

# Ingeniería para el Procesado Masivo de Datos

Dr. Pablo J. Villacorta

## Tema 1. Introducción a las Tecnologías Big Data

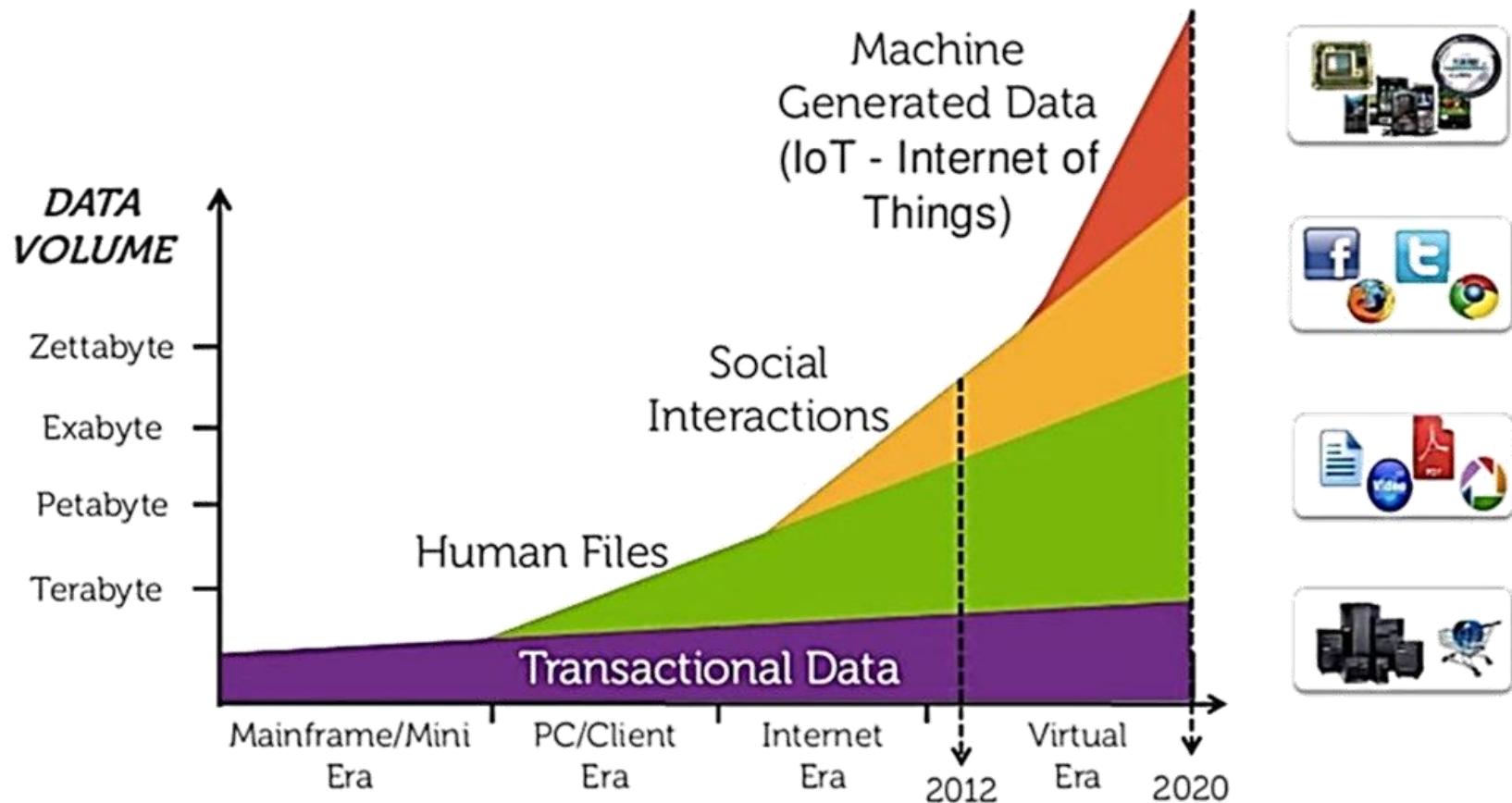
Noviembre de 2025

# Objetivos del tema

- ▶ 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.

# Motivación de las tecnologías Big Data

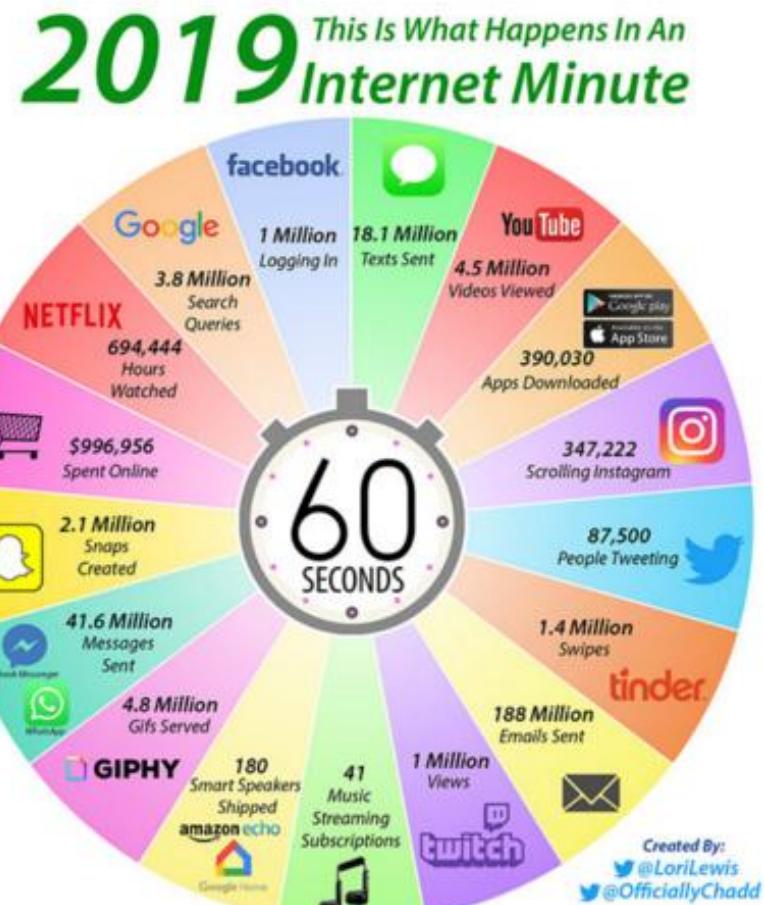
- ▶ Vivimos en una sociedad **interconectada**: la **era del cliente**
- ▶ El 90 % de los datos existentes se generaron en los últimos 2 años



# Fuentes (orígenes) de datos en la actualidad

- ▶ **Interacciones persona – persona**
  - ▶ Email (texto), redes sociales (imágenes, videos, texto), foros de internet (texto)
  - ▶ Datos no estructurados al contener la vida cotidiana de las personas
- ▶ **Interacciones persona – máquina**
  - ▶ Navegación en internet:
    - ▶ Logs generados cuando navegamos por sitios web
    - ▶ Datos generados cuando hacemos transacciones bancarias, compras online (viajes, entradas, hoteles, Amazon).
    - ▶ Representan interacciones digitales de un cliente con una empresa
  - ▶ Datos estructurados o semi-estructurados
- ▶ **Interacciones entre máquinas**
  - ▶ Datos generados por sensores (Internet of Things, IoT)
  - ▶ Datos perfectamente estructurados al estar generados por máquinas

# Motivación de las tecnologías Big Data



# Motivación de las tecnologías Big Data

## 2019 *This Is What Happens In An Internet Minute*



## 2020 *This Is What Happens In An Internet Minute*



# Motivación de las tecnologías Big Data

## 2020 *This Is What Happens In An Internet Minute*



## 2021 *This Is What Happens In An Internet Minute*



# Motivación de las tecnologías Big Data



# La Transformación Digital en torno al dato

- ▶ Más interacciones digitales que físicas entre personas y sus compañías
  - ▶ Energía, empresas de telefonía, tiendas online, banca online
  - ▶ Cada interacción genera **datos** que hablan de los **clientes** (nosotros)
  - ▶ **¡Nosotros** evolucionamos más rápido que las compañías!
    - ▶ Hueco entre las compañías tradicionales y nuestra forma de vivir
- ▶ **Transformación Digital** para adaptarnos al nuevo cliente:
  - ▶ Centrarse en el cliente (*customer centricity*): mejorar su experiencia, prever sus necesidades y **su comportamiento**: analizando sus datos!
  - ▶ Centrarse en lo que ocurre en los **canales digitales**: genera datos!
  - ▶ Decisiones **guiadas por los datos** (convertirse en **data driven**)
    - ▶ **(Big) Data Science**: los clientes generan **muchísimos** datos

# Motivación de las tecnologías Big Data

- ▶ En la **era del cliente**, éste genera **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)

# ¿Qué es un proyecto Big Data?

- ▶ Un proyecto de datos se considera Big Data cuando implica alguna de las tres V's
  - ▶ **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 BBDD relacionales, no relacionales, datos de imágenes, sonido, etc.) que tienen que ser manejados y cruzados de manera conjunta.

# ¿Qué es un proyecto Big Data?

- ▶ Una definición mejor sería:
  - ▶ **Un proyecto es big data cuando la mejor manera de resolverlo (más rápida, eficiente, sencilla) implica utilizar tecnologías big data**

Causas de esa imposibilidad:

- ▶ 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*).

**Tecnologías Big Data:** 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.

# Aspecto de un cluster en la actualidad

*Mare Nostrum 4  
Barcelona  
Supercomputing  
Center (CSIC)*



# Top 500 de clusters más potentes (Junio 2021)

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	<b>Supercomputer Fugaku</b> - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	<b>Summit</b> - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	<b>Sierra</b> - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	<b>Sunway TaihuLight</b> - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
5	<b>Perlmutter</b> - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	706,304	64,590.0	89,794.5	2,528

# Top 500 (Nov 2023 y Junio 2024)

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	<b>Frontier</b> - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	<b>Aurora</b> - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	4,742,808	585.34	1,059.33	24,687
3	<b>Eagle</b> - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Microsoft Azure United States	1,123,200	561.20	846.84	
4	<b>Supercomputer Fugaku</b> - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
5	<b>LUMI</b> - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107

# Top 500 de clusters más potentes (Nov 2023)

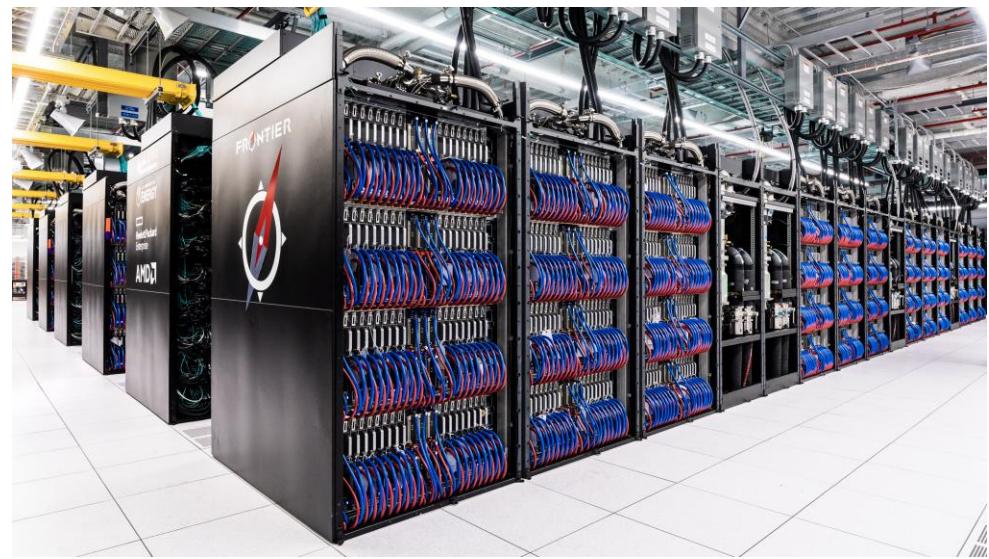


Superordenador Fugaku, cuyo desarrollo empezó en 2014

- Terremotos, tsunamis
- Nuevos materiales
- COVID-19



Portátil Yungai Rev-9 con procesador AMD EPYC 7713 de 64 cores



Superordenador Frontier

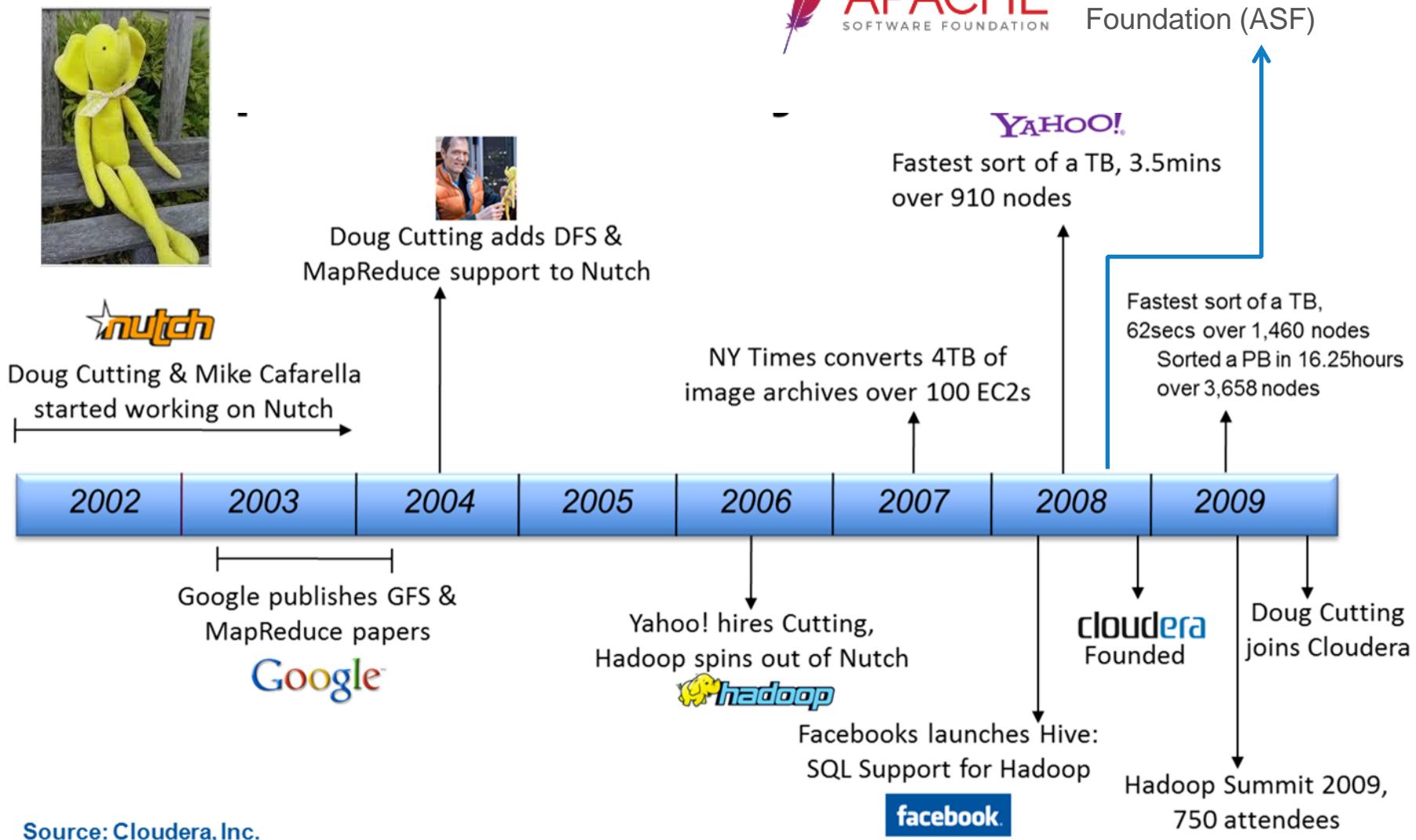
# Historia de Hadoop y Spark

- ▶ **Google (C++) - Almacenamiento + procesamiento usando commodity hardware**
  - 2003 - Google File System (GFS) – **el germen de HDFS**  
<https://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 (era su tesis doctoral en UC Berkeley, grupo 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

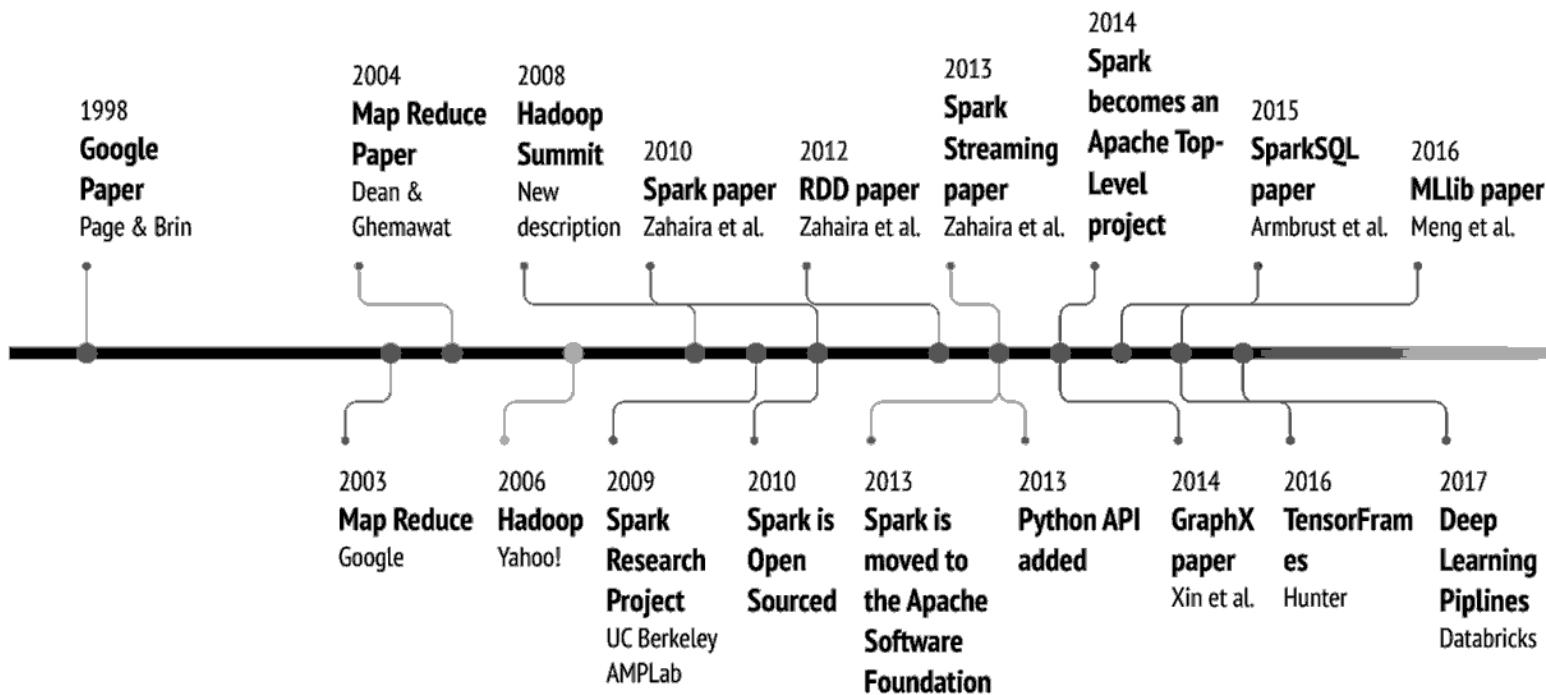


Yahoo dona Hadoop a la Apache Software Foundation (ASF)



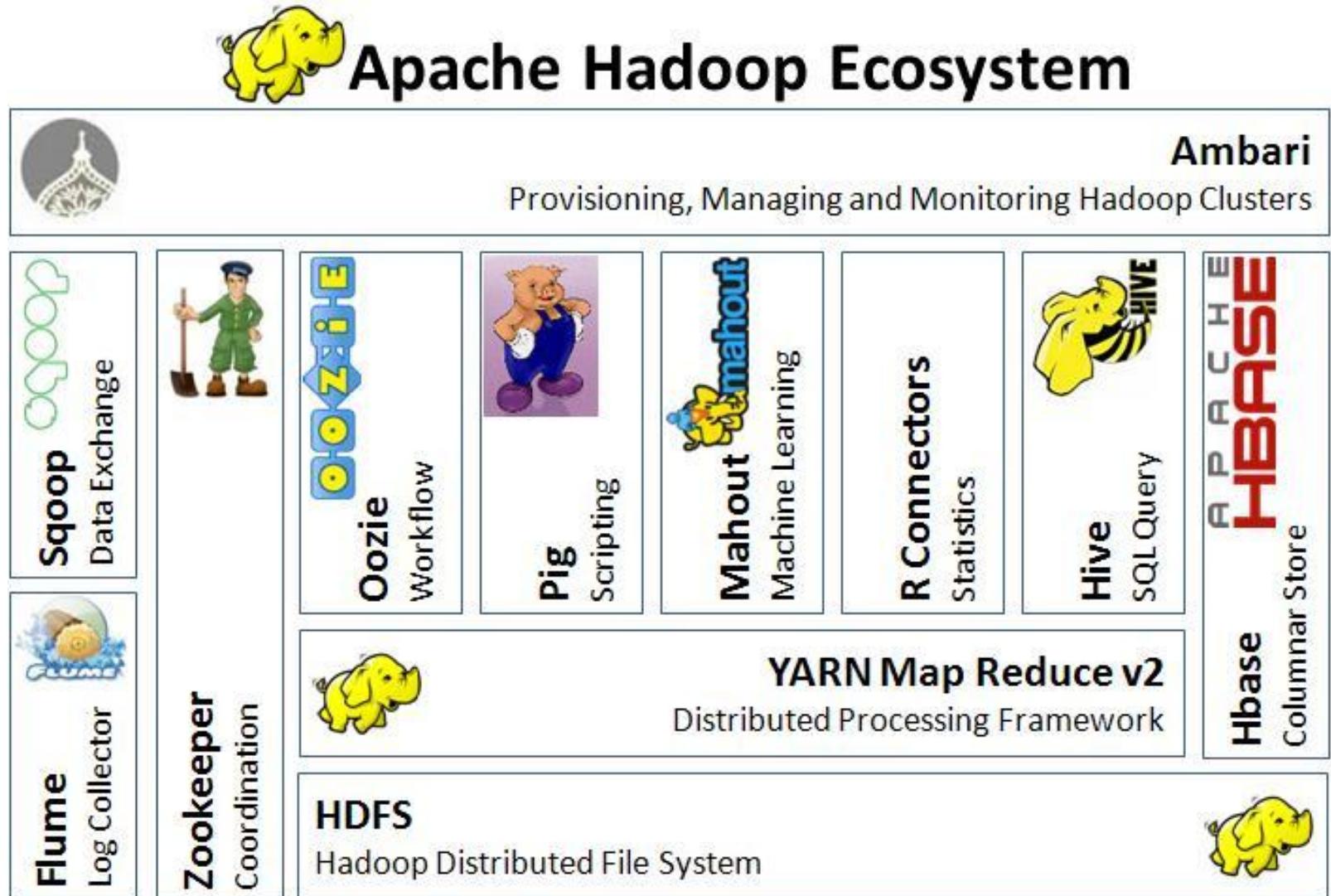
# Historia de Hadoop y Spark

## Apache Spark Timeline



By Favio Vázquez

# Historia de Hadoop y Spark



# Componentes de Hadoop

**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

**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 para ser consumidos por uno o varios procesos externos (por ejemplo trabajos de Spark).

# Distribuciones de Hadoop

- ▶ Todos los componentes de Hadoop se puede descargar e instalar de forma independiente (requiere configuración posterior)
- ▶ Distribución de Hadoop: producto software con todos los componentes de Hadoop pre-instalados en versiones que inter-operan bien entre ellas, y con soporte adicional.
- ▶ Sandbox: máquina virtual que emula un sistema operativo con el software pre-instalado, listo para ejecutar sin requerir instalación.

Característica	Cloudera	Hortonworks	MapR
Componentes	Apache modificados y añadidos	Sólo 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: VM Ware )	Linux y Windows	Linux (Windows: VM Ware)
Año de creación	2008	2011	2009
Observaciones	Es la más extendida. Certificación muy popular.	Única para Windows, y única 100 % open-source	La más rápida y fácil de instalar

...eso es todo por hoy...  
:-)

Cualquier duda, consulta o comentario:  
mensaje a través de la plataforma!

**unir**  
LA UNIVERSIDAD  
EN INTERNET

[www.unir.net](http://www.unir.net)