Introducción a Big Data

M1- Introducción.

Fundación Telefónica Movistar
CURSO DE INTRODUCCIÓN A BIG DATA.

M1. Introducción.

Antecedentes históricos.

El término Big Data se utilizó por primera vez en 1997, cuando dos investigadores de la NASA, Michael Cox y David Ellsworth, se encontraron con un problema de visualización de datos, que implicaba el manejo de grandes volúmenes de datos con herramientas de proceso tradicionales.

Muchos autores concuerdan que en ese momento nació el Big Data.

Las cosas han cambiado desde ese día hasta hoy, incluso lo que conocemos actualmente como Big Data.

Hitos principales.

Como era lógico suponer, las primeras empresas que tuvieron que enfrentarse a este nuevo desafío, fueron las empresas de internet, como Google, Facebook y otras tantas.

El desafío consistía en gestionar grandes volúmenes de datos, en un tiempo razonable para el usuario, ya que por el tipo de servicio que prestaban, la información solo resultaba de utilidad si la misma llegaba a tiempo.

El reto estaba instalado, allá por el 2002, estaba claro que el principal problema era de infraestructura, la actual en ese momento, no les permitía cumplir con los parámetros de servicio demandados. Había llegado el momento de hacer un cambio de paradigma en lo que a almacenamiento y proceso de información se trataba, y desde luego, acompañar este cambio con una enorme inversión monetaria.

Entre las principales inversiones realizadas, destaca la hecha por Google, en concreto con la creación de los índices invertidos.

Un índice invertido es una manera de estructurar la información, que luego será recuperada por un motor de búsqueda, con el objetivo de poder realizar una búsqueda de texto completa.

En este caso, el buscador crea os índices o parámetros de búsqueda, indicando cuáles son los documentos que los contienen, esta manera de indexar permite que el usuario acceda, al introducir los parámetros de búsqueda, a todos los documentos que contienen a dichos parámetros.

Fundamentos del cambio.

En esa época, se registraban, aproximadamente algo mas de 20 mil, millones de páginas web, con un contenido promedio de 20KB, lo que a grandes rasgos nos da un número importante de datos a recorrer.

Sin entrar demasiado en cuestiones técnicas de los equipos, con la velocidad a la que podían procesar datos las computadoras en ese momento, realizar una búsqueda en todas las páginas, llevaría aproximadamente de 4 a 5 meses.

Se pensó también en algo que hoy en día es habitual, pero que requirió su tiempo de maduración, que es el procesamiento paralelo, y lo que hoy es una forma de trabajo habitual sobre todo en Big Data, en ese entonces implicaba una serie de problemas nada fácil de solucionar:

- **Coordinar los equipos mediante un sistema operativo distribuido.**
- **Una compleja depuración de problemas.**
- **Estado de procesos en cada máquina.**
- **Optimización coordinada de los procesos.**
- **Identificación y manejo de datos locales.**

Todos estos ítems, entre otros, debían ser resueltos en tiempo real para darle verdadera utilidad a los datos de búsqueda.

MAPE/ REDUCE

Es un modelo de programación presente en la mayoría de los lenguajes modernos, que da soporte al procesamiento paralelo de grandes volúmenes de datos enfocado en grupos de computadoras.

Este modelo maneja la complejidad de trabajo en clúster (es un conjunto de máquinas que trabajan de manera coordinada para la solución de un determinado problema), fundamentalmente gracias a estas características:

- **Paralelización o distribución automática de actividades.**
- **Manejo coordinado de la gestión de carga de datos.**
- **Manejo optimizado de transferencia entre discos y red.**
- **Gestión de fallos.**
- **Arquitectura robusta.**

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.**

Sin duda, 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.

Luego, en 2007, con el uso de Big Data ya extendido de manera generalizada, no solo en el mundo empresarial, sino también en campo de la ciencia y la técnica, se comenzó a normalizar el tratamiento de grandes volúmenes de datos.

En 2008 aparece Cloudera, una compañía que proporciona software basado en Apache Hadoop que integra varias tecnologías y herramientas destinadas a crear y explotar DATA LAKE y DATA WAREHOUSE.

Un gran porcentaje del software que se utiliza en las plataformas para procesamiento masivo de datos es Open Source, es decir, no requieren el pago de una licencia para hacer uso de ellos.
Sin duda, este fue uno de los pilares para la expansión de esta tecnología, ya que las licencias imponían grandes barreras que limitaban el acceso a este tipo de tecnologías a unos pocos.

Las cuatro V de Big Data.

La clave principal de Big Data es dar respuesta a las cuatro V que caracterizan a este tipo de procesos:

Volumen.

Sin duda un concepto directamente ligado con Big Data, que hace referencia a la enorme cantidad de datos que tenemos que manejar.
Veamos algunas cifras para poner en contexto el tema del volumen:

MASIFICACIÓN USO DE INTERNET

SURGIMIENTO DE LAS REDES SOCIALES

CRECIMIENTO EXPONENCIAL DE DISPOSITIVOS MÓVILES

INTERFACES DE USUARIO MÁS SIMPLES E INTUITIVAS

CAMBIOS EN LAS FORMAS DE PROCESAMIENTO

FUERTE BAJA EN LOS COSTOS DE ALMACENAMIENTO

CADA DÍA CREAMOS 2,5 QUINTILLONES DE BYTES DE DATOS. (2,5 Exabytes)

EL 90% DE LOS DATOS DEL MUNDO DE HOY SE GENERARON EN LOS ÚLTIMOS 2 AÑOS

En 2008, entre todas las CPU del mundo se procesaron aproximadamente 9,57 billones de GB.

En 2009, la compañía McKinsey calcula que una empresa norteamericana promedio, de aproximadamente mil 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.

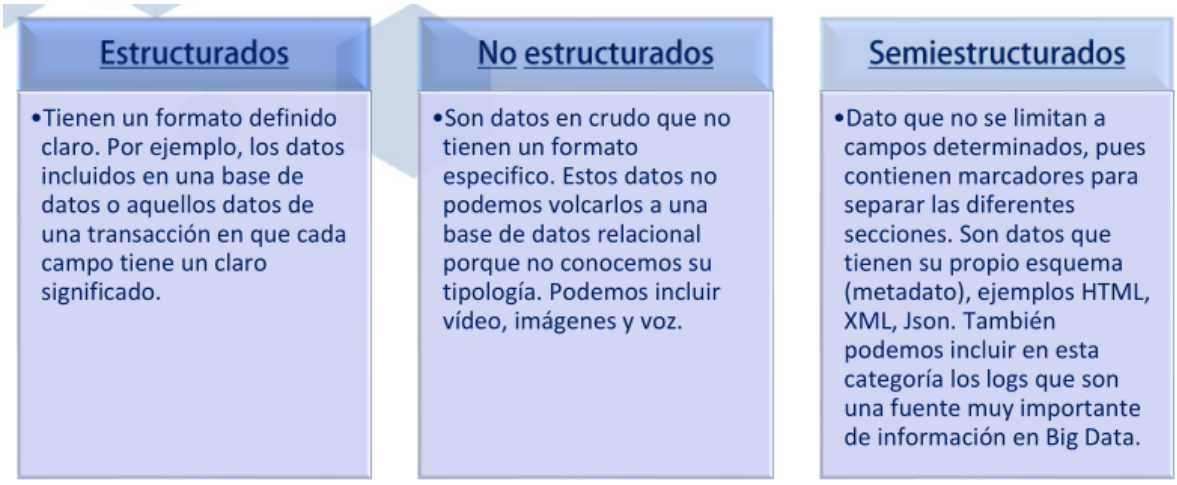
Para ver el proceso y la magnitud, veamos qué sucede en Internet en un sólo minuto. Referencia año 2019.

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, pero no tiene velocidad de respuesta.

Variedad.

Es habitual en Big Data trabajar con diferentes tipos de datos, así como distintas fuentes de información.
Es acá cuando notamos una diferencia fundamental entre todos los datos recolectados, y es básicamente su forma, ya que trabajaremos con datos estructurados, semiestructurados o directamente, no estructurados.
Hoy en día es común tener que realizar modelados o análisis en base a datos completamente diferentes, por ejemplo, geolocalizaciones, imágenes, audios, texto, etc.



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 alguna fuente confiable 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. Al menos en el punto de la recolección.
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

Una más.

Recientemente y ya desde un punto de vista mas orientado al negocio, nos encontramos con muchos autores que hablan acerca del valor.

Si bien utilizamos software Open Source, debemos ser consientes que detrás de toda esta infraestructura, existe un gran costo, por ello es necesario asegurarnos que el proyecto que encaramos sea rentable para la compañía y que no se quede solamente en una mejora técnica, que sin duda será bienvenida, pero, en definitiva, las ganancias son las que mandan.

Siempre la forma mas efectiva de verificar este tipo de situaciones es mediante mediciones, que dependerán de cada caso en particular, por lo que será necesario generar un caso de negocio (BUSINESS CASE) antes de iniciar cualquier proyecto de Big Data.