Introducción al *big data*

PID_00250684

Jordi Casas Roma

Tiempo mínimo de dedicación recomendado: 2 horas

Universitat Oberta de Catalunya

Índice

Introducción

Iniciaremos este módulo con una introducción al concepto de *big data*, centrándonos en el cambio de paradigma que supone la llegada de los datos masivos.

A continuación, veremos una de las primeras definiciones de *big data*, relacionada con las magnitudes del dato. A partir de esta primera definición han surgido muchas más que, en general, amplían la definición original. A partir de esta definición inicial, presentaremos la definición utilizada en este texto y veremos algunos estándares importantes en relación con los datos y la interconexión de sistemas.

Finalmente, presentaremos un pequeño ejemplo que servirá para ilustrar un escenario de uso de tecnologías *big data*, en este caso concreto, en una *smart city*.

Objetivos

En los materiales didácticos de este módulo encontraremos las herramientas indispensables para alcanzar los siguientes objetivos:

1. Conocer los antecedentes que han llevado a la aparición del *big data*.
2. Comprender el cambio de paradigma asociado al *big data*.
3. Conocer los factores que pueden hacer que un problema analítico pueda ser resuelto empleando metodologías y herramientas de *big data*.

1. Antecedentes y contextualización

El término *big data* —terminología anglosajona ampliamente utilizada que se suele traducir por *datos masivos*— apareció a principios del siglo XXI en el entorno de la ciencias, en particular, de la astronomía y de la genética, debido a que ambos campos experimentaron una gran explosión en la disponibilidad de datos. Por ejemplo, en el campo de la astronomía el proyecto de exploración digital del espacio llamado Sloan Digital Sky Survey 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. En el campo de la genética, un ejemplo relevante sería el proyecto del genoma humano. Este proyecto tiene como objetivo encontrar, secuenciar y elaborar mapas genéticos y físicos de gran resolución del ADN humano, y genera una cantidad de datos del orden de 100 gigabytes por persona.

El proyecto Sloan Digital Sky Survey

El proyecto Sloan Digital Sky Survey tiene como objetivo identificar y documentar los objetos observados en el espacio. Podéis acceder a su página web desde el siguiente enlace: http://www.sdss.org

En estos últimos años la explosión de datos se ha generalizado en muchos de los campos que rodean nuestra vida cotidiana. Entre otros, el incremento del número de dispositivos con conexión a Internet, el auge de las redes sociales y el Internet de la cosas (IoT) han provocado una explosión en el volumen de datos disponibles. Además de la gran cantidad de datos, es importante destacar que muchos de ellos son abiertos y accesibles, lo que permite que puedan ser explotados por usuarios o instituciones de cualquier parte del mundo.

IoT

Internet de las cosas o IoT (*Internet of Things*, en inglés) es un concepto que se refiere a la interconexión digital de objetos cotidianos con Internet.

Pero el mero hecho de disponer de una gran cantidad de datos no aporta valor. El verdadero valor de los datos está en su análisis e interpretación.

La aparición de nuevas técnicas y tecnologías de procesamiento de datos surgió a causa de la imposibilidad de procesar la enorme cantidad de datos que se generaban con las técnicas tradicionales. Aunque la mejora y el abaratamiento del *hardware* de los ordenadores permite con las técnicas tradicionales cargar y procesar más datos, el aumento en la cantidad de datos es de varios órdenes de magnitud superiores. Por lo tanto, aunque podamos adquirir más y mejor *hardware*, esto es absolutamente insuficiente para afrontar el aumento masivo de datos. Por ejemplo, imaginemos en cómo debería ser el ordenador de Google para indexar todos los contenidos de la web. Por consiguiente, también fue necesaria la evolución de la tecnología basada en el *software*.

Las grandes empresas de Internet, como Google, Amazon y Yahoo!, se encontraron con varios problemas importantes para continuar desempeñando sus tareas cotidianas. En primer lugar, la gran cantidad de datos que estaban acumulando hacía inviable su procesamiento en un único ordenador. Por tanto, se debía usar procesamiento distribuido para involucrar distintos ordenadores que trabajasen con los datos de manera paralela y así poder procesar más datos en menos tiempo. En segundo lugar, la heterogeneidad de los datos requirió nuevos modelos de datos para facilitar la inserción, la consulta y el procesamiento de datos de cualquier tipo y estructura. Y en tercer lugar, los datos debían de procesarse de forma rápida, aunque hubiera muchos datos para procesar. Por ejemplo, un buscador web no sería útil si devolviera los resultados de nuestra búsqueda después de un tiempo demasiado largo, por ejemplo tras una espera de más de una hora. En consecuencia, estas grandes empresas que gestionaban grandes volúmenes de datos se dieron cuenta de que las técnicas de procesamiento 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.

Muchas de las técnicas que se desarrollaron para dar respuesta a los problemas planteados por los datos masivos utilizan un planteamiento basado en el procesamiento paralelo. Este paradigma se basa en dos pasos principales:

1) En primer lugar, se divide el problema en subproblemas de menor tamaño y complejidad. De esta forma, se pueden distribuir los distintos subproblemas en distintas computadoras de forma que cada una se encargue de un subproblema de forma independiente.

2) En segundo lugar, la solución final del problema se compone a partir de las soluciones parciales de los subproblemas. Una vez cada subproblema es resuelto de forma independiente, se ensamblan todas las pequeñas soluciones resultantes para crear la solución global del problema inicial.

2. El nuevo paradigma de *big data*

Hasta ahora, si un agente o institución deseaba evaluar un fenómeno no podía, generalmente, recoger todos los datos relacionados con él. El motivo era que los métodos de recogida y procesamiento de datos eran muy costosos en tiempo y en dinero. En estos casos, se escogía 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. Este es el paradigma de la **causalidad**, donde se intenta establecer una relación de causa-efecto entre el fenómeno que se analiza y los datos relacionados con él.

Lectura complementaria

Viktor Mayer-Schönberger; Kenneth Cukier (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. *John Murray*

Hoy en día, en cambio, la recogida de datos masivos ha permitido obtener información sobre la muestra completa (o casi) de datos relacionada con el fenómeno que hay que evaluar, es decir, en toda la población. Por ejemplo, si una institución desea analizar los tuits que tratan sobre un tema de interés público, es perfectamente factible que pueda recoger todos los tuits que hablan del tema y analizarlos. En este caso, el análisis no pretende confirmar o invalidar una hipótesis, sino establecer **correlaciones** entre distintas variables de la muestra. Por ejemplo, supongamos que existe una fuerte correlación entre el lugar de residencia de los vecinos de una ciudad y su opinión ante una determinada problemática de la ciudad. En este caso, podemos explotar la relación que existe entre ambas variables aunque no sepamos la causa que induce de la una a la otra.

Tuits

Los tuits son pequeños mensajes de texto, limitados a 280 caracteres (originalmente 140), que permiten a los usuarios de Twitter expresar su estado actual, comunicar noticias o mantener pequeñas conversaciones.

Los datos masivos imponen un nuevo paradigma donde la correlación «sustituye» a la causalidad. Determinar la causalidad de un fenómeno pierde importancia y, en contraposición, «descubrir» las correlaciones entre las variables se convierte en uno de los objetivos principales del análisis.

Este cambio de paradigma provoca que los sistemas de *big data* se centren en encontrar «qué» aspectos están relacionados entre sí y no en «por qué» están relacionados. Estos sistemas pretenden responder cuestiones del tipo: ¿qué pasó?, ¿qué está pasando? y ¿qué pasaría si...?, pero desde un punto de vista basado en las correlaciones, donde no se busca la explicación del fenómeno, sino solo el descubrimiento del fenómeno en sí. En consecuencia, la causalidad pierde terreno a favor de la asociación entre hechos.

2.1. Primera definición de *big data*

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.

* http://gtnr.it/1bKflKH

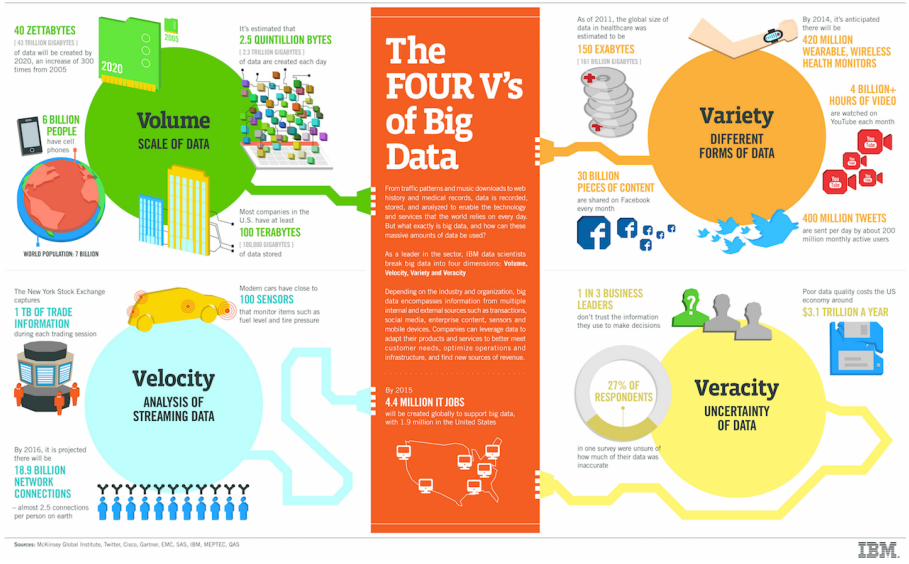
Lectura complementaria

Doug Laney (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*

Esta definición 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.

La figura 1 muestra la interacción de las 4 V de *big data* según IBM: existen grandes volúmenes de datos (*volume*), procedentes de una gran variedad de fuentes (*variety*), de un cierto grado de incertidumbre (*veracity*) y que puede ser necesario procesar para obtener rápidas respuestas (*velocity*).

Figura 1. Interacción de las 4 V de *big data* según IBM



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

A continuación veremos con más detalle las 4 V de la definición de *big data*.

2.1.1. Volumen

Bases de datos relacionales

Las bases de datos relacionales almacenan los datos en tablas y permiten las interconexiones o relaciones entre los datos de distintas tablas. El lenguaje utilizado para consultas y mantenimiento se llama *Structured Query Language* (SQL).

Se estima que el volumen de datos existente en la actualidad está por encima del zettabyte y que crecerá de forma exponencial en el futuro.

Los almacenes de datos tradicionales, basados en bases de datos relacionales, tienen unos requisitos de almacenamiento muy controlados y suelen estar acotados en máximos de crecimiento de unos pocos gigabytes diarios. Si 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 almacenamiento considerando un entorno de *big data*.

Zettabyte (ZB) = 10^{21} bytes = 100000000000000000000 bytes.

2.1.2. Velocidad

En un entorno dinámico, el tiempo que se tarda en obtener la información o el conocimiento frente a determinados sucesos es un factor tan crítico como la información en sí misma. En algunos casos, la información extraída de los datos es útil mientas los datos son «frescos», pero pierde valor cuando los datos dejan de reflejar la realidad. La gestión del tráfico es un ejemplo que requiere decisiones que deben tomarse en un espacio de tiempo breve; esto es, prácticamente en tiempo real. Por tanto, en algunos casos el *big data* debe tratar de proporcionar la información necesaria 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.

Existen dos tipos de velocidades que juntas condicionarán la velocidad final de respuesta ante nuevos datos. Estas son:

- **Velocidad de carga**. Los nuevos datos deben ser preparados, interpretados e integrados con el resto de datos antes de poder ser procesados. Estos procesos incluyen los procesos de extracción, transformación y carga (ETL), que son costosos en tiempo y en recursos *hardware* y *software*.

- **Velocidad de procesamiento**. Una vez los nuevos datos están integrados y listos para su análisis, debemos considerar otros tipos de procesamiento, como la aplicación de funciones estadísticas avanzadas o técnicas de inteligencia artificial. Este procesamiento suele implicar consultas para la extracción del conjunto de los datos de interés, almacenamiento intermedio de dichos datos o aplicación de los cálculos sobre el conjunto de los datos extraídos, por ejemplo.

ETL

Extraer, transformar y cargar (en inglés, *Extract, Transform and Load* (ETL)) es el proceso que permite mover datos desde múltiples fuentes, reformatearlos, limpiarlos, normalizarlos y almacenarlos en el sistema utilizado para su análisis.

2.1.3. Variedad

La estructura de datos se define como la forma en que se encuentran organizados un conjunto de datos. La variedad se refiere a los diferentes formatos y estructuras en que se representan los datos. Según su nivel de estructuración, podemos clasificar los orígenes de datos de la siguiente manera:

- **Orígenes de datos estructurados**. La información viene representada por un conjunto o agrupación de datos atómicos elementales, es decir, datos simples que no están compuestos de otras estructuras. Se conoce de antemano la organización de los datos, la estructura y el tipo de cada dato

Fichero CSV

Un fichero CSV (del inglés *Comma-Separated Values*) es un tipo de documento en formato abierto que permite representar datos en forma de tabla, donde las columnas se separan por comas y las filas por saltos de línea.

elemental, su posición y las posibles relaciones entre los datos. Los datos estructurados son de fácil interpretación y manipulación.

Los ficheros con una estructura fija en forma de tabla, como los ficheros CSV o las hojas de cálculo, son claros ejemplos de orígenes de datos estructurados.

- **Orígenes de datos semiestructurados**. La información viene representada por un conjunto de datos elementales, pero a diferencia de los datos estructurados no tienen una estructura fija, aunque tienen algún tipo de estructura implícita o autodefinida.

  Ejemplos de este tipo de datos son los documentos XML o las páginas web. En ambos casos los documentos siguen ciertas pautas comunes, pero sin llegar a un nivel de estructuración fija.

Fichero XML

Un fichero XML (del inglés *eXtensible Markup Language*) es un tipo de documento semiestructurado, compuesto por datos elementales, pero de definición no previamente conocida, y que incluye etiquetas para describir su propia definición.

- **Orígenes de datos no estructurados**. La información no aparece representada por datos elementales, sino por una composición cohesionada de unidades estructurales de nivel superior. La interpretación y manipulación de estos orígenes de datos resulta mucho más compleja que el de los estructurados o semiestructurados.

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

2.1.4. Veracidad

La gran cantidad de datos y sus orígenes en *big data* provoca que la veracidad del dato deba ser especialmente considerada y se deba aceptar cierto grado de incertidumbre. Este grado tolerado de incertidumbre puede tener origen en la exactitud del dato y en la fiabilidad de su procesamiento (exactitud del cálculo):

- **Exactitud del dato**. Muchos de los datos analizados mediante *big data* son intrínsecamente dudosos, relativos o con un cierto grado de error inherente. Por ejemplo, los datos procedentes de redes de sensores utilizados para medir la temperatura ambiental pueden incluir cierto grado de incertidumbre, dado que generalmente unas pocas mediciones se hacen extensibles a zonas y períodos más grandes.

- **Exactitud del cálculo**. Una parte muy importante de los cálculos en *big data* están basados en métodos analíticos que permiten cierto grado de incertidumbre. 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, que puede ser inferior al 100 % aunque los datos originales se consideren veraces.

Por ejemplo, aunque los comentarios de los usuarios de Facebook sobre una empresa son veraces, el resultado de su análisis mediante técnicas automáticas de procesamiento del lenguaje natural puede tener una fiabilidad por debajo del 100 %.

2.2. Nuestra definición de *big data*

A partir de la definición anterior, han aparecido otras definiciones alternativas que iban añadiendo, progresivamente, más V a las definiciones anteriores. Conceptos como por ejemplo la variabilidad, la validez o la volatilidad se han incorporado en esta definición de *big data* según la propuesta de algunos autores.

El hecho de que existan múltiples definiciones complica la comprensión e identificación de escenarios. La gran mayoría de ellas incluyen lo que se conoce como las 3 V del *big data*, que hemos comentado anteriormente y que son magnitudes físicas del dato.

Por tanto, en aras de tener un enfoque pragmático, en este texto vamos a usar la siguiente definición de *big data*.

Entendemos por ***big data*** el conjunto de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización de conjuntos de datos complejos.

Y entenderemos por conjuntos de datos complejos aquellos que dado su volumen, velocidad o variedad no pueden ser tratados de forma eficiente en un sistema tradicional de análisis de datos.

2.3. Clasificación de NIST

De acuerdo con el NIST*, y en particular dentro de su grupo de trabajo de *big data*, existen tres tipologías de escenarios que requieren el uso de *big data*. Los tipos disponibles se resumen a continuación:

* **https://www.nist.gov/**

NIST

NIST es el acrónimo de National Institute of Standards and Technology, institución americana que estudia, define y promueve estándares tecnológicos.

- **Tipo 1**, donde una estructura de datos no relacional es necesaria para el análisis de datos.
- **Tipo 2**, donde es necesario aplicar estrategias de escalabilidad horizontal para procesar y analizar de manera eficiente los datos.
- **Tipo 3**, donde es necesario procesar una estructura de datos no relacional mediante estrategias de escalabilidad horizontal para procesar y analizar de manera eficiente los datos.

Por tanto, para una determinada necesidad analítica, es posible identificar si estamos en un escenario de *big data* o no, y si es necesario este tipo de tecnologías, hecho que cada vez más se erige como un punto relevante y de partida para la implementación de este tipo de proyectos.

2.4. Estándares en *big data*

A medida que el *big data* ha adquirido mayor importancia para las organizaciones y estas se han empezado a preocupar por cómo llevar a cabo un proyecto de este tipo, ha quedado patente que se necesita interconectar múltiples sistemas y tecnologías. Esta integración e interoperabilidad de sistemas requiere estándares de mercado.

Por ejemplo, dentro del contexto de la inteligencia de negocio y la analítica ya existen estándares como UIMA* (*Unstructured Information Management Architecture*), OWL** (*Web Ontology Language*), PMML*** (*Predictive Model Markup Language*), RIF**** (*Rule Interchange Format*) y XBRL***** (*eXtensible Business Reporting Language*), que permiten la interoperabilidad de analítica de datos en información no estructurada, ontologías de modelos de datos, modelos predictivos, el intercambio de datos entre organizaciones y reglas e informes financieros, respectivamente.

*** https://uima.apache.org/**
**** http://bit.ly/2jdglBH**
***** http://dmg.org/**
****** http://bit.ly/2Bp6dO8**
******* https://www.xbrl.org/**

Desde 2012, varios grupos de trabajo de la comunidad internacional han empezado a trabajar en la creación de estándares, como por ejemplo NIST, TM-Forum, Cloud Security Alliance, ITU, Open Data Platform initiative (ODPi) y Common Criteria Portal.

En el caso de ODPi, su búsqueda de estándares se fundamenta en proponer una configuración mínima de Apache Hadoop que según su criterio incluye solo cuatro componentes: HDFS, YARN, MapReduce y Ambari.

La gran mayoría de estos grupos soporta la adopción efectiva de tecnología *big data* a través del consenso en definiciones, taxonomías, arquitecturas de referencia, casos de uso y *roadmap* tecnológicos.

Taxonomía

Cuando hablamos de taxonomía hacemos referencia a una clasificación u ordenación en grupos de cosas que tienen unas características comunes.

3. Ejemplo de escenario *big data*

A continuación se muestra el contexto de una *smart city* como una situación en la que las técnicas y tecnologías de *big data* pueden ser necesarias para un correcto procesamiento y análisis de los datos.

Supongamos que la ciudad en cuestión recoge los siguientes datos:

- Las cámaras de tráfico recogen imágenes de forma continua, ya sea en formato de vídeo o también identificando los vehículos que circulan en cada vía a través de su matrícula.

- Los sensores de las zonas de aparcamiento exteriores proporcionan información continua sobre su ocupación, indicando en cada momento si cada una de las plazas de la ciudad está vacía u ocupada.

- Una gran cantidad de sensores repartidos por toda la ciudad analizan la calidad del aire en periodos cortos de tiempo, produciendo un análisis continuo de las distintas zonas de la ciudad. En cada análisis se incluyen muchos factores relacionados con la contaminación y los agentes tóxicos del aire.

- Los puntos de transporte urbano basado en el uso compartido de bicicletas informa en cada momento sobre su disponibilidad en cada punto de recogida y devolución de la ciudad.

Pero además, el ayuntamiento ha decidido complementar la información que recoge mediante los distintos sensores de la ciudad con información obtenida a través de Internet y de las redes sociales. Entre otros, el ayuntamiento se plantea:

- Monitorizar las acciones de los usuarios que visitan alguna de las páginas web municipales, registrando información como por ejemplo las páginas accedidas, el dispositivo con el que se accede o la ubicación, cuando está disponible. Adicionalmente, también se quiere analizar y registrar los comentarios de los usuarios en los foros municipales, donde se discute cualquier tema de interés para la ciudad, y de las encuestas en línea que el ayuntamiento realiza periódicamente.

- Recopilar información de la red social Twitter referente al estado del tráfico en cualquier momento y punto de la ciudad. Para ello, obtienen y almace-

nan todos los tuits con información referente a alguna de las principales calles, vías o paseos de la ciudad.

- Almacenar la información de las interacciones de los usuarios en la página municipal de Facebook, como por ejemplo el número de «Me gusta» o los comentarios de los usuarios.

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. El volumen de datos generado diariamente en una gran ciudad puede superar los límites físicos de las bases de datos y herramientas de análisis tradicionales.

2) Velocidad. Los análisis de datos relacionados con el tránsito deben tener respuestas rápidas que permitan detectar y corregir problemas, en la medida de lo posible, de forma casi inmediata. Una respuesta tardía a los problemas de tráfico es sinónimo de no reaccionar.

3) Variedad. Existen distintos orígenes de datos, algunos de ellos no estructurados o semiestructurados, como por ejemplo los comentarios en Facebook o los foros municipales, donde encontramos algunos campos de texto libre para recoger opiniones e impresiones.

4) Veracidad. Existen datos provenientes de redes sociales, de encuestas en línea y de foros que pueden contener faltas de ortografía, abreviaturas e interpretaciones ambiguas. 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. 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.

Resumen

En este módulo didáctico hemos presentado los antecedentes históricos que han llevado a la aparición del *big data*. A partir de los problemas para la gestión de conjuntos de datos complejos aparece la necesidad de almacenar y procesar este tipo de información de manera más eficiente.

Es en este punto donde algunas empresas e instituciones de todo el mundo empiezan a trabajar en soluciones que permitan dar respuesta a análisis de grandes conjuntos de datos, análisis en tiempo real (o casi) y análisis de conjuntos de datos semiestructurados o no estructurados.

Estas iniciativas son las que han acabado por definir lo que hoy en día conocemos como *big data* y que, aunque no tengamos una definición única y consensuada, sí que intuitivamente hay una definición colectiva más o menos general o aceptada, aunque con algunos matices notables.

Después de presentar la definición inicial de *big data* y sus implicaciones, hemos presentado la definición que utilizaremos en este texto, así como algunas indicaciones sobre los estándares existentes y una posible clasificación de problemas que requieren el uso de *big data*.

Glosario

bases de datos relacionales *f pl* Son las bases de datos que almacenan los datos en tablas y permiten las interconexiones o relaciones entre los datos de distintas tablas. El lenguaje utilizado para consultas y mantenimiento se llama SQL (*Structured Query Language*).

datos estructurados *m pl* Son aquellos que siguen un patrón igual para todos los elementos y que además es conocido *a priori*. Por ejemplo, los datos de una hoja de cálculo presentan los mismos atributos para cada fila.

datos semiestructurados *m pl* Son una forma de datos que no contiene una estructura fija predefinida *a priori*, pero que contiene etiquetas u otros marcadores para separar los elementos semánticos y hacer cumplir jerarquías de registros y campos de los datos. Por ejemplo, las documentos JSON o HTML.

datos no estructurados *m pl* Son aquellos que no siguen ningún tipo de patrón conocido *a priori*. Por ejemplo, dos documentos de texto o imágenes.

Extraer, Transformar y Cargar *v* Es el proceso que permite mover datos desde múltiples fuentes, reformatearlos, limpiarlos, normalizarlos y almacenarlos en el sistema utilizado para su análisis.
en Extract, Transform and Load
sigla **ETL**

fichero CSV *m* Es un tipo de documento en formato abierto que permite representar datos en forma de tabla, donde las columnas se separan por comas y las filas por saltos de línea.
en Del inglés, Comma-Separated Values

fichero XML Es un tipo de documento semiestructurado, compuesto por datos elementales pero de definición no previamente conocida, que incluye etiquetas para describir su propia definición.
en Del inglés, eXtensible Markup Language

Internet de las cosas *m* Es un concepto que se refiere a la interconexión digital de objetos cotidianos con Internet. *en* Internet of Things sigla **IoT**

taxonomía *f* Clasificación u ordenación en grupos de cosas que tienen unas características comunes.

tuit *m* Pequeño mensaje de texto, limitado a 280 caracteres (originalmente 140), que permite a los usuarios de Twitter expresar su estado actual, comunicar noticias o mantener pequeñas conversaciones.

Bibliografía

Mayer-Schönberge, Viktor; Cukier, Kenneth (2013). *Big data. La revolución de los datos masivos*. Madrid: Turner Publicaciones.

White, Tom (2015). *Hadoop: The Definitive Guide, 4th Edition*. O'Reilly Media.