Introducción al *big data*

Jordi Conesa i Caralt
José Luis Gómez García

PID_00209840

UOC
Universitat Oberta de Catalunya

Ninguna parte de esta publicación, incluido el diseño general y la cubierta, puede ser copiada, reproducida, almacenada o transmitida de ninguna forma, ni por ningún medio, sea éste eléctrico, químico, mecánico, óptico, grabación, fotocopia, o cualquier otro, sin la previa autorización escrita de los titulares del copyright.

Índice

Introducción 5

1. Orígenes 7

2. Cambio de paradigma de *big data* 10
2.1. Analítica de negocio 11

3. Definición de *big data* 13
3.1. Volumen 13
3.2. Velocidad 15
3.3. Variedad 17
3.4. Veracidad 19

4. Escenario de adopción de *big data* 21

Resumen 25

Introducción

Es difícil definir con rigor qué es *big data*, ya que es un concepto relativamente nuevo que aún está en evolución. Por otra parte, como veremos más adelante, la definición más aceptada no se implementa a partir de lo que es, sino a partir de las características de los datos que pretende analizar.

En este módulo introductorio empezaremos describiendo los orígenes de *big data* y justificaremos por qué el *big data* puede considerarse un nuevo paradigma a la hora de tomar decisiones y no solo una nueva tecnología relacionada con la programación distribuida. Finalmente, se definirá *big data* y se mostrará un ejemplo donde el uso de técnicas *big data* son aconsejables.

1. Orígenes

El término *datos masivos*, que puede considerarse la traducción al castellano de *big data*, aparece por primera vez en el entorno de las ciencias. En particular, en la astronomía y en la genética, motivado por la gran explosión en la disponibilidad de datos que experimentaron estas ciencias durante la primera década del siglo XXI. Ejemplos de ello son el proyecto de exploración digital del espacio llamado Sloan Digital Sky Survey o el proyecto del genoma humano. El primero de ellos generó más volumen de datos en sus primeros meses que el total de los datos acumulados en la historia de la astronomía hasta ese momento. Por otro lado, el proyecto del genoma humano tenía como objetivo encontrar, secuenciar y elaborar mapas genéticos y físicos de gran resolución del ADN humano. Cabe tener en cuenta que el genoma de una persona es del orden de los 100 gigabytes.

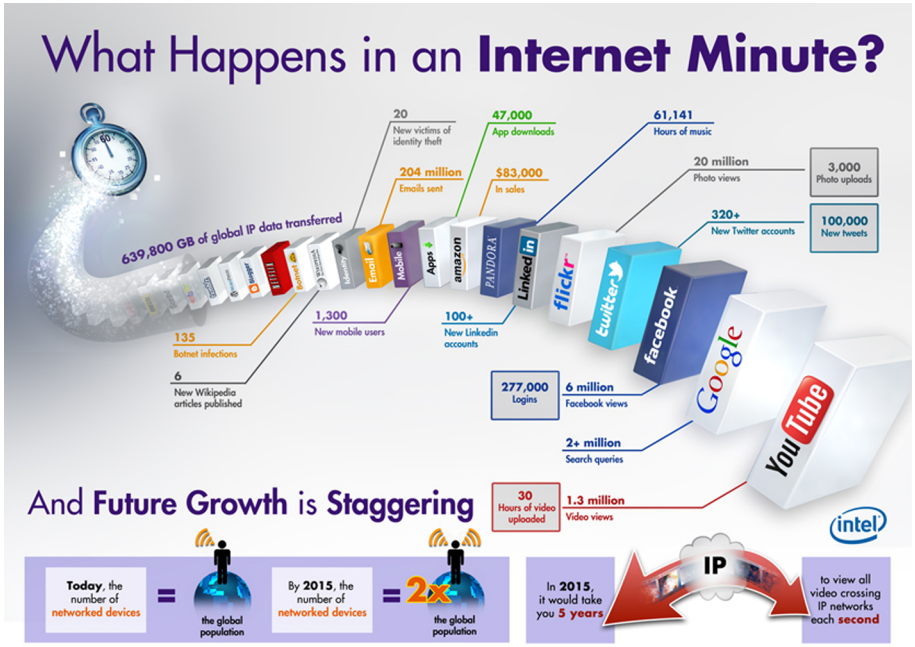
El proyecto Sloan Digital Sky Survey

El proyecto Sloan Digital Sky Survey tiene como objetivo identificar y documentar los objetos observados en el espacio. Este es uno de los estudios más ambiciosos e influyentes que se han realizado en la historia de la astronomía. Mediante el procesamiento de imágenes de gran parte del espectro luminoso, se han obtenido listas de objetos observados, así como varias características y magnitudes astronómicas, tales como posición, distancia, brillo o edad. En algo más de ocho años de operaciones, se han obtenido imágenes que cubren más de la cuarta parte del cielo, creando mapas en tres dimensiones que contienen más de 930.000 galaxias y más de 120.000 cuásares. http://www.sdss.org/

Desde un contexto más general, la explosión de datos en estos últimos años también ha sido una realidad. De hecho, desde mediados de la primera década del siglo XXI, el incremento del número de dispositivos con conexión a internet, junto al auge de las redes sociales, han provocado una explosión en el volumen de datos disponibles. Muchos de estos datos son abiertos y accesibles, lo que permite que puedan ser explotados por cualquier tipo de agente, incluidas las empresas.

Como ejemplo, la figura 1 muestra la cantidad de datos que se mueven en internet cada minuto (datos del 2013).

Figura 1. ¿Qué sucede en un minuto en internet?



Fuente Intel: http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html

Los datos masivos existen, pero disponer de gran cantidad de datos de por sí no aporta valor. El verdadero valor de los datos está en su análisis e interpretación, no en su generación. Por tanto, la aparición de los datos no solo responde a su disponibilidad, sino también a la aparición de tecnologías que permitan procesarlos, analizarlos e interpretarlos.

A medida que fue aumentando el volumen de los datos, se hizo más difícil alojarlos en la memoria que los ordenadores empleaban para procesarlos. Esto motivó la modernización y la evolución de las técnicas y tecnologías de procesado de datos tradicionales. Una parte muy importante de esta modernización vino de la mano de las mejoras en el hardware de los ordenadores y en su abaratamiento, que ayudó de manera decisiva al entorno de las ciencias, al menos en los primeros proyectos de datos masivos. No obstante, con más y mejor hardware no es suficiente. Pensad, si no, en cómo debería ser el ordenador de Google para indexar todos los contenidos de la web. También han sido necesarios cambios en las tecnologías software para procesar una gran cantidad de datos eficientemente.

La evolución de la tecnología basada en software surgió en el seno de grandes empresas de internet, como Google, Amazon y Yahoo! Estas se encontraron con que las técnicas de procesado de datos tradicionales no permitían tratar todos los datos que utilizaban de manera eficiente y tuvieron que crear sus propias tecnologías para poder continuar con el modelo de negocio que ellos mismos habían creado. Las premisas que siguieron para el replanteamiento tecnológico fueron las siguientes:

- Existe gran cantidad de datos que hacen inviable su procesado en un único ordenador. Por tanto, se debe usar procesamiento distribuido para involu-

crar distintos ordenadores que trabajen con los datos de manera paralela. Así podrán procesar más datos en menos tiempo.

- Los datos son heterogéneos y eso requiere nuevos modelos de datos para facilitar la inserción, la consulta y el procesamiento de datos de cualquier tipo y estructura. Estos nuevos modelos de datos han dado lugar a nuevas bases de datos, llamadas NoSQL, que utilizan estructuras de datos distintas a las del modelo relacional y que permiten tratar más eficientemente tipos de datos heterogéneos o muy relacionados.

- Los datos deben procesarse de forma rápida. Aunque haya que procesar muchos datos, su proceso debe ser rápido. Por ejemplo, un buscador web no sería útil si devolviera la búsqueda a nuestra consulta un día (o un minuto) después de haberla realizado.

Por ejemplo, en el caso del proyecto del genoma humano, en el 2012 la empresa Life Technologies presentó su herramienta The Ion Proton, la cual, siguiendo las premisas anteriores, era capaz de secuenciar el genoma completo de una persona en un día. La herramienta utilizaba técnicas de procesamiento paralelo y técnicas estadísticas de comparación, muy usadas en *big data*. De modo resumido, los pasos que seguía dicha herramienta para procesar el genoma humano de una persona en un día eran estos:

1) Dividir el problema en subproblemas de menor tamaño y complejidad: Secuenciadores de ADN digitalizan el genoma por partes, pequeños fragmentos de la secuencia del ADN. Se distribuyen las partes a distintos ordenadores distribuidos de manera que se procesen de forma paralela.

2) Componer la solución final a partir de la integración de las soluciones parciales de los subproblemas: Mediante procesamiento paralelo se ensamblan todas las pequeñas secuencias resultantes de la resolución de los subproblemas para formar la secuencia del genoma completo. En el proceso se ejecutan distintos controles de calidad, que permiten, por ejemplo, arreglar posibles duplicidades y errores de ensamblado y aplicar técnicas de comparación con los genomas de otros individuos para detectar variaciones y resolver ambigüedades en la secuencia individual.

Esta técnica de dividir un problema en problemas más pequeños y de menos complejidad que puedan tratarse de forma paralela y combinar después los resultados finales responde al nombre de *MapReduce* y es una de las técnicas de *big data* más utilizadas.

2. Cambio de paradigma de *big data*

Los datos masivos imponen un nuevo paradigma donde la correlación "sustituye" a la causalidad. Hasta ahora, los métodos de recogida y procesado de datos eran costosos y eso provocaba que al querer evaluar un fenómeno no se pudieran recoger todos los datos relacionados con él. En estos casos se elegía una pequeña muestra aleatoria del fenómeno, se definía un conjunto de hipótesis que comprobar y se estimaba con una cierta probabilidad que, para la muestra elegida, dichas hipótesis eran válidas. Hoy en día el paradigma ha cambiado, ya que es posible recoger datos de forma masiva, siendo capaces de tener información sobre la muestra completa de datos (o casi) relacionada con el fenómeno que hay que evaluar, es decir, toda la población. Por ejemplo, si una empresa quiere analizar los tuits que tratan sobre ella, es perfectamente factible recoger todos los tuits que la mencionan y analizarlos. Al encontrar correlaciones entre distintas variables de la muestra (por ejemplo, los adultos de una región geográfica consumen más productos de la empresa), podemos explotarlas aunque no sepamos la causa. Encontrar y probar la causa puede ser harto complejo y para el negocio no es necesario en absoluto. Eso implica un cambio de paradigma, donde explicar la causalidad pierde importancia respecto a la correlación.

Tal y como se ha comentado, el cambio de paradigma mental provocado por *big data* se basa en que:

- Ya no se trata de que nuestra experiencia o intuición nos indique si algo es plausible y, a posteriori, intentar confirmarlo mediante distintos enfoques, con unos pocos datos recogidos al efecto (la muestra).

- Ahora se trata de aunar la información disponible de toda la población en diversidad de medios (redes sociales, tiendas, clientes, investigación de mercados, vídeos, textos, sensores, etc.) y analizarla mediante diversos métodos estadísticos para descubrir aquellos hechos que realmente impactan en nuestra búsqueda, así como las interrelaciones entre los hechos ocurridos.

Este cambio de paradigma provoca que los sistemas analíticos se centren en encontrar "qué" aspectos afectan a la toma de decisión y no en "por qué" afectan esos aspectos. Al igual que ocurre en los sistemas BI tradicionales, se podrían responder cuestiones del tipo: "qué pasó", "qué está pasando" y "qué pasaría si", pero desde un punto de vista estadístico, no causal, donde no se busca la explicación del fenómeno, sino solo el descubrimiento del fenómeno en sí. En consecuencia, la causalidad entre hechos pierde terreno a favor de asociación (conexión, analogía, paralelismo y reciprocidad de estos hechos).

2.1. Analítica de negocio

El objetivo principal de la analítica de negocio es hacer inferencias, es decir, hacer predicciones o descubrir tendencias, sobre ciertas características de una población, para tomar decisiones que repercutan de manera positiva en el negocio. Dichas inferencias se realizan sobre la base de la información contenida en una muestra de la población elegida de forma aleatoria. La condición de aleatoriedad es esencial para cerciorarse de que la muestra es representativa con respecto a la población.

Al plantear una investigación estadística, el tamaño de la muestra es un factor crucial que tener en cuenta. Si la representatividad es suficiente, cuanto más grande sea la muestra, más exacta será la estimación resultante y la prueba de hipótesis se realizará con un mejor criterio estadístico. Evidentemente, si la muestra abarcara toda la población, la generalización de los resultados obtenidos sería inmediata e indiscutible.

En el entorno *big data*, podemos llegar a utilizar muestras que se aproximan mucho más al total de la población que las aproximaciones tradicionales. Esto es posible tanto porque somos capaces de recoger más datos (observaciones), como porque somos capaces de procesar más cantidad de datos en menor tiempo.

Otra característica, debida a la gran variedad de datos, es que resulta incluso posible analizar datos que en principio no parecían suficientemente relevantes como para ser encuestados, o simplemente los descartábamos por la imposibilidad de recogerlos o por su alta subjetividad.

Estos hechos elevan el análisis estadístico, sobre datos masivos, a nuevos niveles de eficacia. Esto es, al analizar datos procedentes de una muestra más cercana a la población real, podemos descubrir más información y con más fiabilidad. Algunos ejemplos que ilustran este cambio de paradigma son los siguientes:

- Google es uno de los mayores exponentes a la hora de recoger y correlacionar grandes volúmenes de datos. De hecho, almacena todos los criterios de búsqueda utilizados por los usuarios, así como las páginas accedidas tras sus búsquedas, junto con cierta información personal de los usuarios (como por ejemplo la fecha, hora, tipo de navegador, idioma del navegador y dirección IP de cada consulta), las páginas por las que navega, etc.

- Internet de las cosas se basa en que los objetos cotidianos tengan capacidad para conectarse a la red, ya sea para enviar información sobre su funcionamiento o sobre su entorno (mediante sensores integrados) o para recibir datos de otros dispositivos. La aplicación de esta filosofía masivamente aumentaría de manera significativa la información que tenemos sobre el mundo que nos rodea, ya que permitiría digitalizar y distribuir

La dirección IP

La dirección IP (IP es un acrónimo para *internet protocol*) es un número único e irrepetible con el cual se identifica una computadora o dispositivo conectado a una red. Dentro de internet y combinado con las bases de datos de proveedores de acceso a internet, sirve para, de manera aproximada, localizar geográficamente un dispositivo.

información hasta ahora desconocida, que puede dar lugar a correlaciones hasta ahora insospechadas.

- Analizando las palabras clave y los enlaces seleccionados junto con la dirección IP, Google ha sido capaz de predecir, con mayor anticipación que los organismos oficiales, futuras epidemias, como por ejemplo las epidemias de gripe (Google Flu Trends, http://www.google.org/flutrends/intl/es/es/#ES). Todo ello se realiza sin conocer los factores que producen la gripe (causalidad), sino fijándose en que una parte de una población geográficamente cercana (localizada a partir de su dirección IP) busca información sobre síntomas o remedios de la gripe (correlación). Con dicha información, los mecanismos de análisis de datos de Google son capaces de deducir que si muchos vecinos de una determinada zona están interesados sobre las causas o remedios de la gripe, es muy probable que exista un foco de gripe en esa zona.

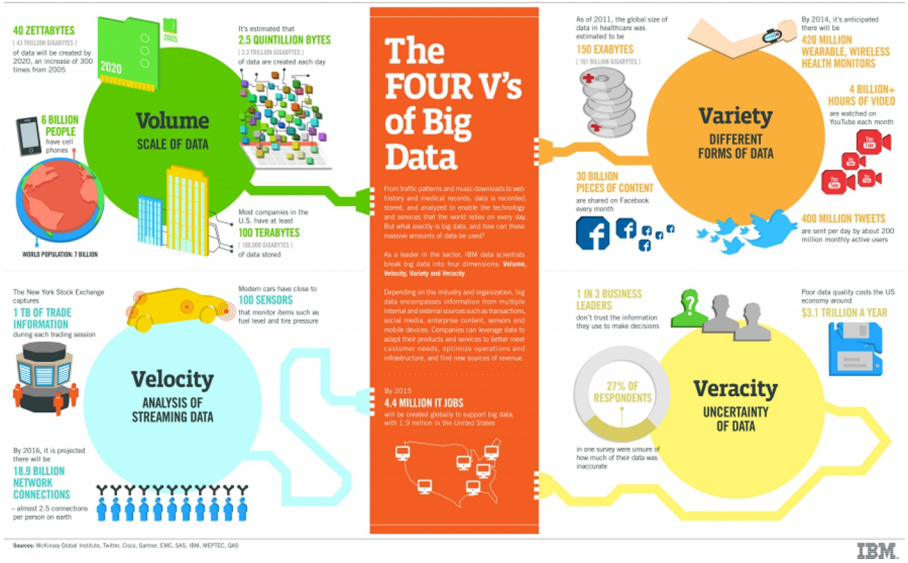
3. Definición de *big data*

Con el término *big data* se pretende describir las tecnologías, técnicas y metodologías relacionadas con el procesamiento de grandes y heterogéneos volúmenes de datos.

En el 2001, el analista Doug Laney de META Group (ahora Gartner) utilizaba y definía el término *big data* como el conjunto de técnicas y tecnologías para el tratamiento de datos, en entornos de gran volumen, variedad de orígenes y en los que la velocidad de respuesta es crítica. Esta definición, a partir de las características del entorno de los datos, se conoce como las 3 V del *big data*: volumen, velocidad y variedad. Hoy en día está comúnmente aceptado que la definición de las 3 V haya sido ampliada con una cuarta V, la veracidad.

El siguiente esquema muestra cómo interactúan las 4 V de *big data* según IBM: existen grandes volúmenes de datos (*volume*), de una confiabilidad cuando menos discutible (*veracity*), procedentes de una gran variedad de fuentes (*variety*) y que puede ser necesario procesar para obtener rápidas respuestas (*velocity*) que ayuden a tomar más y mejores decisiones.

Figura 2. Las 4 V de *big data* según IBM



Fuente Intel: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

A continuación se describen en más detalle las 4 V de la definición de *big data*.

3.1. Volumen

En los últimos años hemos vivido una gran explosión de datos. Se estima que el volumen de datos existente en la actualidad está por encima del zettabyte[1] y que crecerá de forma exponencial en el futuro. A nivel mundial, por poner

(1) Zettabyte (ZB) = 1000000000000000000000 bytes = 10^{21} bytes.

un par de ejemplos ilustrativos, cada día se crean 2,5 trillones de bytes de datos. Además, el 90% de los datos existentes a día de hoy se han creado en los últimos dos años.

En el entorno empresarial, los orígenes de datos tradicionales, como ERP, CRM o aplicaciones de RRHH, tienen unos requisitos de almacenamiento muy controlados y suelen estar acotados en máximos de crecimiento de unos pocos gigabytes diarios. Este es el límite de confort para un *data wararehouse* tradicional. Si tras incluir nuevos orígenes de datos, multiplicamos el volumen de información y sobrepasamos este límite de confort, el rendimiento del sistema podría verse gravemente afectado y, por tanto, habría que replantearse reestructurar el sistema de BI considerando un entorno de *big data*.

Enterprise resource planning

ERP (*enterprise resource planning*): sistemas informáticos de apoyo a la planificación de recursos empresariales. Típicamente manejan la producción, logística, distribución, inventario, envíos, facturas y contabilidad de la compañía de forma modular.

En la figura 3 podemos ver los volúmenes y la complejidad de datos generados por los orígenes de datos más comunes en una empresa. Podemos comprobar que la gran explosión de datos que da lugar al *big data* tiene que ver con:

Customer relationship management

CRM (*customer relationship management*): sistemas informáticos de apoyo a la gestión de relaciones con los clientes, ventas y marketing.

1) la aparición de nuevos orígenes de datos, como son las redes sociales, los vídeos o los sensores RFID;

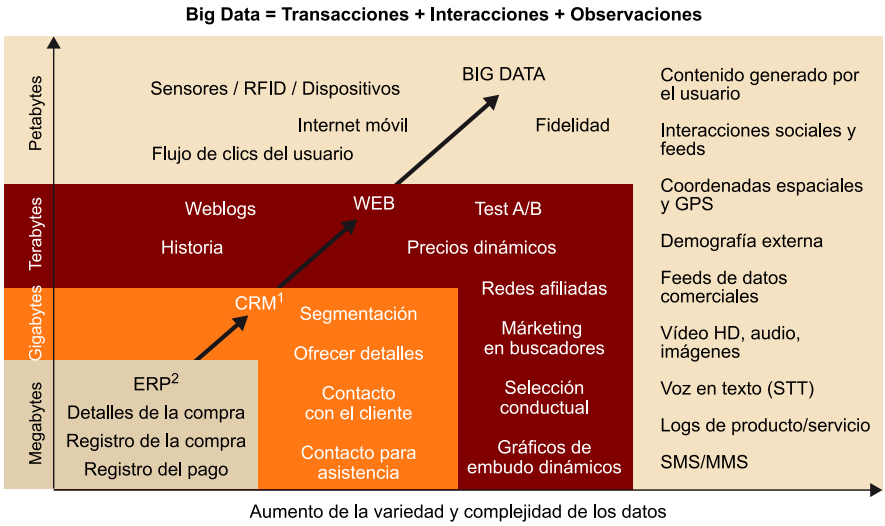
2) la aplicación de procesos analíticos que hasta ahora no se aplicaban de forma masiva, como por ejemplo analizar los textos de los mensajes de los clientes para estimar su sentimiento/opinión sobre la empresa, y

3) la recogida de información de datos que anteriormente eran desechados, como por ejemplo los deslizamientos del ratón en una página web o los recorridos de los coches recogidos por los sistemas de geoposicionamiento (GPS).

RFID

Siglas de *radio frequency identification*, en español identificación por radiofrecuencia. Esta tecnología permite, entre otras cosas, identificar, posicionar y trazar los movimientos de cualquier objeto marcado con una etiqueta RFID.

Figura 3. Incrementos de volumen por origen de datos



1 - Márketing relacional
2 - Sistema de planificación de recursos empresariales

Fuente: http://www.Hortonworks.com

Estos volúmenes adicionales podrían desbordar la capacidad de almacenamiento o de gestión de los *data warehouse* de la empresa. Por ejemplo, si comparásemos el volumen de datos de la información de un tique de compra, frente al volumen de los datos que obtendríamos si monitorizásemos todas las operaciones realizadas por un cajero de un supermercado (denominado comúnmente captura de instante y acción de un TPV), podríamos conocer, por ejemplo, errores frecuentes, velocidad de registro de cada producto, velocidad de conexión por tipo de tarjeta de crédito, etc. Pero, por otra parte, fácilmente estaríamos multiplicando por varios órdenes de magnitud el volumen de datos. Es decir, pasaríamos de volúmenes medios diarios de escala Megabyte a escala Gigabyte, por ejemplo.

3.2. Velocidad

En un entorno tan dinámico como el actual, muchas decisiones deben tomarse con gran rapidez. El tiempo que tardamos en reaccionar frente a un problema es un factor tan crítico como la decisión en sí misma, y actuar tarde puede ser catastrófico. Por ejemplo, un movimiento social contra una empresa, provocado por una noticia malinterpretada, puede afectar muy negativamente a la fidelización de sus clientes si no se detecta a tiempo y se actúa con celeridad. Redes sociales como Twitter pueden propagar rápidamente una información incorrecta, por lo que el volumen de datos se ve incrementado en función del tiempo que se deje pasar, por ejemplo.

Detectar e impedir que se realice un acceso indebido o un fraude, recomendar un determinado producto según la navegación de un cliente, averiguar si un cliente está a punto de darse de baja o descubrir que una pieza de un motor está cercana al final de su vida útil, son otros ejemplos que requieren decisiones que deben tomarse en el momento en el que se produce el detonante que las provoca; esto es, prácticamente en tiempo real. Por tanto, un objetivo del *big data* es tratar de proporcionar la información necesaria para la toma de decisiones en el menor tiempo posible.

Aunque se trata de un objetivo y una característica deseable en *big data*, técnicamente no siempre es posible trabajar en tiempo real, ni siquiera en frecuencias de actualización cercanas.

Los sistemas de BI

Los sistemas de BI en tiempo real suelen estar enmarcados en entornos muy controlados, donde muchas veces el sistema operacional, generador de los datos, está condicionado y diseñado para el acceso a dichos datos.

Existen ciertas barreras que se deben superar por *big data* para mejorar la velocidad frente a los sistemas tradicionales. Estas son:

1) Velocidad de carga. Antes de que un dato sea analizado debe pasar por distintos procesos que lo preparen, interpreten y lo integren con el resto de los datos. Estos procesos incluyen los procesos de extracción, transformación y carga (ETL), que permiten transformar, normalizar y cargar los datos al *data warehouse*, manteniendo, además, ciertos criterios de calidad de datos. Estos procesos son costosos en tiempo y en recursos hardware y software.

ETL

ETL (*extract*, *transform and load*, en castellano extraer, transformar y cargar) es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos, limpiarlos, normalizarlos, y cargarlos en otra base de datos, *data mart* o *data warehouse* para analizar y apoyar un proceso de negocio.

Por otra parte, garantizar la calidad del dato desde multitud de perspectivas posibles puede generar procesos innecesarios o redundantes según el tipo de análisis que realizar. Cuanto más queramos asegurar el grado de calidad, más costosos se vuelven.

Otro aspecto que cabe tener en cuenta es la necesidad de acelerar el acceso a los datos más frecuentes; normalmente, los datos más recientes se reclaman con más frecuencia y se exige una mayor rapidez de respuesta para obtener los informes que los contienen, frente, por ejemplo, a los datos de varios años de antigüedad. Para que esto sea posible, normalmente el administrador de la base de datos mueve los datos a las unidades más rápidas (o a memoria RAM) y genera estructuras para acelerar las operaciones más frecuentes sobre esos datos. Lógicamente, estos procesos de optimización de acceso a datos, al igual que ocurre con procesos ETL o de calidad, consumen tiempo y recursos en la creación del dato, lo que ralentiza la disponibilidad inicial del dato, eso sí, en aras de un acceso más rápido posteriormente.

Técnicas de optimización de acceso a datos

Las técnicas de optimización de acceso en bases de datos suelen estar basadas en generación y actualización de índices, recálculo de estadísticas de acceso, agregación y materialización de vistas.

Estos procesos de ETL, calidad y optimización de datos se llevan a cabo en sistemas con un gran volumen de datos, por lo que cualquier problema que pase inadvertido en sistemas pequeños (segundos de retardo) puede convertirse en minutos u horas en sistemas *big data*.

2) Velocidad de procesamiento. Al operar con los datos, las funciones de consulta permitidas en los sistemas gestores de bases de datos relacionales son básicamente la selección, la proyección, la combinación y la agregación; además, algunas bases de datos relacionales pueden incluir otras funciones básicas aritméticas y estadísticas.

Otros tipos de procesamiento, como la aplicación de funciones estadísticas avanzadas o técnicas de inteligencia artificial, suelen requerir implementaciones a medida, que implican:

- Consulta para la extracción del conjunto de los datos de interés.
- Almacenamiento intermedio de dichos datos.
- Aplicación de los cálculos sobre el conjunto de los datos extraído.
- Explotación y almacenamiento del resultado.

Al no ejecutarse de forma nativa en la base de datos, estas funciones no aprovechan al máximo los recursos del sistema, con lo que se genera una potencial pérdida de rendimiento. Además, es frecuente que se ejecuten en un servidor

distinto al de la base de datos, lo que repercute negativamente en la velocidad y puede producir sobrecarga en la red de datos, ya que los datos (recordemos que son masivos) deben moverse desde el servidor de bases de datos al servidor responsable de almacenar y ejecutar el software especializado para el cálculo.

En estos casos, puede ser recomendable el uso de un sistema de procesamiento distribuido tipo *MapReduce* o gestor de bases de datos de tipo NoSQL con un modelo de datos más adecuado a las necesidades concretas del sistema a implementar.

3.3. Variedad

La variedad se refiere a los diferentes formatos y estructuras en que se representan los datos. Definimos estructura de datos como la forma en que se encuentran organizados un conjunto de datos. Desde la perspectiva de BI, podemos clasificar los orígenes de datos según su nivel de estructuración en:

a) Orígenes de datos estructurados. La información viene representada por un conjunto de datos atómicos elementales[2] o agrupaciones de ellos. Se conoce de antemano la organización de los datos, la estructura y el tipo de cada dato elemental, su posición y las posibles relaciones entre los datos. Los datos estructurados son de fácil interpretación y manipulación.

(2)Un dato elemental es un dato de tipo simple, no compuesto de otras estructuras.

Ejemplos de orígenes de datos estructurados se encuentran en las bases de datos relacionales, en las aplicaciones operacionales (ERP, CRM, aplicaciones de RR. HH.), o en ficheros con una estructura fija en forma de tabla, como por ejemplo ficheros CSV u hojas de cálculo.

Fichero CSV

Fichero CSV (del *inglés Comma-Separated Values*) son un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, donde las columnas se separan por comas (o punto y coma cuando la coma es el separador decimal: España, Francia, Italia...) y las filas por saltos de línea.

b) Orígenes de datos no estructurados. Son aquellos donde la información no aparece representada por datos elementales, sino por una composición cohesionada de unidades estructurales de nivel superior. El valor informacional de estos orígenes de datos tiende a ser mayor que el de los estructurados, pero su interpretación y manipulación resulta mucho más compleja.

Ejemplos de orígenes de datos no estructurados son textos, audios, imágenes o vídeos.

c) Orígenes de datos semiestructurados. Los datos semiestructurados son aquellos que, tratándose de datos elementales, no tienen una estructura fija, aunque tienen algún tipo de estructura implícita o autodefinida; o aquellos en los que no todos los datos presentan una estructura elemental.

Un anuncio de empleo podría entenderse como dato semiestructurado, ya que puede tener algunos campos distintos en función del anuncio: una estructura ligeramente parecida en todos los anuncios y algún campo formado por un texto libre o fotografía (no estructurados).

Las bases de datos tradicionales tienen limitaciones para procesar datos no estructurados. Normalmente permiten almacenar textos, documentos y archivos multimedia, pero no proveen de funcionalidades adicionales para procesar su contenido convenientemente. Habitualmente se requieren aplicaciones

de terceros, o extensiones de la base de datos, para procesar o visualizar la información no estructurada almacenada en las bases de datos tradicionales. Por ejemplo, una dase de datos puede almacenar un documento PDF en formato binario. Consultando la base de datos se podría mostrar el conjunto de bits que lo compone, pero de manera ilegible. Es necesaria una tercera aplicación para que lo interprete y lo muestre de forma legible o permita buscar una palabra dentro del documento.

Algunas de las bases de datos relacionales más populares incorporan capacidades documentales que permiten un tratamiento de textos más eficiente. Por ejemplo, incorporan funciones de búsqueda en textos y documentos y otras funciones de gestión documental y multimedia, que permiten extraer metadatos de los ficheros almacenados, como por ejemplo el nombre de la canción y autor almacenados en un fichero en formato mp3. Aun así, su tratamiento es limitado y se produce de forma no nativa, con lo que ofrece una funcionalidad limitada y un rendimiento mejorable.

Por otro lado, los sistemas de BI tradicionales pueden trabajar con un gran número de orígenes de datos distintos, pero asumen siempre que estos están estructurados. Cuando el origen de datos no es estructurado, la solución más frecuente para su inclusión en un sistema BI es tratar de estructurar sus datos mediante un proceso de ETL.

Estructurar orígenes de datos no estructurados conlleva una pérdida de información, ya que solo se extraen y se almacenan las cuestiones que previamente han sido consideradas relevantes. Pongamos el ejemplo de una radiografía. Podríamos crear programas que descubran fracturas y manchas y establezcan su intensidad y posición. Los datos obtenidos serían datos estructurados y podrían almacenarse en un *data warehouse*. No obstante, si en un futuro se descubre, por ejemplo, que una determinada claridad es relativa a cierta enfermedad, esta característica no estaría contemplada en la base de datos si no se ha almacenado la radiografía original.

Estructurar orígenes de datos semiestructurados complica el proceso de carga, ya que obliga a añadir tantas excepciones como posibilidades de variación tengan los datos. Por ejemplo, a la hora de cargar un CSV, conocemos la posición y el tipo de dato de cada columna, por lo que –salvo errores– podríamos realizar una carga directa masiva a una base de datos relacional. Sin embargo, a la hora de cargar un fichero XML, primero habrá que interpretarlo. Esto se debe a que en XML el orden de los datos no es importante y que la misma información se puede representar de distintas formas. Obviamente, si alguno de los campos de datos fuera no estructurado, sufriríamos también pérdida de información en el proceso de estructuración de dicho campo.

Fichero XML

Fichero XML: fichero semiestructurado, compuesto por datos elementales pero de definición no previamente conocida, sino que incluye etiquetas para describir su propia definición.

3.4. Veracidad

La confianza en la veracidad de los datos es una característica que debe existir en cualquier sistema de apoyo a la toma de decisiones. Tomar decisiones a partir de datos erróneos puede tener consecuencias desastrosas. En *big data*, la gran cantidad de datos y orígenes de estos provoca que la veracidad del dato deba ser especialmente considerada y se deba aceptar cierto grado de incertidumbre. A continuación describiremos en qué consiste este grado tolerado de incertidumbre, que puede tener origen en la veracidad (o exactitud) del dato y en la fiabilidad de su procesamiento (exactitud del cálculo).

En un sistema de BI tradicional se presupone la veracidad de la información, lo que se llama **exactitud del dato**. Para satisfacer este requisito, una gran parte del trabajo, tanto de los desarrolladores como de los usuarios, es asegurar la calidad de los datos; para ello se emplean técnicas y procedimientos como por ejemplo técnicas de limpieza de datos, de enriquecimiento, de mapeo, de control de integridad, de gestión de datos maestros y de modelado de datos. Todo ello, antes de que los datos estén listos para el análisis.

Limpieza de datos

El *data cleansing, data scrubbing* o limpieza de datos, es el acto de descubrir, corregir o eliminar datos erróneos de una base de datos.

Muchos de los datos analizados mediante *big data* son intrínsecamente dudosos, relativos o con un cierto grado de error inherente. Ejemplo de ello son los datos procedentes de redes de sensores utilizados para realizar predicciones de las condiciones climáticas. Son datos en los que unas pocas mediciones se hacen extensibles a zonas y períodos más grandes. En contraste a los sistemas de BI tradicionales, en los repositorios de *big data* raramente se realizan los procesos de calidad de datos; o al menos no se realizan inicialmente, ya que podrían elevar el tiempo de carga o el coste en hardware a niveles inasumibles.

En los sistemas de ayuda a la toma de decisiones, los datos pueden ser generados a partir de procesos de modificación/análisis sobre los datos originales. En el caso de los BI tradicionales, estos cálculos se basan en la agregación sobre datos absolutos. La agregación de datos es un proceso determinista sin margen de interpretación. Si los datos son veraces, su agrupación también lo será. Es lo que se denomina la **exactitud del cálculo**.

No obstante, una parte muy importante de los cálculos en *big data* están basados en métodos analíticos que permiten cierto grado de incertidumbre. Es decir, aunque los datos originales se consideren veraces (los comentarios de usuarios de Facebook sobre una empresa), el resultado de su análisis puede no serlo (la información sobre la opinión de los usuarios sobre la empresa obtenida automáticamente de sus comentarios tiene una fiabilidad por debajo del 100%).

La minería de datos, el procesamiento del lenguaje natural, la inteligencia artificial o la propia estadística permiten calcular el grado de fiabilidad. Se trata de indicadores de la fiabilidad o exactitud de la predicción. Por ejemplo, el error muestral o error de estimación es el error causado al observar una mues-

tra en lugar de la población completa. No es lo mismo tomar una decisión a partir de una muestra del 50% de la población o a partir de solo el 0,5%, por lo que es un indicador que se debe tener en cuenta.

4. Escenario de adopción de *big data*

A continuación se muestra un escenario de ejemplo con el objetivo de describir una situación en la que las técnicas y tecnologías de *big data* pueden ser de ayuda. En este caso no podemos hablar de *big data* estrictamente como volumen de datos, pero sí por la variación, veracidad y velocidad de estos.

Suponed una empresa en la que se realizan análisis periódicamente con una hoja de cálculo que ocupa 1 GB. El PC con el que se opera la hoja de cálculo es capaz de procesarla, pero le lleva casi una hora abrirlo y recalcular los nuevos datos.

La empresa crece y se incorporan otras divisiones, lo que provoca un gran crecimiento del volumen de datos, y a final de año la hoja de cálculo llega a ocupar más de 100 GB. Por ese motivo, se decide incorporar un PC más potente (8 núcleos de 64 bits, 128 GB de memoria, etc.) solo para la ejecución de la hoja de cálculo. Pero nos volvemos a encontrar en la situación de partida, necesitamos una hora para cargar la hoja de cálculo.

A medida que avanza el tiempo, el volumen de datos sigue creciendo, 1.000 GB, 10 TB, etc. Por ello que se decide utilizar un sistema clásico de BI.

Conversión en bytes

Gigabyte = 10^9 = 1.000.000.000 bytes.

Terabyte = 10^{12} = 1.000.000.000.000 bytes.

Petabyte = 10^{15} = 1.000.000.000.000.000 bytes.

Exabyte = 10^{18} = 1.000.000.000.000.000.000 bytes.

Supongamos que el número de usuarios y sobre todo el volumen de datos sigue creciendo 100 TB, 1.000 TB, etc. Llega un punto en el que el sistema gestor de bases de datos empieza a tener problemas de rendimiento. También crecen las necesidades informacionales y analíticas. Ciertos procesos de segmentación y de identificación de patrones requieren varios días de ejecución.

Un día, la empresa se plantea realizar el lanzamiento de un nuevo producto. Para analizar la viabilidad de dicho lanzamiento distintos departamentos realizan las siguientes tareas orientadas a recoger datos sobre la potencial aceptación del nuevo producto:

- El *community manager* envía los mensajes oportunos en las redes sociales para captar la impresión de los internautas sobre el nuevo producto.

- El equipo de marketing realiza distintos tipos de encuesta y dirige distintos *focus group* para analizar el lanzamiento.

- Se utilizan tecnologías RFID en las tiendas de la empresa con el objetivo de trazar los movimientos de sus clientes.

- El administrador web analiza la actividad de los clientes en el portal de ventas y en la página de Facebook de la empresa.

Por tanto, la información con la que se cuenta para analizar el lanzamiento del nuevo producto es:

- Análisis tradicionales de ventas, comportamiento de compras, etc.

- Segmentación de clientes.

- Datos facilitados por institutos de estadística: demográficos, sociales, económicos, etc.

- Los gustos de fans en Facebook y del portal web.

- Opiniones vertidas por los clientes o potenciales clientes (recogidas de la web, de las redes sociales y de los *focus groups*).

- Los desplazamientos de clientes por las tiendas gracias a los *tags* RFID que llevan los productos.

- Resultados de encuestas y focus groups.

Este escenario presenta los siguientes problemas relacionados con las 4 V y que lo hace un buen candidato para aplicar técnicas de *big data*:

1) Volumen. Está lejos de superar los límites físicos de las bases de datos relacionales, aunque se sitúa en el límite aconsejado para el *data warehouse* de la empresa. Utilizar técnicas tradicionales de BI para realizar el análisis podría requerir un cambio de hardware, ya que al tamaño del actual *data warehouse* se deberían añadir los nuevos orígenes de datos para este análisis.

2) Velocidad. Encontramos procesos estadísticos que tardan varios días en ejecutarse. Por otro lado, la construcción de los procesos ETL puede ser muy compleja debido a la gran cantidad y variedad de datos. Además, los tiempos necesarios para ejecutar los procesos ETL pueden ser muy elevados debido a la heterogeneidad de los datos y de sus orígenes.

Focus groups

Técnica cualitativa de estudio de las opiniones o actitudes de un público utilizada en ciencias sociales y en estudios comerciales. Consiste en la reunión de un grupo de personas, entre 6 y 12 normalmente, con un moderador, investigador o analista, encargado de hacer preguntas y dirigir la discusión. Normalmente, el objetivo es evaluar el nivel de aceptación o identificar las características buscadas en un determinado producto o elemento publicitario.

Etiquetas RFID

Tags o etiquetas RFID son la forma de empaquetado más habitual de los dispositivos RFID. Son autoadhesivas y se caracterizan por su flexibilidad, su "delgadez", la capacidad de poder ser impresas con código humanamente legible en su cara frontal y las capacidades de memoria que dependerán del circuito integrado que lleve incorporado.

3) Variedad. Existen distintos orígenes de datos, algunos de ellos no estructurados o semiestructurados, como por ejemplo los comentarios en Facebook o los formularios de los *focus groups*, donde encontramos algunos campos de texto libre (los utilizados para recoger opiniones e impresiones, para proponer mejoras, etc.).

4) Veracidad. Existen datos provenientes de redes sociales (con faltas de ortografía, abreviaturas e interpretaciones ambiguas), de encuestas (donde puede que las respuestas sean anotadas en lugares equivocados o a veces los entrevistados no respondan o den respuestas incorrectas o inapropiadas) y de *focus groups* (que también pueden presentar cierto escepticismo, por tratarse de una pequeña muestra de la población y por las inferencias de experiencias previas o por la presencia de un líder muy marcado en el grupo). El hecho de tratar con estos datos provoca que el grado de incertidumbre sea elevado.

Las 4 V son los síntomas que indican la conveniencia de utilizar un sistema de *big data* para realizar un determinado análisis. El análisis de *big data* difiere ligeramente de los análisis tradicionales, debido a que se analizan todos los datos de las distintas fuentes de datos de manera integrada. Este tipo de análisis puede tener algunas implicaciones en sus resultados. En las siguientes líneas comentamos las posibles ventajas de usar un análisis de tipo *big data* para el ejemplo presentado.

En los sistemas tradicionales se tiende a realizar análisis distintos a partir de cada área (ventas, redes sociales, tiendas, clientes, investigación de mercados, etc.) debido a la dificultad de analizar la combinación de todos los datos de origen. Posteriormente las conclusiones obtenidas de los distintos análisis se combinan en una conclusión final. Uno de los principales problemas de este modo de trabajo es que los datos no se tratan en su conjunto, sino desde islas del conocimiento. Esto puede provocar una pérdida de información acerca de las relaciones que existen entre datos de distintas áreas, que pueden ser relevantes e incluso decisivas para el resultado final.

En contraposición, en *big data* se tiende a analizar los datos combinados de todas las fuentes de información. Al contar con los datos combinados de raíz, se minimiza la pérdida de información y se incrementan las posibilidades de encontrar nuevas correlaciones no previstas.

A continuación vamos a ver qué implicaciones podría tener realizar un análisis u otro en el caso que nos ocupa.

En el caso de utilizar un análisis tradicional, se analizarían los datos de los *focus groups* y de las redes sociales por separado y luego se integrarían sus conclusiones. Supongamos que al evaluar los *focus groups* la aceptación del producto ha sido mayoritariamente afirmativa y que según el análisis de las redes sociales el lanzamiento del producto despertó gran interés. Según estos datos, los ana-

listas podrían concluir que el lanzamiento debe realizarse tal y como se había propuesto, ya que el producto gustó (conclusión del análisis del *focus group*) y su lanzamiento despertó interés (conclusión del análisis de las redes sociales).

En el caso de utilizar un análisis más cercano al *big data* se analizarían los datos de los *focus groups* y de las redes sociales conjuntamente. Supongamos que a través de dicho análisis se descubren dos hechos:

1) Los grupos a los que menos ha gustado el producto (según los *focus groups*), coinciden con los grupos en los que más interés despertó el lanzamiento (según las redes sociales).

2) Los grupos a los que más les gustó el producto coinciden con los de menor penetración en redes sociales, por lo que ni siquiera se habrían tenido en cuenta en la combinación de análisis. Según dichos resultados, parece coherente pensar que los analistas no serían tan optimistas como en el caso anterior y solicitarían algunos cambios en la campaña de lanzamiento antes de sacar el nuevo producto al mercado, ya que parece que la campaña de lanzamiento no está enfocada al público potencial del producto.

Otro punto donde el uso de *big data* puede aportar ventajas en el proceso de análisis es en la búsqueda de información histórica. Ante cualquier nueva acción comercial, es habitual revisar la historia de la empresa para encontrar precedentes similares. Establecer estas similitudes permite tener un punto de referencia contra el que medirse y aplicar las lecciones aprendidas en experiencias pasadas.

Condicionado por el tipo de análisis realizado tradicionalmente, las bases de información de las empresas tienden a recoger datos de los resultados de los análisis realizados en el pasado, pero no de los datos origen utilizados en los análisis. Eso dificulta la búsqueda de situaciones pasadas similares a la actual, ya que solo se podrá comparar análisis contra análisis o, en el caso más extremo, conclusiones contra conclusiones. Elegir experiencias previas que no se ajusten lo suficiente podría reforzar una idea de éxito poco realista, forzando, por ejemplo, un lanzamiento genérico a una escala desmedida.

En *big data* es más natural almacenar todos los datos de origen. Eso facilita que se puedan realizar búsquedas de precedentes similares en el nivel datos, y no solo en el nivel de análisis o conclusiones. En el caso de ejemplo, esto permitiría comparar los datos del lanzamiento actual con los datos de los anteriores lanzamientos. El hecho de trabajar con los datos de origen (no estructurados) permite una comparación más realista y contextualizada.

Resumen

Los cambios acaecidos en los últimos años nos han llevado a un contexto en el que los datos son más masivos que nunca, tanto en volumen como en el tipo de información que recogen y la velocidad a la que se producen. Este gran volumen de datos requiere nuevas tecnologías y nuevas filosofías para almacenar y procesar los datos. Estas nuevas tecnologías permiten recoger y procesar grandes cantidades de datos y, con ellos, realizar análisis más precisos y generalizables. Además, el gran volumen de datos está potenciando el uso la correlación entre datos en contra de construir modelos para intentar explicar la causalidad.

Big data se define como el conjunto de técnicas y tecnologías para el tratamiento de datos, en entornos de gran volumen, variedad de orígenes, y en los que la velocidad de respuesta es crítica. Además, en los sistemas de BI también hay que tener en cuenta otro factor: la veracidad de los datos. Si tomamos una decisión en función de los datos, hay que conocer el margen de error (o exactitud) de estos, ya sea por su origen o por los procesos utilizados para generarlos. Dicho esto, un problema susceptible de ser atacado mediante una aproximación de *big data* es un problema que satisface las cuatro características descritas (volumen, variabilidad, velocidad y veracidad).