inteligencia artificial que desarrolla sistemas que pueden cambiar el comportamiento de éstas, de forma autónoma a partir de su experiencia. De su desarrollo se deriva, de forma simplificada, una visión un tanto optimista o utópica, por la que se tiende a creer que la innovación resolverá el grueso de nuestros problemas y que hay que dejar que la tecnología haga su trabajo.

No obstante, algunos señalan que no todo es positivo en este avance de las máquinas. Así, la denominada «paradoja de la segunda era de la máquina» destaca que, a pesar de que, por ejemplo en Estados Unidos se han alcanzado niveles muy altos de productividad y de creación de riqueza, el empleo no se ha mantenido y las diferencias de renta se han acentuado[76].

Junto con el aprendizaje de máquina, el otro factor que está incidiendo en el debate sobre el papel de la intuición y los datos en la toma de decisiones es la creciente abundancia de datos y la aparición de técnicas como Big Data, especialmente adecuadas para facilitar la toma de decisiones. Todo apunta a que está teniendo lugar un profundo replanteamiento del rol que debe jugar la intuición del directivo en la toma de decisiones.

Acercándonos a la realidad

En el artículo «Sistema del MIT permite a las máquinas procesar datos que ayudan en la toma de decisiones» publicado en Universia el 24 de febrero de 2015[77] se comenta el creciente poder de las máquinas para tomar decisiones.

- ***¿Reflexione sobre la importancia que pueden tener los avances en la capacidad de las máquinas para transmitir el conocimiento que adquieren?***

3.3.2 La Intuición como Forma de Tomar Decisiones

El tema de fondo es si la toma de decisiones se debe basar en la intuición de los directivos o en los datos. Cuando los datos son escasos, caros de obtener, o no están disponibles en formato digital, tiene sentido dejar que las decisiones se tomen en base a la intuición. Los directivos serán los que casi en exclusiva tomen las decisiones, y lo harán basándose en la experiencia que han acumulado y en patrones y relaciones observadas e interiorizadas. La intuición es el nombre dado a este estilo de toma de decisiones.

Para las decisiones de especial importancia, los gerentes responsables suelen estar en lo alto de la organización, o bien son «consultores» externos caros, traídos por su experiencia y resultados. Por ello, irónicamente algunos sostienen que muchas empresas suelen

[76] Erik Brynjolfsson y Andrew McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. Ed. por W. W. Norton & Company. 2014

[77] Universia. *Sistema del MIT Permite a las Máquinas Procesar Datos que Ayudan en la Toma de Decisiones*. http://noticias.universia.com.pa/actualidad/noticia/2015/02/24/1120416/sistema-mit-permite-maquinas-procesar-datos-ayudan-toma-decisiones.html. [Online; publicado el 24 de febrero de 2015]

tomar la mayoría de sus decisiones importantes basándose en la opinión del «hipopótamo» («HiPPO» – «highest-paid person's opinion») – la opinión de la persona mejor pagada[78].

En cualquier caso, todas las organizaciones deberían estar pensando en aprovecharse de las técnicas digitales y de los datos para apoyar su toma de decisiones. Si su empresa no lo hace, lo hará la competencia, y entonces quizás sea demasiado tarde para que intente ponerse al día. Sus competidores estarán siguiendo una senda de crecimiento exponencial, mientras que su empresa si no asume la realidad digital seguirá estando, en el mejor de los casos, en una senda lineal.

El difícil cambio a los datos es una necesidad

Puede afirmarse que el papel de un gerente de alto nivel en un mundo profundamente basado en datos va a cambiar. La pregunta clave que se debería formular sería: ¿dónde realmente aporto valor y de dónde debo apartarme para dejar paso a los datos? Este proceso va a significar un profundo replanteamiento del rol que debe jugar la intuición en la toma de decisiones del directivo.

En la actualidad hay muchos líderes de organizaciones que dicen: «Por supuesto que gestiono basándome en los datos. Recibo los datos y los utilizo como una entrada a mi proceso de decisión final.» Pero en muchas ocasiones los datos se utilizan para tratar de «justificar» opciones que ya habían sido elegidas, y se actúa así a pesar de que la evidencia nos dice que en general, esta forma de tomar decisiones conduce a un resultado peor que si el administrador se basara fundamentalmente en los datos. Este comportamiento se explica porque a los administradores experimentados y exitosos les resulta muy difícil abandonar la idea de que hay algo mágico o insuperable en su intuición personal.

Téngase en cuenta, que los altos ejecutivos, generalmente han llegado a donde están porque son muy buenos en lo que hacen. Y estos ejecutivos confían en su equipo, ya que también son profesionales con experiencias exitosas en la toma de decisiones. Por ello, la realidad es que la consolidación de Big Data coloca a los ejecutivos ante una dificultad real, ¿cómo dejar la intuición y empezar a guiarse por los datos cuando toda su cultura se ha basado en la experiencia y en la práctica? Todos los que son unos expertos en la administración desde un punto de vista práctico (*domain expert*), todos los que han gestionado con éxito una organización o han formado parte de un equipo de alta dirección, realmente creen en su capacidad y en su forma de tomar decisiones porque en buena medida eso es lo que explica su éxito. Pero en cierto sentido, puede que se estén viendo afectados por el conocido sesgo de supervivencia y el miedo al cambio.

Esto explica que la posibilidad de aceptar los datos como instrumento para tomar decisiones plantee reticencias. Aunque un cierto número de altos ejecutivos se sentirán

[78] Andrew McAfee. *The Path to Better Decisions, in One Cute Graph.* http://andrewmcafee.org/2013/05/mcafee-prediction-hippos-decisions-algorithms/. [Online; publicado en mayo de 2013]

favorables al cambio y estarán dispuestos a anular su propia intuición cuando los datos no están de acuerdo con ella, en el mundo de los negocios sigue habiendo directivos que confían demasiado en la experiencia y la intuición y no lo suficiente en los datos.

Por estas razones, un primer obstáculo a superar para dejar la intuición es aceptar una nueva cultura de toma de decisiones. En este proceso, si bien los desafíos técnicos que conlleva la utilización de grandes volúmenes de datos son muy reales, los obstáculos culturales y gerenciales son incluso mayores; empezando por aceptar el nuevo papel del equipo de alta dirección en el proceso de toma de decisiones. Por ello, uno de los aspectos más críticos de la implantación de un proyecto de Big Data es su impacto en cómo se toman las decisiones y quién va a hacer que se cumplan.

Acercándonos a la realidad

En el artículo «La intuición también cuenta» aparecido en el blog de Eduard Punset el 4 noviembre de 2012[79] se destaca la importancia de la intuición.

- ***Reflexione sobre la siguiente frase recogida en el artículo citado «Digamos de antemano que los partidarios de utilizar la razón en lugar de la intuición se equivocan tanto o más que los segundos y que la mente se mueve con menos dificultades en el último caso.»***

3.3.3 Proceso de Transición a la Toma de Decisiones Basada en Datos

Una explicación de la reticencia al cambio, es que se subestime la importancia de los datos. Un buena manera de evidenciar las ventajas que una gestión basada en datos puede conllevar, es reflexionando sobre la evolución experimentada por algunas empresas que han puesto los datos en el centro de su modelo de negocio como es el caso de Facebook, Netflix, Amazon o Google. Tomando como referencia estos ejemplos, habría que propiciar el comienzo de un proceso de transición hacia un mayor uso de los datos.

Los ejecutivos interesados en tomar decisiones en base a los datos deberían comenzar con dos técnicas muy simples:

- Primero, deberían adquirir el hábito de empezar su proceso de toma de decisiones preguntándose «¿de dónde provienen los datos?,», «¿qué dicen los datos?», «¿qué tipo de análisis se ha realizado?» y «¿confiamos en los resultados del análisis de los datos?».
- En segundo lugar, deberían estar dispuestos a dejarse corregir por los datos. Pocas cosas son más poderosas para cambiar una cultura de toma de decisiones que ver a un alto ejecutivo reconocer cuando los datos han refutado una corazonada.

[79] Eduard Punset. *La Intuición También Cuenta*. http://www.eduardpunset.es/19024/general/la-intuicion-tambien-cuenta. [Online; publicado el 4 de noviembre de 2012]

La gestión basada en los datos

La gestión basada en datos nace de una idea muy simple; supone buscar las mejores evidencias que se puedan encontrar, aceptarlas tal como son y actuar en consecuencia. En esencia, el modelo de gestión basada en hechos es un marco de actuación para trasformar los datos en bruto, en información y la información en conocimiento que facilite actuar. Es decir, la clave es transformar los datos en conocimiento y éste en la toma de decisiones y la realización de acciones por parte de la dirección.

El análisis de datos debe servir de ayuda a los directivos, que apenas tienen tiempo, a filtrar los datos relevantes de entre el cúmulo de información disponible, eliminando el problema de la sobrecarga de información y presentando de forma resumida e intuitiva los resultados del análisis.

La gestión basada en datos, es un elemento fundamental de la empresa inteligente. Gracias al Big Data, el presunto colapso por exceso de datos fruto de Internet y otras fuentes externas, se convierte en una oportunidad para las empresas que gestionan basándose en los datos. La gestión basada en los datos consiste en poner los datos en primer plano del proceso de toma de decisiones, y aplicar la técnicas de análisis apropiadas para que sean los datos los que, formulando las preguntas oportunas y utilizando las técnicas apropiadas, nos faciliten la toma de decisiones. Por estas razones puede afirmarse que la toma de decisiones basada en los hechos y en los datos es una decisión racional.

Destacar la importancia de la gestión basada en datos no quiere decir que la gestión basada en la experiencia práctica ya no tenga sentido. Lo que debe destacarse es que la gestión basada en datos es ahora, al menos, igualmente importante e irá a más. Los datos dicen lo que está pasando en realidad, mientras que la gestión basada en la experiencia siempre tiene un sesgo hacia el *statu quo*, lo que hace que sea muy difícil mantenerse al día en presencia de cambios disruptivos[80].

Gestionar en base a los datos y apoyándose en las nuevas herramientas requiere ser capaz de hacer las preguntas correctas. Además, esa habilidad va a ser crecientemente importante en el futuro. Será necesario contar no sólo con conocimientos técnicos, sino también tener un dominio de lo que los clientes están exigiendo, incluso si no saben muy bien lo que quieren. Esta combinación de habilidades técnicas y dominio del negocio serán la clave.

Para tomar decisiones en base a los datos es necesario que alguien conozca el problema que se va a abordar y que pueda identificar los conjuntos de datos que podrían ser útiles en su solución. Una vez llegados a este punto, lo mejor que se puede hacer es acotar el campo de acción de los expertos en el conocimiento práctico, que vienen con ideas

[80] Michael J. Mauboussin. *The True Measures of Success*. Ed. por Harvard Business Review. https://hbr.org/2012/10/the-true-measures-of-success. [Online; publicado en octubre de 2012]

preconcebidas sobre lo que son las correlaciones y relaciones interesantes, y contar con alguien que sea realmente bueno extrayendo la información que tienen los datos.

La industria petrolera y de gas, por ejemplo, dispone de fuentes de datos increíblemente ricas. Gran cantidad de sus brocas tienen sensores, y a medida que perforan, basándose en el tiempo requerido para que las ondas sonoras sean capturadas por una grabadora, se puede tener una idea de lo que hay debajo de la tierra. Esos datos son muy complejos y, hasta ahora, han sido interpretados en su mayoría de forma manual. El problema es que cuando una persona interpreta de forma manual lo que sale de un sensor en una broca o un movimiento sísmico, se le escapa una gran parte de la información, que un algoritmo de una máquina inteligente puede recoger. Por ello, hay que recurrir al aprendizaje de máquina para poder extraer toda la información que tienen los datos.

Acercándonos a la realidad

El artículo «La importancia de los datos en la toma de decisiones para mejorar las escuelas»[81], publicado en el diario El Mercurio en junio de 2012, se centra en la importancia de tomar decisiones basadas en los datos.

- ***En el artículo en cuestión se toma como referencia el sector de la educación, ¿Cree que en el sector educativo la toma de decisiones basada en los datos tiene una especial relevancia?***

Un algoritmo miembro de la comisión directiva

Un caso extremo de toma de decisiones racional sería aquel en el que, disponiendo de información para todo tipo de situaciones, dejásemos que un supercomputador fuera el que tomara las decisiones. En un artículo de 1967 de McKinsey Quarterly, «El gerente y el imbécil», Peter Drucker señaló que un ordenador no toma ninguna decisión; sólo ejecuta las órdenes[82]. Nos obliga a pensar, para establecer los criterios. Si más estúpida es la herramienta, más brillante tendrá que ser el que la utilice. Drucker concluía que la computadora era la herramienta más tonta que jamás había tenido.

¡Cómo han cambiado las cosas! Después de años de promesas, gracias a los avances de la inteligencia artificial, las computadoras están reemplazando a profesionales cualificados en campos como la arquitectura, la aviación, el derecho, la medicina o la geología del petróleo. Y están cambiando la naturaleza del trabajo en una amplia gama de otros empleos y profesiones, entre otras la del directivo.

---

[81] Lorna M. Earl. *La Importancia de los Datos en la Toma de Decisiones para Mejorar las Escuelas*. http://www.educarchile.cl/ech/pro/app/detalle?id=215969. [Online; publicado en julio de 2012]

[82] Martin Dewhurst y Paul Willmott. *Manager and Machine: The New Leadership Equation*. Inf. téc. [Online; publicado en septiembre de 2014]. McKinsey & Company

Este tipo de avances han hecho realidad el caso extremo que servía de título a esta sección. Una empresa de capital riesgo de Hong Kong, Deep Knowledge Ventures ha nombrado a un algoritmo de toma de decisiones como miembro su junta directiva[83].

Deep Knowledge Ventures toma decisiones de inversión sobre empresas de ciencias biológicas, a partir de estudios minuciosos de grandes cantidades de datos. Al igual que otros miembros de la junta, el algoritmo vota sobre si la empresa debe de invertir en una empresa en particular o no. El algoritmo toma sus decisiones analizando las previsiones financieras de las diversas empresas, las diversas pruebas clínicas, los derechos de propiedad intelectual y las rondas previas de financiación. El verdadero objetivo de Deep Knowledge Venture, al decidir incluir un ordenador en la comisión directiva fue llamar la atención sobre la posibilidad real de que un algoritmo inteligente pueda tomar decisiones de inversión de forma independiente. A su vez, la iniciativa llevada a cabo por Deep Knowledge Venture supone un apoyo explícito a la toma de decisiones basada en datos y a las técnicas de Big Data.

Los avances de las máquinas inteligentes nos sorprenderán, pero sólo transformarán la vida de los altos ejecutivos si los avances de gestión lo permiten. Todavía hay una gran cantidad de trabajo por hacer para poder sacarle el máximo provecho a la creciente cantidad de datos y adecuarlos a las máquinas inteligentes y a su creciente potencial de toma de decisiones. Además de eso, sería necesario que los líderes de alto nivel propicien la propia labor de las máquinas inteligentes, lo que en cierto sentido puede ir en contra de un siglo de desarrollo organizacional, en buena medida centrado en el enriquecimiento de la labor del gerente.

Si estas dos cosas suceden, y es probable que así ocurra, por la sencilla razón que las organizaciones de vanguardia que logren ventajas competitivas serán imitadas, el papel del líder de alto rango va a evolucionar. En cualquier caso, es posible que, precisamente en estas condiciones los ejecutivos de la era de las máquinas inteligentes serán capaces, gracias al toque humano, de marcar más claramente la diferencia.

Este toque humano quiere decir saber formular las preguntas oportunas, y tener el vigor para enfrentarse a circunstancias excepcionales y la capacidad para hacer cosas que las máquinas no pueden. Eso incluye tolerar la ambigüedad y saber centrarse en el lado menos técnico de la dirección, saber comprometer y motivar a la organización y fomentar su capacidad de auto-renovación.

¿Sustituirán los robots a los altos directivos?

Como se ha señalado en el punto anterior, el caso de Deep Knowledge Venture pone de manifiesto cómo bajo ciertas circunstancias un ordenador puede tomar decisiones de

[83] Rob Wile. *A Venture Capital Firm Just Named an Algorithm to its Board of Directors — Here's what it Actually Does*. Ed. por Business Insider. http://www.businessinsider.com/vital-named-to-board-2014-5. [Online; publicado el 13 de mayo de 2014]

inversión. El tema que ahora se analiza es de mayor alcance, pues lo que se plantea es si un robot puede sustituir a un directivo[82]. El sorprendente progreso de las posibilidades del software permite que un cerebro digital conduzca un coche por una calle sin chocar con nada ni herir a nadie. Muchas de las tareas que hasta hace poco parecían territorio exclusivo de los seres humanos están siendo realizadas por ordenadores.

En este contexto cabe preguntarse si los robots sustituirán a los altos directivos, o más ampliamente planteado ¿en qué medida las tareas y responsabilidades de éstos se verán afectadas por los ordenadores? En algunas actividades, sobre todo cuando se trata de encontrar respuestas a determinados problemas, el software supera incluso a los mejores gestores. Ante esta situación, un tema clave para los propios directivos es saber si lo adecuado es hacer valer su propia experiencia y capacidades o admitir la nueva situación y reservarse ciertas parcelas en el proceso de toma de decisiones[84].

En cualquier caso, la figura del directivo senior no está ni mucho menos obsoleta, pero deberá saber adaptarse a los nuevos roles que ha de desempeñar. A medida que las máquinas inteligentes progresan, los altos directivos están llamados a crear formas organizativas innovadoras necesarias para poder canalizar el talento humano que circula online por todo el mundo. Así, por ejemplo, todo el proceso *crowd* tiene un gran potencial que no es fácil capitalizar. Los ejecutivos tendrán que hacer hincapié en su capacidad creativa y de liderazgo y en su pensamiento estratégico.

Por otro lado, la labor del directivo con experiencia también será muy relevante cuando se trate de saber cuáles son los problemas a los que nos debemos enfrentar. Los expertos que tienen un profundo conocimiento de un área, son los que saben dónde están las mayores oportunidades y desafíos. Así, por ejemplo la empresa estadounidense PASSUR, que ha aplicado Big Data para gestionar el tráfico aéreo, ha contratado a un buen número de profesionales precisamente porque tienen un amplio conocimiento de las operaciones en los principales aeropuertos de Estados Unidos. Conforme el uso de los datos se va intensificando, el papel de los expertos con experiencia operativa está cambiado. Se les valora no por sus respuestas al estilo HiPPO, sino porque saben qué preguntas se deben hacer.

Los directivos, para seguir siendo competitivos en un entorno digital, deberán superar la dificultad de la mente humana para comprender el significado e implicaciones de las funciones y tendencias exponenciales. En el mundo digital todo sucede mucho más rápido de lo que imaginamos y el buen directivo debe evitar que los acontecimientos lo sorprendan y lo desborden. Y, en todo caso, deberá hacer valer sus fortalezas. Por ejemplo, hasta ahora no se ha visto que un algoritmo pueda negociar con eficacia; motivar y dirigir un equipo; averiguar lo que está pasando en una complicada situación social; o lo que motiva a la gente y cómo se consigue que los empleados se muevan en la dirección

[84] Erik Brynjolfsson y Andrew McAfee. *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Ed. por Digital Frontier Press. 2011

deseada. Estas son habilidades humanas y se van a seguir requiriendo. Pero si la gente que está actualmente administrando a las grandes empresas piensan que no hay nada de la revolución de la tecnología que les vaya a afectar, todo apunta a que se estarían comportando ingenuamente.

Acercándonos a la realidad

Miguel Artime, en Cuaderno de Ciencias publicó el 22 de septiembre de 2014 el artículo «El primer robot de Asimov toma decisiones éticas sorprendentes (y nefastas)»[85].

❖ ***¿En qué sentido las decisiones que toma un robot pueden ser sorprendentes?***

3.4 Nuevas Competencias, Nuevas Capacidades, Nuevos Roles: Chief Data Officer, Data Sciencist, Data Steward

3.4.1 Nuevas Competencias y Nuevas Capacidades

Durante los últimos 30 años, la mayoría de las empresas han añadido nuevas funciones a nivel de director en respuesta a los cambios en los entornos empresariales. Así, el director financiero (CFO – *Chief Financial Officer*), que no existía en la mayoría de las empresas en la década de 1980, se elevó de categoría para atender a la gestión financiera y a las relaciones con los inversores. Posteriormente, se creó el cargo de director de marketing (CMO – *Chief Marketing Officer*) y se convirtió en pieza clave para gestionar los nuevos canales mediáticos, todo lo relacionado con la marca y la relación con el cliente. Para dirigir todos los temas relacionados con la estrategia se creó la figura del CSO (*Chief Strategy Officer*), cuya misión es ayudar a las empresas a abordar los mercados globales cada vez más complejos y en rápida evolución, mediante el diseño de las estrategias más apropiadas.

En la actualidad, la importancia de los datos y su analítica están alterando el panorama de los negocios, y de nuevo las empresas necesitan incorporar en sus equipos de dirección algunos directivos especialmente responsables de la gestión de los datos. La identificación de oportunidades relacionadas con los datos para mejorar los ingresos, aumentar la productividad y, a veces, crear nuevos negocios, implica que las empresas se deban ocupar no sólo de reclutar nuevos talentos y llevar a cabo las inversiones en infraestructura de la información, sino también de procurar cambios importantes en las mentalidades y en la capacitación que permitan impulsar el análisis de datos, lo que suele requerir cambios en el equipo de dirección.

[85] Miguel Artime. *El primer robot de Asimov toma decisiones éticas sorprendentes (y nefastas).* Ed. por Cuaderno de Ciencias. https://es.noticias.yahoo.com/blogs/cuaderno-de-ciencias/el-primer-robot-de-asimov-toma-decisiones-eticas-sorprendentes-y-nefastas-195423828.html. [Online; publicado el 22 de septiembre de 2014]

En esta sección, siguiendo el artículo de Brown, Court y Willmott[86] se analizan las tareas más importantes que deben desempeñar los ejecutivos responsables de la gestión de datos.

3.4.2 Las Seis Tareas Principales del Analista de Datos

La elaboración y aplicación de Big Data y otras estrategias avanzadas de analítica es mucho más que servir datos a un tercero para que investigue las tendencias ocultas. Se trata más bien de lograr un cambio generalizado en la forma en que una empresa realiza su actividad del día a día en todo lo relacionado con los datos. La naturaleza de este cambio normalmente requiere de cambios en el equipo de dirección. Se necesitan profesionales con conocimientos institucionales, que sepan navegar entre las, a veces, turbulentas aguas de la organización, que sepan alcanzar equilibrios, con autoridad, para cuando aparezcan conflictos sobre las competencias a la hora de tomar decisiones, y con la capacidad de demostrar que la dirección está comprometida con una nueva cultura en la que la analítica desempeña un papel muy relevante. Según Brown, Court y Willmott[86], para lograr este cambio cultural y que el poder de los datos de asiente en las organizaciones es conveniente llevar a cabo una acción concertada en la que se tengan en cuenta los seis puntos que se detallan a continuación.

3.4.2.1 Establecer una nueva mentalidad

Los equipos sénior que se embarquen en este viaje necesitan adquirir un conocimiento suficiente del análisis de datos, de forma que puedan entender qué resultados se podrían obtener rápidamente y asumir la idea de que los datos deben ser una pieza central de su negocio. Sólo cuando esa visión de alto nivel está asentada se podrá irradiar una idea a toda la organización que se traduzca en cambios de comportamiento duraderos.

Como ejemplo de cómo proceder para alcanzar este deseado cambio de mentalidad, cabe señalar que los líderes de una gran empresa de transporte pidieron a su director estratégico (CSO) que se hiciera cargo del análisis de datos[86]. El CSO, para aumentar el conocimiento de los directivos de la compañía, organizó visitas a las grandes compañías de gestión de datos. Tras las visitas, pidió a cada unidad de negocio que elaborara una ficha en la que se reflejasen las prioridades en materia de datos de cada uno de ellos, para poder incluirlas en su plan estratégico para el próximo año. Este proceso contribuyó al establecimiento de metas empresariales reales y a la vez logró captar la atención de los ejecutivos de las unidades de negocio sobre cómo podrían utilizar el análisis de datos para mejorar sus resultados. En poco tiempo, estaban compartiendo abiertamente ideas y explorando nuevas oportunidades, lo que ayudó a dinamizar la organización.

---

[86] Brad Brown, David Court y Paul Willmott. *Mobilizing your C-Suite for Big-Data Analytics*. Inf. téc. [Online; publicado en noviembre de 2013]. McKinsey & Company

3.4.2.2 Definir una estrategia de analítica de datos

Como se ha señalado con anterioridad, el análisis de datos no alcanzará su pleno potencial sin una estrategia clara y sin articular los puntos de referencia clave para alcanzar el éxito. Muchas empresas fallan en este área, ya sea porque nadie a nivel del equipo de alta dirección se encarga, de forma explícita, de la redacción de un plan o porque no hay suficiente debate o tiempo dedicado a conseguir establecer un conjunto claro de prioridades aceptadas por los altos responsables.

En una empresa de telecomunicaciones, el CEO estaba dispuesto a seguir impulsando el análisis de datos, para mejorar conocimientos sobre la retención de clientes y la fijación de precios[86]. Aunque la compañía se movió con suficiente agilidad como para contratar a un líder de análisis de datos de alto nivel, el esfuerzo se estancó con rapidez. El equipo de análisis hizo correctamente su trabajo en cuanto al modelado y al análisis, sin embargo, los colegas de las unidades de negocios no le prestaron la atención suficiente a la formación de sus gerentes de nivel medio en el uso de los nuevos modelos de datos, quizás porque no vieron el potencial del proyecto y porque no pensaron que era parte de «sus» prioridades estratégicas. Es decir, el proyecto de Big Data careció de una estrategia de implantación y no se había integrado en la planificación estratégica de la compañía.

Un caso completamente distinto es el de una compañía de consumo norteamericana, en la que el CEO pidió al director de operaciones de servicios digitales y online, un ejecutivo con amplia experiencia en gestión de datos, que creara un plan para implantar Big Data, pero con la condición de que se hiciera en colaboración con el responsable de una unidad de negocio que no estuviese familiarizado con Big Data[86]. Con esta asociación, combinando un experto en datos y un gerente experimentado de primera línea, se aseguraron de que los objetivos establecidos en el plan, estuvieran directamente relacionados con temas de negocio de alto impacto. Por otra parte, dado que los responsables de Big Data compartían el progreso del trabajo con sus colegas de primera línea, su modelo de colaboración se convirtió en un referente para los esfuerzos de planificación de otras unidades de negocio.

Acercándonos a la realidad

En el artículo «5 consejos para introducir con éxito el Big Data en la estrategia empresarial»[87] publicado el 17 de mayo de 2013 en Marketing Directo, se dice que «El Big Data está cada vez más integrado en el ámbito empresarial, pero muchas empresas todavía no controlan a la perfección las fórmulas necesarias para introducir esta estrategia en su quehacer empresarial. A pesar de que un 65 % de los 'marketeros' señalan que los datos van a influir cada vez más en la comunicación de marca, muchos de ellos aún no saben cómo integrarlos en sus estrategias.»

[87] MarketingDirecto. *5 Consejos para Introducir con Éxito el Big Data en la Estrategia Empresarial.* http://www.marketingdirecto.com/actualidad/marketing/5-consejos-para-introducir-con-exito-el-big-data-en-la-estrategia-empresarial/. [Online; publicado el 17 de mayo de 2013]

- ***Analice los requisitos para implantar un proyecto de Big Data que se señalan en el citado artículo y reflexiona sobre su oportunidad.***

3.4.2.3 Determinar qué se debe crear internamente, comprar o alquilar

Otro conjunto de decisiones que requieren de la autoridad y la experiencia de un líder de alto rango está relacionado con la recopilación de datos y la construcción de modelos avanzados de análisis y herramientas diseñadas para mejorar el rendimiento. Las demandas de recursos son a menudo considerables y dado que en el mercado hay un buen número de proveedores externos capaces de proporcionar datos básicos, modelos y herramientas, cuando se trata de poner en marcha un proyecto de Big Data, se necesita la experiencia de algún alto directivo para decidir qué cosas se compran, cuáles se alquilan o subcontratan y cuales se desarrollan internamente.

3.4.2.4 Contar con expertos en análisis de datos

Bajo casi cualquier escenario estratégico, las organizaciones necesitarán más expertos en análisis de datos que puedan lograr que las empresas prosperen en un entorno de cambio rápido. El nuevo entorno también requiere habilidades de gestión que conllevan un número creciente de científicos de datos que generen los modelos de predicción o de optimización que soportarán el crecimiento. Un líder con experiencia en gestión de datos deberá coordinar la contratación de estos expertos y procurar que se coordinen de la mejor manera posible con el resto de los miembros de la organización.

Acercándonos a la realidad

Analice la información contenida en el artículo «¿Qué características debe reunir un profesional del Big Data?» publicado por Canal Basics el 20 de septiembre de 2013[88].

- ***Reflexione sobre lo señalado en el artículo citado sobre la importancia de mantener un pensamiento crítico y una naturaleza inquisitiva cuando se aborda un proyecto de Big Data.***

3.4.2.5 Movilizar los recursos

Las empresas a menudo se ven sorprendidas por el importante esfuerzo de gestión que conlleva la movilización de recursos humanos y de capital necesarios para crear nuevas herramientas de soporte para la toma de decisiones y ayudar a los gerentes de primera línea a poner en marcha modelos de analítica avanzada. Contar con un directivo de alto nivel responsable de la gestión de los datos es vital para romper las barreras institucio-

[88] Canal Basics. *¿Qué Características Debe Reunir un Profesional del Big Data?* http://www.confirmasistemas.es/es/contenidos/canal-basics/que-caracteristicas-debe-reunir-un-profesional-del-big-data. [Online; publicado el 20 de septiembre de 2013]

nales que con frecuencia obstaculizan la puesta en práctica de proyectos de análisis de datos.

3.4.2.6 Crear capacidades en la primera línea de gestión

Las soluciones analíticas sofisticadas que los estadísticos y los científicos inventan, deben integrarse en las herramientas de primera línea de forma simple y atractiva, para que los gerentes y empleados de primera línea se sientan cómodos usándolas todos los días. El esfuerzo para adoptar un programa de datos debe comprender también la formación y capacitación formal y en el puesto de trabajo. Conforme las empresas van avanzando en la puesta en práctica de una estrategia de datos, muchas optan por considerar que necesitan aumentar la capacidad ejecutiva, siendo frecuente crear una unidad centralizada dedicada al análisis de datos.

3.4.2.7 Caso del algoritmo de Page-Brin: Una capacidad adicional del analista de datos, el ingenio

Además de todos los condicionantes citados, el analista de datos tiene que tener una elevada dosis de ingenio[89]. El ingenio trabaja planteando, de forma adecuada, los problemas y es fundamental en la búsqueda de las mejores formas de resolverlos. Tomemos el caso de la patente estadounidense número 6.285.999, cedida a la Universidad de Stanford y presentada por los cofundadores de Google, Larry Page y Sergey Brin, junto con otros colaboradores. Se trata de una solución verdaderamente ingeniosa a un problema de Big Data. El problema consistía en ordenar unas mil millones de páginas web activas, con el fin de permitir que un usuario pudiera buscar de forma efectiva entre esas páginas mediante consultas sencillas. Un algoritmo de búsqueda eficiente sería aquel que presentase al usuario las páginas ordenadas según una palabra o frase clave, y según su relevancia en relación con dicha palabra o frase.

Para resolver este problema se puede pensar en la World Wide Web como una red descomunal de páginas web, los patrones de referencia entre páginas, actuando como los enlaces y las páginas como los nodos. Tomando esto como una entrada, hay que calcular algún tipo de rango, basado en la prominencia o *status* de cada una de las páginas en 24 horas o menos, todo ello con el hardware existente.

El procedimiento de resolución consiste en calcular un «peso» aproximado para cada página, basado en el número relativo de enlaces de entrada (veces que es citado) y salida (citas realizadas). De esta forma lo que se pretende es simular el comportamiento de un usuario que está navegando por la web. También incluye un factor que mide la probabilidad de que el usuario se aburra y abandone. Con esa información se construye una matriz enorme pero con muchos ceros, lo que permite que pueda fácilmente invertirse y que diariamente muestre la importancia de cada página en la web.

[89] Mihnea C. Moldoveanu. «Algorithmic Foundations for Business Strategy». En: *SSRN* 2210077 (2013)

3.4.3 Nuevos Roles: Chief Data Officers, Data Sciencists y Data Stewards

Las limitaciones para ejercer la capacidad de liderazgo en materia de datos, puede que socaven los esfuerzos de muchas empresas. Para superar este problema hay que pensar en nuevas estructuras de gestión, en nuevas funciones y en una nueva división del trabajo.

Por las razones apuntadas, las empresas y profesionales deberían ponerse como objetivo adquirir competencias para Big Data, esto es, para el análisis de flujos de datos en tiempo real, mediante fuentes multiestructuradas y con herramientas capaces de analizar grandes volúmenes de datos. Los trabajadores de TI están adquiriendo nuevos roles, haciendo de puente entre la tecnología y el negocio. Se necesitan capacidades avanzadas de gestión de información/análisis y experiencia en negocios. Los trabajos de Big Data requieren analistas de datos y diversos profesionales, entre otros con los perfiles descritos a continuación.

3.4.3.1 Chief Data Officer

El Chief Data Officer (CDO), que podríamos traducir como Director de Datos, de una organización, es el técnico responsable de la gobernabilidad de los datos en toda la empresa y de la utilización de la información como un activo[90]. Para cumplir con sus obligaciones coordina el procesamiento, el análisis y la minería de datos así como la gestión de la información[91]. El CDO obtiene información del Director de Tecnología (CTO), del director de márketing (CMO), del responsable de estrategia (CSO), del Consejero Delegado (CEO), entre otros.

El CDO es el responsable de determinar qué tipo de información va a capturar, retener y explotar la empresa y con qué fines (Estela Gómez Rielo, 2014). El CDO no tiene necesariamente por qué coincidir con el director del sistemas de información cuya responsabilidad, entre otras, es la de almacenar los datos y facilitar su posterior tratamiento.

El responsable de la gestión global de los datos no se empezó a reconocer como un miembro de la alta dirección hasta finales de la década de los 80 del siglo pasado. En aquellas fechas fue cuando algunas organizaciones empezaron a reconocer la importancia de la tecnología de la información, así como las implicaciones del *Business Intelligence*, la integración de datos, la gestión de datos maestros y el procesamiento de datos como activo para el funcionamiento diario de los negocios. La función del responsable de los datos incluye la definición de las prioridades estratégicas para la compañía en el área de sistemas de información, la identificación de nuevas oportunidades de negocio en el área de

[90]Wes Hunt. *Why Data Needs a Leader.* http://asmarterplanet.com/blog/2014/08/data-needs-leader.html. [Online; publicado el 7 de agosto de 2014]

[91]Estela Gómez Rielo. *El Laberinto de la Información. ¿Dónde Está la Salida?* http://mundocdo.blogspot.com.es/2013/12/gobernanza-de-la-informacion.html. [Online; publicado el 10 de octubre de 2015]

los datos, la optimización de la generación de ingresos a través de los datos, y en general la consideración de los datos como un activo estratégico para la compañía.

La función de la gestión de datos se ha visto impulsada en los últimos años por un conjunto de factores entre los que cabe destacar:

- El aumento de las arquitecturas orientadas a servicios; (SOA – *Service Oriented Architecture*). Una arquitectura orientada a servicios es un paradigma de arquitectura concebido para diseñar y desarrollar sistemas distribuidos (un sistema distribuido es una colección de computadoras separadas físicamente y conectadas entre sí por una red de comunicaciones; cada máquina posee sus componentes de hardware y software que el programador percibe como un solo sistema) y su misión es satisfacer los objetivos de los negocio. Entre las características de las arquitecturas SOA cabe destacar la facilidad y flexibilidad de integración con sistemas que se han venido utilizando en el pasado,la alineación directa con los procesos de negocio, reduciendo costes de implementación, la capacidad para propiciar la innovación de servicios a los clientes y la capacidad de adaptarse rápidamente ante cambios en el entorno.

  Las arquitecturas orientadas a servicios facilitan la creación de sistemas de información altamente escalables y ajustados al negocio de la organización, y su estructuración facilita la prestación de servicios, comúnmente servicios web.

- La integración de sistemas a gran escala, y los mecanismos de almacenamiento de datos heterogéneos.

Para llevar a cabo las tareas mencionadas resulta necesario contar con una persona de alto nivel, que tenga conocimiento del negocio, que domine las técnicas, que sepa liderar equipos y tenga capacidad para gestionar e implementar todo lo relacionado con los datos. Esta persona, además deberá saber aprovechar las oportunidades de generar ingresos. El CDO será el responsable de explicar al equipo directivo el valor estratégico de los datos y su importante papel como un activo empresarial y fuente de ingresos.

En fechas más recientes, con la aceptación del concepto de «*Data Science*», el responsable de los datos se ha empezado a considerar como una persona clave para el diseño de la estrategia de la empresa. Debe tenerse en cuenta que esta persona tiene un papel muy relevante en la cuantificación de las diferentes líneas de negocio y en consecuencia, en la definición de la estrategia de crecimiento de la compañía.

Como síntesis de lo señalado cabe afirma que los datos necesitan un líder, y éste es el CDO. El CDO debe tener como misión formular e implementar una estrategia que ponga los datos en el centro del negocio. Para ello, el CDO debe prestarle una especial atención

a lograr alinear plenamente sus iniciativas con el negocio, debe diseñar una detallada estrategia de datos y debe saber priorizar los proyectos[92].

Ejemplos de la generalización de la figura del CDO

A título personal uno de los CDO más conocidos es el estadounidense John Bottega. Este profesional de los datos , es un ejecutivo de estrategia y gestión de datos con más de 30 años de experiencia en la industria. A lo largo de su carrera, John Bottega ha ocupado diversos cargos en funciones de gestión de datos de una empresa. Fue el CDO en Citi y en Bank of America y en la actualidad es el CDO de la Reserva Federal de Nueva York y presidente del Consejo de EDM (Enterprise Data Management).

Por otro lado, cabe señalar que en numerosas ciudades estadounidenses, como por ejemplo, San Francisco, Chicago, Filadelfia y Baltimore tienen Directores de Datos. El ejército de Estados Unidos, también con un Oficial Jefe de Datos y la Comisión Federal de Comunicaciones de Estados Unidos tiene varios Consejeros de datos. Asimismo, varios departamentos del gobierno de Estados Unidos tienen CDO. En el campo del sector público europeo, cabe indicar que en Francia, en septiembre de 2014, Henri Verdier, fue nombrado *Administrateur Général des Données* (CDO) de la Administración francesa, siendo la primera vez que un país nombró un CDO.

Con carácter general, debe señalarse que como consecuencia de la crisis financiera internacional de 2008, numerosos grandes bancos y compañías de seguros crearon el cargo de CDO. El objetivo de estos nombramientos fue tratar de asegurar la calidad de los datos y la transparencia de todos los asuntos relacionados con la regulación y gestión de riesgos, así como la presentación de informes analíticos.

Acercándonos a la realidad

En el artículo «The Role of the Chief Data Officer in Financial Services» publicado por la consultora Capgemini, se analizan las principales tareas que debe desarrollar un CDO en el caso de los servicios financieros.

- ***Reflexione sobre las conclusiones del trabajo (página 25 del artículo referenciado) y evalúe si las entidades financieras que conoce llevan a cabo este tipo de tareas.***

3.4.3.2 Data Scientist

La figura del científico de datos (*data scientist*) como puesto de trabajo ha ido creciendo de forma paralela al crecimiento del análisis de datos. El científico de datos no está ligado exclusivamente a los proyectos de Big Data, pero su papel encaja con algunas de las tareas relacionadas con la complejidad de los datos y los análisis propios del Big Data,

[92]Marc Teerlink y col. *The New Hero of Big Data and Analytics: The Chief Data Officer*. Inf. téc. [Online; publicado en 2014]. IBM Global Business Services

especialmente si se compara con los roles tradicionales del analista de datos[93]. En cierto modo el científico de datos representa una evolución del rol empresarial de analista de datos (el analista de datos es el responsable de inspeccionar, limpiar y transformar datos con el objetivo de resaltar información útil) y de hecho, la formación profesional de ambos tiene muchos puntos en común. El científico de datos requiere una buena base en el área de computación y son muy convenientes conocimientos de análisis de datos, estadística y matemáticas, pero también se requiere un buen conocimiento del negocio.

Si se pretende establecer una diferenciación entre el científico de datos y el analista de datos, cabría señalar que el científico de datos, además de los conocimientos técnicos propios del analista de datos debe tener un profundo conocimiento del negocio, así como la capacidad de comunicar los resultados a los líderes empresariales y a los responsables de TI, de forma que la organización aprecie la contribución que el análisis de datos puede hacer a la hora de superar los retos empresariales. El científico de datos debe procurar que quede claro para todos, que el análisis de datos desempeña un papel fundamental a la hora de afrontar los problemas del negocio y que se están abordando correctamente los problemas que tienen un mayor impacto y valor para la organización.

Por otro lado, si bien un analista de datos tradicional que está colaborando en cómo mejorar la atención al cliente puede centrarse sólo en los datos de una sola fuente - el sistema de CRM de la compañía, por ejemplo - un científico de datos generalmente explorará y examinará los datos de múltiples fuentes dispares. El científico de datos analizará todos los datos entrantes con el objetivo de descubrir algunos temas que antes estaban ocultos, y comprobará si pueden contribuir a proporcionarle algún tipo de ventaja competitiva a la empresa. Por lo tanto, un científico de datos no se limita a recopilar y a presentar datos, sino que los analiza en profundidad y recomienda la mejor manera de utilizarlos.

El científico de datos toma la información, esto es los datos, como la base de su investigación. Empieza por preguntarse por el verdadero significado de los datos y seguidamente comienza a hacerse preguntas del tipo ¿cómo evolucionarán eso datos en el futuro? ¿qué hacer si...? Esto es, el científico de datos, a partir de la información disponible se cuestiona los supuestos hasta ahora aceptados y los procesos existentes. Una vez depurados los resultados de su análisis, el científico de datos deberá tener la capacidad de comunicar las conclusiones de su trabajo y las recomendaciones que se deriven del mismo, a los directivos de las áreas de negocio directamente afectadas y a los máximos responsables de la compañía, siempre respetando la estructura de liderazgo de la organización. Así, pues, el científico de datos debe aunar ciertas características del investigador científico, con otras de un gerente que domina el arte de la comunicación y las relaciones entre los miembros de la organización[94].

---

[93] IBM. *What is a Data Scientist?* http://www-01.ibm.com/software/data/infosphere/data-scientist/. [Online; consultado el 10 de diciembre de 2015]

[94] Aroa Pérez. *'Data Scientist', el Trabajo más Sexy.* http://www.misapisportuscookies.com/2012/11/data-scientist-trabajo-sexy/. [Online; consultado el 19 de noviembre de 2012]

Por estas razones algunos autores han elegido el de «*data scientist*» como el trabajo más sexy de la década[95]. Es un trabajo que requiere las características de un profesional polivalente, que combina diferentes capacidades. Debe tener capacidad analítica y formación en estadística, pero también tener curiosidad, porque el científico de datos debe saber crear significado a partir de unos datos de los que, en un principio, no resulta obvio inferir consecuencias útiles para la organización y obtener valor a partir de los datos. Un científico de datos quiere construir cosas, no sólo dar consejos.

Acercándonos a la realidad

En el artículo «Data scientist, el trabajador tech que todas las empresas buscan»[96] publicado el 16 de noviembre de 2012, se comentan algunas de las características de este perfil profesional.

- ***¿En qué sentido se suele hablar de los data scientists como «mitad artista y mitad analista»?***

3.4.3.3 Data Steward

Un administrador o auxiliar de datos (que sería la traducción más próxima al termino anglosajón) es la persona responsable de la gestión de todos aquellos elementos relacionados con los datos[97]; también conocidos como elementos críticos de los datos, hacen referencia no solamente de los datos en sí, sino también de sus metadatos asociados (los metadatos son datos que describen otros datos, pues actúan cómo índices para ubicar otros datos o para describirlos). Los administradores de datos llevan a cabo una serie de tareas especializadas que requieren participar activamente en la definición de procesos, políticas, directrices y responsabilidades de la administración de todos los datos de las organizaciones.

La primera tarea que debe realizar un administrador de datos consiste en identificar los elementos que va a gestionar o administrar. El trabajo realizado por el administrador de datos finalmente se plasma en una serie de normas, estándares y controles y en la entrada efectiva de datos en el sistema. Para confeccionar el conjunto de normas, estándares, y controles, el administrador de datos colabora estrechamente con los siguientes grupos de profesionales:

---

[95] Thomas H. Davenport y D.J. Patil. *Data Scientist: The Sexiest Job of the 21st Century*. Ed. por Harvard Business Review. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/. [Online; publicado en octubre de 2012]

[96] TICbeat. *Data Scientist, el Trabajador Tech que Todas las Empresas Buscan*. http://www.ticbeat.com/innovacion/data-scientist-trabajador-tech-todas-empresas-buscan/. [Online; publicado el 16 de noviembre de 2012]

[97] Office. *Understanding the Role of Data Stewards in Data Management*. https://support.office.com/en-au/article/Understanding-the-Role-of-Data-Stewards-in-Data-Management-ae3352f3-4389-45e8-a682-7fd6edb92524. [Online; consultado el 11 de diciembre de 2015]

- Analistas de normalización, para, de forma consensuada, elaborar las normas que se integran en el glosario empresarial.
- Arquitecto de datos y/o con los creadores de modelos, para definir los estándares a los que se deben ajustar los datos.
- Analistas de calidad, para concretar los controles que garanticen la calidad de los datos.
- Miembros del equipo de operaciones, para que los datos estén alineados con los intereses y reglas de negocio.

Aunque las tareas comentadas son las más relevantes que debe desempeñar un administrador de datos, también hay que mencionar las funciones de custodia de datos. Los términos en los que se define la custodia de los datos se deben establecer de forma concreta cuando las organizaciones van a intercambiar datos. Hay que especificar los términos del intercambio para que este se pueda realizar de manera precisa y consistente entre los sistemas informáticos.

4 Introducción al Trabajo con Datos

Como indicamos en la sección 2.3.2, donde describimos los componentes de la capa de gestión de los datos en un sistema de Big Data, cuando vamos a realizar un trabajo con datos, debemos comenzar por identificar las fuentes de las que obtendremos los datos en bruto.

Tras la identificación de las fuentes de datos que son relevantes para nuestro problema, comenzaremos a definir e implementar los procesos que nos permitan extraer los datos de las fuentes originales, para posteriormente proceder a su almacenamiento o carga en nuestros sistemas. Una vez allí, podemos comenzar a trabajar con estos datos.

Estos procesos de extracción y carga de los datos no son triviales. Esto se debe sobre todo a que los datos que proceden de las distintas fuentes probablemente no tendrán la misma estructura que el sistema en el que se desean cargar. Esto implica el desarrollo de procesos intermedios que los transformen, preparándolos para la carga. En conjunto, estos procesos se conocen como procesos ETL –*«Extract, Transform and Load»*– o procesos de Extracción, Transformación y Carga.

Como en cualquier producto o servicio, la calidad final de éste es directamente proporcional a la calidad de las materias primas utilizadas para su elaboración. Del mismo modo, la calidad del conocimiento obtenido a partir de los datos será directamente proporcional a la calidad de los datos utilizados para el análisis. De manera que antes de cargar los datos en nuestro sistema, debemos garantizar la calidad de los datos usados en este proceso.

A lo largo de este capítulo realizaremos una revisión de los procesos que hay que seguir antes de cargar los datos en nuestro sistema. Primero haremos una revisión de las fuentes de datos, para posteriormente describir los procesos ETL y de control de la calidad.

4.1 Fuentes de Datos

Independientemente de cual sea el alcance de nuestro proyecto, cuando trabajamos con datos nos enfrentamos a la tarea de decidir con qué datos queremos o necesitamos trabajar para obtener un valor añadido. Debido a la explosión en la variedad de fuentes de información disponibles en la actualidad, cada vez es más frecuente hacer uso de datos que no han sido generados internamente por nuestros sistemas de información. Por ello, las fuentes de datos pueden clasificarse, atendiendo al origen de los datos con los que trabajaremos, en internas o externas.

Las **fuentes de datos internas** pueden dividirse a su vez en dos tipos:

- **Bases de datos transaccionales (sistemas** OLTP**)**: aquellas que han sido generadas por nuestros sistemas OLTP (definidos en el capítulo 1). Es decir, todos aquellos sistemas que procesan las transacciones que se producen en la operativa de nuestro negocio o que utilizamos para facilitar la ejecución del mismo. También se conocen como los sistemas operacionales o transaccionales. Entre estos sistemas podemos contar con paquetes software más genéricos (como los ERP, CRM, SCM, etc.) o aplicaciones desarrolladas a medida (bien por departamentos internos o proveedores externos). Normalmente estos datos se encuentran almacenados en sistemas gestores de bases de datos relacionales estándar.
- **Sistemas de información analíticos (sistemas departamentales)**: previsiones, presupuestos, planes de proyecto, etc. En general se trata de información muy particular de cada departamento o línea de negocio que se utiliza en la actividad diaria de los equipos de trabajo de las empresas. Estos datos se encuentran en formatos menos estándar, como son hojas de cálculo, ficheros de texto o pequeñas aplicaciones de software personalizadas (por ejemplo, un desarrollo a medida que captura información de las diferentes máquinas que forman una cadena de producción). En este último caso la información suele estar más estructurada, aunque la base de datos en la que residen los datos puede no ser tan estándar como en los paquetes comerciales anteriormente citados.

Cualquier compañía necesita contextualizar la información que generan internamente sus sistemas con la realidad del mundo exterior mediante la incorporación de **fuentes de datos externas**. Por ejemplo:

- **Información de profesionales expertos** que pueden proporcionar análisis sectoriales o tendencias de mercado. Estas fuentes de información pueden, en

muchos casos, ser compradas a terceros (por ejemplo, Nielsen vende estudios de mercado en el sector de la distribución de gran consumo e IMS de la industria farmacéutica).

- **Datos que permitan** enriquecer la información que tenemos de nuestros clientes. Tradicionalmente se ha recurrido a la información socio-demográfica proporcionada por organismos públicos o privados, aunque desde hace unos años los organismos públicos proporcionan en mayor o menor medida información de muy diversa índole a los ciudadanos a través de sus páginas web (parte de lo que se conoce como *Open Data*). Recientemente se utiliza, cada vez con mayor intensidad, los datos procedentes de las redes sociales (*Social Media*) que nos proporcionan información muy valiosa acerca de los gustos y comportamientos de nuestros clientes o de nuestro grupo potencial de clientes objetivo.

- **Información procedente de máquinas y sensores** que, cada vez más, se encuentran conectadas a la red. Se trata de lo que denomina la «internet de las cosas» o M2M (acrónimo de *«Machine to Machine»*), es decir los millones de sensores, actuadores o cualquier tipo de máquina de diferente propósito, que capturan información que posteriormente procesan y almacenan. La variedad de estos objetos es muy grande, desde sensores medioambientales que capturan información meteorológica, concentraciones de gases o de partículas como el polen; a sensores de presencia que permiten contabilizar el número de personas o vehículos que se encuentran en un lugar o que pasan por un determinado lugar.

- **Dispositivos móviles** que son algo que siempre llevamos encima y aunque hay muchos aspectos a considerar –legales, regulatorios, etc.– pueden convertirse en una fuente de información muy valiosa para las empresas. Sólo por citar un ejemplo, los teléfonos inteligentes (*«smartphone»*) incorporan muchos sensores en su interior que están constantemente capturando información. Por ejemplo, cuando utilizamos una aplicación de mapas para guiarnos a una dirección, necesitamos activar el GPS del teléfono, lo que estará constantemente generando información acerca de nuestra posición. Esta información, previamente anonimizada, puede ser de gran utilidad para la empresa que proporciona la aplicación de mapas.

Habitualmente la información proporcionada por las fuentes de datos externas se suele obtener a través de APIs que permiten el consumo automático de las mismas desde aplicaciones software. En el caso de la información sectorial proporcionada por especialistas sectoriales, lo normal es que la ésta se proporcione en forma de ficheros de texto, con cierta estructura, de manera que su integración en nuestros sistemas sea simple.

Existen muchos factores que contribuyen a la complejidad de la integración de la información, entre los que podemos citar:

- **El número de fuentes de datos distintas** de las que cargamos la información. Independientemente de que sean internas o externas, acceder a distintas fuentes de datos requiere ciertas habilidades y el conocimiento de varias tecnologías. Además, cuanto mayor es el número de fuentes de datos que necesitemos integrar, más probable será que los datos no estén normalizados, es decir, encontraremos codificaciones distintas. A mayor número de fuentes de datos, mayor número de modelos de información transaccional que necesitamos conocer y documentar, para entender el significado de cada uno de sus elementos. Además, la definición de los distintos componentes de nuestro sistema de información no es siempre consistente en distintas aplicaciones de nuestra compañía.

- **La variedad de la información** con la que debemos trabajar. La información procedente de nuestros sistemas internos es, normalmente, estructurada. Sin embargo, es cada vez más frecuente que nos encontremos con información no estructurada como correos electrónicos, informes, vídeos, comentarios en redes sociales, etc.

- **El volumen de información** con el que debemos trabajar. Cuando trabajamos con fuentes de datos internas, salvo casos muy específicos, el volumen de información que tenemos que integrar es manejable. Sin embargo, cuando comenzamos a integrar fuentes de datos externas el volumen de datos comienza a ser un problema de consideración, especialmente para las tecnologías tradicionales. Retomando el ejemplo de los mapas, la aplicación de mapas estará generando información como la latitud y la longitud cada cierto tiempo (20 o 30 segundos por ejemplo). Esta información acumulada para los recorridos, aunque sean de tan sólo 15 minutos, de un número considerable de usuarios (que utilicen la aplicación a diario en el mundo) puede llegar a suponer un volumen considerable de datos que además se generan de forma constante.

En ocasiones tendremos que analizar si la información de la que disponemos es la que necesitamos para alimentar los modelos de negocio que hemos definido anteriormente. En este punto, muchas veces descubrimos que no disponemos de la información necesaria para completar el modelo de negocio que habíamos diseñado. Esta circunstancia nos puede llevar a modificar nuestras aplicaciones transaccionales para conseguirla.

4.1.1 Open Data

Una de las posibles fuentes de datos externas provienen del *Open Data*. El Open Data surge de la idea de permitir que los datos puedan ser accedidos y distribuidos de forma libre y sin restricciones, en especial los datos públicos que se han generado con los impuestos de todos los contribuyentes. Esto puede generar riqueza a una escala inesperada. Además, esta liberación de datos públicos puede animar a empresas y organismos privados a abrir (parte de) sus datos produciéndose así un efecto bola de nieve. El objetivo por tanto, es similar al de otros movimientos «*Open*», si bien el caso de *Open Data*

es más reciente y está ganando popularidad en el contexto de la Web. En este sentido, cabe destacar iniciativas de datos abiertos como las del Gobierno de Estados Unidos[98] y del Gobierno del Reino Unido[99], que pretenden poner a disposición de los ciudadanos los datos que producen sus gobiernos.

En esta misma línea, la Unión Europea ha apostado por los *Open Data* en un portal Web[100] en el que se publican los datos producidos por gobiernos de la UE, proyectos realizados con financiación pública, agencias públicas, etc. Sin embargo, existen otros repositorios de datos públicos abiertos que no están directamente relacionados con datos gubernamentales, como los que podemos encontrar en DataHub[101].

En todos estos repositorios podemos encontrar datos en diferentes formatos (incluyendo datos estructurados, semi-estructurados y sin estructura), ya que estos repositorios no establecen normalmente limitaciones en este sentido. Por tanto, estas fuentes de datos podrán servir para alimentar los procesos de análisis del Big Data con datos de libre acceso que nos pueden servir para mejorar los almacenes de datos que dan soporte a la toma de decisiones empresarial.

El acceso a estos repositorios de datos abiertos puede variar en sus mecanismos dependiendo del repositorio e incluso de la propia fuente de datos. De forma que nos podemos encontrar con:

- Sitios FTP para la descarga de ficheros.
- Servicios SOAP o REST para acceder a los datos con la posibilidad de realizar filtrado sobre los mismos usando parámetros en las llamadas al servicio.
- Servicios SPARQL para consultar Datos Vinculados.
- APIs ODBC o JDBC para el acceso a bases de datos mediante consultas SQL.

Aunque estos datos son de libre acceso, si se pretende hacer un uso empresarial de los mismos es necesario revisar las licencias para que el uso que hagamos de estos datos no incumpla ninguna normativa o restricción legal. El uso no comercial de estos datos suele estar libre de restricciones, y la única limitación es la obligación de citar la fuente de los datos en caso de redistribuirlos o publicarlos de alguna forma. Por ejemplo, para el caso del repositorio de datos de la Unión Europea nos encontramos el siguiente aviso:

[98] *The Home of the U.S. Government's Open Data.* https://www.data.gov. [Online; consultado el 11 de diciembre de 2015]

[99] *Opening Up Government.* https://www.data.gov.uk. [Online; consultado el 11 de diciembre de 2015]

[100] *European Union Open Data Portal.* https://open-data.europa.eu/en/data/. [Online; consultado el 11 de diciembre de 2015]

[101] *The Datahub.* https://datahub.io/es/. [Online; consultado el 11 de diciembre de 2015]

Acercándonos a la realidad

Vamos a considerar el caso de una empresa de venta online de material de ferretería. En nuestra tienda virtual se detecta que una parte importante de las ventas se realizan a empresas del sector turístico. Por tanto, interesa conocer los ciclos de mayor actividad en estas empresas, ya que es posible que haya una correlación entre los periodos de menor actividad con las actividades de reparaciones y por tanto la compra de material en nuestra tienda.

❖ ***Localice fuentes de datos abiertas que puedan usarse para conocer cuando tienen la mayor actividad las empresas del sector turístico que son clientes de nuestra tienda virtual, según el país donde se localicen.***

Por ejemplo, para conocer los meses de mayor ocupación hotelera podemos hacer uso de fuentes de datos fiables como son las bases de datos gubernamentales. Ya que nuestra empresa tiene base en España suponemos que nuestros clientes son españoles y europeos. Por tanto se pueden usar como fuente datos abiertas las siguientes:

- Encuesta de ocupación hotelera en España del INE[102].
- Ratio de ocupación hotelera a nivel europeo procedente de Eurostat[103].

4.2 Extracción, Limpieza, Transformación y Carga

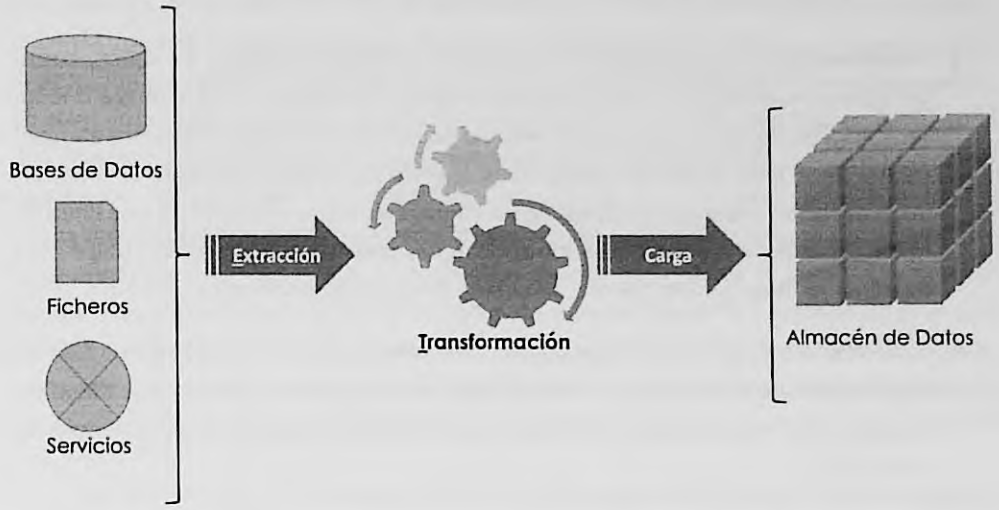
Una vez hemos estudiado las fuentes de datos, es necesario extraer los datos que necesitemos, transformarlos y cargarlos en nuestro repositorio de información. A estos procesos se los denomina procesos ETL –*Extract, Transform and Load*– (ver Figura 24), los cuales son necesarios para acceder a los datos de las fuentes de información y cargarlos de forma correcta en el almacén de datos (o datawarehouse). El proceso ETL se divide en 4 subprocesos:

1. **Extracción**: La extracción nos permite establecer los mecanismos por los que acceder a las diferentes fuentes de datos de interés de la empresa y extraer la información relevante para el analista de datos. Esta extracción podría realizarse de datos internos o externos en cualquier formato (e.g. bases de datos relacionales, ficheros de texto, hojas de cálculo, correos electrónicos).

[102] http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft11%2Fe162eoh&file=inebase

[103] http://ec.europa.eu/eurostat/web/tourism/data/database

Figura 24 – Proceso de carga de datos en el almacén de datos



2. **Limpieza:** Los datos, procedentes de las bases de datos operacionales y los sistemas transaccionales, pueden contener información sin depurar y deben ser limpiados. Por ejemplo si queremos procesar una dirección introducida en nuestra base de datos manualmente, podemos necesitar procesos de limpieza. Dada la dirección «Avda. de los Juegos Olímpicos 1, 2ºF, 29560, Álora, Málaga, España», podemos necesitar dividir esta información y homogeneizarla. La división de esta dirección podría ser la siguiente:

    - Tipo: Avda
    - Nombre: de los Juegos Olímpicos
    - Número: 1
    - Piso: 2º
    - Puerta: F
    - Código Postal: 29560
    - Población: Álora
    - Ciudad: Málaga
    - País: España

    Este proceso debería ser capaz de homogeneizar y corregir algunos errores:

    - Substituir el tipo de vía por Avenida (asumiendo Avda como su abreviatura).
    - Eliminar referencias o reparadores como «º».

- Usar código postal correcto de la población: 29500.
- Almacenar el nombre de los países en inglés (Spain).

3. **Transformación**: La transformación permite cambiar el modelo de datos, consiguiendo el paso efectivo de un sistema de gestión de datos transaccional a uno informacional. Por ejemplo, nuestro modelo de almacén de datos (o datawarehouse) podría almacenar la lista de poblaciones, usando su código postal para diferenciarlas, y de forma separada los datos de los clientes donde la dirección es una línea de texto. En este caso, sería necesario concatenar nuestros valores individuales para construir la dirección del cliente: «Avenida de los Juegos Olímpicos 1 2 F».

4. **Carga**: La fase final será la carga de los datos en el datawarehouse. En esta fase es importante que se verifique, como veremos en la siguiente sección, la calidad de los datos resultantes ya que en ellos se basarán los sistemas de apoyo a la toma de decisiones.

4.3 Calidad de los Datos

Aunque no sea el único factor determinante, es indudable que la calidad final de cualquier producto o servicio, es directamente proporcional a la calidad de las materias primas utilizadas para su elaboración. Del mismo modo, la calidad del conocimiento obtenido a partir de los datos, será directamente proporcional a la calidad de los datos utilizados para el análisis.

Si tenemos en cuenta que las organizaciones utilizan como base para tomar sus decisiones, tanto operativas, como estratégicas, el análisis de los datos, bien generados por sus sistemas transaccionales, o bien procedentes de fuentes externas, la calidad de los datos de partida se convierte en un asunto de vital importancia. Dicho de otro modo, si los datos en los que basamos nuestros análisis, no tienen la calidad adecuada, ya sea porque no son fieles a los hechos que representan o porque son incompletos, las conclusiones obtenidas serán erróneas y por tanto las decisiones tomadas, inadecuadas.

Diferentes estudios y encuestas han tratado de cuantificar el efecto que la mala calidad de los datos utilizados puede tener en los resultados de una compañía. Dos estudios realizados en diferentes años[104,105], cuantificaron el impacto negativo, que la utilización de datos de mala calidad, en una empresa de gran tamaño entre un 8 % y un 12 % de las ventas brutas.

[104] Thomas C. Redman. «The Impact of Poor Data Quality on the Typical Enterprise». En: *Commun. ACM* 41.2 (feb. de 1998), págs. 79-82. ISSN: 0001-0782. DOI: 10.1145/269012.269025. URL: http://doi.acm.org/10.1145/269012.269025

[105] Anders Haug, Frederik Zachariassen y Dennis van Liempd. «The costs of poor data quality». En: *Journal of Industrial Engineering and Management* 4.2 (2011), págs. 168-193. ISSN: 2013-0953. DOI: 10.3926/jiem.2011.v4n2.p168-193

Acercándonos a la realidad

Sin embargo, aunque las pérdidas siempre pueden cuantificarse económicamente, la forma en la que estas se manifiestan en la empresa es muy distinta. A veces se traducen en la denegación de créditos a clientes, otras en la prescripción incorrecta de medicamentos, en el incorrecto etiquetado de productos alimenticios o en accidentes fatales. El Instituto de Finanzas Internacionales y la consultora McKinsey & Company (2011)[106] citaron como uno de los factores clave en la crisis financiera global, que comenzó en 2007, la inadecuada arquitectura de datos y las tecnologías utilizadas para dar soporte a la gestión del riesgo financiero. Durante esta crisis, muchos bancos, sociedades de inversión y compañías de seguros perdieron miles de millones de dólares, llevando a algunos de ellos a la quiebra. Además del impacto económico directo, supuso el inicio de la recesión económica, con millones de hipotecas ejecutadas, pérdida de empleos, la quiebra de fondos de pensiones y la pérdida de confianza de los ciudadanos en el sistema financiero.

- ***Revise el estudio sobre el Big Data como clave para la gestión de riesgos financieros[107].***

Por lo tanto, es importante diseñar y poner en práctica metodologías que permitan cuantificar el impacto de la falta de calidad de los datos y nos ayuden a identificar aquellos factores que afectan a esta calidad, todo ello para poder tomar las medidas adecuadas que nos permitan aumentar la calidad de los datos empleados.

Todo ello ha llevado a las compañías a poner el foco en la calidad de datos, considerando los datos como un recurso tan importante y crítico como otros, tales como las personas, el capital, las materias primas o las instalaciones. Las empresas más adelantadas ya cuentan con un responsable de esta tarea: el *Chief Data Officer* (CDO) o Director de Datos.

La calidad de los datos es fundamental, como afirma Bill Inmon[108], considerado el padre del *data warehousing*, en un artículo aparecido en «Business Intelligence Network» sobre calidad de datos:

> *Las organizaciones actúan bajo la premisa de que la información de que disponen es precisa y válida. Si la información no es válida, entonces no pueden responder de las decisiones basadas en ella.*

Consecuentemente, es necesario asegurar que la calidad de los datos sea máxima. Si en los datos hay errores, estos se propagarán a lo largo de toda la organización y serán muy difíciles de localizar. Los errores en los datos pueden provenir de:

---

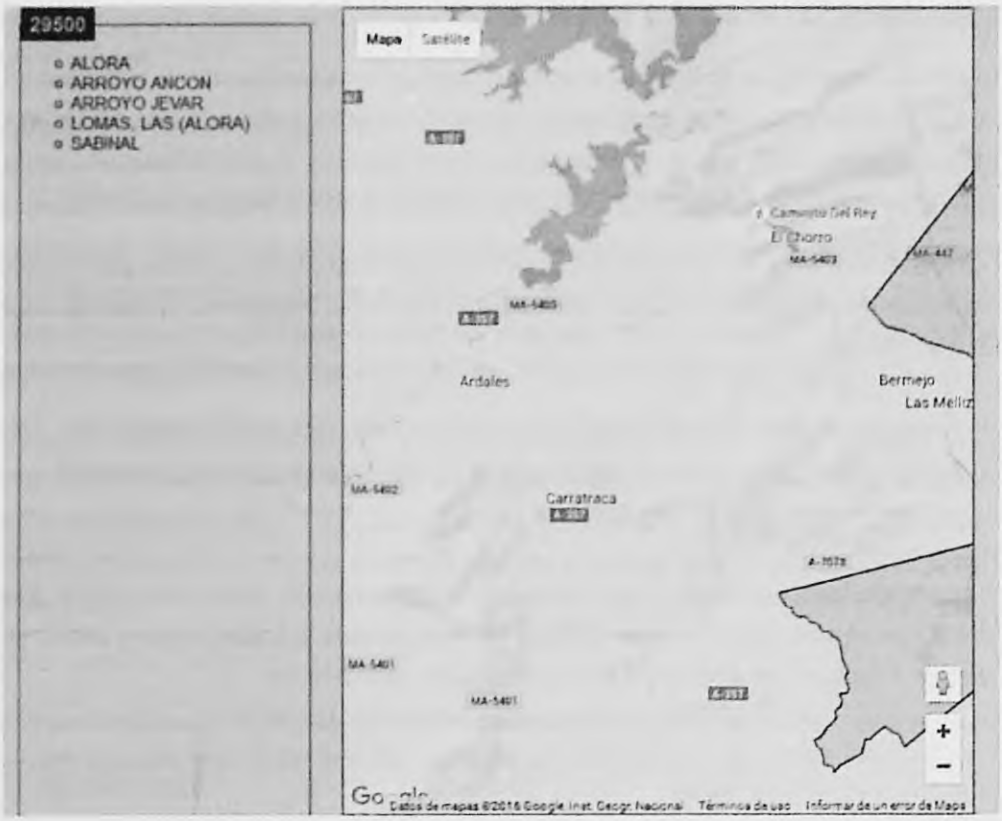
[106] James Manyika y col. *Big data: The next frontier for innovation, competition, and productivity*. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. [Online; consultado el 7 de enero de 2016]

[107] http://www.economistinsights.com/sites/default/files/RetailBanksandBigData.pdf

[108] Bill Inmon. *Elusive Data Quality*. http://www.b-eye-network.com/view/8629. [Online; consultado el 7 de enero de 2016]

- El origen de los datos, es decir los sistemas transaccionales de los que recuperamos los datos. El punto de entrada de los datos es muy importante, ya que, si los datos están mal, la inexactitud se propaga por todo el sistema y es muy difícil de encontrar y corregir una vez propagado. Muchos de estos errores se producen porque los usuarios pueden introducir datos sin ningún tipo de control. No es una buena opción corregir estos errores en el proceso ETL sin proceder a la modificación de las aplicaciones de origen, puesto que los errores se seguirán produciendo en futuras cargas. Esta opción, de corrección posterior a la captura de los datos, a pesar de ser mucho más rápida inicialmente es mucho más costosa a largo plazo. Mitigar los errores en la entrada de datos, normalmente pasa por definir estrategias que permitan reducir los errores, por ejemplo:
    - Automatizar al máximo la entrada de datos, evitando que los usuarios introduzcan datos libremente, como cuando se permite que elijan entre distintos valores o cuando se calculan valores –por ejemplo el código postal de una dirección–. Por ejemplo (ver Figura 25), la aplicación, dado un código postal nos podría mostrar las posibles localidades y el mapa de la zona incluida en el código postal[109].

Figura 25 – Ejemplo de cálculo de poblaciones en base a su código postal. Poblaciones incluidas en el código postal 29500.



[109] http://www.codigospostales.com/mapa.cgi?codigo=29500

- Establecer reglas de entrada de datos. Si etiquetamos el sexo como H para hombre y M para mujer, que la aplicación no admita otros valores.
- Homogeneizar el formato de los datos. Es decir, incluir en el proceso de captura de datos la transformación en tiempo real de los mismos. Por ejemplo, la aplicación puede cargar siempre el dato en mayúscula, independientemente de que el operador utilice mayúsculas o minúsculas para el apellido de un cliente.

- Los procesos de integración. Los procesos ETL deben trabajar con datos de diferente procedencia, lo que significa tener que trabajar con bases de datos alimentadas por distintas personas y aplicaciones que pueden no tener los mismos criterios de carga. Un ejemplo muy habitual y simple es la homogeneización de las fechas. Diferentes aplicaciones o diferentes configuraciones de un mismo sistema gestor de bases de datos, pueden trabajar con formatos de fecha distintos: *dd/mm/aa*, *dd/mm/aaaa*, *mm/dd/aa* o formato largo *día mes año*.

- El propio *data warehouse*. Todos los procesos de cálculo y agregación que se hacen en nuestro propio datawarehouse, pueden ser una fuente de errores que hay que vigilar continua y proactivamente.

Asumir que la calidad de nuestros datos es buena puede ser un error fatal. Normalmente, cuando se construye un sistema central de gestión de datos, la mayoría de las organizaciones se enfocan en identificar que datos necesitan analizar, los extraen de las fuentes de origen y finalmente proceden a cargarlos. Generalmente no se piensa en la calidad de los datos, permitiendo la carga de los errores. Las comprobaciones se deberán llevar a cabo, de forma manual o automatizada, teniendo en cuenta distintos niveles de detalle y variando los periodos de tiempo. Así por ejemplo, se puede comprobar que los datos cargados coinciden con los de las fuentes de datos origen; por ejemplo, comprobando que las ventas totales o el número de pedidos coinciden diariamente con la información cargada en el *data warehouse*.

El proceso de calidad de los datos debe ser continuo e iterativo para conseguir la mejora en la calidad de los datos. Este proceso nos puede ayudar a mejorar nuestros sistemas transaccionales, corregir errores en el *data warehouse*, mejorar el proceso ETL o incluso mejorar los modelos de negocio por parte de los usuarios.

Acercándonos a la realidad

Según lo que plantea la norma ISO 9000: 2000, la calidad se podría definir como «el grado en el que un conjunto de características inherentes cumple con los requisitos, esto es, con la necesidad o expectativa establecida, generalmente implícita u obligatoria».

❖ ***¿Qué beneficios podrán obtener las empresas que le dan importancia a la calidad de sus datos?***

En general, les permiten obtener beneficios claves para agregar valor al negocio y diferenciarse del resto de sus competidores, mediante:

- La minimización los riesgos en sus proyectos, especialmente en los relacionados con Tecnologías de la Información.
- El ahorro de tiempo y recursos, haciendo un mejor uso de la infraestructura tecnológica y sistemas para explotar su información.
- La toma de decisiones de negocio oportunas, en base a información confiable, validada y limpia.
- La adaptación a estándares o regulaciones internacionales sobre el manejo de información, permitiendo facilidad al momento de ejecutarlas.
- La mejora de la confianza, buenas relaciones e imagen de la empresa antes sus clientes frente a la competencia.

El problema de la calidad de los datos, debe ser considerado como un asunto estratégico, al que debemos asignar objetivos, recursos y planificación. La responsabilidad de la calidad de los datos no corresponde sólo a los departamentos de tecnología, sino también a los responsables de negocio, así como a cada uno de los responsables de los distintos procesos y las aplicaciones que los soportan.

4.3.1 Técnicas de Profiling

Previamente a la construcción del sistema ETL, una tarea fundamental consiste en la correcta identificación de los orígenes de datos que permitirán el llenado de nuestro sistema de almacén de datos. Esta tarea se llevará a cabo mediante las técnicas de profiling.

El *profiling* de datos es el análisis estadístico de los **valores de los datos** que contiene un conjunto o base de datos para evaluar su consistencia, originalidad y lógica.

El proceso de *profiling* de datos no sirve para identificar los datos erróneos o imprecisos, sino que «sólo» se utiliza para comprobar que los valores datos cumplen con una serie de reglas de negocio previamente establecidas. De este modo, se obtendrá una idea clara de lo difícil que será utilizar los datos existentes para otros propósitos en nuestra organización. Se utiliza también para generar métricas de evaluación de la calidad de los datos.

Por tanto, las técnicas y herramientas de *profiling* sirven para evaluar el contenido real, la estructura y la calidad de los datos, mediante la exploración de las relaciones que existen entre los valores a lo largo y ancho de la base de datos.

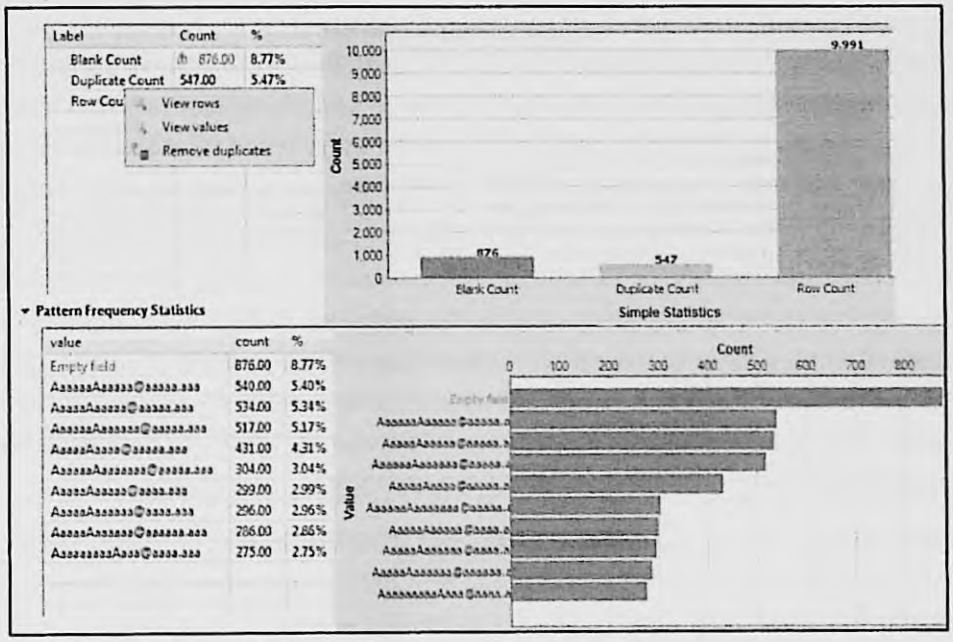
Acercándonos a la realidad

En la actualidad, existen herramientas automáticas para realizar profiling de datos. Un ejemplo de bastante éxito es «*Talend Open Studio*»[110]. Esta herramienta nos permite realizar tareas específicas de profiling de datos como:

- Comprobar los rangos de los datos por categoría, para analizar los campos de las bases de datos por texto o valor numérico;
- Aplicar reglas de negocio para identificar valores que sobrepasan, o no alcanzan, ciertos umbrales predefinidos. Por ejemplo, filtrar valores de fechas entre 1 de enero de 2010 y 31 de diciembre de 2016;
- Identificar valores de datos que no cumplen con especificaciones estándares de códigos, como identificadores comerciales de productos («Stock-keeping unit» o SKU, «Global Trade Item Number» o GTIN, etc.), direcciones de e-mail («`nombre@dominio.com`»), códigos postales, etc. Un ejemplo ilustrativo de la operación de profiling sobre el dato «dirección de e-mail» se puede observar en la Figura 26, donde se calcula el porcentaje de las cuentas por categorías (Blank, Duplicate y Row) y por patrón de dirección (*Pattern Frequency Statistics*.)

El *profiling* de datos es un proceso que **implica aprender de los datos.** En este proceso se emplean técnicas de descubrimiento y análisis para encontrar características de los datos

Figura 26 – Ventana típica de Talend Open Studio para el profiling sobre el valor del dato «dirección de e-mail»



[110]Talend. *Talend Open Studio for Data Quality*. https://www.talend.com/resource/data-profiling.html. [Online; accedido el 12 de enero de 2016]

que puedan ser estudiadas por un analista de negocio para determinar si el dato encaja con los objetivos de negocio.

Las fases de los procesos de evaluación de la calidad de los datos podrían resumirse en tres etapas:

1. ***Profiling* de datos**: la aplicación de técnicas de *profiling* de datos ayuda a descubrir y definir los requisitos de calidad de los datos. El software de *profiling* de datos se utiliza como inicio del proceso de descubrimiento, pero no de evaluación. Se trata de encontrar las reglas y requisitos que ayudarán a llevar a cabo una evaluación posterior más exhaustiva de la calidad de datos.

2. **Determinación de los requisitos de calidad de los datos**: gracias a la visibilidad proporcionada por la acción del *profiling* de datos, se está en disposición de empezar a definir algunas de las reglas de calidad que los datos deberán cumplir. El objetivo es ser capaces de comparar la calidad de los datos con respecto a lo establecido por el conjunto de criterios aprobados.

3. **Evaluación de la calidad de los datos**: una vez realizado el *profiling* de datos y descubiertos los requisitos o normas de calidad de los datos, es momento de aplicar las reglas para proceder a la evaluación de la calidad de los datos. En este proceso se registran los datos que se considera que han superado la prueba de calidad y los que no.

El *profiling* de datos tiene dos objetivos distintos, uno estratégico y otro táctico. Como objetivo estratégico, en el momento en el que se identifica una fuente de datos, debemos realizar un *profiling* básico para evaluar si dicha fuente de datos es adecuada para incluir sus datos en nuestro almacén de datos (o data warehouse) y ofrecer una decisión acerca de si seguir adelante o no con dicha fuente. Idealmente, esta evaluación debería realizarse inmediatamente después de identificar una fuente de datos candidata, durante el proceso de análisis de los requisitos de negocio.

Como objetivo táctico, el proceso de *profiling* de datos ofrece a los diseñadores de los procesos ETL una guía de cuántos procesos de limpieza de datos (*data cleansing*) van a tener que realizarse. Además, este proceso los protege de posibles retrasos en los plazos del proyecto debidos a resultados inesperados del sistema por tener que gestionar datos incorrectos. Por tanto, el *profiling* de datos debe realizarse lo antes posible.

Tradicionalmente, el *profiling* está relacionado con el uso de los datos en entornos en los que el propósito para el que se van a usar está bien definido por adelantado. En proyectos de Big Data, los datos pueden tener que usarse de manera distinta a la que se pensó originalmente, y por tanto, hay que realizar el *profiling* de forma consecuente para poder abordar la reutilización de los mismos. Esto se debe al hecho de que la mayoría de

los proyectos de Big Data tienen como objetivo realizar análisis exploratorios sobre datos mal definidos y por tanto, tratan de averiguar cómo utilizar o reutilizar dichos datos. Para conseguir esto debe articularse, cuantificarse y poner a disposición de los usuarios distintas medidas de calidad, entre las que cabe destacar la completitud, la correctitud o la coherencia:

- **Completitud**: esta dimensión es una medida para conocer si se tienen todos los datos necesarios para responder a las consultas. Para evaluar si un conjunto de datos es completo, hay que empezar por definir las consultas que se quieren realizar y determinar los campos necesarios y el porcentaje de registros necesarios para responder cómodamente a esas consultas. Si los datos se consideran incompletos, los registros que falten se pueden reparar, eliminar, marcar o ignorar dependiendo de los casos de uso del análisis.
- **Correctitud**: esta dimensión mide la exactitud de los datos. Para descubrir si los datos son correctos, hay que estimar qué significa que los datos sean incorrectos. Esto depende estrictamente del contexto empresarial. Por ejemplo, en aquellos casos en que los datos tiene que ser únicos, los datos duplicados se consideran incorrectos.
- **Coherencia**: esta dimensión mide si los datos tienen sentido de forma aislada y determina si los registros se relacionan unos con otros de manera consistente siguiendo la lógica interna del conjunto de datos.

4.3.2 El Muestreo de Datos

El muestreo de los datos, también conocido como *«Sampling»*, es una técnica que se realiza para estimar la calidad de los datos mediante el análisis estadístico de un subconjunto de estos, en lugar de analizar el conjunto completo.

Uno de los criterios más importantes a la hora de seleccionar una muestra es su *representatividad*, que determina cuánto se asemeja la muestra al conjunto completo de datos. La representatividad tiene que ser alta si queremos obtener un resultado exacto. El tamaño de la muestra tiene también un impacto importante en la exactitud en la representación. El muestreo de enormes volúmenes de datos balancea el compromiso entre coste y calidad del *profiling* de datos, ya que es muy costoso y complejo realizar el *profiling* del conjunto de datos completo.

No obstante, debido a la complejidad del análisis a realizar, no todos los tipos de datos y métodos de recolección de datos van a requerir muestreo.

Acercándonos a la realidad

- ***¿Se le ocurre alguna situación en la que no se requiera realizar el muestreo previo de los datos?***

En los sistemas de recomendación en «*E-Commerce*» y de análisis de flujos de clics («*Clickstream*», como se verá en el Capítulo 11), los datos tienen que ser analizados en su totalidad, sin opción a realizar muestreo, ya que el muestreo podría introducir ciertos sesgos y reducir la exactitud de los resultados.

Acercándonos a la realidad

Según Kimball [111], el *profiling* de los datos se debe realizar varias veces y con intensidad variable a lo largo del proceso de desarrollo de un almacén de datos. La evaluación de los perfiles debe llevarse a cabo tan pronto como se identifiquen las fuentes de datos candidatas y justo después de la adquisición de los requisitos de negocio. El propósito general es el de clarificar si se disponen los datos correctamente y en un nivel de detalle suficiente de manera que se puedan detectar las anomalías existentes desde un estado temprano. Si este no es el caso, entonces el proyecto debería ser cancelado.

4.4 Preparación de los Datos

El propósito fundamental de la preparación de los datos es la manipulación y transformación de los datos sin refinar para que la información contenida en el conjunto de datos pueda ser descubierta o estar accesible de forma más fácil. La preparación de datos engloba a todas aquellas técnicas de análisis de datos que permiten mejorar la calidad de un conjunto de datos, de modo que las técnicas de extracción de conocimiento mediante minería de datos puedan obtener mayor y mejor información (mejor porcentaje de clasificación, reglas con más completitud, etc.)

Los datos pueden estar dispersos en la empresa y almacenados en formatos distintos. También pueden contener incoherencias como entradas que faltan o entradas incorrectas. Por ejemplo, los datos pueden mostrar que un cliente adquirió un producto incluso antes de que se ofreciera en el mercado o que el cliente compra regularmente en una tienda situada a 2.000 kilómetros de su casa. Por ello, es necesaria una correcta preparación.

La preparación de los datos, aunque parece sencilla, es un proceso que junto con la selección de datos consume el 70 % del esfuerzo en los proyectos de minería de datos de nueva implantación. En concreto, hay que tratar de asegurar los siguientes aspectos:

- Que los datos tengan la calidad suficiente, es decir, que no contengan errores, redundancias ni falta de valores.
- Que los datos sean los necesarios: quizá habrá datos que no nos hagan falta y quizá tengamos que añadir otros nuevos. De hecho, no es común que los datos que necesitamos realmente hayan sido recogidos por el sistema con el propósito de llevar a cabo justamente el tipo de estudio que queremos emprender.

[111] Ralph Kimball. *The Data Warehouse ETL Toolkit*. Wiley, 2004

- Que los datos estén en la forma adecuada: muchos métodos de construcción de modelos requieren que los datos estén en un formato determinado, que no tiene por qué coincidir necesariamente con el formato en que se encuentran almacenados.

4.4.1 Operaciones de Preparación de Datos

Dentro de las operaciones de preparación de los datos, cabe destacar dos técnicas principales: la normalización y la **discretización.**

La normalización consiste en situar los datos sobre una escala de valores equivalentes que permita la comparación de atributos que toman valores en dominios o rangos diferentes. Si no hay normalización previa, los métodos de análisis (o de minería de datos) tienden a quedar sesgados por la influencia de los atributos con valores más altos, hecho que distorsiona el resultado.

La *discretización* consiste básicamente en establecer un criterio por medio del cual se puedan dividir los valores de un atributo en dos o más conjuntos disjuntos. Por ejemplo, para información referente a un rango de edades entre 10 y 80 años, se podría discretizar entre los mayores y menores de 30 años, es decir, jóvenes y adultos, respectivamente. Discretizar datos es útil para:

- Reducir el coste computacional pues se trabaja sobre un conjunto de valores menor.
- Reducir la velocidad en el proceso de aprendizaje automático.
- Los valores discretos necesitan menos memoria para ser almacenados.
- Se reduce el tamaño del modelo resultante. Cuando se trabaja con datos continuos, los modelos de clasificación que se obtienen son mayores. Por ejemplo, los árboles de decisión acostumbran a tener un factor de ramificación más alto cuando se trabaja con datos continuos que cuando se hace con datos discretos. Describiremos más en detalle estos modelos en el Capítulo 8.
- Mejoran la comprensión al describir el elemento utilizando menos términos.

Lo que se busca con la discretización es mantener la información asociada al atributo que se discretiza, aunque tiene el inconveniente de la pérdida de información aportada por los valores continuos; el objetivo es por tanto, obtener una división en intervalos a partir de un atributo numérico continuo, de manera que a cada intervalo se le pueda asociar una etiqueta.

4.4.2 Tratamiento de la Falta de Datos

Uno de los problemas más habituales en el tratamiento previo de los datos es la ausencia de valores para un atributo determinado. En estos casos se suele optar por sustituir el valor que falta por: la media de los valores que presenta el atributo en los datos o el valor más frecuente. Por ejemplo, si tenemos un dato que es la edad del cliente, en caso de no disponer del dato para un cliente concreto podríamos usar la edad media de nuestros clientes o el valor de edad más frecuente.

❖ ***¿Se le ocurre en qué situación puede ser más útil cada una de las alternativas planteadas para la substitución de valor?***

Cada alternativa puede llegar a dar resultados diferentes, especialmente cuando el número de valores que faltan es elevado. En general, estos métodos demasiado sencillos llevan a introducir sesgo en los datos. Es común que en los casos para los cuales no se ha recogido una observación se asigne un valor especial (por defecto), por ejemplo, para el caso de las edades de los clientes podríamos usar -1 o 999. Es importante no tratarlos como un valor numérico más, sino como lo que realmente son: ausencia de información con respecto a un atributo en un caso observado.

4.4.3 Limpieza de Datos

La limpieza de datos básicamente consiste en eliminar aquellos que sean erróneos o redundantes. Los datos recogidos a mano o procedentes de la fusión de varias bases de datos suelen mostrar factores de distorsión importantes que hay que limpiar, entre los que cabe destacar:

- **Datos incompletos**: en aquellos datos definidos inicialmente como «no obligatorios» o de «formato libre», puede suceder que al realizar la entrada de datos correspondiente queden incompletos.
- **Datos redundantes**: a veces se repiten los valores de los datos referentes al mismo atributo, especialmente cuando fusionamos distintas bases de datos. Suele suceder cuando no se lleva el control de las redundancias, por ejemplo, cuando se cargan varias veces los datos de contacto de un mismo cliente o usuario.
- **Datos incorrectos o inconsistentes**: caso muy común cuando el tipo de valores que puede recibir un atributo no está controlado porque ha sido declarado como «texto libre», o bien está definido como un tipo determinado (cadenas alfanuméricas por ejemplo) pero no se han mantenido los procesos de control de errores necesarios.
- **Errores de transcripción**: muy típicos, como por ejemplo mayúsculas y minúsculas.

- **Datos envejecidos**: ciertos datos se convierten en incorrectos porque no han sido actualizados de la manera adecuada. Por ejemplo, es mejor introducir la fecha de nacimiento que la edad del sujeto, pues con el paso del tiempo, si no se actualiza esta edad, el sujeto siempre tendrá la misma. En cambio, con la fecha de nacimiento, siempre podemos actualizar los datos a la fecha actual.

- **Variaciones en las referencias a los mismos conceptos**: por ejemplo, un abogado puede ser considerado como profesional liberal o como autónomo.

- **Datos sesgados**: puede darse este problema con datos que reflejan en conjunto un valor determinado. También en el contexto en el que los valores procedan de un conjunto de datos muy determinado. Por ejemplo, los datos procedentes de una encuesta de conocimiento en nuevas tecnologías de comunicación, aquellos valores procedentes de las encuestas realizadas en las Escuelas de Ingenierías de Telecomunicaciones estarán sesgadas hacia un grado de conocimiento alto.

Al llevar a cabo el proceso de limpieza en entornos de Big Data, se debe tener en cuenta el compromiso entre calidad de los datos y tiempo empleado en su limpieza.

❖ ***Reflexione sobre el por qué de esta afirmación. ¿Es rentable la limpieza exhaustiva de datos cuando se manejan grandes volúmenes de estos?***

Por lo general, es necesario realizar una secuencia de tareas automáticas de limpieza de datos de forma iterativa hasta que se hayan explorado todos los datos y se hayan localizado las posibles anomalías. Estas tareas de limpieza deben hacer referencia a los atributos y a las reglas de negocio relevantes, para así alcanzar una mayor calidad de datos.

Los datos de mala calidad pueden surgir en las distintas fases del proceso de gestión de los datos, es decir, recogida, entrega, almacenamiento e integración:

- Problemas en la limpieza del Big Data derivados de la **recogida de datos**: pueden surgir inconsistencias en los datos debido a los errores en los métodos de recolección de los mismos. Estos métodos pueden consistir en la entrada manual de datos de redes sociales (donde no se utilizan palabras estándar), la introducción de datos duplicados y los errores de cálculo.

- Problemas en la limpieza del Big Data derivados de la **entrega de datos** inadecuada: esto puede deberse a la conversión incorrecta de los datos después de que hayan pasado por un sistema de entrada de datos integrado con Hadoop, es decir, problemas de almacenaje, transmisión, etc.

- Problemas en la limpieza del Big Data derivados del **almacenamiento de los datos**: pueden surgir datos inconsistentes como resultado de problemas en el almacenamiento físico y lógico de los datos. Ocurre cuando los datos se almacenan por un largo periodo de tiempo y tienden a corromperse, lo que se conoce como degradación de bits (del inglés «*bit rot*»).

- Problemas en la limpieza del Big Data derivados de la **integración de los datos**: integrar datos de fuentes heterogéneas contribuye de forma significativa a la pérdida de calidad de los datos, dando lugar típicamente a casos de registros inconsistentes.

4.4.4 Transformación de los Datos

En los **procesos ETL**, tras la extracción de los datos de la fuente o fuentes de origen llega la segunda fase: la transformación. La fase de transformación de un proceso ETL consiste en la **aplicación de una serie de funciones o reglas de negocio sobre los datos extraídos para convertirlos en datos que, a continuación, serán cargados en la nueva fuente**.

Para entender la necesidad de un proceso de transformación debemos tener en cuenta que en un proceso ETL se manejan fuentes diversas, algunas de ellas de fuera de la propia organización: información bursátil de una Web ajena a la empresa, cualquier tipo de descarga de Internet, un paquete de Office, etc.

Esta variedad de bases de datos, en ocasiones de varios países, con diferentes idiomas y distintas unidades de medida, imposibilita o dificulta la posibilidad de realizar comparaciones si con anterioridad no se realizan conversiones y cambios de formato. De ahí la necesidad de los procesos de transformación.

Esta función corresponde al desarrollador o analista del proceso ETL en cuestión. La definición de las transformaciones a realizar se realiza en función de un análisis previo y de la fase de limpieza que, como ya hemos señalado y profundizaremos más adelante, se trata de un proceso separado pero estrechamente ligado al de transformación.

La transformación de los datos, tras la extracción de los mismos y como paso previo a su carga, no puede considerarse una fase secundaria ni prescindible. Sin un buen trabajo de transformación de datos no sería posible realizar comparaciones y análisis. Dicho de otro modo, se renunciaría a uno de los grandes beneficios para las organizaciones al implementar un proceso ETL.

Existen una serie de mecanismos de transformación que pueden ser empleados en esta etapa dependiendo de la naturaleza de los datos a tratar:

- **Datos numéricos a categóricos**: los datos categóricos son los que toman valor en un conjunto finito valores. Por ejemplo, nuestros datos pueden recoger la edad de nuestros clientes. Pero el análisis de datos que vamos a realizar sólo

requiere conocer el rango de edades de los mismos. Este tipo de transformación podría modificar estos datos de forma que se cambiara la edad de los clientes por valores categóricos:

- 18-30 Años: Jóvenes.
- 31-65 Años: Adultos en edad laboral
- 66 Años en adelante: Clientes en edad de jubilación.

- **Datos categóricos a numéricos**: disponemos de datos que aparecen descritos mediante valores categóricos y necesitamos disponer de los valores numéricos correspondientes. Por ejemplo, si hemos obtenido valores de descuento aplicado a la venta de varios productos, y estos descuentos están categorizados (descuento de cliente frecuente, descuento de cliente ocasional, descuento de familiar), podemos obtener el descuento real para cada categoría (20 %, 10 %, 5 %).
- **Simplificación de valores**: dividir por ejemplo los sueldos por mil o un millón.
- **Agrupación de valores continuos**. Esta transformación aplica operaciones de agregación en los datos, por ejemplo, computar resultados mensuales y anuales de los datos de bolsa diarios.
- **Normalización de datos**: poner los valores numéricos en un intervalo determinado. Típicamente, el valor de un atributo se transforma para que se encuentren en el rango entre cero y uno. Esto se hace para eliminar ciertos efectos no deseados en ciertos atributos del análisis.
- **Añadir una etiqueta**: que indique a qué clase pertenece un registro. Por ejemplo, si estamos gestionando las ventas, podemos recuperar datos sobre las devoluciones y añadir una etiqueta («producto devuelto») para reflejar aquellos registros de productos que han sido devueltos por el cliente.
- **Conversión de unidades**: por ejemplo, convertir millas en kilómetros por hora o viceversa. Algo muy habitual cuando se extraen datos de países con unidades métricas distintas. Otro caso sería la conversión de diferentes monedas (libras, euros...) en un único valor estándar.
- **Selección de columnas para su carga posterior**: por ejemplo, hacer que las columnas con valores nulos no se carguen.
- **Agregación de columnas**: añadir una columna con la procedencia de determinados automóviles sería un ejemplo.
- **Dividir una columna en varias**: esta acción resulta de gran utilidad para, por ejemplo, separar en tres columnas (una para el nombre y otras dos para los apellidos) la identificación de una persona que antes estaba en un solo campo.

- **Traducir códigos**: por ejemplo, si la fuente de origen almacena una «H» para hombres y una «M» para mujeres, dar las instrucciones necesarias para que en destino se guarde un «1» para hombres y un «2» para mujeres.

- **Obtener nuevos valores calculados**, como por ejemplo la suma total de ventas de cada tienda en un día.

- **Unir datos de varias fuentes**. En el proceso de transformación será necesario combinar datos de varias de las fuentes presentes en la empresa. Así, puede que necesitemos unir los datos del departamento de recursos humanos con el departamento de ventas para poder disponer de los detalles de los vendedores que hicieron cada una de las ventas. Sin esta unión de datos tendríamos únicamente un código de empleado, pero no sus datos personales, oficina en la que trabaja, etc.

- Lookups: es cuando se toma un dato y se lo compara con otro tipo de datos, cruzando información. Por ejemplo, capturar un código de cliente de una base de datos y cruzarlo con otra base de créditos concedidos para saber si dicho cliente disfruta o no de ese préstamo.

- **Generalización**: en esta transformación, los datos detallados se sustituyen por un valor agregado a menor nivel de detalle. Por ejemplo, datos detallados como las ventas de cada cliente para su localización física (la calle en la que vive), pueden reemplazarse por abstracciones superiores, como ciudad o estado, según el análisis que se vaya a realizar. De esta forma podemos pasar de analizar las ventas en una dirección concreta a su análisis en una ciudad o estado completo.

El proceso de transformación de los datos en uno de los bloques de construcción fundamentales y un paso vital en el proceso de descubrimiento de conocimiento de análisis del Big Data. En este sentido, las transformaciones de datos predominantes en el Big Data son: normalización, agregación, generalización y unión de varias fuentes de datos.

En el Capítulo 11 se presenta un caso de uso en el que se aplican las técnicas descritas en este capítulo, aplicadas a la transformación y carga en el Big Data. Como parte de dicho dicho Capítulo 11 se muestra como se transformarán los datos para su posterior carga en HCatalog.

5 Recolección de Datos

5.1 El Modelo de Negocio

Cuando trabajamos con datos, independientemente del área de actividad en el que nos movamos, siempre pretendemos medir algo. Así por ejemplo en el entorno empresarial, cuando se trabaja en proyectos de *Business Intelligence*, es muy común medir resultados de un determinado área de negocio: Ingresos, gastos, ventas, resultado neto, etc. Para posteriormente comparar con algún indicador de referencia que nos permitirá identificar si los resultados son buenos, malos o los esperados y así poder tomar las acciones adecuadas.

Para poder trabajar con datos en problemas del mundo real, necesitamos ser capaces de trasladar las particularidades y la lógica del problema que estamos tratando, o de nuestro negocio (o al menos de aquellas partes que nos interesa estudiar), a la realidad tecnológica en la que almacenaremos y gestionaremos los datos. Es decir, si volvemos a tomar como ejemplo el mundo empresarial, deberemos ser capaces de trasladar la realidad de nuestro negocio (todos aquellos entes que participan y todos aquellos procesos que intervienen en el desarrollo del negocio) a las estructuras técnicas, más o menos interrelacionadas, donde se almacenaran los datos.

Para tal fin, el primer paso consistirá en acotar que partes de nuestro negocio serán objeto de nuestro estudio, es decir, qué queremos medir y cómo lo vamos a hacer. Para ello, debemos definir cuál es nuestro modelo conceptual del problema, que puesto que estamos hablando de un entorno empresarial, suele denominarse modelo de negocio, y cuáles son las métricas que vamos a utilizar para medirlo.

Definir el modelo de negocio nos facilitará la siguiente tarea, la de definir el modelo de datos que nos permita utilizar los datos en beneficio del negocio.

Acercándonos a la realidad

Para introducir el concepto de modelo de negocio, supongamos que un minorista, propietario de un pequeño comercio físico, decide comercializar sus productos en Internet, mediante la creación de una tienda de comercio electrónico.

Si bien es cierto que algunas de las decisiones iniciales pueden ser tomadas de forma arbitraria, basadas más en el conocimiento del negocio, en la intuición o en el sentido común, más que en los datos, ya que en el arranque no se dispondrá de ninguno; antes de poner en marcha la tienda de comercio electrónico, es necesario identificar las particularidades del negocio online, para ser capaces, posteriormente, de medir su efecto en el negocio una vez que se ponga en marcha.

Debemos comenzar por entender cómo funciona nuestro negocio en Internet. Como en el mundo físico, nuestra tienda vende productos, que previamente habremos comprado a nuestros proveedores, a los clientes que visitan la tienda online. Si bien es cierto que esto no deja de ser una transacción típica de cualquier establecimiento físico, nuestro modelo de negocio online tiene ciertas peculiaridades, propias del comercio electrónico, que debemos tener en cuenta. Puesto que el propósito es ilustrar la necesidad de conocer el modelo de negocio para ayudarnos en la definición de nuestro modelo de datos, no haremos una relación exhaustiva de estas diferencias, simplemente citaremos algunas.

Así por ejemplo, la peculiaridad más evidente de cualquier transacción online, es que esta se produce de forma virtual, no presencial, lo que significa que será necesario contemplar procesos y procedimientos específicos asociados a esta particularidad:

- El pago de las compras realizadas en el mundo físico normalmente se realiza mediante dinero en metálico o mediante tarjeta de crédito. En el mundo online, el pago en metálico no es posible, debido a que la transacción no es presencial. El pago mediante tarjeta de crédito si lo es, sin embargo, también debido a que la transacción no se realiza en presencia física del cliente, es necesario establecer mecanismos y procedimientos que garanticen la autenticidad de transacción, es decir debemos asegurarnos que quien está realizando el pago, es realmente el dueño de la tarjeta y no se trata de un uso fraudulento de la misma. Además, el mundo online permite la realización de pagos de transacciones por otros métodos, como la transferencia bancaria o los pagos a través de otros agentes (como PayPal) entre otros mecanismos. Estos métodos de pago propios del mundo online, tienen asociados unos costes específicos y unos procesos especiales. ¿Cómo afectan estos costes de transacción asociados a diferentes métodos de pago a la rentabilidad? ¿Es nuestro ingreso por cada producto suficiente para soportar estos costes?

- Una vez finalizado el proceso de pago de la compra, se debe iniciar un proceso de entrega a domicilio. Evidentemente este proceso lleva asociado un coste. ¿Debemos cargar este coste al cliente? ¿Debemos hacerlo siempre? ¿En qué condiciones podemos asumir este coste y no repercutírselo al cliente?

- Los clientes no están físicamente cerca de nuestro local, no pasan físicamente por la puerta de nuestro local, ni pueden entrar a preguntarnos en caso de duda. Por el contrario, nuestros potenciales clientes están en Internet, lo que significa que pueden estar físicamente en cualquier lugar del mundo, lo que nos obliga a hacernos nuevas preguntas: ¿pueden nuestros productos ser atractivos para clientes de otros países? Si es así, ¿debemos traducir los contenidos de nuestra tienda a otros idiomas? ¿Cuáles? ¿Podremos identificar la procedencia de una transacción? ¿Cómo?

- Los mecanismos publicitarios de promoción de nuestros productos en Internet también son específicos del canal de comercialización online. Si los potenciales clientes están en internet, ¿cuáles deben ser los canales en los que lancemos nuestras campañas de márquetin? ¿Qué mecanismos de promoción están disponibles en Internet? ¿Qué canales existen? ¿Cómo podemos hacer un seguimiento del rendimiento de estos diferentes mecanismos y canales? ¿Cuáles son los más eficientes?

Para poder contestar a estas y otras preguntas, es necesario ser capaces de capturar y almacenar determinada información, que en el caso de las transacciones físicas, ni siquiera existe. En las transacciones online si queremos analizar la rentabilidad de un producto, al igual que en el mundo físico tendremos que tener en cuenta el coste del producto y los asociados al almacenaje, sin embargo tendremos que incorporar a este cálculo otros costes específicos como los costes asociados al cobro o los costes de entrega, si decidiésemos incluir estos en el precio de venta del producto.

Dada la globalidad de la venta a través de Internet, será necesario incorporar en nuestro modelo de datos, información relativa a la procedencia de las transacciones. Para tal fin será necesario que nuestro modelo contemple la capacidad para almacenar información procedente de la aplicación que se encarga de servir las páginas de nuestra tienda a los clientes (el web server), que guarda información acerca de las direcciones IP que consultaron nuestra tienda, lo que nos permite identificar los países desde donde se han realizado visitas a nuestra tienda.

Finalmente para contestar a las últimas preguntas relativas a la promoción publicitaria, necesitaremos ser capaces de almacenar información relativa a dos tipos de eventos:

- Por un lado al tráfico generado por nuestra tienda, es decir cuánta gente la visita, cuántas de las visitas se interesan por algún producto (hacen clic en un

determinado producto), cuántos de los que hacen clic acaban comprando finalmente un producto, cuántos usuarios repiten una compra, con qué frecuencia.

- Por otro, a la información relativa al funcionamiento de las campañas publicitarias en diferentes canales y medios. Información que, en la mayoría de los casos, procederá de plataformas de terceros como Ad Servers de publicidad, plataformas de campañas en motores de búsqueda, etc.

En definitiva, como acabamos de ver, construir un modelo conceptual acerca del funcionamiento de nuestro problema o negocio, nos permite identificar particularidades del mismo, que se deberán contemplar, para poder recabar la información necesaria que nos permita utilizar los datos en beneficio del negocio.

Una vez descrito el modelo de negocio, el siguiente paso en el trabajo con los datos, será la creación del modelo de datos que soporte nuestro modelo de negocio.

5.2 El Modelo de Datos

Una vez definido el modelo de negocio, la siguiente tarea a la que nos enfrentamos consiste en trasladar este modelo –algo que utiliza un lenguaje y conceptos propios del mundo empresarial– al lenguaje y los conceptos propios del mundo de la tecnología. En definitiva, necesitamos hacer entender a los técnicos o tecnólogos nuestro negocio y sus particularidades e interrelaciones. Para conseguirlo, necesitaremos construir un modelo de datos.

Un modelo de datos es una representación visual de aquellas personas, lugares, conceptos y objetos que son del interés para un negocio (o un problema en general). Estos modelos son usados para facilitar la comunicación entre el mundo de los negocios y el mundo técnico. Un modelo de datos está compuesto por símbolos que representan los conceptos que deben ser comunicados y sobre los que trabajar. El modelo de datos podría considerarse como el mapa, o el plano de los datos. A menudo se considera el modelo de datos como los planos que los arquitectos utilizan para construir las casas. Los planos de las casas consiguen plasmar un problema técnico y complejo, en un conjunto de diagramas relativamente simples, que cualquier profano puede entender. De forma similar, el modelo de datos plasma un problema complejo (un negocio o una parte de él) en un diagrama, que puede ser entendido tanto, por el personal técnico, como por el personal del ámbito empresarial o de negocio.

Del mismo modo que el propietario de una casa (o el constructor) estará completamente involucrado en el diseño de la misma, los responsables de un determinado área de negocio estarán involucrados en el diseño del modelo de datos. Se consigue así que las bases de datos y las aplicaciones que se generen realicen una operación relacionada con el negocio de forma correcta.

Si queremos construir una casa recurriremos a un arquitecto para que realice un diseño de la misma y nos proporcione los planos correspondientes. El arquitecto, antes de empezar a trabajar, debe identificar cuáles son nuestras necesidades, objetivos y expectativas. ¿Qué uso daremos a la casa? Vivienda habitual, segunda vivienda, oficinas, etc. ¿Qué tipo de vivienda estamos buscando? Una vivienda unifamiliar, una vivienda en un edificio de varias alturas, etc. ¿Cuál es nuestro presupuesto? En definitiva tendrá que identificar cuál es el alcance y los objetivos del proyecto.

Una vez que el arquitecto conoce los detalles y el alcance del proyecto, generará diferentes planos: desde los planos a más alto nivel que contemplan el diseño de la casa, hasta los de más bajo nivel que detallan la situación y medidas de los pilares de la casa, pasando por todos los planos intermedios, tales como el diseño y disposición de las diferentes estancias, hasta los de fontanería o electricidad. Podemos, por tanto identificar diferentes niveles de profundidad y detalle. De esta forma podríamos disponer de un modelo conceptual (el concepto de casa), un diseño lógico (los planos de detalle, por ejemplo, los de fontanería, electricidad y diseño de interiores) y un diseño físico (los que permitirán a los obreros la ejecución de la construcción de la vivienda). De forma similar, podemos definir tres diferentes niveles en el modelo de datos:

- **Modelo conceptual:** identifica las relaciones de más alto nivel entre las distintas entidades. Es un modelo completamente abstracto e independiente de la implementación. La Figura 27 muestra un ejemplo de Modelo Entidad Relación, en el que se modelan productos y clientes y cómo se relacionan para almacenar los pedidos de estos clientes.

Figura 27 – Esquema típico de modelo conceptual entidad relación

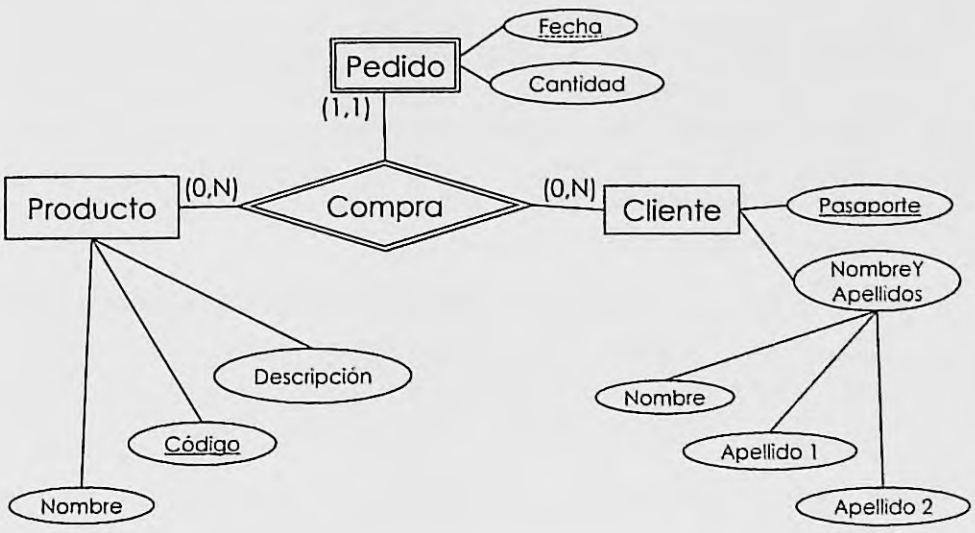


Figura 28 – Modelo de datos lógico (relacional) equivalente al modelo de datos entidad relación

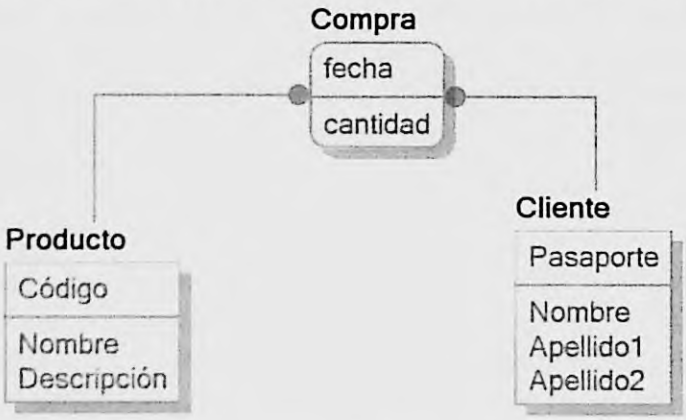
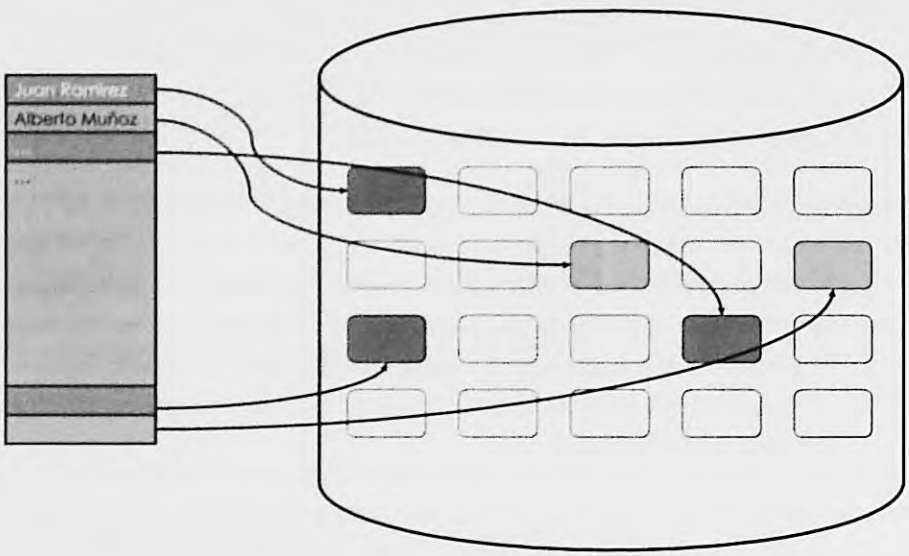


Figura 29 – Modelo de datos físico para el almacenamiento de los datos



- **Modelo lógico**: ilustra las entidades, atributos y relaciones específicas que intervienen en una función de negocio. Sirve como base para la creación del modelo de datos físico. Este es un modelo comprometido con un modelo de implementación concreto, por ejemplo, el Modelo de Datos Relacional. La Figura 28 muestra el modelo lógico (Modelo de Datos Relacional) equivalente al modelo entidad relación mostrado en la Figura 27.

- **Modelo físico**: representa una aplicación específica de un modelo de datos lógico. Las bases de datos se almacenan en ficheros del disco duro. El modelo físico describe como los datos que siguen el modelo lógico son almacenados en el disco y como acceder eficientemente a estos datos (Figura 29).

El modelo relacional y lenguaje de consultas estructurado

En la mayoría de los casos cuando nos referimos al modelo de datos, estamos haciendo referencia al modelo lógico. En el modelo lógico –basado en el modelo relacional desarrollado por E. F. Codd en el año 1970[112]– trabajaremos con entidades, atributos, relaciones, claves y referencias. Las bases de datos en las que se apoyan la mayoría de las aplicaciones de gestión son el modelo físico que surge de la aplicación del modelo lógico definido. El modelo relacional utiliza un lenguaje específico para hacer consultas a los datos almacenados conocido como lenguaje *«SQL»*, acrónimo de *«Structured Query Language»*, o lenguaje de consultas estructurado.

Pasamos a describir los elementos que podemos encontrar en un modelo relacional, como el ejemplo de modelo de datos que se ilustra en la Figura 28. Las **entidades** son la representación de elementos con identidad propia e independientes. En nuestro ejemplo serían productos (Mesa de ordenador color natural, papelera negra) o clientes concretos (Juan Ramirez, Alberto Muñoz). Estas entidades se pueden agrupar en tipos de entidades que tiene características comunes. En nuestro caso los tipos de entidades serían Producto, Compra y Cliente. Las entidades del modelo lógico se convierten a un modelo físico que define cómo se almacenarán los datos en disco. En el ejemplo de la Figura 29, vemos como para el entidad Cliente los nombres pueden almacenarse en diferentes bloques de disco.

Las entidades están descritas por **atributos**, que son las propiedades asociadas a las entidades. Por ejemplo, la entidad «Juan Ramirez» tiene como nombre «Juan» y Apellido1 «Ramirez». Al igual que para las entidades podemos tener tipos de atributos que agrupan estas características. En nuestro caso tendría que para el tipo de entidad Cliente disponemos de los tipos de atributos Pasaporte y Nombre, Apellido1 y Apellido2.

Las entidades están conectadas entre sí mediante **relaciones**. En nuestro ejemplo la relación compra nos permite conectar los productos, clientes y pedidos. Así por ejemplo, podríamos tener una relación entre Juan Ramirez, papelera negra y un pedido del 18 de febrero de 2016. Nuevamente, estas relaciones pueden agruparse en *tipos de relación*. Los *tipos de relación* pueden tener diferentes cardinalidades. Los *tipos de relación* y sus cardinalidades, representan las reglas de negocio del mundo real. Así por ejemplo, un cliente puede realizar muchos pedidos, sin embargo un pedido sólo puede haber sido realizado por un sólo cliente.

5.2.1 Modelo de Datos Multidimensional

El modelado multidimensional es posiblemente la técnica para la presentación de datos analíticos más utilizada en los sistemas de almacenes de datos actuales, debido a que tiene la capacidad de abordar simultáneamente dos requisitos fundamentales:

[112] Edgar F. Codd. «A Relational Model of Data for Large Shared Data Banks». En: *Communications of the ACM* 13.6 (1970), págs. 377-387

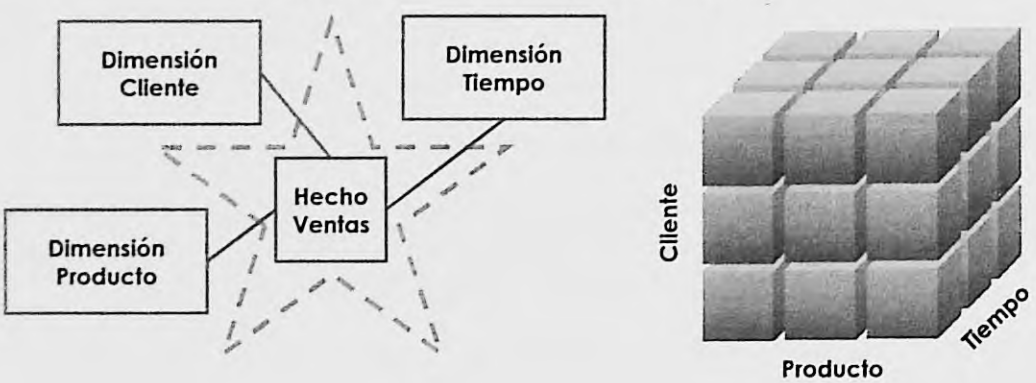
- Permite la obtención y representación de los datos de manera entendible por los usuarios finales o expertos de negocio.
- Permite la consulta rápida de los datos.

Los modelos de datos multidimensionales permiten definir el contexto de análisis de datos en términos de negocio. Estos modelos son la base para el desarrollo de almacenes de datos. Los modelos multidimensionales incluyen dos elementos clave:

- Dimensiones. Perspectivas respecto a las que queremos analizar los datos. Por ejemplo la dimensión tiempo, que nos permite definir en que ámbito temporal queremos analizar los datos. Una dimensión producto nos permitiría analizar los datos por productos, sus categorías o el fabricante de los mismos. La dimensiones son principalmente descriptivas, incluyendo normalmente muchos atributos textuales (nombres, descripciones, categorías, etc.). Puede haber tantas dimensiones como sean necesarias.
- Hechos. Los hechos definen los elementos que se quieren medir. Los hechos suelen contener pocos atributos numéricos (por ejemplo cantidad vendida, euros facturados, etc.).

Estos modelos multidimensionales suelen representarse en forma de cubos (para modelos tridimensionales), como se ilustra en la Figura 30 para un modelo multidimensional con las dimensiones Tiempo, Producto y Cliente. Estos modelos conceptuales pueden traducirse a diferentes modelos lógicos. Dado que el modelo relacional es muy conocido y existen sistemas gestores de bases de datos relacionales ampliamente utilizados en las empresas, se suele recurrir a una estructura de diseño relacional muy típica para un almacén de datos que es el «*Modelo en Estrella*» (parte izquierda de la Figura 30). Este modelo en estrella traduce un modelo multidimensional a una serie de tablas relacionales. Se trata de un patrón de diseño ya conocido desde los inicios de las bases de datos

Figura 30 – Esquema estrella versus cubo OLAP



relacionales y al que Kimball[113] dio popularidad por su uso en la mayoría de los sistemas de Business Intelligence de la actualidad.

Acercándonos a la realidad

Supongamos que queremos realizar el modelo de datos conceptual para un almacén de datos. Para ello vamos a comenzar definiendo las dimensiones de nuestro modelo:

- Dimensión Cliente: SWID, Año de Nacimiento, Ciudad, Estado, País y Género.
- Dimensión Producto/Página de Entrada: URL, Categoría
- Dimensión Tiempo: Fecha, Mes, Año.

Respecto los hechos, nos centraremos en la cantidad vendida cada día a un cliente de un producto. Este modelo conceptual podemos desarrollarlo usando la herramienta Indyco Builder[114]. La Figura 31 muestra un posible modelo multidimensional para este ejemplo.

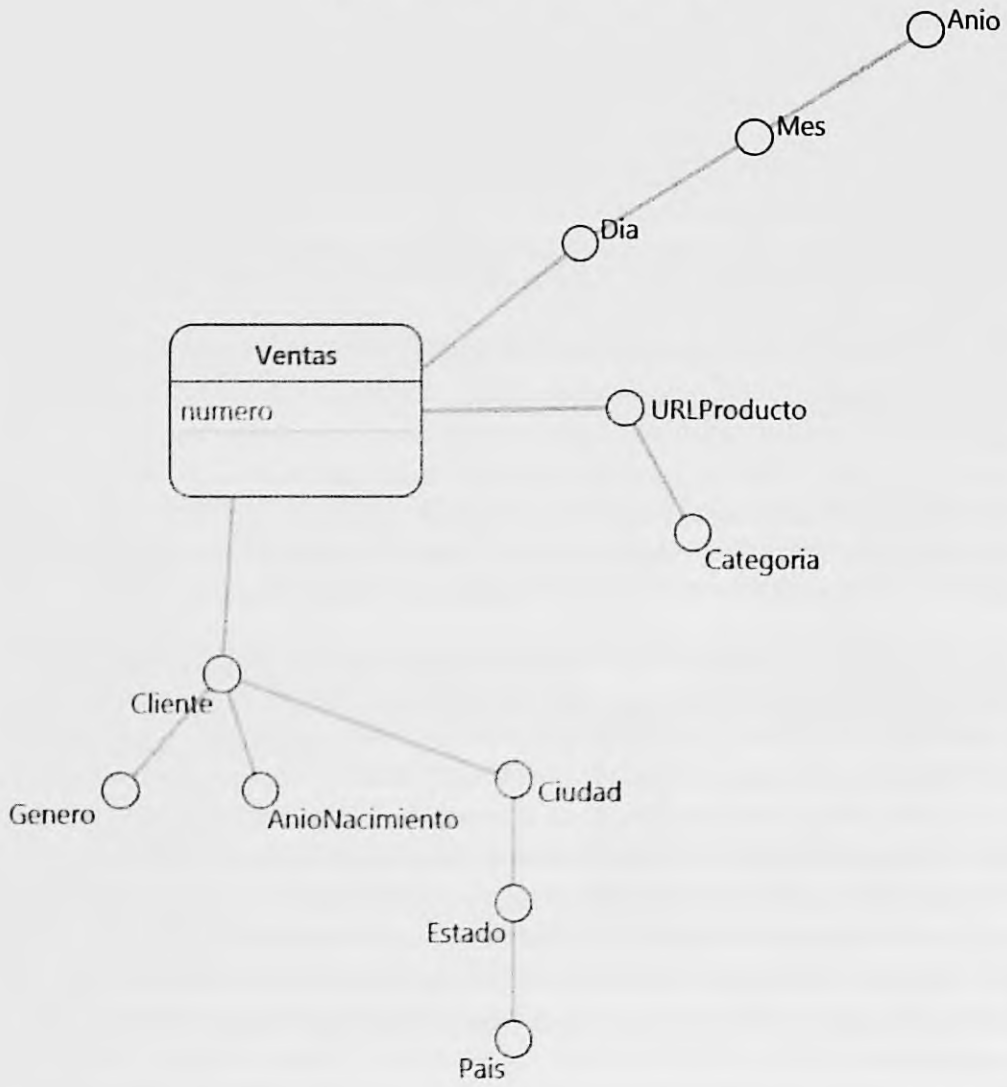
Una vez definido el modelo conceptual multidimensional, de manera independiente a la plataforma de implementación, se debe desarrollar el modelo lógico. Existen dos tendencias dependiendo de la tecnología subyacente a utilizar: el modelo multidimensional OLAP (MOLAP) y el modelo relacional OLAP (ROLAP). MOLAP se basa en el almacenamiento de las dimensiones en matrices, lo qu epermite un acceso arápido a las operaciones OLAP. Por contra, ROLAP se basa en el almacenamiento en tablas de un modelo de datos relacional. Debido a razones de eficiencia y escalabilidad (las herramientas MOLAP tradicionalmente tienen dificultades para consultar con modelos con dimensiones muy altas, del orden de millones de miembros), normalmente se hace uso del modelo relacional como guía de diseño. Sin embargo, podemos destacar las siguientes ventajas de usar MOLAP:

- Consultas rápidas debido a la optimización del rendimiento de almacenamiento, la indexación multidimensional y la memoria caché.
- Ocupa menor tamaño en disco en comparación con los datos almacenados en base de datos relacional debido a técnicas de compresión.
- Automatización del procesamiento de los datos agregados de mayor nivel.
- Muy compacto para conjuntos de datos de pocas dimensiones.
- El modelo de almacenamiento en vectores/matrices proporciona una indexación natural.

[113] Ralph Kimball. *The Data Warehouse ETL Toolkit*. Wiley, 2004

[114] http://www.indyco.com/download/

Figura 31 – Modelo Multidimensional



- Eficaz extracción de datos lograda gracias a la pre-estructuración de los datos agregados.

El entorno «Big Data/Almacén de datos/Business Intelligence» incluye tanto esquemas de estrella como cubos OLAP, y ambos tienen un diseño lógico común con dimensiones reconocibles, aunque difieran en su la implementación física.

Los cubos OLAP proporcionan una serie de optimizaciones de cálculo, indexación y rendimiento, además de funcionalidades analíticamente más robustas que aquellas disponibles con SQL. La desventaja, sin embargo, es que se paga un alto precio en términos de rendimiento como consecuencia de la inclusión de estas capacidades (se genera una gran carga computacional), sobre todo con grandes conjuntos de datos.

Acercándonos a la realidad

Cuando queremos realizar análisis complejos utilizando nuestros datos, nuestro modelo de datos puede resultar insuficiente o en cualquier caso ineficiente, por lo que necesitamos construir otros mecanismos de acceso a los datos que nos permitan contestar a las preguntas que nos hacemos.

❖ ***Pero, ¿Por qué puede resultar insuficiente o ineficiente nuestro modelo de datos?***

Una respuesta global consiste en que los datos de mediciones y cálculos, referentes a cierta perspectiva de un negocio se encuentran, en el mejor de los casos, replicados y dispersos en diferentes tablas en nuestro modelo.

Por tanto, la consulta de estas mediciones suele ser compleja y costosa produciendo una respuesta lenta. Además, es posible que no se consiga reflejar la perspectiva de nuestro negocio, por falta de relación entre algunas de estas mediciones.

Estas entidades adicionales que nos pueden ayudar a entender mejor nuestro modelo de negocio suelen ser:

- Las tablas de hechos («*fact table*»): aquello que queremos medir o analizar.
- Las tablas de dimensiones («*Dimension Table*»): cómo lo queremos medir o analizar.

5.2.1.1 Hechos

En un modelo multidimensional, la tabla de hechos almacena las mediciones de rendimiento resultantes de sucesos de los procesos empresariales de una compañía (por ejemplo la cantidad vendida de un producto en un día).

Consiste en almacenar los datos de medición de bajo nivel que resultan de un proceso de negocio en un único modelo multidimensional. Dado que los datos de medición aportan considerablemente el mayor volumen de datos, no deberían replicarse en varios lugares (tablas) para múltiples funciones de organización referentes a la actividad de la empresa, sino que deberían integrarse en un único repositorio.

De este modo, al permitir a los usuarios en múltiples y diferentes departamentos acceder a un único repositorio centralizado para cada conjunto de datos de medición (hechos), se asegura un uso consistente de los datos en el conjunto global de la empresa. El término «*hecho*» representa una medida de negocio. Imagínese que está viendo los productos que se venden y anotando la cantidad y unidad de venta por cada producto en cada transacción de venta. Cada fila en la tabla de hechos corresponde a un evento de medición.

Es importante tener en cuenta la idea de que «un evento de medición en el mundo físico tiene una relación de uno a uno con una sola fila en la tabla de hechos» correspondiente. Este es un principio fundamental para el modelado multidimensional.

Los hechos más útiles son los de naturaleza numérica y aditiva, tales como la cantidad de ventas en cierta unidad monetaria, por ejemplo, dólares, pesos o euros. En referencia a la tabla de hechos mostrada en la Figura 32[115], si no hay actividad de venta de un producto dado, no es necesario poner ninguna fila en la tabla. A pesar de su simplicidad, las tablas de hechos constituyen por lo general el 90 % o más del total de espacio consumido por un modelo multidimensional. Las tablas de hechos tienden a contener gran número de filas o tuplas, pero suelen contener un número reducido de columnas.

5.2.1.2 Dimensiones

Las tablas de dimensiones complementan las tablas de hechos. Las tablas de dimensiones contienen el contexto asociado a un evento de medición. Describen el «quién, qué, dónde, cuándo, cómo y por qué» asociado con el evento.

Como se ilustra en la Figura 32, las tabla de dimensiones a menudo tienen muchas columnas o atributos, aunque tienden a tener un menor número de filas que las tablas de hechos. Cada dimensión se define por una sola clave primaria (PK), que sirve como base para ser referenciada en la tabla de hechos a la que esté unida.

Los atributos de dimensión sirven como la fuente primaria para definir las consultas al modelo multidimensional (atributos a consultas, como agrupar los datos, etc.). En una solicitud de consulta o informe, los atributos son identificados por las palabras. Por ejemplo, cuando un usuario quiere ver las ventas en dólares por parte de una marca en concreto, la marca debe estar disponible como un atributo de dimensión.

Todas las tablas de hechos tienen dos o más claves externas (véase la Figura 32) que se conectan a las claves principales de las tablas de dimensiones. Por ejemplo, la clave de producto (id_Producto) en la tabla de hechos Ventas coincide con una clave de producto específica en la tabla de la dimensión Producto. Cuando todas las claves en la tabla de hechos coinciden con sus respectivas claves primarias en las tablas de dimensión, las tablas satisfacen la restricción de integridad referencial. Por lo general, hay un conjunto de dimensiones que identifican de manera única cada fila en la tabla de hechos.

Los atributos de las tablas de dimensiones juegan un papel vital en un sistema Big Data/Almacén de datos/BI, debido a que son la fuente de prácticamente todas las restricciones y también de las secciones en los informes generados. Los atributos deben ser palabras reales en lugar de abreviaturas crípticas o códigos alfanuméricos.

[115] Ralph Kimball. *The Data Warehouse ETL Toolkit*. Wiley, 2004

5.2.1.3 Esquema Estrella

La característica principal de un esquema multidimensional en estrella es su sencillez y simetría (véase el ejemplo de la Figura 32). Obviamente, los usuarios de negocios se benefician de esta simplicidad porque los datos son más fáciles de entender y navegar.

Se pueden observar cientos de casos en la industria en los que los usuarios asumen rápidamente el modelo multidimensional como el más apropiado para su negocio. Además, la reducción del número de tablas y el uso de los descriptores significativos de negocio hacen que sea más fácil la navegación entre dichas tablas y es menos probable que se produzcan errores. La simplicidad de un modelo multidimensional también tiene ventajas de rendimiento.

Los modelos multidimensionales son fácilmente extensibles para adaptarse a los cambios. El marco predecible de un modelo multidimensional resiste a cambios inesperados en el comportamiento del usuario. Todas las dimensiones son puntos de entrada simétricamente iguales en la tabla de hechos.

Otra forma de entender la naturaleza complementaria de las tablas de hechos y dimensiones es verlas desde el punto de vista de un «informe». Como se ilustra en la Figura 33,

Figura 32 – Tablas de hechos (Ventas) y dimensiones (Producto, Cliente, Almacén, Tiempo y Promoción) en un esquema de estrella

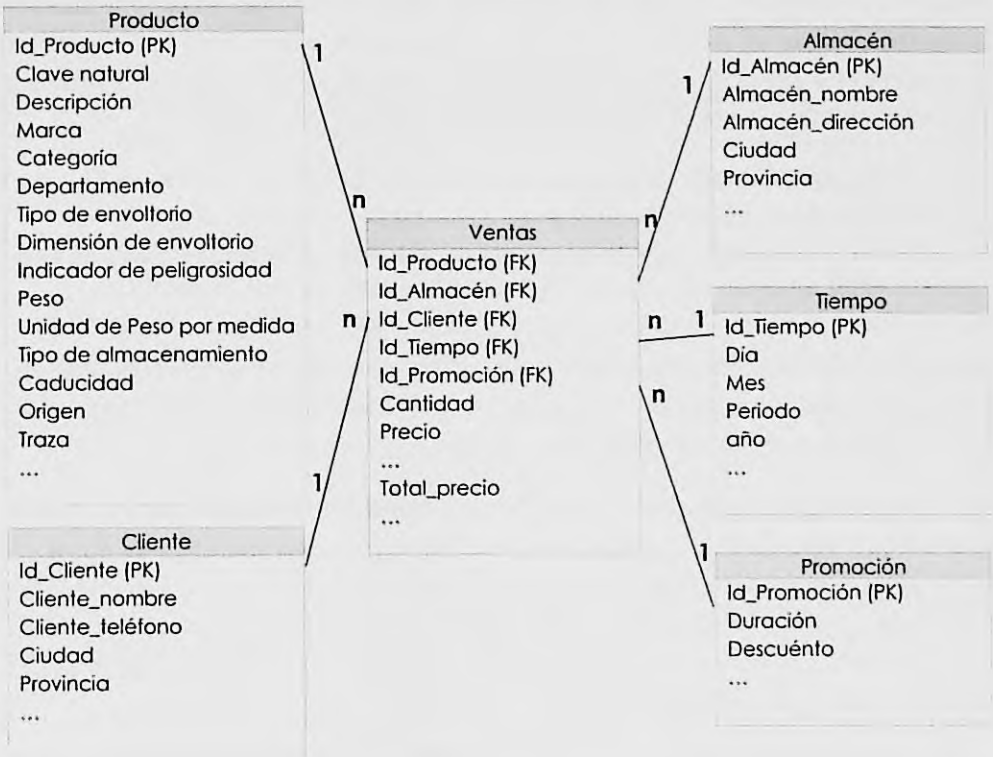
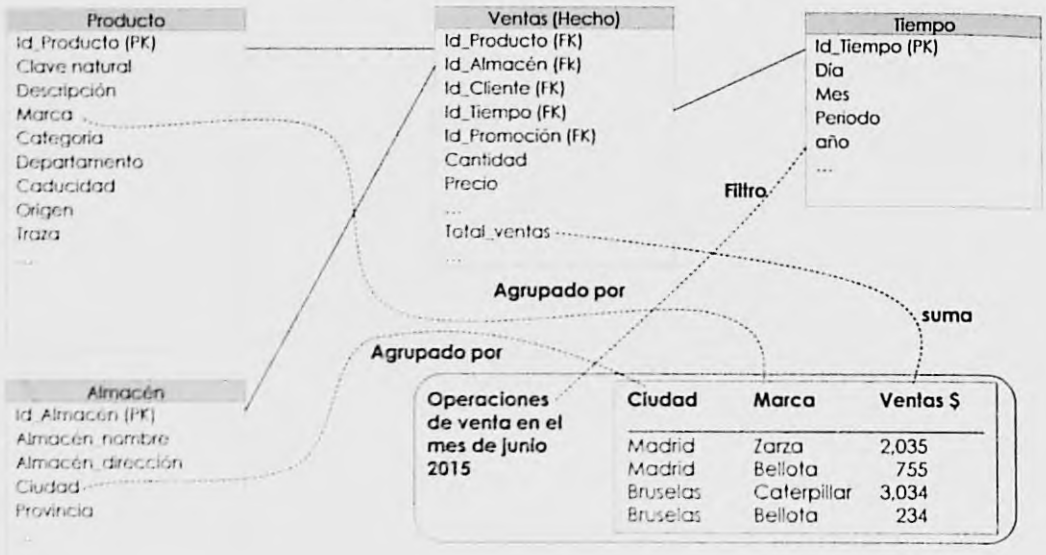


Figura 33 – Atributos de dimensión y hechos en formato de informe



los atributos de dimensión actúan como filtros de informes y etiquetado, mientras que las tablas de hechos suministran los valores numéricos y calculados.

5.3 Modelo Físico: Desarrollo y Carga

Tal y como explica Ralph Kimball en su libro «The Datawarehouse ETL Toolkit»[116], los sistemas ETL construyen o «cargan» un almacén de datos con información útil y de calidad. La construcción de un sistema de este tipo es una actividad que no es visible para los usuarios finales, pero que fácilmente consume el 70 % de las necesidades de recursos para el desarrollo y mantenimiento de un sistema de gestión de datos.

Acercándonos a la realidad

- ***Los procesos de carga ¿son un mero traspaso de información de un sistema a otro?***

Son mucho más, pues ayudan a validar y acomodar los datos en el modelo. Unos procesos de carga mal definidos o mal validados, pueden hacer inutilizable un sistema impecablemente diseñado, aunque mal alimentado por unos procesos mal construidos.

Una vez completados los procesos de extracción y transformación de los datos, la etapa final en todo proceso ETL es, por tanto, la **estructuración física y la carga de datos** en los modelos multidimensionales de destino para su presentación en vistas o gráficas de negocio. La misión principal de los sistemas ETL es la de acomodar la información, previamente extraída y transformada en las tablas de dimensiones y de hechos, siguiendo el modelo multidimensional subyacente.

[116] Ralph Kimball. *The Data Warehouse ETL Toolkit*. Wiley, 2004

En este sentido, las tablas de hechos son típicamente grandes y conllevan mucho tiempo de carga, pero la preparación para su exposición en paneles de visualización es normalmente sencilla. Cuando las tablas de dimensiones y de hechos en un modelo multidimensional se actualizan, entonces se realiza la indexación (o indización), se cargan con datos agregados apropiados y se asegura la calidad. Finalmente, se notificará mediante un mensaje del sistema a los usuarios expertos en inteligencia del negocio sobre dicha publicación/actualización de nuevos datos.

5.3.1 Fase de Carga

La fase de carga es por tanto el momento en el cual los datos transformados son cargados en el sistema de destino. Consiste en mover los datos desde las fuentes operacionales, o el almacenamiento intermedio, hasta el almacén de datos para su carga en las tablas correspondientes.

El proceso de carga puede consumir mucho tiempo cuando se maneja gran cantidad de datos (como es el caso en entornos Big Data). Por eso es necesario automatizar todos los procesos involucrados en la carga de los almacenes de datos.

Por lo general, la primera carga de datos es la que requiere el manejo de un mayor volumen de datos, por lo que se planteará para aquellos momentos en los que los picos de trabajo en general sea menor en las fuentes de datos, intentando así aliviar el sistema de carga y compatibilizar tareas. Se pueden diferenciar por tanto dos etapas en el proceso de carga:

- **Carga inicial.** Que conllevará **datos históricos** ya almacenados en las fuentes de datos. Dado que se necesita el almacenamiento de la historia de los datos, se deberá planear una carga masiva de estos. Esta primera carga se puede subdividir a su vez en bloques de datos, aliviando así el sistema de BI en general.
- **Mantenimiento periódico.** Las subsiguientes cargas del almacén de datos o «refrescos» se realizarán siguiendo el propio ciclo de negocio. Es importante detectar los periodos de menos actividad en las fuentes de datos para realizar la carga en estos momentos.

Una vez en fase de mantenimiento periódico, existen 2 formas de desarrollar la carga:

- TAL («trunc and load»): limpia el repositorio de datos y carga de nuevo toda la información con el nuevo contenido.
- **Incremental:** se utiliza cuando únicamente se carga información nueva o información que necesita ser actualizada.

La frecuencia del mantenimiento periódico (o carga) está determinada por la cantidad de datos y los requisitos de los usuarios. Se deben determinar las «ventanas de carga» más convenientes para no saturar la base de datos. Ocasionalmente se puede considerar el archivo o incluso la eliminación de datos obsoletos que ya no interesan para el análisis.

A continuación, enunciamos los conceptos y fases fundamentales de la carga. No obstante, una descripción detallada de la secuencia de pasos de un proceso de carga típico se puede encontrar en el libro de *Trujillo, Mazón y Pardillo*, «*Diseño y Explotación de Almacenes de Datos*»[117].

Ventana de carga

Para establecer una ventana de carga es necesario evaluar el tiempo disponible para todo el proceso ETL. Es decir, se requiere planificar, comprobar y monitorizar la carga de trabajo de las fuentes de origen. Un caso típico es planificar la ventana de carga para los periodos de menor actividad del usuario (noche, madrugada).

Por tanto, es necesario definir una estrategia de carga en función del volumen de datos, la infraestructura técnica disponible, la novedad de los datos o la frecuencia de refresco, y los requisitos propios de usuario. Estos factores influirán en el tamaño, en términos de duración, de la ventana de carga.

Indexación

En cuanto al proceso de indexación en el almacén de datos, una buena práctica es la de calcular los índices después de la carga ya que se agiliza la generación de éstos respecto a si se hace durante la carga, pues se debe añadir un tiempo adicional a la ventana de carga. En el caso de generar índices únicos, se deben gestionar las restricciones de integridad del almacén de datos, es decir, deshabilitarlas antes de la carga del almacén y volverlas a activar antes de la creación de índices. Es decir, la comprobación de las relaciones existentes entre dimensiones y hechos (por ejemplo los productos incluidos en una venta) puede desactivarse durante el proceso de carga, y realizar esta comprobación al finalizar el proceso.

Integridad de los datos

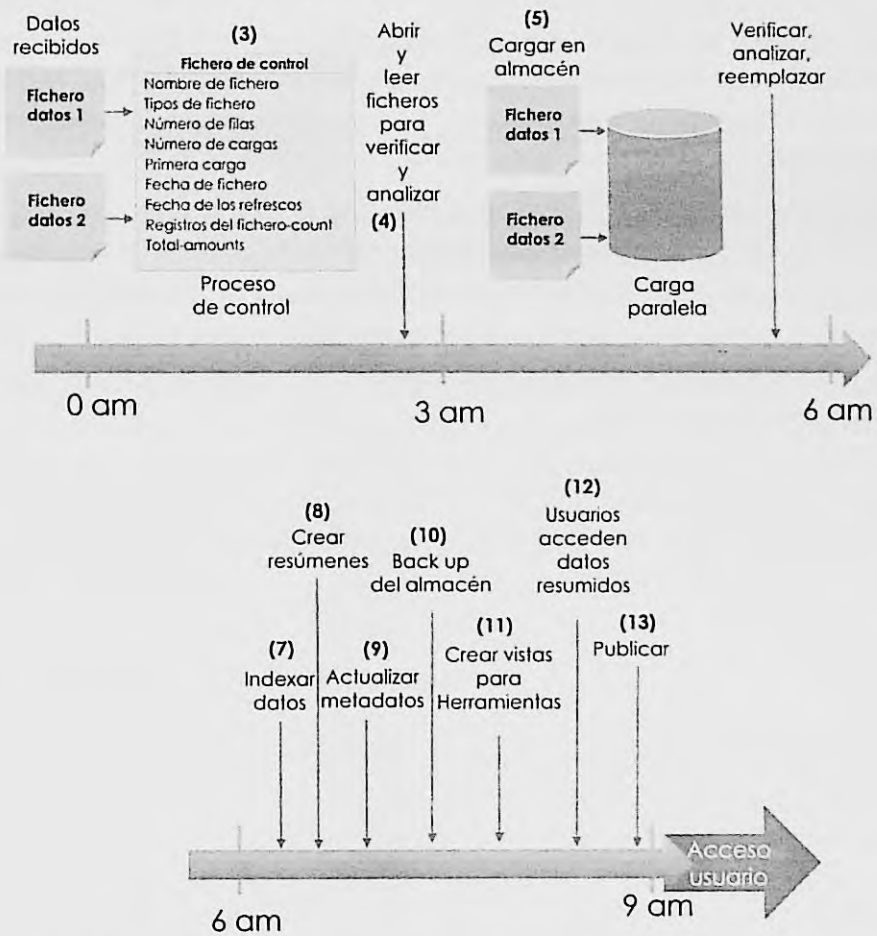
Una vez generadas las claves, sólo queda comprobar la integridad de los datos cargados. Una primera tarea de verificación consiste en cargar los datos en un fichero o tabla intermedia y comprobar los totales del almacén de datos con los totales antes de la carga. Tras esto, se debe comprobar también que no hayan surgido errores de carga (volcados en los ficheros «`log`»), que el proceso haya terminado satisfactoriamente (normalmente se muestran mensajes de éxito) y considerar también aspectos de seguridad en cuando a accesos no deseados durante la carga.

[117] J. C. Trujillo, N. Mazón y J. Pardillo. *Diseño y explotación de almacenes de datos. Conceptos básicos de modelado multidimensional.* Editorial Club Universitario, 2010

En este sentido, el proceso de carga puede durar hasta 24 horas para finalmente disponer los datos para acceso por parte de los usuarios finales. Por ello, es necesario llegar a un compromiso entre este tiempo de carga y el acceso de usuarios, que pasa normalmente por utilizar tablas temporales de datos parciales para agilizar la frecuencia de la vista o refresco de datos en los paneles de visualización de cara a los usuarios.

En resumen, siguiendo los pasos del ejemplo, tenemos que en primer lugar (1) se obtienen los requisitos de usuario y a continuación se diseña el ciclo de carga (2) en base a estos. Como se puede ver en la Figura 34, el siguiente paso (3) consiste en actualizar un fichero de control con los datos de la carga a realizar. Entonces, comienza el proceso de control (4) y se realiza la carga en el almacén de datos (5). Este paso se realizaría a las 3 de la mañana hora nocturna, tras el cual se pasa a verificar, analizar y reemplazar los datos cargados (6).

Figura 34 – Ejemplo de proceso de carga. Ejemplo tomado del libro de *Trujillo, Mazón y Pardillo, «Diseño y Explotación de Almacenes de Datos»*[118]



[118] J. C. Trujillo, N. Mazón y J. Pardillo. *Diseño y explotación de almacenes de datos. Conceptos básicos de modelado multidimensional*. Editorial Club Universitario, 2010

A las 6 am, la indexación de datos comienza (7), se crean resúmenes (8) y se actualizan los metadatos (9). Tras ello, se realizan las copias de seguridad del almacén de datos (10) y se crean las vistas necesarias para las herramientas de BI especializadas. De esta forma, los usuarios pueden ya acceder a los datos resumidos (12), publicarlos (13) y a partir de las 9 am, comenzar el acceso típico al sistema de almacén de datos.

Acercándonos a la realidad

En este punto, es importante señalar que la carga de datos provenientes del Big Data en un modelo multidimensional no varía en esencia respecto a las técnicas clásicas en los procesos ETL de almacenes de datos (*data warehouse*), ya que las mayores diferencias tienen lugar en las etapas de extracción y transformación.

- ***¿Por qué las mayores diferencias tienen lugar en los procesos ETL en almacenes de datos y Big Data se realiza en las etapas de extracción y transformación?***

El hecho de contar con datos de naturaleza estructurada en el contexto de almacenes de datos y con datos no estructurados, por lo general, en entornos Big Data, hace que los procesos de extracción y transformación (con las implicaciones de almacenamiento, acceso, obtención de datos, profiling, curado, etc.) sean diferentes en ambos casos.

En el Capítulo 11 se presenta un caso de uso en el que se aplican las técnicas descritas en este capítulo. Como parte de dicho dicho Capítulo 11 se muestra el diseño de modelos de datos en un caso de uso real (el análisis del «*Clickstream*» o «Huella Digital» en e-commerce). El Clickstream es el proceso de recogida, almacenamiento y análisis de las direcciones o enlaces Web sobre las que un usuario o visitante hace click y accede a lo largo de su visita a un sitio Web.

6 Almacenamiento de Big Data

6.1 Introducción al Almacenamiento Masivo de Datos

Una de las primeras cuestiones que surgen a la hora de trabajar con Big Data es: ¿dónde almacenamos los datos? El almacenamiento de la información es el primer paso antes de poder comenzar a procesarla o a realizar una analítica de los datos.

En la actualidad, son muy numerosas las fuentes y orígenes de datos que están generando grandes volúmenes de datos de tipos muy diversos de una forma continuada. Evidentemente, el hecho de contar con una cantidad tan grande de datos introduce nuevos desafíos a la hora de almacenarlos y posteriormente procesarlos, ya que se puede disponer de un gran volumen de datos que exceda el tamaño de un disco duro de un ordenador o que llegue a una velocidad mayor que la velocidad de escritura de uno de estos dispositivos de almacenamiento.

6.2 Sistemas de Ficheros Distribuidos

Dadas algunas de las características del Big Data, que incluyen el gran volumen de datos y la elevada velocidad a la que se producen; los sistemas de ficheros diseñados para el almacenamiento de Big Data están sujetos a nuevos desafíos, que los sistemas de ficheros clásicos no tenían que afrontar. El principal desafío es que la mayoría de sistemas de ficheros clásicos funcionan sobre un único soporte físico, y esto constituye una limitación importante en lo que respecta a las dos uves anteriores: no habrá espacio suficiente para almacenar todos los datos y además el disco no dispondrá de la velocidad suficiente para poder almacenar todos los datos según llegan.

6.2.1 El Enfoque de Google: GFS

Como ya comentamos anteriormente, los sitios web son una de las principales fuentes de Big Data. Precisamente, los motores de búsqueda son una de las aplicaciones que tienen que lidiar con este tipo de datos, con el fin de indexar todos los sitios web para poder atender las peticiones de búsqueda de sus usuarios.

Google es el motor de búsqueda más utilizado en Internet, además de ser la página web más visitada del mundo. El producto surgió en 1997 como parte de la tesis doctoral de Larry Page y Sergey Brin, dos estudiantes de doctorado de Stanford. La compañía Google Inc. se fundó en 1998, y desde entonces el número de servicios que ha venido prestando a sus usuarios ha crecido de un modo imparable.

En el año 2000, Google da a luz a su producto para anunciantes AdWords y a la barra de herramientas de Google (Google Toolbar) para el navegador, lo que permite a los usuarios realizar búsquedas de un modo más cómodo. En 2001 inicia su servicio de búsqueda de imágenes, y en 2002 añade el servicio de noticias Google News, el de búsqueda de productos Google Product Search y herramientas para desarrolladores.

En 2003, el abanico de servicios de Google sigue creciendo al introducir su servicio de publicidad dirigida AdSense, el sistema de libros online Google Books y la plataforma Blogger para el alojamiento de blogs. Es evidente que llegados a este punto, las necesidades de almacenamiento de datos de Google han crecido de un modo espectacular, y es entonces cuando Google publica en una conferencia científica el diseño del sistema de ficheros distribuido que emplea la empresa para poder hacer frente a estas necesidades: es el Google File System o GFS[119].

Este sistema de ficheros presenta como principal característica que es distribuido, es decir, funciona sobre varios ordenadores, cada uno de ellos con sus propios discos duros. Los ficheros se almacenan repartidos entre estos discos duros, lo cual aumenta la capacidad de almacenamiento (al haber más discos se pueden almacenar más datos) y además también incrementa la velocidad de lectura y de escritura (al haber varios discos, se pueden acceder a varios de ellos a la vez).

6.2.2 El Nacimiento de Hadoop

En 2002, Doug Cutting y Mike Cafarella estaban trabajando en Nutch, un proyecto para lanzar un motor de búsqueda de código abierto. Durante el desarrollo, ambos ingenieros repararon en que necesitaban algún mecanismo para poder escalar la cantidad de datos que almacenaban. Cuando Google liberó en 2003 la especificación de GFS; los desarrolladores de Nutch aprovecharon para integrarlo en su buscador. Para ello, implementaron una versión de código abierto de GFS que denominaron NDFS, que estuvo lista en 2004.

[119] Sanjay Ghemawat, Howard Gobioff y Shun-Tak Leung. «The Google File System». En: *19th ACM Symposium on Operating Systems Principles*. 2003, págs. 29-43

En 2005, Doug y Mike se dieron cuenta de que NDFS podría utilizarse en aplicaciones distintas a Nutch. Por ello, separaron el proyecto NDFS y, junto con MapReduce (una herramienta que veremos en el siguiente capítulo), crearon Hadoop[120], y renombraron el sistema de ficheros a HDFS (*Hadoop Distributed File System*).

En 2006 Doug Cutting se encontraba trabajando en Yahoo! Inc, por lo que migró la infraestructura del buscador de Yahoo! a Hadoop, pudiendo así hacer frente al importante crecimiento del volumen de datos. Desde ese momento, Yahoo! se convirtió en uno de los principales contribuyentes al proyecto Hadoop, que en 2008 se convierte en proyecto de primer nivel de Apache, lo que garantiza el compromiso de la comunidad de desarrolladores con el mismo.

En la actualidad, más de la mitad de las empresas del Fortune 50 emplean Hadoop en su infraestructura[121]. La aplicación con más datos soportada por Hadoop es posiblemente Facebook, que en 2012 reveló que empleaba un sistema de ficheros HDFS con más de 100 PB (petabytes, miles de terabytes) de datos[122], creciendo a un ritmo aproximado de 0,5 PB por día.

6.2.3 Propiedades de HDFS

Como se comentó anteriormente, el sistema de ficheros distribuido de Hadoop (HDFS, *Hadoop Distributed File System*) es una implementación de código abierto del sistema GFS que presentó Google en 2003. Por este motivo, ambos sistemas comparten las mismas características.

Tanto GFS como HDFS son sistemas de ficheros que están enfocados al almacenamiento de Big Data, lo que implica que están optimizados para almacenar ficheros muy grandes. Por esta razón, mientras que los sistemas de ficheros clásicos tienen tamaños de bloque de entre 4 KB y 128 KB, GFS y HDFS están diseñados para trabajar con bloques mucho mayores, del orden de 64 MB o 128 MB (aunque esta cantidad es configurable, como veremos más adelante). Por esta razón, estos sistemas de ficheros no son especialmente eficientes cuando se trabaja con ficheros muy pequeños, sino que despliegan su potencial al emplear ficheros más grandes.

---

[120] Se dice que Doug Cutting decidió llamar así al proyecto basándose en el nombre de un juguete de un elefante amarillo que tenía su hijo, lo que también habría dado lugar al logotipo de Hadoop.

[121] PRNewswire. *Altior's AltraSTAR - Hadoop Storage Accelerator and Optimizer Now Certified on CDH4 (Cloudera's Distribution Including Apache Hadoop Version 4)*. http://www.prnewswire.com/news-releases/altiors-altrastar---hadoop-storage-accelerator-and-optimizer-now-certified-on-cdh4-clouderas-distribution-including-apache-hadoop-version-4-183906141.html. [Online; publicado el 18 de diciembre de 2012]

[122] Andrew Ryan. *Under the Hood: Hadoop Distributed Filesystem reliability with Namenode and Avatarnode*. https://www.facebook.com/notes/facebook-engineering/under-the-hood-hadoop-distributed-filesystem-reliability-with-namenode-and-avata/10150888759153920. [Online; publicado el 13 de junio de 2012]

No obstante, la principal propiedad de HDFS viene descrita en su nombre: el sistema es **distribuido**. Esto quiere decir que los ficheros se almacenarán repartidos entre varias máquinas, cada una con uno o más discos duros. Cada una de estas máquinas recibe el nombre de nodo, mientras que un conjunto de varios nodos conectados entre sí recibe el nombre de cluster de Hadoop.

El hecho de que el sistema de ficheros sea distribuido presenta varias ventajas adicionales. La principal ventaja es que proporciona escalabilidad horizontal. Esto significa que el *cluster* es capaz de mejorar su rendimiento y capacidad simplemente añadiendo nuevos nodos. Esta propiedad contrasta con la escalabilidad vertical de los enfoques clásicos, en los que un sistema solo podía mejorar su rendimiento si se actualizaban sus componentes (por ejemplo, se añadían discos de más capacidad, procesadores más rápidos, etc.). La principal ventaja de la escalabilidad horizontal radica en su coste: se pueden añadir nuevos equipos relativamente económicos y la capacidad del *cluster* seguirá creciendo.

En el caso de HDFS, añadir nuevos equipos implica aumentar la capacidad de almacenamiento del *cluster*; es decir, si se duplica el número de nodos, entonces la cantidad de datos que este puede almacenar también se duplica[123]. Por otro lado, la incorporación de nuevos nodos también incorpora otra ventaja: cuando se lean y escriban ficheros, estas operaciones se pueden hacer de forma paralela, es decir, varios ordenadores podrán leer y escribir datos al mismo tiempo. Esto reducirá los tiempos de lectura y escritura, pudiendo hacer frente a mayores velocidades de entrada y salida de datos.

Aún más interesante es el hecho de que la distribución se realiza a nivel de bloque, no de fichero. ¿Qué significa esto? Si almacenamos un fichero que ocupa por ejemplo 10 bloques, estos bloques se almacenarán en nodos distintos. Esto tiene la ventaja de que se pueden almacenar ficheros que sean más grandes que cualquiera de los discos duros que tengamos, lo que es algo impensable en los sistemas de ficheros tradicionales. Además, también podemos beneficiarnos de la mejora de la velocidad que señalábamos antes incluso cuando solo estamos leyendo un único fichero.

- ***Hasta el momento hemos señalado algunas de las ventajas del almacenamiento distribuido. ¿Se le ocurre algún inconveniente que tenga asociado este sistema? En caso afirmativo, ¿puede plantear alguna solución para eliminar o al menos aliviar estos inconvenientes?***

No obstante, el almacenamiento distribuido también acarrea algunos problemas que deben ser estudiados y solventados. Uno de estos problemas es la tolerancia a fallos: normalmente la tasa de fallos de un nodo es muy baja, sin embargo, al contar con un número muy elevado de nodos, las probabilidades de que uno de ellos falle crece sustancialmente. Por ejemplo, si la probabilidad de que falle un nodo en un año es de una entre mil

[123] En realidad, esta es una afirmación aproximada. En principio, esto solo ocurriría si todos los nodos tienen iguales características. Además, el rendimiento del *cluster* no crece linealmente con su tamaño, sino algo más despacio debido al coste que implica gestionar más nodos.

(1‰), lo normal sería que este fallo no ocurriera si solo tenemos una máquina. Sin embargo, si contamos con un *cluster* con 1.000 nodos de idénticas características, la estadística indica que es probable que uno de estos mil nodos fallen.

El problema de que un nodo falle es que puede perderse el acceso a los datos que tiene almacenados de forma temporal, o incluso estos pueden perderse de forma irrecuperable. Evidentemente, este es un escenario que se debe evitar. Tanto GFS como HDFS proporcionan una solución a este problema, que consiste en la **replicación** de los bloques. Esto quiere decir que cada bloque está almacenado de manera redundante en más de un nodo del *cluster*. Cada uno de estas apariciones del bloque se denominan **réplicas**.

En caso de que un nodo del *cluster* deje de funcionar (o incluso quede inservible), aún se podría acceder a los datos que estaban almacenados en él, ya que habrá réplicas guardadas en otros nodos. Por defecto, HDFS utiliza un **factor de replicación** de 3, lo que significa que para cada bloque se almacenan 3 réplicas en nodos distintos. Además, este factor de replicación fuerza a que este número de réplicas esté disponible para cada bloque en todo momento. Es decir, si un nodo falla, los bloques que él contenía deben ser replicados de nuevo en otros nodos para garantizar que se cumple el factor de replicación.

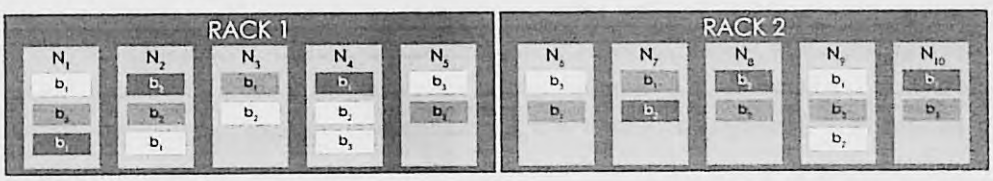
Por último, el sistema de replicación de HDFS se puede configurar para que tenga en cuenta la **topología** del *cluster*, es decir, la forma en la que los equipos están conectados entre sí. En el caso de que se defina una topología en Hadoop, HDFS tratará de almacenar dos réplicas en el mismo *rack*, mientras que la tercera réplica se almacenará en un *rack* diferente o, si hay posibilidad, incluso en otro centro de datos. Esto permitiría que, en caso de quedar sin disponibilidad todo un *rack* o centro de datos (por fallos de corriente, incendios u otras causas de fuerza mayor), aún se podría seguir accediendo a los datos.

La Figura 35 sintetiza de forma gráfica las propiedades descritas en esta sección. En ella se muestran ficheros distintos con intensidades distintas de color, y para cada uno de los bloques, cómo estos se distribuyen entre diferentes nodos. Se puede observar que cada bloque cuenta con tres réplicas en nodos distintos, teniendo en cuenta además la topología del *cluster*.

6.3 Tecnologías de Bases de Datos No Relacionales

Anteriormente ya vimos el modelo de base de datos relacional, que emplea el lenguaje SQL (del inglés *Structured Query Language*). A continuación vamos a ver tecnologías al-

Figura 35 – Ejemplo de la distribución de los bloques de los ficheros entre nodos en HDFS



ternativas a las bases de datos relacionales, mencionando algunas de las herramientas más relevantes o extendidas en cada una de estas tecnologías.

6.3.1 Introducción a las Bases de Datos No Relacionales

Las tecnologías alternativas a las bases de datos relacionales han recibido recientemente el nombre de NoSQL, del inglés *not only SQL* (no solo SQL). La mayoría de herramientas NoSQL emplean lenguajes distintos de SQL para realizar consultas, aunque existen algunas implementaciones que sí utilizan SQL.

Ahora bien, ¿cuáles son las principales diferencias entre las bases de datos SQL y las NoSQL? Evidentemente, las diferencias dependerán de las herramientas concretas a comparar, si bien podemos establecer algunas que se darán en casi todos los casos:

- Las bases de datos relacionales están diseñadas para almacenar datos estructurados, mientras que las no relacionales pueden almacenar también datos semiestructurados.
- Las bases de datos relacionales están ideadas para ser transaccionales (OLTP), garantizando los principios de ACID (atomicidad, consistencia, aislamiento y durabilidad); mientras que las bases de datos no relacionales no suelen ofrecer estas garantías.
- Las bases de datos relacionales permiten combinar relaciones empleando la sentencia JOIN de SQL, algo que la mayoría de bases de datos NoSQL no permiten.
- Las bases de datos relacionales no suelen escalar bien a la hora de almacenar grandes cantidades de datos (y no suelen soportar escalabilidad horizontal), mientras que muchas bases de datos NoSQL sí presentan esta escalabilidad.

A continuación se presentan con más detalle los principales modelos de bases de datos no relacionales: el modelo clave-valor, el documental, el orientado a grafos y el orientado a columnas.

6.3.2 Bases de Datos Clave-Valor

La premisa de las bases de datos clave-valor es extremadamente sencilla: los datos son valores identificados por una clave. Las claves no pueden repetirse, y a cada una de ellas le corresponde un valor.

Evidentemente, este modelo de datos no es adecuado para resolver todos los problemas. Sus principales desventajas es que no soporta JOIN ni claves ajenas, y normalmente la única consulta permitida es acceder a un valor usando su clave, lo que se realiza de forma muy eficiente.

La principal ventaja de este enfoque, además de su simplicidad, es que permite realizar una distribución de los datos fácilmente: basta con almacenar los pares clave-valor en nodos distintos, obteniendo así escalabilidad horizontal. Además, también suelen soportar redundancia de datos (que, como vimos anteriormente, es una funcionalidad importante cuando los datos están distribuidos), lo que se puede conseguir simplemente almacenando cada par clave-valor varias veces.

Una de las herramientas más conocidas que implementan el paradigma clave-valor es Project Voldemort[124], que a continuación presentaremos con más detalle. Otra opción extendida es Redis[125], que presenta una gran flexibilidad en las claves que puede almacenar (pueden ser estructuras de datos como listas, conjuntos, etc.), pero funciona cargando todos los datos en memoria principal. Por esa razón, el acceso a los datos es mucho más rápido, pero la capacidad de almacenamiento es más limitada y además introduce algunos desafíos adicionales, lo que lo convierte en una opción poco adecuada para almacenar Big Data.

6.3.2.1 Voldemort

Project Voldemort es una base de datos clave-valor de código abierto. Esta base de datos es empleada por LinkedIn para el almacenamiento de muchos de los servicios críticos que proporciona.

Su funcionamiento es extremadamente sencillo, aunque al mismo tiempo posee algunas características que lo colocan en una buena posición para ser un almacén de Big Data.

Escalabilidad. En Voldemort, los datos pueden estar distribuidos entre diferentes nodos, de forma similar a como funciona HDFS. Además, para garantizar la disponibilidad, se realiza de forma automática una replicación de los datos en varios nodos, de tal modo que si uno de ellos deja de funcionar no se pierda información. El proceso de distribución, replicación y gestión de fallos es completamente transparente para el usuario y las aplicaciones; y se pueden configurar políticas para la distribución de los datos, similar al concepto de topología que ya mencionamos al hablar de HDFS. Para decidir los equipos donde se almacenarán cada una de los valores, se aplicará una determinada función *hash* a la clave. Los nodos son independientes entre sí, y no existe ningún servicio que constituya un cuello de botella.

Rendimiento. Realizar una búsqueda por valor o insertar un nuevo valor es una operación muy eficiente. Normalmente, en los sistemas relacionales el tiempo de operación se ve incrementado por dos razones fundamentales: el uso de JOIN para combinar dos o más relaciones y las garantías de transaccionalidad ACID. En Voldemort, los desarrolla-

[124] Project Voldemort. *Voldemort - a Distributed Database*. http://www.project-voldemort.com/voldemort/. [Online; consultado el 30 de mayo de 2015]

[125] Redis. *Redis*. http://redis.io. [Online; consultado el 31 de mayo de 2015]

dores aseguran que se pueden esperar entre 10 y 20 mil operaciones por segundo, en función de las características de los nodos. Además, Voldemort hace *caching*, es decir, carga algunos datos que se consultan frecuentemente en memoria principal para agilizar su consulta.

Operaciones. Las operaciones soportadas por Voldemort son muy limitadas, en concreto, son las siguientes:

- **Recuperar un valor**, empleando el comando `value = store.get(key)`.
- **Introducir o actualizar un valor**, con el comando `store.put(key, value)`.
- **Borrar un par valor**, mediante la instrucción `store.delete(key)`.

A pesar de la simplicidad de estas operaciones, las claves y los valores pueden llegar a ser muy complejos. Por ejemplo, ambos pueden ser estructuras de datos como listas o mapas, o documentos JSON o XML.

6.3.3 Bases de Datos Documentales

Otro paradigma en las bases de datos no relacionales es el orientado a documentos, o documental. En estos sistemas de bases de datos cada registro es un documento que normalmente no tiene que estar sujeto a una estructura fija. Se trata por tanto de un enfoque adecuado para almacenar datos semiestructurados. De hecho, muchas bases de datos orientadas a documentos son en realidad almacenes de ficheros XML o JSON, como los que vimos al hablar de datos semiestructurados.

La principal diferencia con las bases de datos relacionales es que no es necesario definir a priori la estructura de los datos. De hecho, cada documento puede tener una estructura diferente, y es frecuente que esta estructura cambie con el tiempo. Esta característica perfila este tipo de bases de datos NoSQL como una elección adecuada para aplicaciones en las que el modelo de datos está en constante evolución.

❖ ***Cuando hablamos de las bases de datos clave-valor vimos que Voldemort puede almacenar documentos JSON como valor. ¿Cuál cree que es la diferencia entre una base de datos clave-valor y una orientada a documentos?***

La diferencia fundamental entre una base de datos clave-valor y una base de datos documental es que en las primeras, tanto la clave como el valor son completamente opacos para el sistema de bases de datos, y es el cliente o la aplicación la que tiene que saber interpretarlos. Es decir, en una base de datos clave-valor, incluso cuando un valor sea un documento JSON, el sistema de base de datos no entenderá qué campos lo conforman ni su significado, simplemente se limitará a devolver ese valor cuando se le solicite. Sin

embargo, una base de datos orientada a documentos sí tiene conocimiento sobre el contenido del documento y su estructura, de tal forma que se pueden hacer búsquedas no solo por clave, sino por cualquier campo del documento, algo impensable en una base de datos clave-valor.

La base de datos orientada a documentos más extendida en la actualidad es MongoDB[126].

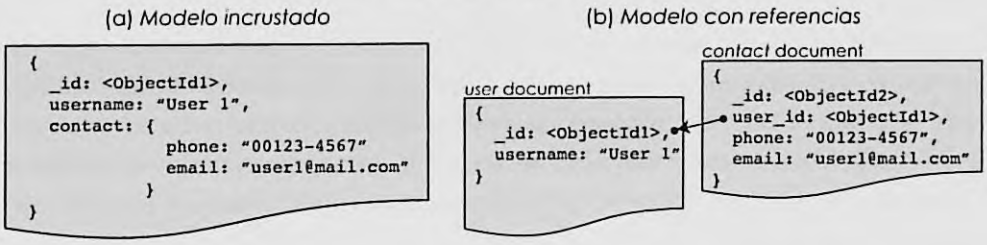
6.3.3.1 MongoDB

Como acabamos de indicar, MongoDB es la base de datos documental más extendida actualmente, y cuenta con más de 10 millones de descargas desde su lanzamiento en 2009 hasta mediados de 2015. A continuación describiremos sus principales características.

Modelo de datos. MongoDB almacena los datos en documentos BSON[127] (Binary JSON), que es una extensión de JSON que permite nuevos tipos de datos (como fechas y datos binarios) y, al igual que JSON, soporta un esquema variable. Todos los documentos tienen una clave llamada `_id`, que es un valor asignado automáticamente por MongoDB para identificar a un documento. Además, los documentos pueden contener subdocumentos incrustados (Figura 36a) o referencias a otros documentos (Figura 36b, que es un enfoque más parecido al que se emplea en bases de datos relacionales). No obstante, es importante señalar que MongoDB no soporta JOIN, por lo que en este último caso debe ser la aplicación la encargada de gestionar cómo se combinan los diferentes documentos.

La principal ventaja del modelo incrustado es que requiere menos consultas, pues al recuperar un documento se recuperan también los subdocumentos que contiene; si bien tiene el inconveniente de que los documentos pueden crecer mucho en tamaño con el paso del tiempo. El modelo de referencias tiene la ventaja de que es más flexible que el de subdocumentos incrustados; con el inconveniente de requerir más operaciones de lectura para recuperar un registro completo con todas sus referencias (pues habrá que acceder a varios documentos). Se puede encontrar más información y consejos de diseño en la documentación de modelos de datos de MongoDB[128].

Figura 36 – Modelos de datos en MongoDB



[126] MongoDB. *MongoDB*. https://www.mongodb.org. [Online; consultado el 1 de junio de 2015]
[127] BSON Spec. *BSON - Binary JSON*. http://bsonspec.org. [Online; consultado el 1 de junio de 2015]
[128] MongoDB. *Data Model Design for MongoDB*. http://docs.mongodb.org/master/MongoDB-data-models-guide.pdf. [Online; publicado el 29 de mayo de 2015]

Rendimiento. Por defecto, el campo `_id` de los documentos está indexado, lo que significa que se puede realizar una búsqueda por ese campo de manera eficiente, algo similar a lo que ocurría en el caso de la recuperación por clave en las bases de datos clave-valor. No obstante, MongoDB permite definir índices en cualquiera de los campos del documento, lo que permite realizar consultas más eficientes por aquellos campos que estén indexados. Se puede encontrar más información sobre los tipos de índice, cómo definirlos y su funcionamiento en la documentación de índices de MongoDB[129].

Escalabilidad. MongoDB permite escalar horizontalmente, dividiendo los datos almacenados en varios nodos de un *cluster* mediante un concepto denominado *sharding*. Esto lo que permite es que cada nodo (también llamado *shard*) almacene un subconjunto de los datos de forma independiente. Para decidir en qué nodo se almacena cada documento, se debe definir una *shard key*, lo que puede hacerse de dos formas distintas:

- ***Sharding* por rango**, donde se define que un determinado rango de valores esté alojado en el mismo *shard*. Esto suele garantizar que valores cercanos estén alojados en el mismo nodo.
- ***Sharding* por** hash, donde se calcula una función *hash* del valor para la *shard key*, lo que permite una distribución más homogénea de los valores entre los diferentes nodos.

Se puede encontrar más información sobre este concepto de distribución de datos en la documentación de *sharding* de MongoDB[130].

Además, MongoDB también permite la replicación de los datos, evitando así que se pueda perder información en caso de que un nodo deje de estar accesible. En MongoDB se denomina *replica set* a un conjunto de nodos que almacenan el mismo conjunto de datos, y que contienen un nodo primario donde se escriben todos los datos y varios nodos secundarios que propagan las operaciones que se han realizado sobre el primario, de tal modo que al final todos tengan los mismos datos. Se puede encontrar más información sobre el proceso de replicación de datos en la documentación de MongoDB[131].

Operaciones. Cuando hablamos de las bases de datos clave-valor avanzamos que las operaciones que soportan son muy sencillas: la inserción o actualización de un valor por clave, la consulta de un valor por clave y la eliminación de un par clave-valor. En una base de datos orientada a documentos como MongoDB el abanico de operaciones es mucho más amplio. Por ejemplo, podremos consultar documentos que tengan un determinado

[129] MongoDB. *Indexes and MongoDB.* http://docs.mongodb.org/master/MongoDB-indexes-guide.pdf. [Online; publicado el 29 de mayo de 2015]

[130] MongoDB. *Sharding and MongoDB.* http://docs.mongodb.org/master/MongoDB-sharding-guide.pdf. [Online; publicado el 29 de mayo de 2015]

[131] MongoDB. *Replication and MongoDB.* http://docs.mongodb.org/master/MongoDB-replication-guide.pdf. [Online; publicado el 29 de mayo de 2015]

valor para un campo y al igual que hacíamos con SQL, podríamos emplear los operadores lógicos AND y OR para forzar a que deban cumplirse varias condiciones o que solo tenga que darse una de ellas respectivamente. Se puede encontrar información más detallada en la documentación de MongoDB[132].

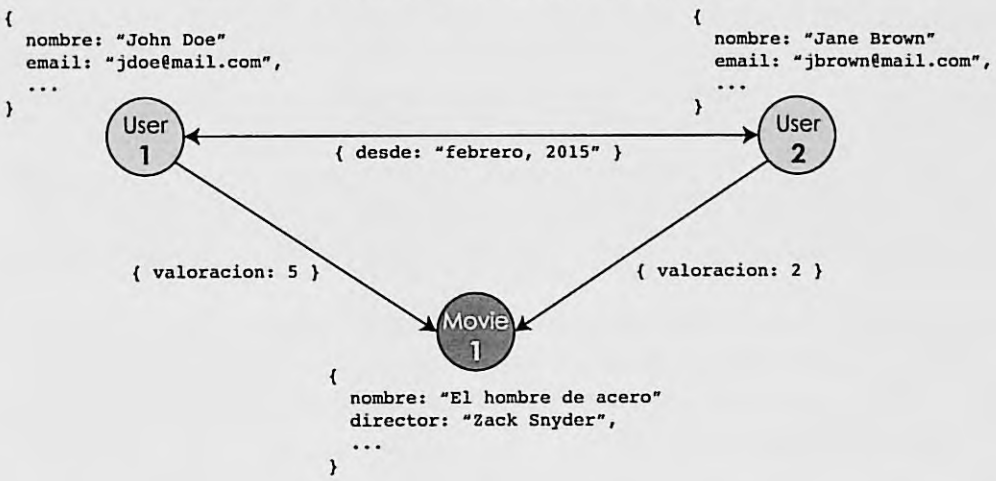
Además, MongoDB proporciona herramientas de agregación, que permiten realizar transformaciones sobre los datos para poder calcular determinados resultados. Esto puede ser equivalente a las agregaciones que tiene SQL, por ejemplo, para contar documentos que tengan una determinada condición, calcular medias y otros valores estadísticos, etc. Se puede encontrar más información sobre las operaciones de agregación en la documentación de MongoDB[133].

6.3.4 Bases de Datos Orientadas a Grafos

Otro tipo de bases de datos no relacionales son las orientadas a grafos. Un grafo es una estructura compuesta por nodos[134] (también llamados vértices) y enlaces (también llamados aristas). De forma simplificada, un nodo representa una entidad o un registro, mientras que un enlace representa una relación entre dos entidades.

La Figura 37 muestra un ejemplo de grafo. En este grafo, podemos ver dos nodos que representan a los usuarios «John Doe» y «Jane Brown», además de otro nodo que representa a la película «El hombre de acero». Como se puede observar, estos nodos son equivalentes a los documentos que ya teníamos en las bases de datos documentales co-

Figura 37 – *Ejemplo de grafo representando una red social*



[132]MongoDB. *MongoDB CRUD Operations.* http://docs.mongodb.org/master/MongoDB-crud-guide.pdf. [Online; publicado el 29 de mayo de 2015]

[133]MongoDB. *MongoDB Aggregation and Data Processing.* http://docs.mongodb.org/master/MongoDB-aggregation-guide.pdf. [Online; publicado el 29 de mayo de 2015]

[134]No se deben confundir con los nodos a los que hacíamos referencia al hablar de las máquinas que conforman un *cluster*.

mo MongoDB, donde cada nodo tiene una serie de campos que, además, no tienen por qué estar sujetos a una estructura fija (es decir, cada nodo puede tener campos distintos).

Además, entre estos nodos hay relaciones. Por ejemplo, entre «John Doe» y «Jane Doe» hay una relación de amistad, que además tiene un campo que indica el momento desde el que son amigos: «febrero de 2015». Entre cada usuario y la película hay otra relación que indica la valoración que ese usuario ha puesto a la película. Como se puede observar, las relaciones también pueden tener campos, y estos tampoco tienen por qué ajustarse a ninguna estructura.

Es posible que en este momento estemos viendo la similitud entre el grafo de la Figura 37 y la relación entre documentos que ya vimos en la Figura 36b, cuando hablamos de MongoDB. No obstante, hay una diferencia principal: en las bases de datos orientadas a grafos, el propio sistema de base de datos tiene en cuenta las relaciones, por lo que permite hacer consultas que incluyan JOIN, mientras que en las bases de datos orientadas a documentos esto no era posible, y era la aplicación la que tenía que realizar esta operación.

La base de datos orientadas a grafos más extendida en la actualidad es Neo4j[135].

6.3.4.1 Neo4j

En esta sección vamos a ver con más detalle Neo4j, puesto que se ha convertido en un estándar *de facto* de las bases de datos orientadas a grafos, además de ser el sistema que da soporte a empresas como InfoJobs, el líder europeo en búsqueda de empleo en Internet.

Modelo de datos. El modelo de datos de Neo4j es un grafo, como el que ya vimos en la Figura 37. Tanto los nodos como los enlaces pueden ser de varios tipos, y pueden contener campos asociados. Por esta razón, el modelo de datos combina características de las bases de datos documentales y de las bases de datos relacionales: permiten alojar datos semiestructurados y al mismo tiempo relacionar varias entidades empleando JOIN.

Rendimiento. Neo4j permite crear índices sobre campos de algunos nodos, de tal forma que la recuperación de nodos por ese campo sea más eficiente. Esta opción es similar a la que podemos tener en SQL, y vimos que también era posible en MongoDB, si bien Neo4j no ofrece tantas opciones de índices.

Escalabilidad. Neo4j se ofrece en dos ediciones: la *Community* y la *Enterprise*, siendo esta última de pago. Precisamente porque está enfocada a ser utilizada en entornos reales empresariales, la edición *Enterprise* ofrece replicación de los datos en varios nodos de un

[135] Neo4j. *Neo4j, the World's Leading Graph Database.* http://neo4j.com. [Online; consultado el 2 de junio de 2015]

cluster. No obstante, hay que señalar que no se realiza una distribución de los datos, es decir, no se permite que cada nodo almacene un subconjunto distinto de los datos, como sí hacía MongoDB mediante el *sharding*. De esta forma, la cantidad de datos a almacenar está limitada a la capacidad del servidor.

Operaciones. Neo4j dispone de un lenguaje de consulta denominado Cypher, que permite realizar inserciones, borrados, modificaciones y recuperaciones de datos. En lo que respecta a las recuperaciones, el funcionamiento más sencillo involucra consultar aquellos nodos que cumplan una determinada condición, como ya hacía MongoDB. No obstante, podemos además emplear los enlaces para consultar las relaciones que existen entre nodos, o incluso los nodos que están relacionados con uno dado.

6.3.5 Bases de Datos Orientadas a Columnas

Como ya vimos anteriormente, Google introdujo en 2003 GFS, el sistema de ficheros distribuido que soportaba los servicios de la compañía. El principal inconveniente de este sistema es que a todos los efectos en un sistema de ficheros, por lo que no está diseñado para recuperar datos eficientemente, ya que la información no está indexada.

En 2006, Google presenta BigTable[136], que es una base de datos que funciona sobre GFS. La principal ventaja de este enfoque es que este sistema va a presentar algunas de las ventajas de las bases de datos, pero además el sistema de ficheros subyacente proporcionará escalabilidad horizontal y replicación de los datos.

BigTable es una base de datos orientada a columnas. Este término hace referencia al hecho de que los valores de una misma columna se almacenan físicamente juntos en disco, lo que incrementa la eficiencia de realizar operaciones de agregación sobre columnas.

Su modelo de datos es relativamente similar al relacional: se pueden crear tablas con una serie de grupos de columnas que son fijos para cada tabla, y que reciben el nombre de «*column families*». Cada registro es una fila, que tiene una clave denominada «*row key*», similar a la clave primaria en SQL. La base de datos construye un índice automáticamente sobre esta clave con el fin de optimizar las recuperaciones de filas.

Los grupos de columnas contienen a continuación columnas identificadas por una clave llamada «*qualifier*», que tendrán un valor asociado para cada registro. Además, en BigTable se almacena todo el histórico de valores para un *qualifier* dado, mediante un *timestamp* que identifica en qué momento del tiempo se asignó un determinado valor a una columna. Estas columnas sí pueden variar entre diferentes registros, lo que hace de BigTable una solución adecuada para almacenar datos semiestructurados.

[136] Fay Chang y col. «BigTable: a Distributed Storage System for Structured Data». En: *7th Symposium on Operating System Design and Implementation*. 2006

Apache HBase[137] es una implementación de código abierto de BigTable que forma parte del ecosistema de Hadoop.

6.3.5.1 HBase

Modelo de datos. El modelo de datos de HBase es idéntico al de BigTable que hemos explicado con anterioridad. Los registros se almacenan en filas, y aunque existen grupos de columnas que están fijos en la estructura de la tabla, las columnas que puede tener cada registro sí son variables, lo que permite almacenar datos semiestructurados.

Rendimiento. HBase indexa de forma automática las claves de cada fila (*row keys*), por lo que la recuperación de un registro dada su clave se puede hacer de forma eficiente. No obstante, HBase no permite crear otros índices. Puesto que la base de datos es orientada a columnas, HBase también ofrece buen rendimiento cuando se quieren calcular agregados para una columna en muchos registros.

Escalabilidad. HBase funciona sobre HDFS, el sistema de ficheros distribuido de Hadoop, por lo que los datos estarán distribuidos entre varios nodos proporcionando así escalabilidad horizontal; y además estarán replicados para evitar pérdidas de datos.

Operaciones. HBase permite crear tablas y añadir, actualizar y eliminar registros en estas tablas.

6.3.6 Resumen de Opciones NoSQL

En esta sección hemos visto cuatro enfoques distintos de bases de datos que son alternativos al modelo relacional: las bases de datos clave-valor, las orientadas a documentos, las orientadas a grafos y las orientadas a columnas.

En la Figura 38 se muestra una pequeña tabla comparativa en la que se pueden observar las principales diferencias entre estas distintas alternativas.

6.3.7 Un Nuevo Paradigma: NewSQL

En los últimos años ha comenzado a surgir un nuevo enfoque de bases de datos que busca combinar la transaccionalidad y las garantías ACID de las bases de datos relacionales con la escalabilidad de las no relacionales, ya que en muchos entornos empresariales ambas características son requeridas.

Este enfoque ha recibido recientemente el nombre de NewSQL, un término acuñado por Matthew Aslett en 2011 para referirse a los nuevos sistemas de bases de datos que estaban surgiendo para hacer frente a los que ya estaban establecidos en el mercado[138].

[137] Apache. *Apache HBase*. http://hbase.apache.org. [Online; consultado el 3 de junio de 2015]

[138] Matthew Aslett. *How will the database incumbents respond to NoSQL and NewSQL?* http://cs.brown.edu/courses/cs227/archives/2012/papers/newsql/aslett-newsql.pdf. [Online; publicado el 4 de abril de 2011]

Una de las bases de datos NewSQL más relevantes es Spanner[139], introducida por Google en 2012 y que automáticamente se consideró la sucesora de BigTable. A diferencia de esta última, Spanner sí ofrece soporte para transacciones de forma distribuida.

6.4 Almacenamiento de Big Data en la Nube

Durante los últimos años, ha habido un rápido crecimiento de los servicios en la nube (o *cloud*), que involucran principalmente tres modelos distintos de servicios: el software como servicio (*software-as-a-service*, SaaS), la plataforma como servicio (*platform-as-a-service*, PaaS) y la infraestructura como servicio (*infrastructure-as-a-service*, IaaS).

Del gran abanico de opciones que ofrece la nube, una parte importante de ellas se dedican al almacenamiento. Además, el almacenamiento en la nube (o *cloud storage*) involucra los tres modelos de servicios anteriores. Por ejemplo, Google Drive, Microsoft OneDrive o Dropbox son ejemplos de SaaS que prestan un servicio de almacenamiento a los usuarios.

No obstante, algunas de las herramientas más interesantes de almacenamiento en la nube se encuentran en las capas de PaaS e IaaS. Los principales competidores de servicios en la nube (como Amazon Web Services o Google Cloud) ofrecen numerosas herramientas que permitirán desde obtener discos duros virtuales, hasta desplegar rápidamente bases de datos SQL y NoSQL.

Esta sección proporciona una visión a estas herramientas, explicando la funcionalidad de cada una de ellas y proporcionando una guía para comenzar a utilizarlas. Es importante señalar que el lector debe contar con una cuenta en estas plataformas si desea utilizar estas herramientas, y que su uso puede conllevar gastos económicos. Por este motivo, se recomienda al lector que se informe bien de los servicios que ofrecen las capas gratuitas

Figura 38 – *Comparación entre tecnologías de bases de datos*

	Relacionales *MySQL*	**Clave-Valor** *Voldemort*	**Documentos** *MongoDB*	**Grafos** *Neo4j*	**Columnas** *HBase*
Modelo de Datos	Tablas y relaciones	Pares clave-valor	Documentos BSON/JSON	Nodos y enlaces	Tablas con columnas no fijas
Relaciones (JOIN)	✔	✘	✘	✔	✘
Transacciones ACID	✔	✘	✘	✔	✘
Índices	✔	Solo en la clave	✔	✔	Solo en la clave de fila (rowkey)
Datos semiestructurados	✘	✔	✔	✔	✔
Escalabilidad horizontal	✘	✔	✔	✘	✔
Replicación de datos	✔	✔	✔	✔	✔

[139]James C. Corbett y col. «Spanner: Google's Globally-Distributed Database». En: *10th Symposium on Operating System Design and Implementation*. 2012

de cada proveedor y, si decide realizar pruebas, controle el gasto en todo momento para evitar imprevistos.

6.4.1 Almacenamiento de Ficheros

Una de las funcionalidades más básicas de los servicios en la nube involucra el almacenamiento de ficheros. Es probable que el lector esté familiarizado con herramientas como Dropbox, en las que se pueden subir ficheros y sincronizarlos desde diferentes equipos.

Los principales proveedores de *cloud* también ofrecen herramientas ideadas para que compañías puedan almacenar sus ficheros en la nube. A continuación se enumeran las principales opciones, indicando cómo utilizarlas y en qué casos son más adecuadas.

6.4.1.1 Amazon S3

Amazon S3 (Simple Storage Service) es un servicio para almacenar ficheros en la nube. La principal ventaja que ofrece es la posibilidad de acceder a estos ficheros desde las instancias creadas en Amazon, tanto para leer datos como para escribirlos.

Cuando accedamos a la consola de Amazon Web Services y escojamos el servicio S3, se nos mostrará una pantalla como la de la Figura 39. El concepto detrás de Amazon S3 es realmente sencillo: consiste en crear cubos (*buckets*) y a continuación subir objetos (ficheros y directorios) a estos cubos. Nosotros podemos decidir las zonas de disponibilidad (*availability zones*) de cada uno de nuestros cubos, y Amazon se encargará automáticamente de la replicación y distribución de los datos para evitar que pueda perderse información.

Para comenzar, haremos clic en el botón «*Create Bucket*». Se mostrará un formulario como el de la Figura 40, donde se nos solicitará un nombre para el cubo y una zona de disponibilidad. Escogeremos los valores que deseemos y haremos clic en el botón «*Create*».

A continuación, se nos mostrará la pantalla de la Figura 41, que contiene un listado de todos los cubos creados en nuestra cuenta. En la barra de la derecha se pueden configurar algunas opciones avanzadas, incluyendo permisos de diferentes usuarios[140] sobre el cubo.

Si hacemos clic en el cubo que acabamos de crear, se nos carga una nueva pantalla (Figura 42) donde podemos observar que se nos indica que el cubo está vacío. Además, como podemos observar, en la parte superior tenemos botones para subir ficheros («*Upload*») y para crear un nuevo directorio («*Create Folder*»). El proceso para crear directorios y subir ficheros es realmente sencillo.

[140] Amazon Web Services permite tener varios usuarios asociados a una cuenta, con diferentes permisos, empleando su herramienta IAM.

Además, si seleccionamos uno o más ficheros y hacemos clic en el botón «*Actions*», podremos ver una serie de operaciones que podemos realizar sobre estos ficheros (Figura 43). Algunos ejemplos son abrir los ficheros, descargarlos, hacerlos públicos, renombrarlos, borrarlos, copiarlos / pegarlos y modificar sus propiedades avanzadas (como el cifrado del fichero, sus permisos y sus metadatos).

Es importante señalar que Amazon permite acceder a los cubos desde las instancias de Amazon EC2. Además, también proporciona una interfaz de consola y una API para poder integrar S3 con nuestras aplicaciones.

Por último, podemos borrar el cubo desde la pantalla con el listado de cubos, si bien previamente deberemos haber eliminado todo el contenido del cubo. Es importante borrar el cubo cuando no lo vayamos a necesitar, pues en caso contrario se nos facturaría el servicio.

Figura 39 – Pantalla de inicio de Amazon S3

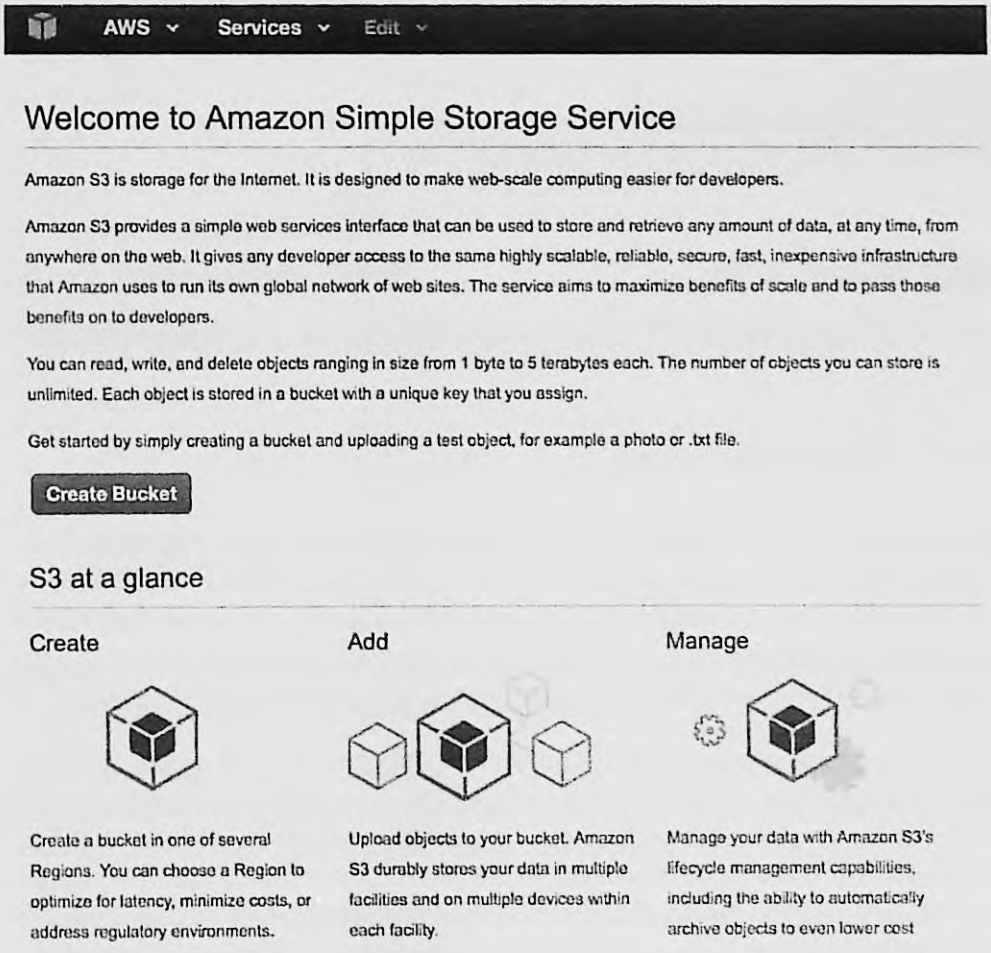


Figura 40 – Creación de un cubo en Amazon S3

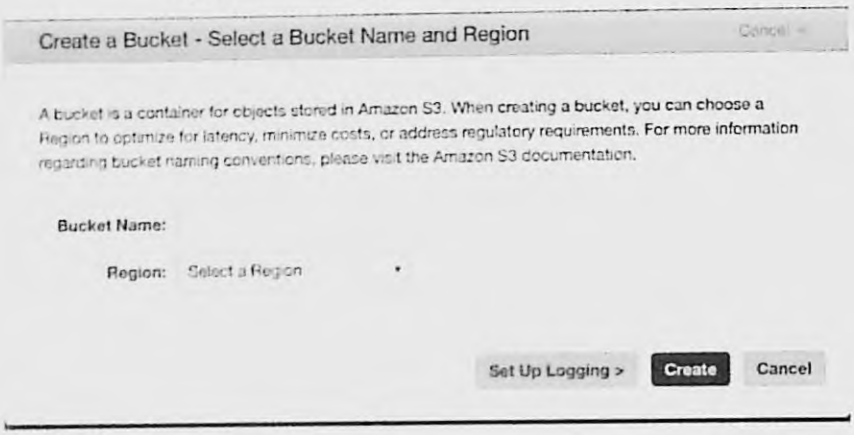


Figura 41 – Listado de cubos en Amazon S3

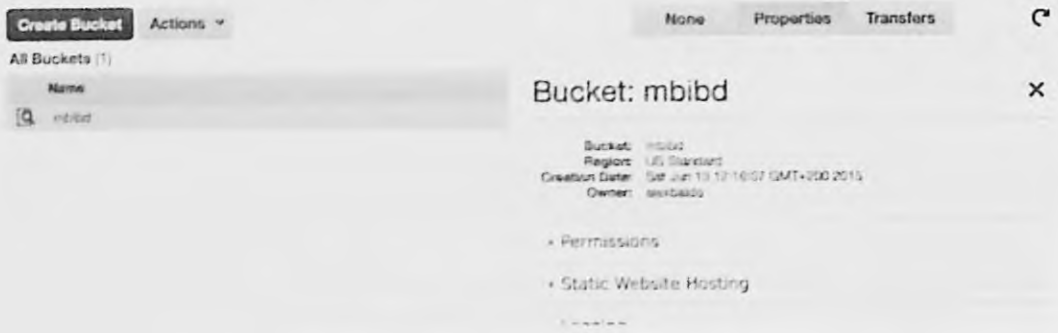
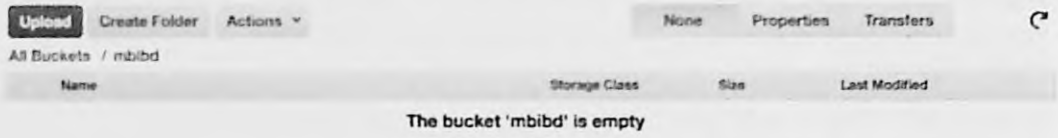


Figura 42 – Detalles del cubo en Amazon S3



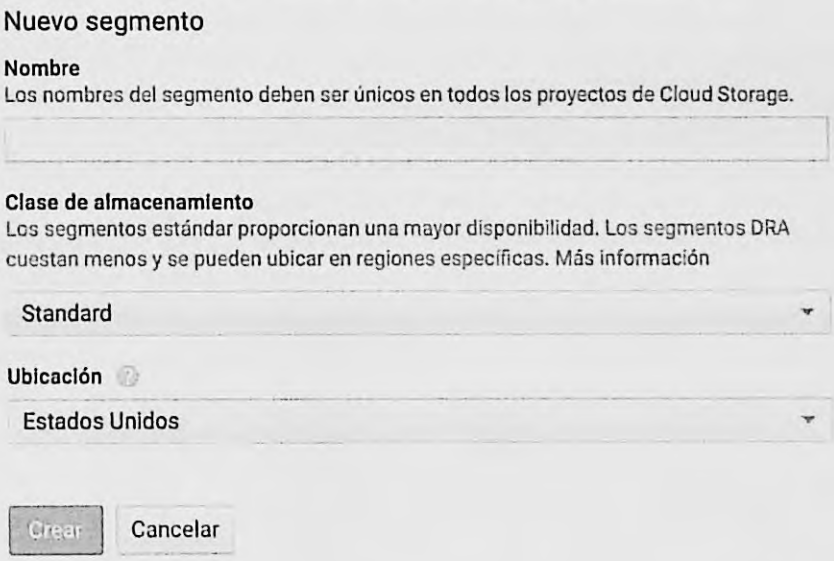
6.4.1.2 Google Cloud Storage

Google también proporciona un servicio de almacenamiento en la nube. Para poder hacer uso de Google Cloud Storage, primero deberemos crear una cuenta de Google Cloud y a continuación crear un proyecto, tras lo cual podremos abrirlo y accederemos al panel de control del proyecto. En la barra lateral de la izquierda, haremos clic en la sección «Almacenamiento» y a continuación en «Cloud Storage» y «Explorador de Storage».

Como podemos observar, Cloud Storage permite almacenar objetos no estructurados (ficheros y directorios) en segmentos. Un segmento es un concepto similar al de cubo o *bucket* que ya vimos en Amazon S3. Para comenzar, haremos clic en el botón «Crear un segmento», y se nos mostrará la pantalla de la Figura 44. A continuación escogeremos un nombre para el mismo, y como podemos observar, podemos decidir también la ubicación (tal y como ocurría con la zona de disponibilidad de Amazon S3) donde se almacenará el

Figura 43 – Operaciones con ficheros en Amazon S3

Figura 44 – Creación de un segmento en Google Cloud Storage



segmento. Además, también se puede escoger la clase de almacenamiento, en función de si queremos primar el precio o la disponibilidad de los datos.

Tras crear el segmento, se mostrará la pantalla de la Figura 45, que como podemos observar es muy similar a la vista que teníamos en Amazon S3. Desde este panel, se pueden crear directorios, subir ficheros y eliminarlos, modificar sus propiedades, etc.

Al igual que Amazon S3, Google proporciona una API para poder leer y escribir ficheros en Google Cloud Storage desde aplicaciones externas, lo que resulta de gran utilidad para

Figura 45 – Contenido del segmento en Google Cloud Storage

los desarrolladores. Para obtener más información sobre el uso de la API, se puede hacer clic en el enlace «Acceso al almacenamiento» en la barra de la izquierda.

Por último, cuando no lo necesitamos podemos eliminar el segmento que hemos creado. En este caso, no es necesario que esté vacío previamente, puesto que Google Cloud Storage eliminará automáticamente su contenido. Si no eliminamos el segmento, el almacenamiento utilizado se facturará, por lo que incurriríamos en costes.

6.4.2 Bases de Datos SQL

Los principales proveedores de servicios en la nube también proporcionan herramientas para crear instancias de bases de datos relacionales. Normalmente estas herramientas funcionan creando instancias (máquinas virtuales) de forma transparente para el usuario, donde se despliega el servidor de bases de datos relacionales escogido (MySQL, PostgreSQL, Oracle, SQL Server, etc).

La principal ventaja que ofrecen estas herramientas es su capacidad de escalar, es decir, de poder mejorar las máquinas en tiempo real si es necesario para que estas puedan soportar más tráfico o almacenar mayores cantidades de datos. No obstante, no siempre se soporta una escalabilidad horizontal, por lo que no todas estas herramientas son realmente capaces de almacenar Big Data.

6.4.2.1 Amazon RDS

Amazon RDS (Relational Database Service) es el servicio de Amazon Web Services para crear bases de datos relacionales. Una vez escogido el servicio en la consola de AWS, se nos mostrará una pantalla como la de la Figura 46. Para comenzar a utilizar RDS, haremos clic en el botón «*Get Started Now*».

A continuación se nos solicitará que escojamos el motor de base de datos relacional que queremos utilizar, en una pantalla como la de la Figura 47. En este caso, Amazon nos permite escoger los principales motores de bases de datos relacionales: MySQL, PostgreSQL, Oracle Database y Microsoft SQL Server. Seleccionaremos uno de ellos, por ejemplo, MySQL Community Edition y haremos clic en el boton «*Select*».

En el siguiente paso, se nos preguntará si la base de datos que vamos a utilizar es para un servicio en producción. De responder afirmativamente, Amazon desplegará la base de datos utilizando varias zonas de disponibilidad y con un almacenamiento que permite realizar consultas más rápidas y tener una mayor disponibilidad de los datos, aunque todo ello con un mayor coste (Figura 48). Para continuar, seleccionaremos la opción «*No*» y haremos clic en el botón «*Next Step*».

A continuación deberemos configurar la base de datos (Figura 49). El servicio nos preguntará la versión del motor a utilizar, el tipo de instancia a utilizar (que determinará el procesador y la memoria del servidor de bases de datos), el tipo de disco duro, la capacidad de almacenamiento (con un mínimo de 5 GB y un máximo de 3072 GB), el nombre de la instancia y los datos del usuario administrador de la base de datos. Después de configurar estos aspectos, haremos clic en el botón «*Next Step*».

Por último, se pueden configurar aspectos más avanzados de la base de datos (Figura 50), como configuración de red y seguridad, copias de seguridad, mantenimiento automático, etc. La opción más interesante es «*Database Name*», dentro de la sección «*Database Options*», ya que si introducimos un nombre se creará una base de datos automáticamente con ese nombre. Para finalizar, haremos clic en el botón «*Launch DB Instance*».

Tras completarse este procedimiento, haremos clic en la opción «*View Your DB Instances*». Se nos cargará el listado con las instancias de bases de datos creadas (por el momento, solo habrá una), tal y como se muestra en la Figura 51. Al principio es posible que la columna *Status* tenga el valor «*creating*», lo que indicará que la instancia aún se está configurando y deberemos esperar un poco más. Una vez arrancada la instancia, podríamos acceder a la base de datos con una herramienta gráfica e introduciendo los datos de

Figura 46 – Pantalla de inicio de Amazon RDS

Figura 47 – Selección de motor de bases de datos Amazon RDS

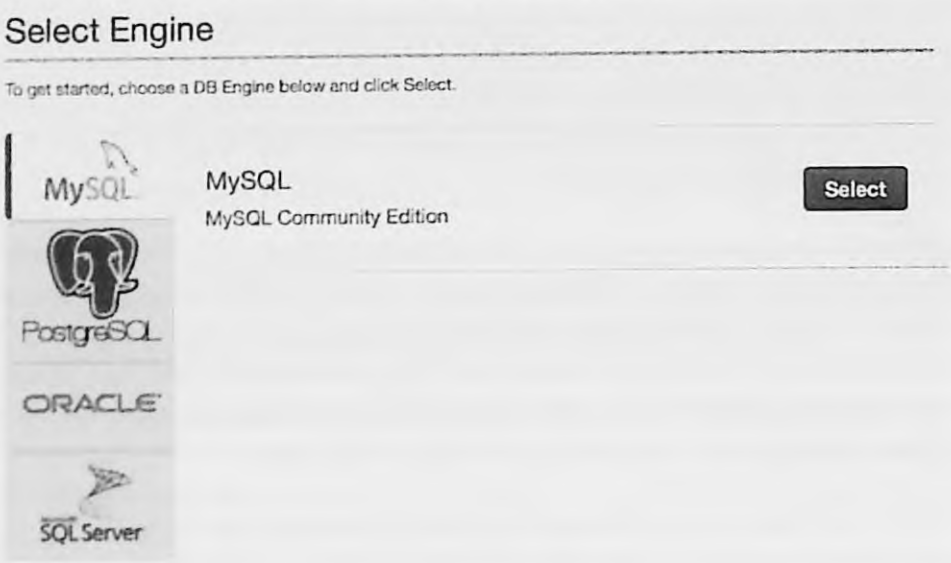
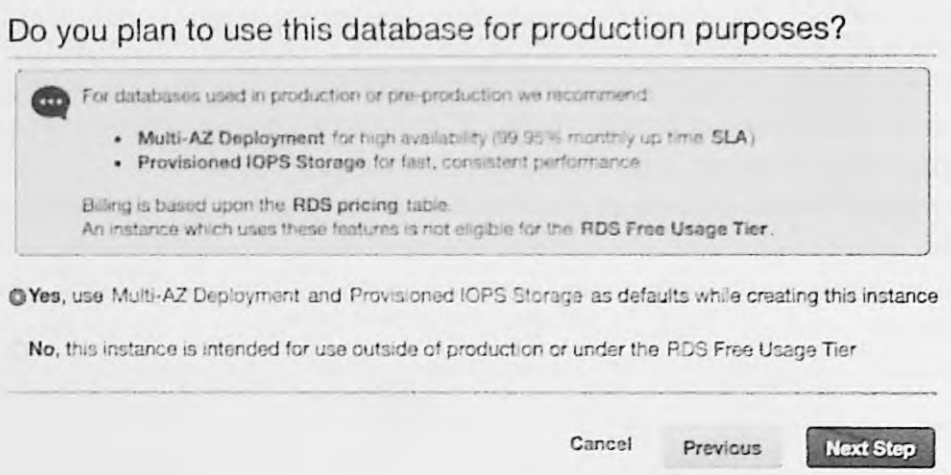


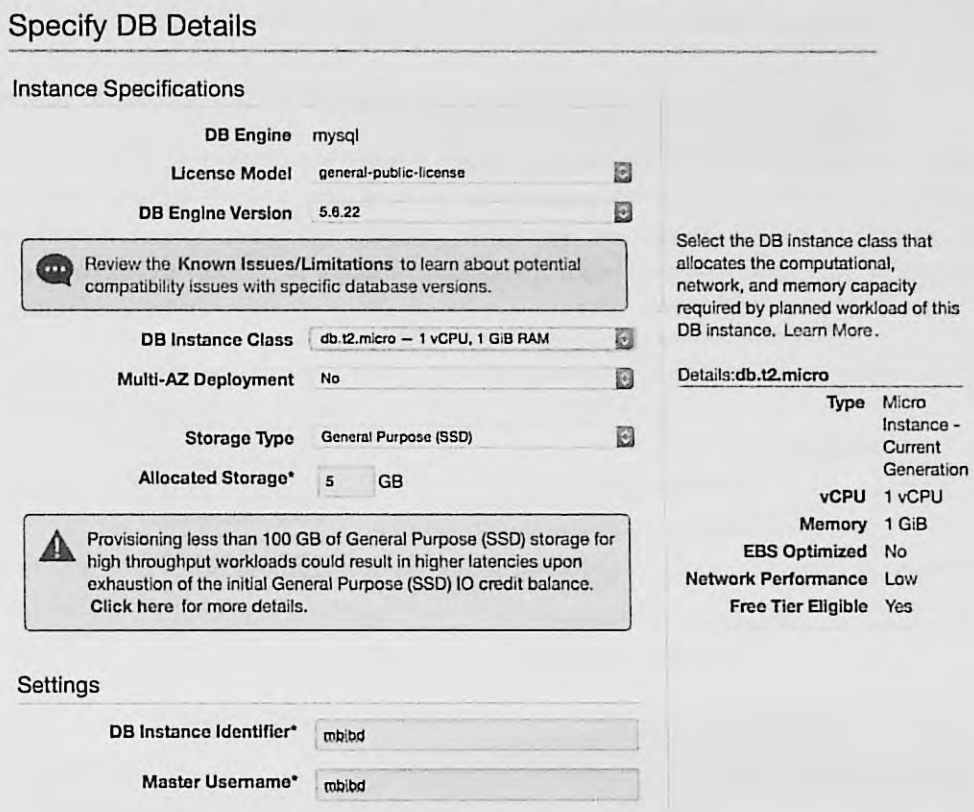
Figura 48 – Selección de características para producción en Amazon RDS



la instancia. El host es el valor indicado por el *Endpoint* en RDS, y los demás valores (nombre de usuario, contraseña y nombre de la base de datos) los especificamos nosotros durante la creación de la instancia.

Como se puede observar, Amazon RDS simplifica enormemente el proceso de crear bases de datos relacionales, sin embargo, no ofrece escalabilidad horizontal, aunque sí es posible mejorar las características de la instancia (es decir, de la máquina virtual) en cualquier momento.

Figura 49 – Creación de la base de datos en Amazon RDS



6.4.2.2 Google Cloud SQL

Google Cloud también ofrece Cloud SQL, un servicio para desplegar fácilmente instancias de bases de datos relacionales, que gestiona automáticamente la replicación de los datos y el mantenimiento del motor de bases de datos.

Tras acceder a la web de Google Cloud y pulsar en la opción «Cloud SQL» dentro de la sección «Almacenamiento» de la barra lateral izquierda, se nos cargará un mensaje como el de la Figura 52. Para comenzar, haremos clic en el botón «*Create instance*», y a continuación se mostrará la pantalla de la Figura 53, en la que podremos asignar un nombre a la instancia de Cloud SQL, una región y una cantidad de memoria.

En las opciones avanzadas se podría escoger la versión del motor de base de datos (Cloud SQL solo funciona con MySQL), el tipo de facturación, las copias de seguridad y otras configuraciones más específicas. Finalmente haremos clic en el botón «Crear» para proceder con la creación de la instancia. Este proceso podrá alargarse durante unos minutos.

Figura 50 – Opciones avanzadas de base de datos en Amazon RDS

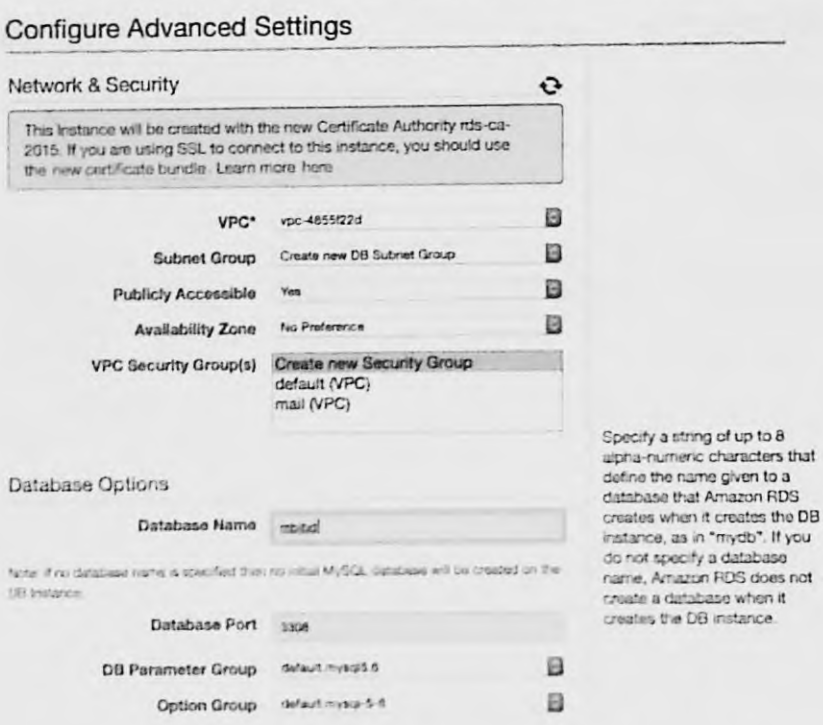


Figura 51 – Listado de instancias de bases de datos en Amazon RDS

Cuando haya finalizado la creación de la instancia, esta aparecerá en la lista, tal y como se observa en la Figura 54. Si hacemos clic sobre el ID de la instancia, se podrá acceder a la pantalla de detalles de la misma, que se corresponde con la mostrada en la Figura 55.

Como se puede observar, este panel nos proporciona información estadística sobre el uso de la instancia de bases de datos. Además, en la parte inferior, el enlace «Cómo conectarse a una instancia de Cloud SQL» nos indica cómo se puede integrar esta base de datos con nuestras aplicaciones. Además, la pestaña «Bases de datos» en la parte superior nos permitirá crear nuevas bases de datos y gestionar las existentes.

Por último, cuando no necesitemos la instancia de bases de datos podremos borrarla, con el fin de que no se facturen cargos.

6.4.3 Bases de datos NoSQL

Como ya vimos en el capítulo anterior, existen numerosas bases de datos que suponen una alternativa al modelo relacional. Normalmente estas bases de datos proporcionan una mejor escalabilidad (que suele ser horizontal), pero renunciando a algunas de las propiedades de las bases de datos relacionales, como la transaccionalidad ACID o las operaciones JOIN.

Los principales proveedores de servicios en la nube ofrecen algunas herramientas de bases de datos NoSQL, que pasamos a describir a continuación.

6.4.3.1 Amazon DynamoDB

Dynamo DB es una base de datos NoSQL de Amazon que emplea un modelo de datos parecido al de un almacén clave-valor, con algunas características propias de las bases de

Figura 52 – Pantalla inicial de Google Cloud SQL

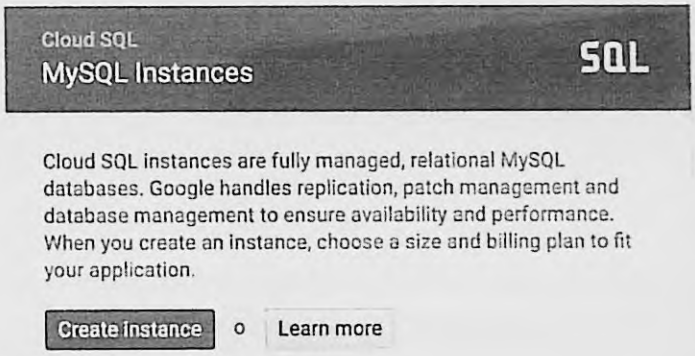


Figura 53 – Creación de una instancia de Google Cloud SQL

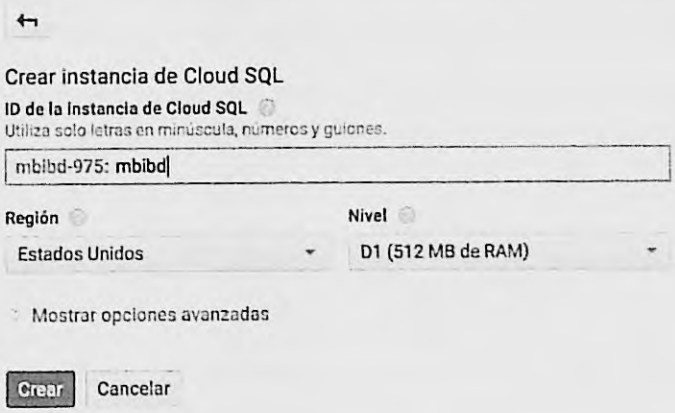


Figura 54 – Lista de instancias Google Cloud SQL

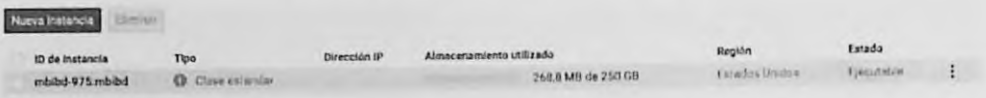
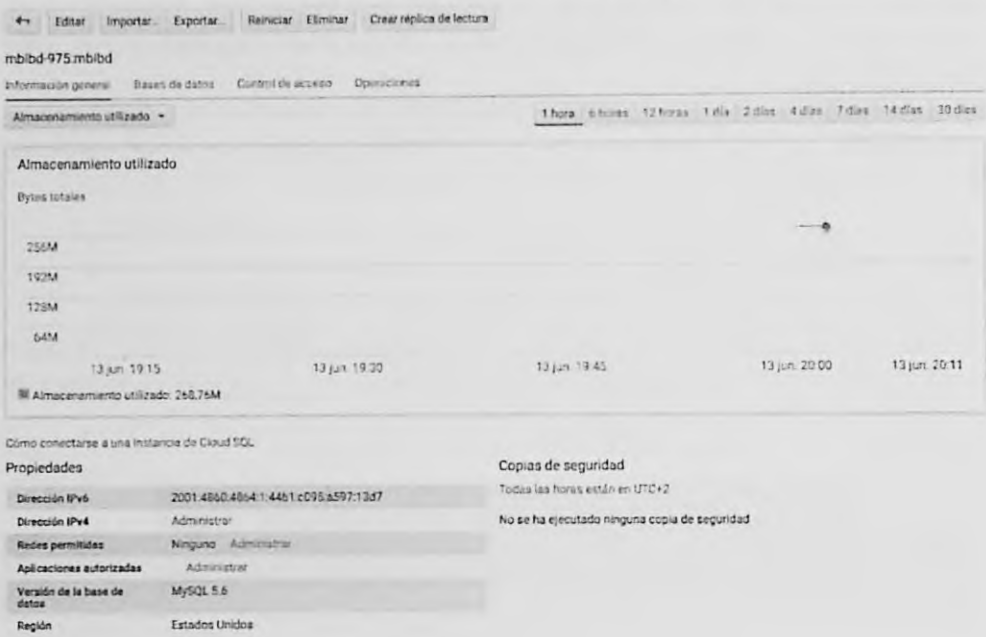


Figura 55 – Detalles de una instancia de Google Cloud SQL



datos orientadas a documentos. El usuario puede crear tablas, siendo cada una de estas tablas una colección de objetos, que a su vez tienen atributos.

En cada tabla se debe definir un campo como clave primaria, que todos los objetos de esa tabla deben tener. No obstante, los demás atributos no se definen a priori, lo que permitirá almacenar datos semiestructurados. También se pueden definir índices secundarios en otros atributos, lo que permite recuperar objetos más eficientemente.

Además, Dynamo DB es escalable horizontalmente, lo que permite almacenar un mayor volumen de datos simplemente añadiendo más equipos al *cluster*.

Una vez que accedamos al servicio Dynamo DB en la consola de AWS, se nos mostrará la pantalla de la Figura 56. En ella, podremos comenzar a utilizar la herramienta creando una tabla, para lo que haremos clic en el botón «*Create Table*».

En este momento, se mostrará el formulario de la Figura 57. En el primer paso del formulario, debemos escoger el nombre de la tabla que vamos a crear. Además, debemos escoger cuál será la clave primaria. Como se puede observar, se ofrecen dos opciones:

crear una clave primaria *hash* o una clave primaria *hash and range*. En ambos casos se construirá un índice *hash* por el primer campo, y opcionalmente se construirá un índice secuencial por el segundo campo si este existe. Tal y como se indica, se debe buscar que el campo *hash* sea equilibrado, es decir, que existan aproximadamente el mismo número de accesos a cada elemento, para distribuir la carga entre los diferentes equipos.

Tras hacer clic en el botón «*Continue*», en el siguiente paso (mostrado en la Figura 58) podremos crear índices secundarios. Estos índices tienen las mismas características que la clave primaria, es decir, pueden ser de tipo *hash* o *hash and range*. El campo «*projected attributes*» nos permite seleccionar los campos que recuperaremos cuando hagamos una búsqueda por ese índice. Por defecto están incluidos todos, pero en caso de que no sean necesarios, eliminar campos permitirá ganar en eficiencia.

En la siguiente pantalla, la mostrada en la Figura 59, Amazon nos pregunta el tamaño aproximado que tendrán nuestros objetos y la cantidad de lecturas y escrituras que esperamos que tengan por segundo, así como si necesitamos «consistencia estricta» en la lectura. Con los valores que introduzcamos, Amazon determinará el número de equipos necesarios que soportará la base de datos, e incluso proporcionará una estimación del coste del servicio. En el siguiente paso se pueden definir alarmas automáticas cuando se estén agotando las unidades de lectura o escritura por hora, mientras que en el último se mostrará un resumen de la configuración escogida para la tabla a crear. Bastará con hacer clic en el botón «*Create*» para que comience el proceso, que podrá llevar unos minutos.

Tras completarse el proceso, la tabla se mostrará como activa en el listado de tablas, tal y como se muestra en la Figura 60. Si la seleccionamos y hacemos clic en el botón «*Explore*

Figura 56 – Pantalla de inicio de Dynamo DB

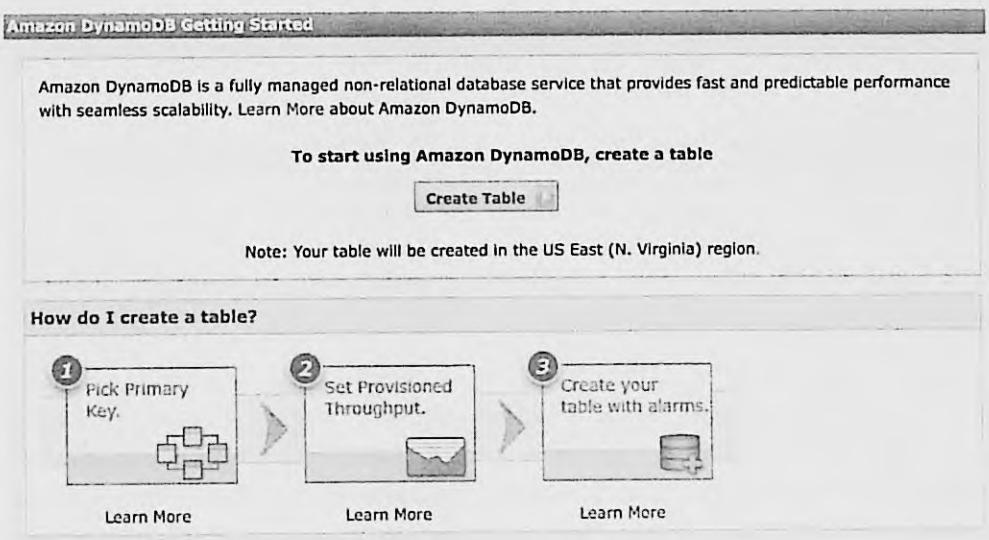


Figura 57 – Creación de una tabla en Dynamo DB

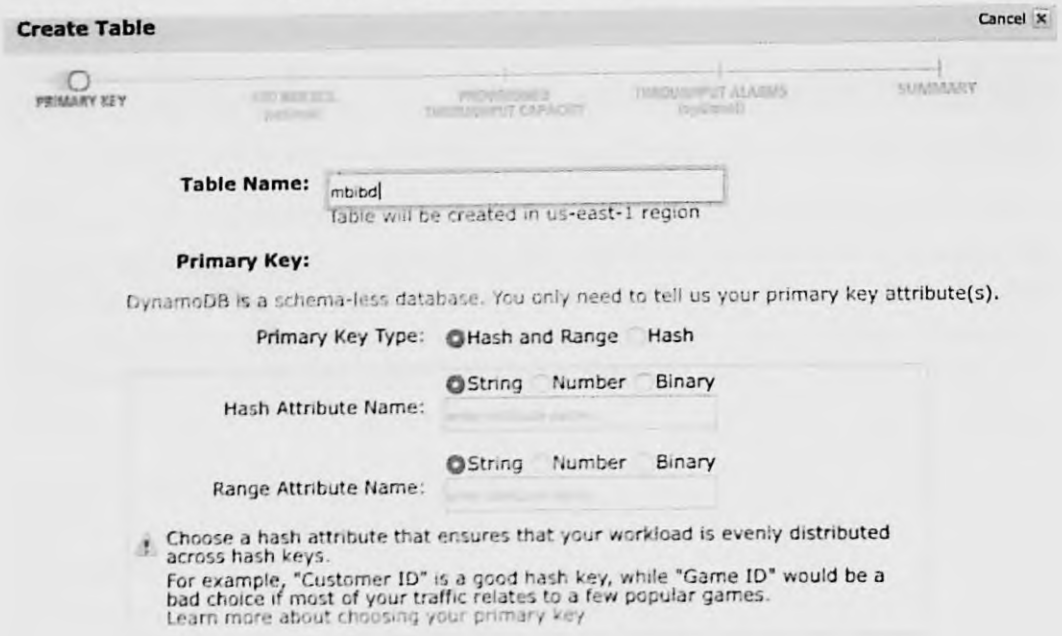


Figura 58 – Selección de índices en Dynamo DB

Table» de la parte superior se nos mostrará una nueva pantalla con los objetos contenidos por la tabla, que de momento estará vacía.

Si hacemos clic en el botón «*Create Item*» se cargará una interfaz para añadir un nuevo objeto a nuestra tabla, como se muestra en la Figura 61. Los campos definidos en la clave primaria aparecen por defecto, al ser obligatorios. El resto de campos los puede definir el usuario manualmente, y no tienen que estar sujetos a ninguna estructura. Como se puede observar, el formato de objeto es similar a JSON, puesto que se trata de una colección de claves y valores. Al pulsar el botón «Save», el objeto se guardará.

Figura 59 – Provisionamiento en Dynamo DB

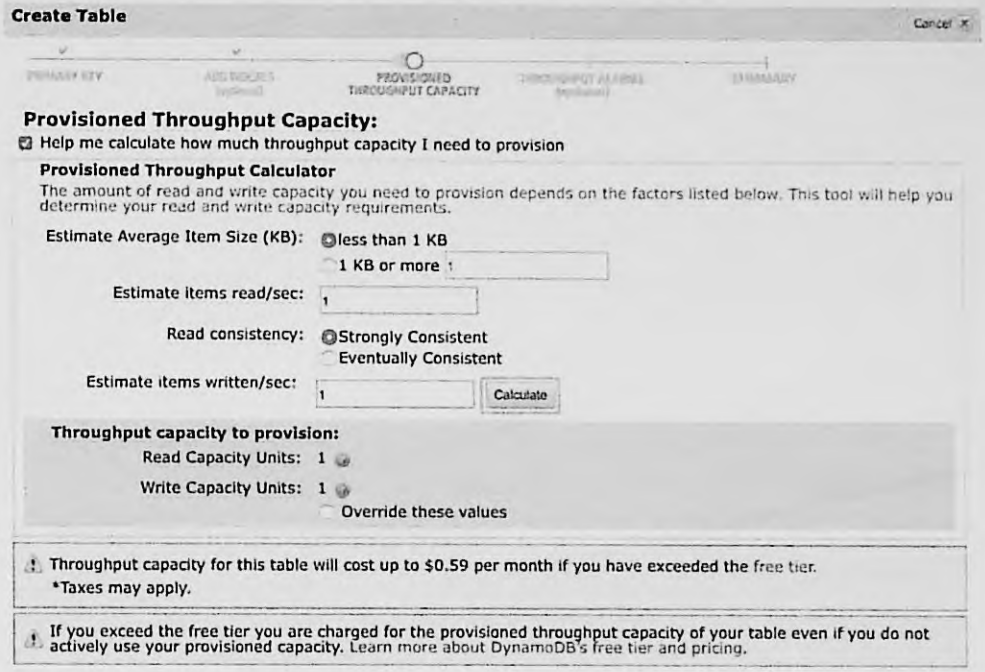


Figura 60 – Lista de tablas en Dynamo DB

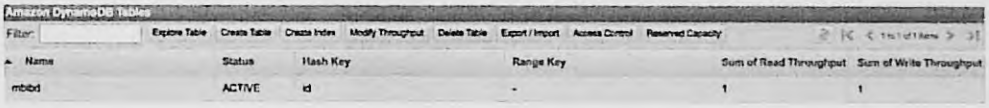


Figura 61 – Creación de objeto en Dynamo DB

Amazon Dynamo DB dispone de una API para que la base de datos se pueda integrar fácilmente con nuestras aplicaciones, permitiendo su uso en el mundo real.

7 Procesamiento de Big Data

7.1 Introducción al Procesamiento Masivo de Datos

Cuando disponemos de Big Data, evidentemente nuestra principal preocupación no es almacenarlo, sino ser capaces de procesarlo y extraer valor de negocio de esta gran cantidad de datos.

En el capítulo anterior ya mencionamos los principales retos que introducen las llamadas «3 Vs» en el almacenamiento de Big Data. No obstante, es importante señalar que estos retos son similares cuando queremos realizar un procesamiento de estos datos.

Evidentemente, necesitaremos sistemas que sean capaces de hacer frente a los desafíos que introducen las 3 Vs del Big Data, permitiendo realizar un procesamiento distribuido de los datos (que en ocasiones tendrá que ser en tiempo real), y que sean capaces de ofrecer suficiente flexibilidad como para poder procesar datos que sean semiestructurados o carezcan completamente de estructura.

7.1.1 Paradigmas de Procesamiento de Big Data

En general, se suelen distinguir dos enfoques cuando se habla de procesamiento de Big Data, en función del factor temporal en el que se lleva a cabo este procesamiento. Estos enfoques son el procesamiento por lotes (*batch processing*) o el procesamiento en tiempo real (*stream processing*).

7.1.1.1 Procesamiento por lotes (*batch*)

En el procesamiento de datos por lotes, el factor tiempo no es crítico. Este tipo de procesamiento se lleva a cabo cuando se cuenta con una gran cantidad de datos, que suelen ser históricos, y se desea realizar algún análisis de los mismos.

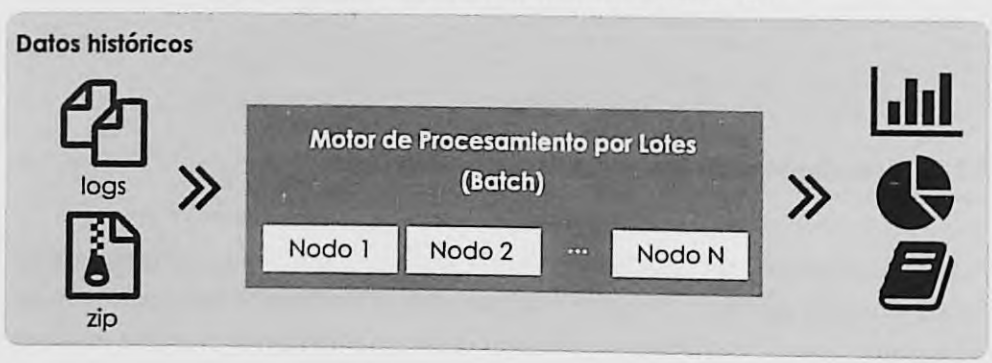
Normalmente, el gran volumen de datos requiere que este procesamiento se lleve a cabo de forma distribuida, lo que además permite agilizarlo, ya que varios nodos pueden estar ejecutando operaciones de forma paralela. No obstante, esto implica que los programas que se desarrollen deben ser capaces de realizar este procesamiento distribuido. Además, como ya ocurriera con el almacenamiento de datos, el hecho de emplear un sistema distribuido introduce nuevas dificultades: ¿qué ocurre si un nodo deja de funcionar mientras está procesando datos? ¿Cómo se agregan los diferentes procesamientos realizados por cada nodo?

Un ejemplo de procesamiento de datos por lotes es el que se produciría si quisiéramos evaluar el comportamiento de un sistema en un periodo de tiempo pasado. Por ejemplo, supongamos que disponemos de un conjunto de *logs* (registros) de un servidor HTTP para nuestro sitio web. Cada una de las líneas en estos *logs* indican el equipo desde el que se establece la conexión, la fecha y la hora, el usuario que está conectado (en caso de que esté disponible), el recurso solicitado al servidor, un código de estado y el tamaño del documento devuelto.

Una compañía podría desear, por ejemplo, realizar un procesamiento periódico de estos *logs* con el fin de extraer analíticas básicas. Al final de cada día se podrían obtener el número de errores producidos, la lista de equipos con más conexiones, la lista de recursos más y menos solicitados, etc. Este procedimiento se podría requerir con diferentes periodicidades, de tal forma que a final de mes se pueden procesar todas las entradas referidas a ese periodo con el fin de obtener un gráfico con las conexiones diarias.

Una ilustración de este ejemplo se muestra en la Figura 62, donde se puede observar cómo un conjunto histórico de datos se procesa para generar una salida. El conjunto de

Figura 62 – Ejemplo de procesamiento por lotes de *logs*



entrada puede contener ficheros con tipos variados: texto, archivos comprimidos, orígenes de bases de datos, etc. El motor de procesamiento por lotes está conformado por un *cluster* con varios nodos, con el fin de paralelizar las tareas.

Si bien las posibilidades son casi infinitas, podemos reconocer una característica: el tiempo de procesamiento no es un factor especialmente apremiante. Esto no significa que la duración de la tarea de procesamiento de datos sea irrelevante: evidentemente si una tarea se ejecuta diariamente no podemos permitirnos que esta se alargue más de un día. Sin embargo, sí es cierto que no es un factor crítico. En muchos casos será indiferente que el proceso tarde varios segundos, unos minutos o incluso una hora. Si realizamos un procesamiento de datos para obtener estadísticas a final de mes, es posible que un proceso que tarda un día en ejecutarse no sea un problema.

En el caso del procesamiento de datos por lotes, el principal desafío al que hay que enfrentarse es el volumen de datos. Los datos históricos pueden conformar un conjunto demasiado grande, y en la mayoría de los casos será necesario distribuir los datos en un *cluster* formado por varios nodos.

En este capítulo veremos el paradigma MapReduce, que fue introducido por Google precisamente para dar respuesta a esta necesidad de procesado de datos por lotes.

7.1.1.2 Procesamiento en tiempo real (*stream*)

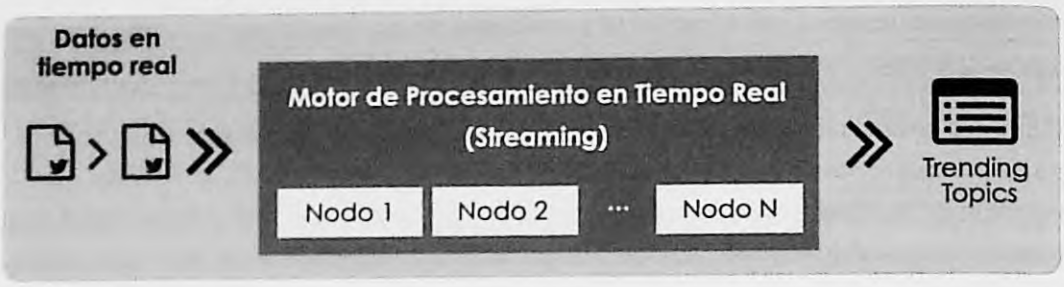
En el procesamiento de datos tiempo real, no solo puede haber un gran volumen de datos, sino que este se genera a una gran velocidad. Evidentemente esto introduce un desafío adicional, ya que el factor tiempo es crucial en el procesamiento de los datos: los datos deben procesarse a medida que van llegando, o en caso contrario se acumularán y el sistema de procesamiento de datos no será funcional.

Evidentemente, este procesamiento también se lleva a cabo de manera distribuida, ya que una sola máquina sería incapaz de recibir los datos y procesarlos a medida que van llegando.

Un ejemplo de este tipo de procesamiento de datos lo podemos encontrar en redes sociales. Un caso muy habitual, y que será familiar para aquellos lectores que sean usuarios de la red social Twitter, es el cálculo de los *trending topics* (temas que son tendencias en un determinado momento).

Para aquellos que no estén familiarizados con esta red social, a continuación se resume su funcionamiento. Twitter es una red social de *microblogging*, donde los usuarios pueden publicar mensajes (tuits) que tengan un máximo de 140 caracteres, y que pueden incluir fotos, vídeos, enlaces, etc. Además, estos mensajes pueden incluir menciones a otros usuarios, que comenzarán con el símbolo «@»; y *hashtags* (etiquetas), que comenzarán con el símbolo «#».

Figura 63 – Ejemplo de procesamiento en tiempo real de tuits



Twitter calcula en tiempo real las tendencias que se producen en cada país o región, o incluso a nivel global. Para ello, detecta las palabras y *hashtags* más repetidas en cada momento.

En la actualidad, en Twitter se publican aproximadamente 500 millones de tuits diarios, habiendo alcanzado el espectacular record de 143.199 tuits en un solo segundo, algo que ocurrió en agosto de 2013[141]. Twitter debe ser capaz de procesar todo este volumen de datos a gran velocidad para poder proporcionar su lista de *trending topics* a cada momento.

La Figura 63 ejemplifica el caso de procesamiento de tuits en tiempo real. Como se puede observar, se introducen numerosas publicaciones en el sistema, que llegan secuencialmente, y el motor de procesamiento en tiempo real debe procesarlos para proporcionar una salida.

En este capítulo veremos la herramienta Storm, precisamente desarrollada por Twitter para poder hacer frente a este problema de procesamiento de datos.

7.2 Procesamiento de Big Data por Lotes

Como ya se explicó en la sección anterior, en el procesamiento de datos por lotes el tiempo de ejecución no suele ser un factor especialmente determinante. Como en el ejemplo de los *logs*, lo habitual será disponer de un gran volumen de datos históricos, que pueden tener diferentes formatos o estructuras, y que se desean procesar con el fin de obtener determinado valor: estadísticas, analíticas y, en general, conclusiones de negocio. No obstante, podemos asumir que este procesamiento tarde minutos, horas o incluso periodos mayores de tiempo en completarse.

7.2.1 El Enfoque de Google: MapReduce

En el capítulo anterior vimos cómo Google introdujo en el año 2003 el sistema de ficheros GFS (Google File System), con el fin de dar solución a sus necesidades de almacenamiento

[141] DMR. *By the Numbers: 150+ Amazing Twitter Statistics*. http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/10/. [Online; consultado el 19 de junio de 2015]. 2015

de Big Data. Esta gran cantidad de datos provenía de los servicios que ofrecía la compañía y que no se limitaban únicamente al motor de búsqueda, sino que incorporaban la búsqueda de imágenes, Google Books, Blogger, etc.

En el año 2004, Google sigue poniendo nuevos servicios a disposición de sus usuarios. Quizás el más representativo surgido ese mismo año es el servicio de correo Gmail, que cuenta desde finales de mayo de 2015 con más de 900 millones de usuarios registrados[142].

Evidentemente, no solo el almacenamiento de los datos supone un desafío para Google. Con GFS, Google es capaz de almacenar datos en un sistema de ficheros distribuido, lo que proporciona escalabilidad horizontal (es decir, se puede lograr más capacidad de almacenamiento añadiendo más nodos de una forma casi lineal) y además, con el fin de evitar los principales inconvenientes de la distribución de ficheros, GFS es tolerante a fallos y dispone de replicación de datos a nivel de bloque.

No obstante, Google no solo necesita almacenar sus ingentes cantidades de datos. Este almacenamiento es el primer paso para, a continuación, realizar un procesamiento de los mismos. Uno de los ejemplos más evidentes de este procesamiento es el que permitiría a Google construir los índices que constituyen la piedra angular de su motor de búsqueda.

Por esta razón, en 2004 Google presenta en una conferencia científica su famoso artículo en el que describe el paradigma de programación MapReduce[143]. Este modelo de programación está pensado para aprovechar las ventajas de GFS, por lo que permite realizar un procesamiento de los datos aprovechando el hecho de que estos se encuentran distribuidos en varias máquinas.

A lo largo de este capítulo se describirá de forma más extensa el paradigma MapReduce y se ilustrará con algunos ejemplos.

7.2.2 El Paradigma MapReduce

MapReduce es un paradigma, es decir, una forma de pensar a la hora de resolver un determinado problema. Como se señaló en la sección anterior, este paradigma surge para dar respuesta a las necesidades de Google de llevar a cabo un procesamiento masivo de sus datos.

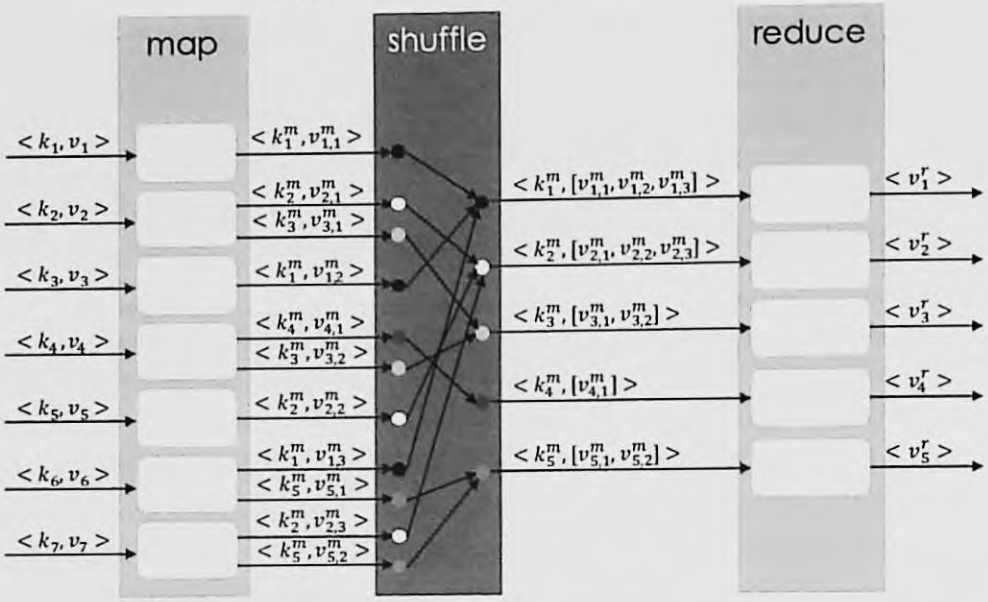
Este procesamiento de datos se lleva a cabo de manera distribuida, de tal forma que las tareas concretas a realizar se dividen en varias subtareas que se reparten entre diferentes nodos de un *cluster*. Cada nodo ejecuta su parte de forma paralela a los demás y finalmente los resultados se agregan.

---

[142] Google. *Gmail now has over 900M users! Thanks for helping us get here.* https://plus.google.com/+Gmail/posts/AjktcDswdKh. [Online; consultado el 19 de junio de 2015]. 2015

[143] Jeffrey Dean y Sanjay Ghemawat. «MapReduce: Simplified Data Processing on Large Clusters». En: *6th Symposium on Operating System Design and Implementation*. 2004

Figura 64 – Ciclo del paradigma MapReduce



Además, una ventaja esencial de este paradigma es que el desarrollador no tiene que preocuparse por la forma en la que las tareas se dividen entre los diferentes nodos del *cluster*, ni tampoco debe prever qué ocurrirá si uno de estos nodos falla. El sistema gestiona automáticamente estos supuestos, de tal forma que el desarrollador puede centrarse en construir los programas que se encargarán de procesar los datos.

El paradigma MapReduce recibe su nombre del hecho de que los programas que emplean este paradigma están constituidos por dos tipos de rutinas o bloques de código que ejecutan una acción determinada: las rutinas *map* y las rutinas *reduce*, que se ejecutan en este orden. Además, entre la rutina *map* y la rutina *reduce* el sistema ejecutará automáticamente una fase denominada *shuffle*. A continuación se describen las características básicas de todos estos componentes, y finalmente en la Figura 64 se muestra de forma genérica el ciclo completo del procesamiento de datos empleando MapReduce.

7.2.2.1 Rutina *map*

La rutina *map* tiene como objetivo realizar un mapeo (*mapping*) de los datos de entrada, es decir, un cambio de su dominio. Fundamentalmente esta rutina servirá para realizar una conversión de la entrada y, si fuera necesario, un filtrado de los datos.

Aunque posteriormente veremos algún ejemplo más concreto, vamos a ilustrar esta definición de la rutina *map* con el caso del análisis de *logs* que ya presentamos en la sección anterior. Para ello, supongamos que queremos calcular el número de recursos que han producido errores o alguna anomalía (es decir, no se han podido entregar correctamente al usuario que los ha solicitado) a lo largo del último mes. Debemos tener en cuenta

que, según el estándar HTTP, el código de estado que se refiere a peticiones satisfechas correctamente es el 200, por lo que los demás códigos de estado indicarán algún error o anomalía.

La entrada a nuestra rutina *map* podría estar formada por ficheros de texto que contengan los *logs* de los últimos días, y además por ficheros comprimidos en ZIP que almacenen los *logs* que tengan más antigüedad.

Nuestra rutina *map* podría realizar las siguientes tareas:

- Para los ficheros comprimidos en ZIP, descomprimirlos y obtener los ficheros de texto que almacena. Esto sería un ejemplo de mapeo o cambio de dominio, puesto que se transforma la entrada.
- Para cada fichero de texto de entrada, eliminar los registros que contienen un código de estado HTTP 200, es decir, en los que no hay ninguna anomalía. Esto sería un ejemplo de filtrado de datos, similar a la sentencia WHERE en SQL.
- Para cada registro válido (indicando una anomalía), devolver únicamente el recurso solicitado y el código de error. Esto constituiría de nuevo un ejemplo de mapeo de datos.

La rutina *map* recibe como entrada una tupla clave-valor ($< k, v >$) y devuelve un conjunto de tuplas clave-valor ($< k^m, v^m >$), de longitud arbitraria m. Esta rutina se ejecutará varias veces, una por cada tupla clave-valor de entrada. Más adelante veremos un ejemplo con más detalle de este funcionamiento.

7.2.2.2 Fase *shuffle*

Esta fase se ejecuta automáticamente y no requiere intervención por parte del desarrollador. En ella, todas las tuplas generadas como salida por las ejecuciones de la rutina *map* se ordenan y se agrupan por clave. De esta manera, de la ejecución de la fase *shuffle* resultarán tuplas que contendrán una clave y un conjunto de valores.

Por ejemplo, si las diferentes ejecuciones de la rutina *map* devolvieron las tuplas clave-valor $< k_i^m, v_{i,1}^m >, < k_i^m, v_{i,2}^m >, \ldots, < k_i^m, v_{i,n}^m >$, tras la fase *shuffle* estas tuplas darán lugar a la tupla $< k_i^m, [v_{i,1}^m, v_{i,2}^m, \ldots, v_{i,n}^m] >$.

La funcionalidad de la fase *shuffle* se asemeja a la instrucción GROUP BY en SQL, y también puede realizar las veces de JOIN, ya que permite agrupar por clave valores procedentes de ficheros o conjuntos de datos distintos. Más adelante veremos ejemplos más concretos de este funcionamiento.

7.2.2.3 Rutina *reduce*

La rutina *reduce* recibe una tupla generada en la fase *shuffle*, por lo que estará recibiendo todos los valores asociados a una misma clave. A continuación, realiza un procesamiento sobre estos valores y finalmente devuelve un valor $< v^r >$.

El objetivo de la fase *reduce* es llevar a cabo una agregación de los datos que han sido mapeados con anterioridad. Esto podría ser equivalente a algunas funciones de agregación en SQL, como SUM, COUNT, AVG, etc.

7.2.3 Ejemplo de MapReduce

En esta sección veremos un par de casos de ejemplo en los que se emplea MapReduce para resolver un problema. El primer ejemplo es un clásico a la hora de explicar el funcionamiento de MapReduce, y consiste en explicar cómo contar palabras en un documento. En el segundo ejemplo, veremos un caso que involucra un conjunto de datos más complejo en el que se realizará un JOIN.

7.2.3.1 Contar palabras con MapReduce

El ejemplo de contar palabras es el «Hola Mundo» del MapReduce, es decir, el ejemplo que se utiliza siempre para explicar el funcionamiento de este paradigma. La descripción del problema a resolver es realmente sencilla: disponemos de un conjunto de documentos de texto, cada uno de los cuales está evidentemente formado por secuencias de palabras. Lo que buscamos es obtener finalmente un listado en el que se indique, para cada palabra que aparezca en estos documentos, el número de ocurrencias de la misma, es decir, el número de veces que esta palabra aparece en estos documentos.

En primer lugar se ejecutará la rutina *map*. Cada una de estas ejecuciones recibirá como entrada una tupla clave-valor, donde la clave será la ruta del documento y el valor será su contenido, es decir, el texto.

La rutina *map* puede ignorar la clave, pues la ruta del documento es irrelevante para nuestro problema. No obstante, sí debe llevar a cabo un procesamiento del valor que recibe como entrada. En concreto, lo que hará la rutina *map* es, en primer lugar, dividir este documento en palabras. A continuación, para cada palabra w devolverá una tupla clave-valor $< w, 1 >$, es decir, un par donde la clave sea la palabra y el valor sea siempre el número 1.

A continuación la fase *shuffle* agrupará todos los valores para cada clave, que en este caso se corresponde con la palabra. Por tanto, el resultado de la fase *shuffle* será un conjunto de tuplas clave-valor, donde la clave sea la palabra y el valor sea una lista de unos, donde cada 1 representa una ocurrencia de esa palabra. Por tanto, este resultado tendrá el siguiente formato: $< w, [1, 1, \ldots, 1] >$. Cada una de estas tuplas resultantes de la fase *shuffle* se introducirán como entrada a una ejecución de la rutina *reduce*, que básicamen-

te se limitará a contar los unos y a devolver un par que contenga la palabra y el número de ocurrencias de la misma: $< w, n_w >$, que es justo lo que buscábamos.

Existen variaciones que se pueden realizar sobre este ejemplo. Por citar una posible variación, la rutina *map* podría almacenar en memoria el número de ocurrencias de cada palabra e imprimirla al final, en lugar de imprimir una tupla $< w, 1 >$ cada vez que lee una palabra. Esta opción generará una salida más reducida, aunque por otro lado requiere más memoria para poder ejecutarse. En este caso habría que realizar una pequeña modificación a la rutina *reduce*, que en lugar de contar el número de unos debería sumar todos los números que recibe como entrada junto con cada palabra.

7.2.3.2 *MovieLens*: valorando películas

A continuación vamos a describir un ejemplo más elaborado que resolveremos utilizando MapReduce. Para ello, vamos a utilizar el conjunto de datos *MovieLens*[144]. Concretamente emplearemos el conjunto denominado «MovieLens 20M», que ha sido lanzado en abril de 2015 y que contiene 20 millones de reseñas a 20 mil películas por parte de 138 mil usuarios.

Si observamos el fichero con la documentación de este conjunto de datos, nos indica que está compuesto por cuatro ficheros. No obstante, nosotros únicamente utilizaremos dos de ellos. El primer fichero que utilizaremos será `movies.csv`, que contiene la siguiente estructura:

```
movieId,title,genres
```

Como se puede observar, este fichero contiene tres campos separados por comas, que son el identificador de la película (`movieId`), el título de la película (`title`) y una lista de géneros (`genres`), separados con el símbolo «|».

Por ejemplo, a continuación se muestran las primeras filas del fichero `movies.csv`:

```
1,Toy Story (1995),Adventure|Animation|Children|Comedy
2,Jumanji (1995),Adventure|Children|Fantasy
3,Grumpier Old Men (1995),Comedy|Romance
4,Waiting to Exhale (1995),Comedy|Drama|Romance
5,Father of the Bride Part II (1995),Comedy
```

El segundo fichero que emplearemos en nuestro ejemplo es `ratings.csv`, que contiene las valoraciones realizadas por los usuarios y se ajusta al siguiente formato:

[144] GroupLens. *MovieLens*. http://grouplens.org/datasets/movielens/. [Online; consultado el 20 de junio de 2015]. 2015

```
userId,movieId,rating,timestamp
```

Como se puede observar, este fichero contiene el identificador del usuario que realiza la reseña (`userId`), el identificador de la película reseñada (`movieId`) que se corresponde con el valor que podíamos encontrar en el fichero `movies.csv`, la valoración del usuario a esa película en una escala entre 0,5 y 5 estrellas, con incrementos de 0,5 (media estrella); y por último la fecha y hora en la que se realizó la reseña en formato UNIX.

A continuación se muestran las primeras filas del fichero `ratings.csv`:

```
1,2,3.5,1112486027
1,29,3.5,1112484676
1,32,3.5,1112484819
1,47,3.5,1112484727
1,50,3.5,1112484580
```

En el problema que vamos a plantear, buscamos obtener como salida final un listado de películas que incluya su título y la valoración media obtenida.

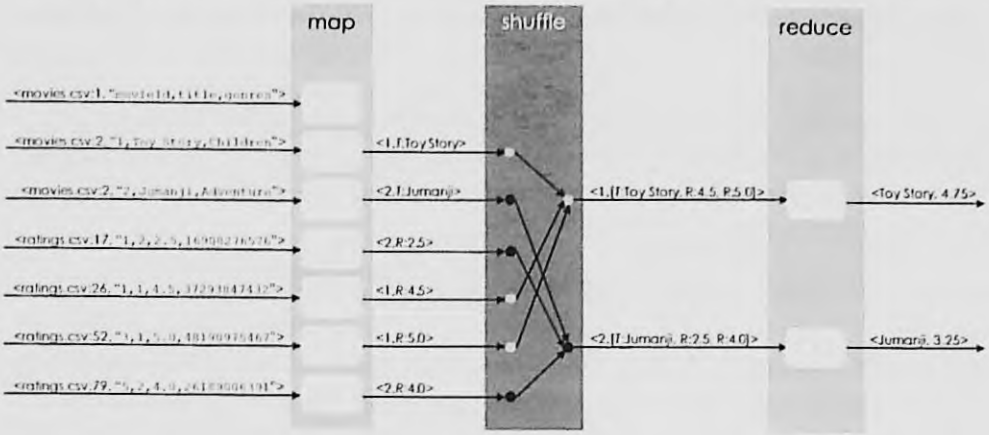
Para ello, la rutina *map* recibirá como valor de entrada una línea de uno de los ficheros. La clave podría ser un valor arbitrario, como el nombre del fichero y el número de línea. Las ejecuciones de la rutina *map* que reciban como entrada la primera línea de los ficheros deberán obviarla, pues esta simplemente indica los campos del fichero CSV y no contienen datos útiles: esto sería un ejemplo de un filtrado.

Con respecto a la funcionalidad principal, las ejecuciones de la rutina *map* que reciban como entrada una línea del fichero `movies.csv` deberán generar como salida una tupla clave-valor, en la que la clave sea el identificador de la película (`movieId`) y el valor sea el título de la película (`title`), junto con una marca que indique que esa tupla hace referencia al fichero de películas. Por otro lado, las ejecuciones de la rutina *map* que reciban una línea del fichero `ratings.csv` deberán producir como salida una tupla clave-valor en la que la clave sea de nuevo el identificador de película y el valor sea la valoración del usuario, indicando con una marca que la tupla hace referencia al fichero de valoraciones.

Como se puede observar, todas las tuplas resultantes tendrán como clave el identificador de la película (`movieId`), independientemente del fichero del que procedan, pues este es el campo común a los dos ficheros. A continuación, la fase *shuffle* agrupará los valores que pertenezcan a la misma clave. Esta es, de hecho, una forma válida de llevar a cabo un JOIN empleando el paradigma MapReduce.

La rutina *reduce* recibirá toda la información referente a una película: su título y el conjunto de valoraciones de los usuarios. Lo que debe hacer es generar como salida una tupla que contenga el título de la película y la media de las valoraciones.

Figura 65 – Solución al problema de *MovieLens* empleando el paradigma MapReduce



En la Figura 65 se muestra gráficamente la resolución conceptual a este problema empleando MapReduce. Los datos utilizados son ficticios, pero iguales en estructura a los reales.

- ***¿Se le ocurre qué procedimiento deberíamos seguir si quisiéramos obtener la valoración media para cada género de película?***

7.2.4 MapReduce en Hadoop

Hasta ahora hemos explicado el paradigma de programación MapReduce, pero no hemos dado detalles de cómo estos programas se ejecutan en un *cluster* de varios nodos.

Como avanzamos en el capítulo anterior, el núcleo de Hadoop está compuesto fundamentalmente de dos herramientas: HDFS y MapReduce. Ambas están integradas de tal forma que se pueda aprovechar el hecho de que los datos estén almacenados de forma distribuida para realizar un procesamiento también distribuido de los mismos.

En concreto, en Hadoop se crean varias tareas *map* (*map tasks*) y varias tareas *reduce* (*reduce tasks*). Estas tareas ejecutarán la rutina *map* y la rutina *reduce* respectivamente, pudiendo ejecutarla una o más veces. Es decir, una tarea *map* o *reduce* puede recibir una o más tuplas clave-valor.

Todo este conjunto de tareas que sirven para procesar ficheros de entrada y generar una salida recibe el nombre trabajo MapReduce (*job*). Es relevante mencionar que tanto la entrada al trabajo MapReduce como la salida que este genere estarán almacenadas en el sistema de ficheros HDFS. No obstante, las salidas temporales de las tareas *map* se almacenarán en el sistema de ficheros local.

- ***¿Se le ocurre por qué la salida de las tareas*** map ***no se almacenan en*** HDFS***?***

La razón de que la salida de las tareas *map* no se almacenen en HDFS es que estos resultados son temporales, y solo tienen valor hasta que se introducen a las tareas *reduce*.

Almacenar estos datos en HDFS incrementaría la carga del sistema de ficheros de forma innecesaria, además de que sometería a replicación unos datos que serán eliminados en el futuro cercano.

Los nodos que forman parte del *cluster* de Hadoop pueden tener diferentes roles a la hora de ejecutar tareas MapReduce. Uno de estos nodos hará las veces de *JobTracker* (rastreador de trabajos) y tendrá la responsabilidad de comprobar que el trabajo MapReduce se completa correctamente.

Para ello, lo primero que hará el *JobTracker* será dividir el trabajo MapReduce en varias tareas *map* y varias tareas *reduce*. El número de tareas *map* suele coincidir con el número de bloques ocupados en HDFS por los ficheros de entrada al trabajo[145]. El número de tareas *reduce* es 1 en muchas ocasiones, aunque se puede configurar por el usuario.

Tras decidir el número de tareas, el *JobTracker* asignará estas tareas a los diferentes *TaskTrackers*, que son otros nodos del *cluster* de Hadoop, que periódicamente informarán al *JobTracker* del estado de la ejecución.

Los *TaskTrackers* normalmente serán también *datanodes* en HDFS, lo que permite aprovechar el llamado «principio de localidad». Este principio consiste en que normalmente las tareas *map* que ejecute un *TaskTracker* recibirán como entrada bloques de HDFS que estén almacenados en esa misma máquina. De este modo, se evita tráfico de red innecesario ya que no se tienen que mover datos de un nodo a otro para ser procesados por la rutina *map*.

No obstante, este mismo principio de localidad no ocurre con las tareas *reduce*, ya que estas tienen que recibir todos los valores asociados a una misma clave, y es posible que se hayan generado valores con la misma clave en tareas *map* ejecutadas en nodos distintos (de hecho, este será un caso muy frecuente, basta pensar en el ejemplo de contar palabras que vimos anteriormente).

Precisamente para evitar un tráfico de red demasiado intensivo en la fase *shuffle*, MapReduce soporta una rutina adicional: la rutina *combine*. Esta operación hace las veces de *reduce* pero se ejecuta a nivel local. De esta manera, se puede realizar en cada *TaskTracker* una primera agregación de los valores generados por las tareas *map* ejecutadas en ese nodo, lo que permite reducir el tráfico de red. Posteriormente, la rutina *reduce* volverá a realizar otra agregación de estos valores previamente agregados por las tareas *combine*.

Normalmente, la rutina *reduce* se podrá reutilizar como rutina *combine* cuando cumpla las propiedades conmutativa y asociativa. Esto es lo que ocurre en el ejemplo de sumar palabras que vimos con anterioridad: cada tarea *map* generará multitud de tuplas que contengan la misma clave con el valor 1. En la fase *shuffle* estas tuplas deben moverse entre los diferentes nodos del *cluster* de tal forma que todas las que contengan la misma

[145] Hadoop Wiki. *How Map and Reduce operations are actually carried out*. http://wiki.apache.org/hadoop/HadoopMapReduce. [Online; consultado el 21 de junio de 2015]. 2015

clave vayan a parar a la misma tarea *reduce*. No obstante, introducir una rutina *combine* permitiría que se realizara la suma de ocurrencias de cada palabra para todas las tuplas generadas por tareas *map* dentro de un mismo nodo. De este modo, en la fase *shuffle* habrá como máximo tantas tuplas con la misma clave como el número de *TaskTrackers*. Además, como la suma cumple las propiedades conmutativa y asociativa, se puede reutilizar la rutina *reduce* como *combine*.

- ***¿Qué cree que ocurrirá si durante la ejecución de una tarea*** map ***falla el*** TaskTracker ***que la está ejecutando?***

Otra responsabilidad del *JobTracker* es asegurarse de que las tareas y finalmente el trabajo se ejecutan correctamente. En caso de que un *TaskTracker* falle, el *JobTracker* deberá crear tareas adicionales y asignarlas a otros *TaskTrackers* que sí estén operativos. En algunas ocasiones, si lo que falla es la ejecución de una tarea *reduce*, es posible que haya que volver a ejecutar tareas *map* de nuevo. Si el *JobTracker* falla, entonces el trabajo MapReduce no se ejecutará.

7.3 Procesamiento de Big Data en Tiempo Real

Cuando se realiza procesamiento de tiempo real (también denominado «en *streaming*»), no solo hay que tratar con un gran volumen de datos, sino que el componente crítico es su velocidad. Por lo general, el sistema de procesamiento de datos recibirá como entrada un flujo de datos que llegará muy rápidamente, y deberá ser capaz de almacenar y procesar estos datos en tiempo real.

7.3.1 El Enfoque de Twitter: Storm

Como mencionamos anteriormente, Twitter es una de las principales fuentes de Big Data en tiempo real, ya que se escriben 500 millones de tuits diarios en la actualidad. Twitter debe ser capaz no solo de almacenar estos datos según se van publicando, sino además de realizar determinados procesamientos de los mismos. Uno de los procesamientos más evidentes es el cálculo de *trending topics*, una lista que se actualiza de forma continua en función de las principales tendencias que se publican en la red social. Esta lista, además, se publica para diferentes regiones: existe una lista global, otra por países e incluso para algunas ciudades.

Storm nació en la empresa BackType, una compañía que desarrollaba productos para obtener analíticas que permitieran a los negocios comprender el impacto de su negocio en las redes sociales, estudiando tanto los datos históricos como los que se iban generando en tiempo real[146].

[146]Nathan Marz. *History of Apache Storm and Lessons Learned*. http://nathanmarz.com/blog/history-of-apache-storm-and-lessons-learned.html. [Online; consultado el 2 de julio de 2015]. 2015

Los desarrolladores de la compañía BackType pronto se dieron cuenta de que sus productos de análisis de datos en tiempo real tenían un enfoque de ingeniería demasiado caótico: tenían que programar todo el sistema distribuido a mano, por lo que la lógica de negocio (la parte de la aplicación que realmente realizaba el análisis de datos) constituía una porción muy pequeña del código de la aplicación, que se dedicaba casi en exclusiva a tener que manejar toda la infraestructura del sistema.

A finales de 2010, Nathan Marz se dio cuenta de que podían simplificar notablemente estas aplicaciones si se consideraba que el flujo de datos era una abstracción distribuida, es decir, que las aplicaciones pueden preocuparse únicamente de los datos sin preocuparse de las máquinas que los almacenan y cómo se distribuyen entre ellas. Concretamente, se le ocurrió que podrían considerarse dos tipos de programas: unos (que denominó «*Spout*») generarían flujos de datos completamente nuevos, mientras que otros (llamados «*Bolt*») se encargarían de recibir como entrada estos flujos de datos, procesarlos y generar otro flujo como salida. Estos programas se ejecutarían de forma inherentemente paralela, tal y como ya hacían las rutinas *map* y *reduce* en Hadoop.

Esto llevó al desarrollo de Storm, un sistema que permitiría a los desarrolladores enfocarse únicamente en el desarrollo de la lógica de los *Spout* y los *Bolt*, sin tener que preocuparse por la gestión del sistema distribuido. Esto es equivalente al caso que ya vimos con MapReduce, donde los programadores solo deben preocuparse por la lógica de la aplicación, resultando completamente transparente para ellos cómo se reparten las tareas entre nodos o lo que ocurre si algún nodo falla.

En mayo de 2011, Twitter inició negociaciones con BackType para adquirir la compañía, siendo Storm, que justo sería presentado durante estas negociaciones, uno de los principales intereses de Twitter por las razones que se comentaron al inicio del capítulo. Además, desde este momento también se comienza a considerar Storm como «el Hadoop del tiempo real», lo que lo coloca a la vanguardia del procesamiento de flujos de Big Data.

El código fuente de Storm fue liberado ese mismo año bajo la licencia *Eclipse Public License*, inmediatamente después de la compra de BlackType por parte de Twitter, convirtiéndose en un éxito al instante. Esto permitió a muchos otros desarrolladores y empresas acceder al producto y contribuir al software conformando la comunidad que ahora sostiene el producto.

Es una plataforma que está programada en Clojure y corre sobre la JVM (*Java Virtual Machine*). Debido a esto se puede programar, en principio, sobre cualquier lenguaje que sea capaz de ejecutarse en la *JVM*, si bien se han ido desarrollando diferentes módulos y paquetes conforme han avanzado las tecnologías para que exista integración con otros lenguajes, como por ejemplo Python.

En la actualidad, Apache Storm es utilizado por numerosas compañías para dar respuesta a sus necesidades de procesamiento masivo de datos en tiempo real[147].

7.3.2 Topología en Storm: *Spouts* y *Bolts*

El principal problema con el que se encuentran todas las plataformas de procesado masivo de datos en tiempo real es la escalabilidad de la solución, es decir, cómo ampliar los recursos para hacer frente a la cantidad de datos crecientes. Una posibilidad consistiría en realizar una escalabilidad vertical, es decir, dotar a nuestro sistema de más recursos: más memoria, más procesador, etc. Sin embargo, esta escalabilidad tiene limitaciones importantes, ya que la cantidad de memoria y la velocidad del procesador están muy limitadas.

Otra opción más adecuada consistiría en realizar una escalabilidad horizontal del sistema, es decir, añadirle más máquinas para que sean capaces de procesar más datos. Al final este enfoque deriva en un problema de tipo productor-consumidor: tenemos una serie de productores (por ejemplo, distintas máquinas) que sirven a un determinado número de consumidores (otro conjunto distinto de máquinas). El principal desafío radica en comunicar esos productores y esos consumidores de manera que los productos se recojan ordenadamente y no se repita el procesado de productos.

7.3.2.1 La solución de Storm

La solución más común a este problema suele ser implementar una cinta (o cola) circular común a todas las maquinas para que todos los consumidores accedan a todos los productos de todos los productores. Evidentemente se añaden una serie de mecanismos de control para garantizar que un producto solo sea procesado por un consumidor, que un productor solo introduce en la cinta si hay espacio, etc. Esta parte suele componer el grueso de todas las implementaciones de cualquier programa de procesamiento masivo de datos en tiempo real.

Gran parte de la generalización del uso de Storm en el procesado masivo de datos, como ya mencionamos anteriormente, se debe a que la plataforma permite al usuario abstraerse de toda la capa encargada de distribuir los datos entre los nodos de procesado. Eso quiere decir que el servicio proporciona una implementación de la cinta y la lógica de productor-consumidor, por lo que el usuario no tiene que preocuparse de ella, permitiéndole centrarse en la lógica de negocio que quiera desarrollar y aumentando así su productividad.

[147] Apache. *Companies Using Apache Storm.* https://storm.apache.org/documentation/Powered-By.html. [Online; consultado el 28 de junio de 2015]. 2015

Además Storm implementa toda una serie de métodos de seguridad para garantizar la estabilidad y escalabilidad del sistema así como la tolerancia a fallos. Para ello se sirve de dos tipos de nodos en su estructura interna: los nodos *Nimbus* y los nodos *Supervisor*.

El nodo *Nimbus* se ocupa de toda la gestión del trabajo, de forma similar a como ya hiciera el *JobTracker* en Hadoop. Cuando se inicia una nueva tarea, este nodo distribuye el código de la aplicación entre los nodos definidos en la topología encargados de ejecutar la tarea (similares por tanto a los *TaskTracker* de Hadoop). Además, sus responsabilidades también incluyen monitorizar el estado del sistema y asignar las tareas a las máquinas. Este nodo es el llamado «nodo maestro» y no ejecuta trabajo, por lo que no los consideramos en nuestro diseño físico cuando desarrollamos aplicaciones para Storm.

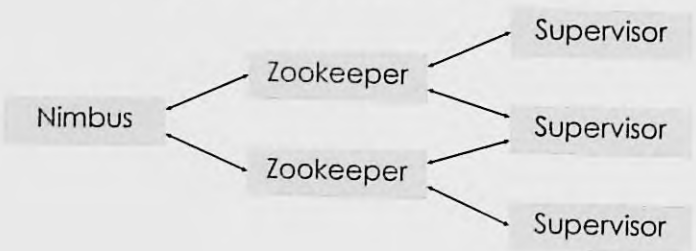
Los nodos *Supervisor* se encargan de iniciar o parar trabajos en funcion de las necesidades y los trabajos asignados por *Nimbus*. Estos nodos son conocidos como nodos esclavos (*Slave nodes*) y son los nodos de los que disponemos a la hora de asignar nodos *Spout* o *Bolt* a nodos físicos.

Además Storm garantiza la tolerancia a fallos incluyendo otro servicio adicional denominado *Zookeeper*, que guarda el estado del trabajo en cada momento para garantizar que si se produce un fallo en un nodo, el trabajo que estaba siendo ejecutado por el mismo se pueda replicar en otro nodo de similares características.

Como podemos observar en la Figura 66, el nodo *Nimbus* y *Supervisor* se relacionan por medio de *Zookeeper*, de manera que en todo momento *Zookeeper* puede guardar el estado, la lista de tareas asignadas a cada nodo, etc.

El lector habrá podido observar muchas similitudes con Hadoop en el comportamiento tolerante a fallos, pero existe una diferencia fundamental: Storm es de hecho más tolerante a fallos que Hadoop. En el capítulo anterior incidimos en que si falla el *JobTracker*, el trabajo MapReduce se cancela y no se puede completar hasta que este nodo vuelva a funcionar. Esto puede ser aceptable en el procesamiento de datos por lotes, donde una tarea puede ejecutarse más adelante con un impacto reducido. Sin embargo, esta opción es inaceptable en una arquitectura de procesamiento de datos en tiempo real, donde cada milisegundo que está cancelada la tarea se pierde información. Por ello, el sistema puede funcionar si el nodo *Nimbus* falla, ya que la gestión de fallos la realiza *Zookeeper*.

Figura 66 – Arquitectura interna de Storm



7.3.2.2 Topologías en Storm

Para indicar a Storm cómo debe asignar y gestionar el trabajo de nuestra lógica de negocio debemos desarrollar una topología siguiendo las indicaciones de Storm. Una topología es esencialmente un grafo (o red) que define las interacciones de los nodos que implementan la lógica de nuestra aplicación (recordemos que Storm se ocupa de toda la gestión del intercambio de datos entre nodos, etc). En la topología disponemos de dos tipos de nodos, denominados *Spouts* y *Bolts*, cuyas funciones veremos con mayor detalle más adelante.

Estos nodos definidos en nuestra topología los conectamos por medio de enlaces de la red en función de las relaciones que deseemos que se formen. Esencialmente los nodos realizarán un procesamiento sobre un flujo de datos (denominado «*stream*»), generando un nuevo flujo que emitirán a través del enlace saliente hacia el siguiente nodo.

En la Figura 67 vemos una topología básica de ejemplo donde tenemos dos nodos de tipo *Spout* y cinco de tipo *Bolt*. El nodo $spout_1$ generará un flujo de datos y lo enviará al nodo $bolt_1$, mientras que el nodo $spout_2$ generará otro flujo de datos, que se transferirá a los nodos $bolt_1$ y $bolt_2$.

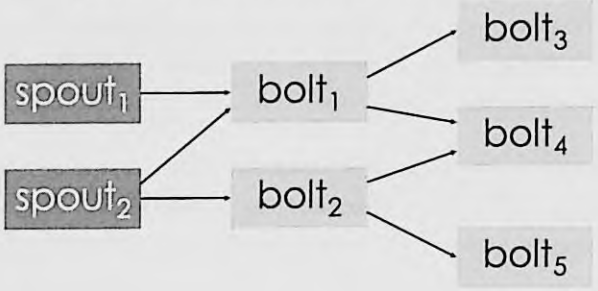
Al mismo tiempo, el nodo $bolt_1$ procesará los datos que recibe de $spout_1$ como entrada, y generará un nuevo flujo de datos como salida que se transferirá a los nodos $bolt_3$ y $bolt_4$, mientras que $bolt_2$ hará lo mismo con los datos recibidos de ambos nodos *Spout* y transferirá su salida a los nodos $bolt_4$ y $bolt_5$.

Los nodos $bolt_3$, $bolt_4$ y $bolt_5$ son nodos terminales que emitirán el resultado obtenido tras toda la secuencia de procesado, o bien lo almacenarán. Como podemos ver, esta topología simple extraerá recursos de dos fuentes distintas de datos, los procesará por medio de los nodos *Bolt* y obtendrá el resultado de tres tareas de procesado diferentes.

7.3.2.3 Nodos *Spout*

Los nodos *Spout* son aquellos encargados de la obtención y el preprocesado de datos. Por un lado deben implementar una rutina encargada de obtener cada tupla de datos y por otro tienen otra encargada de preprocesar esos datos de la manera que consideremos

Figura 67 – Ejemplo de topología básica en Storm



apropiada o incluso simplemente de enviar esa tupla sin procesar a los nodos *Bolt* con los que este conectado. Por convenio este nodo nunca realiza un procesado avanzado de los datos, limitándose en todo caso a realizar un procesado muy simple para filtrar los datos o adecuarlos al formato que utilicemos.

Debido a la función que desempeñan, los nodos *Spout* siempre son nodos de origen, nunca nodos intermedios ni finales en la topología.

7.3.2.4 Nodos *Bolt*

Los nodos *Bolt* son aquellos que se ocupan del procesado intermedio o final de las tuplas de datos. Pueden hacer distintos tipos de procesado: implementación de una función, almacenamiento en una base de datos, filtrado, conversión, etc. En el caso particular de almacenado en una base de datos debe ser de acceso inmediato ya que cualquier demora prolongada en el procesamiento de un nodo *Bolt* puede provocar el agolpamiento de datos en la entrada de ese nodo con la pérdida de datos o de inmediatez que eso supone. En esta misma línea debemos prestar especial atención a los nodos cuya entrada proviene de dos o más nodos ya que son subceptibles de que se formen también cuellos de botella, es decir, de que el flujo de entrada se suceda a una velocidad mayor de la que es capaz de procesar el nodo.

En el caso de que un nodo *Bolt* sea nodo final, el procesado de los datos puede incluir su almacenamiento para servirlos por medio de otra aplicación o para realizar un procesamiento por lotes en el futuro. Por supuesto, se puede implementar cualquier otra lógica que nos interese, pero esta se incluiría en la logica asociada al nodo.

En este caso los nodos *Bolt* sí pueden ser nodos intermedios o finales, pero nunca nodos de origen.

7.3.3 Ejemplo con Storm: Procesado de Tuits

A continuación se plantea un ejemplo para realizar un procesamiento de tuits en tiempo real, mostrando la topología que se debe construir para realizar todo este proceso y las responsabilidades de cada *Spout* y cada *Bolt*.

Supongamos que queremos calcular una sola lista de *trending topics* para Madrid y Barcelona, los dos municipios con más habitantes en España. No obstante, no vamos a tener en cuenta todas las palabras, sino únicamente los *hashtags* o etiquetas, que son aquellos términos que comienzan con el símbolo '#'.

Lo primero que debemos plantear es la topología para nuestro procesamiento de datos. En este caso, hemos optado por emplear una topología que cuenta con dos *Spouts* y dos *Bolts*. Los nodos *Spout* tienen la responsabilidad de iniciar una conexión con Twit-

ter. Twitter ofrece un servicio web denominado *Streaming API*[148], con la que un sistema puede establecer una conexión y recibir como entrada un flujo de datos. Además, este servicio permite recibir un parámetro denominado `locations` que indica el rango de coordenadas de donde se obtendrán los tuits.

Si bien un solo nodo *Spout* podría haberse extraido los tuits de ambas ciudades, se ha optado por utilizar dos para evitar que el *Spout* se sature si el volumen de tuits es grande.

En este flujo de datos, cada tupla producida por Twitter es un documento JSON que incluye, además del contenido del tuit, numerosa información sobre el autor y metadatos de la propia publicación. Nuestro sistema no necesita esta información en absoluto, y son los propios *Spouts* los que pueden encargarse de producir un flujo de datos en el que cada tupla sea únicamente el texto del tuit, ignorando los demás campos. Se puede ver fácilmente la similitud con la rutina *map* en el paradigma MapReduce.

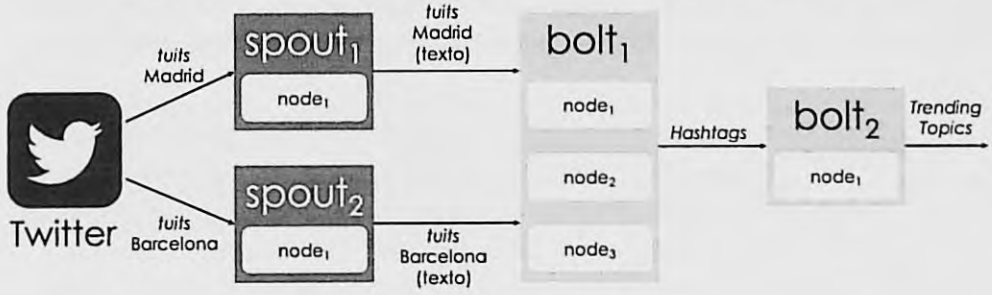
A continuación, esta salida pasará al primer nodo *Bolt*. Este nodo se encargará de extraer de los tuits aquellas palabras que constituyan un *hashtag*, que se generarán como salida. En este caso, como se puede observar, el nodo *Bolt* puede generar cero o varias tuplas de salida por cada tupla de entradas, en función de si el tuit contiene 0, 1 o varios *hashtags*.

Finalmente, el último *Bolt* mantendrá en memoria un contador para cada *hashtag* que indique las veces que ha aparecido en la secuencia de tuits, y devolverá como salida la lista con los n *hashtags* más frecuentes. Esta salida no tiene que producirse necesariamente siempre que se reciba una entrada, sino que podría producirse de forma periódica.

Evidentemente, podríamos pensar que un solo nodo podría llevar a cabo todo este procesamiento. Sin embargo, es importante dividir las tareas todo lo posible, para evitar que los nodos se sobrecarguen de trabajo, ya que en caso contrario se pueden producir cuellos de botella que imposibiliten el procesamiento en tiempo real.

La Figura 68 muestra la topología que resuelve este problema junto con los nodos físicos necesarios para ello.

Figura 68 – Topología y arquitectura física resultantes para procesar tuits



[148]Twitter. *The Streaming APIs*. https://dev.twitter.com/streaming/overview. [Online; consultado el 3 de julio de 2015]. 2015

7.4 Simplificando el Procesamiento de Big Data: Apache Spark

Hasta el momento hemos presentado los dos principales paradigmas en el procesamiento de Big Data: el procesamiento por lotes (*batch*) y el procesamiento en tiempo real (*streaming*).

Para resolver problemas de procesamiento de datos por lotes hemos mostrado el funcionamiento del paradigma MapReduce, que emplea dos rutinas (*map* y *reduce*) con el fin de llevar a cabo este procesado de datos. Respecto al procesamiento de datos en tiempo real hemos empleado Storm para definir una topología que sea capaz de procesar datos de forma escalable.

No obstante, recientemente ha surgido una herramienta que está siendo cada vez más empleada en la comunidad de profesionales del Big Data, ya que permite combinar de forma sencilla procesamiento por lotes y en tiempo real, así como realizar análisis interactivo de datos y ejecutar tareas de aprendizaje automático. Esta herramienta es Apache Spark.

7.4.1 Historia de Spark

El proyecto Spark es relativamente nuevo, ya que fue iniciado por Matei Zaharia en la Universidad de California en Berkeley en 2009, dentro de laboratorio AMPLab.

En 2010 el código de Spark fue liberado y en 2013 pasó a formar parte de la Apache Software Foundation, convirtiéndose en 2014 en proyecto de primer nivel de Apache.

En 2014, Matei Zaharia ganaría el premio ACM a la mejor tesis doctoral por el trabajo de investigación que le llevó al desarrollo de Spark. Además, este mismo año Spark batió un record mundial a la hora de ordenar un conjunto de datos de 100 TB, lográndolo en tan solo 23 minutos[149].

7.4.2 El Ecosistema de Apache Spark

Apache Spark es un proyecto que está compuesto por diferentes herramientas y librerías, que se complementan unas con otras con el fin de ofrecer una solución completa para el procesamiento y análisis de Big Data. Los principales componentes son los siguientes:

- Spark Core: constituye el núcleo de Spark, ofreciendo como principal abstracción el RDD (conjunto de datos distribuido resistente, del inglés *resilient distributed dataset*). Spark Core proporciona numerosas operaciones que se pueden realizar sobre los conjuntos de datos, incluyendo operaciones MapReduce.

[149] Reynold Xin. *Spark Officially Sets a New Record in Large-Scale Sorting*. http://databricks.com/blog/2014/11/05/spark-officially-sets-a-new-record-in-large-scale-sorting.html. [Online; publicado el 5 de noviembre de 2014]. 2014

- Spark Streaming: es una librería de Spark que permite realizar procesamiento de datos en tiempo real. El enfoque empleado por Spark se denomina «*micro-batching*», ya que en realidad lo que hace es realizar el procesado de datos sobre pequeños conjuntos de datos que pertenecen al flujo que se introduce como entrada, agregando posteriormente estos datos procesados con el fin de simular que el procesamiento se lleva a cabo en tiempo real.

- Spark MLib: es una librería de Spark que permite ejecutar tareas de aprendizaje automático. Esto permite aprender modelos a partir de datos que a continuación se pueden emplear para realizar clasificación de datos o recomendación, entre otras muchas cosas.

- Spark SQL: es un módulo de Spark que permite trabajar con datos estructurados, pudiendo hacer consultas SQL sobre los mismos. La funcionalidad de este módulo guarda ciertas similitudes con Hive, ya que permite al usuario realizar consultas SQL que Spark se encarga de traducir a código que se puede ejecutar de forma distribuida.

- Spark GraphX: es una librería de Spark que permite ejecutar operaciones sobre grafos (redes) de forma distribuida, permitiendo por ejemplo calcular el *PageRank* de una serie de sitios web dada su estructura.

Al igual que Hadoop, Apache Spark permite desplegar un *cluster* de varios nodos, lo que permitiría realizar el procesamiento de datos de forma distribuida. Además, Spark también puede integrarse directamente con Hadoop para utilizar el *cluster* ya desplegado, así como acceder a datos de HDFS, HBase, Hive, etc.

Spark promete ser hasta 100 veces más rápido que Hadoop a la hora de realizar un procesamiento de datos. Esto se debe a que Spark realiza en memoria principal todos los cálculos posibles, lo que ahorra accesos a disco duro ganando en eficiencia.

7.5 Procesamiento de Big Data en la Nube

Anteriormente ya presentamos algunas de las opciones de almacenamiento de Big Data en la nube que ofrecen los principales competidores de servicios *cloud*.

No obstante, la variedad de servicios ofrecidos en la nube es inmensa, y no se limitan únicamente al almacenamiento. Todas las compañías mencionadas anteriormente proporcionan servicios que permiten llevar a cabo procesamiento de datos sin necesidad de disponer de una infraestructura propia.

El procesamiento de datos en la nube tiene una serie de importantes ventajas asociadas. Como hemos visto hasta ahora, las tecnologías de procesamiento de Big Data se caracterizan por funcionar sobre un *cluster* de máquinas, proporcionando así escalabilidad horizontal. Esto quiere decir que si el número de datos a procesar se multiplica, nosotros

podríamos multiplicar por el mismo factor el número de nodos de nuestro *cluster* con el fin de que el procesamiento se lleve a cabo en aproximadamente el mismo tiempo.

El principal inconveniente es que normalmente no es factible provisionar con tanta facilidad máquinas físicas en nuestro propio centro de datos. Si quisiéramos, por ejemplo, añadir 100 nodos nuevos a nuestro *cluster*, el tiempo requerido para adquirir las máquinas, recibirlas, instalarlas y configurarlas es por lo general bastante elevado, por lo que no se pueden satisfacer nuestras necesidades de procesamiento con la velocidad deseada. Además, en caso de que el volumen de datos disminuya en el futuro, nos encontraremos con un centro de datos infrautilizado.

El procesamiento en la nube permite provisionar instancias (máquinas virtuales) en pocos minutos, y destruirlas cuando no se necesiten, facturando únicamente por el uso que se haga de ellas. Por este motivo, esta opción resulta especialmente interesante, debido a su flexibilidad y coste.

En esta sección, vamos a mencionar y explicar el funcionamiento de las principales herramientas en la nube que permiten realizar un procesamiento de datos por lotes.

7.5.1 Procesamiento por Lotes

Algunas de las herramientas disponibles en la nube para realizar procesamiento de Big Data permiten llevar a cabo este procesamiento por lotes. La mayoría de estas herramientas lo que hacen en realidad es desplegar una distribución de Hadoop en un *cluster* conformado por varios nodos. El número de nodos y la potencia de los mismos se puede configurar y, normalmente, se puede modificar con el paso del tiempo para adaptarlo a nuestras necesidades.

7.5.1.1 Amazon Elastic MapReduce

Amazon EMR es probablemente una de las primeras soluciones que surgieron para abordar el problema del procesamiento de datos en la nube. Tal y como se ha indicado, esta herramienta lo que hace esencialmente es provisionar un número definido de máquinas y a continuación provisionar Hadoop sobre ellas.

De este modo, al final contaremos con un *cluster* de Hadoop, pudiendo utilizar tanto HDFS como MapReduce (así como otras herramientas de Hadoop) con el fin de llevar a cabo el procesamiento de nuestros datos.

Tras acceder a la consola de Amazon Web Services y escoger el servicio EMR, se mostrará la pantalla de la Figura 69. Para comenzar a emplear esta herramienta, debemos hacer clic en el botón «*Create cluster*».

A continuación se mostrará una pantalla (Figura 70) en la que se podrá configurar el *cluster* de Hadoop que se va a crear. En primer lugar, podremos indicar el nombre del *cluster*,

Figura 69 – Pantalla de inicio de Amazon EMR

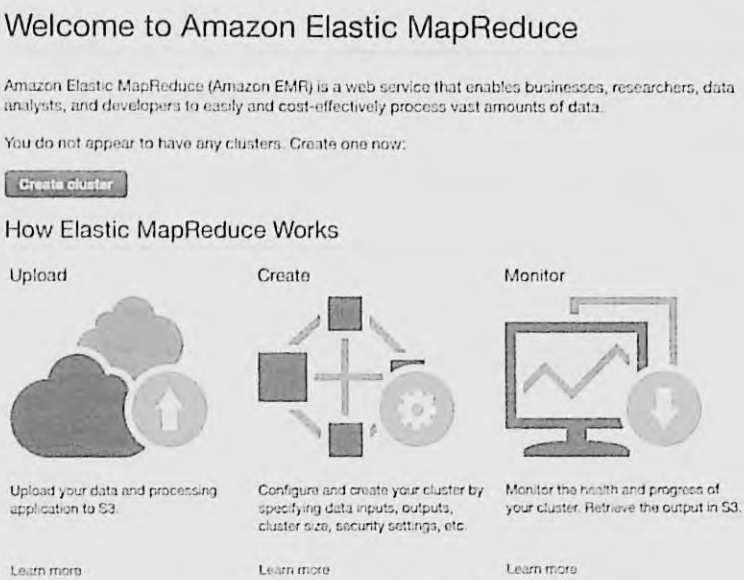
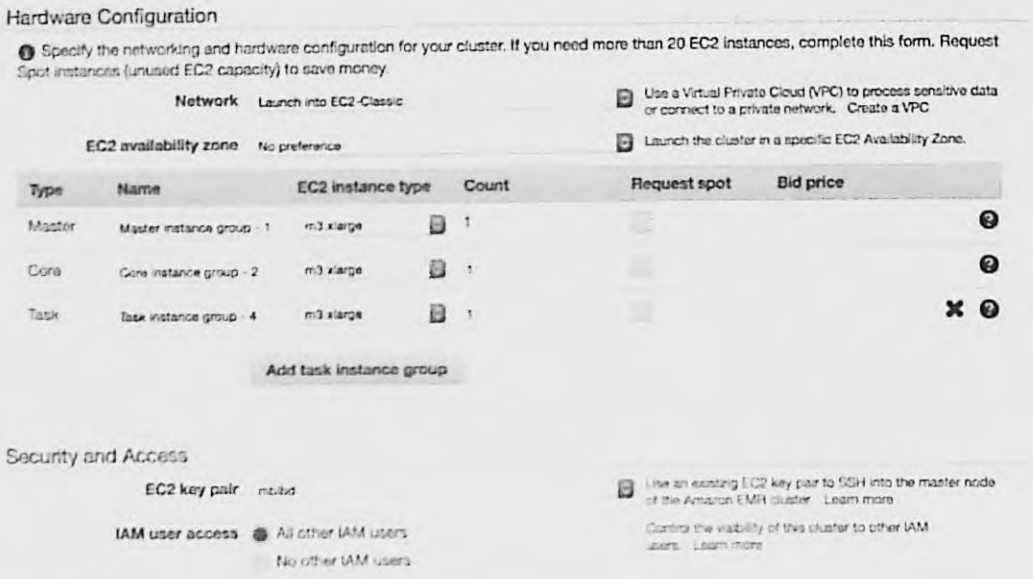


Figura 70 – Configuración general del *cluster* en EMR

Figura 71 – Configuración de software del *cluster* en EMR

configurar la protección de destrucción (que permite evitar que accidentalmente se pueda borrar el *cluster*) y configurar los *logs* (registros).

Figura 72 – Configuración de hardware del *cluster* en EMR



En el siguiente paso se puede configurar el software que se desplegará en el *cluster* de Hadoop. Concretamente, tal y como muestra la Figura 71, se puede escoger la distribución de Hadoop (actualmente están soportadas MapR y la propia de Amazon), la versión del núcleo de Hadoop y las aplicaciones que se instalarán.

Respecto a las aplicaciones, por defecto vienen incluidas Hive, Pig y Hue. Las dos primeras simplifican el desarrollo de aplicaciones MapReduce proporcionando una abstracción basada en SQL y álgebra relacional respectivamente. La última proporciona una interfaz web para simplificar la utilización de Hadoop. Además de estas herramientas, también se pueden seleccionar otras, como Spark o HBase, para que se desplieguen de forma automática en el *cluster* de Hadoop.

A continuación, podremos seleccionar los nodos que formarán parte de nuestro *cluster*, tal y como se muestra en la Figura 72. Estos nodos pueden ser de tipo *master*, *core* o *task*[150].

El nodo *master* es el que lleva a cabo todo el control del *cluster*, incluyendo la monitorización de los demás nodos y la gestión de las tareas. Los nodos *core* son aquellos que sirven tanto para almacenar datos (*datanodes*) como para ejecutar trabajos MapReduce.

Por otro lado, los nodos *task* solo ejecutan trabajos MapReduce, pero no almacenan datos. Estos nodos son por tanto los más adecuados para escalar nuestras necesidades de procesamiento de datos, puesto que pueden ser borrados cuando no hagan falta sin que ello afecte a los datos almacenados.

[150] Amazon. *Elastic MapReduce Developer Guide: Instance Groups.* https://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/InstanceGroups.html. [Online; consultado el 17 de julio de 2015]. 2015

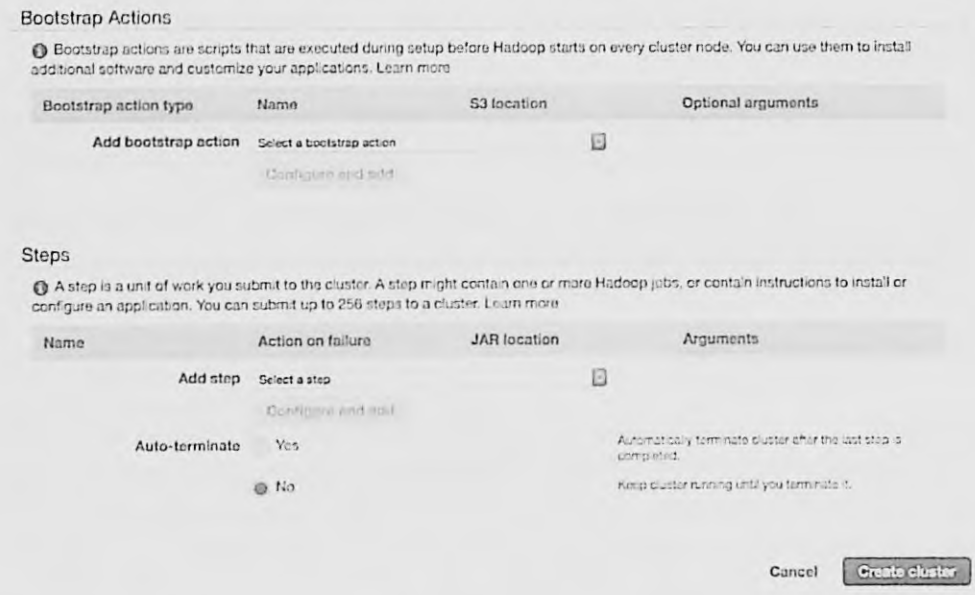
Amazon también proporciona algunas indicaciones a la hora de decidir los datos que se pueden almacenar en un *cluster*. Para calcular el valor aproximado, basta con multiplicar el número de nodos *core* con la capacidad de cada nodo y dividir entre el factor de replicación[151]. Este factor es variable, y por defecto está configurado a 1 si el *cluster* tiene 3 o menos nodos, 2 si tiene entre 4 y 9 nodos y 3 si tiene 10 o más nodos. Como máximo, salvo que lo solicitemos expresamente a Amazon, el máximo número de nodos que podemos asignar a un *cluster* son 20.

Otro paso que debemos seguir antes de crear nuestro *cluster* de Hadoop consiste en crear claves para acceder a los equipos, ya que en caso contrario no podremos conectarnos a ellos. Para crear claves, debemos haber accedido previamente al servicio Amazon EC2 y hacer clic en la opción «*Key Pairs*» en el menú de la derecha.

Después haremos clic en el botón «*Create Key Pair*» y escogeremos un nombre para nuestra clave. Podremos ver que inmediatamente tras su creación se descargará la clave automáticamente en nuestro ordenador. Esta clave que acabamos de crear es la que debemos emplear en la opción «*EC2 key pair*» en el formulario de creación del *cluster*.

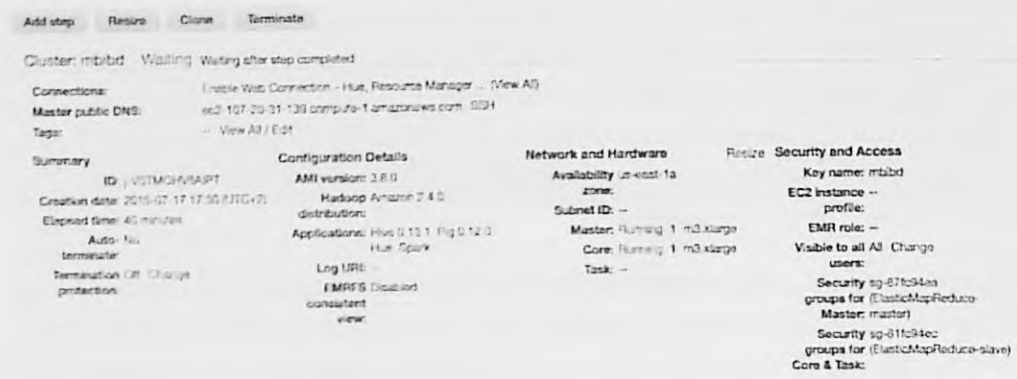
Por último, en las secciones que se muestran en la Figura 73 podríamos definir acciones que se ejecutarán durante la configuración del *cluster* de Hadoop («*Bootstrap Actions*») o después de su creación («*Steps*»). Las primeras hacen posible cambiar la configuración de Hadoop a los parámetros que decidamos. Las segundas permiten ejecutar aplicaciones,

Figura 73 – Configuración de acciones del *cluster* en EMR



[151] Amazon. *Elastic MapReduce Developer Guide: Choose the Number and Type of Virtual Servers.* https://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-plan-instances.html. [Online; consultado el 17 de julio de 2015]. 2015

Figura 74 – Confirmación de creación del *cluster* en EMR



trabajos MapReduce u otros servicios de forma automática tras haber iniciado el *cluster* de Hadoop, sin tener que iniciarlas manualmente.

Para finalizar, haremos clic en el botón «*Create cluster*» para que comience el provisionamiento de las máquinas y el despliegue de Hadoop. Tras hacer clic, se mostrará una pantalla en la que el texto superior indicará el estado «*Starting*» y en la sección «*Network and Hardware*» el estado de las instancias será «*Provisioning*».

Cuando este proceso haya finalizado, se mostrará la pantalla de la Figura 74. En ella, los enlaces «*SSH*» y «*Enable Web Connection*» indican cómo establecer una conexión con el *cluster* para poder utilizarlo. El acceso al *cluster* puede realizarse de dos modos distintos: o bien a través de una consola de comandos, o bien a través de la herramienta Hue, que proporciona una interfaz gráfica para poder trabajar con Hadoop y con sus principales herramientas, como Hive o Pig.

Además, el botón «*Resize*» permite modificar el tamaño del *cluster* añadiendo o eliminando nuevos nodos, en función de la cantidad de datos que deban ser procesados.

Cuando no se vaya a seguir empleando el *cluster*, es importante destruirlo (pulsando el botón «*Terminate*») para evitar que se facture por su uso.

7.5.1.2 Google BigQuery

BigQuery es un servicio de Google que permite realizar análisis de Big Data empleando consultas SQL, obteniendo resultados en poco tiempo. En general, no proporciona una arquitectura tan flexible como la que permite MapReduce, ya que los datos deben ser estructurados, de tal forma que puedan ser almacenados en una tabla.

No obstante, Google BigQuery es una solución muy interesante ya que permite trabajar con datos de igual forma que haríamos con una base de datos relacional (incluso realizando operaciones JOIN entre tablas) pero con volúmenes enormes de datos, que pueden ocupar miles de millones de filas.

Figura 75 – Pantalla de inicio de Google BigQuery

Figura 76 – Esquema del *dataset* Wikipedia en Google BigQuery

Table Details: wikipedia

Schema | Details | Query Table

Schema

title	STRING	REQUIRED	The title of the page, as displayed on the page (not in the URL). Always starts with a capital letter and may begin with a namespace (e.g. "Talk:", "User:", "User Talk:", ...)
id	INTEGER	NULLABLE	A unique ID for the article that was revised. These correspond to the order in which articles were created, except for the first several thousand IDs, which are issued in alphabetical order.
language	STRING	REQUIRED	Empty in the current dataset.

wp_namespace	INTEGER	REQUIRED	Wikipedia segments its pages into namespaces (e.g. "Talk", "User", etc.)
			MEDIA = 202; // =-2 in WP XML, but these values must be >0
			SPECIAL = 201; // =-1 in WP XML, but these values must be >0
			MAIN = 0;
			TALK = 1;

is_redirect	BOOLEAN	NULLABLE	Versions later than ca. 200908 may have a redirection marker in the XML.
revision_id	INTEGER	NULLABLE	These are unique across all revisions to all pages in a particular language and increase with time. Sorting the revisions to a page by revision_id will yield them in chronological order.

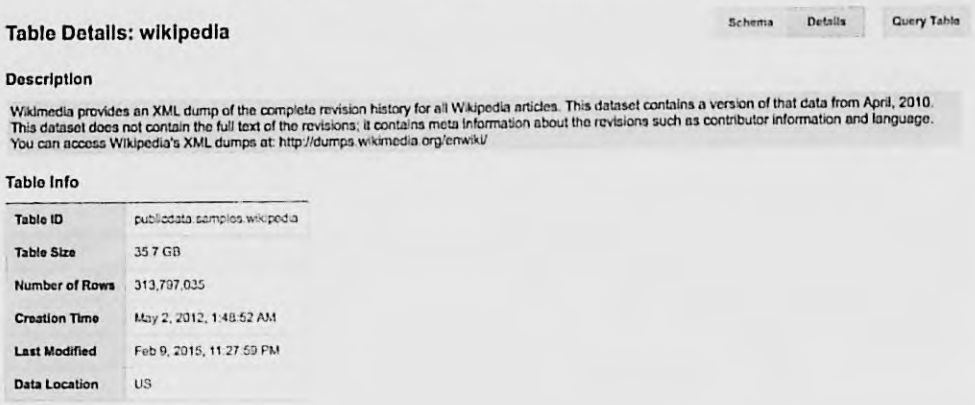
Para comenzar a utilizar BigQuery, accederemos a la consola de Google Cloud y seleccionaremos la opción «BigQuery» dentro de la sección «Big Data» en el menú de la izquierda. Esto cargará una nueva ventana donde se mostrará la pantalla de la Figura 75.

Google BigQuery incluye algunos conjuntos de datos públicos que podremos utilizar con el fin de llevar a cabo un análisis de datos para ilustrar su funcionamiento. Para nuestro ejemplo, vamos a emplear el conjunto de datos de Wikipedia. Este *dataset* contiene el historial completo de revisiones con fecha de abril de 2010. Este historial almacena estadísticas sobre las ediciones para todos los artículos de Wikipedia, incluyendo información sobre la longitud del artículo después de su edición.

Como se observa en la Figura 76, desde Google Query se puede consultar el esquema de la tabla, lo que nos mostrará sus campos, el tipo de cada campo y, en caso de que esté definida, una descripción de los mismos.

Google BigQuery también permite mostrar detalles sobre la tabla con la que vamos a trabajar. En el caso del conjunto de datos de Wikipedia, por ejemplo, podemos ver que la tabla consta de más de 300 millones de filas y ocupa casi 36 GB de almacenamiento, por lo que podemos afirmar que se trata de un conjunto de datos con un volumen importante

Figura 77 – Detalles del *dataset* Wikipedia en Google BigQuery



(aunque Google BigQuery permite trabajar con volúmenes significativamente mayores). Toda esta información se muestra en la Figura 77.

A continuación, vamos a ejecutar una consulta que nos permitirá saber cuáles son los diez artículos más largos de Wikipedia[152]. Para ello, haremos clic en el botón «*Compose Query*», escribiremos la siguiente consulta:

```
SELECT title, num_characters
FROM [publicdata:samples.wikipedia]
WHERE wp_namespace = 0
ORDER BY num_characters DESC LIMIT 10
```

En ella, buscamos obtener el título y el número de caracteres de aquellas entradas que sean de tipo artículo (`wp_namespace = 0`) ordenadas por el número de caracteres de forma descendiente.

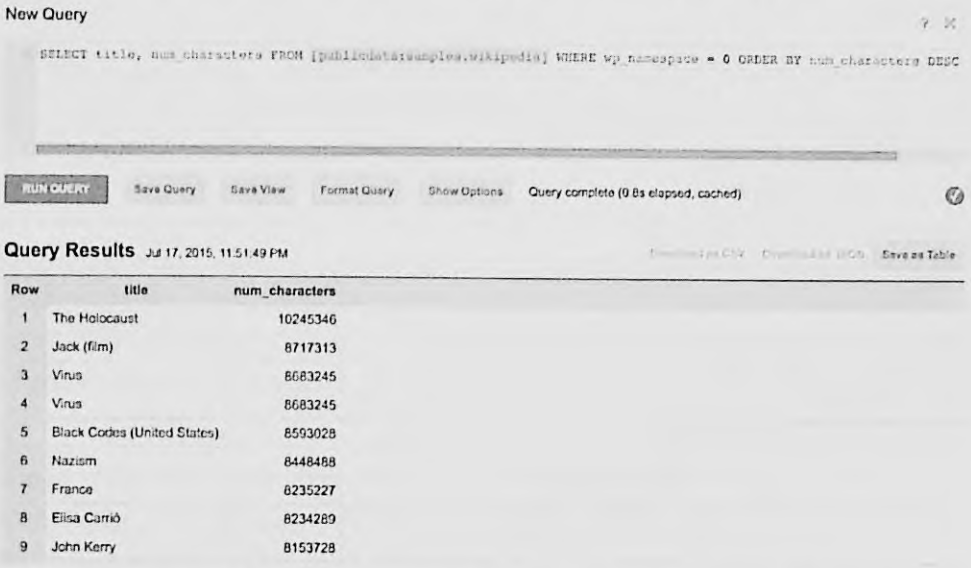
Tras hacer clic en el botón «*Run Query*», se mostrarán los resultados de la consulta (Figura 78). En este caso, se puede observar que el artículo titulado «*The Holocaust*» era el más largo del momento, con más de 10 millones de caracteres. Esta consulta, puesto que ya había sido ejecutada con anterioridad, no ha tardado ni un segundo en devolver una respuesta.

A continuación vamos a ejecutar otra consulta cuyo objetivo es devolver aquellos artículos de Wikipedia que han sido editados más veces:

```
SELECT title, COUNT(*) AS edits
FROM [publicdata:samples.wikipedia]
```

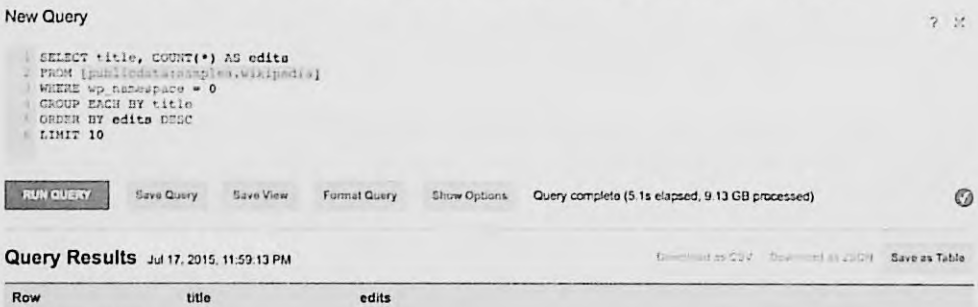
[152] Puesto que los datos son de abril de 2010, el resultado obtenido de esta consulta no se corresponde con la realidad actual.

Figura 78 – Consulta con BigQuery: artículos más largos de Wikipedia



Row	title	num_characters
1	The Holocaust	10245346
2	Jack (film)	8717313
3	Virus	8683245
4	Virus	8683245
5	Black Codes (United States)	8593028
6	Nazism	8448488
7	France	8235227
8	Elisa Carrió	8234289
9	John Kerry	8153728

Figura 79 – Consulta con BigQuery: artículos más populares de Wikipedia



Row	title	edits
1	George W. Bush	43652
2	List of World Wrestling Entertainment employees	30572
3	Wikipedia	29726
4	United States	27433
5	Michael Jackson	23245
6	Jesus	21768
7	Deaths in 2009	20814
8	World War II	20546
9	Britney Spears	20529

```
WHERE wp_namespace = 0
GROUP EACH BY title ORDER BY edits DESC LIMIT 10
```

En este caso agrupamos los artículos por título y contamos las ocurrencias para cada artículo, siendo el número de ocurrencias el número total de ediciones. La Figura 79 muestra los resultados obtenidos tras ejecutar esta consulta. Se puede observar que el artículo editado más veces es aquél que hace referencia al presidente de los Estados Unidos

«George W. Bush». Como curiosidad, este artículo sigue siendo el más editado en la actualidad[153].

Más interesante que los resultados es el tiempo requerido para ejecutar la consulta: tan solo 5 segundos, teniendo que procesar un total de más de 9 GB de datos.

7.5.1.3 Azure Spark

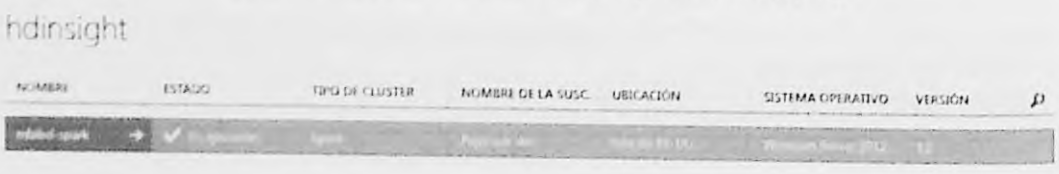
Azure nos permite emplear la plataforma HDInsights para crear un *cluster* de Apache Spark. Para ello, al acceder a HDInsights desde la consola de Azure y seleccionar la opción para crear un nuevo *cluster*, deberemos escoger la opción «Spark» y a continuación introducir los datos que se solicitan: nombre del *cluster*, número de nodos y contraseña de administrador, como se puede ver en la Figura 80.

Tras completar este paso, Azure dará comienzo al proceso de crear el *cluster* de Spark, que podrá llevar varios minutos. A continuación se mostrará la lista de *clusters* de HDInsight (Figura 81), donde podremos ver el *cluster* que acabamos de crear, indicando que es de tipo «Spark».

En la parte inferior de la pantalla podremos ver tres enlaces: «Panel de Spark», «Equipo Portátil Ligero Jupyter» y «Equipo Portátil Ligero Zeppelin». Tras hacer clic en cualquiera

Figura 80 – Panel de creación de un *cluster* de Spark en Azure

Figura 81 – Listado de *clusters* de HDInsight en Azure (Spark)



[153] Wikipedia. *Wikipedia Records*. https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_records#Edits. [Online; consultado el 17 de julio de 2015]. 2015

Figura 82 – *Notebook* Jupyter para Spark en Azure

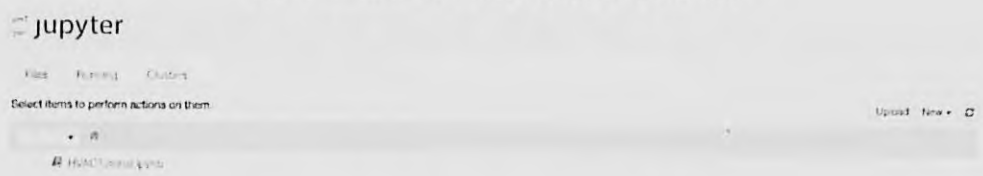
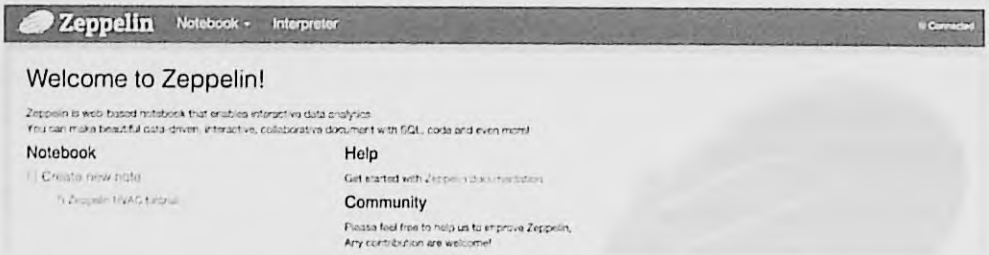


Figura 83 – Hoja en Jupyter para Spark en Azure

Figura 84 – *Notebook* Zeppelin para Spark en Azure



de ellos se nos solicitará el nombre de usuario («admin») y la contraseña que hayamos especificado en el momento de la creación del *cluster*.

El enlace «Equipo Portátil Ligero Jupyter» cargará un *notebook* que permitirá trabajar directamente con Spark (Figura 82). En resumen, lo que esta herramienta permite es ejecutar código (en este caso en Python) para ejecutar trabajos de Spark. En la pantalla que se carga, podemos ejecutar la hoja «*HVACTutorial.ipynb*» para ver cómo funciona. Además, podríamos hacer clic en el desplegable «*New*» para crear una nueva hoja.

Al abrir una hoja, se mostrará la interfaz de la Figura 83, donde podremos editar código Python para conectarnos a Spark con el fin de cargar datos, aplicar transformaciones y ejecutar acciones.

El enlace «Equipo Portátil Ligero Zeppelin» permite algo parecido a Jupyter, pero con una interfaz diferente (Figura 84). Tras hacer clic en el enlace se mostrará una guía de

Figura 85 – Hoja en Zeppelin para Spark en Azure

dos pasos que permitirán acceder al *notebook*. De nuevo, podríamos crear una nueva hoja haciendo clic en el botón «*Create new note*», pero en este caso abriremos la hoja «*Zeppelin HVAC tutorial*».

Tras hacerlo, podremos ver la interfaz de la Figura 85. Como se puede observar, es muy flexible, puesto que permite ejecutar código (en este caso en Scala) para trabajar con Spark, así como ejecutar consultas. Los resultados pueden mostrarse en diferentes formatos, incluyendo gráficas de diversos tipos con el fin de mejorar la visualización y el entendimiento de los datos.

Una vez que hayamos terminado de utilizar nuestro *cluster* de Spark y no vayamos a necesitarlo, deberemos eliminarlo haciendo clic en el enlace «Eliminar» en la parte inferior del listado de *clusters*, para evitar que se siga facturando por él.

8 Análisis de Big Data

8.1 Introducción al Análisis de Big Data

En los dos capítulos anteriores hemos presentado tecnologías para almacenar Big Data y para realizar transformaciones sobre estos datos. No obstante, nos interesa obtener valor de negocio, es decir, extraer información relevante de los datos con el fin de poder comprenderlos y explotarlos.

8.1.1 ¿Qué es el Análisis de Big Data?

Como ya se ha señalado, el análisis de Big Data es el proceso por el que se lleva a cabo un examen de los datos de los que se dispone con el fin de obtener un determinado valor: descubrir patrones ocultos en los datos que no son visibles a simple vista, descubrir tendencias de cambio en algunos aspectos del negocio, entender los principales intereses de los clientes o usuarios, etc.

De nuevo, las 3Vs del Big Data juegan un papel fundamental a la hora de determinar los procedimientos y tecnologías necesarios para llevar a cabo el análisis de los datos. En general, estos desafíos producidos por las características del Big Data han sido señalados en trabajos científicos, como el publicado en 2014 por Fan et al.[154]

[154] Jianqing Fan, Fang Han y Han Liu. «Challenges of Big Data Analysis». En: *National Science Review* 1 (2014), págs. 293-324

8.1.2 Introducción al Aprendizaje Automático

El aprendizaje automático (también denominado *machine learning* en inglés) es un conjunto de técnicas pertenecientes al campo de la inteligencia artificial que permiten descubrir patrones y aprender modelos a partir de los datos.

Algunos ejemplos de aplicaciones de problemas que se pueden solucionar empleando técnicas de aprendizaje automático son las siguientes:

- Detectar automáticamente si un correo electrónico es no deseado (*spam*) a partir de su contenido, basándose en los reportes que han hecho los usuarios sobre correos electrónicos anteriores.
- Predecir si un gasto realizado con una tarjeta de crédito es legítimo o fraudulento, en función del histórico de gastos del portador de la tarjeta.
- Predecir el gasto que va a realizar un usuario en nuestro comercio en función de la información demográfica que disponemos sobre ese usuario.
- Detectar grupos o segmentos de usuarios que tienen intereses similares, de manera que podamos adaptar nuestros servicios y atención a cada uno de estos grupos, mejorando así la percepción de los usuarios.
- Recomendar a un usuario un producto que pueda serle de interés, en función de los productos que sabemos que el usuario ha comprado o de los que ha consultado información.

En los casos de detectar *spam* y predecir fraude en el uso de tarjetas de crédito hablamos de problemas de clasificación. En aprendizaje automático, la clasificación consiste en aprender un modelo a partir de datos que están previamente clasificados o etiquetados, que puede explotarse para predecir la clase de nuevos datos, como pueden ser nuevos correos electrónicos que no sabemos a priori si son *spam*, u operaciones con tarjeta de crédito que en el momento de efectuarse desconocemos si son fraudulentas o no.

El caso de predecir el gasto que va a realizar un usuario es un problema de regresión. Los problemas de regresión son casi idénticos a los de clasificación, con la única diferencia de que lo que se trata de predecir no es una clase, sino un valor numérico continuo.

Otro problema frecuente en aprendizaje automático es el de segmentación, agrupación o *clustering*, como en el caso de detectar grupos de usuarios que hemos mencionado. En este tipo de problemas, los datos no están etiquetados, sino que lo que se busca es agrupar los datos que son similares entre sí. Por último, un problema muy frecuente a la hora de realizar aprendizaje automático sobre Big Data es el de realizar recomendaciones, es decir, tratar de determinar qué *items* interesan a qué usuarios. Los *items* pueden ser productos en un sitio web de comercio electrónico, pero también páginas o incluso otros usuarios en una red social.

8.2 Análisis Predictivo

8.2.1 Introducción al Aprendizaje Supervisado

Se denomina «aprendizaje supervisado» a una rama del aprendizaje automático (*machine learning*) en la que los datos están etiquetados. El propósito de este tipo de tareas de aprendizaje consiste en aprender un modelo a partir de datos que están etiquetados, que conforman el conjunto de entrenamiento. A continuación, el modelo debe poder explotarse para predecir el valor o etiqueta de aquellos datos para los que este valor o etiqueta se desconoce. Se puede encontrar una explicación en vídeo (en inglés) sobre el aprendizaje supervisado en el canal de Youtube de Mathematical Monk[155].

En aprendizaje supervisado se pueden distinguir fundamentalmente dos tareas: la clasificación y la regresión. Ambas son muy similares en su formulación, y difieren en que en la clasificación la etiqueta de los datos es categórica (es decir, es un valor que pertenece a un conjunto de datos acotado), mientras que en la regresión este valor es continuo y puede ser en principio cualquier número real.

En aprendizaje supervisado, se denomina **instancia** a cada uno de los elementos que conforman nuestros datos. Cada una de estas instancias contiene un vector de atributos que representa información sobre la misma, pudiendo cada uno de estos atributos ser numérico o categórico.

Además, cada una de estas instancias contendrá una clase (en problemas de clasificación) o salida (en problemas de regresión). Este valor deberá ser conocido a priori en las instancias que conforman el conjunto de entrenamiento, pero normalmente se desconocerá en otras instancias. Precisamente, el modelo que se ha aprendido automáticamente a partir de los datos se podrá emplear para predecir la clase o salida de una instancia que no esté etiquetada. Por este motivo, estas técnicas se denominan de análisis predictivo, ya que buscan realizar un pronóstico sobre un valor desconocido a partir de un conjunto de datos históricos.

Para ejemplificar estos conceptos, podemos suponer un caso en el que queremos predecir el coste de una vivienda en función de determinadas características de la misma. Cada vivienda de nuestro conjunto de datos sería una instancia y sus atributos serían las características de la misma: su ubicación, año de construcción, tamaño, etc. La clase en este caso sería el coste de la vivienda, que sería el valor que queremos predecir.

Los problemas de clasificación y regresión se pueden formalizar matemáticamente del siguiente modo: disponemos de un conjunto de entrenamiento compuesto por N instancias, siendo la primera (x_1, y_1) y la última (x_N, y_N). x_i hace referencia al vector de atributos de la instancia i, mientras que y_i es el valor de la clase o la salida de la función. Las técnicas de clasificación y regresión lo que buscan es encontrar una función $g : X \rightarrow Y$ donde

[155] Mathematical Monk. *(ML 1.2) What is Supervised Learning?* https://www.youtube.com/watch?v=WpxK_SK2a0. [Online; consultado el 28 de enero de 2016]

X es el espacio de entrada e Y es el espacio de salida. Así, cuando contemos con datos para los que solo dispongamos de su vector de atributos x_j, podremos aplicar la función aprendida con el fin de predecir el valor de la clase o de salida, siendo este $y_j = g(x_j)$.

Idealmente, esta función debe ser capaz de generalizar correctamente, es decir, debe permitir que el modelo aprendido tenga una buena capacidad predictiva sobre aquellos datos que no están etiquetados.

8.2.2 Modelos de Aprendizaje Supervisado

Todos los modelos de aprendizaje supervisado que se pueden aprender deben ajustarse a la definición matemática anterior. No obstante, estos modelos pueden ser de varios tipos dependiendo de cómo sea la función aprendida.

8.2.2.1 Modelos probabilísticos

En los modelos probabilísticos se busca descubrir la distribución de probabilidades que permite inferir el valor de la clase o la salida de la función a partir de los valores de los atributos. Algunos de los modelos probabilísticos que veremos más adelante están basados en la estadística bayesiana.

8.2.2.2 Modelos lógicos

Los modelos lógicos buscan aprender un conjunto de reglas que determinarán el valor de la clase o la salida de la función a partir de preguntas realizadas sobre los atributos. Un ejemplo de este tipo de modelos son los árboles de decisión, que veremos más adelante.

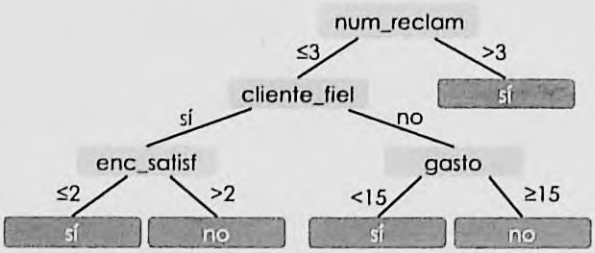
8.2.2.3 Modelos geométricos

En los modelos geométricos, el vector de atributos de cada instancia se representa en un espacio geométrico de M dimensiones, donde M es el número de atributos. A continuación se aprenden hiperplanos que separan regiones del espacio en los que hay valores de la clase distintos. A continuación, cuando vayamos a predecir la clase de una instancia, bastará con determinar su posición en el espacio y la clase que corresponde a esa posición.

8.2.3 Técnicas de Clasificación

A continuación describiremos algunas de las técnicas que permiten aprender modelos de clasificación de los tipos descritos anteriormente. Además de las aquí incluidas, existen otras técnicas tales como las redes de neuronas artificiales o las máquinas de soporte vectorial, que no se describirán en el capítulo.

Figura 86 – Ejemplo de árbol de decisión



8.2.3.1 Árboles de decisión

Un árbol de decisión es un modelo lógico que permite inferir la clase de una instancia mediante la realización de preguntas sobre sus atributos.

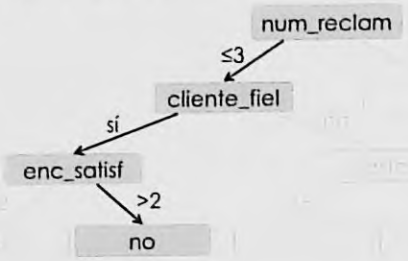
En el árbol, se denomina «raíz» a la pregunta que se realiza en primer lugar. Cada nodo del árbol plantea una cuestión sobre uno de los atributos, mientras que las ramas del árbol son las encargadas de plantear las diferentes respuestas. En cada iteración, se realiza la cuestión indicada en el nodo actual y se escoge la rama que corresponda con la respuesta correcta, llegando a un nuevo nodo. Finalmente, los denominados como «nodos hoja» ya no contienen más preguntas, sino la clase que debe ser devuelta como predicción.

Por ejemplo, la Figura 86 muestra un ejemplo de árbol de decisión que se podría utilizar para predecir si un cliente va a abandonar nuestra empresa de telefonía. Los árboles de decisión son equivalentes a sistemas de reglas, y en este caso concreto, la lógica que sigue este árbol de ejemplo es la siguiente:

1. Si el cliente ha puesto más de 3 reclamaciones, entonces abandonará la compañía.
2. Si el cliente ha puesto 3 o menos reclamaciones, entonces:
    a) Si el cliente pertenece a un programa de fidelización, entonces:
        1) Si ha indicado una nota menor o igual a 2 en la encuesta de satisfacción, entonces abandonará la compañía.
        2) Si ha indicado una nota superior a 2 en la encuesta de satisfacción, entonces permanecerá en la compañía.
    b) Si el cliente no pertenece a un programa de fidelización, entonces:
        1) Si el cliente tiene un gasto mensual inferior a 15 euros, entonces abandonará la compañía.
        2) Si el cliente tiene un gasto mensual igual o mayor a 15 euros, entonces permanecerá en la compañía.

A continuación se muestra una instancia de ejemplo con unos valores determinados para un cliente de la compañía que ha puesto una reclamación, es un cliente fiel, ha manifestado una satisfacción de 4 puntos y tiene un gasto mensual de 31 euros:

Figura 87 – Ejemplo de explotación del árbol de decisión aprendido



num_recl	cliente_fiel	enc_satisf	gasto
1	sí	4	31

Partiendo del árbol de decisión de la Figura 86, podríamos concluir el futuro del usuario en la compañía. Para ello, bastaría con seguir el recorrido respondiendo las preguntas desde la raíz hasta llegar a una hoja. Este proceso se muestra en la Figura 87, concluyendo que el usuario no abandonará la compañía.

8.2.3.2 Naïve Bayes

Naïve Bayes es una técnica de aprendizaje automático supervisado que permite aprender modelos probabilísticos que se pueden explotar para realizar una predicción de clase. Además, en este caso no solo se realizará esta predicción sino que se podrá obtener la probabilidad para cada una de las clases.

Naïve Bayes está basado en el teorema de Bayes, que describe la probabilidad condicional de un evento A dado B en términos de la probabilidad condicional del evento B dado A y la probabilidad a priori de A, de acuerdo con la siguiente fórmula:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Este teorema es de especial interés porque permite obtener la probabilidad de A dado B conociendo la probabilidad de B dado A. Veamos este caso con un ejemplo práctico: imaginemos que sabemos la probabilidad de que aparezca la palabra «viagra» en un correo electrónico no deseado y también en un correo electrónico cualquiera; sabiendo además la probabilidad de que un correo electrónico sea no deseado. Todo esto nos permitiría saber la probabilidad de que un correo electrónico sea *spam* sabiendo únicamente que contiene la palabra «viagra».

Por ejemplo, supongamos que la probabilidad de que un correo electrónico sea *spam* es $P(\text{spam}) = 0,7$, la probabilidad de que la palabra «viagra» aparezca en un correo electrónico cualquiera es de $P(\text{viagra}) = 0,1$; y la probabilidad de que la palabra «viagra» aparezca en un correo electrónico no deseado es $P(\text{viagra}|\text{spam}) = 0,6$. Con esta infor-

mación, podemos calcular la probabilidad de que un correo electrónico sea no deseado sabiendo que en él aparece la palabra viagra:

$$P(\text{spam}|\text{viagra}) = \frac{P(\text{viagra}|\text{spam}) \times P(\text{spam})}{P(\text{viagra})} = \frac{0,6 \times 0,1}{0,7} = 0,086$$

Naïve Bayes se basa en este teorema para entrenar un modelo probabilístico a partir de los datos de entrenamiento. Además, Naïve Bayes funciona bajo la hipótesis de que los atributos son independientes entre sí. Esto quiere decir que cada atributo contribuye de manera independiente a la clase.

Se puede encontrar una explicación en vídeo (en inglés) sobre la clasificación empleando Naïve Bayes y sus propiedades matemáticas en el canal de Youtube de Mathematical Monk[156].

8.2.3.3 k-Nearest Neighbors

k-NN (*k-Nearest Neigbors*, los k vecinos más cercanos) es una técnica de clasificación de las llamadas «vagas» (o «*lazy*» en inglés), ya que no existe realmente una fase de entrenamiento del modelo.

Esta técnica de clasificación tiene un fundamento geométrico, puesto que su funcionamiento se basa en determinar la clase de una instancia simplemente observando la clase de las k instancias que están más próximas a ella en un espacio multidimensional, con tantas dimensiones como atributos tienen las instancias. Se puede encontrar una explicación en vídeo (en inglés) sobre la clasificación empleando *k-Nearest Neigbors* en el canal de Youtube de Mathematical Monk[157].

Este procedimiento requiere la elección de una medida de distancia, que es la que se encarga de decidir cuáles son los vecinos más cercanos. Algunas medidas de distancia típicas son la euclídea, la Manhattan, la coseno, etc.

A continuación se muestra una serie de puntos en un espacio bidimensional:

x	y	class	x	y	class
2	3	+	1	1	x
3	1	+	1	4	x
2	2	+	4	4	x
5	2	+	3	2	x

[156] Mathematical Monk. *(ML 8.1) Naive Bayes Classification.* https://www.youtube.com/watch?v=8yvBqhm92xA. [Online; consultado el 28 de enero de 2016]

[157] Mathematical Monk. *(ML 1.6) K-Nearest Neighbor Classification Algorithm.* https://www.youtube.com/watch?v=4ObVzTuFivY. [Online; consultado el 28 de enero de 2016]

Figura 88 – Ubicación de las instancias

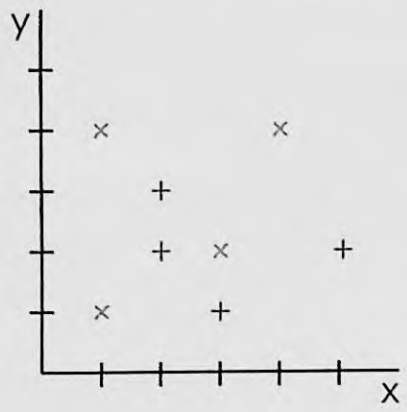
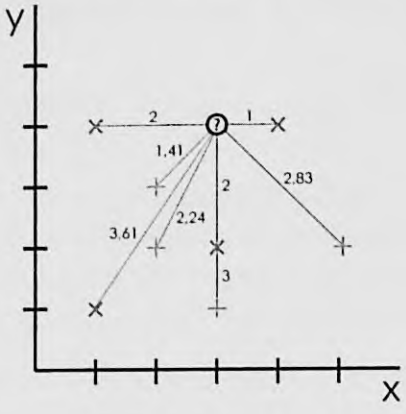


Figura 89 – Clasificación de una instancia



Se ha optado por limitar el problema a dos dimensiones por facilitar la visualización, aunque el procedimiento es extensible a un número arbitrario de dimensiones.

La Figura 88 muestra la situación de cada una de las instancias de ejemplo en el plano. Supongamos que a continuación nos llega una nueva instancia cuya clase es desconocida y queremos predecirla:

```
x     y     class
3     5       ?
```

En primer lugar, debemos definir una medida de distancia que será la que determine cómo de cerca o lejos están dos puntos en el plano. En este caso hemos optado por una distancia euclídea, explicada en la sección 8.3.2.1. Además, también debemos optar por un valor para el parámetro k, que en este caso hemos optado por dejar a 1. En caso de usar un valor de k mayor que 1, se tomará la clase más frecuente de entre todas las de los vecinos más cercanos.

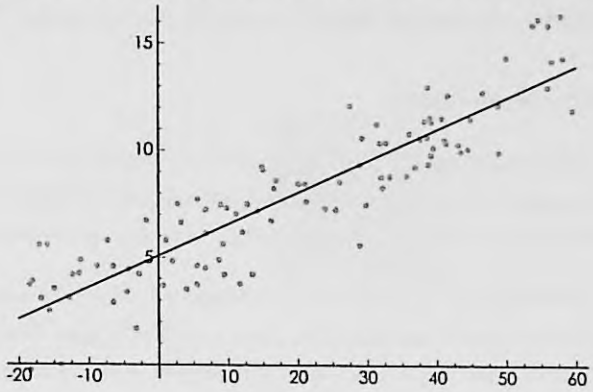
La Figura 89 muestra las distancias desde la nueva instancia hasta todas las instancias del conjunto de entrenamiento. En este caso, se puede observar que el vecino más cercano tiene distancia 1 y es de la clase «x», por lo que este será el resultado de la predicción.

Uno de los inconvenientes de esta técnica es que cuando el conjunto de entrenamiento consta de muchas instancias, el tiempo para realizar una predicción es elevado, puesto que se debe calcular la distancia entre la instancia cuya clase se quiere predecir y todas las instancias del conjunto de entrenamiento.

8.2.4 Técnicas de Regresión

Como ya hemos explicado anteriormente, un problema de regresión es aquél en el que se busca predecir un valor numérico en función de los valores de los atributos de una

Figura 90 – Regresión lineal de un conjunto de instancias en un plano



instancia. Para ello, se busca encontrar una función $g : X \rightarrow Y$ donde X es el espacio de entrada e Y es el espacio de salida.

8.2.4.1 Regresión lineal

La regresión lineal ya se introdujo anteriormente, donde vimos gráficamente como un conjunto de puntos en un plano de dos dimensiones podía aproximarse mediante una recta, tal y como se muestra en la Figura 90. No obstante, de un modo más general, la regresión lineal permite aprender un hiperplano que sea capaz de aproximar puntos en un espacio de M dimensiones.

8.2.4.2 Árboles de regresión

Los árboles de regresión son modelos para resolver problemas de regresión. Son idénticos en su funcionamiento a los árboles de decisión que hemos visto anteriormente, con la principal diferencia de que las hojas no contienen una clase, sino un valor numérico o una función que devuelve un valor numérico.

El concepto de árbol de regresión fue introducido por Breiman en 1984[158], cuando presentó CART. En CART, en cada iteración se escogía aquel atributo que redujera la varianza de los datos, un procedimiento que contrasta con los árboles de decisión en que no se emplea el concepto de ganancia de información, puesto que no se dispone de un valor categórico en la clase. Cuando la varianza en el conjunto de datos es suficientemente pequeña, se crea un nodo hoja que contiene como valor la media de los valores de los datos.

[158] L. Breiman y col. *Classification and Regression Trees*. Wadsworth y Brooks, 1984

Se puede encontrar una explicación en vídeo (en inglés) sobre la construcción de árboles de regresión con CART en el canal de Youtube de Mathematical Monk[159].

8.2.5 Evaluación de Modelos

Los modelos de clasificación y regresión pueden padecer de sobreajuste, es decir, pueden aprenderse demasiado bien las instancias del conjunto de entrenamiento, pero fallando luego cuando se introduzcan en el modelo las instancias cuya clase se desconoce.

Por ello, cuando construimos un modelo de clasificación o regresión normalmente queremos saber qué tal funcionará cuando queramos explotarlo para predecir el valor de una clase con instancias para la que esta no es conocida. Una aproximación para evaluar el funcionamiento de un modelo de clasificación o regresión pasa por ejecutar los siguientes pasos:

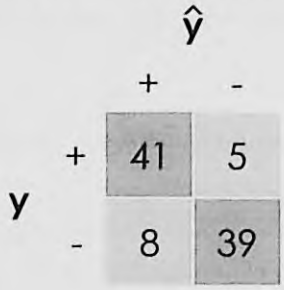
1. Aleatorizar el orden de las instancias, para evitar que la división en conjuntos de entrenamiento y test que realizaremos en el siguiente paso esté sesgada por la ordenación de las mismas.
2. Dividir el conjunto de datos etiquetados en dos subconjuntos: uno de entrenamiento y el otro de test. Normalmente se debe decidir el tamaño de cada uno de los conjuntos. Aunque no existe una regla de oro para decidir cuántas instancias asignar a cada uno de ellos, es frecuente ver problemas en los que se emplea un 70 % de los datos para entrenamiento y un 30 % de los datos para test.
3. Emplear el conjunto de entrenamiento para aprender un modelo de clasificación o regresión.
4. Emplear las instancias del conjunto de test para explotar el modelo. El modelo nos devolverá una predicción para cada instancia, que será un valor categórico o numérico que podremos comparar con el valor real de la clase de dicha instancia.

Al finalizar el proceso, contaremos con un conjunto de instancias para los que conocemos el valor real de la clase y el valor predicho. Con estos valores podremos calcular diversas medidas de calidad de los modelos.

En problemas de regresión, se suelen emplear medidas de error, de tal forma que un modelo es mejor que otro si esta medida de error es menor en el primero de los dos. Existen diversas medidas, pero la más común es el error cuadrático medio, que se calcula según la siguiente fórmula:

[159] Mathematical Monk. *(ML 2.2) Regression Trees (CART).* https://www.youtube.com/watch?v=zvLOpbgtW3c. [Online; consultado el 28 de enero de 2016]

Figura 91 – Ejemplo de matriz de confusión para un problema de clasificación binario



$$e = \frac{\sum_{j=1}^{N} \left(\hat{y}^{(j)} - y^{(j)}\right)^2}{N}$$

Recordemos que $\hat{y}$ es el valor de la predicción mientras que y es el valor real de la clase.

En problemas de clasificación, una medida que nos puede indicar la calidad del modelo es la tasa de aciertos. Esta indica el porcentaje de instancias del conjunto de test para los que la clase se ha predicho correctamente.

No obstante, en problemas de clasificación podemos construir lo que se denomina «matriz de confusión», que es una tabla como la que se muestra en la Figura 91. Si hay K clases, la matriz tendrá dimensión $K \times K$. En esta matriz, las columnas indican dónde se ha clasificado cada instancia (la clase predicha) y las filas indican las clases reales.

Los valores pertenecientes a la diagonal principal de la matriz son los aciertos, es decir, aquellas instancias para las que la clase real y la predicha coinciden. En este ejemplo, la tasa de aciertos de la clasificación sería $\frac{41+39}{41+39+5+8} = 86,02\,\%$.

Además, podemos definir algunas de las métricas que se pueden obtener a partir de la matriz de confusión:

- El número de verdaderos positivos (TP) es el número de instancias de la clase «+» que han sido clasificadas correctamente. En este caso hay 41 verdaderos positivos.

- El número de verdaderos negativos (TN) es el número de instancias de la clase «-» que han sido clasificados correctamente. En este caso hay 39 verdaderos negativos.

- El número de falsos positivos (FP) es el número de instancias de la clase «-» que han sido clasificados incorrectamente como de la clase «+». En este caso hay 8 falsos positivos.

- El número de falsos negativos (FN) es el número de instancias de la clase «+» que han sido clasificados incorrectamente como de la clase «-». En este caso hay 5 falsos negativos.

A partir de estas métricas podemos definir nuevas medidas de calidad:

- La precisión (en inglés «*precision*») se define como el ratio entre el número de verdaderos positivos y el número de instancias clasificadas como positivas, es decir $\frac{TP}{TP+FP}$. En este caso, el valor de la precisión es de $\frac{41}{41+8} = 83,67\,\%$.
- La sensibilidad o exhaustividad (en inglés, «*sensitivity*» o «*recall*») se define como el ratio entre el número de verdaderos positivos y el número de instancias que son en realidad positivas, es decir $\frac{TP}{TP+FN}$. En este caso, el valor de la precisión es de $\frac{41}{41+5} = 89,61\,\%$.
- La puntuación F1 (en inglés «*F1 score*» o «*F measure*») es una medida que combina la precisión y la exhaustividad en un único valor, calculado como su media armónica. Así, la puntuación F1 se calcula como el producto de ambas medidas entre la suma de ambas medidas, es decir: $F1 = 2\frac{precision \times recall}{precision + recall}$. En este caso, el valor de F1 es igual a $2\frac{0,8367 \times 0,8961}{0,8367 + 0,8961} = 86,54\,\%$.

8.2.6 Aplicaciones de las Analíticas Predictivas

En el ámbito del Big Data, son numerosas las aplicaciones prácticas que se pueden realizar a partir de las técnicas de analíticas predictivas. En esta sección se enumeran algunas de estas aplicaciones, donde cabe destacar que suelen estar sujetas a la disponibilidad de un gran volumen de datos y, en algunos casos, a una gran velocidad de entrada de los mismos.

8.2.6.1 Detección de fraude

Esta aplicación se da especialmente en dos escenarios distintos: la detección de transacciones fraudulentas con tarjeta de crédito (por ejemplo, cuando esta ha sido sustraída) y la detección de reportes fraudulentos de siniestros a una compañía aseguradora. No obstante, este campo de aplicación se está extendiendo a nuevos dominios, como la declaración fraudulenta de ingresos y propiedades para evadir impuestos o el uso ilegítimo de un sistema (por ejemplo, mediante la suplantación de identidad de un usuario).

En cualquiera de estos casos, el problema a resolver es de clasificación binaria, donde la clase indica si la instancia es legítima o fraudulenta. Normalmente la cantidad de datos históricos (transacciones de tarjeta de crédito, siniestros dados de alta, etc.) es muy elevada y, además, en este dominio suele primar la velocidad a la hora de realizar el análisis predictivo. Por ejemplo, en el caso de una compañía de servicios financieros, puede ser

necesario que el resultado de la predicción sobre una transacción se obtenga en milisegundos, de tal forma que se pueda aprobar o rechazar automáticamente el gasto.

8.2.6.2 Diagnóstico médico

Es frecuente emplear modelos predictivos con el fin de realizar un diagnóstico del riesgo de que se dé una determinada enfermedad dado el historial clínico de un paciente. Normalmente estos sistemas sirven como soporte a la toma de decisiones del personal médico, por lo que no sustituyen su criterio de expertos. Estos sistemas también suelen ser de clasificación binaria, donde la clase indica si se desarrollará o no una determinada enfermedad.

Normalmente, los modelos probabilísticos en los que se devuelve un valor continuo entre 0 y 1 suelen ser adecuados, pues permiten devolver una medida de certeza del sistema sobre el hecho de que el paciente pueda desarrollar o no la enfermedad.

Otra característica de estos problemas es que es posible que no todos los atributos estén disponibles, y normalmente cada atributo tiene un coste distinto, ya que existen pruebas médicas que son costosas (en términos económicos o de intrusión para el paciente) y que por tanto intentan evitarse salvo que sea imprescindible practicarlas.

8.2.6.3 Clasificación de imágenes astronómicas

En el ámbito científico, se emplean técnicas de aprendizaje automático para realizar clasificación de imágenes capturadas mediante telescopios. Normalmente, estas imágenes se toman con cierta frecuencia y tienen grandes resoluciones, lo que da lugar a conjuntos de datos muy grandes, imposibles de procesar manualmente.

Por este motivo, los humanos etiquetan de forma manual algunas imágenes indicando si en ellas aparece un astro o no. A continuación, se entrena un modelo de clasificación con el fin de que sea el clasificador el que determine si aparecen astros relevantes en las demás imágenes tomadas del espacio.

8.2.6.4 Predicción del rendimiento de anuncios

La publicidad online es una práctica extendida a lo largo de todos los sectores de la industria, y supone un importante flujo de ingresos. En esta tarea entran en juego tres agentes:

- El anunciante tiene un contenido que quiere publicitar: un servicio, un producto, un sitio web, etc.

- El usuario final es aquel agente que puede mostrar interés en el contenido anunciado, en cuyo caso podrá pinchar en el anuncio.
- La agencia de publicidad se encarga de mostrar los anuncios en sitios web relevantes y a usuarios finales que puedan estar interesados.

La idea es maximizar el CTR o «*click-through rate*», que es el porcentaje de clics realizados en un anuncio sobre el total de veces que el anuncio se ha mostrado, es decir, sus impresiones.

La predicción del rendimiento de un anuncio se puede enfocar de dos maneras distintas. Por un lado, se puede considerar un problema de clasificación binaria, donde se busca predecir si un usuario hará clic en un anuncio que se le muestre, dadas las características del anuncio y las del usario. Por otro lado, puede enfocarse como un problema de regresión, donde se busca predecir el CTR de un anuncio dadas sus características.

8.2.6.5 Predicción de abandono

Otro problema frecuente consiste en predecir si un determinado cliente va a abandonar una compañía o no, dadas sus características, las características de los productos o servicios contratados y el histórico del comportamiento de este usuario. Esta aplicación también se denomina «predicción del *churn*», y es muy frecuente sobre todo en el ámbito de las telecomunicaciones.

Esta tarea se puede representar de nuevo como un problema de clasificación binaria, donde la clase indica si el usuario permanecerá en la compañía o la abandonará.

8.3 Análisis de Patrones y Recomendación

8.3.1 Introducción al Aprendizaje No Supervisado

En la sección anterior proporcionamos una definición de lo que se conoce como «aprendizaje supervisado», indicando que es aquél en el que las instancias están etiquetadas con una «clase», de tal forma que las técnicas de clasificación y regresión que vimos buscan aprender un modelo que sea capaz de predecir el valor de esta clase. Se puede encontrar una explicación en vídeo (en inglés) sobre el aprendizaje no supervisado en el canal de Youtube de Mathematical Monk[160].

En el caso del aprendizaje no supervisado, las instancias no están dotadas de una etiqueta. En este tipo de problemas, el objetivo no es predecir una clase, sino detectar patrones que indiquen que varias instancias son similares según una medida de distancia dada.

[160]Mathematical Monk. *(ML 1.3) What is Unsupervised Learning?* https://www.youtube.com/watch?v=IANt56UOaSk. [Online; consultado el 28 de enero de 2016]

Figura 92 – Agrupamiento de instancias

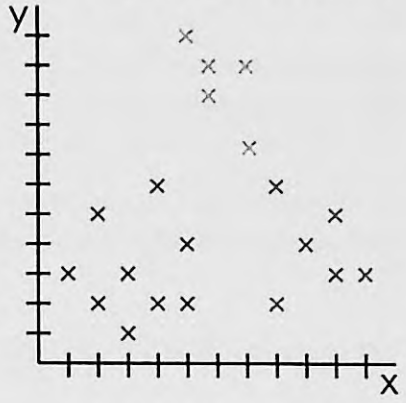
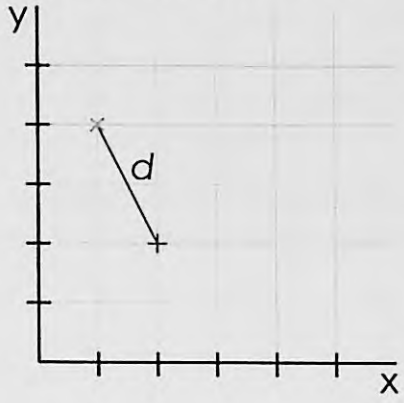


Figura 93 – Distancia euclídea



Este problema recibe el nombre de clustering o *clustering*, aunque en determinadas aplicaciones también se denomina «clustering».

El problema del clustering se puede formalizar matemáticamente del siguiente modo: disponemos de un conjunto de datos compuesto por N instancias, siendo la primera x_1 y la última x_N, donde x_i hace referencia al vector de atributos de la instancia i. Al problema se introduce además una función de distancia d, de tal manera que se define la distancia entre dos instancias como $d_{ij} = d(x_i, x_j)$.

Las técnicas de clustering tienen como objetivo encontrar C grupos o *clusters*[161] así como una función de asignación $c_k = f(x_i)$, que dada una instancia determinará el *cluster* al que pertenece.

La Figura 92 muestra un ejemplo de clustering de instancias en dos dimensiones, donde se puede ver que las instancias han sido agrupadas en 3 *clusters*, cada uno visualizado con un color diferente.

8.3.2 Funciones de Distancia

Como se introdujo anteriormente, uno de los componentes necesarios para realizar la tarea de clustering de instancias es la medida de distancia D. Esta función recibirá como entrada dos instancias y devolverá como salida un valor real que indique la distancia entre ellas. Cuanto menor sea este valor, mayor será la similaridad entre las dos instancias.

Existen numerosas medidas de distancia, siendo algunas más apropiadas que otras para determinados dominios. En esta sección se muestran algunas de las medidas de distancia más comunes.

[161] El concepto de «*cluster*» empleado en aprendizaje no supervisado no debe confundirse con el que hemos empleado anteriormente para hacer referencia a los sistemas distribuidos compuestos por varios nodos.

Figura 94 – Distancia Manhattan

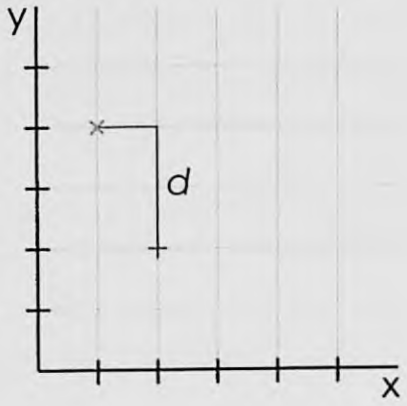
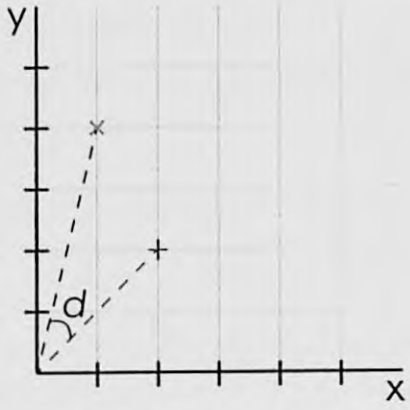


Figura 95 – Distancia coseno



8.3.2.1 Distancia euclídea

La función de distancia euclída devuelve un valor que se correspondería con la longitud del segmento de la línea recta que une dos puntos en un espacio euclídeo. La Figura 93 muestra gráficamente la distancia euclídea entre dos instancias situadas en un plano de dos dimensiones.

La distancia euclídea es ampliamente utilizada para resolver problemas de clustering. No obstante, debemos reparar en que tiene un inconveniente: en el caso de que los atributos estén definidos en intervalos distintos, puede darse el caso de que algunos atributos afecten mucho más a la distancia que otros. No obstante, existe una forma de solventar este inconveniente que pasa por normalizar los valores, ya que al hacerlo ninguno adquiere más relevancia frente a otros a la hora de calcular la distancia.

8.3.2.2 Distancia Manhattan

La función de distancia Manhattan se correspondería a la mínima suma de las longitudes de los segmentos que habría que cruzar en una rejilla para unir los dos puntos. Es similar a la distancia euclídea, pero no permite que la recta que une los puntos sea diagonal, debiendo descomponerse en segmentos verticales y horizontales únicamente. La Figura 94 muestra gráficamente la distancia Manhattan entre dos instancias situadas en un plano de dos dimensiones.

Al igual que en el caso de la distancia euclídea, puede ser necesario normalizar los valores de los atributos para evitar que unos contribuyan más que otros al valor de la distancia.

8.3.2.3 Distancia coseno

La función de distancia coseno se corresponde con el ángulo que forman las dos instancias si se consideran vectores en un espacio geométrico. La Figura 95 muestra gráficamente la distancia coseno entre dos instancias situadas en un plano de dos dimensiones.

La distancia coseno es muy utilizada en análisis de textos, cuando se quiere medir la similaridad entre dos documentos empleando sus vectores de frecuencia de términos.

8.3.3 Técnicas de Agrupamiento

Existen numerosos algoritmos diferentes que permiten llevar a cabo la tarea de clustering de instancias. En esta sección se muestran algunos de los enfoques más comunes.

8.3.3.1 Basadas en conexiones: clustering jerárquico

Los algoritmos de *clustering* basados en conexiones buscan conectar las instancias para ir formando *clusters*, empleando la función de distancia que hayamos definido. Estas técnicas se caracterizan por proporcionar una jerarquía de *clusters*, también llamada «dendograma». Esta jerarquía puede verse como un árbol, donde la raíz contiene un único *cluster* que agrupa a todas las instancias, mientras que las hojas contienen un *cluster* por cada instancia.

Las técnicas de *clustering* jerárquico pueden ser de dos tipos: aglomerativas y divisivas. En las técnicas aglomerativas inicialmente cada instancia es un *cluster* que únicamente la contiene a ella, y estos *clusters* se van uniendo de forma iterativa. Por otro lado, en las técnicas divisivas inicialmente hay un solo *cluster* que contiene a todas las instancias, e iterativamente se va dividiendo en *clusters* más pequeños.

8.3.3.2 Basadas en centroides

Una de las aproximaciones al problema de clustering es el basado en centroides. Un centroide no es más que un punto en el espacio que determina el «centro de gravedad» de cada *cluster*. Tras ejecutar el algoritmo de clustering, dispondremos de una serie de *clusters* que estarán identificados cada uno por un centroide.

Calcular la pertenencia de una instancia a un *cluster* es sencillo: basta con escoger aquel *cluster* cuyo centroide sea el más cercano.

Uno de los algoritmos más extendidos para efectuar *clustering* basado en centroides es K-Medias, más frecuentemente denominado por su nombre en inglés «*K-Means*». El mé-

todo fue presentado por Lloyd en un informe técnico publicado por Bell Labs. en 1957[162], y en 1982 fue publicado en una revista científica[163].

El algoritmo K-Medias tiene varios inconvenientes:

- K-Medias únicamente funciona adecuadamente cuando los datos son linealmente separables. En el caso de grupos que no son linealmente separables, K-Medias no conseguirá buenos resultados.
- K-Medias requiere que se le introduzca como entrada el valor de K, es decir, es necesario definir el número de *clusters* a priori, lo cuál no siempre es posible. Existe una evolución del algoritmo denominada X-Means[164] que trata de averiguar de forma automática el valor de k óptimo.
- El resultado proporcionado por K-Medias es sensible a la inicialización aleatoria de los centroides. Una mala inicialización puede llevar a que el algoritmo converja a mínimos locales, proporcionando una mala clustering de los datos. En 2007 se presentó K-Means++[165], que realiza una inicialización más cuidadosa de los centroides para evitar este efecto.
- Por definición, K-Medias siempre emplea la distancia euclídea. Existen variaciones al algoritmo que permiten trabajar con otras medidas de distancia. Así, por ejemplo, *K-Medians* emplea la distancia Manhattan, mientras que *K-Medoids* emplea funciones de distancia arbitrarias entre las instancias. No obstante, estos métodos no pueden considerarse basados en centroides, pues no se ajustan realmente a la definición de centroide.

8.3.3.3 Basadas en distribuciones

Como hemos visto en la técnica anterior, un conjunto de *clusters* se pueden identificar por sus centroides, que representan el punto medio en el espacio de todas las instancias del conjunto de entrenamiento a las que contienen.

No obstante, esta técnica no es capaz de detectar determinadas distribuciones de puntos en el espacio. Supongamos por ejemplo que disponemos de las instancias representadas en el espacio de dos dimensiones de la Figura 96. En esta figura se puede observar que disponemos de instancias que están situadas de forma horizontal (en el eje de las X) y otras que están situadas de forma vertical (en el eje de las Y). En realidad, estos datos se

[162] S. P. Lloyd. *Least Square Quantization in PCM*. Inf. téc. Bell Telephone Laboratories, 1957

[163] S. P. Lloyd. «Least Squares Quantization in PCM». En: *IEEE Transactions on Information Theory* 28.2 (1982), págs. 129-137

[164] D. Pelleg y A. Moore. «X-Means: Extending K-Means with Efficient Estimation of the Number of Clusters». En: *Proceedings of the 17th International Conference on Machine Learning*. 2000, págs. 727-734

[165] D. Arthur y S. Vassilvitskii. «K-Means++: the Advantages of Careful Seeding». En: *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*. 2007, págs. 1027-1035

pueden asimilar a los que habrían sido generados por dos distribuciones gaussianas (o distribuciones normales), de tal forma que la media de ambas está situada alrededor del punto $(6, 6)$, pero la varianza de cada una de ellas es distinta.

En este caso, la asignación de dos *clusters* mostrada en la figura sería capaz de recoger las dos distribuciones gaussianas que generan estas instancias. Además, se puede observar que existen puntos que pueden pertenecer a más de una distribución (y por lo tanto, a más de un *cluster*). En general, en estos casos decimos que una instancia tiene una probabilidad de pertenencia a cada uno de estos *clusters*.

Para extraer las distribuciones gaussianas que generan a las instancias del conjunto de entrenamiento con la finalidad de hacer clustering, se puede emplear el algoritmo EM o *Expectation-Maximization* introducido en 1977[166].

8.3.3.4 Basadas en densidades

Al visitar las técnicas de clustering basadas en distribuciones hemos visto que estas son capaces de detectar *clusters* que las técnicas basadas en centroides no eran capaces de inferir. No obstante, las técnicas basadas en distribuciones tampoco pueden detectar todo tipo de *clusters*.

La Figura 97 muestra instancias agrupadas en dos *clusters* que no podrían haber sido detectadas por técnicas basadas en centroides o en distribuciones, ya que los datos no son linealmente separables y tampoco están generados a partir de varias distribuciones gaussianas. No obstante, las técnicas basadas en densidades sí son capaces de obtener estas agrupaciones.

Figura 96 – Instancias generadas a partir de dos distribuciones gaussianas

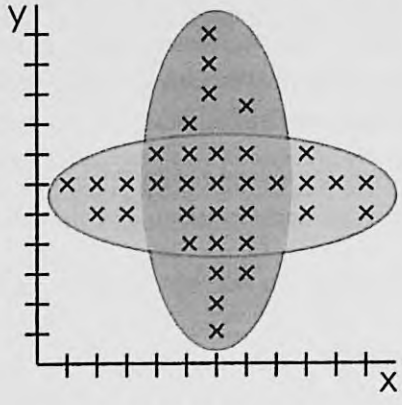
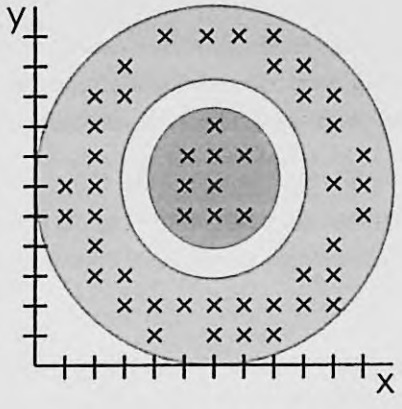


Figura 97 – Agrupación de instancias basada en densidades



[166] A.P. Dempster, N.M. Laird y D.B. Rubin. «Maximum Likelihood from Incomplete Data via the EM Algorithm». En: *Journal of the Royal Statistical Society* B39.1 (1977), págs. 1-38

El algoritmo de *clustering* basado en densidades más extensamente citado y utilizado es con diferencia DBSCAN, introducido en el año 1996[167].

8.3.4 Recomendación

En la actualidad uno de los grandes campos de aplicación del análisis de Big Data es la construcción de sistemas recomendadores. Estos están presentes en muchas de las aplicaciones de Internet, pero destacan especialmente en los sitios web de comercio electrónico.

Hablando en términos muy generales, el propósito de un sistema de recomendación es extraer patrones a partir de los datos para inferir cuáles son los posibles intereses de un usuario concreto, para a continuación sugerirle a ese usuario que adquiera, visite, siga, etc.[168] estos intereses.

La recomendación tiene una relación muy estrecha con el problema de *clustering* que acabamos de presentar, como mostraremos a continuación.

A la hora de construir sistemas de recomendación se presentan fundamentalmente dos enfoques: el filtrado basado en contenidos y el filtrado colaborativo.

8.3.4.1 Filtrado basado en contenidos

En este caso, se busca recomendar al usuario ítems similares a aquellos en los que se ha interesado previamente. En muchos sitios web es frecuente ver secciones de «productos relacionados» o «noticias relacionadas». En este caso, la recomendación es del tipo «si te interesa X, entonces puede interesarte Y», donde Y es un ítem similar o relacionado con X.

En el filtrado basado en contenidos es fácil ver la relación con el problema del *clustering*: los ítems a recomendar serán aquellos que pertenezcan al mismo *cluster* que el de referencia (por ejemplo, el que el usuario está visitando actualmente). Este *clustering* puede realizarse en base a las características del producto, su descripción, etc.

El principal inconveniente del filtrado basado en contenidos es que es difícil que se recomienden al usuario nuevos *items* de interés, ya que las recomendaciones siempre son productos similares a los que el usuario ya conoce.

---

[167] M. Ester y col. «A Density-Based Algorithm for Detecting Clusters in Large Spatial Databases with Noise». En: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. 1996, págs. 226-231

[168] A lo largo de esta sección nos referiremos en todo momento a *ítems*, ya que los conceptos que vamos a ver se pueden aplicar a productos, noticias, contenidos multimedia, etc.

8.3.4.2 Filtrado colaborativo

En el filtrado colaborativo, para realizar recomendaciones se tienen en cuenta las preferencias de usuarios similares a aquél que es recomendado. La idea que alimenta este enfoque es que si dos usuarios A y B tienen intereses similares en una serie de aspectos conocidos (por ejemplo, productos que han valorado con puntuaciones similares o productos que ambos han adquirido), entonces se puede emplear la información del usuario A para inferir información desconocida sobre el usuario B.

Las reglas por tanto de recomendación en el filtrado colaborativo son del tipo «otros usuarios que compraron X también compraron Y».

La ventaja de este enfoque con respecto al filtrado basado en contenidos es que se pueden ofrecer ítems que son completamente distintos al actual, permitiendo así ofrecer un catálogo de recomendaciones mucho más amplio.

Existen muchos algoritmos distintos para realizar recomendación de productos mediante filtrado colaborativo, y algunos de ellos emplean técnicas de *clustering* para determinar grupos de usuarios que tienen intereses similares. Para ello, si tenemos I ítems, podemos modelar cada usuario como un vector de longitud I donde la posición $x_i^{(u)}$ es la valoración que tiene el usuario u del ítem i. Aunque este es un enfoque, los valores no tienen por qué representar necesariamente valoraciones, y podrían por ejemplo ser valores binarios, por ejemplo 0 si el usuario no ha comprado un producto y 1 si lo ha hecho.

Normalmente estos vectores serán dispersos, es decir, la mayoría de sus posiciones serán 0, ya que los usuarios habrán adquirido o valorado un porcentaje muy pequeño de todo el catálogo de ítems. No obstante, existen técnicas que permiten trabajar eficientemente con vectores dispersos, reduciendo así el tiempo del proceso de *clustering*.

8.3.5 Aplicaciones del Aprendizaje No Supervisado

Como ya hemos señalado, el problema de agrupación o *clustering* tiene como principal objetivo inferir patrones a partir de los datos, por lo que son tareas que tienen numerosas aplicaciones prácticas en el contexto del Big Data. En esta sección se enumeran algunas de estas aplicaciones.

8.3.5.1 Segmentación de usuarios

Dentro del análisis de mercado, la clustering de usuarios o clientes es una de las aplicaciones más frecuentes de las técnicas de *clustering*. En concreto, esta tarea busca encontrar grupos de usuarios que comparten ciertas características, ya sean atributos demográficos (edad, sexo, zona residencial...) o intereses sobre productos. Como hemos visto anteriormente, esta aplicación además supone una base para comenzar con el desarrollo de sistemas de recomendación que empleen filtrado colaborativo.

8.3.5.2 Detección de comunidades

Se pueden emplear técnicas de *clustering* para llevar a cabo análisis de redes sociales con el fin de detectar comunidades de usuarios. Estas comunidades son núcleos de personas que mantienen un mayor contacto entre sí, es decir, tienen una mayor densidad. Normalmente estos núcleos se forman porque estos usuarios han estudiado o trabajado juntos, o bien comparten determinados intereses.

8.3.5.3 Clasificación de ítems

En Internet es frecuente que en algunas ocasiones se nos sugieran etiquetas automáticas para clasificar elementos que introducimos en una página web: puede ocurrir al insertar un producto en una página de segunda mano, al introducir una entrada en un *blog* o incluso puede servir para detectar el estilo de una canción en función de su ritmo.

El problema del etiquetado automático de elementos es en realidad un problema de clasificación, como los que vimos en el capítulo anterior, pero que no obstante se puede resolver mediante técnicas de *clustering*. Para ello, basta con disponer de un conjunto de entrenamiento sobre cuyas instancias aplicaremos un algoritmo de *clustering*, etiquetando posteriormente cada uno de los *clusters*. A continuación, cuando se introduzca una instancia nueva, basta con asignarla a un grupo para saber su etiqueta.

Al implementar esta tarea, es frecuente combinar K-Medias con la técnica de clasificación k-NN. Como recordaremos del capítulo anterior, el principal problema de esta técnica «vaga» es que es costoso el calcular la distancia entre la instancia que queremos clasificar y todas las instancias del conjunto de entrenamiento. Por este motivo, al aplicar K-Medias se reduce todo el conjunto de entrenamiento a únicamente los centroides de los K *clusters*. Posteriormente basta con calcular la distancia con estos centroides para ver a qué grupo se asigna la instancia, y en consecuencia qué etiqueta se le atribuye.

8.3.5.4 Análisis genómico

El *clustering* también ofrece aplicaciones interesantes en el ámbito de la bioinformática. En concreto, se pueden emplear estas técnicas para agrupar genes que contengan proteínas con funcionalidades similares, lo que permite mejorar el conocimiento científico sobre el ADN de un organismo.

8.4 Aprendizaje Automático Escalable

En ocasiones es necesario o al menos recomendable el realizar un análisis de los datos de forma distribuida, ya que estos siguen requiriendo mucho espacio para ser almacenados o los algoritmos de aprendizaje automático empleados se pueden paralelizar reduciendo significativamente su tiempo de ejecución.

Durante los últimos años han ido surgiendo herramientas que han empleado las principales plataformas de procesamiento de Big Data (Apache Hadoop, Apache Spark, etc.) con el fin de llevar a cabo este análisis de datos de forma distribuida, abstrayendo notablemente la complejidad de la distribución de los datos, simplificando así el proceso para usuarios y desarrolladores.

En esta sección vamos a mencionar brevemente dos de estas herramientas: Apache Mahout y Apache Spark MLLib.

8.4.1 Apache Mahout

Apache Mahout[169] es un proyecto de software libre de la Apache Software Fundation. El objetivo es proporcionar implementaciones libres (gratuitas y de dominio público) de algoritmos de aprendizaje automático distribuido o escalable. Estos algoritmos incluyen en la actualidad algunos de los que hemos visto durante el capítulo para realizar clasificación, regresión, clustering, recomendación o clustering de la dimensionalidad.

Tradicionalmente estos algoritmos habían sido diseñados para ejecutarse sobre Hadoop MapReduce. En la actualidad, estos algoritmos se conservan, aunque se están introduciendo otros que funcionan sobre otras plataformas como Apache Spark, ya que estas prometen un mejor rendimiento. Además, en la última versión han introducido un nuevo entorno de análisis matemático denominado «Samsara», que soporta el desarrollo rápido de nuevos algoritmos de aprendizaje automático de forma más sencilla.

8.4.2 Apache Spark MLlib

Apache Spark MLLib[170] es una librería muy versatil para realizar tareas de aprendizaje automático de forma escalable. Tiene la principal ventaja de que está integrada al 100 % con la plataforma de procesamiento de Big Data Apache Spark, de la que ya hablamos en el capítulo anterior. Esta integración permite, entre otras cosas, seguir empleando la abstracción de los conjuntos de datos distribuidos o RDD (*Resilient Distributed Dataset*) para aprender y explotar modelos de clasificación, regresión o *clustering*, entre otras cosas.

Los desarrolladores que empleen Spark MLLib podrán programar en Java, Scala y Python, una versatilidad que nos facilita la tarea de análisis de datos, agilizando el tiempo de desarrollo de las implementaciones de los soluciones específicas de aprendizaje automático que necesitemos.

Además, al funcionar sobre Apache Spark, MLLib tiene un rendimiento muy alto comparado con otras plataformas de procesamiento de Big Data, mejorando notablemente en tiempo de ejecución a Hadoop MapReduce. Además el proyecto, que es software libre,

[169] Apache. *Apache Mahout: Scalable Machine Learning and Data Mining*. http://mahout.apache.org. [Online; consultado el 6 de agosto de 2015]. 2015

[170] Apache. *MLlib | Apache Spark*. http://spark.apache.org/mllib/. [Online; consultado el 6 de agosto de 2015]. 2015

cuenta con las contribuciones de una fuerte comunidad, lo que le ha permitido implementar la mayoría de algoritmos que hemos visto en este capítulo y muchos otros para realizar tareas de análisis de datos y aprendizaje automático.

8.5 Análisis de Big Data en la Nube

Como ya hemos visto en los capítulos anteriores, existen compañías que prestan servicios de almacenamiento y procesamiento de datos en la nube. La principal ventaja de estos servicios es que permiten que los clientes paguen únicamente por los recursos que necesitan a cada momento, evitando así tener que provisionar servidores físicos en sus propios *datacenters*.

Si bien los servicios de almacenamiento y procesamiento de datos en la nube existen desde hace algunos años, recientemente ha venido surgiendo un nuevo abanico de productos que permiten realizar análisis de datos en la nube. Algunos, como los que permiten realizar análisis de datos mediante consultas SQL, ya los vimos anteriormente. No obstante, existen numerosas herramientas que permiten aplicar sobre nuestro Big Data algunas de las técnicas de aprendizaje automático que acabamos de explicar.

A continuación vamos a mostrar cómo algunas de las herramientas más extendidas de aprendizaje automático en la nube permiten sacar aún más valor de nuestros datos sin tener que dedicar un gran esfuerzo en el desarrollo de soluciones para analizarlos, proporcionando un enfoque desde un punto de vista práctico y mostrando directamente cómo realizar algunas aplicaciones clásicas del aprendizaje automático, tales como la predicción o la detección de anomalías.

8.5.1 Predicción

Como ya hemos visto con anterioridad, el problema de predicción o clasificación consiste en averiguar para una determinada instancia, caracterizada mediante una serie de atributos, la etiqueta o categoría que le correspondería. Para ello se cuenta con un conjunto de entrenamiento, que son instancias para las que se conoce con anterioridad su etiqueta, y con estos datos se procede a entrenar un modelo de clasificación: un árbol de decisión, un modelo probabilístico, etc.

BigML[171] es una compañía dedicada exclusivamente a los servicios de aprendizaje automático en la nube. Uno de los servicios más destacados de la empresa es precisamente el análisis predictivo.

Tras completar el registro en BigML e iniciar sesión se mostrará la pantalla de la Figura 98, donde se recoge el listado de orígenes de datos disponibles. Como podemos observar, por defecto aparecen algunos conjuntos de datos de forma automática, lo que nos va a permitir probar la plataforma con estos conjuntos de datos.

[171] BigML. *BigML.com is Machine Learning for Everyone*. https://bigml.com. [Online; consultado el 10 de agosto de 2015]. 2015

Además, se pueden añadir nuevos conjuntos de datos. Estos se pueden importar directamente en formato CSV o en ARFF de Weka. También se puede conectar con servicios externos como Dropbox, Amazon S3, Azure, Google Drive, etc. con el fin de poder importar conjuntos de datos que ya están almacenados en otra plataforma en la nube.

Si hacemos clic en uno de estos orígenes de datos se mostrará su descripción, tal y como se puede observar en la Figura 99. En esta pantalla, cada fila muestra un atributo de los datos. La primera columna muestra el nombre del atributo y la segunda su tipo. Las siguientes columnas muestran valores de ejemplo para las primeras instancias en el conjunto de datos, de tal forma que podemos adquirir una idea de los valores que puede tomar cada atributo.

Hasta ahora hemos hablado de orígenes de datos en lugar de conjuntos de datos. La principal diferencia entre estos conceptos dentro de BigML es que un origen de datos es un fichero (por ejemplo, un CSV) mientras que un conjunto de datos son las instancias que contiene el fichero.

Si hacemos clic en el botón de la parte superior derecha denominado «*Configure source*» se desplegará el formulario de la Figura 100. En este formulario podremos indicar, por ejemplo, el separador de campos del fichero, los símbolos que indican datos valores disponibles o si la primera fila indica los nombres de los atributos o no. Incluso se podría realizar en esta fase un análisis de textos para vectorizarlos en *tokens*, empleando técnicas como el *stemming* para reducir palabras a su raíz semántica o la detección de *stop words* para ignorar palabras sin valor semántico, como las preposiciones o los determinantes.

Una vez configurados todos los aspectos del origen de datos, podemos procesarlo para crear un nuevo conjunto de datos. Para ello, basta con hacer clic en el botón denominado «*Configura dataset*», donde podremos escoger el nombre del conjunto de datos y el

Figura 98 – Listado de los orígenes de datos en BigML

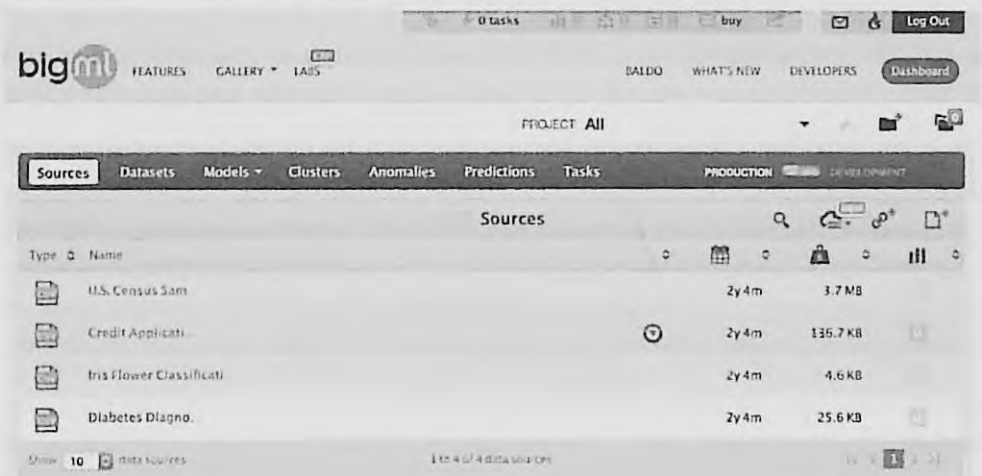


Figura 99 – Detalles de un origen de datos en BigML

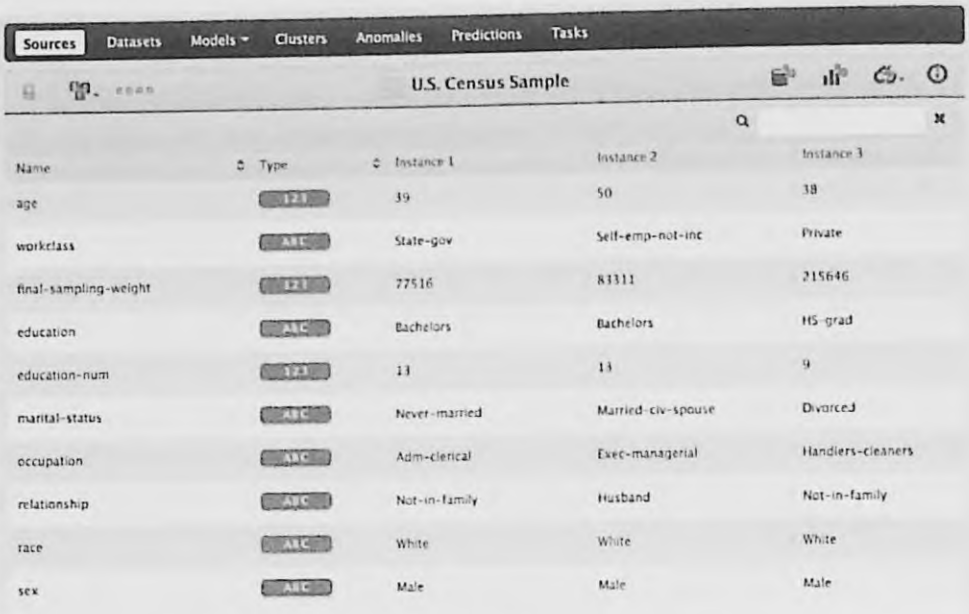


Figura 100 – Configuración de un origen de datos en BigML

tamaño del mismo definido como una fracción del total de datos disponible en el origen de datos. Finalmente haremos clic en el botón «*Create dataset*» para proceder con la creación del conjunto de datos.

Una vez creado el conjunto de datos se mostrará la pantalla de la Figura 101. En esta pantalla se pueden visualizar los histogramas para cada atributo. Estos histogramas muestran las distribuciones de valores de cada atributo y sus frecuencias. Si colocamos el ratón sobre una de las barras se nos mostrará un cuadro de texto con el nombre del valor al que se refieren, como se puede observar en la Figura 101 con el caso del estado civil «*Divorced*».

A continuación, si colocamos el ratón sobre el botón de acciones de la parte superior derecha (el que muestra como icono dos ruedas dentadas), se mostrará un listado de operaciones, donde se pueden escoger numerosas opciones para configurar diversos ti-

Figura 101 – Detalles de un conjunto de datos en BigML

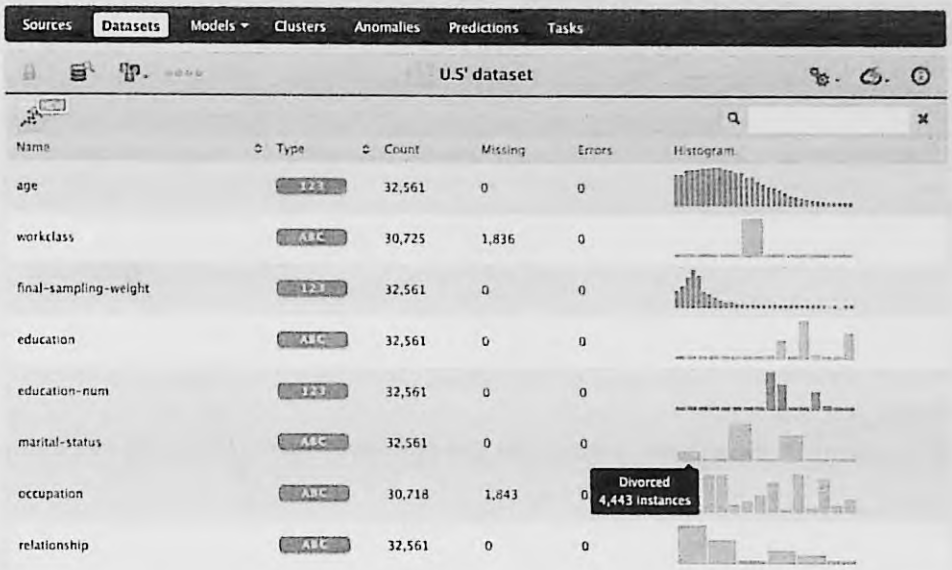
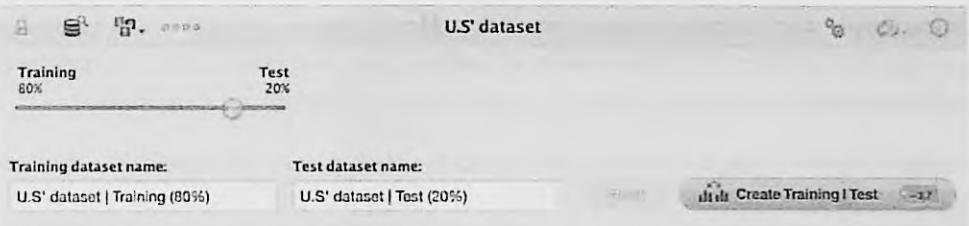


Figura 102 – División de un conjunto de datos en entrenamiento y test en BigML

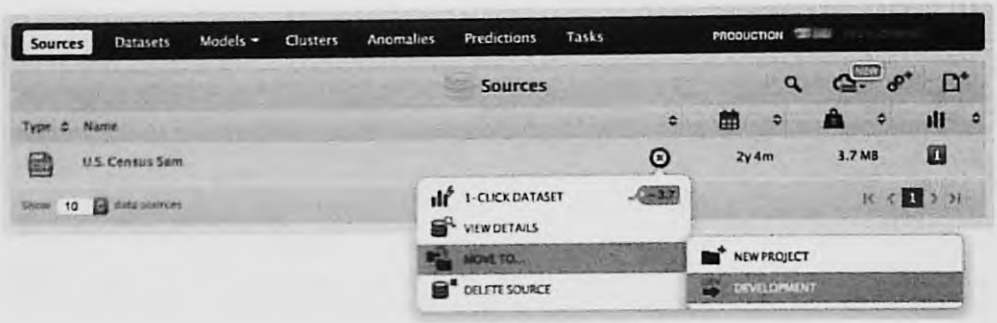


pos de modelos, así como para generar una muestra aleatoria del conjunto de datos u obtener un nuevo conjunto de datos filtrado (es decir, que contengan aquellas instancias que cumplan una condición). En nuestro caso, vamos a escoger la opción «*Training and Test set split*», que nos permitirá dividir el conjunto de datos en un subconjunto de entrenamiento y otro de test.

A continuación se mostrará el formulario de la Figura 102 que nos permitirá escoger el porcentaje de instancias asignado a cada uno de los conjuntos de datos, que por defecto es el 80 % para entrenamiento y el 20 % para test. Además, podemos configurar los nombres de los conjuntos de datos de forma individual. Una vez establecida la configuración deseada, haremos clic en el botón «*Create Training | Test*».

Es posible que llegados a este punto BigML nos informe de que no tenemos crédito para realizar la operación, siempre que no hayamos contratado una suscripción válida. BigML nos permite trabajar en modo desarrollo con conjuntos de datos de hasta 16MB de forma gratuita.

Figura 103 – Traslado de un origen de datos de Producción a Desarrollo en BigML



Para ello, deberemos ir a la pantalla de orígenes de datos en la pestaña «*Sources*». A continuación, si pulsamos en la flecha que se muestra junto al nombre del origen de datos, se cargará un desplegable (Figura 103) que permitirá mover el conjunto de datos a desarrollo, en la opción «*Move To*» y a continuación «*Development*». Después, podremos pinchar en el interruptor «*Production / Development*» para cambiar el modo de funcionamiento actual. A partir de este momento vamos a trabajar en modo de desarrollo para evitar tener que adquirir una suscripción a BigML.

Una vez creados los conjuntos de entrenamiento y test, si vamos a la vista de los conjuntos de datos (pestaña *Datasets*) se mostrará el listado de la Figura 104, donde podemos ver que hay dos conjuntos de datos nuevos. A continuación, pincharemos en el conjunto de entrenamiento que acabamos de crear, el que incluye «*Training*» en su nombre.

Vamos a emplear este conjunto de datos para crear un modelo de predicción, que en BigML será por defecto un árbol de decisión. Para ello, desplegaremos el panel de acciones haciendo clic en el botón que muestra las ruedas dentadas, y a continuación escogeremos la opción «*Configure Model*». Se mostrará el formulario de la Figura 105.

En este formulario, lo primero que debemos hacer es seleccionar la clase que queremos predecir. En este caso escogeremos el atributo «*income-class*», que indica si un determinado ciudadano estadounidense cobra más o cobra menos de 50.000 dólares al año. En la opción de «*pruning*» (poda) dejaremos la opción seleccionada por defecto, en la que BigML se encarga de determinar de forma inteligente la poda más adecuada.

Si hacemos clic en la opción «*Configure*» podríamos seleccionar numerosas opciones avanzadas relativas al entrenamiento del modelo de predicción. Algunas de estas opciones se describen a continuación:

- El número máximo de nodos que puede contener el árbol de decisión entrenado, lo que permite limitar su tamaño.
- Los costes de los atributos: se puede indicar que un atributo es más costoso de evaluar que otros, lo cuál es algo común en algunos dominios como el médico.

Esto permite que el árbol de decisión resultante evite escoger ese atributo salvo que sea fundamental, y tratará de hacerlo lo más tarde posible.

- Los pesos de las clases: esto permite que se premie más acertar unas clases sobre otras, es decir, se puede dar más importancia a los verdaderos positivos o a los verdaderos negativos, en función del dominio.
- Se puede tomar una muestra reducida del conjunto de entrenamiento antes de entrenar el árbol de decisión. Se puede escoger tanto el tamaño de la muestra como el método de muestreo y si incluye o no reemplazo.
- Se puede indicar si realizar o no una reordenación del conjunto de datos, para evitar sesgos si los datos están previamente ordenados.

Finalmente, una vez escogida la configuración deseada para nuestro modelo de predicción, haremos clic en el botón «*Create model*». El procedimiento de entrenamiento del árbol de decisión durará unos segundos. A continuación se mostrará el modelo entrenado, tal y como podemos observar en la Figura 106. Como ya vimos anteriormente, cada nodo del árbol indica una pregunta sobre un atributo de los datos, y los nodos hoja indican el valor para la clase. En este caso, si pulsamos sobre cualquiera de los nodos del árbol se nos muestra el camino de preguntas y respuestas seguidas hasta llegar a ese nodo.

Figura 104 – Listado de conjuntos de datos en BigML

Figura 105 – Creación de un modelo de predicción en BigML

En la Figura 106 hemos seleccionado un nodo hoja, y se puede observar que BigML nos indica que la clase preferida para aquellos individuos no casados con otra persona civil, con un máximo de 26 años, con una ganancia de capital de 7.136 dólares como máximo y con un nivel educativo igual o inferior a 12 (lo que según la definición del conjunto de datos es alguien asociado a la academia pero sin un título universitario) es aquella que indica que el individuo cobrará menos de 50.000 dólares al año.

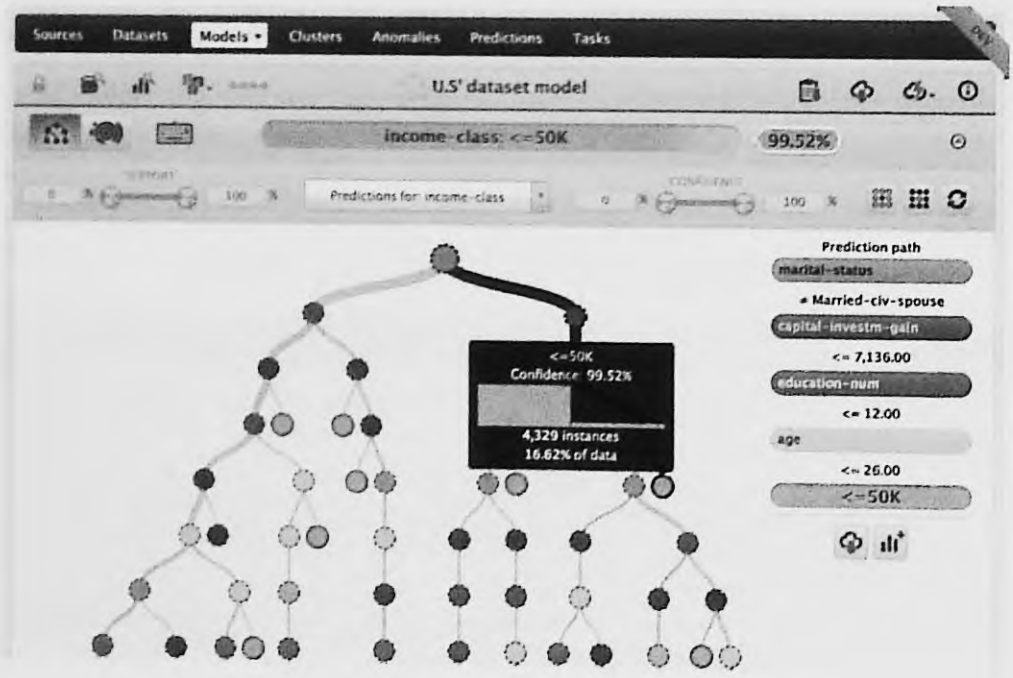
Además, BigML también nos indica un valor de confianza, que no es más que el porcentaje de aciertos sobre el conjunto de entrenamiento para las instancias que cumplían las características indicadas por esta hoja.

Si hacemos clic en el botón «*Sunburst*» en la parte superior izquierda se mostrará una visualización alternativa del árbol de decisión, con una forma circular (Figura 107). El coloreado de este gráfico se puede escoger en función de la confianza de la predicción, del valor de la clase o de un coloreado arbitrario de los atributos.

En la parte superior derecha, el botón «Model Summary Report» permite generar un fichero con estadísticas de las distribuciones de los diferentes atributos y cómo cada uno de ellos contribuye al valor predicho de la clase.

Por último, el botón de acciones abre el listado de operaciones. Si hacemos clic en la opción «*Predict question by question*» se cargará la interfaz de la Figura 108 que permite contestar preguntas sobre instancias una a una, mostrando en cada momento la predicción más probable y el valor de confianza.

Figura 106 – Visualización del árbol de decisión en BigML



Si por el contrario escogemos la opción «*Predict*», se cargará la pantalla de la Figura 109 que permitirá dar valor a cada uno de los atributos individualmente, mostrando en la parte superior la predicción realizada por el modelo y el nivel de confianza. Esta opción es por tanto idéntica a la anterior en funcionalidad, variando únicamente en la interfaz que permite introducir valores a los diferentes atributos de la nueva instancia.

Por último, podemos evaluar el rendimiento del árbol de decisión sobre un conjunto de test. Para ello escogeremos la opción «*Evaluate*» del desplegable de acciones. Esto cargará la pantalla de la Figura 110, donde debemos seleccionar el modelo aprendido y el conjunto de test. Si abrimos el desplegable «*Configure*», se muestran opciones avanzadas para la evaluación, lo que incluye un mapeo de campos (suponiendo que ambos conjuntos de datos tuvieran una estructura diferente) y la posibilidad de evaluar sobre una muestra de este conjunto. Para continuar, haremos clic en el botón «*Evaluate*».

Lo primero que podemos ver tras realizar la evaluación de un modelo es ver la matriz de confusión resultante (Figura 111), donde se pueden observar los verdaderos y falsos positivos y negativos, así como la precisión y la exhaustividad para cada clase. También se puede ver la precisión y exhaustividad media, y la puntuación F1 media.

En la parte inferior se muestran los resultados para la precisión, la exhaustividad y la puntuación F1. Estos valores pueden mostrarse agregados o por clases. Además, se pueden comparar con dos modelos base: el llamado «*mode*» que escoge siempre la clase más fre-

Figura 107 – Visualización alternativa del árbol de decisión en BigML

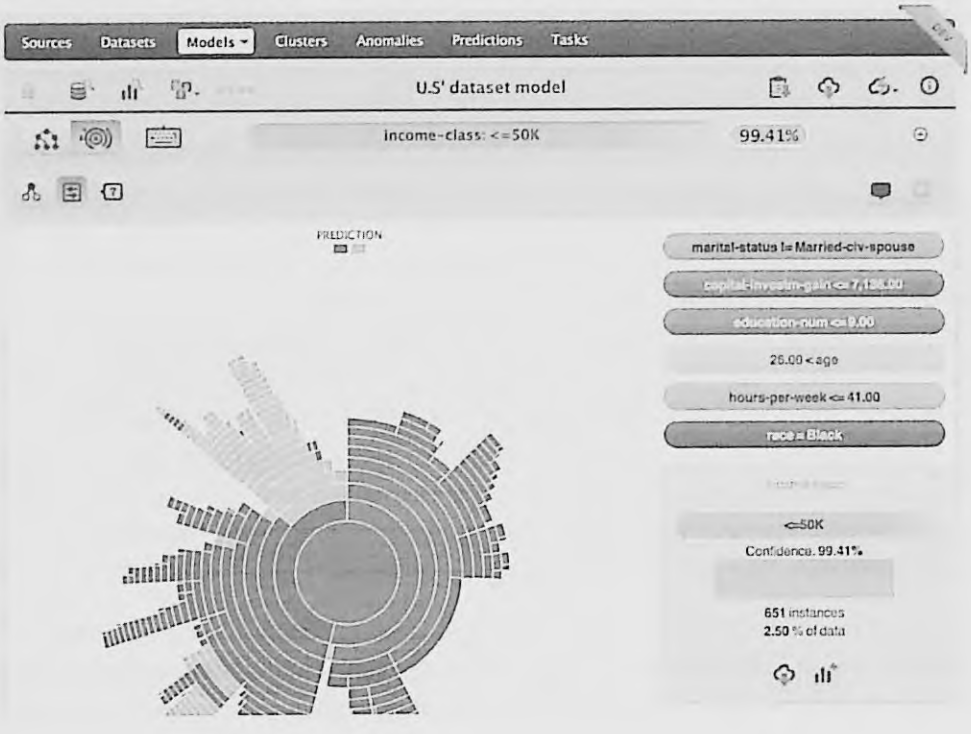


Figura 108 – Predicción de instancias pregunta a pregunta en BigML

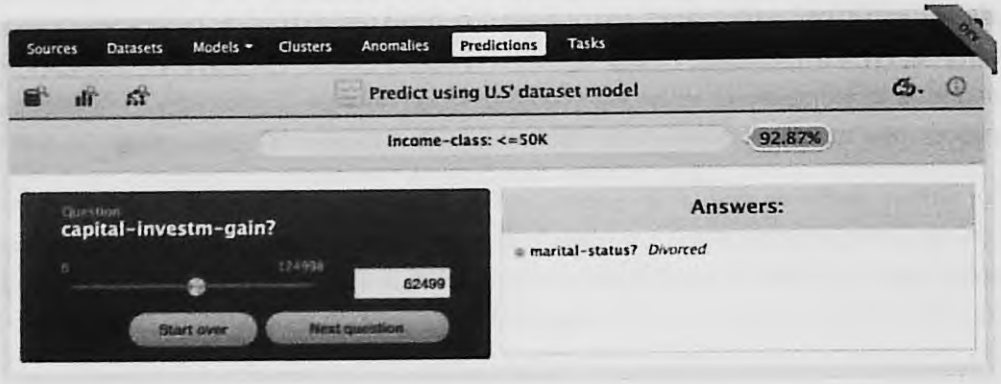
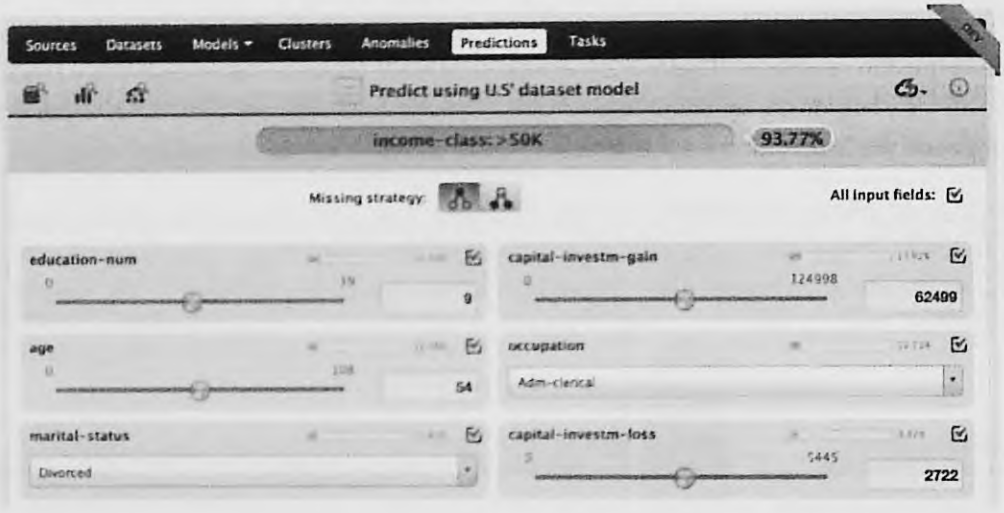


Figura 109 – Predicción de instancias en BigML



cuente y el «*random*» que funciona de forma completamente aleatoria. Esta información se muestra en la Figura 112.

Aunque ahora hemos visto modelos de predicción basados en árboles de decisión, BigML también permite aprender modelos basados en *ensembles*, que son clasificadores formados por varios árboles de decisión. Para ello, desde la vista del conjunto de datos de entrenamiento haremos clic en la opción «*Configure Ensemble*» en lugar de «*Configure Model*». El resto del procedimiento para crear el *ensemble* es prácticamente idéntico al que ya empleáramos al crear un árbol de decisión, si bien se permite configurar también el número de árboles que pueden componer el *ensemble* (hasta 10.000 en producción) y el modo de selección de atributos (escoger todos para todos los árboles o hacerlo de forma aleatoria).

Tras unos segundos que tarda en entrenarse el modelo, finalmente se mostrará una pantalla similar a la de la Figura 113. En ella, se indica el número de árboles de decisión que componen el *ensemble*, y a continuación se muestra de forma resumida cada uno de es-

Figura 110 – Evaluación de un modelo en BigML

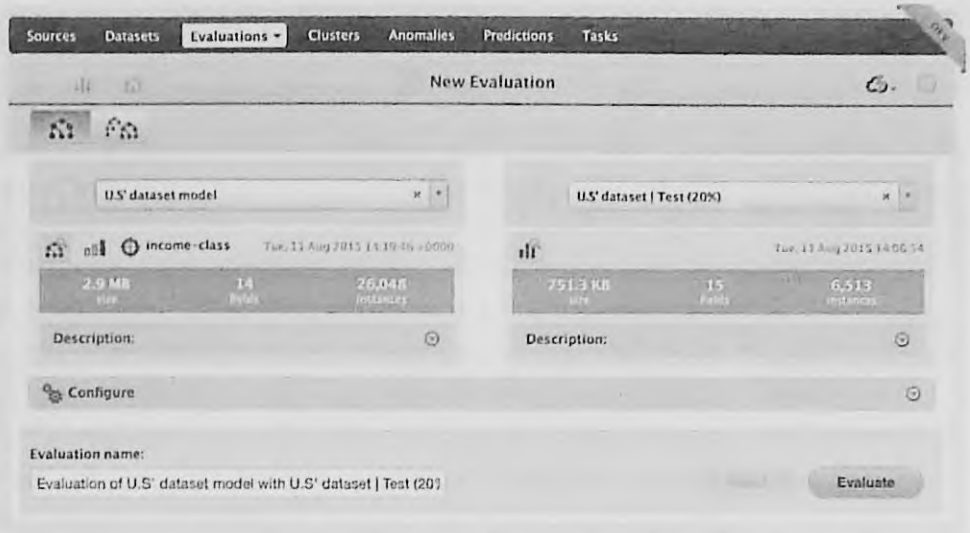
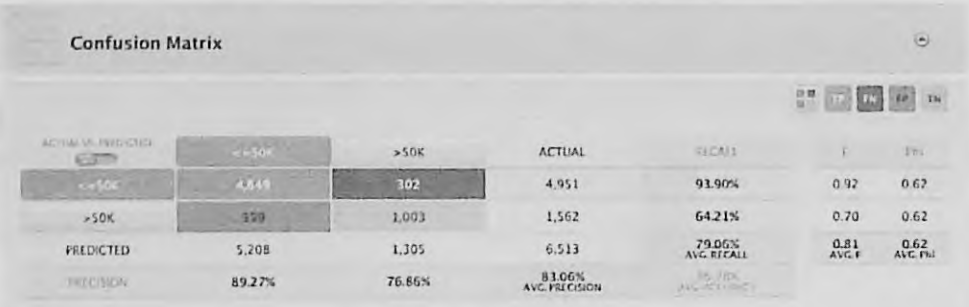


Figura 111 – Matriz de confusión de un modelo en BigML



tos árboles. Haciendo clic en cualquiera de ellos se puede explorar el árbol de decisión correspondiente con un mayor detalle, como ya vimos en las figuras 106 y 107. Finalmente, podríamos proceder con su explotación (es decir, para realizar predicción sobre nuevas instancias) o evaluarlo sobre el conjunto de test como ya hicimos anteriormente.

- ***¿Proporciona mejores resultados el uso de un ensemble en lugar de un*** árbol de decisión ***para la predicción de ingresos sobre el conjunto de datos del censo de los Estados Unidos que hemos utilizado en este ejemplo?***

8.5.2 Agrupamiento

Otro de los problemas de aprendizaje automático que hemos visto en este capítulo involucraba la detección de patrones mediante la agrupación de instancias similares, o *clustering*.

BigML permite realizar tareas de *clustering* en la nube de forma sencilla. Para comenzar, vamos a abrir el conjunto de datos «U.S' dataset» con el que ya trabajamos para realizar

Figura 112 – Resultados de la evaluación de un modelo en BigML

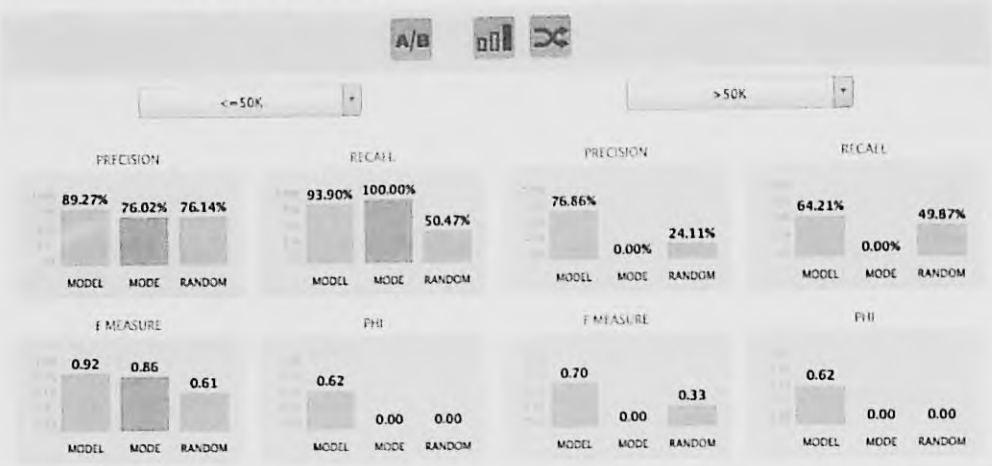
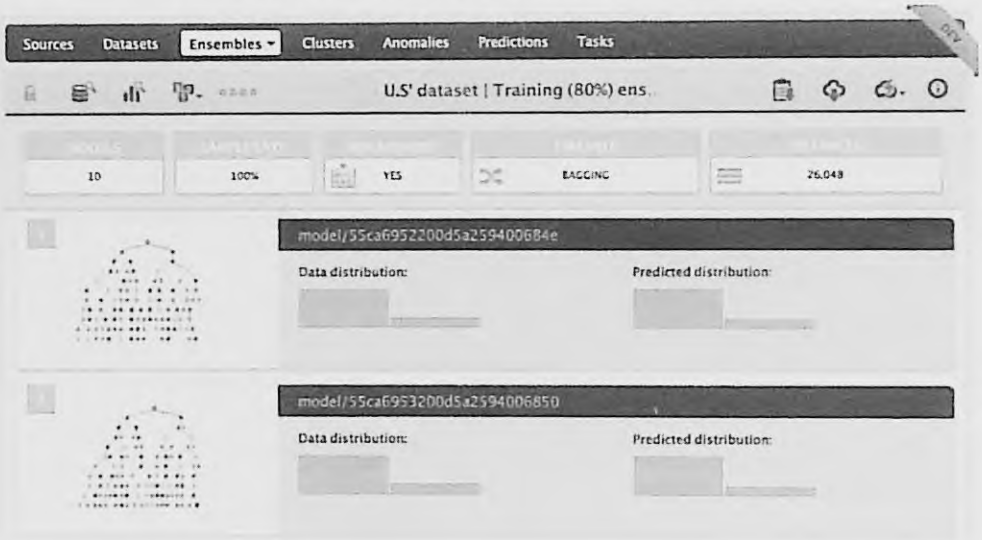


Figura 113 – Visualización del *ensemble* en BigML



predicción. En este caso, vamos a realizar *clustering* para ver si podemos encontrar grupos de población que sean similares.

Una vez cargado el conjunto de datos haremos clic en botón que muestra el desplegable de acciones, que es aquél que contiene el icono de las ruedas dentadas. De las opciones que se muestran escogeremos «*Configure Cluster*».

A continuación se mostrará el formulario de la Figura 114. En este formulario podremos escoger el algoritmo de *clustering*, el valor numérico por defecto para rellenar los valores numéricos vacíos y podemos optar también por entrenar un modelo de predicción para cada *cluster*.

Con respecto a los algoritmos de *clustering* podemos escoger K-Means, en cuyo caso deberemos seleccionar un número predefinido de *clusters*. La otra opción es G-Means, que es una variación de K-Means donde el valor de K se obtiene automáticamente a partir de una serie de parámetros estadísticos[172].

Además, si hacemos clic en el desplegable «*Configure*» podremos configurar otros parámetros más avanzados, como el peso de cada atributo o las características de la muestra con la que se entrenará el modelo. En este caso, vamos a optar por la configuración por defecto (en la que se emplea el conjunto de entrenamiento completo y no existen pesos para los atributos) y a continuación haremos clic en el botón «*Create Cluster*».

Tras unos segundos de operación, se mostrará una visualización de los *clusters* calculados, similar a la que se puede ver en la Figura 115. Si seleccionamos uno de los *clusters*, en la columna de la derecha se mostrará el centroide para ese *cluster*, pudiendo ver en este caso cuáles son los principales perfiles de usuario que ha descubierto el algoritmo de aprendizaje automático.

Si colocamos el ratón sobre el botón de acciones de la parte superior derecha (el que tiene por icono una nube con un rayo), entonces se mostrará el listado de operaciones. Una opción interesante es la que se denomina «*Centroid*». Si hacemos clic en ella se cargará la pantalla de la Figura 116, donde podremos configurar manualmente cada uno de los atributos de una nueva instancia. Según modifiquemos estos atributos, en la parte superior se mostrará el *cluster* asignado a esta instancia, así como la distancia a su centroide.

8.5.3 Detección de Anomalías

Otra aplicación frecuente del aprendizaje automático es la detección de anomalías, que consiste en analizar las instancias de un conjunto de datos para identificar aquellas que son notablemente diferentes al resto.

Figura 114 – Creación de un modelo de *clustering* en BigML

[172] G. Hamerly y C. Elkan. «Learning the K in K-Means». En: *Advances in Neural Information Processing Systems 16*. 2004, págs. 281-288

Figura 115 – Visualización de los *clusters* en BigML

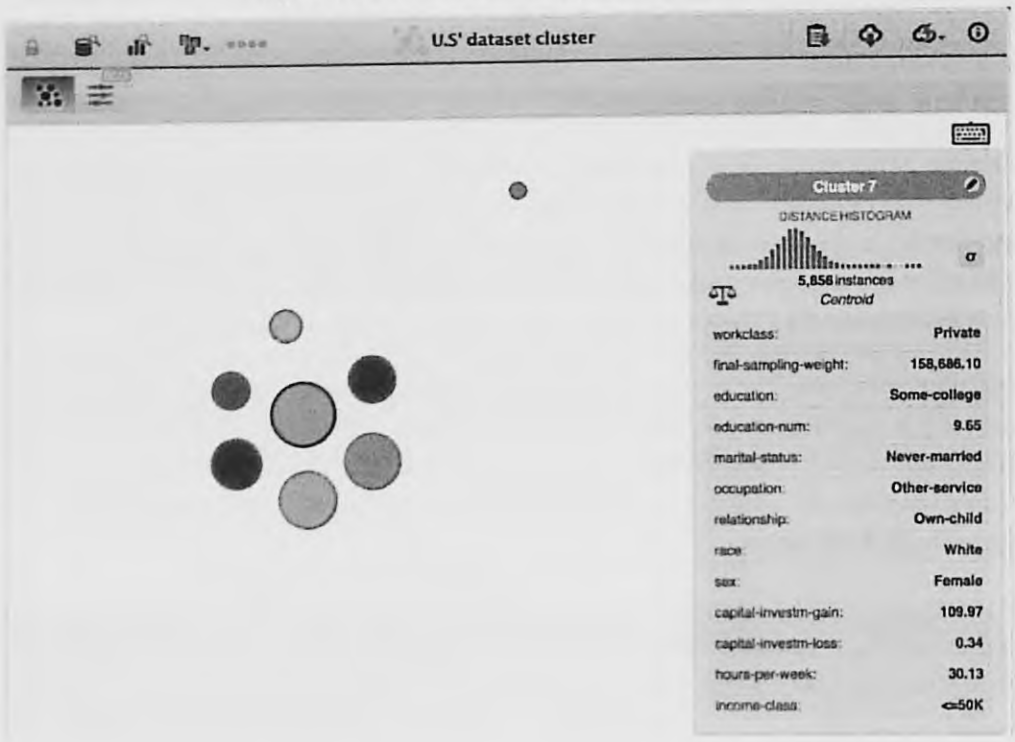
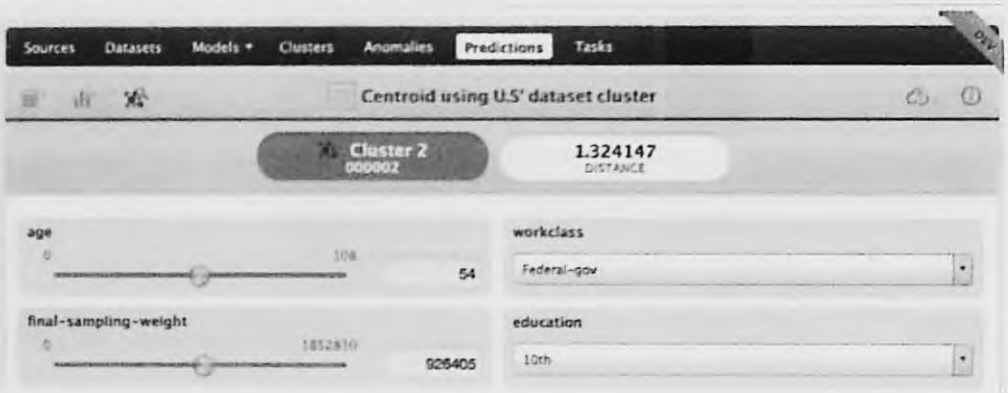


Figura 116 – Asignación de instancias a *clusters* en BigML



La detección de anomalías es un campo que tiene numerosas aplicaciones. Una de las más relevantes es el análisis de fraude en pagos con tarjeta de crédito. Supongamos, por ejemplo, que un cliente de tarjeta de crédito suele realizar siempre sus compras con tarjetas en comercios del territorio español. Además, suele comprar siempre alimentos y ropa, y no gasta más de 200 o 300€ por cada transacción.

En un momento dado, la entidad recibe una orden de pago efectuada en un comercio electrónico en Alemania, para comprar tecnología por valor de 800€. Evidentemente, esta operación es una anomalía en el historial de compras del cliente, y podría indicar un

caso de fraude (robo o falsificación de la tarjeta). En este caso, podría saltar la alarma en la entidad bancaria, que podría decidir bloquear el pago o contactar con el cliente.

En capítulos anteriores no hemos mencionado directamente técnicas de detección de anomalías, aunque las hemos aproximado. Evidentemente, podríamos dar una solución a este problema empleando clasificación binaria, donde la clase puede ser «no-anomalía» o «anomalía». El principal problema es que en muchos casos no habrá suficientes instancias de la clase «anomalía», o es posible que no haya ninguna en absoluto, lo que limita el empleo de esta técnica.

Cuando explicamos la técnica DB-SCAN de *clustering* basado en densidades sí hablamos del concepto de instancias anómalas, que eran aquellas que no pertenecían a ningún *cluster*. Este enfoque puede resultar interesante, pues nos permite detectar datos que no son cercanos a ninguno de los grupos que hemos identificado.

BigML permite realizar tareas de detección de anomalías. A continuación vamos a volver a hacer uso del conjunto de datos de ingresos de ciudadanos estadounidenses para detectar instancias anómalas. En este caso, una instancia anómala será aquella que haga referencia a un ciudadano particular, que se aleje del resto en cuanto a su información personal, académica y a su nivel de ingresos.

Lo primero que haremos será abrir el conjunto de datos «U.S' dataset». Una vez cargado haremos clic en el botón que muestra el desplegable de acciones, que es aquél que contiene el icono de las ruedas dentadas. De las opciones que se muestran escogeremos «*Configure Anomaly*».

Una vez seleccionada la opción, se mostrará el formulario de la Figura 117. La configuración básica incluye el número de anomalías a detectar y el tamaño del bosque que utiliza para construir el modelo. Finalmente, tras haber completado la configuración deseada, haremos clic en el botón «*Create anomaly detector*».

Tras unos segundos que tardará en completarse el entrenamiento del modelo, se muestra la pantalla de la Figura 118, donde se pueden ver las 10 anomalías más señaladas. Junto a cada anomalía, podemos ver el «porcentaje de anómala» y los principales atributos que contribuyen a hacer de esa instancia una anomalía.

Si hacemos clic en una de estas anomalías, en el lateral derecho se mostrarán los atributos de la instancia seleccionada. Para cada atributo se muestra su histograma, es decir, la gráfica que muestra la frecuencia de cada valor (o rango de valores en caso de atributos numéricos). En naranja se muestra el valor que toma la instancia detectada como anómala.

Si miramos cuidadosamente el resultado de la instancia que se ha considerado la anomalía más señalada de todas, nos encontramos con que se trata de una mujer (la mayoría de las instancias hacen referencia a ciudadanos varones), que trabaja para el gobierno fe-

Figura 117 – Creación de un modelo de detección de anomalías en BigML

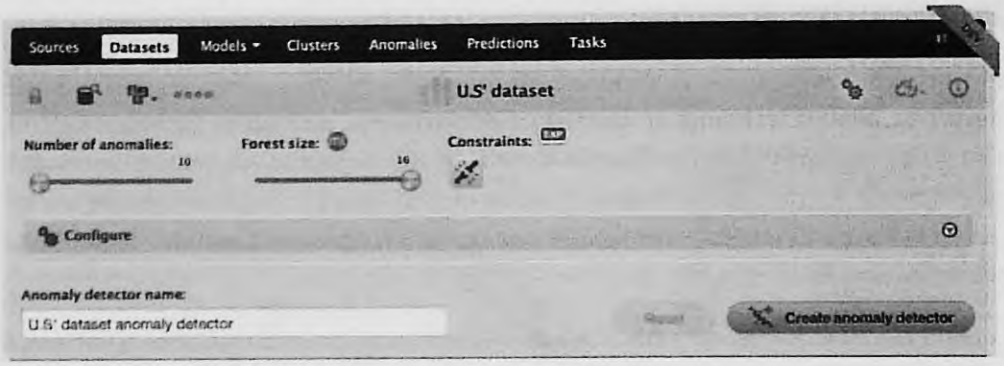
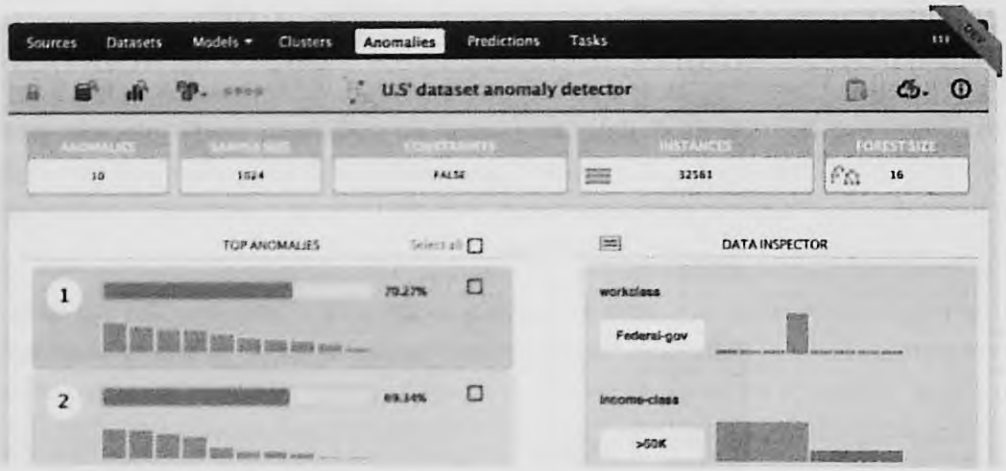


Figura 118 – Visualización de las anomalías en BigML

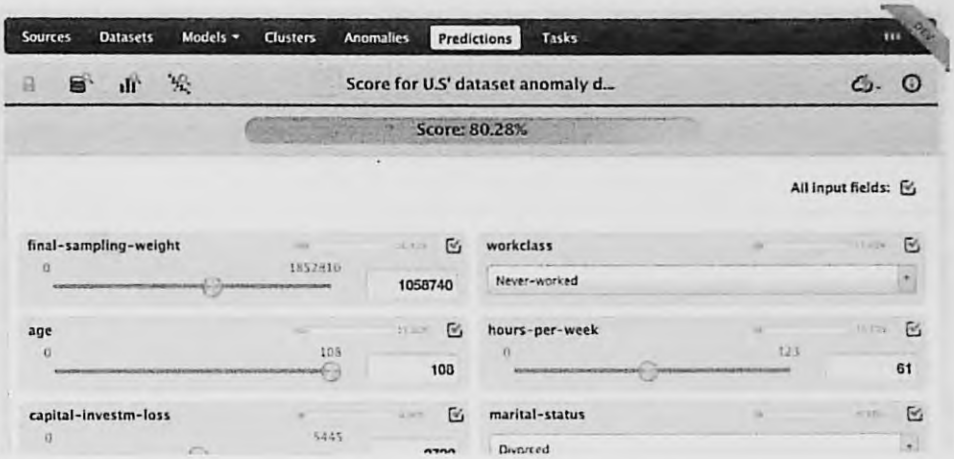


deral de los Estados Unidos (la mayoría de los ciudadanos del conjunto de datos trabajan en el sector privado) e ingresa anualmente más de 50.000 dólares, algo que tampoco es el caso más frecuente.

Si colocamos el ratón sobre el botón de acciones de la parte superior derecha (el que tiene por icono una nube con un rayo) se mostrará el listado de operaciones. Resulta interesante la opción denominada «*Anomaly Score*». Si hacemos clic en ella se cargará la pantalla de la Figura 119, donde podremos configurar manualmente cada uno de los atributos de una nueva instancia y finalmente hacer clic en el botón «*Score*» en la parte inferior, y el sistema empleará el modelo entrenado para detectar si esta es una anomalía o no, y su porcentaje de anómala.

En el ejemplo de la Figura 119 hemos optado por indicar el caso de un hombre muy mayor (de 108 años) con ingresos anuales superiores a 50.000 dólares, que no ha trabajado nunca a lo largo de su vida y con un nivel educativo muy avanzado, ya que dispone de un doctorado. Evidentemente, este caso es detectado por el sistema con una anomalía muy clara, con una puntuación del 80,28 %.

Figura 119 – Deteccion de anomalías en BigML



9 Visualización y Consumo de Datos

9.1 Visualización y Análisis de Datos: Fundamentos y Herramientas

9.1.1 Visualización de la Información

La visualización de la información consiste en la presentación y la interacción con grandes colecciones de datos. Se usa profusamente en áreas como la biotecnología, inteligencia de negocios, minería de datos, periodismo y servicios de inteligencia.

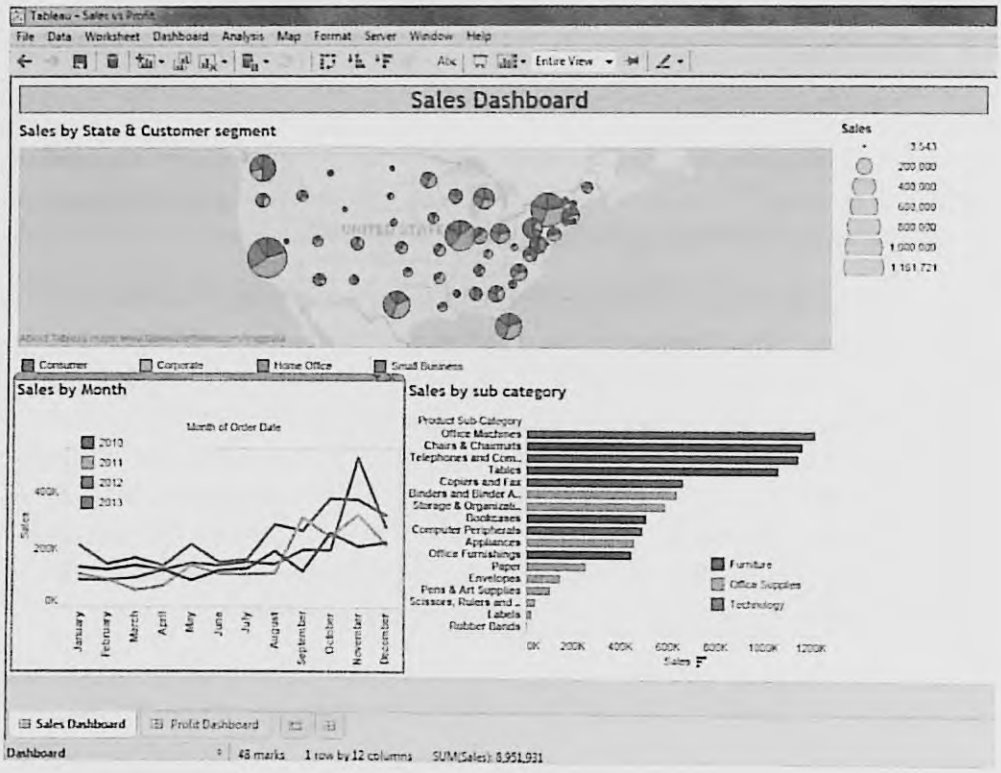
El principal objetivo de la «visualización de los datos» es representar el conocimiento de manera intuitiva y eficaz mediante el uso de diferentes gráficos. Por tanto, se usa para la transmisión de información de manera sencilla, aunque proporcionando el conocimiento intrínseco de grandes conjuntos de datos complejos. Para ello, se requiere el uso de gráficos estéticamente atractivos y funcionales. A su vez, la información a representar se abstrae mediante formas esquemáticas, incluyendo atributos o variables para las unidades de información, de manera que se obtiene información muy valiosa sobre los propios datos de cara a ser utilizada en análisis posteriores. De esta manera, se consigue un marco de análisis mucho más intuitivo que los típicos enfoques numéricos.

La visualización de datos en entornos empresariales o académicos va tomando cada vez más importancia, debido al gran número de análisis a realizar a diario por los expertos en todas las áreas de conocimiento y a la gran cantidad de información valiosa y útil de que se dispone. Hoy en día es imposible realizar tareas de análisis directamente sobre representaciones numéricas del conjunto de datos obtenido por las corporaciones. No se concibe la monitorización de grandes sistemas de manera no visual, es decir, sin la utilización de grandes paneles de gráficos, que además se actualizan normalmente con información en tiempo real. En entornos Big Data, donde el volumen de información a

analizar y monitorizar es de gran escala, la tarea de visualización de los resultados del análisis se convierte en imprescindible.

Por ejemplo, el portal de comercio electrónico eBay[173] tiene cientos de millones de usuarios activos y millones de productos vendidos cada mes y, por tanto, genera una gran cantidad de datos. Para analizar y extraer información comprensible de todos estos datos, eBay utiliza la herramienta de visualización de Big Data «Tableau»[174], que tiene la capacidad de transformar grandes cantidades de datos complejos en imágenes intuitivas. La Figura 120 muestra un ejemplo de diferentes medidas representadas en un mapa (ventas, beneficios, pedidos y costes de envío), evolución de medidas a lo largo del tiempo (beneficios, ventas, precio unitario) y una representación de las ventas dependiendo de la categoría de los productos. Estos gráficos son además interactivos lo que permite modificar aspectos de su visualización o acceder al detalle de los datos. Por ejemplo, los empleados de eBay pueden visualizar tipos de búsquedas relevantes que realizan los usuarios, y supervisar los últimos comentarios de los clientes y llevar a cabo análisis de sentimientos, preferencias y tendencias globales.

Figura 120 – Ejemplo de panel de monitorización de ventas mediante la herramienta Tableau



[173] http://www.ebay.com/

[174] *Business Intelligence and Analytics | Tableau Software*. http://www.tableau.com. [Online; consultado el 15 de diciembre de 2015]

Así pues, la visualización de los datos nos permite establecer relaciones, patrones e historias de una forma mucho más práctica e intuitiva. De esta forma, la visualización justifica el dicho popular universal «*una imagen vale más de mil palabras*». Las técnicas de visualización de datos se utilizan fundamentalmente con dos objetivos:

- Aprovechar la habilidad humana de extraer patrones a partir de imágenes.
- Ayudar al usuario a comprender más rápidamente patrones descubiertos automáticamente por un sistema de extracción del conocimiento o KDD («*Knowledge Discovery in Databases*», descrito en el Capítulo 1).

9.1.2 Fundamentos para la Visualización de Datos

Formalmente, en cuanto al análisis de datos, el fundamento matemático es el método estadístico. Este método se basa en la creación de una matriz rectangular en la que las filas representan los casos, sujetos y observaciones, y las columnas representan las puntuaciones de los atributos, variables o mediciones que permiten comparar el grado de similitud o diferencia de los actores en relación a sus atributos.

Así por ejemplo, en un análisis de datos de redes (ya sean redes bioquímicas, redes de interacción, redes de comunicación, etc.), la matriz tiene los mismos elementos en las filas que en las columnas, dado que lo que analiza son las relaciones (vínculos) entre actores (nodos) y no sus atributos. El objeto de análisis, en este caso, son los vínculos que tiene cada uno de los nodos para descubrir qué vínculos tienen y cómo se comportan cada uno de ellos dentro de una muestra seleccionada. Los resultados, por tanto, son un muestreo estadístico.

En cambio, por ejemplo, en el análisis de datos de las redes sociales, el objeto son los vínculos que se producen entre los usuarios, su tipo y su comportamiento. Para ello se utiliza un censo poblacional ya que si se ha elegido un nodo, para que el análisis sea correcto, deben ser analizados la totalidad de sus vínculos con el resto de usuarios.

La visualización, desde el punto de vista matemático, puede ser descrita como una representación no modificada de la información con técnicas de presentación de grafos, entendiendo por grafo un conjunto de puntos, cuyos vértices están unidos por aristas (líneas), la unión de los cuales forman una estructura de red. La importancia en la visualización de este tipo de estructuras no está en la forma geométrica resultante, sino en la manera en la que los vértices se conectan.

Desde un enfoque informático, la visualización de los datos se describe como la organización de la información en categorías, en un grado creciente de complejidad, bajo ciertos esquemas o estructuras (línea, árbol, estrella, anillo, malla, etc.), de acuerdo al orden y jerarquía en que se vinculan sus componentes y bajo unos criterios de dependencia, subordinación y recorrido.

En este contexto, el desarrollo continuo de la informática nos ha llevado a reconocer que la información es algo modificable y sujeto a cualquier tipo de transformación posterior. Esta libertad en su tratamiento y la mejora de la usabilidad de las herramientas, han propiciado lo que conocemos como superabundancia de información, con una carga visual y cognitiva que supera la capacidad humana para procesarla de modo eficiente en un espacio de tiempo breve. Es por ello que el análisis de datos no estructurados y generados de manera ininterrumpida adquiere una gran importancia, aportando información en tiempo real especialmente útil para las empresas e instituciones. Este es el contexto que abarca el Big Data.

9.1.2.1 Características básicas de la representación gráfica de la información

A principios de la década de los ochenta, el Profesor Edward Tufte explicó que los principios de diseño de la información gráfica deben apoyar las tareas de análisis, mostrando en última instancia la comparación o la causalidad[175]. En su libro de 1983, «La representación visual de la información cuantitativa»[175], Edward Tufte define una serie de «gráficos» y de principios para la visualización gráfica eficaz de manera que la excelencia en los gráficos estadísticos se compone de ideas complejas comunicadas con claridad, precisión y eficiencia. La representación gráfica debería por tanto seguir los siguientes principios:

- Mostrar los datos.
- Inducir al espectador a reflexionar sobre lo sustancial (en lugar de sobre la metodología, el diseño gráfico y la tecnología de la producción gráfica).
- Evitar distorsionar lo que los datos tienen que decir.
- Presentar una gran cantidad de datos en un espacio pequeño.
- Dar coherencia a grandes conjuntos de datos.
- Animar a realizar comparativas visuales de diferentes conjuntos de datos.
- Revelar los datos en varios niveles de detalle, desde una visión amplia a una vista con todos los detalles.
- Servir a un propósito razonablemente claro: descripción, exploración, tabulación o decoración.
- Estar estrechamente integrada con las descripciones estadísticas y verbales de un conjunto de datos.

Los gráficos pueden ser más precisos y reveladores que los cálculos estadísticos convencionales. Sin embargo, no aplicar los principios citados puede dar lugar a gráficos engañosos, que distorsionan el mensaje o que apoyen una conclusión errónea.

[175] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2001

9.2 La Visualización para el Análisis de Datos

La idea básica subyacente en toda tarea de visualización de datos es la de permitir al usuario comprender lo que está pasando. Desde el enfoque de la minería de datos por lo general, implica la extracción de información «oculta» de una base de datos. Este proceso de comprensión puede ser, a veces, bastante complicado[176].

En la mayoría de las operaciones que se realizan de forma usual sobre una base de datos estándar, lo que el usuario ve es algo que ya se sabía que existía en la base de datos. Así, por ejemplo, un informe que muestra la distribución de las ventas por producto y región es sencillo de entender para el usuario, ya que intuitivamente se sabe que este tipo de información ya existe en la base de datos. Si la empresa vende productos diferentes en las distintas regiones, no hay ningún problema en traducir una muestra de esta información en conocimiento relevante para el proceso de negocio. La minería de datos, por el contrario, extrae información, que el usuario no conoce a priori, de una base de datos . Esta información pueden ser, por ejemplo, relaciones desconocidas hasta el momento entre los datos que se están analizando. Estas relaciones implícitas son las joyas que la minería de datos espera localizar. Esta información, desconocida por el usuario, es susceptible de mostrarse de manera visual para facilitar su comprensión.

Hay muchas maneras de representar información gráficamente. Las visualizaciones que se utilizan deben orientarse a maximizar el valor de la información que se visualiza. Para ello, es necesario que entendamos las necesidades del espectador y diseñemos la visualización teniendo en cuenta su usuario final.

9.2.1 Necesidad de Visualización de la Minería de Datos

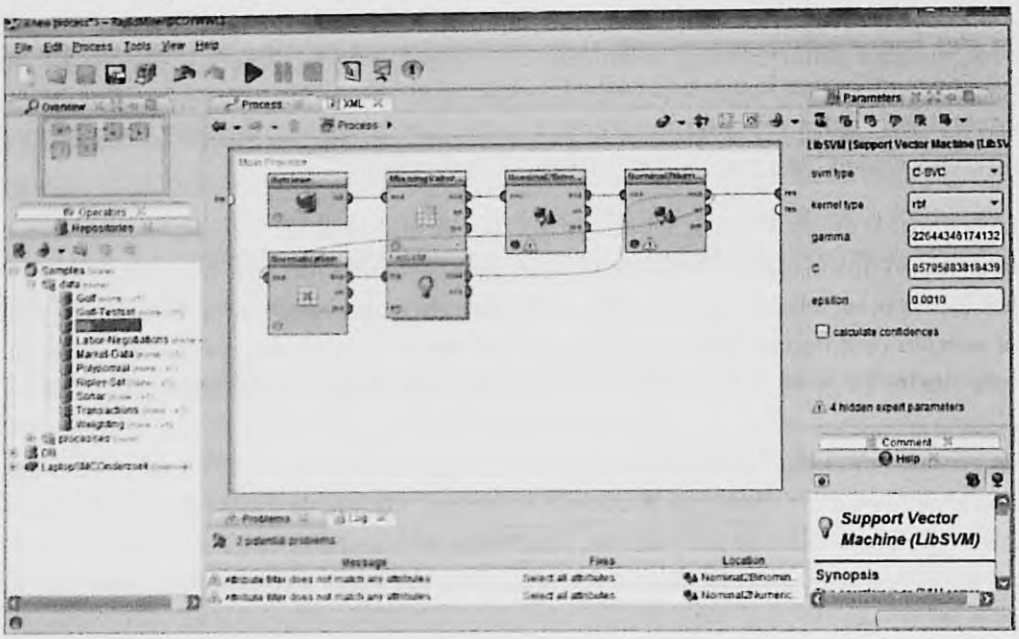
La visualización de modelos de minería de datos debe cubrir dos objetivos clave: la comprensión y la confianza. La comprensión es, sin duda, la motivación fundamental detrás de la visualización del modelo (por ejemplo, los modelos básicos de predicción).

La forma más interesante de utilizar un modelo de minería de datos es conseguir que el usuario entienda lo que está pasando para que pueda actuar directamente y con la mayor rapidez posible. La visualización debe permitir al usuario discutir y explicar la lógica que existe detrás del modelo a sus propios colegas, clientes y otros usuarios. Además, el modelo debe entenderse de modo que las acciones que se toman en base a los resultados obtenidos se puedan justificar ante los demás.

Comprender también implica contexto. Si el usuario puede entender lo que se ha descubierto en el contexto de los problemas de su negocio, entonces va a confiar en lo que descubre y a utilizar los resultados obtenidos en beneficio del negocio. En este sentido, hay dos requisitos que se deben cumplir:

[176] Kurt Thearling y col. *Information Visualization in Data Mining and Knowledge Discovery*. Ed. por Usama Fayyad, Georges Grinstein y Andreas Wierse. Morgan Kaufman, 2001

Figura 121 – Vista gráfica del generador de *workflows* de RapidMiner



1. Es necesario realizar la visualización de los resultados obtenidos con la minería de datos de una manera que se muestren los hechos relevantes.

2. Se debe permitir al usuario interactuar con la visualización para que pueda contestar a las preguntas pertinentes y despejar las dudas que se le planteen.

Por lo que respecta al primer requisito, se cuenta con soluciones creativas incorporadas en un buen número de productos comerciales y de Open Source para la visualización de modelos de minería de datos, como: Weka[177], RapidMiner[178] (en la Figura 121 se puede observar la vista típica del generador de *workflows* de aplicación de minería de datos de RapidMiner), Tableu (mostrada en el apartado anterior) y QlikView[179]. Los indicadores clave de rendimiento («*Key Performance Indicators*» - KPIs, que serán definidos en la Sección 9.5.2), su evolución y los modelos financieros (por ejemplo: beneficios, costes y retorno de la inversión) dan al usuario un sentido del contexto y facilitan que la presentación de los resultados se ajuste a la realidad.

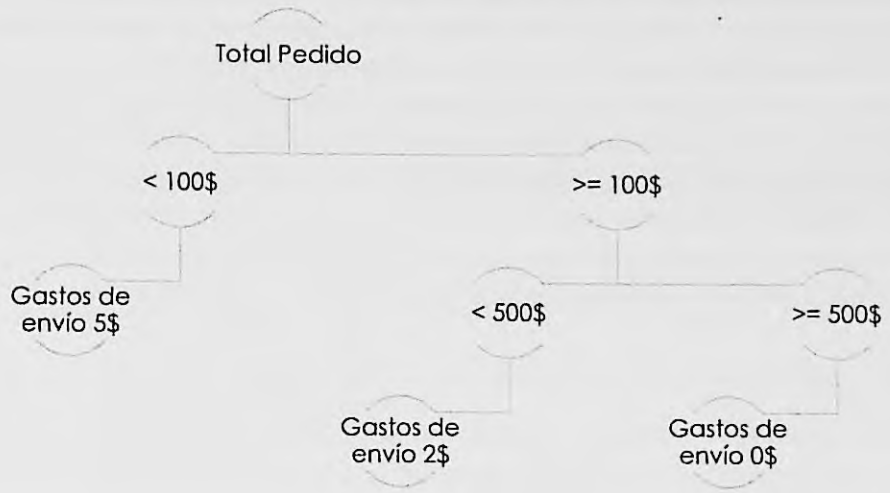
Respecto al segundo requisito, sin embargo, aún no se ha resuelto de manera adecuada. La interacción es, para muchos usuarios, la clave de la visualización en la minería de datos. La manipulación de los datos y la visualización de los resultados de manera dinámica,

[177] *Weka 3: Data Mining Software in Java*. http://www.cs.waikato.ac.nz/ml/weka/. [Online; consultado el 15 de diciembre de 2015]

[178] *RapidMiner – #1 Open Source Predictive Analysis Platform*. https://rapidminer.com. [Online; consultado el 15 de diciembre de 2015]

[179] *Business Intelligence | Data Visualization Tools | Qlik*. http://www.qlik.com. [Online; consultado el 15 de diciembre de 2015]

Figura 122 – Ejemplo de árbol de decisión



permite al usuario tener una idea de como cambian estos y comprobar si está pasando algo contrario a la intuición. La interactividad ayuda a lograr este objetivo. Por ejemplo, la visualización de un árbol de decisión (ver ejemplo en la Figura 122) es un tipo de modelo fácilmente representable incluso a gran escala, pero lo que realmente quiere hacer el usuario es arrastrar y soltar los segmentos en un mapa con el fin de poder responder al interminable número de preguntas del estilo, «qué pasaría si». Por ejemplo, que efecto económico habría tenido si no se hubiera cobrado gastos de envío a todos los pedidos de más de 100$.

La integración con herramientas de apoyo a la toma de decisiones (por ejemplo, OLAP) permitirá a los usuarios visualizar los resultados de la minería de datos. En la actualidad existen herramientas integrales para el «*Business Intelligence*» que intentan abordar este problema de la interacción con los datos de manera satisfactoria, como las plataformas SAP[180] y Pentaho[181].

9.3 Nuevos Paradigmas de Visualización de Datos

Cuando se hace referencia al papel fundamental que desempeña la visualización de datos, no solo nos referimos a las herramientas de visualización utilizadas en un contexto estrictamente empresarial, si no que su uso se amplía concediendo a la interfaz gráfica un rol protagonista en ámbitos aparentemente tan alejados de los negocios como el del arte y la cultura.

[180] *Data Visualization Design Studio | BI | Analytics | SAP*. http://www.sap.com/pc/analytics/business-intelligence/software/design-studio/index.html. [Online; consultado el 15 de diciembre de 2015]

[181] *Pentaho | Data Integration, Business Analytics and Big Data Leaders*. http://www.pentaho.com. [Online; consultado el 15 de diciembre de 2015]

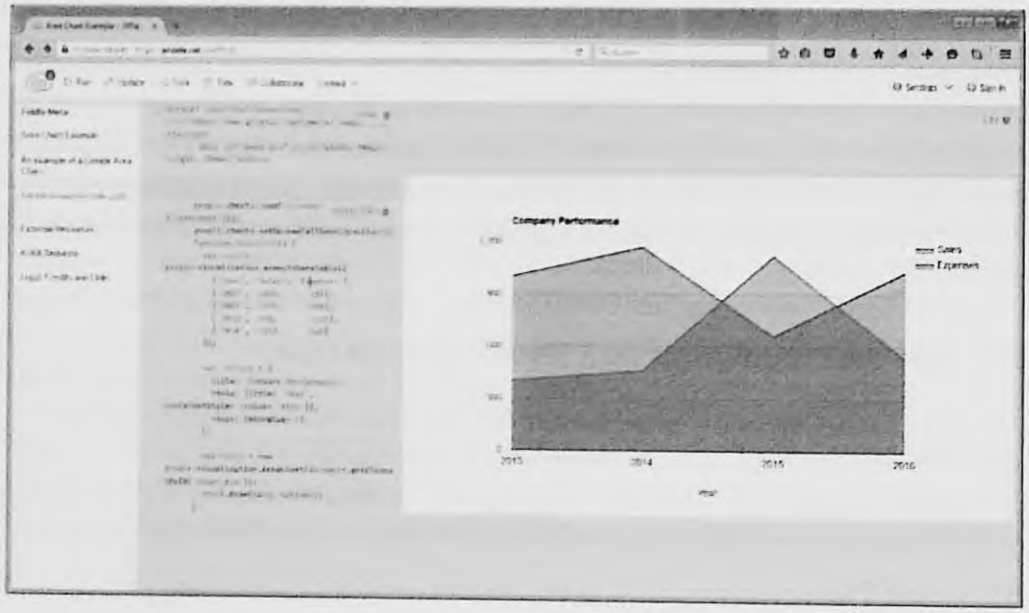
La tendencia actual en las herramientas de visualización se basa en gran medida en el uso de librerías software para el soporte al desarrollo de visualizaciones «a medida», sin las restricciones del tipo de gráficos disponibles en las aplicaciones de escritorio. Si bien estas librerías software requieren cierto conocimiento de programación, su uso se acerca cada vez más al usuario final. En este sentido, cuentan con el soporte de manuales para usuarios «no expertos». Estas guías de introducción incluyen multitud de ejemplos sencillos e interactivos. Prácticamente todas estas librerías software para la visualización mantienen comunidades abiertas de desarrolladores que generan continuamente documentación de soporte. Todo esto facilita el empleo de las nuevas herramientas de visualización, incluso a usuarios sin conocimiento de programación informática.

9.3.1 Librerías para la Visualización de Datos

Existe una amplia gama de librerías disponibles para ayudar al desarrollador a crear sus propias visualizaciones. Un conjunto de las más utilizadas y populares son las siguientes:

- Google Chart **Tools**[182]. Es una herramienta de Google Developers que permite la creación de gráficas en forma de imágenes en diferentes formatos, como por ejemplo PNG. Su funcionamiento se basa en peticiones HTTP a una determinada URL. Esta librería incluye ejemplos para diversos tipos de visualizaciones y la posibilidad de probarlas en un entorno interactivo[183]. Así por ejemplo podemos

Figura 123 – Ejemplo de representación de datos con Google Chart



[182] *Charts | Google Developers*. http://chart.apis.google.com. [Online; consultado el 15 de diciembre de 2015]

[183] https://jsfiddle.net/mhft5t3t/

cargar una representación de áreas (Figura 123), y comprobar como podemos cambiar la representación si cambiamos los datos usados (Figura 124).

Figura 124 – Ejemplo de representación de datos con Google Chart tras modificar los datos usados

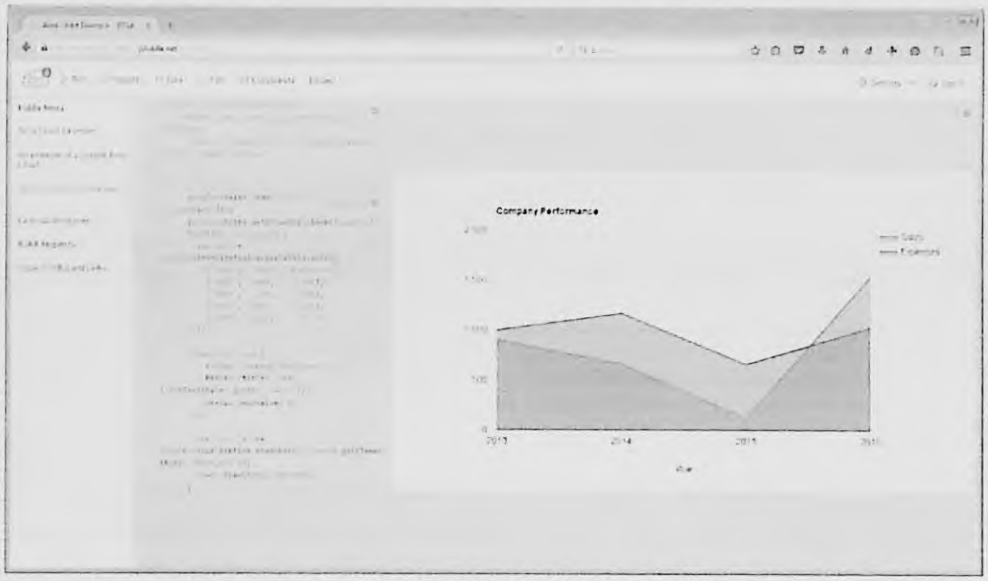
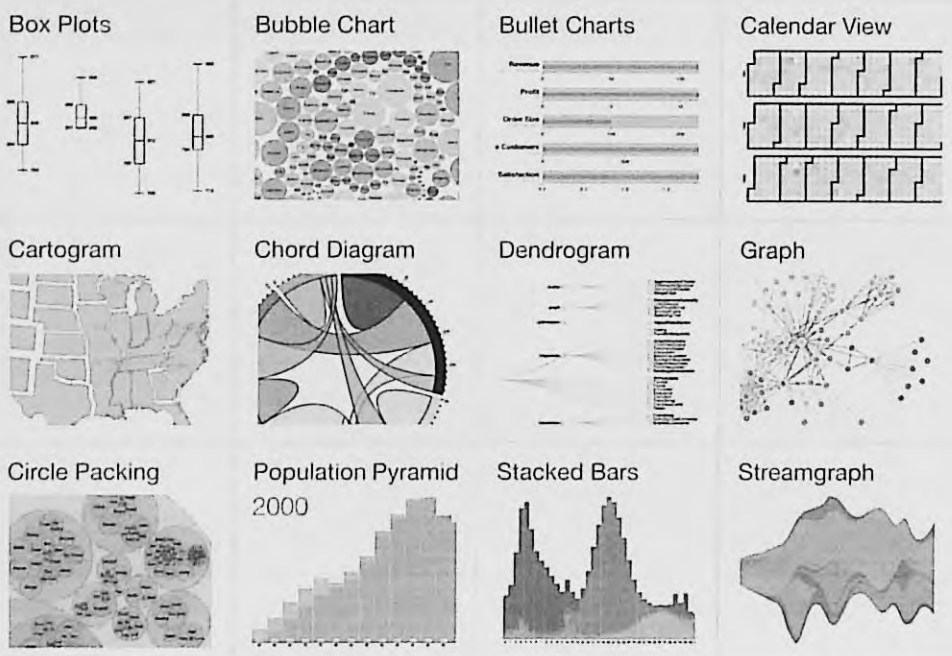


Figura 125 – Selección de gráficos de la galería de D3.js



- D3.js[184]. Es una biblioteca (JavaScript) que permite crear visualizaciones complejas y gráficos interactivos. Esta librería permite manipular documentos basados en datos usando estándares abiertos de la Web (como HTML, CSS y JavaScript). La gran ventaja de las librerías basadas en JavaScript reside en que los navegadores pueden crear visualizaciones complejas sin depender de un software propietario. Esta librería dispone de ejemplos muy atractivos, como los mostrados en la Figura 125, sobre los que el usuario puede actuar de manera dinámica.

Figura 126 – Ejemplo de carga de datos en Recline

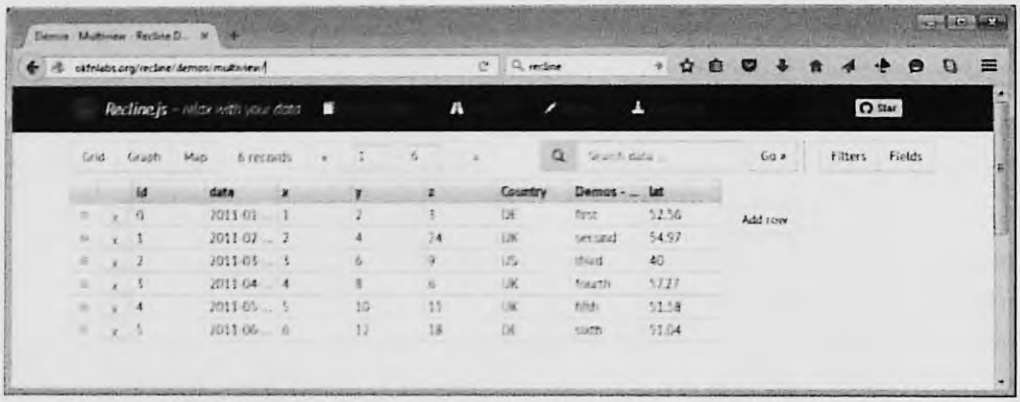
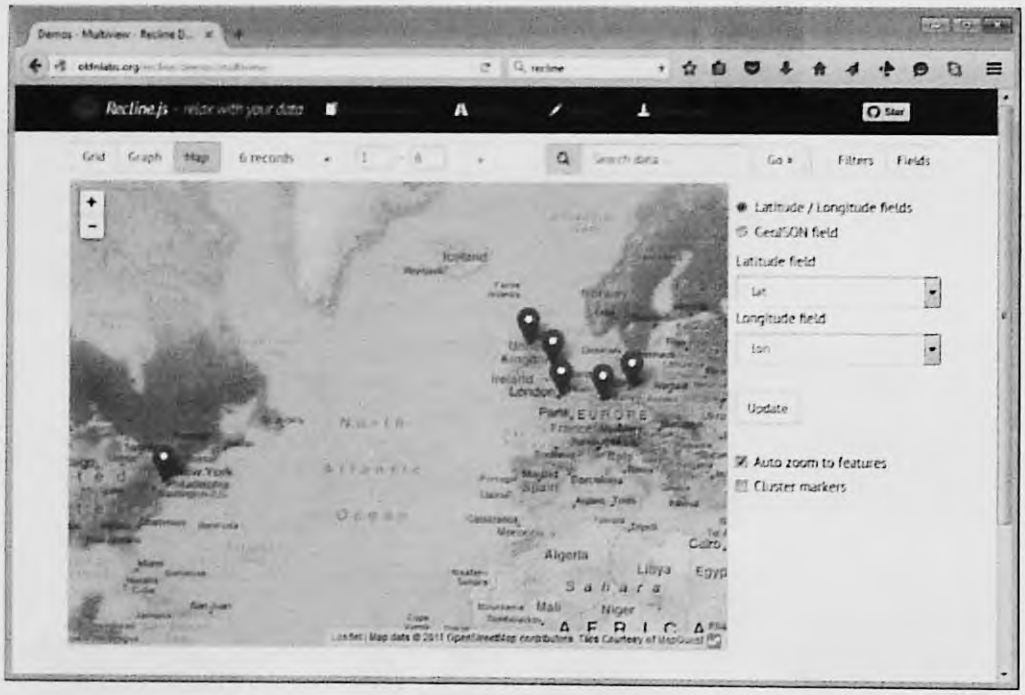


Figura 127 – Ejemplo de visualización de datos en un mapa en Recline



[184] *D3.js – Data-Driven Documents*. http://d3js.org. [Online; consultado el 15 de diciembre de 2015]

- Recline.js[185]. Es una biblioteca para el desarrollo de aplicaciones basadas en HTML y JavaScript. Está diseñada para facilitar la integración, por lo que es fácil de integrar en otros sitios Web y aplicaciones. Orientada a desarrolladores sin grandes conocimientos de programación, utiliza una interfaz sencilla para la vista (y edición) de datos. Esta biblioteca ofrece una serie de ejemplos que evidencian de su capacidad visual[186] donde podemos modificar los datos de entrada (Figura 126) y obtener gráficos de diferente tipo, como por ejemplo representaciones sobre un mapa (Figura 127).

9.3.2 Herramientas Comerciales Orientadas a Visualización

En el terreno de las aplicaciones comerciales para el tratamiento y visualización de datos en entornos Big Data, existen importantes soluciones integrales que ofrecen módulos para la visualización. La utilización de estos módulos es interactiva y están especialmente diseñados para usuarios expertos en inteligencia de negocio, sin necesidad de conocimientos de programación informática. Dos herramientas de éxito en el mercado actual son Pentaho y Zoomdata, que se describen a continuación.

- **Pentaho**[181] supone ya una solución de referencia en el Business Intelligence, ya que vincula firmemente la integración de datos con completas analíticas de negocio para el Big Data, soportando Hadoop, NoSQL y bases de datos analíticas. Esta herramienta ofrece una solución completa de análisis del Big Data que soporta todo el proceso de análisis de datos desde los procesos ETL y la integración de los datos al análisis en tiempo real y la visualización del Big Data. Pentaho da soporte completo desde el acceso a los datos a la toma de decisiones.

- **Zoomdata**[187] (véase la Figura 128) ofrece una solución integral como aplicación de *Business Intelligence* para entornos Big Data. Mientras otras herramientas utilizan componentes propietarios para el servicio de datos, Zoomdata utiliza la capacidad de las tecnologías Apache Spark[188] (descrito en la Sección 7.4) y Spark DataFrame[189] para el cálculo de datos a través de consultas incluyendo la agregación, filtrado y manejo de datos en caché. Esto permite a Zoomdata utilizar los recursos de los *clusters* Spark para realizar operaciones de inteligencia de negocio y visualizaciones en tiempo real, a gran velocidad y escala. En

[185] *Recline Data Explorer and Library*. http://okfnlabs.org/recline/. [Online; consultado el 15 de diciembre de 2015]

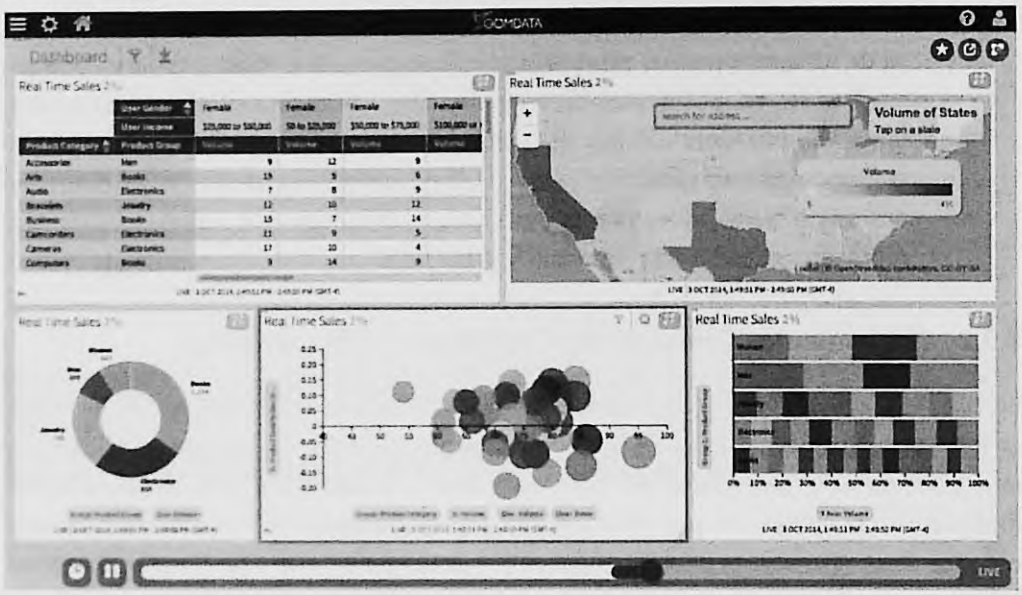
[186] http://okfnlabs.org/recline/demos/multiview/

[187] *Big Data Exploration, Visualization, Analytics*. http://www.zoomdata.com. [Online; consultado el 15 de diciembre de 2015]

[188] *Apache Spark – Lightning-Fast Cluster Computing*. https://spark.apache.org. [Online; consultado el 15 de diciembre de 2015]

[189] *Spark SQL and DataFrames*. https://spark.apache.org/docs/1.5.2/sql-programming-guide.html. [Online; consultado el 15 de diciembre de 2015]

Figura 128 – Ejemplo de vista del panel principal de visualización de Zoomdata



cuanto a la cantidad de gráficos que ofrece cuenta con la gran mayoría de los ofrecidos por D3.js[184], por lo que se pueden componer paneles de visualización con diferentes vistas de datos.

9.4 Diseño de Informes

El diseño de informes permite ofrecer informes de distintos tipos según el uso que se vaya a hacer de los mismos. Nos podemos encontrar con informes de operaciones en una empresa, incluyendo posibles datos financieros como meros resúmenes de la actividad financiera de la empresa. Sin embargo, es posible diseñar informes que permitan extraer de estos datos operacionales información de utilidad para las personas encargadas de la toma de decisiones. Estos informes darán soporte a su trabajo diario de seguimiento de las actividades de la empresa y de las acciones o decisiones tomadas previamente.

Por tanto, el uso de informes es una parte fundamental de las actividades de mejora en la inteligencia de negocio y la gestión del conocimiento. El diseño de informes es fundamental para ofrecer la información que se ha estructurado a través de procesos ETL, de forma que las personas encargadas de la toma de decisiones puedan realizar estas tareas de forma eficiente y basándose en datos de calidad. Estos informes se han ofrecido tradicionalmente a través de versiones impresas de los mismos. Con el auge de las tecnologías informáticas y de las comunicaciones en la actualidad se accede a ellos mayoritariamente a través de la Intranet de la empresa y se presentan con un gran soporte gráfico.

En definitiva podemos decir que el objetivo del diseño de los informes es producir vistas de los datos que sintetizan las operaciones de la empresa (extraídos de diferentes fuentes

de información), que sean comprensibles por aquellas personas a cargo de los procesos de toma de decisiones.

9.4.1 Componentes del Sistema de Informes

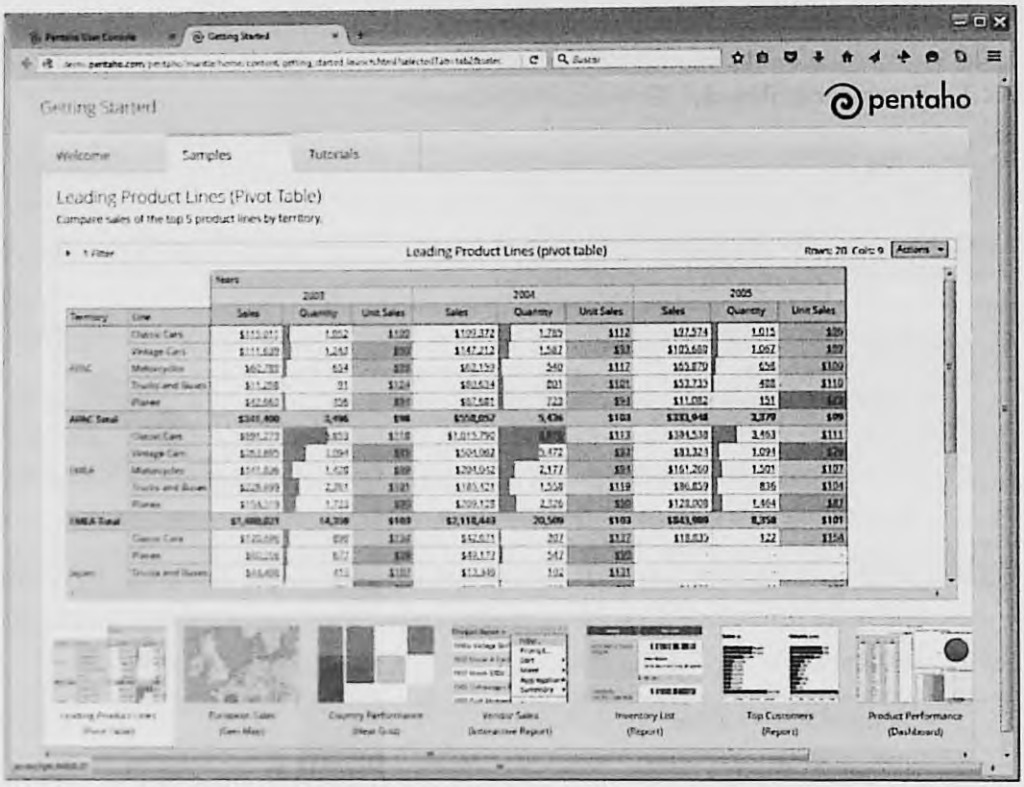
Aunque cada sistema de generación de informes puede tener una estructura distinta, que dependerá en gran medida de las necesidades de la empresa y las de fuentes de información disponibles, es posible identificar algunos elementos en común:

- **Instrumentación.** Son los dispositivos que toman medidas y generen eventos.
- **Provisión de datos.** Los sistemas que recogen datos necesitan normalmente otro componente encargado de su provisión. La provisión de datos puede realizarse con diversas estrategias: envío directo, transferencia periódica o cuando se acumule una cantidad determinada de datos.
- **ETL.** Los datos proporcionados por el componente anterior pueden ser analizados para controlar y asegurar su calidad y guardarse en el almacén de datos de la empresa.
- **Repositorio de datos.** Los datos que se generan y que serán usados en los informes podrán estar almacenados en cualquier soporte como ficheros de datos, aunque lo más frecuente es que estén en sistemas estructurados como bases de datos relacionales, almacenes de datos, etc.
- **Lógica de negocio.** Este componente se encarga de procesar los datos almacenados para generar las métricas que serán usadas en los informes.
- **Publicación.** El sistema construirá los informes a partir de los datos disponibles y ofrecerá algún mecanismo para su uso. La publicación en un repositorio permitirá a los usuarios acceder a estos informes. Estos informes pueden igualmente enviarse directamente a los usuarios.
- **Calidad.** Normalmente estos sistemas incluyen mecanismos que permitan conocer la calidad de los informes, como por ejemplo: encuestas, auditorías, etc.

Pentaho ofrece una herramienta de demostración online de su sistema de generación de informes. Para ello es necesario registrarse en su Web[190]. Una vez que introducimos nuestros datos de contacto tenemos la posibilidad de acceder al sistema con la cuenta de invitado (usuario y clave «*guest*»). La Figura 129 muestra un ejemplo de informe de ventas por año, territorio y línea de producto. Estos informes permiten al usuario además explorar los datos que han dado lugar a la representación mostrada. Así, pulsando sobre

[190] *Experience Pentaho Business Analytics | Pentaho Hosted Demo*. http://www.pentaho.com/hosted-demo. [Online; consultado el 15 de diciembre de 2015]

Figura 129 – Ejemplo de las perspectivas del cuadro de mando integral con Pentaho



los enlaces de la tabla de datos podemos acceder al detalle de venta para una región y línea de producto en un año concreto. Cómo se puede ver en la imagen para los datos de muestra de esta demo disponemos de diferentes tipos de representaciones.

9.5 El Cuadro de Mando Integral y los KPIs

9.5.1 El Cuadro de Mando Integral

Conforme las empresas se van transformando para competir en un contexto en el que la información es clave para mantener una situación de ventaja competitiva, la explotación de elementos intangibles se torna más decisiva incluso que la capacidad de invertir o gestionar los elementos tangibles. Se denomina Cuadro de Mando Integral (CMI) a una herramienta de gestión empresarial muy útil para medir la evolución de la actividad de una compañía y sus resultados, desde un punto de vista estratégico.

Esta aproximación se basa en tomar medidas que vayan más allá de las medidas financieras. Las métricas que no son financieras nos pueden permitir realizar predicciones sobre el rendimiento futuro, produciendo algo más que informes de la situación actual.

El objetivo es enlazar las acciones inmediatas que realicemos con los objetivos generales de la empresa a medio-largo plazo. El cuadro de mandos integral se basa en la visión de la empresa desde cuatro perspectivas (Figura 130):

Figura 130 – Perspectivas del cuadro de mando integral

- **La formación y el crecimiento.** Esta perspectiva incluye al empleado y su formación, dado que los empleados son una fuente de conocimiento de gran valor para la empresa. Es por tanto de vital importancia, en un contexto altamente competitivo y en constante evolución, que esta fuente de conocimiento (los empleados) se mantenga actualizada y por tanto implicada en un proceso continuo de aprendizaje y reciclaje.

- **Los procesos internos.** Esta perspectiva se refiere a métricas relacionadas con los procesos internos de la empresa y por tanto deberán ser los gerentes que conozcan estos procesos los que las definan.

- **Los clientes.** Esta perspectiva se centra en los clientes y las medidas que permitan medir el grado de satisfacción con nuestra empresa. Estos indicadores son especialmente relevantes, ya que si los clientes no están satisfechos con nuestros productos o servicios es posible que se busquen otra empresa que ofrezca un producto o servicio similar.

- **Finanzas.** Esta aproximación incluye perspectivas complementarias a la métrica del área financiera, facilitando información que permita relacionar las finanzas con las otras tres perspectivas.

La visión global que ofrezca el cuadro de mando deberá propiciar una estrategia de gestión basada en los siguientes cuatro procesos:

- **Compartir la visión.** Este proceso permite a los gestores llegar a un consenso sobre la visión y estrategia de la empresa. Se trata de traducir los objetivos de

alto nivel de los directivos (ser los mejores del sector, ser el principal proveedor, etc.) a términos operacionales que permitan realizar acciones concretas. Para ello es necesario expresar estos grandes objetivos junto a un conjunto de medidas, que describan el éxito a largo plazo y que sean aceptadas por todos los ejecutivos senior.

- **Comunicar y enlazar**. Este proceso permite a los gestores comunicar su estrategia hacia arriba y abajo en la organización, así como enlazar esta estrategia con objetivos departamentales e individuales. Tradicionalmente los departamentos son evaluados en términos de desempeño financiero y los incentivos individuales se basan en base a objetivos financieros a corto plazo. Los CMI dan la oportunidad a los gestores de hacer llegar la estrategia a largo plazo a todos los niveles de la empresa, para que comprendan esta estrategia y se alineé con los objetivos a corto plazo.

- **Plan de negocio**. Este proceso permite a las empresas combinar sus planes financiero y de negocio. Cuando se hace uso de los objetivos planteados en el CMI como base para decidir las prioridades y donde destinar recursos, es posible enfocarse en aquellas iniciativas que van en la dirección de los objetivos a largo plazo. Los objetivos estratégicos a largo plazo deberán traducirse a objetivos y medidas operacionales concretas para unidades, departamentos e individuos.

- **Realimentación y aprendizaje**. Este proceso proporciona a las empresas la capacidad para el aprendizaje estratégico. Con el uso de CMI, la empresa puede monitorizar los resultados a corto plazo desde las cuatro perspectivas antes citadas y evaluar la estrategia en base a medidas de desempeño a corto plazo. Es posible por tanto diseñar las estrategias de forma que se vayan adaptando en tiempo real.

El propio proceso de crear un CMI fuerza a las empresas a integrar su plan estratégico y los procesos financieros, por lo que se asegura que los presupuestos den soporte a las estrategias. Los usuarios de los CMI deberán definir métricas de progreso en cada una de las cuatro perspectivas y establecer objetivos para cada una de estas métricas. En base a estas métricas se definirán las acciones concretas a realizar para cumplir con los objetivos planteados y se definirán hitos intermedios que permitan realizar un seguimiento y determinar si se va avanzando en la dirección correcta.

❖ ***Los tres primeros procesos (compartir la visión, comunicar y enlazar, y plan de negocio) son vitales para implementar la estrategia, pero ¿son suficientes en un entorno altamente competitivo y cambiante como en el que comúnmente se desarrollan las actividades de negocio?***

Estos tres procesos deben concebirse como instrumento que proporciona un proceso de aprendizaje, en el que los objetivos permanecen constantes. Por ello es muy importante

recurrir a la realimentación y al aprendizaje para detectar cualquier posible desviación sobre el plan establecido y poner en marcha iniciativas para corregir estas desviaciones.

9.5.2 Los KPIs

Un KPI (del inglés «*Key Performance Indicator*») es un indicador clave de rendimiento (o desempeño), que mide lo bien o mal que está funcionado un proceso empresarial. Estos indicadores deben estar asociados a un objetivo que se habrá fijado de antemano. Estas medidas pueden expresarse como un porcentaje de cumplimiento de dicho objetivo. Los KPIs deberán cumplir con los siguientes requisitos:

- **Relevancia.** Deben ser medidas importantes para el éxito de la operación comercial. Si no hay relación directa entre desempeño óptimo y la métrica, esta última no es un KPI.
- **Medible.** Los KPIs deben ser claramente medibles, cuantificables y manejables por el personal responsable de la organización.

El uso de los CMIs aborda una forma de conectar los planes estratégicos a largo plazo con la medida del funcionamiento de la empresa a corto plazo. Sin embargo, existen opiniones[191] que indican que esto no es suficiente sin una metodología para definir las métricas o KPIs que se usen en los CMI.

9.6 Información Geográfica

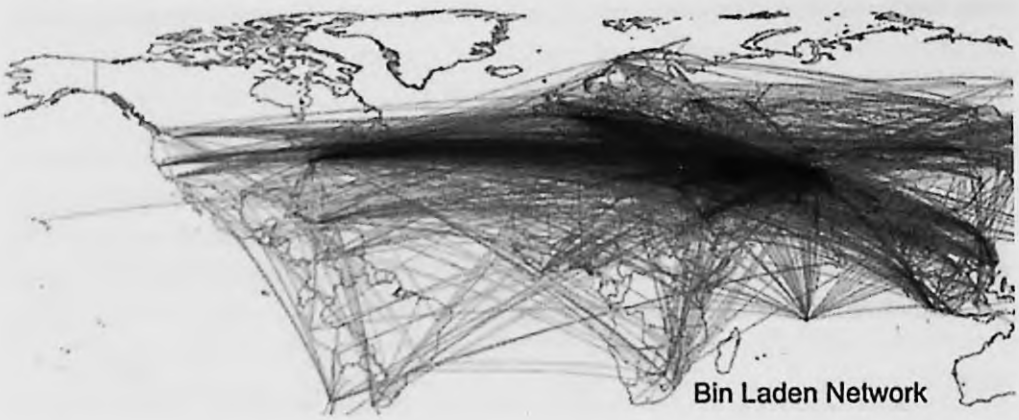
El término Big Data saltó de los foros más técnicos y comenzó a emerger en las noticias en 2011 cuando las técnicas de inteligencia tradicionales no eran capaces de localizar a Osama Bin Laden. Fue un estudio de análisis espacio-temporal de datos provenientes de noticias el que arrojó ciertas pistas sobre su localización[192]. Se analizaron noticias en las que aparecía mencionado Osama Bin Laden y sus localizaciones, incluyendo la relación espacial entre aquellas localizaciones incluidas en la misma noticia. El resultado fué un mapa de conexiones que apunta a un punto de un radio de 200 kilómetros. Lo interesante de este caso es que, además, la localización que se infirió en este estudio (Abbottabad)[193] sólo se menciona en una noticia de forma directa (Figura 131).

[191] KPI Library. *Why the Balanced Scorecard IS NOT a KPI Measurement Tool.* http://kpilibrary.com/topics/why-the-balanced-scorecard-is-not-a-kpi-measurement-tool. [Online; consultado el 15 de diciembre de 2015]

[192] Kalev H. Leetaru. «Culturomics 2.0: Forecasting Large-Scale Human Behavior Using Global News Media Tone in Time and Space». En: *First Monday* 16.9 (2011)

[193] Kalev H. Leetaru. «Culturomics 2.0: Forecasting Large-Scale Human Behavior Using Global News Media Tone in Time and Space». En: *First Monday* 16.9 (2011)

Figura 131 – Mapa de relaciones espaciales de noticias relacionadas con Osama Bin Laden



9.6.1 Sistemas de Información Geográfica

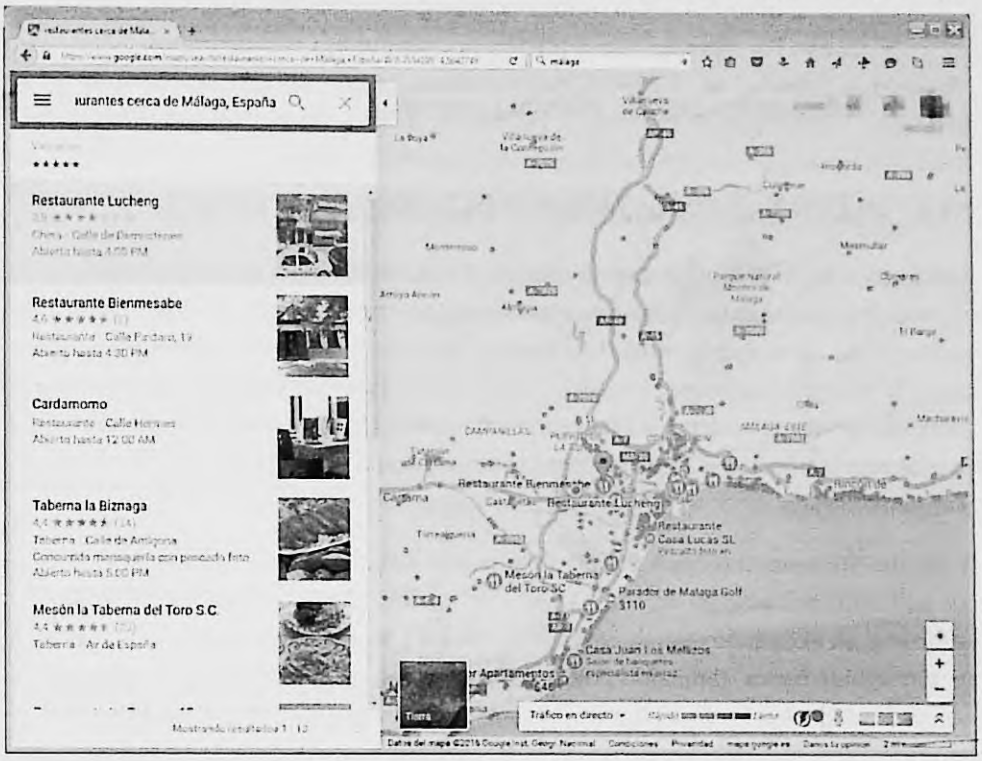
Al igual que un procesador de texto se usa para escribir documentos en un ordenador, podemos usar un Sistema de Información Geográfica (SIG) para gestionar la información espacial en un ordenador. Los Sistemas de Información Geográfica se componen de:

- **Datos Espaciales**. Estos datos incluyen posiciones espaciales, forma de los objetos referenciados, sistemas de coordenadas usados, etc.
- **Hardware**. Hardware capaz de almacenar los datos espaciales, gestionar su acceso, procesarlos y mostrarlos.
- **Software**. El acceso y uso de los datos espaciales requiere de software especializado que sea capaz de acceder a los datos espaciales y realizar operaciones sobre los mismos (localizarlos, calcular distancias, calcular relaciones espaciales entre diferentes objetos, etc.).

Los SIG requieren tener acceso a mapas digitales, sobre los que se posicionará la información espacial. Estos sistemas suelen usar el mapa como una capa base sobre la que se van añadiendo capas de datos espaciales, de forma que podemos ir mostrando u ocultando capas de los datos sin cambiar los datos usados. Por ejemplo, podemos buscar en Google Maps «Málaga», seleccionar la vista de «Restaurantes» y posteriormente queremos ver el tráfico. Esto nos muestra un mapa de fondo (el mapa de la zona que almacena y gestiona Google) y dos capas de datos superpuestas (Figura 132): puntos mostrando las localizaciones de los restaurantes y líneas de colores indicando el tráfico típico para el ámbito temporal indicado.

Los sistemas de Información geográfica hacen uso de datos espaciales. Los datos espaciales representan información sobre la ubicación física y la forma de objetos geométricos. Estos objetos pueden ser ubicaciones de punto u objetos más complejos como países,

Figura 132 – Ejemplo de capas de datos espaciales sobre la ciudad de Málaga en Google Maps



carreteras o lagos. La información asociada a un dato espacial la podemos dividir en dos elementos:

- **Datos geográficos**. Estos datos incluyen la ubicación del objeto en el espacio y su forma geométrica. En el ejemplo mostrado anteriormente estos datos incluirán la localización de los restaurantes y su forma geométrica se ha identificado como un punto para representar a cada uno de los restaurantes, y líneas para representar las carreteras.

- **Datos no geográficos**. Estos datos incluyen información sobre el objeto que no dependen de su posición o forma. Así, por ejemplo, si se desear analizar la escolarización por ciudades, usando Google Maps se mostraría cada ciudad, y los datos no geográficos serían el año y la cantidad de niños en edad de escolarización en esa ciudad. En el caso de los restaurantes de Málaga, la información no geográfica sería el nombre del restaurante, la valoración de los clientes, la foto, etc.

Los datos geográficos normalmente utilizan datos vectoriales como pares de valores X e Y indicando coordenadas. Estos datos vectoriales pueden expresarse mediante tres tipos de objetos espaciales:

- **Puntos**. Por ejemplo, ciudades, accidentes geográficos puntuales y hitos.
- **Líneas**. Por ejemplo, líneas telefónicas, carreteras y vías de trenes.
- **Polígonos**. Por ejemplo, edificios y parques.

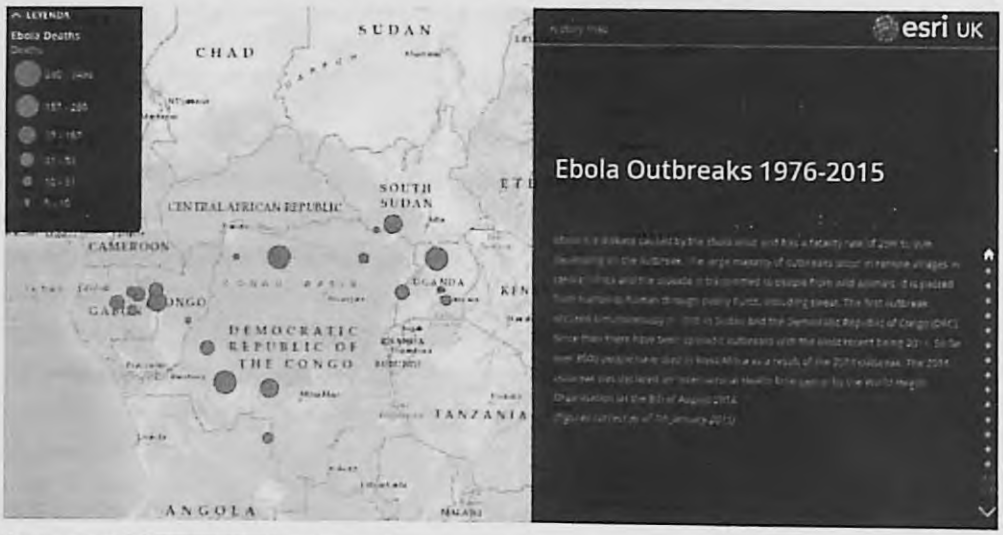
9.6.2 Visualización de Información Geográfica en el Big Data

Hemos visto cómo la representación espacial de los datos facilita la interpretación de los mismos. Este hecho gana mayor importancia en el contexto del Big Data ya que en muchas ocasiones los datos que tenemos que analizar incluyen muchas dimensiones y tamaños muy grandes que hacen que el análisis directo de los datos sea complejo. Por tanto, una aproximación complementaria a las representaciones visuales (gráficos, diagramas, etc.) es usar representaciones visuales como capas de datos en un sistema de información geográfico si estos datos llevan asociados datos espaciales.

El uso del SIG para el análisis del Big Data puede aplicarse en numerosas áreas de las que podemos destacar: modelado y análisis del cambio climático, análisis de localizaciones (como en el caso de Osama Bin Laden), ventas y comercio electrónico, financiación del terrorismo, banca, campañas políticas y elecciones, respuesta a desastres naturales, industria de la aviación, supervivencia de enfermedades, seguros y análisis de fraude.

Como caso de estudio ilustrativo, podemos destacar el uso de herramientas SIG y Big Data, para hacer un seguimiento del Ébola. La Figura 133 muestra una herramienta de Esri[194] que permite analizar la historia de los avances del Ébola (desde 1976 hasta 2014). Esta herramienta muestra datos agregados incluyendo tanto datos oficiales como datos

Figura 133 – Mapa histórico sobre la progresión del virus del Ébola



[194] *Esri – GIS Mapping Software, Solutions, Services, Map Apps, and Data*. http://www.esri.com. [Online; consultado el 16 de diciembre de 2015]

de redes sociales que permiten identificar los brotes incluso antes de que la OMS (Organización Mundial de la Salud) declare la alerta.

Como parte del caso de uso que se presenta en el Capítulo 11 se mostrará como visualizar los datos de ClickStream y los resultados del análisis. Concretamente, se realizarán representaciones de los datos con información geográfica sobre mapas usando Microsoft Excel y diagramas de barras para mostrar los datos de las edades de los clientes.

10 Seguridad y Gobernanza

10.1 Introducción

Tradicionalmente, la seguridad relacionada con los datos almacenados y gestionados por una compañía ha sido un asunto de vital importancia, por la transcendencia y repercusión que el acceso indebido, la filtración o la sustracción de datos, bien pertenecientes a clientes o usuarios, bien trascendentes y confidenciales para la compañía, puede tener tanto desde un punto de vista legal, como de competencia y reputación.

La gestión de la seguridad, sin embargo, se limitaba a un entorno controlado. Habitualmente las empresas almacenaban sus datos de forma centralizada, gestionados por equipos específicos (los responsables de sistemas y redes). Además los datos almacenados, se puede decir que también procedían de entornos controlados, sistemas operacionales de la compañía, datos solicitados a los clientes o datos específicos obtenidos con el consentimiento y el permiso de los propietarios de los mismos (como por ejemplo los datos personales solicitados a los clientes a través de formularios de la web corporativa). Lo que se traduce en que la mayoría de los datos eran datos estructurados.

Big Data ha hecho evidente que la gestión de petabytes de datos en un *cluster*, compuesto por decenas, cientos o incluso en muchos casos, miles de servidores enlazados, que sirven como repositorio de información para múltiples aplicaciones, plantea muchas cuestiones relacionadas con la protección de activos muy valiosos y apreciados como son la información y los datos.

Los problemas relacionados con la seguridad y la privacidad no han hecho más que magnificarse con el volumen, la velocidad y la variedad de los datos procedentes de Big Data,

así como el despliegue de estos sistemas en la nube, la diversidad de las fuentes de datos de las que se nutren, la entrada continuada de datos (*streaming*) o la necesidad de utilizar sistemas distribuidos.

Los métodos tradicionales de gestión de la seguridad y la privacidad ya no son aplicables, o al menos necesitan ser revisados y adaptados a las nuevas características de Big Data, tanto desde un punto de vista tecnológico como legal.

En la primera parte del capítulo, analizaremos la seguridad tanto desde el punto de vista legal, como técnico. Identificaremos que es un dato de carácter personal, por qué es importante protegerlo y su impacto en la privacidad. Asimismo analizaremos cómo se puede abordar la protección de los datos desde un punto de vista técnico, introduciendo como el cifrado de datos puede ayudarnos en esta tarea.

En la segunda parte del capítulo abordaremos otra parte importante en el trabajo con datos, todas aquellas tareas relacionadas con la gestión de los datos, concepto que se conoce con el término gobernanza o gobierno de los datos.

10.2 Seguridad

10.2.1 Big Data, Riesgos Tecnológicos y Jurídicos

La mayoría de los proyectos que hoy se utilizan para la gestión, almacenamiento o procesamiento de Big Data, surgieron como consecuencia de una necesidad tecnológica para solucionar los problemas que las tecnologías utilizadas en aquel momento no podían solucionar y casi siempre como consecuencia de proyectos académicos o procesos de innovación, nunca pensando en la creación de una solución comercial para ser vendida a clientes.

10.2.1.1 Riesgos tecnológicos intrínsecos al Big Data

Garantizar la seguridad y privacidad de sistemas de Big Data implica asegurar la infraestructura de los sistemas, los entornos de computación distribuida y los propios repositorios de datos. Para proteger los datos en sí mismos, es necesario garantizar que la difusión de la información se realiza preservando la privacidad y protegiendo los datos sensibles, mediante la aplicación de sistemas que permitan encriptarlos y la aplicación de políticas y mecanismos de control de acceso. La gestión de grandes volúmenes de datos demanda soluciones escalables y distribuidas, que permitan asegurar los almacenes de datos, permitiendo la realización de auditorías eficientes y garantizando la procedencia de los datos. Así mismo, es necesario verificar la integridad de los datos recibidos en «*streaming*» procedentes de diferentes fuentes (como los datos recibidos desde los teléfonos móviles de millones de usuarios que ejecutan un juego o una aplicación). Además, esta capacidad de recibir ingentes cantidades de datos en tiempo real puede aprovecharse

para llevar a cabo análisis en tiempo real de los incidentes de seguridad, de manera que podamos monitorizar y asegurar la buena salud de la infraestructura.

Teniendo en cuenta lo señalado de acuerdo con la CSA (Cloud Security Alliance)[195] podemos clasificar los retos tecnológicos a los que se enfrenta el Big Data en lo referente a la seguridad y la privacidad en los cuatro grandes grupos:

1. Seguridad de la infraestructura.
    a) Aseguración de los procesos de computación realizados en entornos de programación distribuida.
    b) Implantación de mejores prácticas relativas a la seguridad en entornos de almacenamiento de datos no relacionales (NoSQL, NewSQL, etc).
2. Privacidad de los datos.
    a) Preservar la privacidad de los procesos de minería de datos y análisis.
    b) Seguridad basada en el encriptado de los datos.
    c) Control de acceso granular.
3. Gestión de los datos.
    a) Asegurar los almacenes de datos y los log de las transacciones.
    b) Auditorias granulares.
    c) Procedencia de los datos.
4. Integridad de los datos y seguridad reactiva.
    a) Validación (y filtrado) de los puntos de entrada de datos.
    b) Monitorización de la seguridad en tiempo real.

10.2.1.2 Cuestiones legales en el tratamiento de Big Data

La materia prima que utiliza el Big Data son los datos. Datos de muy variada procedencia y tipología, entre los que sin duda se encuentran los datos personales o datos de carácter personal. El problema que surge con este tipo de datos es que en la actualidad la mayoría de los países disponen de una regulación para este tipo de datos y normalmente su protección está considerada dentro de los derechos fundamentales o constitucionales de los ciudadanos.

Dicho lo anterior, podemos identificar al menos un conflicto de intereses en lo que al tratamiento de datos se refiere. Por un lado tenemos los datos y los derechos asociados

[195] *Cloud Security Alliance*. https://cloudsecurityalliance.org/. [Online; consultado el 11 de diciembre de 2015]

a ellos y por otro el uso de los mismos, unas veces con fines operativos y de mejora de la eficiencia y otros con fines comerciales y de desarrollo. Gestionar este tipo de conflictos siempre pasa por legislar para tratar de encontrar un equilibrio entre ambos lados.

Hablar por tanto de Big Data y seguridad, desde un punto de vista legal, es hablar de privacidad y protección de datos de carácter personal. Y básicamente podemos adelantar que, el tratamiento de datos a través de técnicas de Big Data, pasa por no utilizar datos de carácter personal. Sin embargo, esto no significa que no podamos utilizar Big Data cuando en nuestros ficheros existan datos personales, sino que cuando estos existan deberemos tenerlo en cuenta y aplicar determinados procedimientos que veremos más adelante.

La protección de datos de carácter personal tiene su origen en los años 70[196], cuando la informática comenzaba a entrar en el mundo empresarial y comenzaba a mostrar su capacidad de gestionar datos a partir de los que se podía obtener información relevante.

Hoy en día, esta preocupación sobre los datos de carácter personal se ha extendido a la mayoría de los países del mundo, que tratan de implantar normativas y regulaciones en este sentido. Actualmente existen dos enfoques para el tratamiento de los datos de carácter personal:

- Uno más en la línea europea, en la que se trata el asunto desde un punto de vista de los derechos fundamentales de los ciudadanos.
- Y otro más en la línea de los países anglosajones como EEUU, en el que se hace una aproximación más mercantilista.

El primero es mucho más restrictivo que el segundo en cuanto al uso de los datos de carácter personal. En cualquier caso, en líneas generales, la protección de datos de carácter personal se considera un derecho fundamental de los ciudadanos y como tal tiene tres características fundamentales:

- Es un derecho irrenunciable. Es decir las personas disfrutan de este derecho de por vida y no pueden renunciar a él.
- Es un derecho inalienable. Es decir que forma parte de la persona y no puede ser negado por ninguna otra persona, gobierno o autoridad.
- Es un derecho imprescriptible. Es decir los delitos contra este derecho nunca prescriben.

Este derecho consiste básicamente en el ejercicio de dos facultades:

[196]En 1974, la revelación en la prensa francesa de un proyecto gubernamental de interconexión de todos los ficheros administrativos, en función de un número de identificación único de los ciudadanos (conocido bajo el nombre de proyecto SAFARI), creó cierta revolución y malestar en la opinión pública.

- **Facultad de control**: es decir, el individuo tienen derecho a ejercer la capacidad de decidir quién trata sus datos y para qué.
- **Facultad de disposición sobre los datos**: es decir, poder tomar decisiones sobre los datos y tener la capacidad de ejercer ciertos derechos contra quien posee sus datos (derechos como los de acceso, rectificación o cancelación).

Por tanto, los datos de carácter personal pertenecen a su titular, quien no puede dejar de serlo y las empresas, las administraciones públicas y los gobiernos o cualquier tercero, sólo pueden realizar un tratamiento de los mismos.

El tratamiento de estos datos puede hacerse única y exclusivamente en dos casos:

- Cuando el propietario de los datos nos da la autorización para hacerlo.
- Cuando una ley nos autorice.

El tratamiento de los datos de carácter personal está sujeto a una serie de obligaciones ante su tratamiento:

- En primer lugar los datos siempre son de su titular y este solamente concede el derecho a utilizarlos.
- Los datos sólo podrán ser utilizados para el fin o fines que su titular ha autorizado.
- El derecho a la utilización de los datos lleva implícita la obligación de la guarda y custodia de los datos, así como la garantía de su confidencialidad. Lo que significa que se deberá garantizar que el titular de los datos pueda ejercer los derechos de acceso, rectificación, cancelación y oposición.

Pero, ¿qué es un dato de carácter personal? Estos datos están definidos legalmente como cualquier información numérica, alfabética, gráfica, fotográfica, acústica o de cualquier otro tipo, concerniente a personas físicas identificadas o identificables.

Se considera que una información hace referencia a una persona identificada, cuando nos indica directamente a qué persona hace referencia, sin necesidad de que tengamos que hacer ulteriores averiguaciones. El ejemplo más claro de este tipo de dato podría ser el documento de identidad de una persona.

Se considera que una información hace referencia a una persona identificable cuando no indica directamente a qué persona se refiere, pero aporta información suficiente para poder llegar a averiguar su identidad. El ejemplo más claro de este tipo de dato podría

ser el ADN de una persona: sabemos que corresponde a una persona concreta, pero no conoceremos su identidad hasta que no apliquemos el procedimiento correspondiente.

Como norma general, cualquier dato puede ser considerado como de carácter personal, cuando es atribuible a una persona determinada. Normalmente un dato por si solo es difícil que pueda ser considerado como dato personal, pero hay muchos factores que pueden acabar convirtiendo un dato en dato de carácter personal. Un domicilio cualquiera no es un dato de carácter personal, si somos capaces de asociar ese domicilio con una persona determinada, este se convierte inmediatamente en un dato personal.

Los datos de carácter personal están muy relacionados con los conceptos de privacidad, intimidad y confidencialidad. Los datos de carácter personal, más allá de identificarnos como personas únicas, son capaces de revelar aspectos de nuestra vida que corresponden a nuestra zona privada y personal.

El derecho a la privacidad forma parte de la Declaración Universal de los Derechos Humanos. Esto quiere decir que es un derecho inherente a cada ser humano, que tiene independencia frente a otros factores, no puede transferirse ni se puede renunciar a él. Como el resto de los derechos humanos, el derecho a la privacidad busca garantizar la dignidad del individuo.

Privacidad es aquello que una persona lleva a cabo en un ámbito reservado. Un sujeto, por lo tanto, tiene derecho a mantener su privacidad fuera del alcance de otras personas, asegurándose la confidencialidad de sus cosas privadas.

10.2.2 La Privacidad como Caballo de Batalla

Las personas tenemos y mostramos diferentes facetas de nosotros mismos, en función del contexto en el que nos desenvolvemos en cada momento. Así, no nos comportamos de la misma manera cuando estamos con nuestros hijos o con nuestra pareja que con nuestros compañeros de trabajo o nuestro jefe. Esto nos permite gestionar el grado de intimidad y cercanía que queremos tener en nuestras distintas relaciones. Somos capaces de mostrar diferentes «yo» al mundo que nos rodea en función del contexto. Precisamente esta variedad en nuestras relaciones sociales es importante para que nuestro desarrollo personal sea pleno. Sin embargo para mantener la citada variedad, necesitamos tener control sobre qué parte de nosotros mismos enseñamos a cada persona.

En el ámbito laboral, la privacidad protege a las personas contra la discriminación. Controlar la información que compartimos puede impedir que cualquier empresa pueda aprovechar información que, no siendo relevante para el desempeño profesional, podría influir en la decisión sobre la contratación o no de una persona, o en las condiciones laborales de la misma.

Desde una perspectiva más general puede afirmarse que cuanta más información sobre nosotros tengan los demás, más vulnerables somos ante ellos.

Para identificar la trascendencia que las aplicaciones del Big Data pueden tener desde un punto de vista legal y cómo pueden afectar a nuestra privacidad, debemos entender cómo el tratamiento de datos puede afectarnos y por qué es necesaria la intervención del regulador. Para tal fin, nos ayudaremos de algunos ejemplos.

10.2.2.1 La actividad diaria de las personas como fuente de información

Los seres humanos, sistemáticamente desde que iniciamos nuestra actividad diaria, hasta que nos acostamos estamos generando una gran cantidad datos. No sólo el mundo digital es capaz de generar datos, el mundo físico también, si bien es cierto que es mucho más sencillo recopilar información desde el entorno digital que desde el físico. Por ejemplo, las empresas que nos proporcionan los suministros básicos como la electricidad, el agua o el gas, estarían encantados de saber a qué hora nos levantamos y cuando nos volvemos a acostar, de esta forma, estas empresas podrían dimensionar sus redes correctamente, previendo el consumo que se va a producir y evitando el sobredimensionamiento innecesario de sus recursos. Sin duda alguna, el análisis de esta información, repercutirá en sus costes y, por lo tanto, debería repercutir en nuestro beneficio, en el importe de la factura que pagamos.

Hasta ahora, estos datos los obtenían de forma agregada, por ejemplo conociendo el consumo de todo un barrio o, como mucho, el de un edificio. Para conocer detalles más atómicos de información, es necesario utilizar métodos más sofisticados, lentos y costosos. Sin embargo, la llegada de los contadores digitales, y de la domótica en general, pueden proporcionar información individualizada acerca de, por ejemplo, cuando encendemos la luz, cuando nos metemos en la ducha, cuando encendemos la cafetera para hacernos el café del desayuno o cuando encendemos la calefacción y si cuando lo hacemos, preferimos que la casa esté más o menos caliente. Esta información, además está disponible en tiempo real y en el formato adecuado para ser procesada.

Toda esta información serviría para poder identificar nuestros gustos y preferencias, y así por ejemplo, poder ofrecernos otros productos o servicios que se adecuen a nosotros. Un hecho significativo acerca de las posibilidades de este tipo de negocio entorno a los datos capturados de nuestros hábitos de consumo en casa, es la compra por parte de Google en 2014 de una empresa de domótica de Palo Alto llamada Nest Labs[197] por un precio de 3,2 mil millones de dólares.

10.2.3 El uso del teléfono como fuente de información

Otro caso muy representativo, es el de los operadores de telefonía móvil. Lo operadores de telefonía disponen de mucha información acerca de los usuarios. Una parte de esta información ha sido proporcionada por los propios clientes, en el momento de contratar el servicio, como consecuencia del proceso de formalización del contrato de prestación

[197] *Nest Labs*. https://nest.com/. [Online; consultado el 11 de diciembre de 2015]

del mismo. Datos como la información personal (residencia, fecha de nacimiento, datos familiares, etc.) o la cuenta bancaria, son algunos ejemplos de esta información.

Sin embargo, la mayor cantidad de información se genera mediante la propia prestación del servicio. La cobertura disponible es ofrecida por un gran número de celdas de telecomunicación geográficamente distribuidas y perfectamente geolocalizadas, a las que nuestros teléfonos se conectan, a medida que nos vamos desplazando. Esto significa que el operador sabe, con bastante precisión, donde nos encontramos en cada momento o las rutas que realizamos en nuestro quehacer cotidiano. Sabe dónde vivimos, dónde trabajamos, dónde pasamos nuestras vacaciones, si tenemos una segunda vivienda, dónde disfrutamos los fines de semana, o incluso si somos seguidores de algún equipo de fútbol, ya que cada dos domingos realizamos un desplazamiento a la zona donde se encuentra el estadio de fútbol del equipo local.

Las llamadas telefónicas son otra fuente de información, los operadores de telefonía necesitan registrar la información de todas las llamadas, número de teléfono al que se llama, duración de la llamada, hora de inicio, etc., son algunos ejemplos de los datos que registran, con el fin de poder prestarnos el servicio y posteriormente facturarnos. Esta información también proporciona a la operadora una valiosa información acerca de nosotros. Con esta información se puede inferir, por ejemplo, quienes son nuestros familiares y amigos: si recibo y hago llamadas a determinados números de teléfono en horas que no son de oficina, como los fines de semana o a por las tardes/noches, lo más probable es que se trate de mis familiares y amigos.

El tráfico de datos que desde nuestro móvil se genera como consecuencia del uso de aplicaciones instaladas en él o cuando estamos sentados delante de nuestro ordenador navegando por Internet, también son una valiosa fuente de información acerca de nosotros y no son sólo las operadoras de telefonía las que disponen de esta información. Los grande proveedores de servicios a través de Internet, como Google, Apple, Facebook, Amazon, Twitter o incluso los AdServer que sirven la publicidad, disponen de una grandísima cantidad de información acerca de nuestros comportamientos, por el uso que hacemos de estos servicios, de forma consciente o inconsciente.

Además de la información que las empresas son capaces de obtener como consecuencia de la prestación de los servicios, nosotros estamos constantemente proporcionando información. Las redes sociales son una inmensa fuente de datos proporcionada, bien directa o indirectamente, por nosotros o por nuestros familiares, amigos y conocidos.

En definitiva, en esta sociedad hiperconectada, podríamos decir que estamos constantemente proporcionando datos, que pueden ser convertidos en información útil, acerca de nosotros. En muchos casos somos nosotros mismos los que cedemos esos datos de forma consciente (por ejemplo, cuando le comunicamos nuestro número de teléfono a un servicio de atención al cliente, nuestra dirección para recibir una factura, o cuando aceptamos ciertos tipos de cookies al navegar por Internet) pero, en la mayoría de los

casos, los datos los recopilan las empresas sin que seamos conscientes de que los están recogiendo y, sobre todo, sin saber para qué se recogen.

Acercándonos a la realidad

¿De dónde y cómo consiguen los datos las empresas? Un grupo austriaco de desarrolladores creó en 2013 un juego para educar y concienciar a la gente en todo lo relacionado con la privacidad y cómo las empresas hacen negocio con los datos.

El juego se llama DataDealer[198], y puede encontrar encontrar un vídeo acerca del juego[199]. Es un juego gratuito publicado bajo licencia Creative Commons y en él se explica a la perfección como las empresas hacen dinero con los datos. De hecho, el objetivo del juego es conseguir la mayor cantidad de datos personales con el fin de generar la mayor cantidad de puntos en el juego.

- ***¿En qué medida nos estamos convirtiendo los consumidores en el producto y el valor más preciado para las empresas?***

10.2.3.1 Los gobiernos, la información sobre las personas y la legislación

Sin embargo, las empresas no son las únicas que recaban información de los usuarios. Los gobiernos también son capaces de hacerlo. La peculiaridad del caso de los gobiernos es que lo que pueden hacer o dejar de hacer es muy diferente dependiendo del momento y de la situación en la que nos encontremos. Lo que en un determinado momento es legal o socialmente aceptable, puede dejar de serlo en otro. Para ilustrar esta afirmación, podemos recordar cómo determinados acontecimientos, como los atentados del 11 de septiembre de 2001, han afectado tanto a la legislación, como a la percepción y aceptación de la sociedad ante el cambio que esta legislación supone en lo que a la perdida de privacidad se refiere.

Otra peculiaridad interesante en lo que respecta a la legislación y la regulación es la territorialidad, especialmente cuando hablamos de servicios prestados en Internet. Las leyes son locales, gestionadas y dictadas por los gobiernos locales, pero es muy común, que los servicios sean prestados por empresas que no están localizadas en el país en el que se recibe el servicio. La mayoría de los servicios prestados a través de Internet, especialmente los grandes y más populares, como Facebook, Twitter, Google o Apple tienen sus sedes, y por tanto sus entidades jurídicas, en EEUU, mientras que el servicio es prestado en cualquier parte del mundo. Ante esta situación, ¿quién es el regulador competente? ¿Qué jurisdicción es aplicable? ¿Son las leyes aplicables en nuestro país capaces de defender nuestros derechos ante potenciales violaciones de estos?

---

[198] *DataDealer*. https://datadealer.com/. [Online; consultado el 11 de diciembre de 2015]

[199] Data Dealer. *Data Dealer: Privacy? Screw that. Turn the Tables!* https://www.youtube.com/watch?v=x2eCAgQ1DTo. [Online; consultado el 14 de diciembre de 2015]

España y la Unión Europea tienen una normativa en materia de protección de datos de carácter personal y privacidad de las más estrictas, sin embargo ¿pueden defendernos de disputas frente a empresas de EEUU? ¿Son los lobbies asociados a estos servicios capaces de influir en la modificación de las leyes? En este punto debemos recordar que la Unión Europea ha aplazado en dos ocasiones la aprobación de nuevas y estrictas normas en materia de privacidad e incluso ya ha rebajado el nivel de exigencia en algunas de las normas originalmente propuestas.

Demasiadas cuestiones con respuestas poco claras. Es evidente que es necesaria la participación del regulador, para garantizar la protección en materia de privacidad, sin embargo ¿qué regulador? Si no existe un regulador mundial, ¿cómo podemos asegurar que nuestra privacidad está a salvo? Sin duda el derecho y las leyes deben estar al servicio de la protección de los ciudadanos, pero cada vez más es más relevante la autoprotección en materia de privacidad.

Desde un punto de vista de la gestión de nuestros derechos individuales, el Big Data, plantea grandes interrogantes que requerirán una regulación seria y eficaz, sobre todo en torno a la gestión de la privacidad.

Acercándonos a la realidad

El uso de los datos por parte de empresas y gobiernos es un hecho innegable y no es nada nuevo, llevan haciéndolo mucho tiempo, sin embargo sólo recientemente los desarrollos tecnológicos, que han permitido capturar y procesar una mayor cantidad de información o algunos acontecimientos, como el caso WikiLeaks o la filtración de información por parte del empleado de la NSA, Edward Snowden, han hecho saltar algunas alarmas en torno a la importancia que la privacidad puede tener.

A continuación se hace referencia a dos artículos de dos periódicos españoles en los que se tratan estos temas:

- En septiembre de 2015 el periódico digital El Confidencial, publicó el artículo «Así es como las grandes empresas venden tus datos en Internet»[200].
- En febrero de 2015, el periódico El País publicó el artículo «Hacienda analizará redes sociales para las investigaciones contra el fraude»[201] en el que se pone de manifiesto el anuncio de la Hacienda española de su intención de hacer uso de las redes sociales para conocer, no solo el «yo» que les comunicamos ofi-

[200] El Confidencial. *Así Es Como Venden tus Datos Personales en Internet.* http://www.elconfidencial.com/tecnologia/2015-09-14/asi-es-como-venden-tus-datos-personales-en-internet_1011071/. [Online; publicado el 14 de septiembre de 2015]

[201] El País. *Hacienda Analizará Redes Sociales para las Investigaciones contra el Fraude.* http://economia.elpais.com/economia/2015/02/19/actualidad/1424366635_632028.html. [Online; publicado el 19 de febrero de 2015]

cialmente al hacer nuestras declaraciones de impuestos, sino también el «yo» que le contamos al resto del mundo a través de las redes sociales.

En 2014 un grupo de jóvenes españoles inició la publicación de un Comic creado por ellos llamado «The Private Eye»[202], cuyo tema principal es precisamente el de la privacidad[203].

La historia transcurre en el año 2076 y en ese momento, después de que la nube hubiese sufrido el mayor descalabro posible y dejase todos los datos y las vergüenzas del mundo al descubierto, la privacidad se ha convertido en el bien más preciado de las personas. A pesar del mensaje apocalíptico, téngase en cuenta que se trata de una ficción y no pretende reflejar un realidad, sin embargo sí que sirve para reflexionar sobre el tema de la privacidad.

❖ ***¿Cree usted que nuestra privacidad está en peligro y lo que se ha dado en llamar el apocalipsis de los datos es un riesgo real?***

10.2.4 Fuentes Públicamente Disponibles vs. Fuentes Accesibles al Público

¿Qué ocurre con la información que está públicamente disponible en Internet? Pues también es aplicable el derecho a la protección de datos. Recordemos que el dueño de los datos es siempre el titular de los mismos, independientemente de si esos datos están en una fuente públicamente disponible o no y sólo podremos utilizar sus datos si estamos legitimados para ello, en alguno de los supuestos antes citados. Es importante remarcar que además de contar con el consentimiento, sólo podremos hacer uso de los mismos para los fines para los que los hemos solicitado.

La ley orgánica de protección de datos (LOPD) sólo contempla una excepción a esta norma, son las fuentes accesibles al público. Con el término fuentes accesibles, la ley española de protección de datos de carácter personal hace referencia a aquellas fuentes de información que pueden ser utilizadas para obtener datos de carácter personal y tratarlos, sin necesidad de disponer del consentimiento del titular del dato.

La LOPD española por ejemplo, en su artículo 3.J define las fuentes accesibles al público como aquellos ficheros cuya consulta puede ser realizada, por cualquier persona, no impedida por una norma limitativa o sin más exigencia que, en su caso, el abono de una contraprestación. La ley en el mismo artículo establece que sólo tendrán la consideración de fuentes de acceso público, exclusivamente:

- El censo promocional.

[202] Panel Syndicate. *The Private Eye*. http://panelsyndicate.com/comics/tpeye. [Online; consultado el 14 de diciembre de 2015]

[203] El País. *EE UU se Rinde al Cómic Español 'The Private Eye'*. http://tecnologia.elpais.com/tecnologia/2014/01/03/actualidad/1388761965_786451.html. [Online; publicado el 3 de enero de 2014]

- Los repertorios telefónicos en los términos previstos por su normativa específica.
- Las listas de personas pertenecientes a grupos de profesionales que contengan únicamente los datos de nombre, título, profesión, actividad, grado académico, dirección e indicación de su pertenencia al grupo.
- Asimismo, tienen el carácter de fuentes de acceso público los diarios y boletines oficiales.
- Los medios de comunicación.

Por tanto, es importante enfatizar en el hecho de que «un dato esté accesible al público» no significa que «el dato pueda ser utilizado por ser público». Si en el proceso de recolección de información no se siguen los pasos correctos, se corre el riesgo de sanción, cuya cuantía dependerá de qué tipos de datos se hayan recogido, de cómo haya sido el proceso de recolección y de qué se haya hecho con ellos.

Existen casos de empresas sancionadas por estas prácticas. En el año 2009 la Agencia Española de Protección de Datos sancionó a la empresa Tick Tack Ticket con una multa por importe de 30.001 € por enviar correos electrónicos con fines comerciales sin consentimiento a cerca de 40.000 direcciones de e-mail[204].

10.2.5 El Reto de la Individualización de los Datos Estadísticos

Si vamos a trabajar con Big Data, lo más razonable es no trabajar con datos de carácter personal o al menos minimizar al máximo la cantidad de datos personales con los que vamos a trabajar. ¿Significa esto que debemos renunciar al uso de Big Data cuando necesitemos trabajar con este tipo de datos? La respuesta claramente es no, pero sin embargo, necesitamos seguir una serie de directrices que nos facilitarán el trabajo con este tipo de datos.

Con el fin de minimizar el uso de datos de carácter personal o incluso no utilizar este tipo de datos, es necesario disociar los datos. La disociación es el término legal que se utiliza para definir el proceso de anonimización de los datos de carácter personal. Es decir, puesto que un dato de carácter personal es aquel que nos permite asociarlo a una persona, el proceso de disociación será aquel que nos permite desvincular la persona del dato, de manera que ya no sea posible realizar tal asociación.

¿Por qué hay que disociar los datos? La utilización de este procedimiento, con carácter previo al acceso y tratamiento de los datos, permite eximir al mismo del cumplimiento

[204] Agencia Española de Protección de Datos. *La AEPD Sanciona a una Empresa por Utilizar un Sorteo para Recabar 40.000 Direcciones de E-mail y Enviarles Spam*. https://www.agpd.es/portalwebAGPD/revista_prensa/revista_prensa/2009/notas_prensa/common/sept/240909_AEPD_sanciona_empresa_utilizar_sorteo_Spam.pdf. [Online; publicado el 24 de septiembre de 2009]

de las obligaciones que establece la ley. Esto nos permite, por ejemplo, la comunicación de los mismos sin la necesidad de recabar el previo consentimiento del afectado. Simplificando tremendamente el proceso.

Para que un procedimiento de disociación pueda ser considerado suficiente, será necesario que de la aplicación de dicho procedimiento, resulte imposible asociar un determinado dato con un sujeto determinado. En algunos supuestos, debe actuarse con especial cautela, ya que en ocasiones un dato aparentemente disociado puede asociarse a una determinada persona física. Por ejemplo si tuviésemos un fichero con los datos de los habitantes de una pequeña localidad, a pesar de que eliminásemos el nombre de las personas, si entre los otros datos aparece por ejemplo la profesión y encontrásemos un dato con la profesión «cartero», si en la localidad, por su reducido tamaño, sólo hay un cartero, seriamos capaces de asociar el dato a la persona.

Al aplicar la disociación debemos tener en cuenta que:

- El proceso de disociación debe ser irreversible, es decir, debe ser imposible poder volver a asociar un dato disociado con la persona identificada o identificable de la que se disoció.

- La disociación no exime de la obligación de haber obtenido los datos personales de forma lícita. Es decir puesto que la disociación consiste en anonimizar datos de carácter personal, la obtención de los datos personales tuvo que haberse hecho de forma correcta, es decir con la pertinente autorización del titular.

- En muchas ocasiones el análisis de los datos (aunque estén perfectamente disociados) nos proporciona información que nos permite generar perfiles o segmentar nuestros datos. En el momento en el que seamos capaces de asociar un dato a un perfil o a un segmento de forma permanente, este data volverá a convertirse en un dato de carácter personal y por tanto deberá volver a contar con la legitimidad para su uso, es decir o bien necesitaremos el consentimiento del titular del dato, o bien estamos amparados por una ley que nos permite utilizar el dato.

10.2.6 Protección de Datos Confidenciales mediante el Cifrado

Uno de los grandes retos a los que nos enfrentamos cuando trabajamos con datos es la necesidad de protegerlos. Proteger los datos, requiere que seamos capaces de construir sistemas seguros, capaces de soportar intentos de violación de su integridad. El problema radica en que construir sistemas seguros no es fácil, ya que nos enfrentamos a diferentes vulnerabilidades potenciales. Nuestros datos podrían verse comprometidos si un atacante:

- Accede físicamente al hardware, bien a las máquinas en las que están alojados los datos o bien a los discos físicos donde está almacenada la información (copiar los discos por ejemplo).
- Aprovecha vulnerabilidades del hardware o del software, bien del sistema operativo o bien del software de gestión de los datos, para poder acceder a los datos.
- Accede al sistema operativo, lo que le permitiría acceder a los discos de almacenamiento o a la memoria de la máquina, pudiendo tener acceso a la información allí almacenada, lo que sin duda incluirá los datos gestionados por el software instalado en la máquina.
- Es capaz de acceder y leer la red que sustenta las comunicaciones entre nodos. Normalmente los sistemas hacen uso de las redes de comunicación para intercambiar información. Supongamos un *cluster* de Hadoop con múltiples nodos o un sistema más tradicional, en el que tenemos una base de datos a la que accede un servidor de aplicaciones. Un atacante podría «escuchar» el tráfico que pasa por la red, viendo el contenido que pasa por ella.
- Otra vulnerabilidad importante en todo sistema, lo constituyen las personas que gestionan los sistemas. Cualquier sistema necesita de personal especializado que lo gestione y lo mantenga, lo que significa que tienen acceso a los datos que residen en él. Administradores de sistemas o de bases de datos necesitan los privilegios suficientes que les da acceso a los datos.

Ante tantas potenciales vulnerabilidades, ¿cómo podemos asegurar nuestros datos? Una posible solución, a muchos de los problemas descritos anteriormente, consistiría en ser capaz de hacer que nuestros datos fuesen ininteligibles para los potenciales atacantes, de manera que, a pesar de que nuestros sistemas se viesen comprometidos, los datos no lo estuviesen, puesto que a pesar de poder acceder a ellos, estos no tendrían ningún sentido para un atacante. La solución para este escenario se basa en el cifrado o encriptación de los datos.

Una solución de este estilo habría evitado multas millonarias a empresas que sufrieron ataques que consiguieron comprometer información de sus clientes, en muchos casos información confidencial. Algunos casos son muy conocidos por su trascendencia en los medios de comunicación: Sony Playstation Network en 2009, Netflix[205] o AOL.

[205] CNN Money. *5 Data Breaches: From Embarrassing to Deadly.* http://money.cnn.com/galleries/2010/technology/1012/gallery.5_data_breaches/index.html. [Online; publicado el 14 de diciembre de 2010]

10.2.6.1 Cifrado o encriptación

La criptografía es la ciencia de aplicar las matemáticas para cifrar y descifrar datos. Cifrar, o encriptar, es el proceso de transformar un determinado texto inteligible (que normalmente se denomina texto plano) mediante un proceso de aplicación de una clave (conocida como clave de cifrado), en un conjunto de datos ininteligibles (conocido como texto cifrado). Para cualquiera que no posea la clave de cifrado, el texto no tendrá ningún sentido, mientras que el poseedor de la clave podrá descifrar el mensaje cifrado, convirtiéndolo en el mensaje original.

Por lo tanto, el proceso de cifrado está basado en un proceso de aplicación de una función o un algoritmo matemático junto con una clave (que dado el algoritmo y el texto plano, nos devuelve el texto cifrado).

En la actualidad en criptografía podemos distinguir tres métodos de cifrado:

- **Criptografía simétrica**. Hablamos de criptografía simétrica, cuando se utiliza una única clave para cifrar y descifrar un mensaje, es decir cuando la clave de cifrado, es la misma, que la de descifrado. Este es el tipo de cifrado que ha dominado la historia de la criptografía hasta hace un par de décadas.

- **Criptografía asimétrica**. En la criptografía asimétrica se utilizan dos claves distintas, una para el cifrado y otra para el descifrado. La clave de cifrado es pública, esto es, conocida por todo el mundo; la de descifrado (clave privada) solamente es conocida por su propietario. Es decir, cualquiera puede cifrar un mensaje y enviárselo a un destinatario, pero solamente el destinatario, mediante su clave privada, podrá descifrar los mensajes que le llegan. Podemos asimilar la criptografía de clave pública a un buzón de correos. Cualquiera puede introducir una carta en el buzón, pero solamente el poseedor de la llave del buzón podrá abrirlo para acceder a su contenido.

- **Criptografía híbrida**. La criptografía híbrida utiliza las ventajas de cada uno de los sistemas definidos anteriormente. Puesto que el cifrado mediante clave simétrica es más rápido y eficiente, podemos cifrar nuestro mensaje mediante este procedimiento y para enviar la clave privada a nuestro receptor, utilizaremos el sistema asimétrico. Puesto que la clave privada del sistema simétrico, por muy compleja que sea, no será muy grande, transmitirla mediante un sistema asimétrico de clave pública-privada no será mayor problema y de esta forma nos aseguramos que el envío de la clave privada del sistema simétrico, se realiza a través de un mecanismo seguro. El proceso para usar un sistema criptográfico híbrido, para por ejemplo enviar un archivo, es el siguiente:

    - Generar una clave pública y otra privada (en el receptor).

- Cifrar un archivo de mediante criptografía simétrica con una clave privada.
- El receptor nos envía su clave pública.
- Ciframos la clave privada, que hemos usado para encriptar el archivo, con la clave pública del receptor.
- Enviamos el archivo cifrado (simétricamente) y la clave del archivo cifrada (asimétricamente y solo puede ver el receptor).

Los protocolos criptográficos SSL utilizados por los navegadores en conexiones seguras (HTTPS) utilizan esta combinación híbrida.

10.2.6.2 Cifrado o encriptado de datos

El cifrado de datos se puede aplicar de dos formas, ambos modos de encriptación se pueden utilizar de forma independiente o conjunta:

- Encriptación de datos en el momento del almacenamiento («*at rest*» o «en reposo»). El cifrado en reposo implica que los datos se almacenan cifrados en los dispositivos físicos (los discos). El cifrado de datos cuando estos se almacenan se puede realizar de cuatro formas:
  - Discos cifrados.
  - Cifrado de discos completos.
  - Cifrado del sistema de ficheros a través del sistema operativo.
  - Cifrado a nivel de aplicación.
- Encriptación de datos en el momento del envío («*in flight*» o «al vuelo»). El cifrado al vuelo, también conocido como cifrado *over-the-wire*, se aplica a los datos en el momento en el que se envían a través de una red.

  El transporte de los datos entre diferentes nodos o entre diferentes elementos de una red de forma segura requiere del uso de protocolos específicos para tal fin. *Transport Layer Security* (TLS – en español «seguridad de la capa de transporte») y su antecesor *Secure Sockets Layer* (SSL – en español «capa de conexión segura») son protocolos criptográficos que proporcionan comunicaciones seguras por una red. Estos protocolos utilizan los métodos criptográficos suelen utilizar la criptografía asimétrica e híbrida.

Cuando activamos la autenticación de usuarios en Hadoop mediante Kerberos, lo que se conoce con el nombre de Hadoop Security, esto no protege los datos mientras viajan por la red; todo el tráfico de red va en claro.

Para hacernos una idea de la importancia que puede tener proteger los datos que viajan por la red, debemos echar un vistazo al uso que Hadoop hace de la misma. Hadoop realiza interacciones a través de la red mediante:

- **Llamadas Hadoop RPC.** Estas llamadas las realizan los clientes Hadoop usando la API, los trabajos MapReduce, y entre los servicios Hadoop (JobTracker, TaskTracker, NameNodes, DataNodes).

- **Transferencia de datos HDFS (*Hadoop Distributed File System*).** Se realiza durante la lectura o la escritura en HDFS; la realizan los clientes Hadoop usando la API, los trabajos MapReduce, y entre los servicios Hadoop. La transferencia de datos HDFS utiliza sockets TCP/IP directamente.

- **MapReduce Shuffle.** La fase de mezcla de datos en un trabajo MapReduce es un proceso de transferencia de datos, de las tareas Map, a las tareas Reduce. Esta transferencia se realiza habitualmente entre distintos nodos del clúster. La mezcla se realiza mediante el protocolo HTTP.

- **Interfaces Web.** Los *demonios* de Hadoop, proveen interfaces web para los usuarios y los administradores, para permitir monitorizar su trabajo y el estado del clúster. Las interfaces web usan el protocolo HTTP.

- **Operaciones FSImage.** Estas operaciones son transferencias de metadatos entre el *NameNode* y el *Secondary NameNode*. Se realizan mediante HTTP.

El problema del cifrado, es que tarde o temprano, necesitaremos realizar operaciones con los datos y en ese momento procederemos a descifrarlos, lo que deja los datos accesibles para cualquier atacante. Si los datos están cifrados en el almacenamiento, en el momento en el que la aplicación que los gestiona los requiere al sistema de ficheros, estos son desencriptados y cargados en su forma original (descifrados) en memoria. Si los datos se encriptan al vuelo, para ser enviados de forma cifrada a través de la red, cuando procedemos a su almacenamiento, tendremos que desencriptarlos para copiarlos al sistema de ficheros. Aunque hayamos optado por ambos sistemas (cifrar al almacenar y al enviar), en el momento en que necesitemos realizar cualquier operación sobre los datos: ordenarlos, filtrarlos, aplicar un algoritmo, etc., necesitaremos ponerlos en su estado original.

La solución a este problema pasaría por encontrar un método que permitiese la ejecución de operaciones sobre datos cifrados, sin necesidad de descifrarlos. Sin embargo, el cifrado de los datos, introduce un aumento en los requerimientos de procesamiento, reduciendo así el nivel de desempeño de las aplicaciones. En este sentido, hay un concepto en el que

se han realizado grandes avances y que tiene un prometedor futuro: los esquemas de cifrado homomórfico[206].

Las aplicaciones que tiene este sistema son enormes, por ejemplo las corporaciones podrían dejar en la nube sus datos de manera segura y además realizar operaciones a éstos, y obtener un resultado. De hecho, se han desarrollado motores de búsqueda en los cuales el buscador no conoce la consulta que realiza el usuario, ni el resultado de la misma, simplemente devuelve el resultado y luego es el usuario quien descifra el resultado para obtener la respuesta a la consulta realizada con total privacidad.

En definitiva, es una forma completamente distinta de cifrar y tratar datos para la que los computadores estándares actuales de 64 bits, no están capacitados. Sin embargo, hay algunas implementaciones prácticas de sistemas que son capaces de ejecutar operaciones sobre datos encriptados, sin necesidad de desencriptarlos.

10.3 Gobernanza de los Datos

La gobernanza de datos es un conjunto de actividades encaminadas a asegurar la calidad de los datos. Controlar la calidad de los datos incluye entre otras las siguientes tareas:

- Incrementar la consistencia y confianza en la toma de decisiones.
- Reducir el riesgo de sanciones por no cumplir con la regulación establecida, como por ejemplo la Ley Orgánica de Protección de Datos (LOPD).
- Mejorar la seguridad de los datos.
- Maximizar el potencial de generación de ingresos de los datos.
- Designar un sistema para medir la calidad de la información.
- Habilitar mejoras en la planificación por parte del personal supervisor.
- Minimizar o eliminar la necesidad de repetir trabajo.
- Optimizar la efectividad de los empleados.
- Establecer un nivel base del rendimiento para motivar los esfuerzos de mejora.
- Reconocer y mantener las ganancias.

[206] La palabra homomorfismo viene de «homo» y «mórfica», que significa de la misma forma. Es decir, la posibilidad de aplicar distintos procesos un mismo dato y obtener el mismo resultado. Un ejemplo simple es la multiplicación, si $a \times b = c$ también sabemos que $\log(a) + \log(b) = \log(c)$, por tanto, estas operaciones son homomórficas. El cifrado homomórfico consiste en realizar operaciones sobre datos cifrados y posteriormente descifrar el resultado, obteniendo lo mismo que si realizamos esas mismas operaciones (o equivalentes) sobre los datos originales. De esta forma se pueden realizar operaciones sobre datos encriptados, sin necesidad de desencriptarlos.

Figura 134 – Áreas de enfoque de la Gobernanza de Datos

La Gobernanza de Datos concierne a aquellas personas o grupos («*Data Stakeholders*») que estén interesados en cómo los datos se crean, recolectan, procesan, manipulan, almacenan y finalmente, se ponen a disposición para su uso o se retiran o eliminan. La toma de decisiones por parte de los *Data Stakeholders* se toman de forma centralizada por la oficina de gobernanza de datos, donde el personal estará organizado en diferentes roles y responsabilidades (Figura 134).

En cierto sentido, la gobernanza de datos es un sistema de toma de decisiones sobre derechos de acceso y control de los procesos relacionados con la gestión de la información. Este sistema se basa en un modelo común que describe quién, cuándo y bajo qué circunstancias puede realizar ciertas acciones sobre los datos y qué métodos puede usar para llevar a cabo estas acciones.

Cuando se habla de gobernanza de datos, se tienen en mente: los esquemas organizacionales, las reglas de negocio, los derechos de decisión, la contabilidad, la monitorización y el control.

Aunque nos podemos encontrar con sistemas de gobernanza de datos con objetivos dispares (integración de datos, gestión centralizada de datos, etc.) todos ellos suelen tener tres objetivos en común:

- Definir, recolectar y alinear reglas.
- Resolver incidencias.

- Monitorizar y forzar que se cumplan las reglas establecidas.

Las empresas se suelen ver obligadas a pasar de una gobernanza informal a la gobernanza de datos formal cuando se da alguna de estas circunstancias:

- La empresa crece tanto que la gestión tradicional no puede dar solución a las actividades en diferentes departamentos relacionadas con los datos.
- El sistema de datos de la empresa se complica tanto que la gestión tradicional no puede dar solución a las actividades en diferentes departamentos relacionadas con los datos.
- Los arquitectos de datos, equipos de aplicaciones orientadas a servicios u otras actividades horizontales de la empresa, necesitan el soporte de un programa inter-departamentales que haga posible una visión global de ésta en relación a las decisiones relativas a los datos, en vez de visiones aisladas o parciales.
- Cambios regulatorios o contractuales que fuerzan el uso de la gobernanza de datos.

Para llevar a cabo la gobernanza de datos de forma adecuada hay que tener en cuenta los siguientes aspectos: organización, personal, procesos, documentación y monitorización. Un desarrollo de los aspectos citados se recoge en la Figura 135.

Figura 135 – Organización de la Gobernanza de Datos

Otro tema a considerar es dónde se localiza la gobernanza de datos dentro de la organización de la empresa. Si la gobernanza de datos se ubica en un área de la empresa dedicada al desarrollo de aplicaciones se corre el riesgo de que esta gobernanza esté sesgada hacia el cumplimiento de los requisitos de las aplicaciones. Igualmente, si esta tarea recae en los gestores de proyectos es posible que la gobernanza de datos se descuide debido a las temporizaciones y limitaciones estrictas en el desarrollo de los proyectos. En cualquier caso, lo más importante es que, se ubiquen donde se ubiquen, a los responsables de la gobernanza se les proporcione los niveles adecuados de liderazgo y de implicación de los usuarios de los datos en la empresa.

10.3.1 Enfoques de Gobernanza de Datos

Es posible enfocar la gobernanza de datos hacia diferentes aspectos de la gestión de la información: Las políticas de datos, las arquitecturas de datos y las estrategias generales de datos.

En cuanto a las políticas de datos, las principales actividades a realizar por un sistema de gobernanza de datos son las siguientes:

- Identificar los usuarios, establecer los derechos de decisión y clarificar las métricas de datos y su seguimiento.
- Establecer directivas de calidad de los datos.
- Controlar la calidad de los datos.
- Informar del estado de las actividades centradas en mejorar la calidad de los datos.

Los requisitos impuestos por regulaciones externas son los elementos que provocan normalmente el inicio de la gobernanza de datos. Las empresas del sector financiero y de la salud por ejemplo gestionan datos de gran "sensibilidad" que obligan a establecer mecanismos de gobernanza de datos. La gobernanza de datos se inicia normalmente en las capas directivas (aproximación de arriba a abajo) y se implementa usando las tecnologías de la información y los recursos de negocio disponibles.

Por lo general, los programas de gobernanza de datos comienzan con un ámbito empresarial, pero en muchas ocasiones se limitan a tipos específicos de datos de los que gestiona la empresa, como: los datos identificativos (número de la seguridad social), información financiera (números de tarjetas de crédito o cuentas bancarias) y registros médicos. Esta gobernanza de datos incluye técnicas para localizar datos sensibles, protegerlos y gestionar las políticas y control de seguridad, acceso y auditoría a los mismos.

En cuanto a la arquitectura de datos, es necesario usar una arquitectura de seguridad de los datos a nivel empresarial con varios componentes para proteger estos datos sensibles,

tanto en el momento de uso, como cuando son almacenados. Entre las actividades de esta gobernanza de datos, además de identificar a los usuarios, establecer los derechos de decisión y clarificar las métricas de datos, cabe destacar las siguientes actividades:

- Ayudar a proteger los datos sensibles a través del soporte a la gestión de acceso y del establecimiento de requisitos de seguridad.
- Alinear las arquitecturas de seguridad de datos y las diferentes iniciativas de gestión de datos de la empresa.
- Ayudar a medir el riesgo y definir controles para gestionar este riesgo.
- Ayudar a combinar requisitos regulatorios, contractuales y arquitecturales.

Una organización cuyo objetivo sea reducir los costes y aumentar la eficiencia operacional se verá inducida al uso de una gobernanza de datos orientada a simplificar la arquitectura de integración de datos. El coste de los cambios y la falta de agilidad son los principales síntomas de una arquitectura de datos que no usa los estándares de integración.

Como estrategia general, la gobernanza de datos ayuda a la organización a usar una aproximación holística a la gestión de los datos en el contexto de los procesos empresariales, dando soporte a las necesidades de integración de datos. En este tipo de gobernanza de datos nos encontramos con el modelado de datos, diseño de bases de datos y distribución del almacenamiento de los datos. Las tecnologías usadas incluyen SOA (*Service Oriented Architecture*) y DaaS (*Data as a Service*). Las actividades típicas en este modelo son:

- Identificar los usuarios, establecer los derechos de decisión y clarificar las métricas control de la calidad de los datos.
- Asegurar la consistencia de los modelos de datos y las definiciones de datos.
- Dar soporte a las políticas y estándares de la arquitectura de los datos.
- Dar soporte a programas de metadatos, SOA y gestión de datos maestros y empresariales.
- Atraer la atención hacia las problemáticas derivadas de la integración de datos.

La gobernanza de datos centrada en las necesidades analíticas se encuentra normalmente enmarcada en el proceso de maduración de los sistemas de apoyo a la toma de decisiones. Las actividades que nos podemos encontrar son:

- Identificar a los usuarios, establecer los derechos de decisión y clarificar las métricas de control de la calidad de los datos.

- Establecer las reglas para el uso y definición de los datos.
- Establecer puntos de control hacia la gobernanza de datos.
- Clarificar el valor de los repositorios de datos y los proyectos relacionados con el uso de los datos.

10.3.1.1 La gobernanza de los datos en entornos del Big Data

Los principios que deben regir la gobernanza de los datos se pueden sintetizar diciendo que los datos deben gestionarse de forma práctica y rápida, es decir siendo conscientes de la "productividad" de los procesos de gestión de los datos y cuidando la calidad de los datos.

El rol de la alta dirección radica en equilibrar la rapidez con que se tratan los datos, esto es, la productividad con la que se trabaja con los datos y la calidad de los mismos. Es importante detectar cuándo una debe primar sobre la otra (calidad sobre productividad o a la inversa). Esta función es la principal tarea que deben llevar a cabo los responsables máximos de los datos en la empresa.

La velocidad a la que el valor agregado de los datos y los resultados de una interacción con el cliente llega a los jefes de producción, ventas y marketing o a los equipos de finanzas tiene un impacto significativo en el negocio y en su posición competitiva.

Asimismo, la calidad de la información influye en esos mismos consumidores de información ya que confían en los datos más que en su propia intuición. En algunos casos, si la calidad se degrada la confianza en los datos disminuye tanto como para hacer que la información resulte prácticamente inservible.

A veces, la inversión en personal e infraestructuras se vuelve insuficiente, simplemente porque las prácticas que rigen el funcionamiento y el mantenimiento de los sistemas que capturan, integran y distribuyen la información no se mantienen al día con el aumento del volumen, la variedad y la velocidad con que se reciben los datos. Por todo esto se necesita implantar un enfoque de gobernanza de los datos en el contexto del Big Data.

10.3.1.2 Recomendaciones básicas para la gobernanza del Big Data

Los principios básicos que deben regir la gobernanza de los datos (productividad y calidad) sirven a los líderes de la organización, al proporcionar una base para la toma de decisiones y un medio para comunicar su intención a los niveles inferiores de producción. El objetivo de gobernar la información para una organización es mover información de forma rápida y práctica, manteniendo la calidad tan alta y tan segura como sea práctico.

Los orígenes de los principios que rigen la gobernanza de datos se remontan a las buenas prácticas de gestión de datos que se desarrollaron en las décadas anteriores a la aparición

del Big Data. Son por tanto una extensión de las prácticas, que a menudo constituyen la base de muchos de los requisitos normativos, tales como SOX, HIPAA, Dodd-Frank[207] y otros. Estos principios se extienden para incluir aquellos necesarios para el Big Data.

Si una organización aún no ha puesto en marcha un programa de gobernanza de datos basados en estos principios, podría considerar las siguientes recomendaciones como un buen punto de partida:

- **Agilidad**: El enfoque de las prácticas de gobernanza de datos debe permitir respuestas ágiles a los cambios en la tecnología, las necesidades del cliente y los procesos internos.
- **Compatibilidad**: Es necesario poder compatibilizar los datos introducidos en la organización por una persona o un proceso respecto a las políticas y regulaciones.
- **Manejo de la calidad**: Dado que la información y los datos son el núcleo del negocio, la calidad del contenido de esa información es de suma importancia para un éxito continuado.
- **Fomentar la participación**: Las personas de una organización son el medio para obtener la calidad, la seguridad y la gobernabilidad/cumplimiento de los datos.
- **Trazar la información**: La comprensión plena del negocio y el flujo de información a través de todos los procesos permitirán a la organización alcanzar los principios básicos de gestión de los datos con productividad y calidad. Esto requiere la captura y registro de datos en reposo y datos en movimiento en toda la organización.
- **Administrar el significado**: Los datos son el idioma del negocio. A tal efecto, la comprensión del lenguaje que utiliza y su gestión reduce la ambigüedad, redundancia e inconsistencia, lo cual se relaciona directamente con la calidad de la información.
- **Manejo de la clasificación**: Es fundamental para el negocio clasificar la fuente general de datos y el contenido intrínseco tan pronto como sea posible, pues de esta forma se facilita la gestión del ciclo de vida de la información, el control de acceso y el cumplimiento normativo.
- **Proteger la información**: La protección de la calidad de los datos y el acceso es esencial para ser capaces de mantener la confianza de los clientes.

[207] Adam Hartung. *Regulations Work: Benefits Of SOX And Dodd-Frank.* http://www.forbes.com/sites/adamhartung/2015/08/16/regulations-work-benefits-of-sox-and-dodd-frank/. [Online; consultado el 12 de enero de 2015]

- **Fomentar el servicio**: Asegurar el uso adecuado y la reutilización de datos requiere asignar a algún empleado de la organización para que se responsabilice de esta tarea. Este rol no puede ser automatizado y requiere la participación activa de un miembro de la organización empresarial para servir como administrador de los datos.

- **Administrar las necesidades a largo plazo**: Las políticas y normas son el mecanismo por el cual la administración comunica sus necesidades de negocio a largo plazo. Son esenciales para un programa de gobierno de datos eficaz.

- **Manejar la retroalimentación o feedback**: Como complemento de las políticas y normas, la retroalimentación permite comunicar a toda la organización las políticas, las normas y los posibles conflictos con los nuevos requerimientos en materia de gobernanza de datos. Esta actividad forma parte del proceso central para impulsar mejoras en la política de datos.

- **Fomentar la innovación**: La gobernanza no debe perjudicar la innovación. El gobierno de datos puede y debe adaptarse a las nuevas ideas y al crecimiento de la organización. Esto se logra a través de la gestión de los entornos de infraestructura como parte de la arquitectura.

- **Controlar el contenido de terceros**: Los datos de terceros juegan un papel cada vez mayor en Big Data, por lo que los controles deben ser los adecuados a las circunstancias. Se deben tener en cuenta la normativa aplicable para las regiones geográficas en que se opera.

10.3.2 Integridad de los Datos

Las empresas que usan sus datos como elemento clave para mejorar su competitividad en el mercado global reconocen que la calidad de éstos es esencial. Por tanto, sus sistemas de control de calidad van más allá de establecer una comprobación rutinaria. En estos casos la gobernanza de datos les proporciona un sistema de control de calidad con procedimientos y políticas para asegurar que sus datos están disponibles, funcionales, seguros y con una integridad demostrable.

La integridad de los datos es fundamentalmente un problema de diseño, pero los sistemas de información no son inmunes a los riesgos de integridad de los datos, en contraposición a los documentos en papel para los que asegurar la integridad es más sencillo. La integridad de los datos se refiere a que los datos están completos, son consistentes y precisos a lo largo de todo el ciclo de vida de los mismos. Para asegurar la integridad de los datos se deben establecer políticas de documentación que controlen cada modificación de los datos, incluyendo normas como las siguientes:

- Cada modificación realizada en un documento debe ser firmada y fechada.

- Esta modificación debe también permitir el acceso a la información original, antes del cambio realizado.
- En ocasiones puede ser necesario documentar también el motivo de la modificación.

Este proceso para asegurar la integridad de los datos podemos dividirlo en cuatro fases:

1. **Perfilado de datos o** data profiling: Las herramientas de perfilado de datos analizan los repositorios de datos con respecto a métricas de medición de la calidad que definen lo que son buenos o malos datos. La creación de perfiles de datos no es una tarea que se realice sólo al principio del proceso, sino que es necesario mantener durante el ciclo de vida de los datos para analizar las tendencias en la calidad de los datos.
2. **Estandarización de datos**: Para validar y corregir los datos de acuerdo a estándares de la industria o propios de la empresa. Esto incluye elementos como formatos de nombres, mayúsculas, direcciones, fechas, etc.
3. **Enriquecimiento de datos**: Este proceso permite añadir información adicional a los datos existentes, como podría ser la información geo-referenciada.
4. **Monitorización de los datos**: Es necesario tener un procedimiento para hacer seguimiento de la calidad de los datos a lo largo del ciclo de vida de los mismos. Este proceso de control permite descubrir áreas donde es necesario incluir nuevos mecanismos de corrección.

Las soluciones de integridad de los datos son aplicaciones complejas que ayudan a las empresas a mejorar significativamente la completitud y corrección de sus datos. Estas soluciones software permiten eliminar problemas como la redundancia o duplicidad en los datos, la falta de consistencia o la estandarización entre sistemas compartiendo o usando la misma información, pérdida de datos o datos incompletos. En esta categoría de soluciones podríamos englobar los sistemas de gestión de la calidad, gobernanza de datos, gestión de datos maestros y perfilado de datos.

La integridad de los datos es un problema de importancia en la actualidad ya que se generan grandes cantidades de datos por lo que asegurar esta integridad puede resultar complicado en muchos casos. Por tanto, la falta de integridad o la ausencia de seguridad sobre la misma hacen que sea difícil obtener valor de grandes cantidades de datos.

10.3.2.1 Integridad de los datos y el Big Data

La integridad de los datos también es importante en el contexto del Big Data, ya que si vamos a reutilizar datos o usar datos masivos de nuestra empresa, sólo podremos sacar

valor real al análisis de los mismos si podemos asegurar cierto nivel de integridad de los datos analizados. Por tanto, podríamos decir que hay cuatro aspectos clave para sacar valor de los datos analizados en el Big Data:

- Asegurar la precisión de los datos. ¿Son los datos precisos? Aunque pueda resultar un proceso complejo y artesanal, podemos indicar algunos pasos comunes que podemos realizar, como: utilizar fuentes de datos de origen de confianza, utilizar datos de socios e intentar evitar la utilización de fuentes de datos desconocidas.

- Asegurar la seguridad de los datos. ¿Son los datos seguros? Será necesario aplicar técnicas que protejan estos datos en la medida de lo posible, usando por ejemplo: normas de certificación ISO/IEC 27001[208] para una mayor protección de los datos, encriptación de datos, además de tomar políticas de gestión y actualización de contraseñas.

- Asegurar la disponibilidad de los datos. ¿Están los datos disponibles todo el tiempo? Para ello podemos implantar diversas capas de replicación de los datos que nos aseguren el acceso a un conjunto mínimo de datos para estos procesos de análisis.

- Asegurar la escalabilidad. ¿Podemos analizarlos de forma escalable? Es necesario tener en cuenta qué modelo vamos a usar en cuanto a capacidad de computación. Si nos decidimos por una infraestructura propia debemos asegurarnos la capacidad (física y económica) de ampliación en caso de ser necesario si las fuentes de datos analizadas crecen mucho. Si decidimos usar recursos en la nube, debemos asegurarnos que tenemos capacidad económica para asumir la necesidad creciente en recursos en la nube conforme crezcan las fuentes de datos analizadas. Por ello, en ambos casos debería planificar la evolución de los sistemas usados por nuestras aplicaciones de análisis del Big Data.

[208] ISO Standard. *ISO/IEC 27001 - Information security management.* http://www.iso.org/iso/home/standards/management-standards/iso27001.htm. [Online; consultado el 12 de enero de 2015]

11 Aplicaciones Reales a Negocio

11.1 Introducción

En este último capítulo, describimos un caso de uso real en el que aplicamos los conceptos y técnicas desarrolladas a lo largo de este libro. En concreto, nos centramos en el análisis del *«Clickstream»* o «Huella Digital» en e-commerce. El Clickstream es el proceso de recogida, almacenamiento y análisis de las direcciones o enlaces Web sobre las que un usuario o visitante hace clic y accede a lo largo de su visita a un sitio Web. La secuencia de pasos o clics de un visitante en una Web se conoce también como la «Huella Digital», cuyo análisis es de gran relevancia en el entorno empresarial del comercio electrónico, ya que ayuda a descubrir información sobre las preferencias de los clientes, dando pie a estrategias de mercadotecnia personalizada en Internet.

El Clickstream es ampliamente estudiado en la actualidad ya que procesa la gran mayoría de los datos que se generan respecto a los movimientos de los usuarios en la Web y el comercio electrónico en particular. De hecho, los datos que se generan en los servicios Web, se almacenan en ficheros de traza o «`.log`» referentes a la actividad de los usuarios. Estos ficheros alcanzan volúmenes lo suficientemente grandes para requerir la aplicación de técnicas del Big Data, tanto para su gestión y modelado, como para su análisis.

En este sentido, el análisis de las visitas a portales de comercio electrónico se ha convertido en elemento fundamental en este sector, ya que provee a los expertos de una fuente de información exhaustiva acerca de las motivaciones de millones de posibles usuarios/clientes en la Web.

El Clickstream de los usuarios en cada momento de acceso y para cada Web o tienda virtual genera una cantidad de datos enorme en los servidores. A día de hoy, estos datos sólo son manejables en su conjunto mediante técnicas de análisis del Big Data. Este caso de uso de análisis de datos de comercio electrónico representa un escenario de uso del Big Data, debido al gran volumen de datos y su tipología (normalmente datos no estructurados), lo que supone un desafío para su gestión por parte de las técnicas de inteligencia de negocio (o «Business Intelligence»).

Acercándonos a la realidad

El análisis de Clickstream está encaminado a responder preguntas en el área de negocio como:

- ***¿Cuál es la mejor ruta de clics para que un visitante encuentre información sobre un producto y termine comprándolo?***
- ***¿Qué tipos de productos añade un cliente a su cesta de compras y cuáles añadirá en un futuro?***
- ***¿Dónde debería invertir para mejorar la tasa de conversión de mi tienda online? Entendiendo como tasa de conversión el ratio entre los visitantes absolutos a una Web y los visitantes que finalmente adquieren un producto.***

11.1.1 Introducción a los Datos

Este caso de uso trata sobre el análisis de datos de acciones de usuarios en un portal Web. Para ello, nos vamos a basar en una plataforma integrada para el tratamiento de Big Data como la que nos ofrece la Sandbox de Hortonworks[209], la cual se puede instalar como máquina virtual en un computador personal. Vamos a tomar un conjunto de datos de los ejemplos disponibles en la distribución HDP 2.X de Hortonworks Sandbox, que nos permite comenzar a trabajar en este caso de uso aunque no seamos propietarios de un servidor Web del que capturar esta información. Estos datos varían ligeramente dependiendo de la versión de Sandbox que usemos. En concreto, se tratan de datos del sitio Web Omniture, entre los que se incluyen los conjuntos de datos siguientes:

- Omniture logs: ficheros de traza (`.log`) del sitio Web con Información como URL, *timestamp* (marca temporal), dirección IP, dirección IP geocodificada e ID de usuario (SWID).
- Users: datos de los usuarios, incluyendo fechas de nacimiento y género.
- Products: datos sobre los productos que aparecen en el portal Web.

[209] *Hortonworks Sandbox - The easiest way to get started with Enterprise Hadoop*. http://hortonworks.com/products/hortonworks-sandbox/. [Online; consultado el 26 de enero de 2016]

Figura 136 – Datos de usuarios de Omniture. El identificador de usuario se expresa mediante código alfanumérico (SWID)

SWID	BIRTH_DT	GENDER_CD
0001BDD9-EABF-4D0D-81BD-D9EABFCD0D7D	8-Apr-84	F
00071AA7-86D2-4EB9-871A-A786D27EB9BA	7-Feb-88	F
00071B7D-31AF-4D85-871B-7D31AFFD852E	22-Oct-64	F
0007967E-F188-4598-9C7C-E64390482CFB	1-Jun-66	M
000B90B2-92DC-4A7A-8B90-B292DC9A7A71	13-Jun-84	M
000C1856-994E-476B-8C18-56994E676B29	29-Dec-80	U
000F36E5-9891-4098-9B69-CEE78483B653	24-Mar-85	F
00102F3F-061C-4212-9F91-1254F9D6E39F	1-Nov-91	F
0010C6F2-8C04-450E-90C6-F28C04B50E97	20-Jun-02	U
0011C945-28C4-4D6F-B1E6-6CA7EFC14548	13-Nov-87	F
001704E0-6CD8-429A-8E0A-89024019CA6A	10-Jul-91	M
001720A4-44E3-43F0-BA8F-2F2E0D7B6275	8-Nov-90	M
001834AA-7451-49EA-B0E4-ED4A71C97AD1	21-Feb-88	M
0018B3DA-4763-460A-B67E-1D5168E4DEB6	21-Apr-02	U
00193DB8-1FE8-440B-8217-BAA2AA5FDD64	31-Jul-70	M
001AFDA9-18D4-4FE8-B4F1-ADD932E0ACB2	16-Jan-81	U
001BFE35-555B-48E1-9ED3-A4BE7677C36C	16-Feb-82	M
001F01F3-11D8-470D-BF34-EA95941E525F	11-Dec-00	F
001F59ED-88C4-4BDC-B90F-2C17AE4A7422	2-Feb-97	U
001F604A-6CA8-45B0-BC56-50800092DF74	29-Jun-88	M
0020446F-975A-4854-B007-B9A7506AA976	20-Jun-91	F
00222A1D-4D9E-489F-ADA3-F28F5B6408AE	11-May-76	F
00229E92-894D-48F5-88BF-812EB4215E29	1-Jul-60	F
0023787E-A7BC-4DEB-A2E6-397BC1C20355	8-Nov-59	F
00240276-1021-4D61-8628-36AE345D8E83	28-Apr-90	M
0027226E-688F-432A-B91A-E10D1DB27CF6	17-Nov-81	M
00277C39-7033-46D2-A77C-397033B6D236	13-Jul-89	U
002974DC-5D3A-4F64-879E-6A9D28A3F421	15-Apr-55	F
002A53AD-3E72-4E72-B8AF-8EBCC7EA84DE	10-Dec-79	M
002A7C6A-E6C0-4760-9493-4B9BB2D0079E	12-Sep-80	F
002D36E9-ACC6-40CD-B0AF-E205FE7F0AE3	5-Mar-58	U
002FF013-AB8F-4FE8-AE3E-EF19A8E0603D	10-Apr-48	U
002FF3AD-A960-4B19-B037-3E4284092F9B	24-Oct-85	F
0033E5A9-878A-48D7-B65A-F1495F4D60A6	10-Nov-89	F
00347AAC-88E2-4016-8D33-9D4AD985A2D2	17-Sep-88	F
0035CFF3-C980-4CE6-AA7B-DB871422545F	1-Nov-60	U
0035FC8F-911D-40C3-832C-61B701A5B6A7	6-Aug-86	M
00367F81-A361-4441-AC03-4D2FF64A4E71	6-May-85	M
00383010-3615-11D4-820C-00A0C9E58E2D	10-Jan-76	M
00393215-4155-4B1C-A7A2-3E84A8ECDBCC	3-Mar-80	F
003E3300-8395-4031-8CF6-34DC480C4A51	21-Dec-82	M
00408C37-8A9A-4D15-808C-378A9A5D15BC	1-Sep-02	U
004171A7-7FDD-487E-A205-300B47240595	25-Apr-85	F
0041F4E3-ED8D-4DDC-A23B-2D8DD18192E9	4-Sep-97	U
0044AF02-16EC-42B0-96DD-52679773A9D6	17-Jul-06	U
0044FFF2-C154-41A5-B6F2-897D56592C98	26-Sep-89	M
004524A8-F0AB-4C1F-B8D6-A6A525B6C11E	5-Jan-74	U

En primer lugar se revisan los datos de *logs* que suelen almacenarse en ficheros de texto con columnas separadas por tabuladores o TSV (*tab separated values*). El conjunto de datos (*dataset*) de Omniture log contiene alrededor de 4 millones de entradas de datos, representando en su conjunto unos 5 días de tráfico clickstream. No obstante, la práctica común en las grandes organizaciones es la de procesar semanas completas, meses e incluso años de datos.

En el caso de los datos de usuario, la Figura 136 muestra una selección de éstos donde se puede observar la estructura TSV con el identificador de usuario SWID, su fecha de nacimiento (BIRTH_DT) y género (GENDER_CD). Estos datos suelen almacenarse me-

Figura 137 – Datos de productos de Omniture

```
url category
http://www.acme.com/    books
http://www.acme.com/SH55126545/VD55149415   movies
http://www.acme.com/SH55126545/VD55163347   games
http://www.acme.com/SH55126545/VD55165149   electronics
http://www.acme.com/SH55126545/VD55166807   computers
http://www.acme.com/SH55126545/VD55170364   home&garden
http://www.acme.com/SH55126545/VD55173061   handbags
http://www.acme.com/SH55126545/VD55177927   clothing
http://www.acme.com/SH55126545/VD55179433   shoes
http://www.acme.com/SH55126554/VD55147564   outdoors
http://www.acme.com/SH5568487/VD55169229    automotive
http://www.acme.com/SH5580165/VD55156528    clothing
http://www.acme.com/SH5580165/VD55173281    tools
http://www.acme.com/SH5582037/VD5582082 accessories
http://www.acme.com/SH5584743/VD55162989    grocery
http://www.acme.com/SH5584743/VD55178549    clothing
http://www.acme.com/SH5585921/VD55178554    clothing
http://www.acme.com/SH5585921/VD55179070    clothing
http://www.acme.com/SH5587637/VD55129406    clothing
http://www.acme.com/SH5587637/VD55134536    shoes
http://www.acme.com/SH5587637/VD55137665    shoes
http://www.acme.com/SH5587637/VD55167939    shoes
http://www.acme.com/SH5587637/VD55178312    shoes
http://www.acme.com/SH5587637/VD55178699    shoes
http://www.acme.com/SH559026/VD5568891  handbags
http://www.acme.com/SH559026/VD5582785  handbags
http://www.acme.com/SH559040/VD55175948 handbags
http://www.acme.com/SH559044/VD5586386  handbags
http://www.acme.com/SH559056/VD55178907 handbags
http://www.acme.com/SH559056/VD55179132 handbags
http://www.acme.com/SH559056/VD55181666 handbags
```

diante identificadores alfanuméricos y generados de manera aleatoria como en el caso del SWID, con el fin de proteger la identidad del usuario real. De esta forma se evita el uso directo de información sensible del usuario, garantizando el cumplimiento de las leyes internacionales de protección de datos y en aras de una buena gobernanza de datos, como ya se especificó en el capítulo 10.

En cuanto a la información sobre productos (Figura 137), se trata de un subconjunto muy sencillo de datos que incluyen la URL en la que se encuentra definido el propio producto y la categoría a la que pertenece. El atributo "category" es muy utilizado para los análisis posteriores en los que se pretende clasificar las preferencias de los clientes y las tendencias de mercado. Además dota de estructura o jerarquía a la propia tienda online y ayuda a la organización de los productos con la intención de facilitar la búsqueda por parte del cliente.

11.2 ClickStream: de la Recolección al Procesamiento

Como ya avanzamos en la introducción, el Clickstream o flujo de clics no es sólo otra fuente de datos que se extrae, se limpia y se vierte en el entorno de almacén de datos y Business Intelligence, sino que supone una colección de fuentes de datos en constante evolución. Existe una gran cantidad de formatos de ficheros `.log` y APIs para capturar los

datos de clics de los usuarios de portales Web. Estos formatos de ficheros `.log` tienen componentes opcionales de datos que, si se usan, pueden ser muy útiles en la identificación de los visitantes, las sesiones, la afinidad, las preferencias y tendencias en la conducta del usuario o visitante de la tienda online.

11.2.1 Recolección

El primer paso en nuestro caso de uso consiste en la recolección de datos. Para ello descargamos el fichero `RefineDemoData.zip`[210] comprimido en formato `.zip` en nuestra máquina local y descomprimimos el contenido. Los ficheros extraídos tendrán una extensión de fichero `.tsv.gz`, aunque en algunas ocasiones, si se realiza la descarga desde cierto tipo de sistema, la extensión puede cambiar a `.tsv.gz.tsv`. Si esto ocurriera, se deben volver a cambiar las extensiones a `.tsv.gz` previamente a la carga de ficheros en la Sandbox.

Conviene señalar que este tipo de obtención de datos sencillo es aplicado al caso de uso concreto de Clickstream. La practica usual es obtener ficheros consecutivos en fecha y hora mediante descargas frecuentes desde repositorios de datos remotos.

11.2.1.1 Carga de ficheros en la Sandbox

Dentro del panel principal de la Sandbox, utilizaremos la herramienta *«HCatalog»* para crear nuevas tablas desde los ficheros de datos. Para esto, hacemos clik sobre el icono de *«HCat»* en el menú principal y una vez dentro, pulsamos sobre «*Create a new table from file*» para crear una nueva tabla, tal y como se muestra en la Figura 138.

Figura 138 – Herramienta «HCatalog» para crear tablas desde ficheros de datos

[210] *Hortonworks Sandbox - Omniture Logs Data Set.* https://s3.amazonaws.com/hw-sandbox/tutorial8/RefineDemoData.zip. [Online; consultado el 26 de enero de 2016]

Figura 139 – Pop-up de selección de ficheros en HCatalog

Dentro del panel de creación de tabla desde fichero escribimos el nombre de la tabla, en nuestro caso «omniturelogs», en la caja de texto «*Table Name*». A continuación, hacemos clic sobre el botón de selección de fichero «*Choose a file*» para que aparezca la aplicación de subida de ficheros.

Aparecerá entonces una ventana a modo de «*pop-up*» como se muestra en la Figura 139, a partir de la cual seleccionaremos el fichero desde el buscador tras pulsar sobre el botón «*Upload-file*».

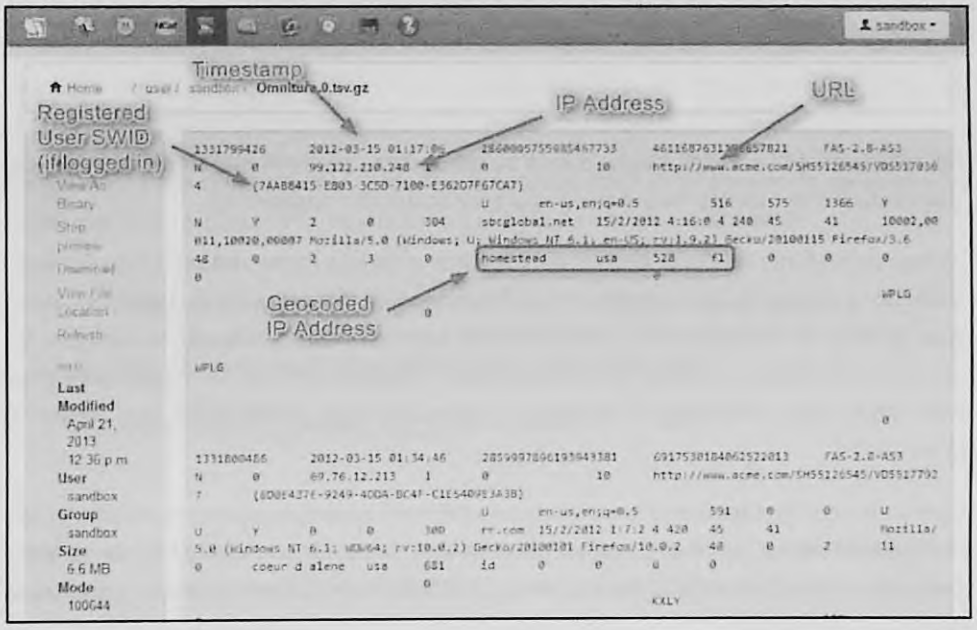
Seguidamente, volverá a aparecer la ventana pop-up de selección de fichero en la cual ya aparece «`Omniture.0.tsv.gz`» para ser seleccionado. La siguiente ventana en aparecer es nuevamente la de creación de nueva tabla desde fichero, ya con el fichero seleccionado `Omniture.0.tsv.gz` en precarga y con el cuadro de diálogo preparado para realizar la creación y carga de datos en la tabla. Una vez en este paso, realizaremos las siguientes acciones:

1. Desactivar la opción de lectura de cabeceras de columnas («Read column headers» en la Figura 140) para crear la nueva tabla utilizando nombres de columna por defecto (col_1, col_2, etc.).
2. Nos dirigimos al fondo de la página y pulsamos sobre el botón «*Create Table*». Aparecerá un indicador de progreso de carga mientras se crea la tabla.
3. Una vez terminado el proceso, la tabla «omniturelogs» ya está creada y aparecerá en la lista de tablas de HCatalog como indicamos anteriormente.

Tras esto, repetiremos este mismo procedimiento de recolección de datos para las tablas de usuarios y de productos. Para ello utilizaremos los ficheros «`users.tsv.gz`» y

Figura 140 – Selección de fichero desde el panel de creación de tabla

Figura 141 – Datos (raw data) de Omniture Log cargados en entorno Hadoop mediante File Browser

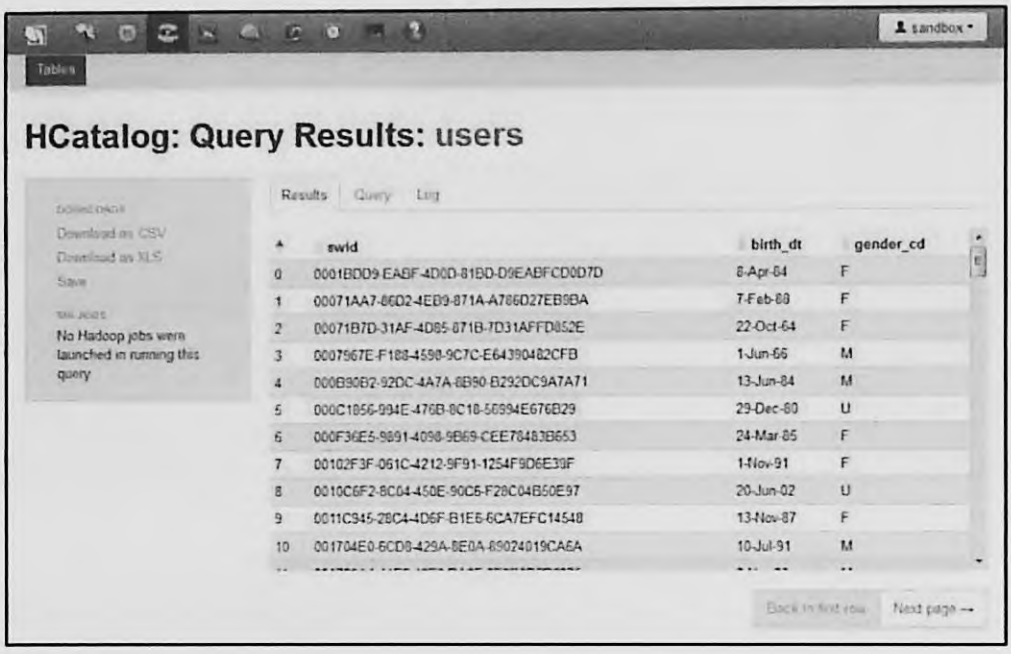


«`products.tsv.gz`» indicando los nombres de tabla «users» y «products», respectivamente. Al crear estas tablas, debemos habilitar la opción de mantener las cabeceras de columnas activadas («*Read column headers*»), para de este modo genere las tablas utilizando los datos de la primera fila como nombres de columna.

11.2.2 Almacenamiento

Una vez cargados los datos, pasamos a la fase de organizar, entender y modelar toda la información para ponerla a disposición de las aplicaciones de Business Intelligence. Para

Figura 142 – Vista de datos de la tabla «users» mediante HCatalog



ello, realizaremos una exploración previa de estos datos para ver qué significado tienen y para estudiar el modelo más indicado para su posterior tratamiento.

En una inspección a fondo, podemos ver los datos en crudo (*«raw data»*) de los ficheros mediante la aplicación de navegación de ficheros «*File Browser*» de la Sandbox, situada en la barra de herramientas. Si hacemos clic sobre el fichero «Omniture.0.tsv.gz», se mostrarán los datos en crudo, donde se pueden inspeccionar la información sobre las visitas Web: URL, timestamp, dirección IP, IP geocodificada, ID de sesión, etc. (véase la Figura 141).

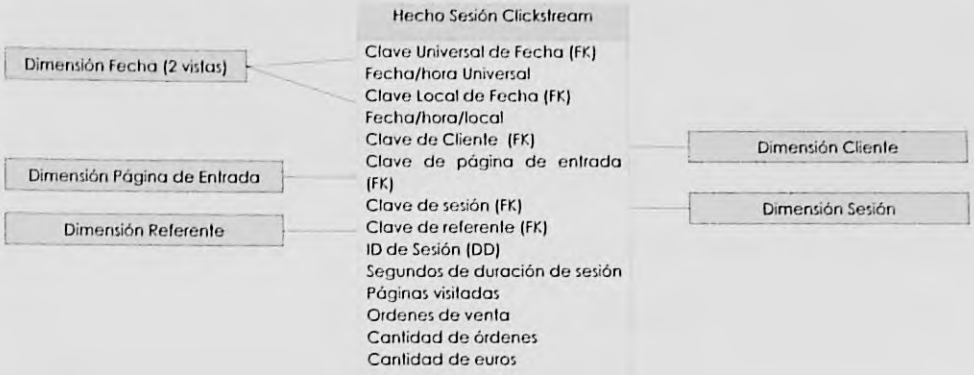
Del mismo modo, para echar un vistazo a la tabla de usuarios podemos utilizar la herramienta HCatalog, mediante la que podemos abrir la ventana con los datos de «users» referentes al identificador software de usuario (SWID), la fecha de nacimiento y el género (Masculino/Femenino), tal y como se muestra en la Figura 142.

Este proceso se repetirá al realizar la navegación sobre la tabla «products» donde aparecerán listados en HCatalog los datos sobre categorías y URLs de descripción de productos.

11.2.3 Modelo Multidimensional

Una vez cargados los datos, podemos definir un modelo multidimensional que nos sirva de ayuda para el análisis que nos ocupa, así como para otras tareas de análisis en el futuro. Sin embargo, vamos a utilizar un esquema simplificado con el fin de mostrar esta fase aplicada al caso de uso seleccionado.

Figura 143 – Tabla de hechos para sesiones únicas



En este modelo, por tanto, las dimensiones apropiadas para esta primera tabla de hechos son: la fecha, la hora del día, el cliente, la página, sesión y el referente. Tras esto, se puede agregar un conjunto de atributos de medidas (en la tabla de hechos) para esta sesión como: segundos de sesión, páginas visitadas, pedidos realizados, cantidad de unidades vendidas y euros facturados. El diseño completo se muestra en la Figura 143, el cual atiende al típico esquema de estrella comentado en el Capítulo 5.

Vamos a analizar qué atributos podemos completar para cada dimensión en base a los datos que hemos cargado en HCatalog simplificando un poco el modelo:

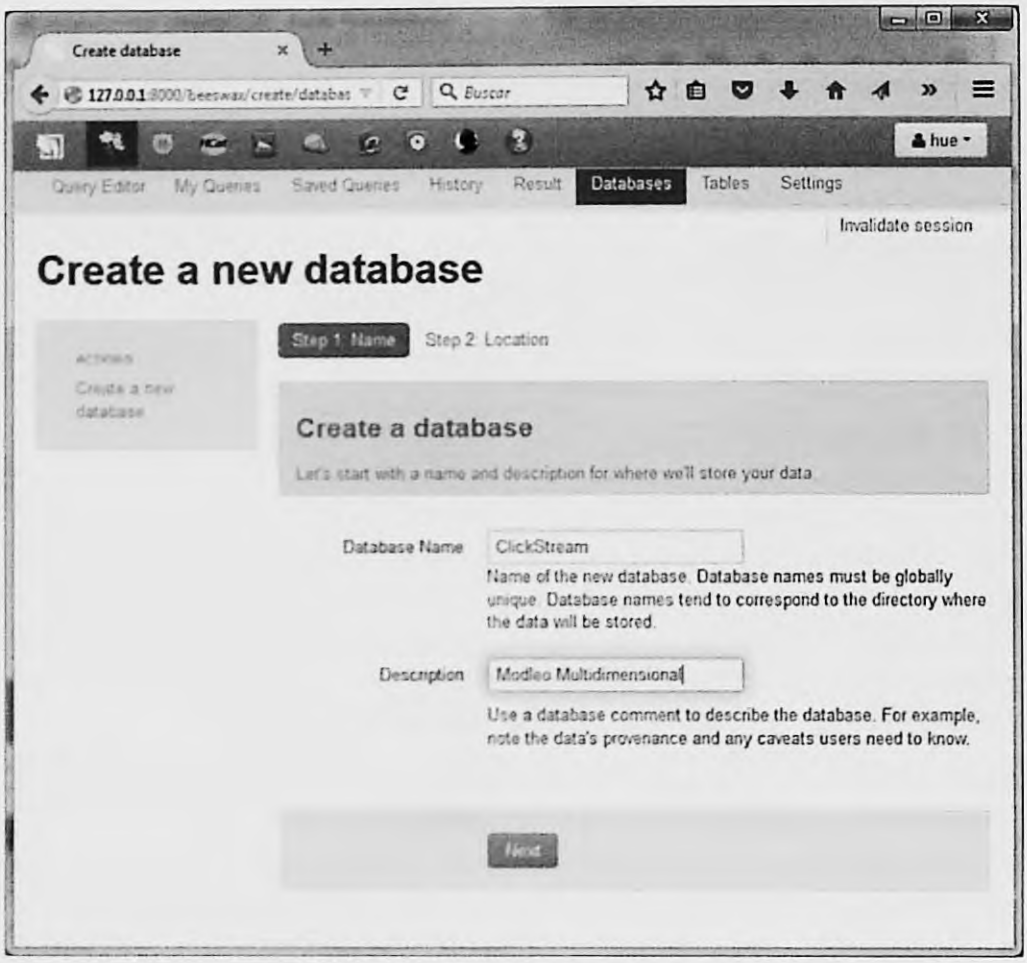
1. Dimensión Cliente: SWID, Año de Nacimiento, Ciudad, Estado, País y Género.
2. Dimensión Producto (Página de Entrada): URL y Categoría.
3. Dimensión Tiempo (Fecha): Fecha, Mes y Año.

Para crear nuestro modelo multidimensional usando Hive. En primer lugar creamos la base de datos ClickStream como se muestra en la Figura 144. Del mismo modo, las tablas las podemos crear usando el interfaz Web con la aplicación HCatalog, como hemos mostrado en la sección anterior (es posible generar bases de datos y tablas introduciendo directamente los comandos mediante una consola o terminal). Será necesario seleccionar en primer lugar la base de datos que vamos a usar para estas tablas.

A continuación se llevarán a cabo los siguientes pasos para la creación de la tabla:

1. Name. Nombre y descripción de la tabla.
2. Record Format. Formato que se usará para almacenar los datos contenidos en la tabla, pudiendo elegirse dos opciones: Delimited (Data files use delimiters, like commas (CSV) or tabs.) o SerDe (Enter a specialized serialization implementation).

Figura 144 – Creación de la base de datos multidimensional



3. Serialization. En este paso seleccionaremos el formato usado para organizar los diferentes campos de la tabla en su almacenamiento en disco.

4. File Format. En este paso seleccionaremos el formato del fichero o ficheros generados, pudiendo elegirse entre tres opciones: TextFile, SequenceFile, InputFormat.

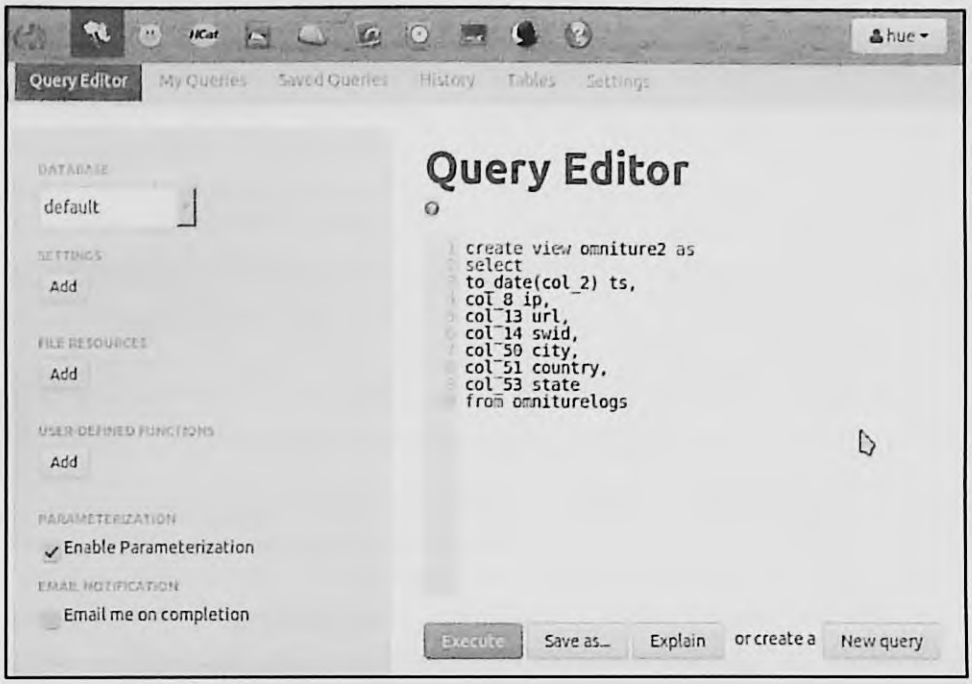
5. Location. Hive decide un espacio por defecto en el HDFS para almacenar los datos de las tablas, pero podemos cambiarlo para indicar una localización diferente.

6. Columns. En este paso podremos añadir los diferentes campos de la tabla indicando para cada uno de ellos el nombre de la columna/campo y su tipo de datos.

11.2.4 Procesamiento

Pasamos describir el procesamiento de los datos Clickstream necesario para el análisis que se va a realizar en el siguiente apartado. Avanzamos aquí que nos centraremos en utilizar librerías de algoritmos de minería de datos para hacer predicciones de clickstream

Figura 145 – Ventana de Hive para la edición de consultas. Creación de tabla «omniture2»



mediante modelos de «regresión lineal» (visto en el Capítulo 8). En concreto, realizaremos predicciones sobre el número de visitantes a un sitio Web determinado en el siguiente periodo de tiempo y para cada país o región.

La primera etapa de procesamiento consiste en generar una vista con todos los datos de Omniturelogs que necesitaremos para nuestro análisis. Seleccionamos en esta consulta los datos referentes a: la dirección IP, la URL, el SWID, la Ciudad, el País o región y el Estado. La consulta Hive que nos proporciona esta vista se muestra en la Figura 145.

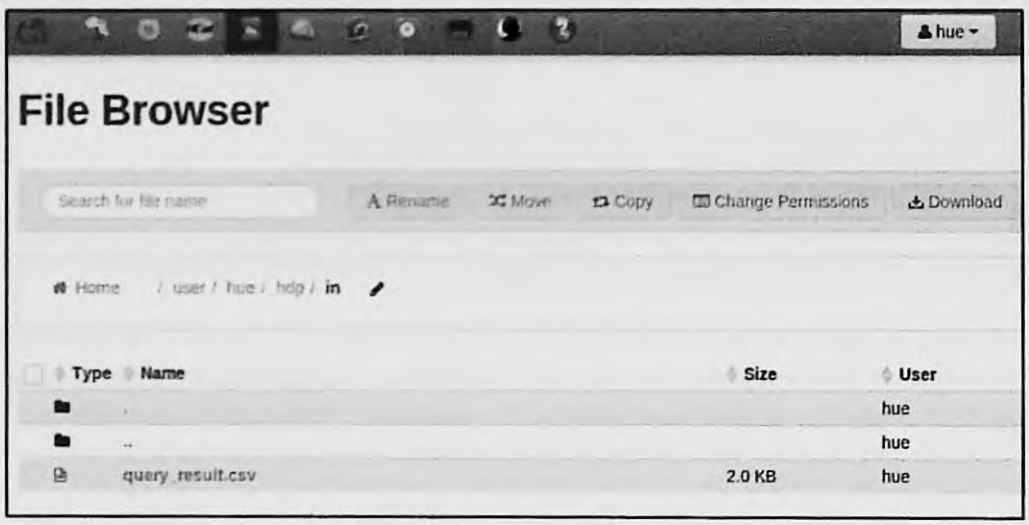
A partir de la vista generada realizaremos ahora una selección de los datos que nos interesan para entrenar el modelo de regresión, básicamente: el país, el timestamp y el número de entradas/visitas. La base de datos de Onminure contiene información desde el 1 de Marzo de 2012 al 15 de Marzo de 2012. Para aquellos países para los que no dispongamos de datos de visitas pondremos el valor «0» y borraremos aquellos conjuntos de datos con pocos valores, ya que introducirían ruido al modelo de regresión. Una muestra del resultado se muestra en la Figura 146.

Finalmente, necesitamos guardar los datos obtenidos para el siguiente apartado, por lo que haremos clic en «Download as CSV» para almacenarlos en HDFS. Como muestra la Figura 147, lo podemos guardar en el directorio «/user/hue/hdp/in» de HDFS con el nombre de fichero «query_result.csv».

Figura 146 – Ventana de Hive que muestra el resultado de la consulta referente a la vista «omniture2»

	country	ts	_c2
0	aus	2012-03-01	2
1	aus	2012-03-03	27
2	aus	2012-03-04	35
3	aus	2012-03-05	15
4	aus	2012-03-06	46
5	aus	2012-03-12	33
6	aus	2012-03-15	16
7	can	2012-03-01	14
8	can	2012-03-03	31
9	can	2012-03-05	55

Figura 147 – Carga de fichero de resultados en HDFS



11.3 ClickStream: del Análisis a la Visualización

En este apartado, continuamos con el caso de uso Clickstream a partir de los datos ya cargados y procesados anteriormente. Como ya se explicó al inicio de este capítulo, las analíticas de visitas en comercio electrónico se han convertido en parte fundamental en este sector, ya que proveen a los expertos en el área de una fuente de información exhaustiva y almacenable acerca de las motivaciones de millones de posibles usuarios/clientes en la web. La gran mayoría de los datos que se generan respecto a los movimientos de los usuarios en la Web y el comercio electrónico en particular, provienen de lo que conoce-

mos como el Clickstream o la huella digital, es decir, las direcciones accedidas a lo largo de una Web.

En este caso de uso, nos centramos en el análisis predictivo mediante técnicas de regresión como las explicadas en el Capítulo 8 sobre el análisis del Big Data. Concretamente utilizaremos herramientas o librerías de algoritmos de minería de datos basadas en el lenguaje *«R»* y sus algoritmos para Big Data en la librería *«RHadoop»* [211]. Un manual completo sobre la instalación y configuración de esta librería se puede encontrar en el enlace de manuales de la Hortonworks[212], donde además se proporcionan las guías rápidas de ejemplos y el código fuente en el que nos basamos para este caso de estudio.

El análisis consistirá en predecir en número de visitantes a cierta tienda de comercio online para el siguiente periodo y por cada país o región, para lo cual utilizaremos el algoritmo de regresión lineal (explicado en el Capítulo 8).

11.3.1 Análisis Predictivo

Para el análisis predictivo, el conjunto de datos inicial contiene los números de clics realizados en la tienda online por país o región, desde el día 3 de marzo hasta el día 15 de marzo. Utilizando estos datos, el algoritmo realizará una predicción de las visitas para el día 16 de marzo para cada región. Utilizamos para este ejemplo las funciones de R para análisis predictivo: «`lm`» (linear model) para generar el modelo y «`predict`» para realizar la predicción (el material necesario se encuentra disponible en la Web de tutoriales de Hortonworks).

El resultado se almacena en un fichero de salida nombrado por defecto «`part-00000`», el cual se sitúa en el directorio HDFS especificado «`/user/hue/hdp/out`». Este fichero de resultados se puede inspeccionar mediante el navegador de ficheros de la Sandbox, tal y como muestra la Figura 148. La predicción consiste básicamente en las zonas geográficas o regiones de donde "procederían" (1ª columna) las visitas y su cantidad (2ª columna).

11.3.2 Visualización

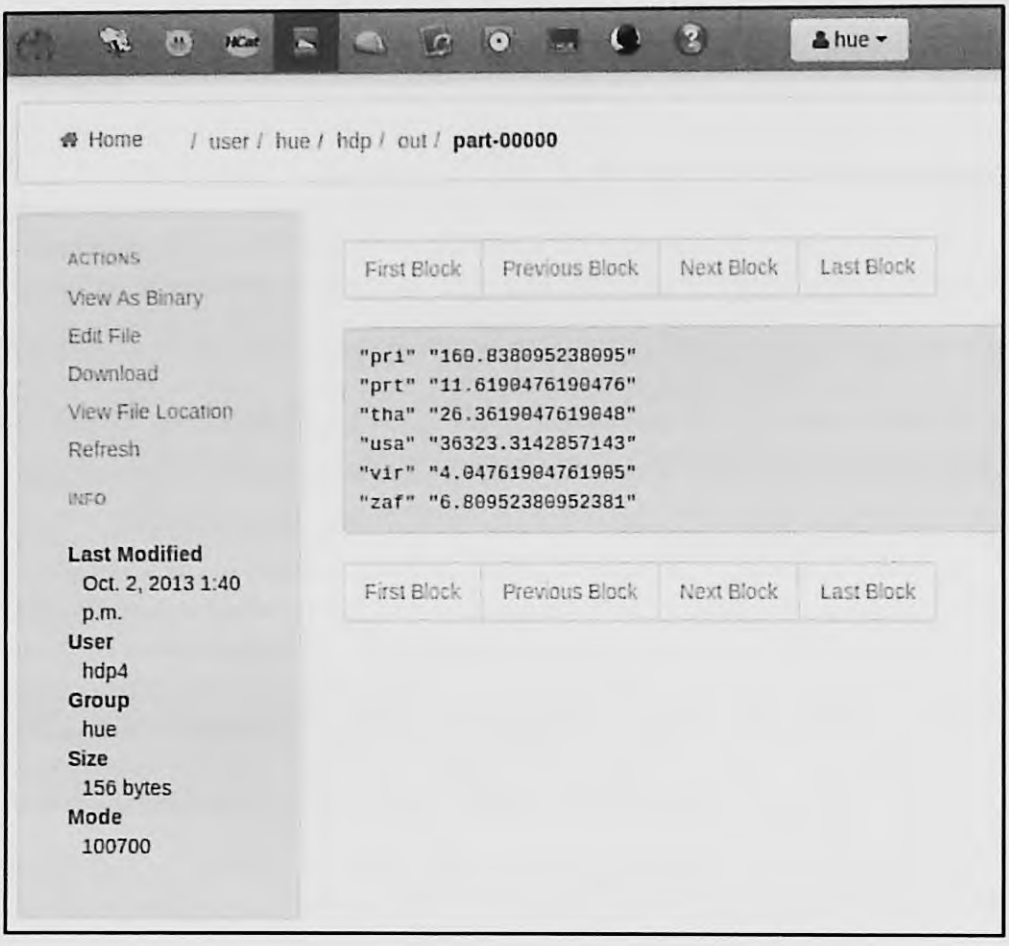
En este apartado haremos algunos ejemplos de visualización de datos de Clickstream. Nos centramos en los datos de logs de visitas de Omniture, a los cuales les añadimos los datos de usuarios y productos cargados previamente en las tablas de «users» y «products», respectivamente. Como herramienta de visualización utilizaremos Microsoft Excel, por lo que explicaremos también el proceso de conexión y carga de datos.

Previamente, como parte del procesamiento y preparación de los conjuntos de datos, crearemos un conjunto unificado para la unión de las tres tablas mediante la herramienta

[211] *Revolution Analytics*. https://github.com/RevolutionAnalytics/RHadoop/wiki. [Online; consultado el 23 de enero de 2016]

[212] *Using RHadoop to predict website visitors*. http://hortonworks.com/hadoop-tutorial/using-rhadoop-to-predict-visitors-amount/. [Online; consultado el 26 de enero de 2016]

Figura 148 – Fichero de resultados de predicción de Clickstream



Hive. Por tanto, en el editor de consultas de esta herramienta introducimos y ejecutamos la consulta para unión de estos datos.

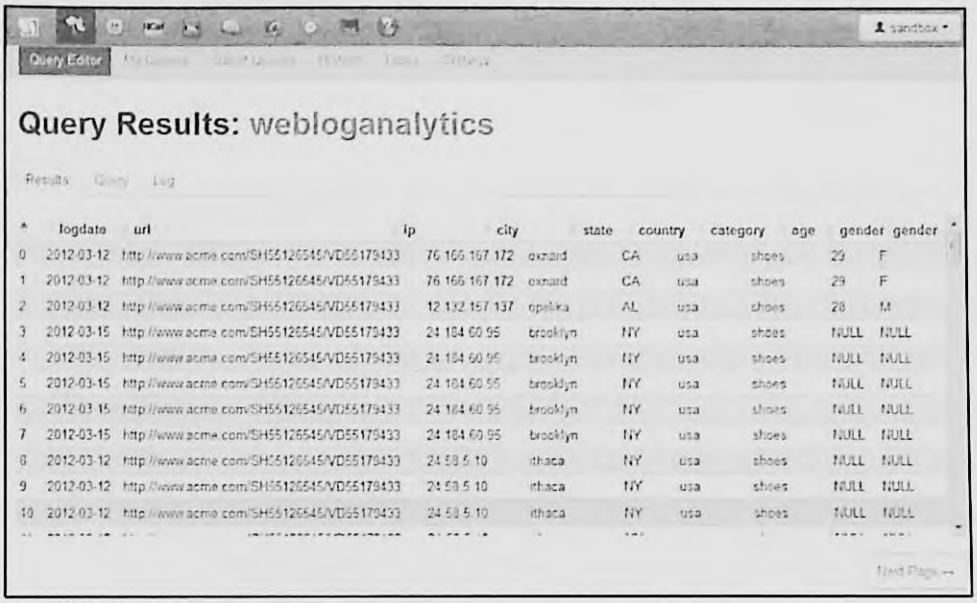
El resultado se puede visualizar en la pestaña de «*results*» del mismo editor para la tabla «*webloganalytics*», tal y como se muestra en la Figura 149. De forma que en cada línea tendríamos los datos de una visita, el usuario que ha visitado el producto y los detalles del producto visitado.

11.3.2.1 Conexión a los datos mediante Microsoft Excel

Para esta actividad nos centramos en la versión Microsoft Excel Professional Plus 2013[213], aunque los pasos a seguir en otras versiones actuales de Excel son similares.

[213] *Office Professional Plus 2013*. https://products.office.com/es-es/professional-plus/office-professional-plus-2013. [Online; consultado el 26 de enero de 2016]

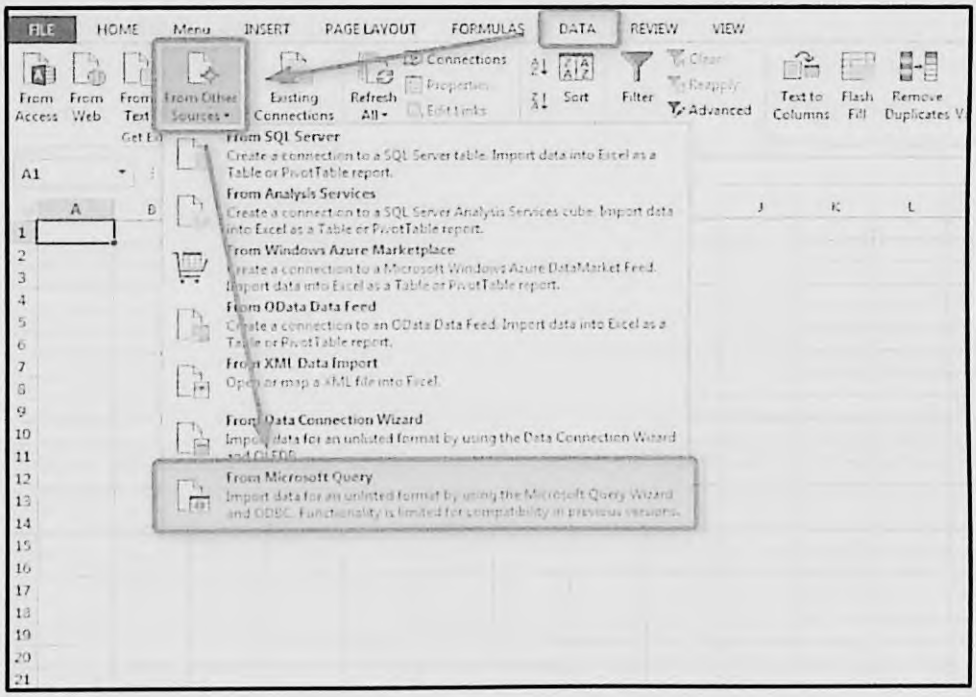
Figura 149 – Resultado de consulta Hive para la generación de la tabla «webloganalytics»



Query Results: webloganalytics

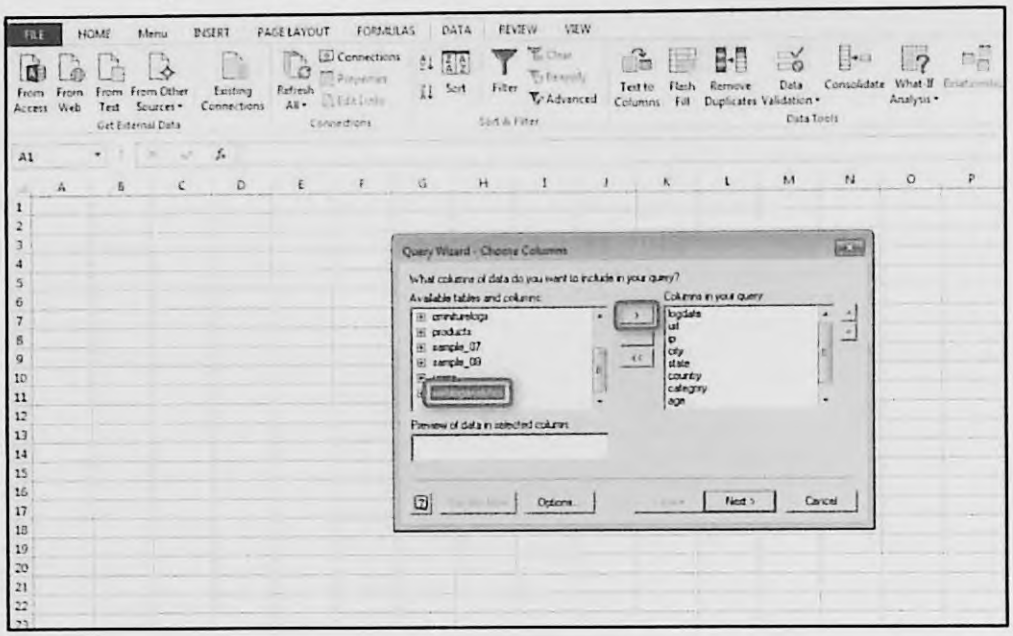
*	logdate	url	ip	city	state	country	category	age	gender	gender
0	2012-03-12	http://www.acme.com/SH55126545/VD55179433	76.166.167.172	oxnard	CA	usa	shoes	29	F	
1	2012-03-12	http://www.acme.com/SH55126545/VD55179433	76.166.167.172	oxnard	CA	usa	shoes	29	F	
2	2012-03-12	http://www.acme.com/SH55126545/VD55179433	12.132.157.137	opelika	AL	usa	shoes	28	M	
3	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes	NULL	NULL	
4	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes	NULL	NULL	
5	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes	NULL	NULL	
6	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes	NULL	NULL	
7	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes	NULL	NULL	
8	2012-03-12	http://www.acme.com/SH55126545/VD55179433	24.58.5.10	ithaca	NY	usa	shoes	NULL	NULL	
9	2012-03-12	http://www.acme.com/SH55126545/VD55179433	24.58.5.10	ithaca	NY	usa	shoes	NULL	NULL	
10	2012-03-12	http://www.acme.com/SH55126545/VD55179433	24.58.5.10	ithaca	NY	usa	shoes	NULL	NULL	

Figura 150 – Tipo de conexión de Microsoft Excel con otras fuente de datos



El primer paso consiste en abrir una nueva hoja de Microsoft Excel y seleccionar como tipo de datos «otras fuentes» y consulta Microsoft, tal y como muestra la Figura 150. Es decir, realizaremos la secuencia de pasos: «`Data >From Other Sources >From Microsoft Query`».

Figura 151 – Ventana de carga de tablas en Excel desde Hortonworks



Se muestra entonces una ventana emergente con las fuentes de datos disponibles, donde debe aparecer y seleccionarse Hortonworks ODBC (que debe instalarse previamente). Este *driver* de Hortonworks nos posibilita el acceso desde Microsoft Excel a nuestras fuentes de datos Hadoop.

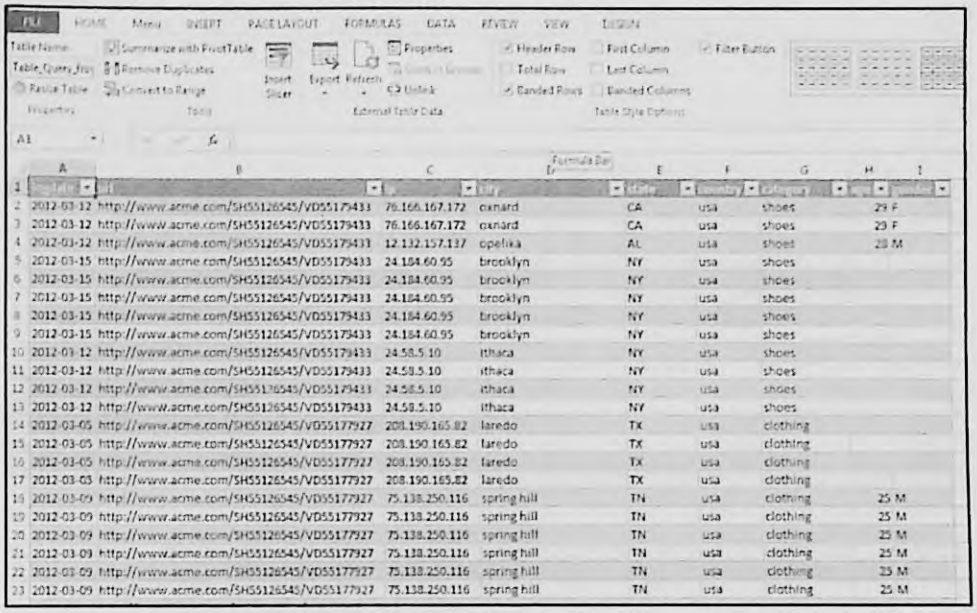
Una vez establecida la conexión con la fuente de datos en la *Sandbox*, aparece la ventana emergente de consultas (Query Wizard) mostrada en la Figura 151. Seleccionaremos entonces la tabla previamente generada «webloganalytics» en la columna de tablas disponibles y la incluimos en la comuna de tablas para consultar (Colums in your query) mediante el botón de la flecha «>». Para continuar, hacemos clic en el botón «*Next*».

Aparece entonces un segundo cuadro de diálogo para el filtrado de datos, en el cual pulsaremos de nuevo sobre el botón «*Next*», ya que no es necesario filtrar para este ejercicio de visualización. Lo mismo haremos en la siguiente ventana de ordenación de datos: pulsamos de nuevo «*Next*». Finalizamos en el Query Wizard pulsando el botón de finalización «Finish», pero teniendo seleccionada la opción de «Return data to Microsoft Excel».

Pasaremos a un nuevo cuadro de diálogo para efectuar la importación de los datos. En este cuadro pulsamos «*OK*» para realizar la importación con la opción de tabla por defecto. Tras esto, se cargan los datos en la tabla de Excel (véase la Figura 152) y ya están preparados para su visualización.

Este proceso de conexión con la fuente de datos es similar en otras herramientas de visualización, como ya vimos para el caso de «*Tableau*» o «*Gephi*» (descritas en el Capítulo 9), por lo que el caso de uso se puede hacer extensible a estas herramientas.

Figura 152 – Carga de datos de *webloganalytics* en Excel



	A	B	C	D	E	F	G	H	I
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	2012-03-12	http://www.acme.com/SH55126545/VD55179433	76.166.167.172	oxnard	CA	usa	shoes	29 F	
3	2012-03-12	http://www.acme.com/SH55126545/VD55179433	76.166.167.172	oxnard	CA	usa	shoes	29 F	
4	2012-03-12	http://www.acme.com/SH55126545/VD55179433	12.132.157.137	opelika	AL	usa	shoes	28 M	
5	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes		
6	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes		
7	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes		
8	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes		
9	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes		
10	2012-03-12	http://www.acme.com/SH55126545/VD55179433	24.58.5.10	ithaca	NY	usa	shoes		
11	2012-03-12	http://www.acme.com/SH55126545/VD55179433	24.58.5.10	ithaca	NY	usa	shoes		
12	2012-03-12	http://www.acme.com/SH55126545/VD55179433	24.58.5.10	ithaca	NY	usa	shoes		
13	2012-03-12	http://www.acme.com/SH55126545/VD55179433	24.58.5.10	ithaca	NY	usa	shoes		
14	2012-03-05	http://www.acme.com/SH55126545/VD55177927	208.190.165.82	laredo	TX	usa	clothing		
15	2012-03-05	http://www.acme.com/SH55126545/VD55177927	208.190.165.82	laredo	TX	usa	clothing		
16	2012-03-05	http://www.acme.com/SH55126545/VD55177927	208.190.165.82	laredo	TX	usa	clothing		
17	2012-03-05	http://www.acme.com/SH55126545/VD55177927	208.190.165.82	laredo	TX	usa	clothing		
18	2012-03-09	http://www.acme.com/SH55126545/VD55177927	75.138.250.116	spring hill	TN	usa	clothing	25 M	
19	2012-03-09	http://www.acme.com/SH55126545/VD55177927	75.138.250.116	spring hill	TN	usa	clothing	25 M	
20	2012-03-09	http://www.acme.com/SH55126545/VD55177927	75.138.250.116	spring hill	TN	usa	clothing	25 M	
21	2012-03-09	http://www.acme.com/SH55126545/VD55177927	75.138.250.116	spring hill	TN	usa	clothing	25 M	
22	2012-03-09	http://www.acme.com/SH55126545/VD55177927	75.138.250.116	spring hill	TN	usa	clothing	25 M	
23	2012-03-09	http://www.acme.com/SH55126545/VD55177927	75.138.250.116	spring hill	TN	usa	clothing	25 M	

11.3.2.2 Visualizando Clickstream

Como ya comentamos en el Capítulo 9, la visualización de los datos nos puede ayudar a optimizar nuestra información y a realizar análisis. En el caso de Clickstream, la visualización de los flujos de clics pueden ayudar a los analistas de sitios de comercio electrónico a optimizar sus tiendas online, con el fin de convertir sus visitas en ventas reales. Dentro de este análisis podemos hacer tareas como las siguientes:

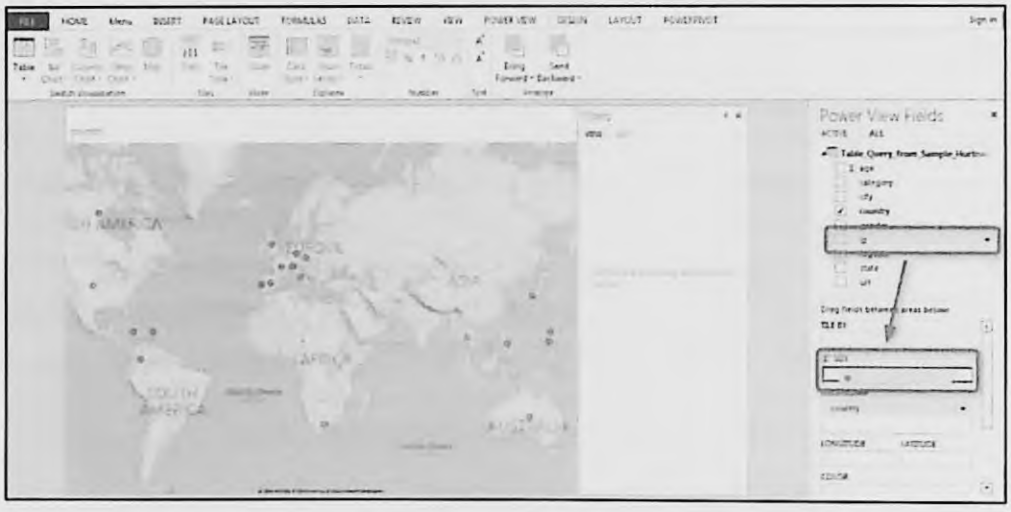
- Analizar los datos de clics por localización de origen.
- Filtrar los datos por categoría de producto.
- Hacer gráficas sobre el perfil del usuario de la tienda online: edad, género, etc.
- Observar un cierto segmento de clientes.
- Identificar aquellas páginas con mayor ratio de rebote o abandono rápido.

Empezamos por seleccionar en Microsoft Excel una nueva vista de selección de datos con los que trabajar. Pulsamos la secuencia: «`Insert >Power View`». Los campos de selección de vista («*Power View Fields*») aparecen entonces en la parte derecha de la ventana con la tabla desplegada en la parte derecha (Figura 153).

Para empezar, podemos echar un vistazo a los países de origen de los visitantes a la tienda online analizada. En el área de «*Power View Fields*», dejaremos seleccionada únicamente

Figura 153 – Vista Excel de selección de campos

Figura 154 – Visualización de datos por localización de los usuarios que han visitado nuestro portal Web en una vista de mapa



la etiqueta de «*country*», por lo que la tabla se actualizará para mostrar sólo los datos del atributo seleccionado, es decir, los datos de países. Mediante la vista de mapas («*Map*») se mostrarán aquellos países con información en nuestra base de datos (Figura 154). El siguiente paso podría ser seleccionar la suma de direcciones IP por país/estado. Para ello, se arrastra el campo de IP sobre la caja «*SIZE*».

Del mismo modo, arrastramos el campo «*country*» desde el panel de «*Power View Fields*» hasta el área de filtros «*Filter*» y seleccionamos el país «*usa*». A continuación, arrastramos también el campo «*state*» a la caja de «*LOCATIONS*» y borramos el campo de «*country*» de esta misma caja haciendo clic en la flecha hacia abajo y presionando en borrar campo «*remove field*».

Figura 155 – Vista de mapa de Estados Unidos con la suma de IPs por estado

Ya es posible utilizar los controles de la herramienta de «Map» para hacer zoom en la zona de Estados Unidos. Para comprobar el resultado, movemos el puntero del ratón sobre cada estado para ver la cuenta de diferentes IPs contabilizadas (véase la Figura 155).

También podemos incluir información sobre los productos, por lo que podemos visualizar las categorías de productos vistos por los usuarios en cada estado. Simplemente, debemos arrastrar el atributo «*category*» en el campo «*COLOR*» del panel de «*Power View Fields*». Como resultado, obtenemos un mapa sobre el que podemos situar el puntero de ratón en un estado determinado y ver las categorías más visitadas. Por ejemplo, en la Figura 156 se puede ver cómo el mayor número de páginas vistas en Florida se corresponde con los productos de prendas de vestir, seguido por los productos de calzado.

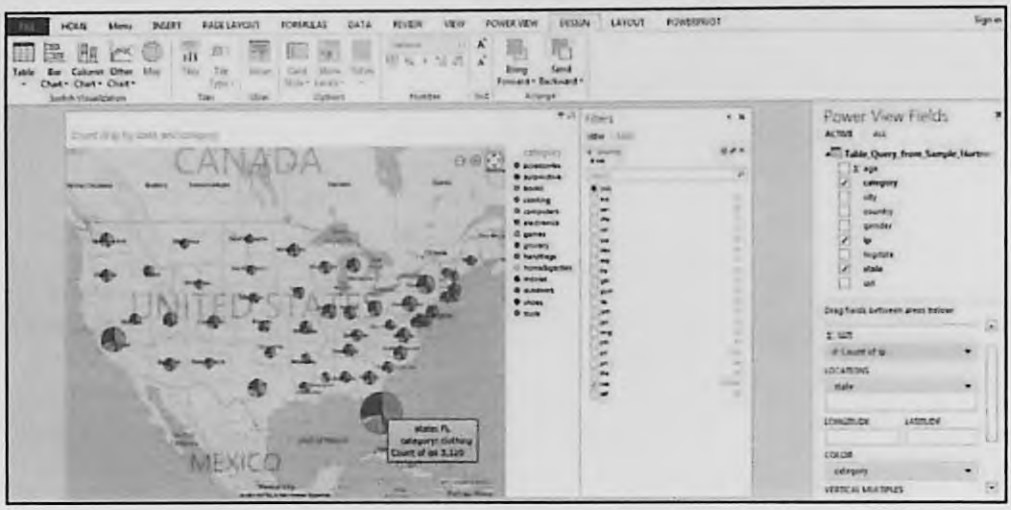
11.3.3 Análisis a Partir de la Visualización

En este punto, vamos un paso más adelante y analizamos nuevos aspectos de nuestros datos de Clickstream. En concreto vamos a analizar los datos relativos a las prendas de vestir por edad y por género del visitante, con el fin de optimizar este segmento de clientes.

Para ello, volvemos a generar una vista pulsando la secuencia: «`Insert >Power View`», por lo que se abre una nueva vista de informe («*report*»). Para acomodar los datos necesarios, realizaremos los siguientes pasos:

1. En el área de «*Power View Fields*», seleccionamos «*ip*» y «*edad*» (*age*). Los demás campos permanecerán sin seleccionar.
2. Arrastramos «*category*» hacia el área de filtros (*Filters area*). Tras esto, seleccionamos el atributo de prendas de vestir «*clothing*».

Figura 156 – Vista del mapa de Estados Unidos incluyendo los datos de productos visitados por estado

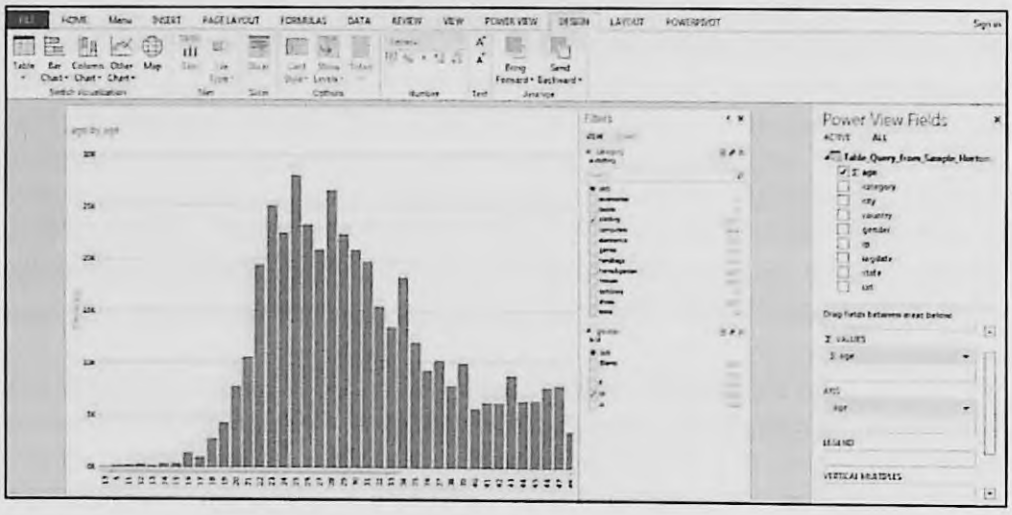


3. Arrastramos el género del visitante (gender) hacia el área de filtros (*Filters area*). Tras esto, seleccionamos el atributo de género «*M*» (*male*).

4. Desde el menú inicial, seleccionamos gráfico de columnas (*Column Chart*) y seguidamente «*Clustered column*».

Para terminar de generar el gráfico de barras, arrastraremos el atributo de «*edad*» (*age*) a la caja *AXIS*. Seguidamente, borraremos el atributo «*ip*» de la caja de *AXIS* desplegando el menú de la flecha hacia abajo y pulsando en borrar campo (*Remove Field*).

Figura 157 – Gráfico de barras correspondiente a la edad de los visitantes



Como resultado, la Figura 157 muestra el gráfico generado según el cual, la mayoría de los hombres que compran prendas de vestir en la tienda analizada tienen una edad entre 22 y 30 años. Con esta información, ya podemos optimizar nuestros contenidos para este segmento de población.

Como segundo análisis, podemos asumir que nuestros datos contienen información sobre las direcciones de páginas (*URLs*) que tienen un alto porcentaje de rebote (*bounce rate*), es decir, si estas páginas de nuestra tienda fueron las últimas que vio el visitante y a través de las cuáles salieron de nuestro negocio online. Por tanto, si filtramos estas *URLs* de nuestro segmento de edad, podremos saber qué páginas debemos optimizar para hacer que nuestros visitantes continúen interesados.

«La información es la gasolina del siglo XXI, y la analítica de datos el motor de combustión». Peter Sondergaard, Vicepresidente Senior de Gartner.

Glosario

ACID

es el acrónimo que hace referencia a las propiedades de transaccionalidad de algunos sistemas de bases de datos: atomicidad, consistencia, aislamiento y durabilidad. 186, 187, 194, 205

API

La interfaz de programación de aplicaciones, abreviada como API1 (del inglés: Application Programming Interface), es el conjunto de subrutinas, funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción. 143, 332

árbol de decisión

es un modelo de aprendizaje automático que se emplea en problemas de clasificación, donde siguiendo una estructura en forma de árbol se baja por las ramas en función de los valores de los atributos hasta llegar a un nodo hoja que contiene la clase. 246–248, 251, 266, 270–276

árbol de regresión

es un modelo de aprendizaje automático que se emplea en problemas de regresión, donde siguiendo una estructura en forma de árbol se baja por las ramas en función de los valores de los atributos hasta llegar a un nodo hoja que contiene la salida. 251

atributo

es una propiedad de las instancias de un conjunto de datos. 245–247, 249, 250, 255–258, 263, 266–268, 270–273, 275–280

BigML

es una herramienta online que permite realizar tareas de aprendizaje automático directamente en la nube, permitiendo subir grandes cantidades de datos y visualizar de forma gráfica los modelos. 266–280

BigQuery

es un servicio de Google Cloud que permite realizar analíticas de datos empleando sentencias SQL sobre grandes cantidades de datos, que pueden llegar a ocupar varios petabytes. 236–239

bit rot

A menudo se define al bit rot como el evento por el cual las pequeñas cargas eléctricas de un bit de memoria se dispersan, posiblemente alterando el código de un programa. 160

Bolt

es un tipo de nodo en Storm que se encarga de procesar los datos que recibe como entrada de nodos Spout o bien de otros nodos Bolt, generando una nueva salida. 224, 226–229

cinta

también denominada "cola", es una estructura de datos en la que los datos siempre se introducen por un extremo (el final de la cola) y se extraen por el otro extremo (el inicio de la cola). 225

clase

en problemas de clasificación, es el valor para el que se desea aprender un modelo, con el fin de poder averiguarlo en función de los demás atributos de las instancias. 244–256, 270–272, 274, 278

Cloud SQL

es un servicio en la nube de Google que permite crear bases de datos relacionales, similar a RDS de Amazon. 203–205

Cloud Storage

es un servicio de almacenamiento en la nube de Google, similar a S3 de Amazon. 198, 199

cluster (aprendizaje automático)

es un conjunto de instancias que guardan cierta relación entre sí o que se encuentran próximas en el espacio. 257, 259–261, 263, 264, 276–278

cluster (sistema distribuido)

es un conjunto de máquinas diseñado para colaborar entre ellas con el fin de aumentar la capacidad o el rendimiento con respecto al caso en el que solo se cuenta con una sola máquina, normalmente distribuyendo el almacenamiento o el procesamiento de los datos. 183–185, 190–192, 206, 213, 215, 216, 221, 222, 231–236, 239–241

clustering

también denominada "segmentación" o "agrupación", es el proceso por el cuál se obtienen *clusters* a partir de un conjunto de instancias, con el fin de agruparlas para que aquellas que sean similares pertenezcan al mismo grupo. 244, 256–265, 276, 278

código abierto

es la denominación que reciben aquellas herramientas de software que por su licencia permiten que cualquier usuario o desarrollador tenga acceso al código de la misma, permitiendo en algunos casos su modificación y reutilización. 182, 183, 187, 193

CRM

CRM proviene de la sigla del término en inglés customer relationship management, y puede poseer varios significados: 1) Administración basada en la relación con los clientes; 2) Software para la administración de la relación con los clientes. 142

CSS

Hoja de estilo en cascada o CSS (siglas en inglés de cascading style sheets) es un lenguaje usado para definir y crear la presentación de un documento estructurado escrito en HTML. 289

D3.js

D3.js (o simplemente D3 por las siglas de Data-Driven Documents) es una librería de JavaScript para producir, a partir de datos, infogramas dinámicos e interactivos en navegadores web. 288, 291

data cleansing

El data cleansing, data scrubbing o limpieza de datos, es el acto de descubrimiento, corrección o eliminación de datos erróneos de una base de datos. 154

Data marts

Sistema de almacenamiento de datos especializado (de un departamento o área temática concreta), procedentes de una o varias fuentes, con una estructura óptima que permite analizar los datos que contiene desde distintas perspectivas. 354

Data Warehouse

Sistema, utilizado en los entornos corporativos, que se caracteriza por integrar, depurar y agregar información de una o varias fuentes de datos distintas, para posteriormente procesarla y permitir su análisis desde diferentes perspectivas de forma rápida. 152, 154, 156, 354, 362

detección de comunidades

es el procedimiento por el que, en una red social, se busca encontrar comunidades de usuarios que tengan ciertas características en común, o bien tengan lazos muy fuertes entre ellos. 263

Dimension Table

Dimension Table, o Tabla de Dimensión, es la que contiene atributos (o campos) que se utilizan para restringir y agrupar los datos almacenados en una tablas de hechos cuando se realizan consultas sobre dichos datos en un entorno de almacen de datos o data mart. plural. 173, 174

distancia coseno

es el ángulo que forman los dos vectores definidos por las instancias en el espacio. 258

distancia euclídea

es la distancia en línea recta entre dos puntos en el espacio. 250, 257, 258, 260

distancia Manhattan

es la distancia en que habría que cruzar en una rejilla para unir los dos puntos en el espacio (es decir, no puede contener segmentos diagonales). 258, 260

Dynamo DB

es un servicio de Amazon que proporciona una base de datos NoSQL consistente en un almacén clave-valor con determinadas características de una base de datos orientada a documentos. 206–209

EMR

acrónimo de Elastic MapReduce, es un servicio de Amazon para desplegar *clusters* de Hadoop en la nube. 232–236

enriquecimiento de datos

El enriquecimiento de datos permite completar las bases de datos con toda la información adicional, que resulte necesaria para conseguir unos objetivos de análisis para la ayuda a la toma de decisiones. 143

ERP

Los sistemas de planificación de recursos empresariales (ERP, por sus siglas en inglés, enterprise resource planning) son sistemas de información gerenciales que integran y manejan muchos de los negocios asociados con las operaciones de producción y de los aspectos de distribución de una compañía en la producción de bienes o servicios. 142

escalabilidad horizontal

es una propiedad de aquellos sistemas que son capaces de mejorar sus capacidades cuando se distribuyen entre varios ordenadores (por ejemplo, permitiendo disfrutar de más almacenamiento, rendimiento, etc). 184, 186, 193, 194, 200, 202, 215, 225, 231

fact table

Fact Table, o Tabla de Hechos, es la tabla central de un esquema dimensional (en estrella) y contiene los valores de las medidas de negocio. plural. 173, 174, 337, 354

filtrado basado en contenidos

es un enfoque para la recomendación en la que los ítems recomendados están basados en otros en los que el usuario ha mostrado interés o ha adquirido. 262

filtrado colaborativo

es un enfoque para la recomendación en la que los ítems recomendados están basados en las preferencias de usuarios similares a aquél al que se le muestra la recomendación. 262, 263

FK

Foreing Key, o Clave Foránea, se utiliza en el modelado entidad-relación para identificar una columna o grupo de columnas en una tabla (tabla hija o referendo) que se refiere a una columna o grupo de columnas en otra tabla (tabla maestra o referenciada). Las columnas en la tabla referendo deben ser la clave primaria u otra clave candidata en la tabla referenciada. plural. 174

FTP

Abreviatura de File Transfer Protocol, el protocolo para intercambiar archivos en Internet. El FTP utiliza los protocolos de Internet TCP/IP para permitir la transferencia de datos, de la misma manera que el HTTP en la transferencia de páginas

web desde un servidor al navegador de un usuario y el SMTP para transferir correo electrónico a través de Internet. 145

función de distancia

es una función que recibe como entrada dos instancias y devuelve como resultado una medida de distancia entre ellas. 256–260

GFS

es un sistema de ficheros distribuido presentado por Google en el año 2003, que ha servido de referencia para posteriormente desarrollar una variante de código abierto denominada Hadoop Distributed File System (HDFS). 182–184, 193, 214, 215

Google Chart

Google Chart es una aplicación de Google para realizar estadísticas web, de fácil uso para desarrolladores de software web. 288

GTIN

Global Trade Item Number o GTIN es el número mundial de un artículo comercial, el cual es utilizado para identificar de manera única a cualquier producto o ítem sobre el cual existe una necesidad de obtener una información específica y al cual se le debe asignar un precio. 153

Hadoop

Proyecto de software libre administrado por la fundación Apache que consta de diferentes elementos de software que permiten el almacenamiento y procesamiento de grandes volúmenes de datos en sistemas distribuidos. 159

hash

en ocasiones denominado "resumen", hace referencia al resultado de aplicar una función de transformación sobre un dato o un valor, presentando por lo general mayor aleatoriedad que el dato original. 187, 190, 206, 207

HBase

es una base de datos NoSQL de Apache incorporada a Hadoop, que presenta un modelo de datos basado en columnas y almacena los datos sobre el sistema de ficheros HDFS. 193, 194, 231, 234

HDFS

es un sistema de ficheros distribuido y de código abierto perteneciente al núcleo de Hadoop y basado en las especificaciones publicadas por Google en su sistema de ficheros GFS. 183–185, 187, 194, 221, 222, 231, 233

HTML

HTML, siglas de HyperText Markup Language («lenguaje de marcas de hipertexto»), hace referencia al lenguaje de marcado para la elaboración de páginas web. 289, 290, 353, 360

instancia (aprendizaje automático)

es una muestra o un elemento de un conjunto de datos. 245–247, 249–254, 256–261, 263, 264, 266, 267, 269, 272, 273, 275–280

instancia (nube)

es el nombre que reciben en ocasiones las máquinas virtuales desplegadas en la nube. 196, 198, 200–205, 232, 235

integridad de los datos

El término integridad de datos se refiere a la corrección y complementación de los datos en una base de datos. 325, 326, 359

JavaScript

JavaScript (abreviado comúnmente JS) es un lenguaje de programación interpretado, orientado a objetos, basado en prototipos, imperativo, débilmente tipado y dinámico. Se utiliza principalmente en su forma del lado del cliente (client-side), implementado como parte de un navegador Web permitiendo mejoras en la interfaz de usuario y páginas Web dinámicas. 289, 290, 353, 360

JDBC

Java Database Connectivity (JDBC) es un derivado de ODBC inspirado en el mismo pero ofrecido como una interfaz de programación de aplicaciones para el acceso a bases de datos desde el lenguaje de programación Java. 145

JobTracker

es un tipo de nodo en Hadoop MapReduce que coordina la ejecución de trabajos MapReduce, dividiendo las tareas *map* y *reduce* entre los diferentes nodos TaskTracker. 222, 223, 225, 226

JOIN

es una sentencia del lenguaje de consulta de bases de datos SQL que permite combinar el contenido de dos tablas mediante un valor que tengan en común. 186, 187, 189, 192, 205, 217, 218, 220, 236

JSON

es el acrónimo de *JavaScript Object Notation*, es un formato para el intercambio de datos que por su estructura es muy ligero y fácil de procesar por un ordenador. 188, 189, 209, 229

k-Nearest Neigbors

es un modelo de aprendizaje automático geométrico, en el que la predicción de la clase o salida de una instancia se realiza basándose en la clase o salida de las k instancias más cercanas a ella. 249

Lookups

Es una técnica de búsqueda de información en una tabla de una base de datos usando como semilla un dato de entrada. 162

LOPD

La Ley Orgánica 15/1999 de 13 de diciembre de Protección de Datos de Carácter Personal, (LOPD), es una Ley Orgánica española que tiene por objeto garantizar y proteger, en lo que concierne al tratamiento de los datos personales, las libertades públicas y los derechos fundamentales de las personas físicas, y especialmente de su honor, intimidad y privacidad personal y familiar. 318

lotes (procesamiento)

es un enfoque para el procesamiento de datos que no se realiza en tiempo real, sino sobre información histórica (lotes de datos) previamente disponible. 211–214, 226, 228–230, 232

Mahout

es un proyecto de Apache que se integra con Hadoop y permite realizar tareas de aprendizaje automático sobre los datos. 264

map

es la primera fase del paradigma de programación MapReduce, en la que se puede realizar un "mapeo" (cambio de dominio) y filtrado de los datos. 216–224, 229

MapReduce

es un paradigma de programación publicado por Google en el año 2004 que permite el procesamiento por lotes de grandes cantidades de datos, distribuidos en un *cluster* de varias máquinas. 183, 213, 215, 216, 218–224, 226, 229, 230, 233–236, 264, 265

Microsoft SQL Server

es una de las bases de datos relacionales comerciales más relevantes. 200

MLLib

es un subproyecto de Apache Spark que permite, utilizando las librerías de Spark, realizar tareas de aprendizaje automático sobre los datos. 264, 265

MySQL

es una base de datos relacional de código abierto cuyo uso está muy extendido sobre todo para dar soporte a aplicaciones web. 200, 204

Naïve Bayes

es un modelo de aprendizaje automático probabilístico, en el que la predicción de la clase de realiza en base al teorema de Bayes y observando con qué probabilidades los atributos determinan el valor de la clase. 248, 249

NewSQL

este término hace referencia a nuevas tecnologías de bases de datos que tratan de combinar la escalabilidad de sistemas NoSQL con las garantías de transaccionalidad ACID de las bases de datos relacionales. 194, 195

nodo

es un ordenador perteneciente a un sistema distribuido (también denominado *cluster*). 183–187, 190, 192, 194, 212, 213, 215, 216, 221, 222, 224–229, 231, 232, 234, 236, 240

normalización

El proceso de normalización de bases de datos consiste en designar y aplicar una serie de reglas a las relaciones obtenidas tras el paso del modelo entidad-relación al modelo relacional. Las bases de datos relacionales se normalizan para: Evitar la redundancia de los datos, Disminuir problemas de actualización de los datos en las tablas y Proteger la integridad de los datos. 26, 40, 144, 157, 162

NoSQL

este término hace referencia a las tecnologías y sistemas de bases de datos que emplean un modelo de datos alternativo al relacional, como pueden ser las clave-valor, documentales, basadas en grafos, etc. 185, 186, 188, 195, 205, 206

ODBC

Open DataBase Connectivity (ODBC) es un estándar de acceso a las bases de datos desarrollado por SQL Access Group (SAG) en 1992. El objetivo de ODBC es hacer posible el acceder a cualquier dato desde cualquier aplicación, sin importar qué sistema de gestión de bases de datos (DBMS) almacene los datos. 145, 344, 357

OLAP

es el acrónimo de On-Line Analytical Processing o procesamiento analítico en línea. Hace referencia a aquellos sistemas orientados al procesamiento analítico de grandes cantidades de datos con el fin de extraer información útil. 171, 172

OLTP

es el acrónimo de OnLine Transacion Procesing, o procesamiento de transacciones en línea. Hace referencia a aquellos sistemas diseñados para responder de forma inmediata a las peticiones del usuario. Estas peticiones suelen ser consideradas transacciones que implican una interacción con datos. 142

Open Data

Los Datos Abiertos (Open Data) son datos que pueden ser libremente utilizados, reutilizados y redistribuidos por cualquier persona. 143–145

Oracle Database

es la principal base de datos relacional disponible de forma comercial. 200

PostgreSQL

es una base de datos relacional de código abierto. 200

profiling

El perfilado de datos (Data Profiling) es un proceso en que se examina una fuente de datos existente para recolectar estadísticas e información sobre esos datos. 152–156, 180, 325

rack

también llamado "armario", es un espacio físico de un centro de cálculo o centro de datos en el que se apilan varios servidores informáticos, que comparten un sistema de alimentación y en algunos casos también equipo de red, etc. 185

RDS

acrónimo de Relational Database Service, es un servicio de Amazon Web Services que permite desplegar bases de datos relacionales en la nube, de tal forma que no es necesario que el usuario tome decisiones con respecto a la arquitectura física. 200–203

Recline.js

Una biblioteca simple pero potente para la creación de aplicaciones de datos en JavaScript y HTML. 290

recomendación

es un conjunto de instancias que guardan cierta relación entre sí o que se encuentran próximas en el espacio. 244, 261–264

reduce

es la segunda fase del paradigma de programación MapReduce, en la que se realiza una agregación de los datos resultantes de la fase *map* tras su ordenación y agrupación por clave. 216, 218–224, 229

regresión lineal

es una técnica de regresión que dado un conjunto de instancias en el espacio, aprende la recta (o el hiperplano) que mejor las aproxima. 251

REST

La Transferencia de Estado Representacional (Representational State Transfer) o REST es un estilo de arquitectura software para sistemas hipermedia distribuidos como la World Wide Web. 145

rutina

es un conjunto de instrucciones de código que al ejecutarse llevan a cabo una determinada tarea. 216–224, 227, 229

S3

acrónimo de Simple Storage Service, es un servicio de almacenamiento en la nube de Amazon, que permite subir ficheros y directorios que posteriormente se pueden descargar o a los que se puede acceder desde instancias de Amazon EC2. 196–199, 266

salida

en problemas de regresión, es el valor para el que se desea aprender un modelo, con el fin de poder averiguarlo en función de los demás atributos de las instancias. 245, 246, 250

SCM

La administración de redes de suministro (en inglés, Supply chain management, SCM) es el proceso de planificación, puesta en ejecución y control de las operaciones de la red de suministro con el propósito de satisfacer las necesidades del cliente con tanta eficacia como sea posible. 142

shard

es el nombre que recibe un nodo en un *cluster* de MongoDB, que permite que esta base de datos pueda escalar horizontalmente. 190

shuffle

es una fase intermedia del paradigma de programación MapReduce, en la que la salida de la rutina *map* se ordena y agrupa por clave antes de ser introducida a la rutina *reduce*. 216–218, 220, 222, 223

SKU

Stock-keeping unit o SKU (en castellano número de referencia) es un identificador usado en el comercio con el objeto de permitir el seguimiento sistémico de los productos y servicios ofrecidos a los clientes. 153

SOAP

SOAP (siglas de Simple Object Access Protocol) es un protocolo estándar que define cómo dos objetos en diferentes procesos pueden comunicarse por medio de intercambio de datos XML. 145

Spark

es un proyecto de Apache que integra diferentes soluciones para el procesamiento de datos por lotes o en tiempo real, así como para realizar analíticas y aprendizaje automático. 230, 231, 234, 239–242, 264, 265

SPARQL

SPARQL es un acrónimo recursivo del inglés SPARQL Protocol and RDF Query Language. Se trata de un lenguaje estandarizado para la consulta de grafos RDF, normalizado por el RDF Data Access Working Group (DAWG) del World Wide Web Consortium (W3C). 145

Spout

es un tipo de nodo en Storm que se encarga de la gestionar la entrada del flujo de datos, es decir, que está conectada directamente al origen de datos. 224, 226–228

SQL

es el acrónimo de Structured Query Language o Lenguaje estructurado de consulta. Es el lenguaje de programación utilizado para el acceso y modificación de los datos en los sistemas de gestión de bases de datos relacionales. 169, 172

Storm

es un sistema de código abierto de Apache para el procesamiento distribuido de datos en tiempo real. 214, 223–227, 230

stream

es la traducción literal al inglés de "flujo", y hace referencia a aquellos casos en los que los datos se introducen al sistema en tiempo real con el fin de ser procesados, analizados, etc. 211, 223, 227, 229

SWID

SWID, SoftWare IDentifier. Identificador software mediante código unitario, comúnmente utilizado apara identificar a usuarios de forma unívoca. 330, 337

TAL

Trunc and Load (TAL), o Corta y Carga, es una estrategia de proceso de carga que consiste en limpiar el repositorio de datos y cargar de nuevo toda la información con el nuevo contenido en el almacén de datos.. 177

TaskTracker

es un tipo de nodo en Hadoop MapReduce que ejecuta tareas *map* y *reduce* sobre los datos. 222, 223, 225

tupla

es una lista ordenada de datos que tienen alguna relación entre sí, de tal modo que quedan agrupados como si fueran un único valor. 217–223, 227–229

visualización

La visualización científica es la transformación de datos científicos y abstractos en imágenes. 281–283, 285, 287, 291

INTRODUCCIÓN AL

BIG DATA

Big Data permite aprovechar la inmensa cantidad de datos que se generan cada día, especialmente a raíz de la eclosión de las redes sociales online, del crecimiento exponencial de dispositivos y de las redes de sensores. Estos datos debidamente utilizados hacen posible que el proceso de toma de decisiones sea más objetivo, y se base menos en la intuición. Con Big Data se pueden detectar tendencias, realizar predicciones de sucesos futuros, o extraerpatrones del comportamiento de los usuarios, para adaptar mejor los servicios a sus necesidades.

El contenido del libro se ha estructurado de forma que se ofrezca una visión global de todos los temas que forman parte de un análisis de Big Data. El libro se ha concebido con un carácter eminentemente aplicado y en el último capítulo se presenta un caso de estudio completo. Por todo ello, el libro puede utilizarse como manual de referencia para realizar una primera toma de contacto con el análisis de Big Data.

El libro puede ser de interés tanto para lectores que no tengan una formación técnica, como para aquellos con formación o amplia experiencia en el mundo de las TIC. Los temas se presentan vía ejemplos de forma que fácilmente se puedan visualizar las posibilidades que ofrece Big Data. Pero los principios teóricos esbozados y el uso de las herramientas propuestas en el libro, constituyen una rigurosa introducción al manejo de los datos.

9 788415 793946

Más información en: www.ingebook.com
978-84-15793-94-6