Introducción: ¿Qué es Big Data?

Módulo 1. Conceptos básicos del Big Data

Introducción al Big Data

Antecedentes

El concepto de Big Data

En 1997 los investigadores de la NASA Michael Cox y David Ellsworth utilizaron por primera vez el término Big Data. La situación se originó al afrontar un problema de visualización de datos, cuando comenzaron a experimentar dificultades para manejar grandes volúmenes de datos. Se puede decir que en ese momento nació el término Big Data.
Desde aquel año han ocurrido muchas cosas. Lo que hoy en día entendemos como Big Data y su aplicación ha cambiado mucho en su uso.

Principales hitos

Algunas de las primeras empresas que se enfrentaron al fenómeno de datos masivos fueron las compañías de Internet, como Google, Facebook, Yahoo! o X.
Estas compañías tenían el gran reto de gestionar grandes volúmenes de datos en un tiempo razonable, ya que la información sólo les era de utilidad si llegaba a tiempo.

En el año 2002, en el contexto tecnológico que se desarrollaba, la infraestructura de la que disponían no les permitía conseguir sus objetivos, ya que debían manejar demasiados datos para los que la infraestructura existente no daba un soporte. Por ello se plantearon realizar una gran inversión en plataformas novedosas de gestión de datos.
Entre estas compañía, fue Google quién realizó el mayor avance en esta área y en concreto para resolver el problema de la creación de índices invertidos, estructuras de datos del estilo:

<Palabra>:<lista de documentos que contienen Palabra>.

¿Qué motivó a Google realizar este avance?

Veamos algunas cifras para entenderlo.
En aquellos años existían aproximadamente 20+ miles de millones de webs con una media de 20 Kilobytes = 400+ terabytes.

Una máquina puede leer aproximadamente entre 30 y 35 MB por segundo. Esto nos lleva a que necesitaría 4 meses para leer las webs, y necesitaría aproximadamente +1000 Discos duros, sólo para almacenar los datos en crudo.
Si en lugar de una máquina utilizamos 1000 máquinas la tarea podría realizarse en menos de 3 horas, pero esta solución origina otros problemas como:

Coordinación y comunicación entre las diferentes máquinas.

- Gestión y tolerancia a los fallos que se generan en el proceso. Depuración de problemas.
- Estado y reporte de cada tarea/actividad en cada máquina.
- Optimización de los procesos.
- Manejo de los datos locales.

Todos estos inconvenientes debían solucionarse en un tiempo que permitiera sacar un beneficio a la información existente, para realmente ser de utilidad.

Map/Reduce, fue la solución que generaron entre 2003 y 2004.

Consiste en un modelo de programación para dar soporte a la computación paralela sobre grandes colecciones de datos en grupos de computadoras. Es decir, Map/Reduce maneja la

complejidad del trabajo en clúster (un clúster es un conjunto de máquinas que trabajan de manera coordinada en la solución de un problema) gracias a sus principales características:

- Distribución automática de actividades a cada máquina del clúster (paralelización).
- Manejo de la gestión de la carga.
- Optimización de las transferencias de información necesaria a disco y en la red.
- Manejo de fallos de máquinas.
- Robustez.

Con esta nueva infraestructura de procesamiento, Google describió como crear un framework, permitiendo manejar un volumen de datos antes desconocido de manera sencilla, ocultando toda la complejidad y la tolerancia a fallos en una serie de librerías.
La base del desarrollo de Map/Reduce se puede resumir en una serie de publicaciones clave, en materia de infraestructura de Big Data, difundidas por Google:

- En 2003 publica un artículo en el que describe el sistema de archivos distribuidos que utilizan, GFS (Google File System).
- En 2004 publican un artículo sobre el tratamiento masivo de datos en clúster. Map/Reduce: Simplified Data Processing on Large Clusters.
- En 2006 publican un artículo sobre BigTable, describe como tener un Sistema distribuido de almacenamiento para datos estructurados. A Distributed Storage System for Structured Data.

Estas tres publicaciones fueron claves para el futuro del Big Data, ya que en ese momento Doug Cutting estaba desarrollando un motor de búsqueda Nutch y se estaba encontrando con problemas de escalabilidad. Al leer los artículos publicados por Google, entendió que eran la solución a su problema de escalabilidad en el buscador Nutch.
La solución fue tan poderosa que derivó en un proyecto conocido con el nombre de Apache Hadoop (nombre del elefante amarillo que tenía su hijo).
Hadoop, es el sistema de código abierto que más se utiliza para almacenar, procesar y analizar grandes volúmenes de datos (cientos de terabytes, petabytes o incluso más), estructurados o no, archivos de registro, imágenes, video, audio, comunicación, etc. Tiene una gran comunidad que sustenta su desarrollo e innumerables proyectos asociados. Estos proyectos cuando tienen suficiente madurez se acaban incorporando como parte de las distribuciones de Hadoop.

A partir del año 2007, se comenzó a extender el uso del Big Data de manera generalizada en el mundo empresarial y de negocios para el tratamiento de grandes volúmenes de datos.
En 2008 aparece Cloudera.
Esta compañía la fundan un antiguo ejecutivo de Oracle, y tres ingenieros que provienen de las empreas Google, Yahoo! y Facebook. Doug Cutting se une al proyecto un año más tarde.
La clave: un software accesible.
Una gran parte del software que se utiliza en las plataformas de tratamiento masivo de datos es *Open source,* es decir, que no hay que pagar licencias por ello.
Esto ha sido clave para la expansión de esta tecnología. La barrera de las licencias que antes imponían las grandes corporaciones se terminó y la tecnología está al alcance de cualquiera.

Las V'S que definen Big Data

La clave del Big Data es dar respuesta a las tres grandes V. Doug Laney fue el primero que def nió el reto de las tres V: Volumen, Velocidad y Variedad.

Volumen

Este concepto está directamente relacionado con el Big Data y hace referencia a la cantidad de datos que tenemos que manejar.
Para hacernos una idea del incremento de data disponible en los últimos años veamos las siguientes cifras:

2008. Las CPUs del mundo procesaron 9,57 zettabytes de información.

2009. "La próxima frontera para la innovación y productividad" (Compañía McKinsey).

2010. La cantidad de información crece exponencialmente conforme avanza la tecnología (Eric Schmidt).

En 2008, entre todas las CPU del mundo se procesaron 9,57 zettabytes de información, lo que significan 9.570.000.000.000 gigabytes (9,57 billones de gigabytes).

En 2009, la compañía McKinsey es la autora de "La próxima frontera para la innovación y productividad", en la que cifra que una compañía americana de aproximadamente 1000 empleados almacenará aproximadamente 200 terabytes de información al año.

En 2010, en una conferencia expuesta por Eric Schmidt, Google, se aportó el impresionante dato de que la cantidad de datos generados en la actualidad en dos días es mayor que la generada por toda la civilización hasta el 2003. Esto significa que la cantidad de información crece exponencialmente conforme avanza la tecnología. Además, todo lo que nos rodea (móviles, redes sociales, así como la imparable digitalización) generan un proceso ya imparable de digitalización, donde los datos no pararán de crecer.

Velocidad

La velocidad es otra característica fundamental. Y es que debemos ser capaces de conocer la información a la velocidad a la que se genera y lo más relevante, tratar y procesarla durante el periodo que sea válida para tener el producto actualizado y obtener así su máximo provecho. Un ejemplo claro de esto sería si un usuario sube una foto a una red social y esta no está disponible para el resto de los usuarios hasta varias horas después. Seguramente este sistema no resultará interesante para los usuarios aunque permita volumen, porque no tiene velocidad de respuesta.

Variedad

En Big Data es habitual trabajar con un número amplio de fuentes de información, que pueden ser fuentes estructuradas, semiestructuradas o no estructuradas, teniendo cada una diversos formatos de tipos de datos. Por ejemplo texto, voz, vídeo, etc.

La información puede clasificarse en diferentes tipos de datos:

Estructurados	No estructurados	Semiestructurados
Tienen un formato definido claro. Por ejemplo, los datos incluidos en una base de datos o aquellos datos de una transacción en que cada campo tiene un claro significado.	Son datos en crudo que no tienen un formato específico. Estos datos no podemos volcarlos a una base de datos relacional porque no c onocemos su tipología. Podemos incluir vídeo, imágenes y voz.	Datos que no se limitan a campos determinados, pues contienen marcadores para separar las diferentes secciones. Son datos que tienen su propio esquema (metadato), ejemplos HTML, XML, Json. También podemos incluir en esta categoría los logs que son una fuente muy importante de información en Big Data.

Aunque estas son las tres primeras V's según ha pasado el tiempo se han aceptado otras igualmente relevantes, las vemos a continuación.

Veracidad

Cuando operamos con muchas fuentes que generan gran cantidad de datos a gran velocidad, es lógico asegurar el grado de veracidad que tienen para así conseguir una maximización de los beneficios en su explotación.

Es decir, no tiene mucho sentido tratar datos obtenidos a través de 5G y que la información no sea veraz por tener una gran distorsión. Esto nos daría como resultado un producto que no cumple con las expectativas.

Por esta razón, es necesario:

1. Realizar una limpieza de los datos.
2. Asegurar la fiabilidad de las fuentes de información. La fiabilidad es más o menos importante en función del negocio, pasando de ser crítica a no vital en función de qué aplicación concreta estemos analizando.

Valor

La última "V" es el valor. A pesar de ser un software Open Source, poner en marcha toda esta infraestructura resulta bastante caro. Por ello, hay que asegurar que el proyecto genera valor para la compañía. Un proceso para certificar esto es medirlo. Dependerá de cada caso concreto, considerándose indispensable generar un caso de negocio (*Business case*) antes de iniciar un nuevo proyecto de Big Data. Si no, existe el riego de que las expectativas nunca se cumplan. "Es imposible cumplir con lo que no se conoce".

Conceptos importantes

Relacionado con el concepto de Big Data hay dos ejes de actuación importantes. El primero, maneja y almacena los datos de manera masiva; y el segundo la parte analítica, que tiene como objetivo extraer conocimiento de los datos, lo cual permite una toma de decisión basada en datos. El manejo de los datos masivos se ha ido viendo a lo largo del documento, así que a continuación se explicará la parte analítica.

Análisis de datos

La analítica avanzada está muy unida al estudio de los datos y a una rama de la inteligencia artificial, el aprendizaje automático (Machine Learning). Esta consiste en la suma del tratamiento masivo de información conjunto con la aplicación de algoritmos de aprendizaje automático de las máquinas.
La idea está clara, una vez que hemos solucionado los inconvenientes de infraestructura, podemos analizar los datos de forma masiva con el fin de encontrar patrones, definir modelos y responder preguntas.

Business Intelligence vs Big Data

Una de las diferencias más importantes entre Business Intelligence (BI) y Big Data es que en BI preguntamos a los datos qué ha pasado y buscamos en ellos el por qué, algo parecido a un estudio forense que explique qué ocurrió y por qué. Estos datos, por ejemplo, nos sirven para la elaboración de informes.
En el caso de Big Data, preguntamos a los datos qué es lo que va a ocurrir con mayor probabilidad.
Es decir, aplicamos técnicas de gestión y almacenamiento de los datos para tomar mejores decisiones y movimientos estratégicos de negocio para intentar anticiparnos al futuro.

El Data Scientist

El perfil del científico de datos o Data Scientist es el responsable de analizar y cuestionar el gran volumen de datos obtenidos. Es la persona que limpia y asimila los datos para extraer su Valor mediante la aplicación de técnicas matemáticas, estadísticas y de aprendizaje automático.
Este perfil es el que contestará a las preguntas de los directivos de una empresa, incluso a aquellas preguntas que no sabían que tenían.
Lo importante, es definir qué pregunta le quiero hacer a mis datos, para que los científicos de datos faciliten la respuesta.

El papel del científico de datos

Este profesional dedica su tiempo a tareas de análisis estadístico más tradicional, a encontrar patrones de comportamiento aplicando algoritmos de minería de datos y a construir modelos predictivos aplicando técnicas de aprendizaje automático.
Su objetivo principal es la **extracción de conocimiento generalizable** a partir de los datos. Asimismo, incorporar las técnicas y métodos del trabajo de la investigación científica, es intensiva en procesamiento estadístico, reconocimiento de patrones, visualización y modelización de la incertidumbre, entre otras técnicas.

Bajo estas premisas un científico de datos debe tener tres habilidades que son importantes, y que lo convierten en un profesional completo y competente:

1. Conocimiento del negocio.
2. Grandes conocimientos de estadística y matemáticas.
3. Habilidades 'hacking' es decir capturar, tratar, transformar los datos en conocimiento mediante la aplicación de técnicas científicas y de aprendizaje automático.

Otro autor como Michael Gualtieri, define a un científico de datos como una persona que realiza nuevos descubrimientos. Hace hipótesis y trata de validarlas.
En este caso, busca conocimiento mediante la realización de hipótesis que debe validar.

Algunas de las tareas que realiza un científico de datos son:

- Visualizan datos e informes para buscar patrones en los datos, , esto es muy similar a BI.
- La diferencia es que los científicos de datos buscan algoritmos que expliquen y generalicen estos patrones mediante la creación de modelos, por eso es importante tener conocimientos profundos de estadísticas y aprendizaje automático.
- Responden a preguntas y modelan que es lo que va a ocurrir, basándose en los datos pasados.
- Demuestran mediante la confrontación de hipótesis.

Por otra parte, las **actividades** que hace un científico de datos son:

1. **Definición** de la pregunta a contestar o caso de negocio (negocio).
2. **Identificación** de las fuentes de datos.
3. **Entender** los datos.
4. **Extraer** los datos relevantes.
5. **Construir los conjuntos** de datos en los que basarse.
6. **Limpiar** los datos.
7. **Estudios estadísticos**.
8. **Modelado**.
9. **Iteración**, es difícil acertar a la primera. Un científico de datos se encuentra con que el primer modelo no cumple con las expectativas y debe entender por qué y que pasos debe seguir, el científico de datos debe poder contestar a las siguientes preguntas para entrar en el modelo iterativo:

¿Tenemos los datos que necesitamos? es decir, necesitamos más variables, o necesitamos menos, qué nuevas variables incorporamos o quitamos.
¿Son las variables independientes?
¿Tienen la escala adecuada?
¿Tengo suficientes datos o necesito más?
¿Es mi hipótesis la adecuada para contestar a la pregunta definida?
¿Estoy usando las técnicas adecuadas?

Es significativo tener personas de soporte en la organización que ayuden al científico de datos para obtener la máxima productividad. En la organización ¿tenemos expertos en ETL, en bases de datos y expertos integradores de la información? Es importante asignar cada actividad al experto correspondiente para poder tener la mayor productividad posible.

Tipos de estudios

En este apartado se comenta con más detalle los tipos de estudio que un científico de datos debe poder realizar. Son seis tipos, y están ordenados por nivel de complejidad, entendiendo que lo lógico es afrontar y conocer los primeros para ir obteniendo información de valor:

Descriptivos
Como indica su nombre, describe las características de los datos que forman parte del estudio (también llamadas variables) con el objetivo de estar al tanto de los datos.
En este primer momento se aplican técnicas descriptivas de análisis estadístico tanto univariable como bivariable.

Exploratorios
Encuentran y establecen conexiones entre los datos. Buscamos correlaciones, linealidad y relaciones entre las variables. Es importante resaltar que correlación no implica causalidad por lo que estos 'modelos' no son buenos para generalizar o predecir.

Inferenciales
Tienen como objetivo probar teorías, se ven muy afectadas por la muestra de los datos (no contamos con una muestra de todo el mundo, solo una parte) y su incertidumbre. Es el objetivo de los modelos estadísticos.

Predictivos
La idea es basarnos en los datos que tenemos para predecir el futuro, es difícil, debemos encontrar las variables adecuadas, el modelo y la técnica.

Causales
Tratan de explicar que ocurre a una variable cuando cambias otra. Suelen requerir el uso de estudios aleatorios.

Mecanísticos
Entender los cambiso exactos en las variables que provocan cambios en otras variables. Son realmente complejos y requiren mucha especialización estadística y física (en donde este tipo de modelos son más habituales).

Herramientas de un científico de datos

Hemos revisado funciones, habilidades, y tipos de estudios que debe conocer y llevar a cabo un científico de datos. La mayoría de estos perfiles están relacionados con los ámbitos de matemáticas, ingenierías etc.
A continuación, se numeran las principales herramientas que permiten a un científico de datos,realizar estudios estadísticos más profundos:

R
Lenguaje de programación con muy buenos algoritmos y soporte estadístico. Es Open Source y tiene una comunidad muy activa que va generando nuevos métodos y algoritmos que pueden ser incorporados mediante la instalación de nuevos paquetes.

SciPy
Conjunto de librerías científicas de Python. Python, otro lenguaje de programación convertida en herramienta de referencia para los científicos de datos, fácil de programar, exportable y como R, es Open Source y tiene una comunidad muy activa que va generando nuevos métodos y algoritmos mediante la instalación de nuevos paquetes.

MATLAB/ Gnu Octave
Realmente potentes, MATLAB es de pago y GNU Octave es Open Source. Son excelentes herramientas para el modelado y prototipado de datos. En el momento de poner el análisis en marcha debemos trasladar los datos a un lenguaje más apropiado (C, java, Python).

Herramientas de visualización
En Python hay numerosas librerías que dan soporte a la visualización tanto estadística como georreferenciada. Actualmente existen diversas herramientas con grandes facilidades para la visualización de datos como pueden ser Qlik o Tableau.
Aunque los científicos de data no las suelen usar habitualmente, son herramientas útiles a conocer para conseguir una mejor visualización de datos de manera fácil. Debido a la gran importancia que tiene una visualización clara de los datos, siguen apareciendo constantemente nuevas herramientas, como por ejemplo Kibana, Plotly, Carto o DataWrapper.

Gephi
Una herramienta de visualización de gráficos muy usada para representar datos de redes sociales.

Netlogo
Para hacer simulaciones de grafos.

NetworkX
Es es una herramienta para el análisis de grafos. Su ventaja es que es compatible con Python, pudiendo importar sus paquetes.