Ingeniería para el Procesado Masivo de Datos

Tema 1. Introducción a las tecnologías big data

Índice

Esquema

1.1. Introducción y objetivos

Empezaremos la asignatura motivando los contenidos que se estudiarán en el resto del temario. Repasaremos las necesidades de la sociedad de la información en la actualidad, una era en la que todos estamos interconectados y somos fuentes de datos. Veremos los retos tecnológicos que esto supone y presentaremos formalmente las tecnologías que los solventan.

Los objetivos que persigue este tema son:

- Comprender cuáles son las necesidades actuales de procesamiento de datos, sus causas y cómo son solventadas por las tecnologías *big data*.
- Entender el concepto de clúster de ordenadores y cuáles son las principales tecnologías distribuidas capaces de explotarlo.
- Conocer las herramientas principales que componen el ecosistema Hadoop, cuál es la finalidad de cada una y cómo se relacionan entre sí.

1.2. La sociedad interconectada: la era del cliente

Las tecnologías *big data* surgen para dar respuesta a las nuevas necesidades de la sociedad actual. Vivimos en un mundo interconectado, en el que el 90 % de la información existente, preservada en medios de cualquier tipo, se ha creado en los últimos dos años. El crecimiento de la información producida en el mundo por fuentes de todo tipo, tanto físicas como electrónicas, es exponencial de un tiempo a esta parte. Aunque las estimaciones acerca del volumen divergen, la siguiente gráfica muestra de manera orientativa este fenómeno.

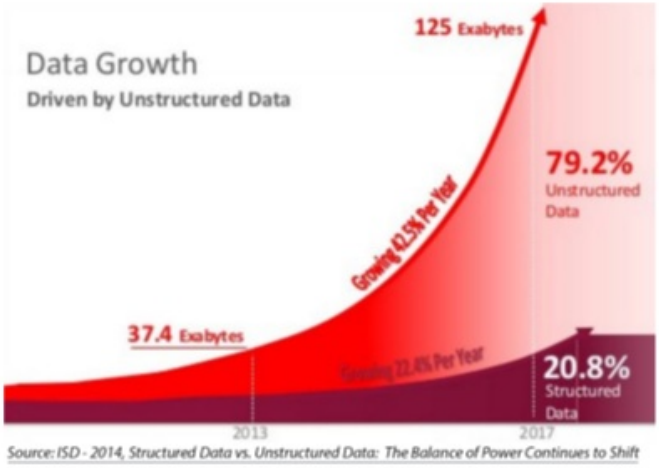


Figura 1. Previsión de crecimiento de los datos generados en todo el mundo. Fuente: Oracle Machine Learning.

Casi el 80 % de los datos que se crean son generados por personas y, por ello, suelen ser datos no estructurados (texto libre, comentarios de personas, tuits, imágenes, sonidos, vídeos). Los 20 % restantes son datos estructurados generados por máquinas [datos de *logs*, sensores, Internet de las cosas (IoT), en general] con el fin de ser procesados generalmente por otras máquinas.

Fuentes de datos en la actualidad

Ideas clave

Existen principalmente tres tipos de situaciones que generan datos en la actualidad:

- La **interacción entre humanos** a través de un sistema informático que registra información mientras se produce la interacción. Ejemplos claros son el correo electrónico, los foros de Internet o las redes sociales, donde los datos los generamos los humanos al interactuar entre nosotros utilizando dichos medios. Suelen ser datos no estructurados, posteriormente procesados por máquinas.

- La **interacción entre un humano y una máquina.** El ejemplo más claro es la navegación en Internet: los servidores web generan *logs* con información sobre el proceso de navegación. Lo mismo ocurre al efectuar compras en alguna plataforma web de comercio electrónico o en banca *online*, donde cada una de nuestras transacciones queda registrada y será procesada después con el objetivo de estudiar nuestro comportamiento, así como de ofrecernos productos mejores y más personalizados. Tienden a ser datos estructurados o semiestructurados.

- La **interacción entre máquinas.** Varias máquinas intercambian información y la almacenan con el objetivo de ser procesada por otras máquinas. Un ejemplo son los sistemas de monitorización, en los que un sistema de sensores suministra la información recibida a otras máquinas para que realicen algún procesado sobre los datos. Al ser la propia máquina quien la genera, suele ser información estructurada, ya que el *software* se encarga de sistematizarla.

Algunas cifras que resumen la cantidad de datos generados gracias a Internet se recogen en la siguiente imagen. Llama especialmente la atención el crecimiento experimentado por empresas como Netflix o Instagram, frente a la estabilización de los gigantes como Google, Facebook o YouTube.

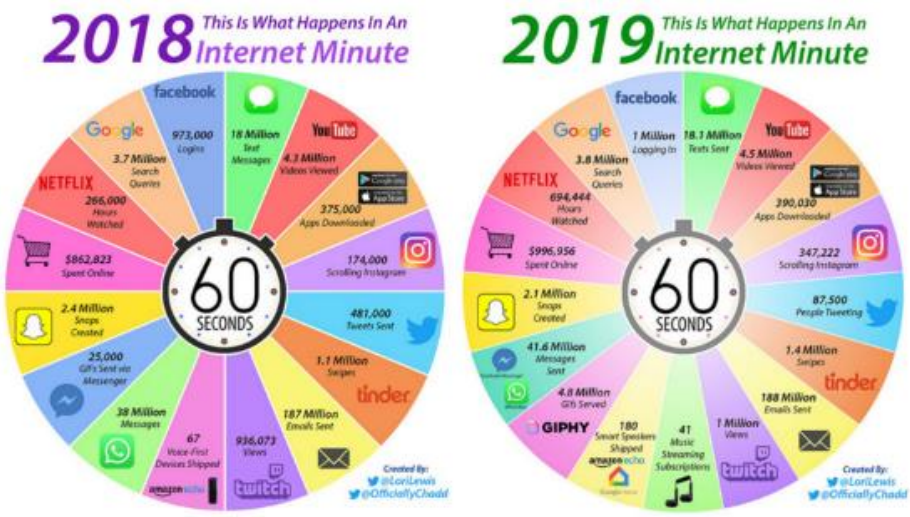


Figura 2. Eventos generados en Internet en un minuto por las empresas digitales. Fuente: Twitter.

La transformación digital en relación con los datos

La conclusión global a la que llegamos es que el mundo ya ha cambiado, y lo podemos confirmar si examinamos hechos como los siguientes:

- La empresa que transporta a más personas en el mundo es Uber, que tiene 0 coches físicos.
- La empresa que más habitaciones reserva es Airbnb, que tiene 0 hoteles físicos.
- La empresa que más música vende es Spotify, que tiene 0 estudios de grabación.
- La empresa que vende más películas es Netflix, que tiene 0 estudios.

Con frecuencia, se llevan a cabo más interacciones digitales que físicas entre las personas y las compañías que nos dan servicio, ya sea de suministro de energía, agua o gas; de telecomunicaciones o telefonía; de venta *online* de productos de todo tipo, o incluso de movimientos y servicios bancarios. Estas interacciones están generando, de forma continuada y masiva, datos muy valiosos que hablan del comportamiento de los clientes y su análisis permite anticipar qué es lo que estos van a demandar. De hecho, estamos evolucionando más rápido que las propias

compañías, hasta el punto de que se ha abierto una brecha entre las empresas físicas tradicionales y los gigantes digitales, como muestra la figura 2.

Con el objetivo de llenar este espacio, surge la **transformación digital**, que persigue esencialmente tres objetivos:

- **Centrarse en el cliente**, es decir, pensar continuamente en lo que necesita y en mejorar (personalizar) su experiencia y sus interacciones con la compañía. Esto requiere recabar y analizar grandes cantidades de datos sobre su comportamiento.
- **Centrarse en canales digitales,** especialmente dispositivos móviles, puesto que las interacciones digitales son las que generan mayor cantidad de datos y, cada vez con más frecuencia, se realizan usando estos dispositivos en vez del PC.
- **Decisiones guiadas por los datos** (*data-driven*), para lo cual es necesaria la ciencia de (grandes) datos (*big data science*).

1.3. Definición de las tecnologías big data

Para acometer estos objetivos, la mayor parte de las tecnologías existentes hasta hace pocos años (principios del siglo xx) no eran suficientes. Ello se debía a la necesidad de procesar, almacenar y analizar datos con ciertas características especiales, las denominadas **tres «v» del *big data***:

- **Volumen**: cantidades de datos lo suficientemente grandes como para no poderse procesar con tecnologías tradicionales.
- **Velocidad**: flujos de datos que van llegando en tiempo real y tienen que procesarse de manera continua según se van recibiendo.
- **Variedad**: datos de fuentes diversas, estructuradas y no estructuradas (sean bases de datos relacionales o no relacionales, datos de imágenes, sonido, etc.), que tienen que ser manejados y cruzados de manera conjunta.

Un proyecto es *big data* cuando implica alguna de las tres «v».

Una definición más ajustada y realista de un proyecto *big data* sería:

Un proyecto es *big data* cuando la mejor manera de resolverlo (más rápida, eficiente, sencilla) implica utilizar tecnologías *big data*.

Podemos definir *big data* como:

Conjunto de **tecnologías** y **arquitecturas** para almacenar, mover, acceder y procesar (incluido analizar) datos que eran muy difíciles o imposibles de manejar con tecnologías tradicionales.

Las causas de esta imposibilidad pueden ser:

- Cantidades ingentes de datos inimaginables hace unos años.
- Datos de fuentes diversas, heterogéneas, poco estructuradas, como documentos o imágenes/sonido, que, aun así, necesitamos almacenar y consultar (NoSQL).
- Datos dinámicos, recibidos y procesados según llegan (flujos de datos o *streams*).

Cabe destacar que, en la definición anterior, se han omitido de manera deliberada palabras como algoritmo, inteligencia, ciencia de datos o cualquier referencia a qué hacer o cómo analizar y explotar dichos datos. Las herramientas *big data* permiten aplicar a datos masivos técnicas que ya existían, pero son tecnologías y no técnicas en sí mismas. Las técnicas de análisis pertenecen al ámbito de la estadística, las matemáticas, las ciencias de la computación y la inteligencia artificial.

La mayoría de estas técnicas y algoritmos han existido desde mucho antes, algunas desde mediados del siglo xx. La sinergia con las tecnologías *big data* consiste en que estas permiten aplicar técnicas de análisis existentes a cantidades de datos mucho mayores, de naturaleza heterogénea. De este modo, logran resultados en menos tiempo y de mucha más calidad, al cruzar datos de diversas fuentes y ser capaces de procesarlos y usarlos para entrenar un algoritmo. Desgraciadamente, se han extendido entre el gran público mitos sobre el término *big data*, tales como los siguientes:

Figura 3. Conceptos erróneos de la finalidad de las tecnologías *big data* extendidos en la sociedad. Fuente: Google Images.

1.4. Origen de las tecnologías big data

La primera empresa que fue consciente del aumento de los datos que se estaban generando en Internet fue Google, ya que su buscador debe ser capaz de indexar las webs nuevas para que puedan ser encontradas. En los albores del siglo xx, Sanjay Ghemawat, Howard Gobioff y Shun-Tak Leung (2003) publicaron un artículo que se hizo mundialmente famoso, en el cual explicaban el sistema de archivos distribuido Google File System (GFS) que habían desarrollado. Los autores presentaron por primera vez la idea de utilizar ordenadores convencionales conectados entre sí (formando un clúster) para poder almacenar archivos que ocupaban más que un solo disco duro.

A esto se le denomina ***commodity hardware***: máquinas no especialmente potentes, similares a las que tienen los usuarios domésticos, pueden conectarse entre sí para trabajar conjuntamente como una sola, a fin de resolver tareas de mayor envergadura. GFS fue la base del sistema de archivos distribuido HDFS que veremos en temas posteriores.

En 2004, Jeffrey Dean y Sanjay Ghemawat publicaron un nuevo artículo, que se popularizó rápidamente, donde explicaban un modelo de programación (MapReduce) aplicable a un clúster de ordenadores para procesar en paralelo archivos almacenados en el sistema de archivos GFS. Google también publicó una biblioteca de programación de código abierto donde implementaba dicho paradigma. Su principal punto fuerte era la abstracción (simplificación) de todos los detalles de *hardware*, redes y comunicación entre los nodos del clúster, para que el usuario pudiera centrarse en el desarrollo de la lógica de la aplicación distribuida de manera sencilla. Durante muchos años, MapReduce fue el estándar de desarrollo de *software big data* a nivel comercial.

En 2009, y motivado por las deficiencias de Hadoop en ciertas tareas, un

investigador llamado Matei Zaharia creó una nueva tecnología *open source* de procesamiento distribuido, llamada Apache Spark, durante la realización de su tesis doctoral en Berkeley. Estudiaremos Spark en profundidad en temas posteriores; por el momento, basta señalar que comparte con MapReduce los principios de ejecutar en un clúster de ordenadores *commodity* y simplificar todos los detalles de redes y comunicación entre los nodos. Desde 2014, MapReduce ha sido reemplazado por Spark en su totalidad. Las herramientas que lo utilizaban como motor de ejecución han sobrevivido gracias a que dicho motor es una pieza intercambiable en muchas de ellas y han sabido adaptarlo a Spark.

El ecosistema Hadoop

La idea básica que hay tras las tecnologías de procesamiento distribuido es la siguiente:

> Es posible procesar grandes cantidades de datos de forma distribuida entre varias máquinas interconectadas (clúster), cada una no necesariamente muy potente (*commodity hardware*). Si se necesita más potencia de cálculo o más capacidad de almacenamiento, basta con añadir más máquinas al clúster.

Siguiendo esta filosofía y teniendo como punto de partida el sistema de archivos distribuido GFS (que en Hadoop se transformó en HDFS o Hadoop Distributed File System) y el paradigma MapReduce, se creó **un conjunto de herramientas *open source* para procesamiento distribuido,** cada una con un propósito específico, pero todas interoperables entre sí, que se denomina el **ecosistema Hadoop** (figura 5).

Figura 4. El clúster MareNostrum 4, en el Barcelona Supercomputing Center (BSC).Cada armario se denomina*rack* y cada bandeja es un ordenador completo (nodo).

Sin intentar ser exhaustivos y a título meramente informativo, damos una breve descripción de cada una:

- **HDFS:** sistema de archivos distribuido que estudiaremos en el tema siguiente.
- **MapReduce:** paradigma de programación para un clúster de ordenadores (forma de estructurar programas y también biblioteca de programación que se ejecuta sobre el clúster). Actualmente ha caído en desuso.
- **Flume:** herramienta para tratamiento de *logs*.
- **Sqoop:** herramienta para migración de grandes cantidades de datos desde bases de datos convencionales a HDFS.
- **Zookeeper:** coordinador.

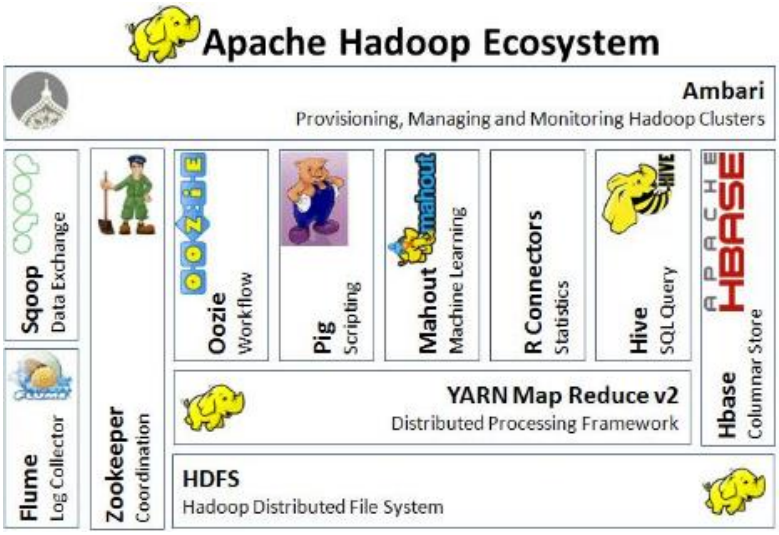


Figura 5. El ecosistema de herramientas *open source* Hadoop para procesamiento distribuido.

- **Oozie:** herramienta para planificación y ejecución de flujos de datos.
- **Pig:** herramienta para programar flujos de datos con sintaxis similar a SQL, pero con mayor nivel de granularidad, cuyo procesamiento se efectúa con MapReduce.
- **Mahout:** biblioteca de algoritmos de *machine learning*. Originalmente programada con MapReduce, tenía un rendimiento pobre, pero actualmente soporta otros *backend* como Spark.
- **R Connectors:** herramientas para conectar MapReduce con el lenguaje de programación R. En desuso, al igual que MapReduce.
- **Hive:** herramienta para manejar datos almacenados en HDFS utilizando lenguaje SQL. En su origen, utilizaba MapReduce como motor de ejecución. Actualmente soporta Spark y Apache Tez.
- **HBase:** base de datos NoSQL de tipo columnar, que permite, entre otras cosas, tener registros (filas) de longitud y número de campos variable.

En este curso, nos centraremos en las siguientes herramientas de Apache, que constituyen el estándar tecnológico *de facto* en la mayoría de las empresas que

utilizan tecnologías *big data*:

- **HDFS (Hadoop Distributed File System):** sistema de archivos distribuido inspirado en el GFS de Google, que permite distribuir los datos entre distintos nodos de un clúster, gestionando la distribución y la redundancia de forma transparente para el desarrollador que vaya a hacer uso de esos datos.

- **Apache Hive:** herramienta para acceder mediante sintaxis SQL a datos estructurados que están almacenados en un sistema de archivos distribuido, como HDFS u otros similares. Las consultas SQL son traducidas automáticamente a trabajos de procesamiento distribuido según el motor que se haya configurado, que puede ser MapReduce, Apache Spark o Apache Tez.

- **Apache Spark:** motor de procesamiento distribuido y bibliotecas de programación distribuida de propósito general, que opera siempre en la memoria principal (RAM) de los nodos del clúster. Desde hace unos años, ha reemplazado totalmente a MapReduce al ser mucho más rápido.

- **Apache Kafka:** plataforma para manejo de eventos en tiempo real, que consiste en una cola de mensajes distribuida y masivamente escalable sobre un clúster de ordenadores. Estos mensajes pueden ser consumidos por uno o varios procesos externos (por ejemplo, trabajos de Spark).

Distribuciones de Hadoop

Al estar constituido Hadoop por un conjunto de herramientas diferentes, cada una requiere su propia instalación y configuración en el clúster para poder operar con otras ya instaladas. Este proceso era tedioso y requería bastantes conocimientos. Con el fin de simplificar esta tarea, surgieron las distribuciones de Hadoop, que son conjuntos de herramientas del ecosistema Hadoop empaquetadas juntas, en versiones compatibles y perfectamente interoperables entre ellas, distribuidas con un

único *software*, razón por la que no hay necesidad de instalarlas por separado.

Empresas como Cloudera, Hortonworks (estas dos fueron competencia mutua durante mucho tiempo, hasta que acabaron por fusionarse en 2018) o MapR nacieron para crear distribuciones de Hadoop y añadirles, en algunos casos, herramientas propietarias totalmente nuevas, que no pertenecían al ecosistema Hadoop, o incluso modificaciones propias del código fuente original de las herramientas de Hadoop, para solucionar fallos o añadir características avanzadas. Todas las distribuciones de Hadoop de estas empresas tienen versiones *open source* y de pago. La siguiente tabla, incluida a título meramente informativo, compara sus características:

	Cloudera	Hortonworks	MapR
Componentes	Apache modificados y añadidos	Solo Apache oficiales	Apache y añadidos
Versiones	*Open source* (CDH) y de pago	Sólo 100 % *open source*	*Open source* y de pago
Sistema operativo	Linux (Windows vía VMWare)	Linux y Windows	Linux (Windows: VM Ware)
Año de creación	2008	2011	2009
Notas adicionales	Es la más extendida. Certificación muy popular.	Única para Windows, única 100 % *open source*	La más rápida y fácil de instalar

Tabla 1. Características de herramientas empresariales.

1.5. Referencias bibliográficas

Dean, J. y Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113. https://doi.org/10.1145/1327452.1327492

Ghemawat, S., Gobioff, H. y Leung, S-T. (2003). The Google File System. A *CM SIGOPS Operating Systems Review*, *37*(5), 29-43. https://doi.org/10.1145/1165389.945450

El papel del big data en la transformación digital

Newman, D. (2015, 22 de diciembre). The role big data plays in digital transformation. *Forbes*. https://www.forbes.com/sites/danielnewman/2015/12/22/the-role-big-data-plays-in-digital-transformation/

En este artículo, el autor explica de qué manera el *big data* puede ayudar a grandes y pequeñas empresas a minimizar costes, maximizar resultados y, en definitiva, ser competitivas en un entorno cambiante.

Historia de Hadoop

Bonaci, M. (2015, 11 de abril). The history of Hadoop. *Medium.com.* https://medium.com/@markobonaci/the-history-of-hadoop-68984a11704

Este artículo analiza en profundidad el contexto en el que nació Hadoop y la evolución que ha experimentado.

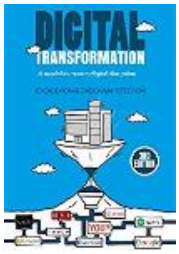
Leading digital: turning technology into business transformation

Westerman, G., Bonnet, D. y McAfee, A. (2014) *Leading digital: turning technology into business transformation*. Harvard Business Review Press.

En este manual, los autores ponen el foco en cómo grandes compañías de la industria tradicional están aprovechando las posibilidades de la era digital como estrategia para alcanzar el éxito.

Digital transformation: a model to master digital disruption

Caudron, J. y Van Peteghem, D. (2018) *Digital transformation: a model to master digital disruption* (3. a edición). BookBaby.



La era digital ha creado multitud de oportunidades para desarrollar nuevos productos y servicios, y abrir nuevas líneas de negocio. Este libro ofrece una metodología para que las empresas que aún no lo han hecho den el salto y aprovechen todas las ventajas que ofrece la transformación digital.

Test

1. En la sociedad actual, la mayoría de los datos que se generan a diario son…

   A. Datos no estructurados generados por las personas.

   B. Datos estructurados generados por máquinas.

   C. Datos estructurados generados por las personas.

2. ¿Qué retos presentan los datos generados por personas en una red social?

   A. Son datos no estructurados (imágenes, vídeos), más difíciles de procesar.

   B. Son datos masivos.

   C. Las dos respuestas anteriores son correctas.

3. El término *commodity hardware* se refiere a…

   A. Máquinas remotas que se alquilan a un proveedor de *cloud* como Amazon.

   B. Máquinas muy potentes que suelen adquirir las grandes empresas.

   C. Máquinas de potencia y coste normales, conectadas entre sí para formar un clúster más potente.

4. Un proyecto se denomina *big data* cuando…

   A. Solo se puede resolver gracias a las tecnologías *big data*.

   B. La forma más eficaz y directa de abordarlo implica tecnologías *big data*.

   C. El problema que resuelve contiene simultáneamente las tres «v».

5. Las tres «v» del *big data* se refieren a:

   A. Volumen, velocidad y variedad.

   B. Voracidad, volumen y velocidad.

   C. Ninguna de las respuestas anteriores es correcta.

6. Lo mejor, si necesitamos más potencia de cómputo en un clúster *big data*, es…

A. Reemplazar algunas máquinas del clúster por otras más potentes.

B. Aumentar el ancho de banda de la red.

C. Añadir más máquinas al clúster y aprovechar todas las que ya había.

7. El sistema de ficheros precursor de HDFS fue…

A. GFS.

B. Apache Hadoop.

C. Apache MapReduce.

8. Una distribución de Hadoop es…

A. Un *software* con licencia comercial para clústeres, difundido por Microsoft.

B. Un conjunto de aplicaciones del ecosistema Hadoop, con versiones interoperables entre sí y listas para usarse.

C. Ninguna de las opciones anteriores es correcta.

9. ¿Qué compañías fueron precursoras de HDFS y MapReduce?

A. Google y Microsoft, respectivamente.

B. Google, en los dos casos.

C. Google y Apache, respectivamente.

10. Definimos *big data* como…

A. Todos aquellos algoritmos que se pueden ejecutar sobre un clúster de ordenadores.

B. Las tecnologías distribuidas de Internet que posibilitan una sociedad interconectada por las redes sociales.

C. Las tecnologías que permiten almacenar, mover, procesar y analizar cantidades inmensas de datos heterogéneos.