UNIVERSIDAD COMPLUTENSE DE MADRID

ntic master School

Tecnologías Big Data y Transformación Digital

Dr. Pablo J. Villacorta

Septiembre de 2020

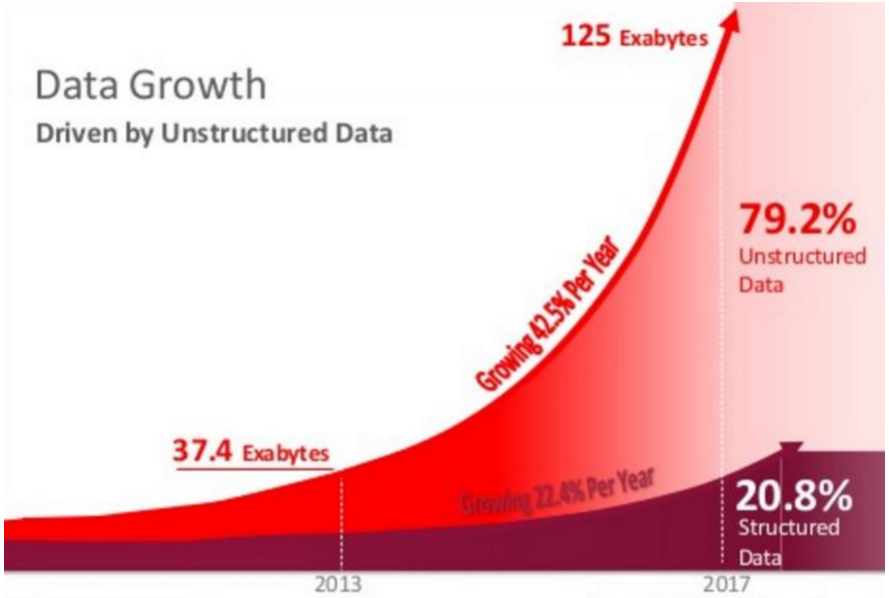
1. Introduction
2. Big Data and its role in Digital Transformation
3. Data Science
4. Deep Learning
5. Artificial Intelligence
6. Distributed Processing

2 Big Data and its role in Digital Transformation

BIG DATA EN TRANSFORMACIÓN DIGITAL

¿POR QUÉ TRANSFORMACIÓN DIGITAL?

- El 90 % de los datos actuales se generó en los últimos 2 años
- 1 Exabyte = 1 M Terabytes



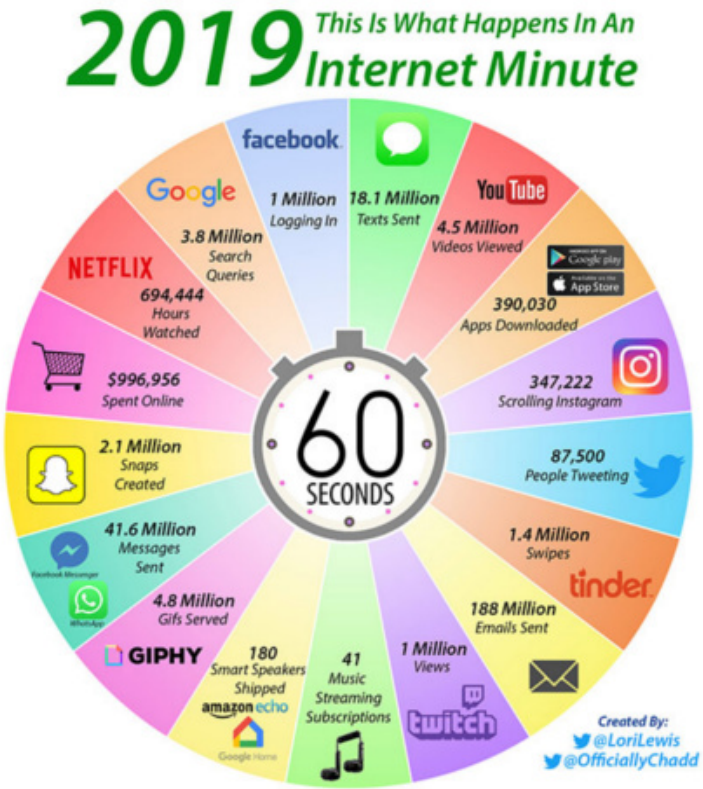
Source: ISD - 2014, Structured Data vs. Unstructured Data: The Balance of Power Continues to Shift

Generados por personas

Generados por máquinas (Internet of Things, logs, sensores…)

LOS DATOS HABLAN DE PERSONAS = CLIENTES

¿POR QUÉ TRANSFORMACIÓN DIGITAL?



~~EL MUNDO ESTÁ CAMBIANDO~~

EL MUNDO HA CAMBIADO

Compañía que mueve más personas	UBER	(Cars = 0)
Compañía que reserva más habitaciones	airbnb	(Hoteles = 0)
Compañía que vende más música	Spotify	(Estudios de grabación = 0)
Compañía que vende más películas	NETFLIX	(Estudios = 0)

Hay más interacciones digitales que físicas con las compañías → **Generan datos masivos**

Los clientes evolucionamos más rápido que las empresas → **¿Qué quieren mis clientes?**

Existe un hueco entre compañías físicas y digitales → **TRANSFORMACIÓN DIGITAL**

OBJETIVOS

1. Centrada en el cliente (customer-centric) : predecir las necesidades del cliente, mejorar su experiencia → **Analizando datos históricos de interacciones de clientes**

2. Canales digitales (sobre todo móviles) → **Interacciones digitales : muchísimos datos**

3. Decisiones guiadas por los datos (inteligencia del dato): **(Big) Data Science**

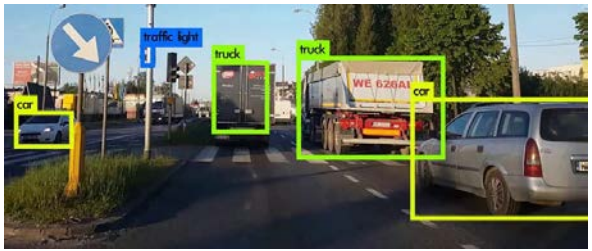
Las tres V's del Big Data: Volumen, Variedad, Velocidad

- ~~*Un proyecto es Big Data cuando tiene alguna de las tres V's*~~
- *Un proyecto es Big Data cuando la mejor manera de resolverlo implica tecnologías Big Data*

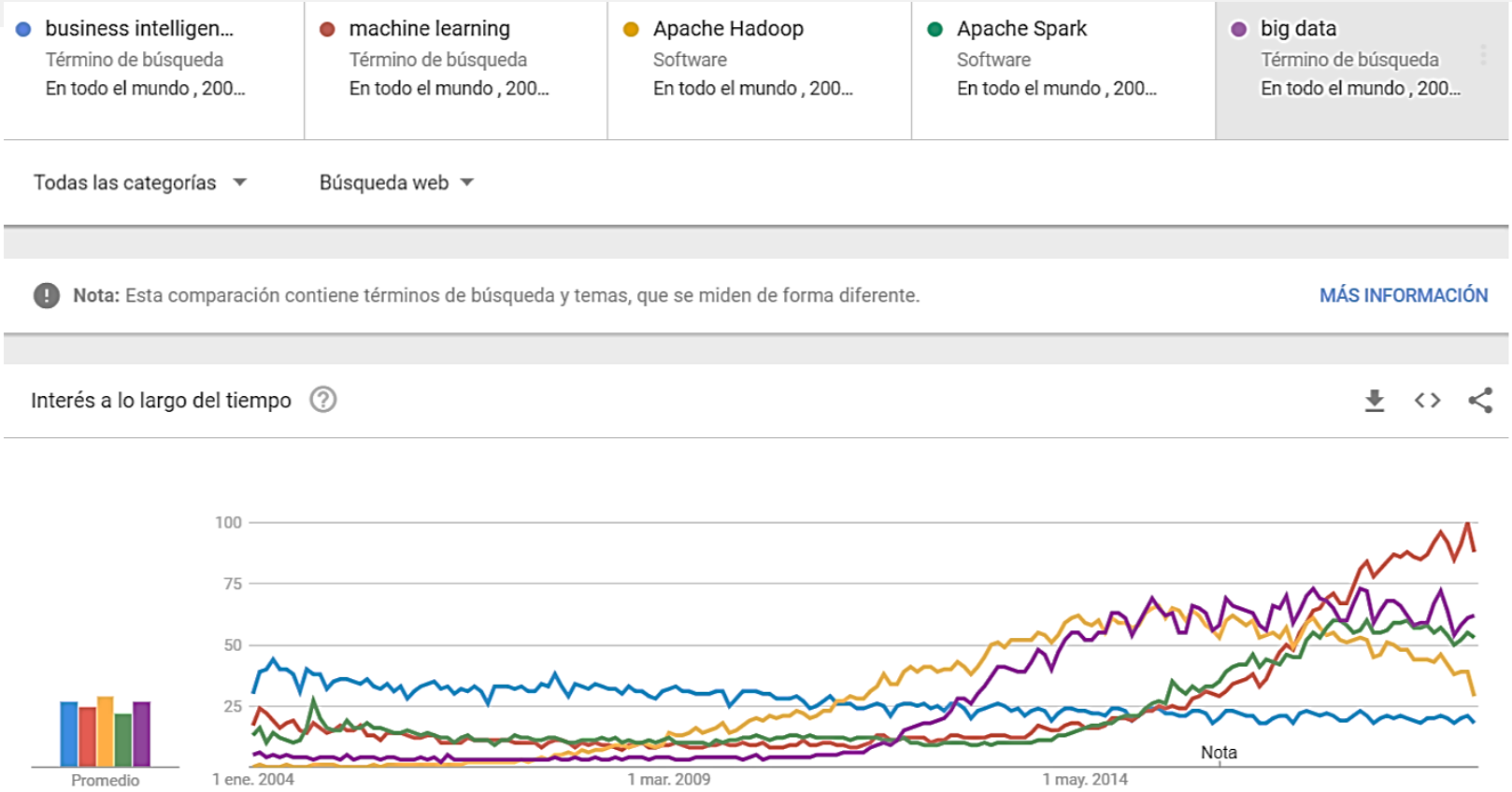
QUÉ ES Y QUÉ NO ES BIG DATA?

- Conjunto de ***tecnologías*** y ***arquitecturas*** para *almacenar, mover, acceder y procesar (incluyendo analizar)* datos que eran imposibles o muy difíciles de manejar antes
  - Grandes cantidades de datos inimaginables hace años
  - Fuentes diversas, heterogéneas, poco estructuradas (documentos, media)
  - Datos que llegan en tiempo real y hay que analizar al vuelo (streaming)

- La definición no incluye nada como *algoritmos, inteligencia, data science* ni nada relacionado con ***qué hacer / cómo analizar y explotar los datos***. Ni fantasías como:



ORIGEN DE LAS TECNOLOGÍAS BIG DATA



ALGUNAS TECNOLOGÍAS BIG DATA

Distributed processing frameworks

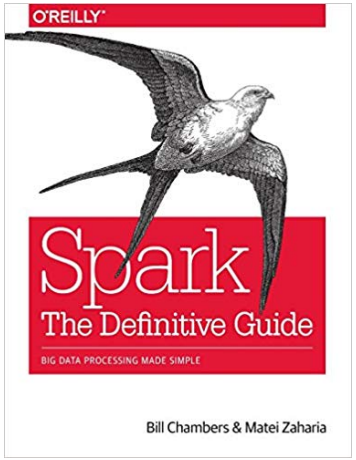
Distributed NoSQL datastores

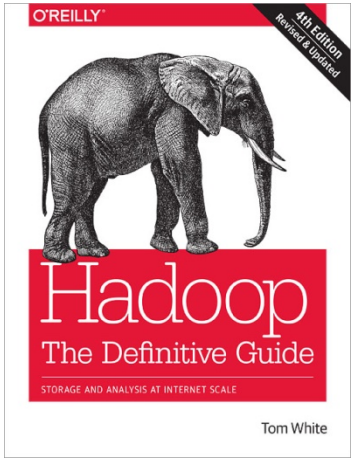
Cloud providers
(infrastructure & more)

Distributed File Systems

Distributed Storage

RECURSOS PARA APRENDER





3 Procesamiento Distribuido

MOTIVACIÓN DE LAS TECNOLOGÍAS BIG DATA

- Grandes cantidades de datos que una sola máquina no puede almacenar ni procesar
  - Procesamiento distribuido entre varias máquinas (clúster), cada una no necesariamente muy potente (***commodity hardware***)
  - Si se necesita más capacidad (datos, memoria o CPU) se añaden nodos
- Datos no estructurados (imágenes, vídeo, documentos) que las BBDD relacionales no pueden manejar
  - Solución: BBDD NoSQL (Hadoop ya incluye una: Apache HBase)

MOTIVACIÓN DE LAS TECNOLOGÍAS BIG DATA

Compañía	Nodos
Yahoo!	42000
LinkedIn	4100
Facebook	1400
NetSeer	1050
Ebay	532
CRS4	400
Powerset / Microsoft	400
…	
Spotify	120

Clusters de Hadoop en grandes empresas

Mare Nostrum 4
Barcelona Supercomputing Center (CSIC)

HISTORIA DE HADOOP Y SPARK

- Google (C++) - Almacenamiento + procesamiento usando commodity hardware
  - 2003 - Google File System (GFS). http://static.googleusercontent.com/media/research.google.com/es//archive/gfs-sosp2003.pdf
  - 2004 - Map Reduce (Simplified Data Processing on Large Clusters). http://static.googleusercontent.com/media/research.google.com/es//archive/mapreduce-osdi04.pdf
- Apache Hadoop (Java)
  - 2002, Doug Cutting desarrolla Nutch. 2006, Hadoop se independiza de Nutch
  - 2008, se hace open-source (incluye una implementación abierta de MapReduce)
  - Adoptado en grandes empresas de todo el mundo a partir del año 2011
- Apache Spark (Scala) – Motivado por procesos iterativos (Machine Learning)
  - 2009 - Matei Zaharia comenzó el proyecto en UC Berkeley's AMPLab
  - 2010 - Open Source
  - 2014 - Forma parte de Apache 2.0. Top Level Project
  - 2015 - Más de 1000 contributors
  - 2016+ La mayoría de clústeres de Hadoop son migrados a Spark.

HISTORIA DE HADOOP Y SPARK

Yahoo dona Hadoop a la Apache Software Foundation (ASF)

Doug Cutting adds DFS & MapReduce support to Nutch

Doug Cutting & Mike Cafarella started working on Nutch

Fastest sort of a TB, 3.5mins over 910 nodes

Fastest sort of a TB, 62secs over 1,460 nodes
Sorted a PB in 16.25hours over 3,658 nodes

NY Times converts 4TB of image archives over 100 EC2s

2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009

Google publishes GFS & MapReduce papers

Yahoo! hires Cutting, Hadoop spins out of Nutch

Facebooks launches Hive: SQL Support for Hadoop

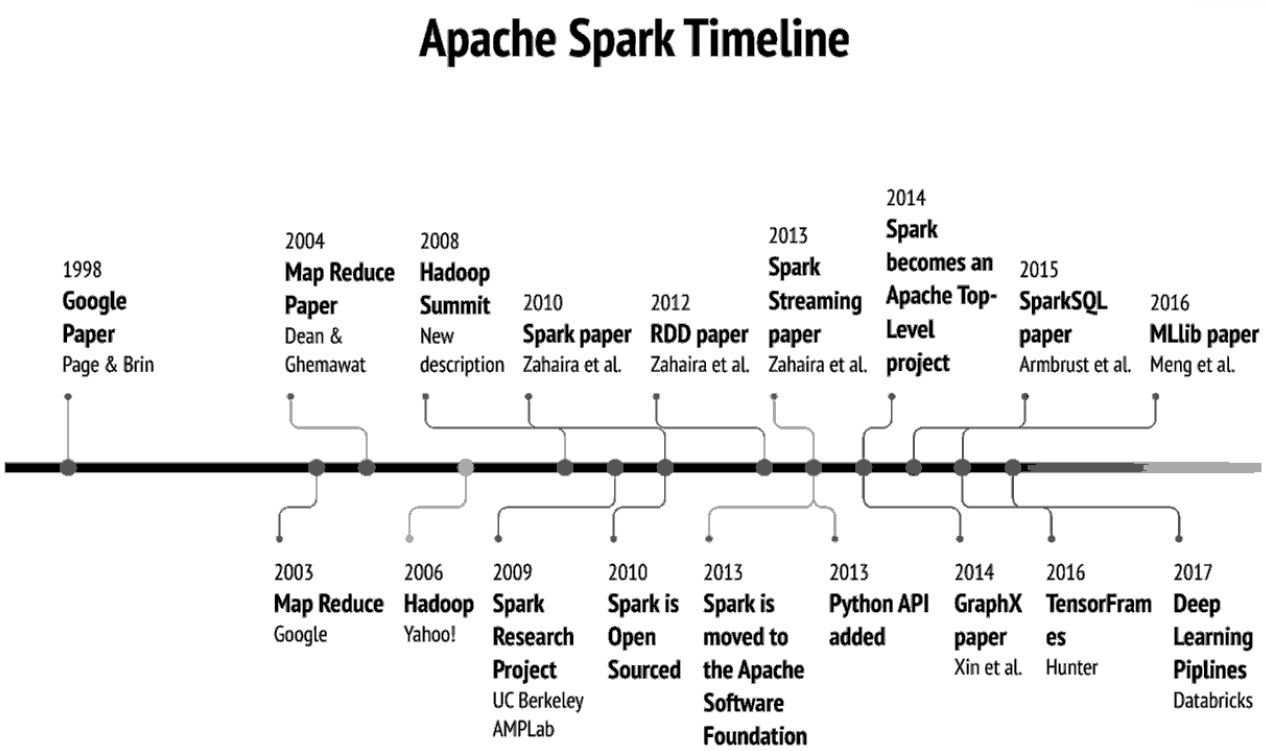
cloudera Founded

Doug Cutting joins Cloudera

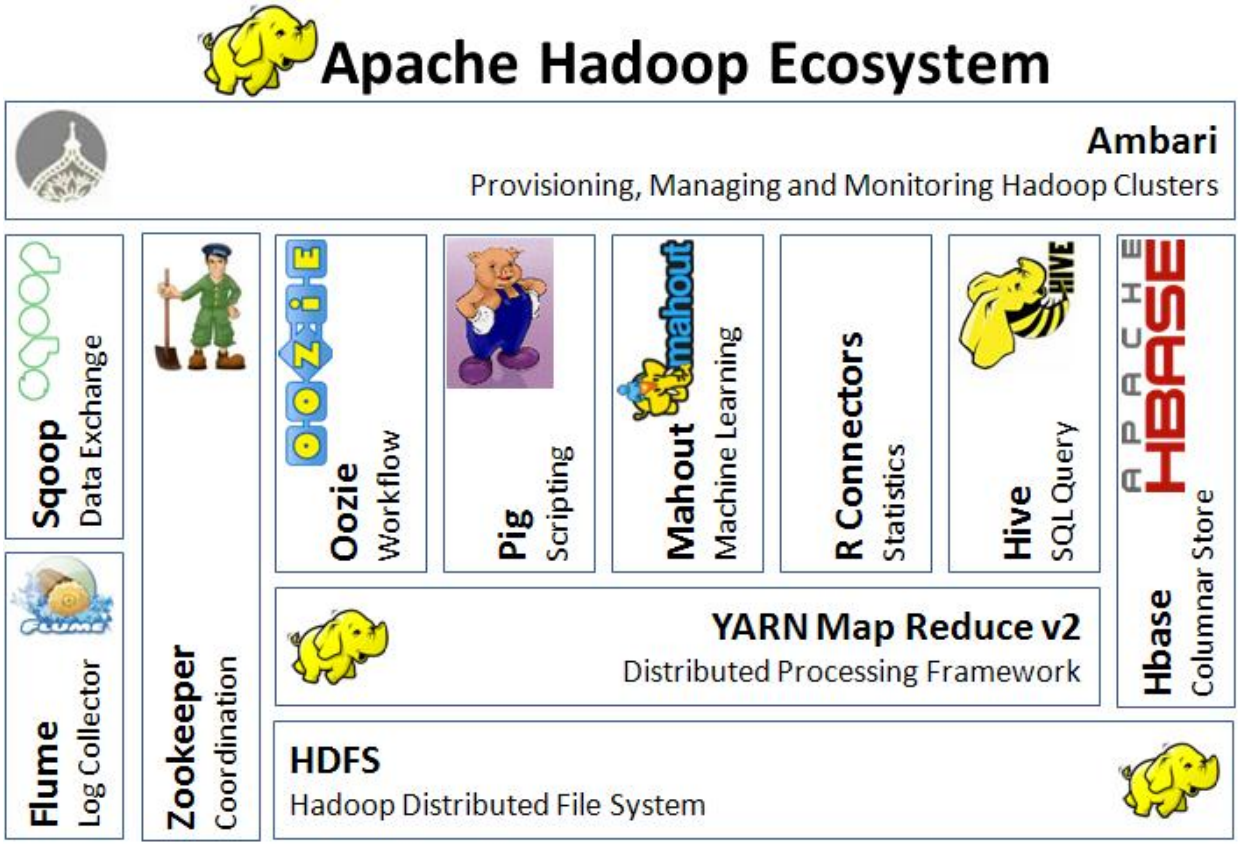
Hadoop Summit 2009, 750 attendees

Source: Cloudera, Inc.

HISTORIA DE HADOOP Y SPARK



HISTORIA DE HADOOP Y SPARK



COMPONENTES PRINCIPALES DE HADOOP

- Hadoop: proyecto de software libre, con licencia Apache, cuya finalidad es prestar una plataforma para la gestión de grandes cantidades de datos.

- 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

- MapReduce: manera de programar y también, motor de ejecución de tareas que corren de forma distribuida en los diferentes nodos del clúster Hadoop. La forma en la que los datos se distribuyen en diferentes subtareas y cómo estas se asignan a cada máquina resulta transparente para el desarrollador.

- Spark: motor de ejecución **en memoria** de tareas que corren de forma distribuida en los diferentes nodos del clúster. La forma en la que los datos se distribuyen en diferentes subtareas y cómo estas se asignan a cada máquina resulta transparente para el desarrollador.

¿Preguntas?

UNIVERSIDAD
COMPLUTENSE
DE MADRID