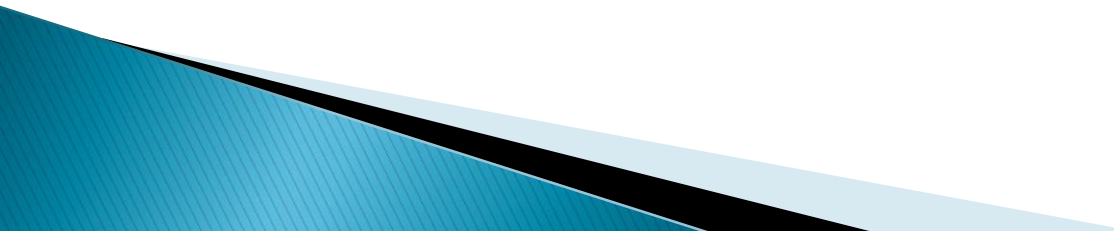


Big Data

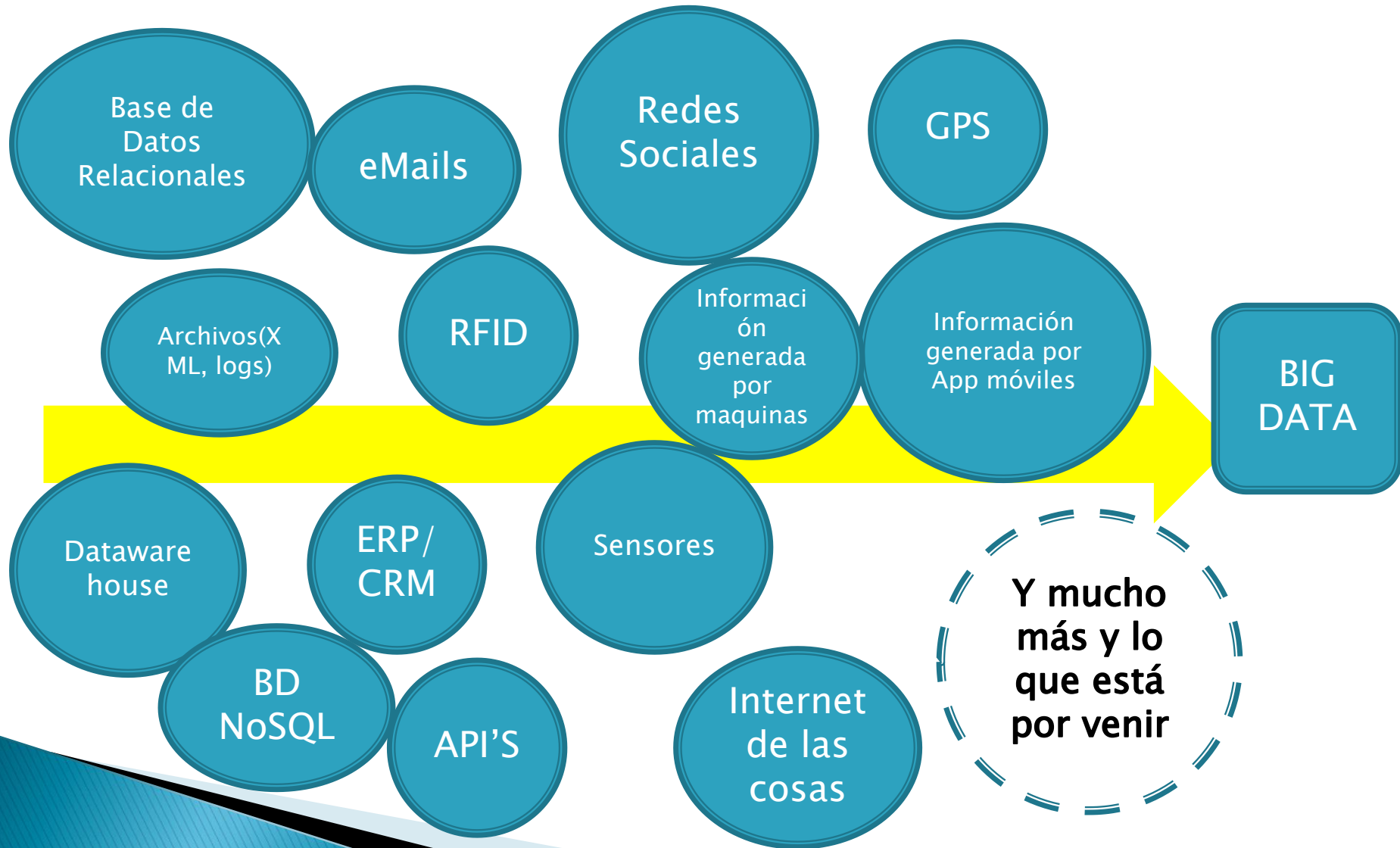
Introducción

- ▶ “Big Data” es desde hacer un par de años una de las grandes tendencias dentro del mundo de la tecnología y del marketing, uno de esos “buzzwords” que en un momento dado empiezan a propagarse y aparecer por todo internet, las grandes empresas se interesan en ello, se crea una industria alrededor y, de repente, todo el mundo sabe lo que es Big Data

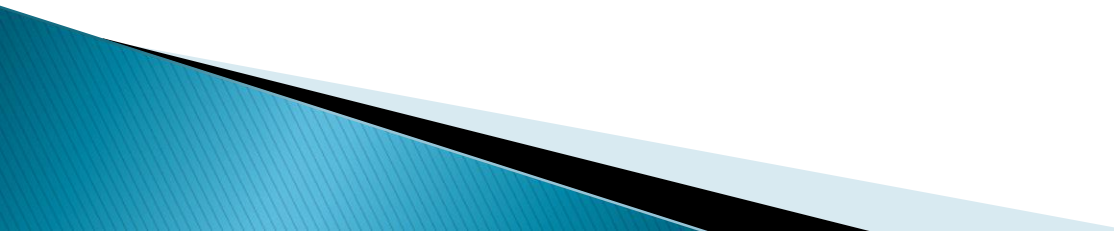
Qué no es Big Data?

- ▶ No es una base de datos enorme
 - ▶ No es un datawarehouse enorme
 - ▶ No es una nueva forma de Business Intelligence
 - ▶ No es llevar las base de datos a la nube
- 

Qué es Big Data?



Definiciones encontradas en Internet

- ▶ Información que tiene un orden de magnitud más grande de lo que estamos acostumbrados.
 - ▶ Información que es muy grande y no se ajusta a las estructuras de las bases de datos actuales.
 - ▶ Es un conjunto de datos cuyo tamaño está más allá de la capacidad de la mayoría de los software utilizados para capturar, gestionar y procesar la información dentro de un lapso tolerable de tiempo.
- 

Las 3 V's

► Volumen

- **Grandes volúmenes de información**
- Se está pasando de hablar en Gigabytes o Terabytes a tamaños de datos de Petabytes, Exabytes o Zettabytes. Volúmenes que se nos escapan

Las 3 V's

► Variedad

- Información de tipos muy diversos
- Ya no solo tenemos información estructurada en Bases de Datos o Archivos.
- Ahora empezamos a tener información con tipos diferentes y totalmente desestructurada

Las 3 V's

▶ Velocidad

- **Velocidad con la que se genera la información**
- La velocidad a la que se genera esta información hace imposible gestionarla con sistemas de base de datos convencionales. Las empresas y las personas ya no quieren estar al día, quieren “estar al segundo”.

Retos Actuales

- ▶ **Dar sentido al gran volumen de datos**
 - Necesitamos las herramientas adecuadas para dar sentido de la abrumadora cantidad de datos generados por la disminución de los costos de hardware y de las fuentes de datos “complejas”.

Retos actuales

- ▶ **La comprensión de una variedad cada vez mayor de datos**
 - Debemos poder analizar datos tanto relacionales como no relacionales. Más del 85% de los datos capturados son desestructurados.

Retos actuales

- ▶ **Habilitación de análisis en tiempo real de los datos**
 - Los nuevos grandes generadores de datos (Twitter, Facebook, ...) están produciendo volúmenes de datos sin precedentes y en tiempo real, lo que no se puede analizar eficazmente mediante procesos por lotes normales.

Utilidades

► Toma de decisiones

- Tomar decisiones en base a datos empíricos y tendencias
- Tomar decisiones en base a corazonadas, instinto o experiencias pasadas

Utilidades

► **Transparencia**

- Compartir y hacer accesible grandes volúmenes de datos a las partes interesadas y de manera oportuna puede crear un enorme valor y aumentar la eficiencia.

Utilidades

► Experimentación

- Una vez recopilados los datos que nos interesan, la experimentación y la exploración de los mismos puede mostrarnos información que a primera vista nunca hubiésemos encontrado o que nunca se nos hubiese ocurrido buscar.

Utilidades

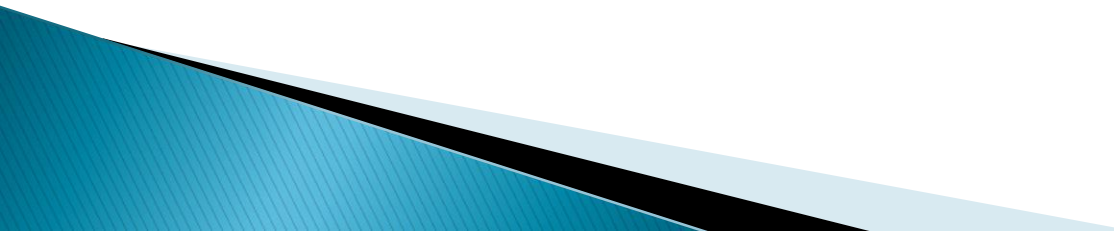
► Innovación

- Permite crear nuevos productos y servicios, mejorar los existentes e, incluso, crear nuevos modelos de negocio..

HADOOP

- ▶ Es una plataforma diseñada para almacenar y analizar grandes volúmenes de datos de diferentes tipos. Basada en Google Map/Reduce y Google Filesystem.

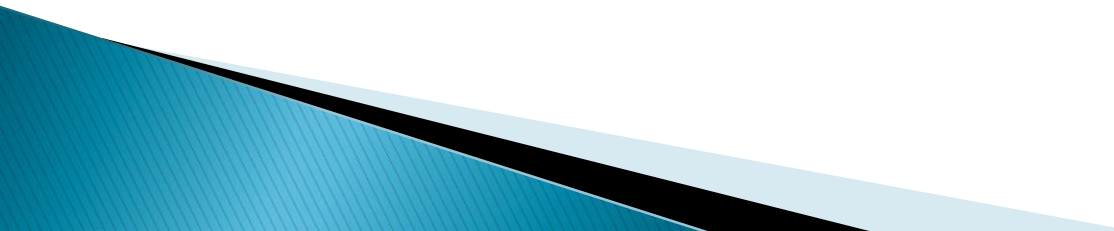
Map/Reduce

- ▶ **Map:** trabajos desarrollados por nosotros. Se distribuyen las tareas en diferentes nodos y se ejecutan en paralelo. Esto genera una información intermedia.
 - ▶ **Reduce:** fusiona la información intermedia y se la ofrece al usuario.
- 

HDFS (Hadoop Distributed File System)

- ▶ **HDFS (Hadoop Distributed File System):**
Sistema de archivos distribuidos, con replicación automática y optimizado para lectura. Cada fichero se partición y se distribuye en todos los servidores.

Otros proyectos alrededor de Hadoop

- ▶ **Hive:** Data Warehouse sobre Hadoop con lenguaje HiveQL (“SQL”).
 - ▶ **Pig:** Lenguaje de script para consulta y análisis de la información. Desarrollado por Yahoo!.
 - ▶ **Sqoop:** Framework para la integración de bases de datos relacionales.
 - ▶ **Flume:** Servicio para recolectar, agregar y mover grandes volúmenes de datos de eventos/logs.
- 

Arquitectura de un sistema BigData



Para experimentar

- ▶ Cloudera VM
 - www.cloudera.com

Gracias por la atención!