Introducción al *big data*

PID_00253881

Xavier Font

Tiempo mínimo de dedicación recomendado: 2 horas

Universitat Oberta de Catalunya

Xavier Font

Doctor en Informática por el Departamento de Arquitectura de Computadores de la UPC, Diploma de Estudios Avanzados por el Departamento de Estadística e Investigación Operativa de la UPC, ingeniero de Organización Industrial por la ETSEIB - UPC y licenciado en Informática por la FIB - UPC. Forma parte del Grupo de Procesado de Señal del ESUP Tecnocampus (grupo de investigación consolidado GRC-1318 y GRC-533). Actualmente es el profesor responsable de la asignatura de Estadística de los Estudios de Ingeniería Informática, Mecánica y Electrónica. Mantiene colaboraciones con el Grupo de Comunicaciones y Procesado de Señal del Imperial College de Londres, donde ha realizado diversas estancias de investigación.

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. La era del *big data* 7

2. Proyecto *big data* 9
2.1. Volumen 9
2.2. Velocidad 10
2.3. Variedad 10
2.4. Veracidad 10
2.5. Definición de *big data* 11

3. Plataformas de procesado *big data* 12
3.1. Apache Hadoop 12
3.2. Cloudera Enterprise Data Hub 13

4. *Big Data Analytics* 14

5. El *big data* en la fábrica inteligente 16

Introducción

La introducción de este tema dentro de esta asignatura pretende, simplemente, poner énfasis en la importancia también del *big data*, como no, en el entorno industrial. Sin embargo, se tratará mucho más ampliamente en la siguiente asignatura. Para ello se introducirá el concepto de *big data* y su importancia en el ámbito general, con lo cual el estudiante podrá inferir fácilmente su implicación en la fábrica inteligente.

El crecimiento explosivo de la potencia y capacidad de procesamiento de los ordenadores, así como el incremento permanente de dispositivos y sensores conectados entre sí, permite entender la importancia del *big data* y su relación-solapamiento con los entornos IdC (Internet de las cosas).

El número de dispositivos conectados a Internet ha excedido el número de individuos que habitan el planeta Tierra. Es fácil imaginar una lista de dispositivos que tenemos a nuestro alrededor: ordenadores, portátiles, tabletas, teléfonos móviles, relojes y pulseras (*wearable devices*), que comparten y generan datos hacia Internet. Se estima que la cifra de dispositivos conectados alcanzará los 50.000 millones hacia el año 2020.

El objetivo de este módulo es introducir el concepto de *big data* en el entorno industrial de la Industria 4.0. Este módulo presentará de una forma práctica las características y aplicaciones dentro del ámbito *big data*.

1. La era del *big data*

Imaginemos nuestra calle no únicamente como el lugar donde residimos sino como un sistema. ¿Quién genera información relativa a nuestra calle? Hay una lista considerable de agentes externos que recogen datos de lo que pasa en nuestra calle: un grupo de datos son los que nos generan dolor de cabeza (por el pago de servicios que hemos contratado):

- Uso del teléfono, móvil e Internet
- Hábitos de consumo eléctrico
- Utilización de gas
- Pago de impuestos municipales (catastro)
- Consumo de agua

Otro grupo de datos que en ocasiones desconocemos pero que también inciden en lo que pasa en nuestra calle (sistema) son:

- Tráfico
- Información meteorológica: temperatura, humedad, presión atmosférica
- Información de calidad del aire: estaciones terrestres (CO2, PM10, entre otros).
- Datos provenientes de satélites de observación de la Tierra
- Estadísticas oficiales: Eurostat, INE, IDESCAT, etc.
- Hábitos de consumo (tarjetas de crédito)
- Interacción en redes sociales
- Otros

Todo este conjunto masivo de datos, que tienen como elemento común nuestra calle, configura un escenario típico de *big data*. En este escenario podemos desgranar algunas características interesantes y que pueden suponer diferentes grados de complejidad.

- Primero: gran cantidad de datos que se generan de forma permanente y por unidad de tiempo (cada segundo 1 Tb).
- Segundo: gran variedad de formatos y tipos de datos. Podemos tener temperatura en grados centígrados, imágenes provenientes del Meteosat, texto escrito en un foro o WhatsApp o enlaces URL.
- Tercero: los datos se generan de forma instantánea y podemos pensar que estamos en un entorno de tiempo real.

En la actualidad, las tendencias hacia una revolución digital en los procesos de manufactura (conocidas como Industria 4.0) se sustentan en diferentes tecnologías: IdC, CPS, *cloud computing* y BDA.

Brevemente, el concepto IdC se refiere a un mundo interconectado con el propósito de recopilar e intercambiar datos. En general, IdC es capaz de ofrecer conectividad avanzada de sistemas físicos, sistemas y servicios.

El acrónimo CPS se refiere a sistemas ciberfísicos (en inglés, *Cyber-Physical System*), que son un mecanismo a través del cual objetos físicos y *software* están íntimamente imbricados, facilitando de esta forma el intercambio de información.

El concepto de *cloud computing* corresponde al uso de servicios computacionales y escalables sobre internet.

El último acrónimo que se especifica es Big Data Analytics (BDA), que intenta resolver de qué forma se pueden procesar los datos con la intención y el propósito de dar la información correcta, para el objetivo correcto, en el momento adecuado.

2. Proyecto *big data*

En los últimos años la potencia de los ordenadores ha crecido exponencialmente. Este fuerte incremento en el rendimiento de procesamiento ha ido en paralelo a la disponibilidad, gestión y almacenaje de grandes cantidades de datos. Todos estos elementos, conjuntamente con en el desarrollo y avance de nuevas técnicas de *machine learning* para la identificación de patrones, han incidido en la importancia de analizar correctamente estos datos para convertirlos en conocimiento o ventaja competitiva para las organizaciones.

Si bien el nombre de *big data* fue inicialmente empleado en un artículo publicado en 2001 y titulado "*3D data management: controlling data volume, velocity and variety*" por Doug Laney, la definición formal ha tenido diversas variaciones y dependiendo del contexto puede también cambiar. El concepto de volumen de datos se asocia a gran volumen de datos, y en este punto podemos preguntarnos: ¿qué magnitud representa? Gb, Tb, Pb, Eb, Zb o Yb[1] ? Para empresas como Google, Facebook, Twitter u organizaciones como la NSA su unidad de medida será la de Zb o Yb. Para empresas medianas quizás estas magnitudes son desconocidas, pero sí pueden tratar problemas de *big data,* que en su caso su unidad de medida estará alrededor de los Tb.

(1)Gigabyte (Gb = 10^9), Terabyte (Tb = 10^{12}), Petabyte (Pb= 10^{15}), Exabyte (Eb= 10^{18}), Zettabyte (Zb= 10^{21}) o Yottabyte (Yb= 10^{24}).

En el artículo publicado por D. Laney se introduce el concepto de las tres V como descripción de los entornos *big data*. Estas tres V aparecen en el título del citado artículo (Volume, Velocity and Variety). Esta presentación inicial de las tres V ha sido extendida y ampliada por otros investigadores (figura 1). De esta forma podemos encontrar la V de veracidad (introducida por IBM) o, por ejemplo, variabilidad o valor. Generalmente la descripción asociada al *big data* viene asociada a las limitaciones de las infraestructuras tecnológicas, de la capacidad de procesamiento y del almacenaje de estos datos.

2.1. Volumen

La capacidad de almacenaje en sistemas tradicionales se ve desbordada por el incremento exponencial de los datos. Podemos pensar en los entornos donde hay un SGBD (Sistema Gestor de Bases de Datos Relacional) que no estaba diseñado para soportar estas cargas de datos ni su variabilidad.

Como ya hemos mencionado anteriormente, el valor de volumen que de forma objetiva identifica un proyecto de *big data* no está escrito y depende de la naturaleza de nuestra infraestructura tecnológica actual. Estos valores pueden moverse por encima de los *petabytes*.

2.2. Velocidad

El concepto de velocidad tiene dos vertientes que deben considerarse. La primera tiene relación directa con el crecimiento de los datos generados. Si cada segundo generamos a través de nuestros sensores un *terabyte*, esta velocidad de carga es muy superior a una entrada de datos del orden de *megabytes*. Adicionalmente deberemos dar respuesta (en muchos casos en tiempo real) a unas necesidades de información y, por tanto, procesamiento de los datos de forma eficiente.

2.3. Variedad

En la descripción de esta V, hay que entender la procedencia y diversidad de los datos generados a través de los diferentes sensores y/o procesos de interés. Muchos datos provienen de un *streaming*; es decir, una cámara o micrófono envía imágenes o audio sin que exista nada más que la imagen en algún formato predeterminado (JPEG, GIFF) o un audio (.mp3) pero no son datos estructurados. El análisis de datos no estructurados como estos o, por ejemplo, el texto incrementan la dificultad de tratamiento.

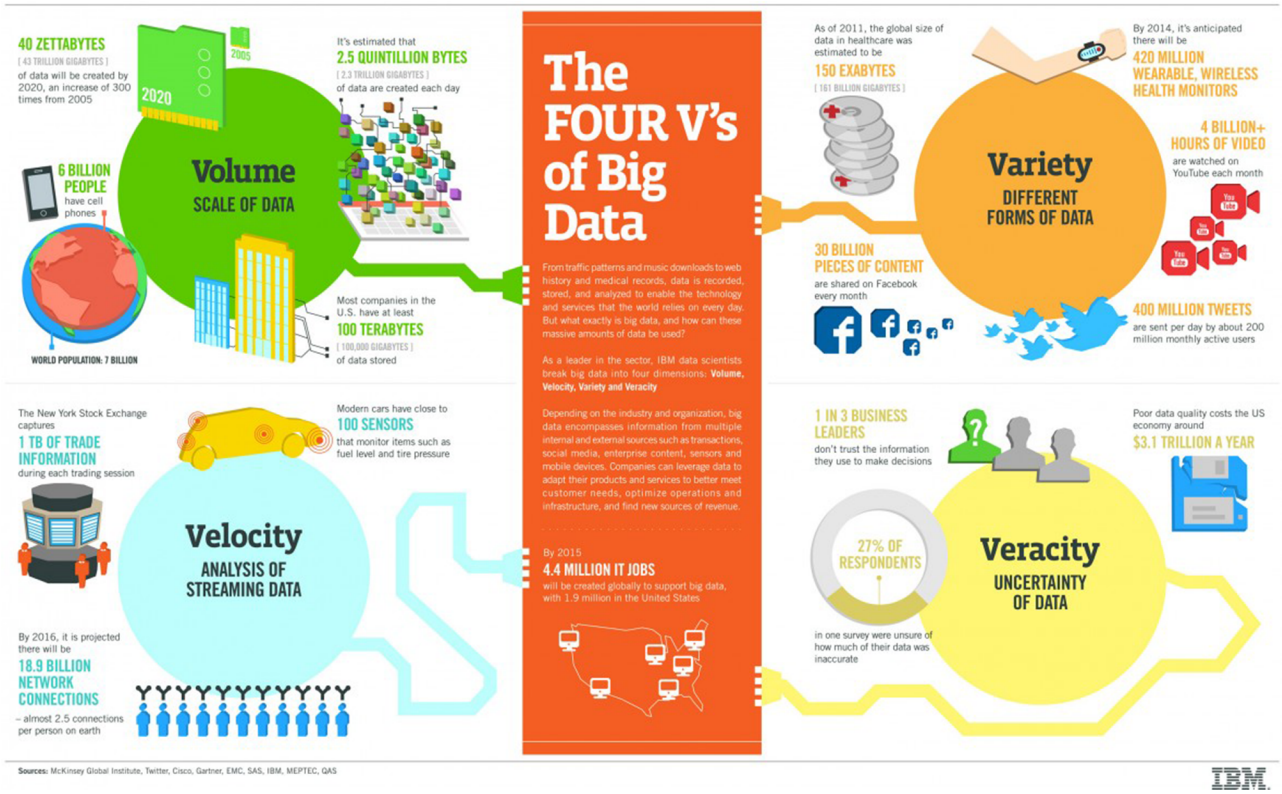
En otras ocasiones los datos son elementales (no están compuestos) y presentan una ventaja evidente en su procesamiento. Ejemplos de datos estructurados son la temperatura de un motor, o diversos parámetros relacionados con el rendimiento de una línea de producción.

Existe una tipología de datos que están a medio camino de los estructurados y de los no estructurados. Esta tipología llamada formalmente "datos semiestructurados" contiene información de cómo interpretar los datos contenidos en el documento. Podemos encontrar ejemplos en los formatos HTML, JSON o XML.

2.4. Veracidad

Una gran cantidad de datos a nuestro alcance no representan absolutamente nada si su calidad está en entredicho. En este sentido es importante determinar la exactitud de los datos, así como la exactitud en los análisis que pretenden identificar patrones y conocimiento estratégico para nuestra organización.

Figura 1. Las 4 V del big data.



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

2.5. Definición de *big data*

El concepto de *big data* se inició como una innovación tecnológica en el ámbito de la computación distribuida para tratar con grandes volúmenes de datos que inundan el día a día de cualquier negocio. Google recibe cada día 3.500 millones de búsquedas. Lo realmente importante es el provecho que obtenemos de los datos almacenados y, en este sentido, cómo analizamos y visualizamos los datos para realizar mejores decisiones y ayudamos a dirigir y definir las estrategias de negocio exitosas.

3. Plataformas de procesado *big data*

Las plataformas y las infraestructuras que dan soporte a proyectos *big data* son en muchos casos complejos y en ocasiones se necesita información del sector industrial para poder ofrecer una mejor solución. Podríamos añadir que pueden aparecer importantes aspectos no incluidos en la definición del concepto de *big data* pero que pueden tener una criticidad o importancia extrema (es decir, seguridad).

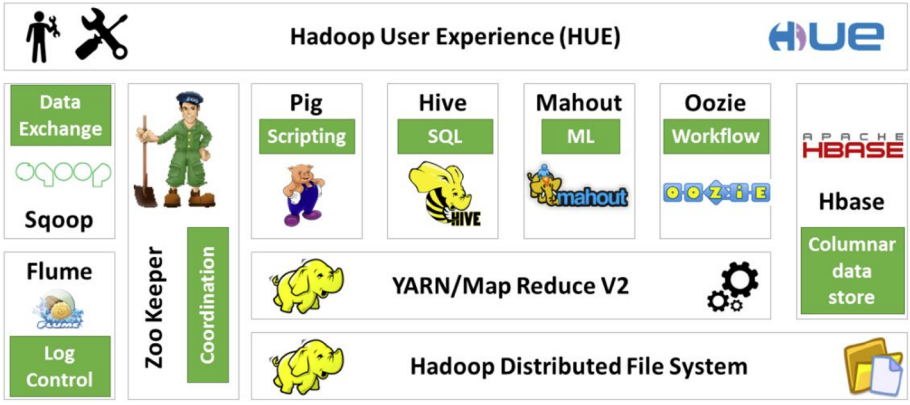
Desde una perspectiva conceptual podemos entender el problema definido por proyectos *big data* como un problema asociado a los sistemas de computación distribuidos (DCS: *Distributed Computer System*). Estos sistemas se componen de diversos elementos *software* que están en múltiples máquinas pero que se ejecutan como un sistema único. En otras palabras, desde la perspectiva del usuario, aunque los elementos están distribuidos se trata como si fuera un único ordenador.

Evidentemente este enfoque presenta dificultades cuando entran en escena indicadores de rendimiento en procesado de datos.

3.1. Apache Hadoop

Utilizado inicialmente por Yahoo! y Facebook, Hadoop es una plataforma de procesado de datos *open source* que permite almacenar y procesar grandes cantidades de datos sobre un conjunto de *clusters*. Los elementos más significativos de esta arquitectura son el sistema de ficheros distribuido (HDFS: *Hadoop Distributed File System*) y el modelo de programación MapReduce.

Figura 2. Hadoop-Ecosystem.



Fuente: https://www.ktexperts.com/hadoop-ecosystem/

A pesar de sus muchas ventajas, tiene algunos elementos mejorables relativos a la seguridad, su falta de ajuste a datos de menor tamaño y sus altos *overheads* en operaciones de entrada y salida. En la figura 2 se aprecia el llamado ecosistema de Hadoop.

3.2. Cloudera Enterprise Data Hub

Esta solución introduce Cloudera Enterprise Data Hub como un producto destinado a convertir los datos en propuestas de valor para el negocio. Ofrece un *framework* basado en Hadoop para entornos IdC y de *big data*.

4. *Big Data Analytics*

Existe un claro interés en los métodos llamados analíticos que en su terminología en ingles los encontramos como: Predictive Analytics o Big Data Analytics. De una forma simplificada podemos definir Big Data Analytics como el conjunto de técnicas y métodos que permiten aprender de los datos. El objetivo de analizar este conjunto enorme y diverso de datos es para descubrir patrones o relaciones que de otra forma pasarían inadvertidos u ocultos. Así, se consigue información relevante que puede ofrecer a las empresas una clara ventaja competitiva respecto a sus rivales. Este conocimiento puede abarcar una lista diversa de industrias y disciplinas (banca y sector financiero, grandes superficies, empresas de seguros, transporte y logística, salud, defensa, telecomunicaciones, energía o producción industrial, entre otros), puede cubrir aplicaciones tan diversas como la identificación de correo basura (*spam*), así como procesos de recomendación para nuestros clientes, o puede predecir tendencias del mercado.

Las organizaciones que ponen a su servicio técnicas analíticas pueden realizar la toma de decisiones de una forma más segura y con criterios más fundamentados.

La relevancia de los métodos analíticos en los entornos *big data* permite crear valor desde la perspectiva del negocio. Por ejemplo:

- Las compañías que usan datos para dirigir el negocio rinden mucho más que las que no lo hacen (es decir, Amazon vs Barnes&Nobles)
- Conseguir resultados en términos de:
  - Incremento de ventas.
  - Reducción de costes.
  - Incremento de la productividad.
  - Predecir amenazas de seguridad.
  - Prevención del fraude.
  - Reducción del coste de adquisición de un nuevo cliente.
  - Optimización de la política de precios.
  - Comprensión semántica de imágenes.
  - Gestión del capital de las redes sociales.
  - Recomendación de productos/servicios.

Sin entrar en detalle de las diferentes soluciones *software* que actualmente se ofrecen y que en la mayoría de casos se adaptan al nuevo entorno de la arquitectura *big data* (Hadoop, MapReduce, etc.), vamos a realizar un pequeño repaso a la tipología de problemas que podemos encontrarnos.

Este conjunto de técnicas proviene de diferentes disciplinas y en algunos casos se solapan entre ellas. Dentro de estas disciplinas podemos encontrar: visualización, reconocimiento de patrones, redes neuronales, *deep learning*, *data mining* o *machine learning*.

Entre las diferentes categorías de técnicas encontramos:

a) Sistemas de aprendizaje supervisado: se basan en hacer predicciones sobre valores y/o etiquetas basándose en el comportamiento o características de los datos que previamente tenemos a nuestra disposición.
- Ejemplos de este tipo de sistemas son: identificar el precio de venta de una vivienda a partir de los metros cuadrados que tiene (aquí queremos predecir el valor de la vivienda).
- Entre las técnicas que ofrecen este tipo de soluciones, cabe destacar:
  - Regresión lineal y logística
  - Clasificación: árboles, *random forest*, *gradient boosting*
  - Redes neuronales, *deep learning*
  - SVM

b) Sistemas de aprendizaje no supervisado: se basan en datos que no tienen etiqueta alguna. El objetivo de este grupo de técnicas es encontrar algún procedimiento que agrupe de forma homogénea los individuos o instancias parecidas (similares) y separe los que no son parecidos.
- Ejemplos de este tipo de sistemas son: agrupar los clientes en función del conjunto de variables que los describen
- Entre las técnicas que ofrecen este tipo de soluciones, destaca:
  - *Clustering*: K-means

5. El *big data* en la fábrica inteligente

La Industria 4.0, como ya se ha comentado ampliamente, representa la integración vertical y horizontal de la empresa, y abre canales de transmisión de datos que hasta la fecha habían estado altamente dificultados o cerrados totalmente. Esta integración ha significado un aumento enorme en la transmisión, almacenamiento y uso de datos de la planta que ahora van a ser tratados por las técnicas de *big data* con la finalidad de mejorar las ideas de producto, influir en el diseño, conseguir mayor eficiencia en la producción, mejorar el mantenimiento y reducir el impacto medioambiental por parte de las fábricas inteligentes.

Hasta nuestros días, todos estos datos, mucho menores en cantidad, que se circunscribían al ámbito que se consideraba directamente concernido en su uso, eran normalmente tratados con técnicas estadísticas muy básicas que no permitían en absoluto encontrar patrones y relaciones que serían de gran utilidad si fuesen convenientemente tratadas, cruzadas y con un número suficiente de datos. *Big data* permite introducir este tratamiento de los datos para extraer de ellos información relevante y estratégica, también, de los datos obtenidos en la planta.