

# 0. INTRODUCCIÓN BÁSICA A BIG DATA

Procesamiento de Datos

IES Clara del Rey

# Índice

- Objetivos ..... 2
- Los antecedentes a Big Data ..... 4
- Qué es Big Data ..... 15
- Qué aporta Big Data a la empresa ..... 21
- La plataforma Big Data ..... 24
- Hadoop: el núcleo de Big Data ..... 31
- Soluciones Big Data ..... 44
- Casos de uso Big Data ..... 48

# Objetivos

- Entender de qué hablan cuando dicen “Big Data” Comprender qué es Big Data: por qué nació, qué realiza concretamente y por qué su gran difusión
- Entender cómo consigue Big Data las mejoras que aporta
- Impacto de esta tecnología en los negocios
- Poder discernir los roles técnicos que participan en un proyecto Big Data
- Conocer las capas de un proyecto Big Data
- Ser capaz de identificar la finalidad de un componente Big Data y sus relaciones
- Introducir el Ecosistema Big Data y su core tecnológico: Apache Hadoop
- Principales distribuidores de productos Big Data
- Mostrar ejemplos reales de aplicación de Big Data

# PARTE PRIMERA

## **LOS ANTECEDENTES A BIG DATA**

# Antecedentes a Big Data

La informática tradicional: de la calculadora a la computación distribuida

El tratamiento de datos se puede entender como un proceso



# Antecedentes a Big Data: de la calculadora a la computación distribuida

- El Tratamiento de Datos es fundamental en el mundo empresarial.

## El análisis de las facturas

### 1. Recolectar



Facturas día anterior

### 2. Almacenar



Programa informático contabilidad

### 3. Procesar



Indicación datos a calcular

### 4. Visualizar

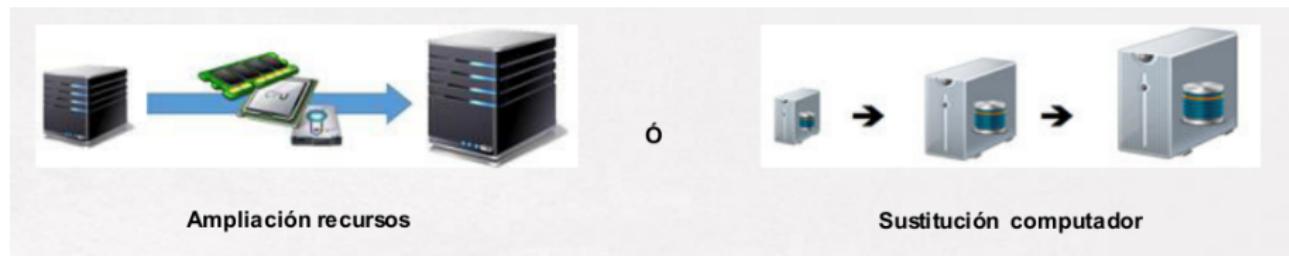


Informes, indicadores, gráficos, ...

# Antecedentes a Big Data: limitaciones de la computación vertical

**Problema:** Aumentan los datos a tratar → aumentan los recursos necesarios

**Solución:** “Crecimiento Vertical” (Escalabilidad Vertical)



# Antecedentes a Big Data: limitaciones de la computación vertical

El Crecimiento Vertical tiene limitaciones:

- El computador no se puede ampliar más.
- No existe un computador tan grande.

**Solución:** Grace Hopper y su *parábola de los bueyes*

# Antecedentes a Big Data: ejemplo de la parábola de los bueyes



# Antecedentes a Big Data: Aparición de la computación distribuida

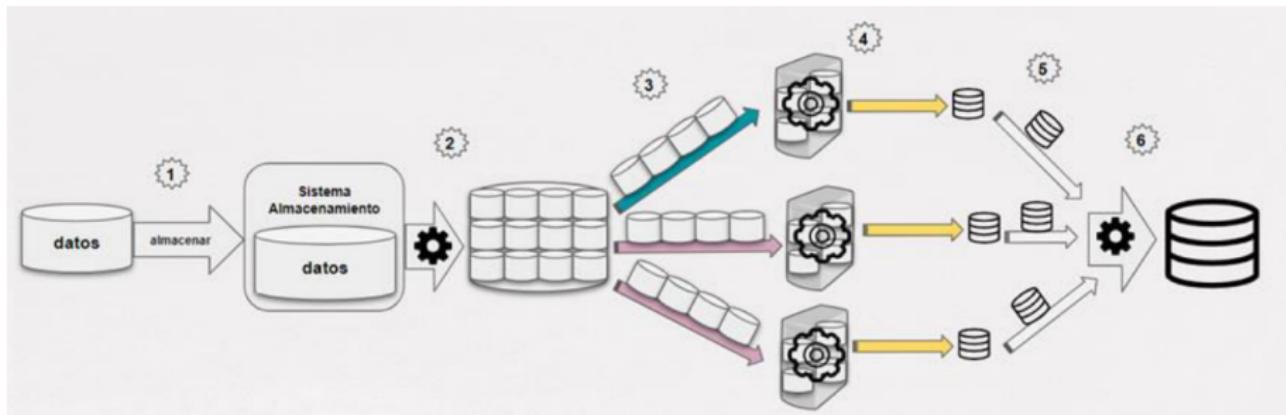
Aparece la “Computación Distribuida” y el “Crecimiento (escalado) Horizontal”



Mayores necesidades → mayor número de computadores

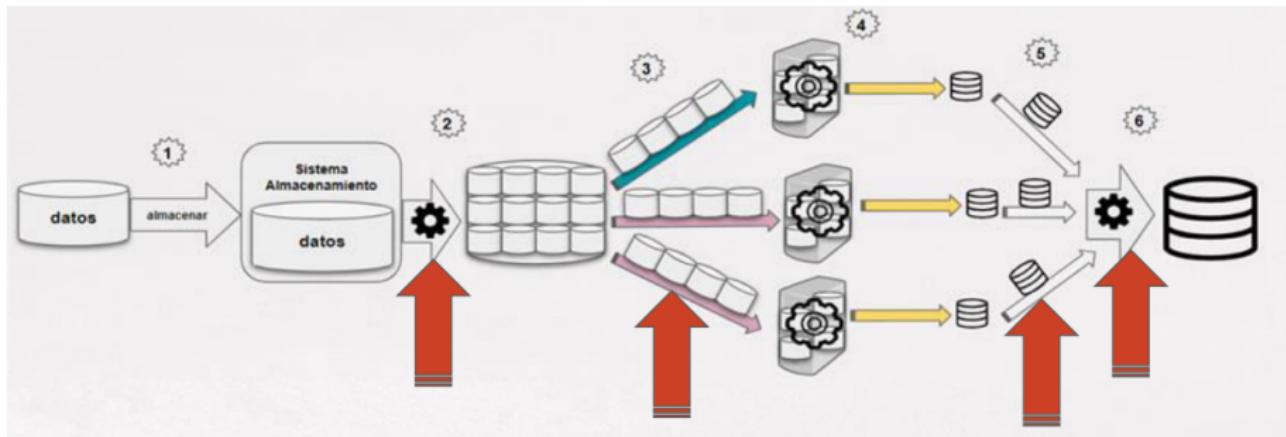


# Antecedentes a Big Data: El tratamiento del dato en la computación distribuida



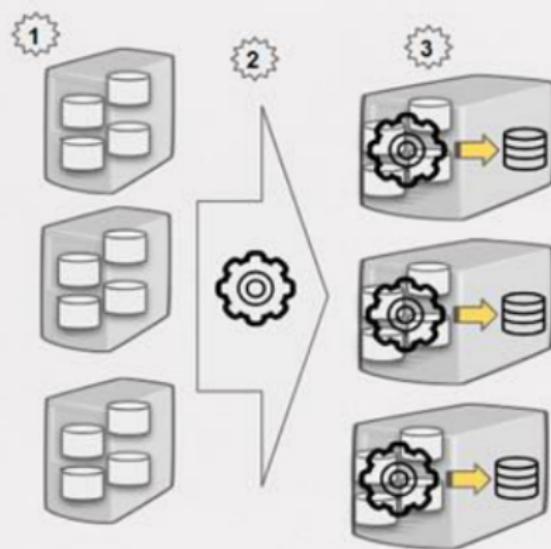
# Antecedentes a Big Data: El tratamiento del dato en la computación distribuida

- El crecimiento no puede ser infinito.
- Aumentan los datos y computadores: Se agrandan los cuellos de botella

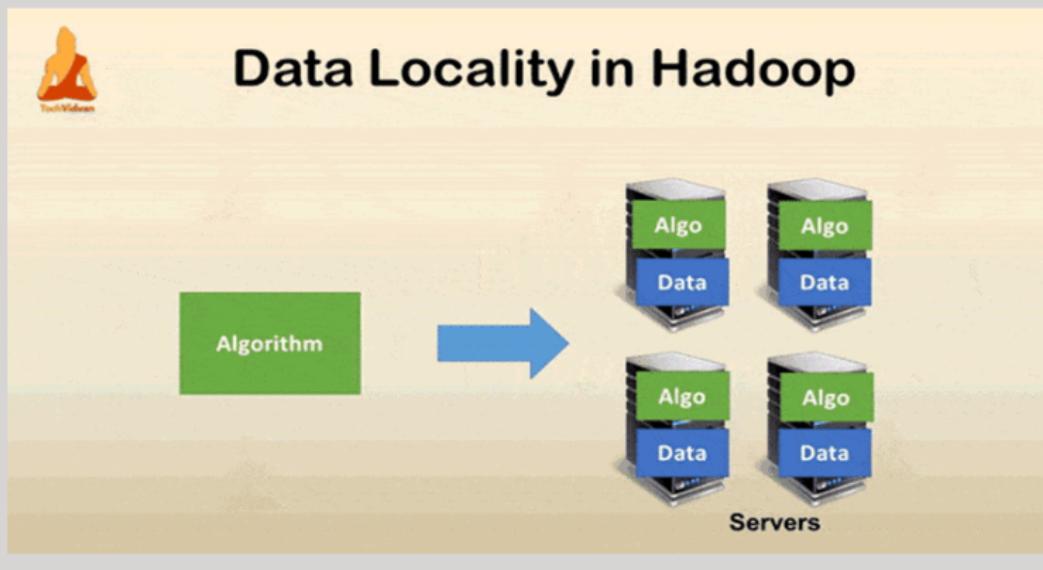


# La clave de Big Data: La localización del dato

- Big Data también es *Computación Distribuida y Almacenamiento Distribuido*
- Big Data realiza el procesamiento allá donde está el dato → “**Data Locality**”



# La clave de Big Data: La localización del dato



# La clave de Big Data: La localización del dato

- Al desaparecer fases, se eliminan los cuellos de botella
- En Big Data, **al almacenar, se distribuyen los Datos**



# La clave de Big Data: La localización del dato

- Big Data aprovecha el “*Crecimiento (escalado) Horizontal*”
- Mayor número de computadoras con mayor almacenamiento y mayor procesamiento



# PARTE SEGUNDA

## QUÉ ES BIG DATA

# Entender qué es Big Data: Son 3 Vs

Las ventajas e innovaciones de Big Data se resumen en 3 V's.



# Entender qué es Big Data: **Volumen**

- Se empieza a hablar de TeraBytes, PetaBytes, ExaBytes, ...
- Los servidores aportan el almacenamiento.
- Un número de servidores ilimitado proporciona almacenamiento ilimitado.

# Entender qué es Big Data: **Velocidad**

- Procesamiento Batch (por lotes) más rápido que lo tradicional.
- **Streaming:** Se pueden procesar los datos según se generan
- Aparece el concepto de **Real Time** (siendo estrictos **Near Real Time**)

# Entender qué es Big Data: **Variabilidad**

- El tratamiento de Datos tradicional utiliza datos estructurados.
- En Big Data se utiliza cualquier dato:
  - **Estructurado**
  - **Semi-estructurado**
  - **Des-estructurado**
- Todo elemento que genere datos es una fuente datos para Big Data.

# Entender qué es Big Data: Hay más V's, compuestas de estas tres

- Al definir Big Data, se pueden encontrar otras V's:
  - **Veracidad**
  - **Valor**
  - **Viabilidad**
  - **Visualización**
- Son evolución de las 3 V's básicas, o son nuevas características.



Volume



Velocity



Variety



Veracity



Value



Variability

# Entender qué es Big Data: Hay más V's, compuestas de estas tres



# PARTE TERCERA

## QUÉ APORTA BIG DATA A LA EMPRESA

## Nuevas oportunidades de negocio

- Gracias a las 3 Vs de Big Data, los Negocios pueden jugar con los datos:
  - Con cualquier tipo de dato.
  - En cualquier momento.
  - Con cualquier cantidad.
- Se dice que hoy en día que **el Dato es el nuevo oro**
- También se dice que “*en Internet cuando algo es gratis, el pago eres tú*”

## Nuevas oportunidades laborales

- Big Data no requiere únicamente perfiles informáticos.
- Con las nuevas oportunidades cobrarán peso todo tipo de perfiles.
- Incluso aparecen nuevos perfiles no técnicos en torno a los datos:  
**Chief Data Officer**

## Perfiles técnicos de trabajo en Big Data

- Administrador / **Big Data Sysadmin**
- Desarrollador / **Data Engineer**
- Analista de Datos / **Data Analytics o Scientist**

# PARTE CUARTA

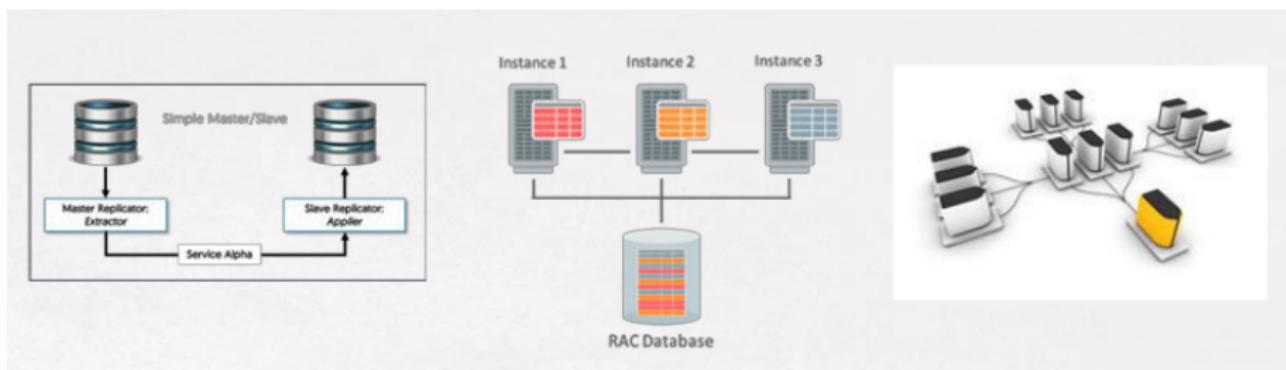
## LA PLATAFORMA BIG DATA

# La plataforma Big Data

- Qué es un Cluster
- Tipos de servidores de un cluster
- Qué se considera una Plataforma Big Data
- Capas de una plataforma Big Data

# Plataforma: Un cluster en Big Data

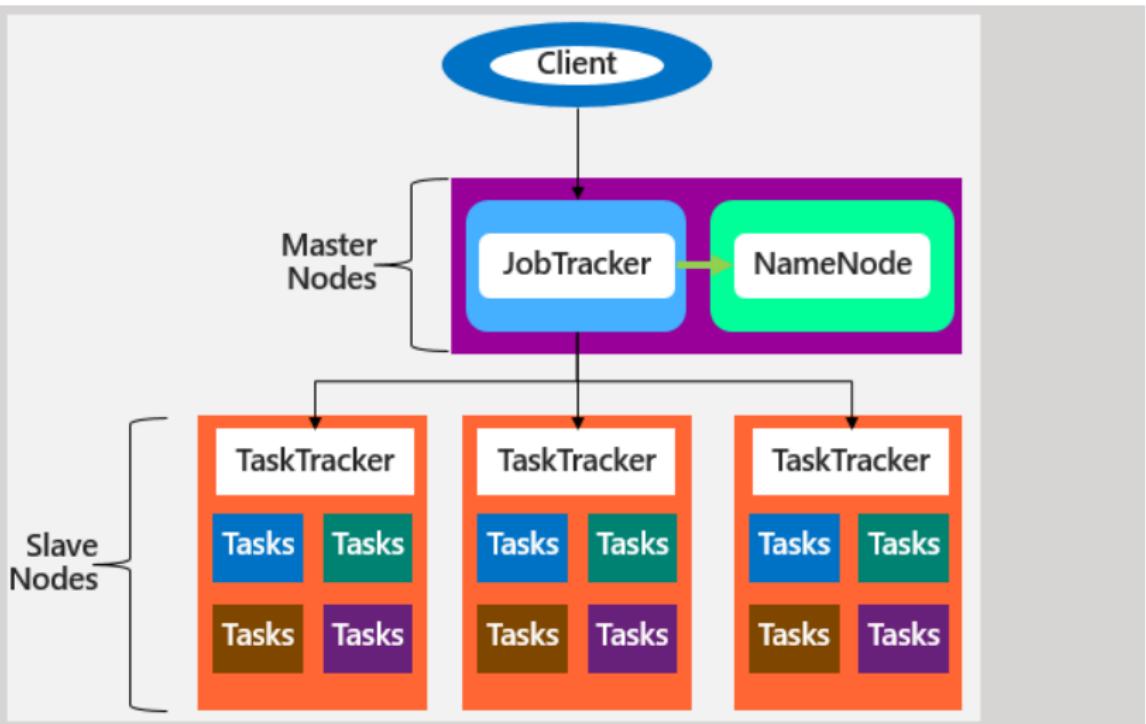
- Un **Cluster** es un conjunto de servidores que trabajan coordinados
- Existen muchos tipos de cluster, según el producto que lo conforme



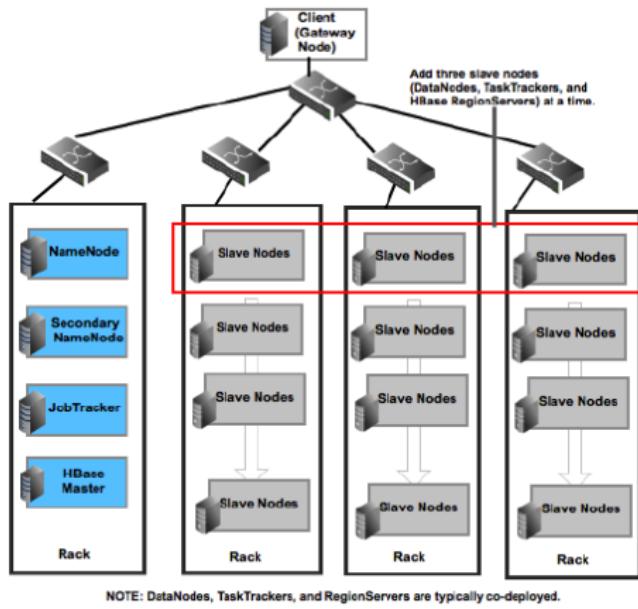
# Plataforma: Tipos de servidores de un Cluster

- Según el producto, el cluster tendrá unos tipos de servidores u otros.
- En Big Data tendremos un Cluster Hadoop
  - **Maestros (Masters)**: coordinan el cluster
  - **Trabajadores (Workers)**: alojan y procesan los datos
  - **Ingesta (Gateway o EdgeNodes)**: conexión del cluster con el exterior
  - **Utilidades (UtilityNodes)**: servicios adicionales
- Al ampliar el número de Workers amplían el procesamiento y almacenamiento del cluster.

# Plataforma: Tipos de servidores de un Cluster



# Plataforma típica de un cluster de Hadoop

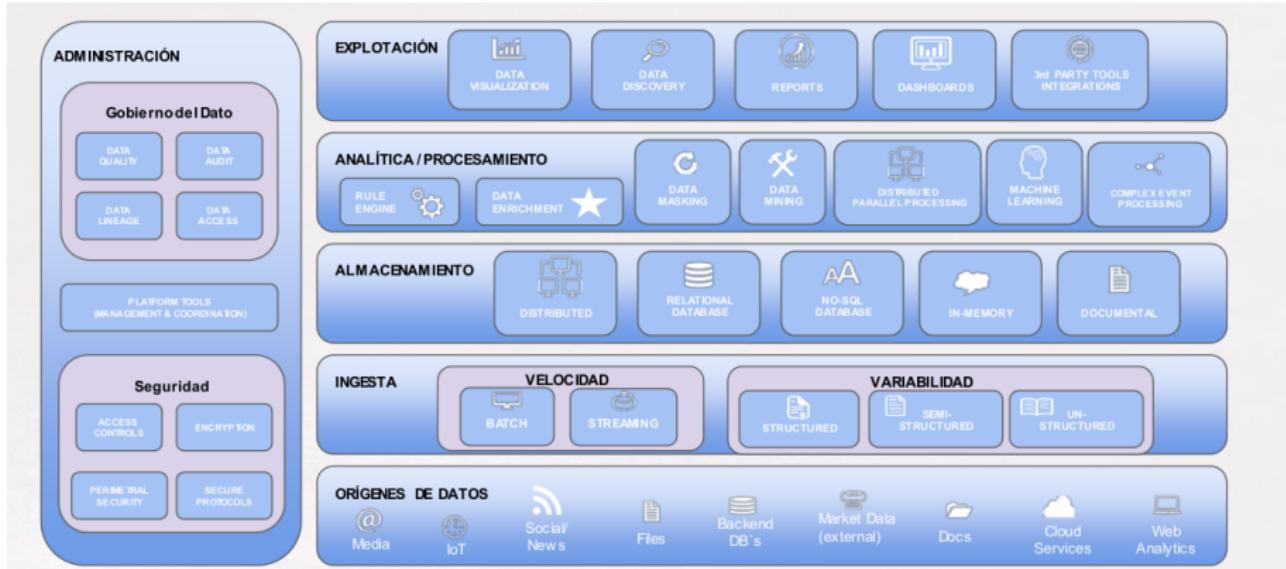


# Qué es una plataforma en un proyecto Big Data

- El **Cluster Hadoop** es el core de una Plataforma Big Data. Pero Big Data no sólo es Hadoop.
- Una **Plataforma Big Data** engloba más tecnologías y se puede considerar un conjunto de capas que agrupan componentes y tecnologías según su finalidad.



# Las capas de una plataforma Big Data



# Ejemplo de capas para Big Data de Bloor Research



# PARTE QUINTA

## HADOOP: EL NÚCLEO DE BIG DATA

# Hadoop como core tecnológico de Big Data

- **Apache Hadoop:** core tecnológico de Big Data
- Historia de Apache Hadoop
- **HDFS y YARN:** Almacenamiento y procesamiento como punto central
- **Ecosistema Hadoop:** diferentes productos para diferentes finalidades
- Ubicar los componentes en sus capas correspondientes
- *Hadoop es al Big Data lo que Linux a los Sistemas Operativos*

# Apache Hadoop: core tecnológico de Big Data

- **Apache Hadoop:** la tecnología que propició cumplir las 3 V's del Big Data.
- P pertenece al proyecto Apache y tiene una de las mayores comunidades activas.



# Historia de Apache Hadoop

- **1997. Lucene** Doug Cutting (D.C.) crea el motor de indexación Lucene.
- **2002. Nutch 1-Mach. Nutch 4-M.**: D. C. crea un buscador distribuido, pero sólo con 4 máquinas.
- **2003. GFS y Map/Reduce**: Google publica cómo almacena y procesa internet (GFS y Map/Reduce)
- **2005. Nutch + Hadoop**: D.C. basado en los whitepapers de Google crea Hadoop.
- **2006 - 2008. Yahoo y Yahoo!!**: Es contratado por Yahoo para desarrollar Hadoop (para adelantar a Google).

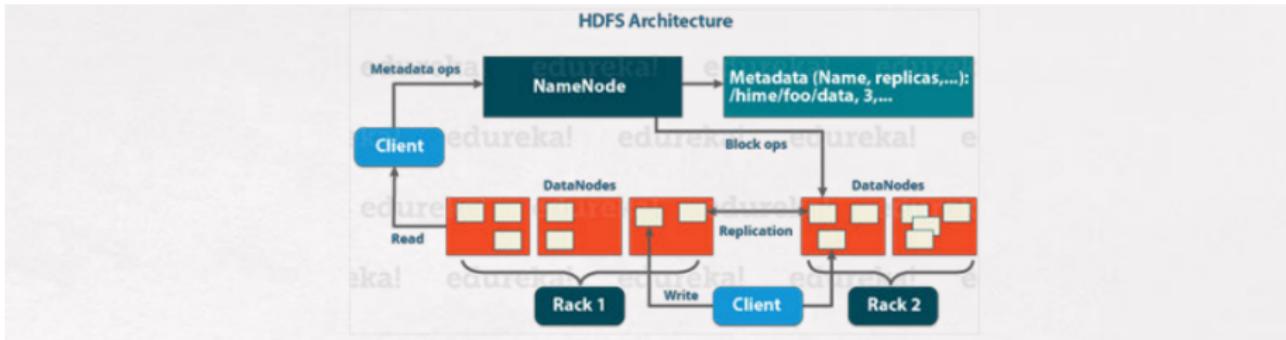
# Historia de Apache Hadoop

- **2009. Cloudera y MapR:** Nacen empresas que dan soporte sobre Hadoop.
- **2010. Hortonworks:** Spin-off de Yahoo de la parte de soporte a Hadoop.
- **2011. Hadoop 2.0 (YARN):** Evolución de Hadoop separando el procesamiento de la gestión de recursos.
- **2012. Spark:** Revolución en el mundo Big Data, el framework de procesamiento estrella hoy en día.

# HDFS y YARN: Almacenamiento y procesamiento como punto central

- **HDFS:** Almacenamiento

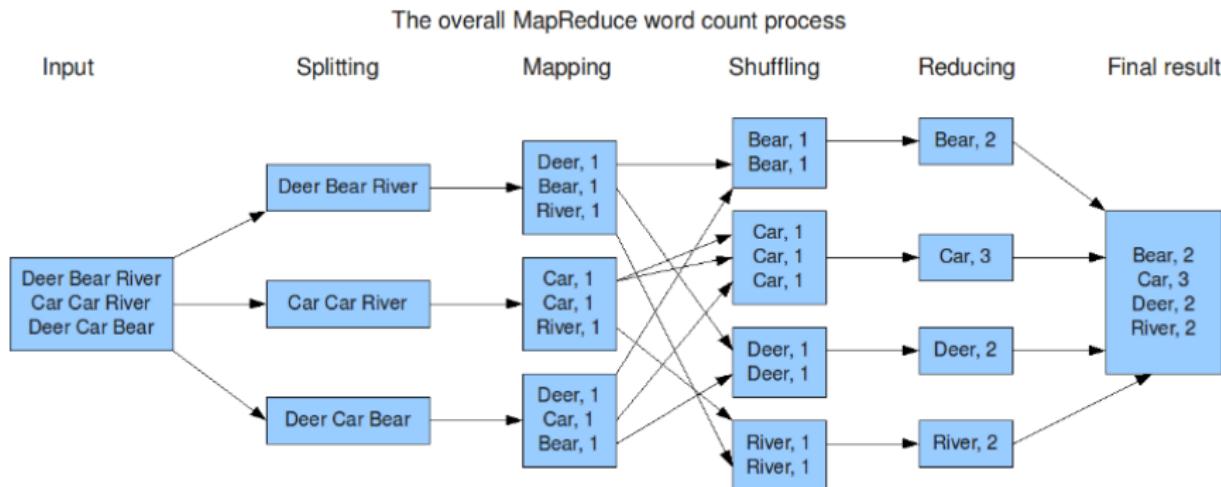
- *Hadoop Distributed File System.*
- Sistema de almacenamiento de Hadoop.
- Apariencia de un único sistema de ficheros, similar a Linux.
- Es altamente escalable y tolerante a fallos.



# HDFS y YARN: Almacenamiento y procesamiento como punto central

## • MapReduce: Procesamiento

- MapReduce permite el procesamiento distribuido sobre HDFS.
- Permite obtener Data Locality (revolución tecnológica de Big Data).
- Es el primer paradigma de programación de Hadoop.
  - Tiene una fase de Mapeo (**Map**), y otra de Agrupación (**Reduce**).

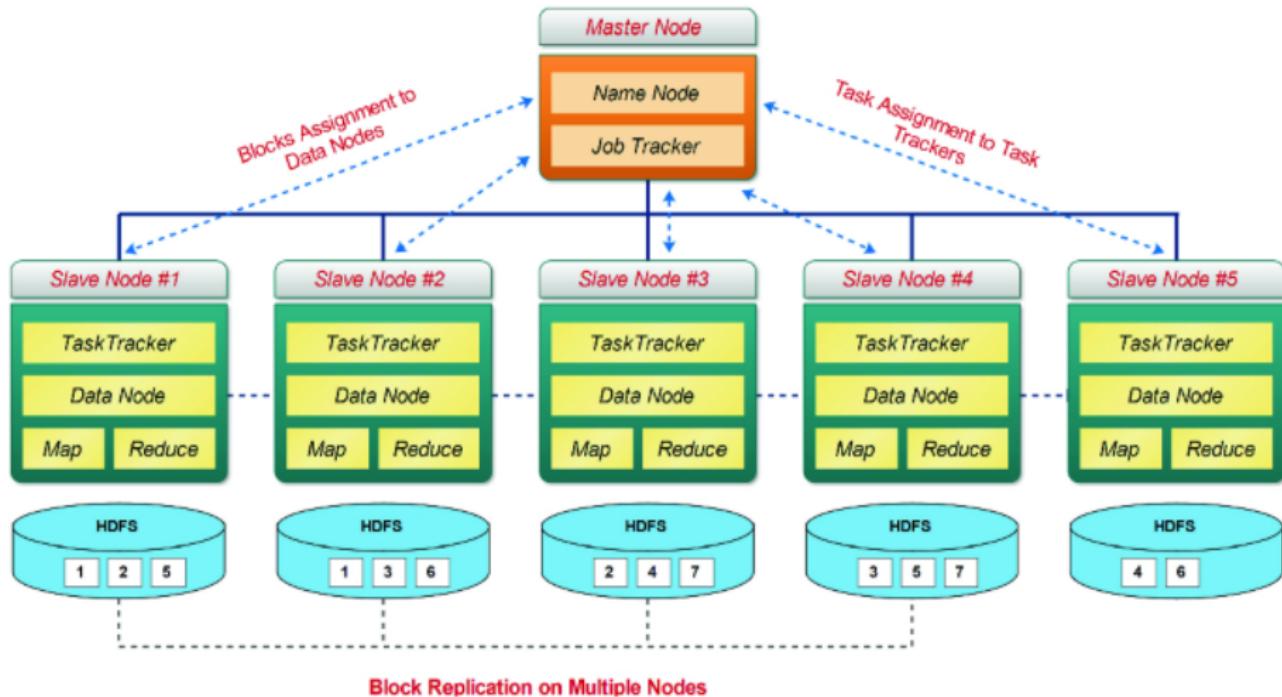


# HDFS y YARN: Almacenamiento y procesamiento como punto central

- **YARN: Gestión de recursos**

- YARN gestiona los recursos (CPU y RAM) del cluster para procesar
- Es tolerante a fallos y permite multitud de ejecuciones en paralelo.
- Gestiona cuotas de ejecución según prioridad, necesidad, disponibilidad,
- ...
- Apareció con MapReduce, pero puede trabajar con otros otros frameworks de procesamiento, como Spark.

# El cluster de Hadoop

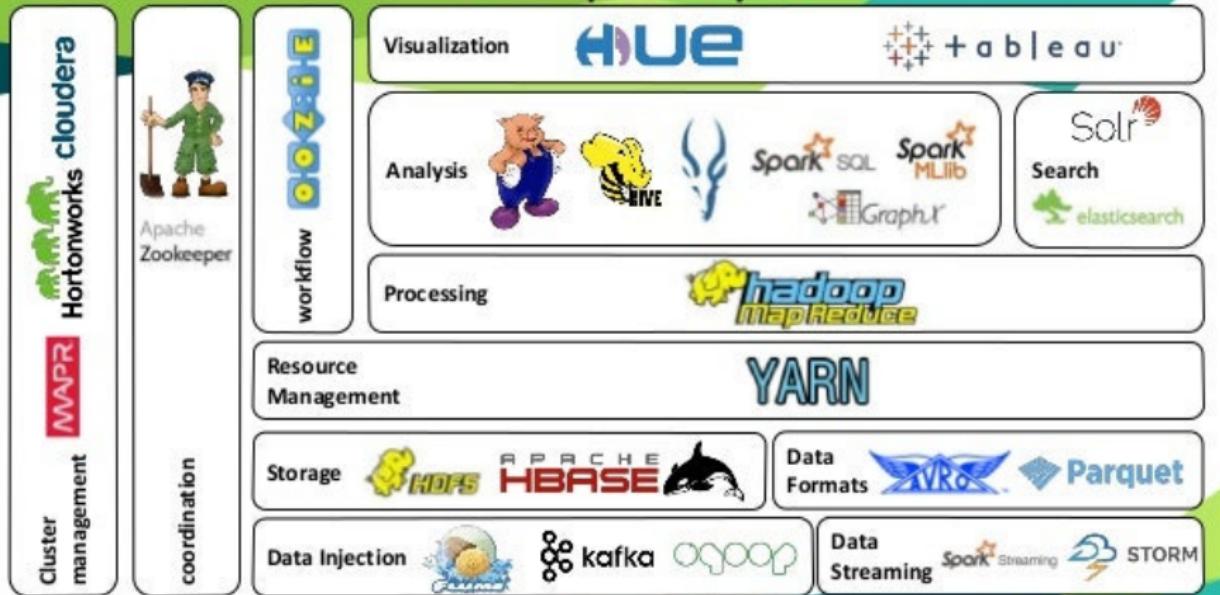


# El ecosistema Hadoop

- Hay una gran cantidad de componentes que trabajan de manera nativa con HDFS y YARN.
- Algunos son también del Proyecto Apache.
- Otros no...
- Pero al conjunto de todos se conoce como “Ecosistema Hadoop”

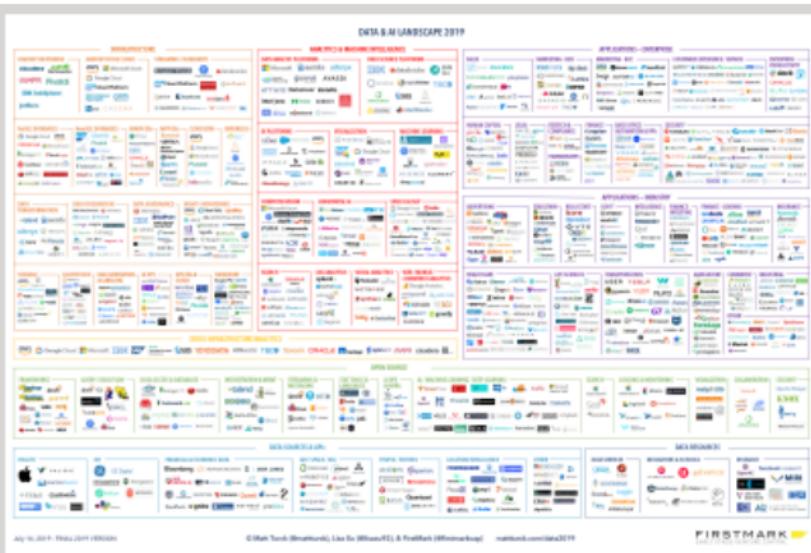
# El ecosistema Hadoop (un ejemplo entre muchos)

## Hadoop Ecosystem



# Ubicar los componentes en sus capas correspondientes

- Hay un gran cantidad de productos en el mundo Big Data (dentro y fuera del ecosistema Hadoop)
- Lo importante es ubicar el producto en las Capas de la Plataforma Big Data



# Ubicar los componentes en sus capas correspondientes



# Hadoop es al Big Data lo que Linux a los Sistemas Operativos

- Apache Hadoop es software libre, pero sin un soporte oficial por parte de Apache, como Linux.
- Hay empresas que ofrecen soporte comercial sobre software libre, tanto en Linux como en Hadoop.
- Hadoop sería comparable a Linux.
- Cloudera y Hortonworks son en Hadoop como Red Hat y Ubuntu en Linux.

# **PARTE SEXTA**

## **SOLUCIONES BIG DATA**

# Proveedores de productos Big Data

- Al instalar Hadoop se instala una Plataforma Big Data.
- Instalar “a mano” Hadoop es casi inviable → hay que recurrir a Plataformas Big Data.
- En Plataformas Hadoop hay dos empresas líderes:

**cloudera®**

Ask Bigger Questions



## Proveedores de productos Big Data: Cloudera

- Tiene componentes propios.
- No soporta cualquier componente de Hadoop.
- Cloudera Manager como administrador (software propio).
- Obligatorio licencia para su uso empresarial: Incluye soporte

# Proveedores de productos Big Data: Hortonworks

- Componentes 100% Apache Hadoop.
- Soporta cualquier componente de Hadoop.
- Ambari Server como administrador (software Open Source).
- No requiere licencia para su uso empresarial. Se puede contratar soporte.

Ambas empresas se fusionaron en **Cloudera**

# PARTE SÉPTIMA

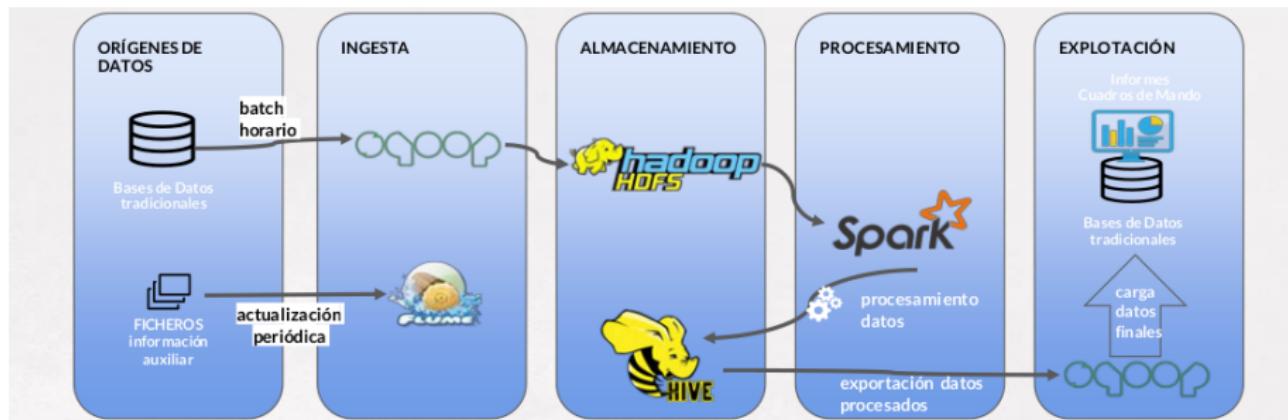
## CASOS DE USO DE BIG DATA

# Ejemplos de aplicación de Big Data

- **Industria:** Mejora de tiempos en procesos de larga duración
- **Operadora Televisión:** Análisis de la calidad de señal de televisión
- **Debate político:** conocer en tiempo real la respuesta de las redes sociales

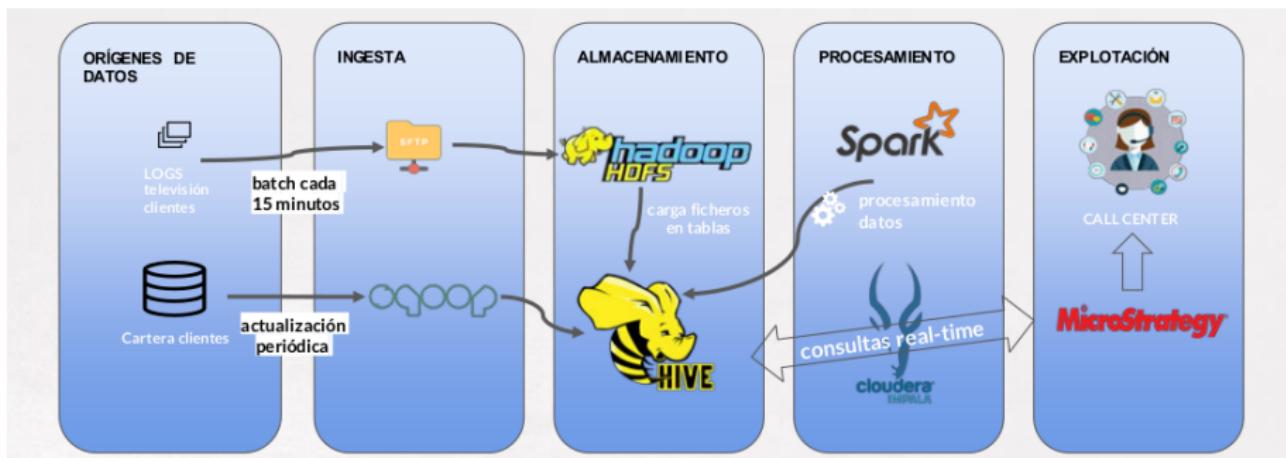
# Industria: Mejora de tiempos en procesos de larga duración

- Procesos que cargan información de unas bases de datos, se procesan y exportan a una base de datos de la que se mostraban gráficos con la información procesada.
- Con un proceso similar, con tecnologías Big Data, se reducen tiempos en un 75%.



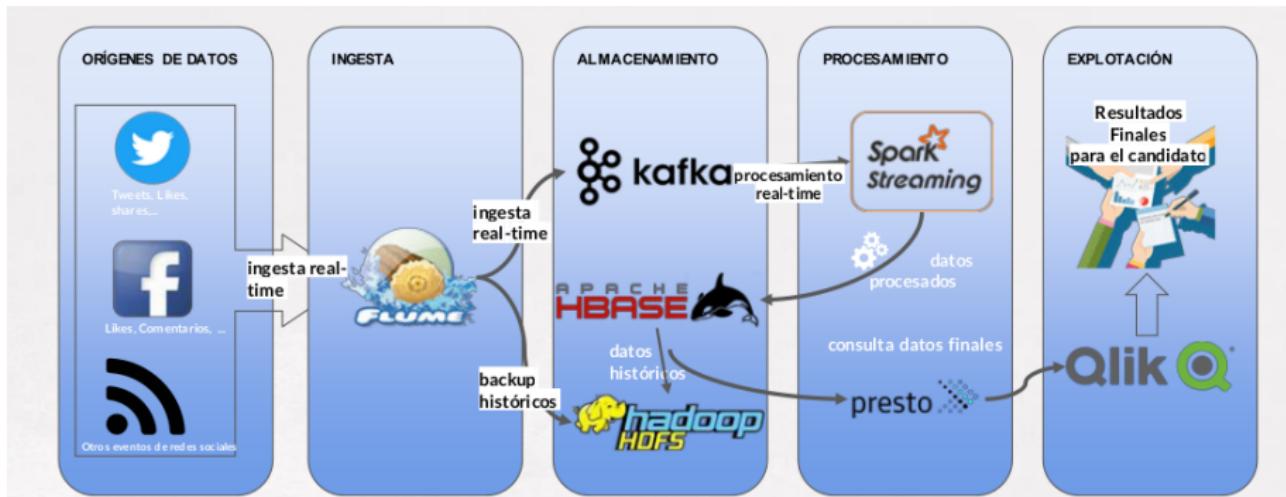
# Operadora de televisión: Análisis de la señal de televisión

- Se recogen los logs de todos los decodificadores de cliente, se almacenan y procesan para que el CallCenter tenga un cuadro de mandos actualizado con los indicadores de la TV de cada cliente: calidad de la señal, cortes, canales, etc.



# Debate político: conocer en tiempo real la respuesta de las redes sociales

- Se muestran a los candidatos resultados en tiempo real de las publicaciones en redes sociales que se van produciendo.



Gracias por su atención

# BIG DATA

