

# Tutorial Big Data CLEI 2019



## Ciudad de Panamá, Panamá

Por: Edwin Montoya  
[emontoya@eafit.edu.co](mailto:emontoya@eafit.edu.co)

---

Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>

## 2. Ecosistema tecnológico de una solución Big Data

## •2. Almacenamiento y Motores

### 1. Almacenamiento:

1. Almacenamiento de datos no estructurados
2. Sistemas de Archivos Distribuidos
3. Bases de datos
4. Data warehouse
5. Datalakes

### 2. SQL y bases de datos relaciones

### 3. Bases de Datos NoSQL

### 4. Big Data, Almacenamiento masivo, motores de procesamiento (Map – Reduce, Spark), ecosistemas

### 5. Plataformas cloud para Big Data, SQL, NoSQL

# ¿Donde guardamos y procesamos tantos datos?

- **En Centros de Datos de mi empresa.**
  - Ej: Bancolombia tiene un cluster de Big Data, con tecnología Hadoop / Spark con aprox 1 PB.
    - 200 SERVIDORES + HADOOP + SPARK
- **En Nube:**
  - Amazon AWS
    - Almacenar: S3, Procesar EC2 / EMR, muchas bases de datos.
    - Machine Learning y Servicios Cognitivos.
  - Microsoft Azure
    - Almacenar: Azure Data Lake, Procesar: HDInsight, Databricks.
    - Azure Machine Learning y Servicios Cognitivos.
  - Google Cloud Platform
  - IBM Cloud

# Tecnología Big Data (1)

- **Tecnologías**

- HADOOP
- SPARK
- FLINK
- STORM
- SAMZA
- .
- .

- **Proveedores**

- \* Cloudera
- \* HortonWorks
- \* MapR
- \* Amazon
- \* Google
- \* Microsoft
- \* IBM

- **Despliegue**

- \* On-premise
- \* Cloud
- \* Hibrido

# Tecnología Big Data (2)

- Open Source
  - R, Python (los principales lenguajes de Big data, ML, IA)
  - Knime
- Comerciales
  - SAS
  - SPSS

## LÍDERES HADOOP / SPARK

- CLOUDERA
  - [www.cloudera.com](http://www.cloudera.com)
- HortonWorks
  - [www.hortonworks.com](http://www.hortonworks.com)
- MapR
  - [www.mapr.com](http://www.mapr.com)

Apache Hadoop  
Software



cloudera



## LIDERES EN NUBE

- Amazon
- Google
- Microsoft
- IBM



Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>

# Zoologico de Hadoop



Apache Ambari



APACHE ZooKeeper™



Apache Flink



Apache Mahout



www.educba.com

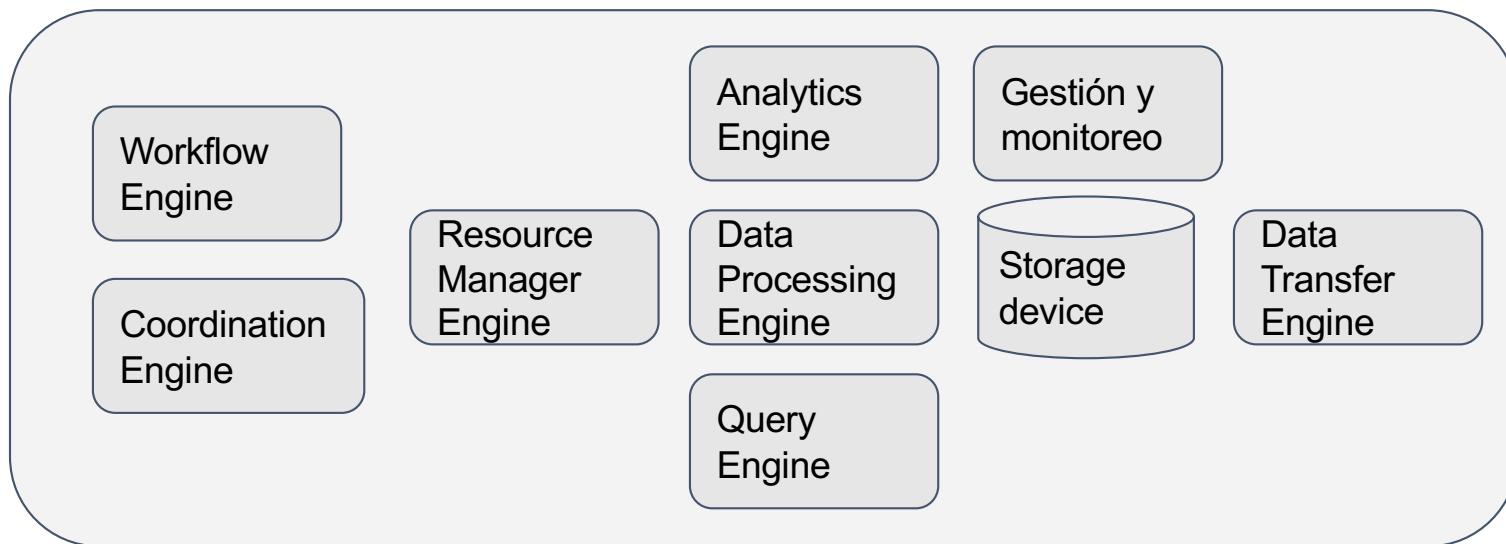
Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>

# Ecosistema tecnológico de una solución Big Data

- Sistema de almacenamiento
- Motor de procesamiento
- Gestor de recursos
- Motor de transferencia de datos
- Motor de consultas (Query Engine)
- Motor analítico (Analytics Engine)
- Motor de flujo de trabajo (Workflow)
- Motor de coordinación

# Ecosistema tecnológico de una solución Big Data



# Sistema de almacenamiento

- Almacenan de forma distribuida y no distribuida
  - Almacenan los datos que luego son procesados
  - Los datos son inmutables: WORM
  - Almacenan datos estructurados, semi-estructurados y sin estructurar
- \* Teorema CAP

# Sistema de almacenamiento

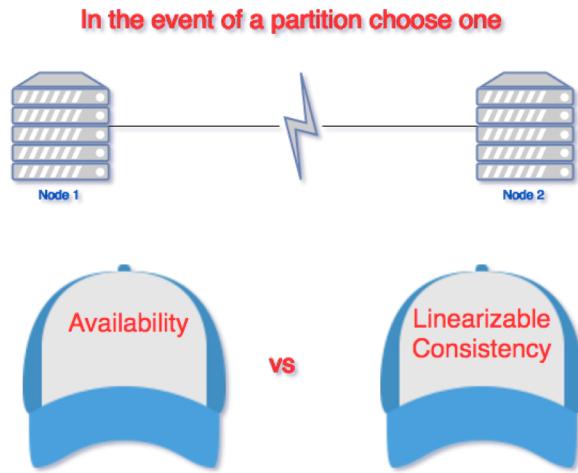
Side-by-side comparison: Object vs. traditional storage			
	OBJECT STORAGE	FILE-BASED STORAGE	BLOCK-BASED STORAGE
<b>Transaction units</b>	Objects, that is, files with custom metadata	Files	Blocks
<b>Supported type of update</b>	No in-place update support; updates create new object versions	Supports in-place updates	Supports in-place updates
<b>Protocols</b>	REST and SOAP over HTTP	CIFS and NFS	SCSI, Fibre Channel, SATA
<b>Metadata support</b>	Support of custom metadata	Fixed file-system attributes	Fixed system attributes
<b>Best suited for</b>	Relatively static file data and as cloud storage	Shared file data	Transactional data and frequently changing data
<b>Biggest strength</b>	Scalability and distributed access	Simplified access and management of shared files	High performance
<b>Limitations</b>	Ill-suited for frequently changing transactional data; doesn't provide a sharing protocol with a locking mechanism	Difficult to extend beyond the data center	Difficult to extend beyond the data center

©2017 TECHTARGET. ALL RIGHTS RESERVED 

Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>

# Teorema CAP



- **Consistencia:** Es la garantía de que cada nodo en un clúster devuelve la misma escritura exitosa más reciente. La consistencia se refiere a que cada cliente que tiene la misma vista de los datos
- **Disponibilidad:** Cada nodo que no falla devuelve una respuesta para todas las solicitudes de lectura y escritura en un período de tiempo razonable. La palabra clave aquí es cada.
- **Tolerancia a la partición:** El sistema continúa funcionando y mantiene su consistencia a pesar de las particiones de red.

# Motor de procesamiento

- Procesa los datos almacenados en el sistema de almacenamiento y también desde fuentes externas (Batch recompute vs Real Time increment)
  - Alta latencia
  - Baja latencia
- Es un framework de programación distribuida/paralela
- Es administrado por el gestor de recursos
- En general es quien prepara, estructura y.... realiza cualquier proceso sobre los datos.
- En una plataforma Big Data pueden existir varios motores y varios tipos de motores de procesamiento según la necesidad.

# Gestor de recursos

- Cada solicitud de procesamiento posee características propias que se deben traducir en términos uso de los componentes y de las capacidades de cómputo de la plataforma de Big Data.
- Se deben poder administrar múltiples solicitudes simultáneamente.
- Un componente debe hacer las veces de kernel, custodio o árbitro de los recursos de cómputo.
- Un gestor de recursos actúa entonces como un ente que planea, prioriza, coordina, asigna y recupera los recursos a cada solicitud según estén disponibles.

# Motor de transferencia de datos

- Es el responsable de transferir los datos desde y hacia la plataforma de Big Data.
- No se ajusta a ningún esquema o estructura de datos de entrada.
- Usualmente proporcionan una o varias funciones de transferencia de datos desde archivos, eventos o fuentes relationales, por lo que es normal que en una plataforma Big Data existan uno varios motores que cubren dichas funciones.
- Algunos motores de transferencia de datos tienen su propio motor interno de procesamiento, el cual se usa exclusivamente para la transferencia de datos desde distintas fuentes de forma simultánea.

# Motor de consultas

- Dado que un motor de procesamiento es de difícil programación, este ó la solución de Big Data deben ofrecer un mecanismo que permita intuir la máquina sobre las tareas de consulta de los datos.
- Generalmente un motor de consultas provee una interfaz de usuario que permite abstraer la complejidad de los motores de procesamiento.
- Usualmente provee uno o más lenguajes de uso común para utilizarlos dentro de dichas interfaces (ejemplo: SQL)

# Motor analítico

- Provee el soporte para ejecutar algoritmos dentro de las categorías de Machine Learning, Deep Learning y Algoritmos estadísticos.
- Opera junto con el motor de procesamiento.
- Puede ser independiente o un componente del motor de consulta.
- Este motor también provee una interfaz de programación y uso de los algoritmos, que puede ser a través de lenguajes de uso común (ejemplos: R, Python y Scala).

# Motor de flujos de trabajo

- Nace como respuesta a la necesidad de planear la ejecución de las tareas que realizan los distintos componentes (internos y externos), los cuales a su vez siguen reglas de prioridad, tiempo, eventos, secuencias, etc.

# Motor de coordinación

- La mayoría de las plataformas o componentes Big Data se ejecutan de forma independiente en distintos servidores.
- Los sistemas distribuidos, deben superar problemas tales como retrasos de mensajes, velocidad del procesador, gestión de la configuración, sincronización, caídas del sistema, detección de fallas, administración de metadatos, maestros, sincronización, etc.

# Motor de coordinación

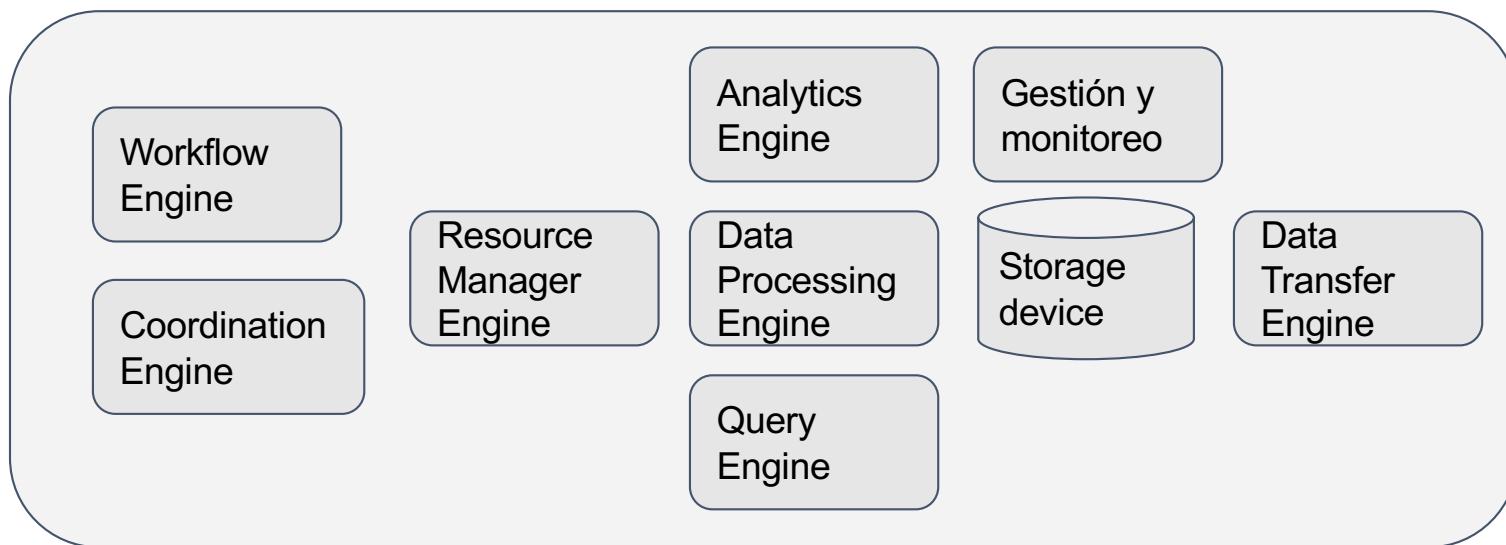
- Este motor es responsable por la monitorización y la coordinación entre las tareas de los distintos componentes de una plataforma Big Data.
- Una solución distribuida de Big Data que deba ejecutarse en varios servidores depende de un motor de coordinación, a fin de garantizar la consistencia operativa en todos los servidores involucrados.

### 3.Oferta de soluciones Big Data

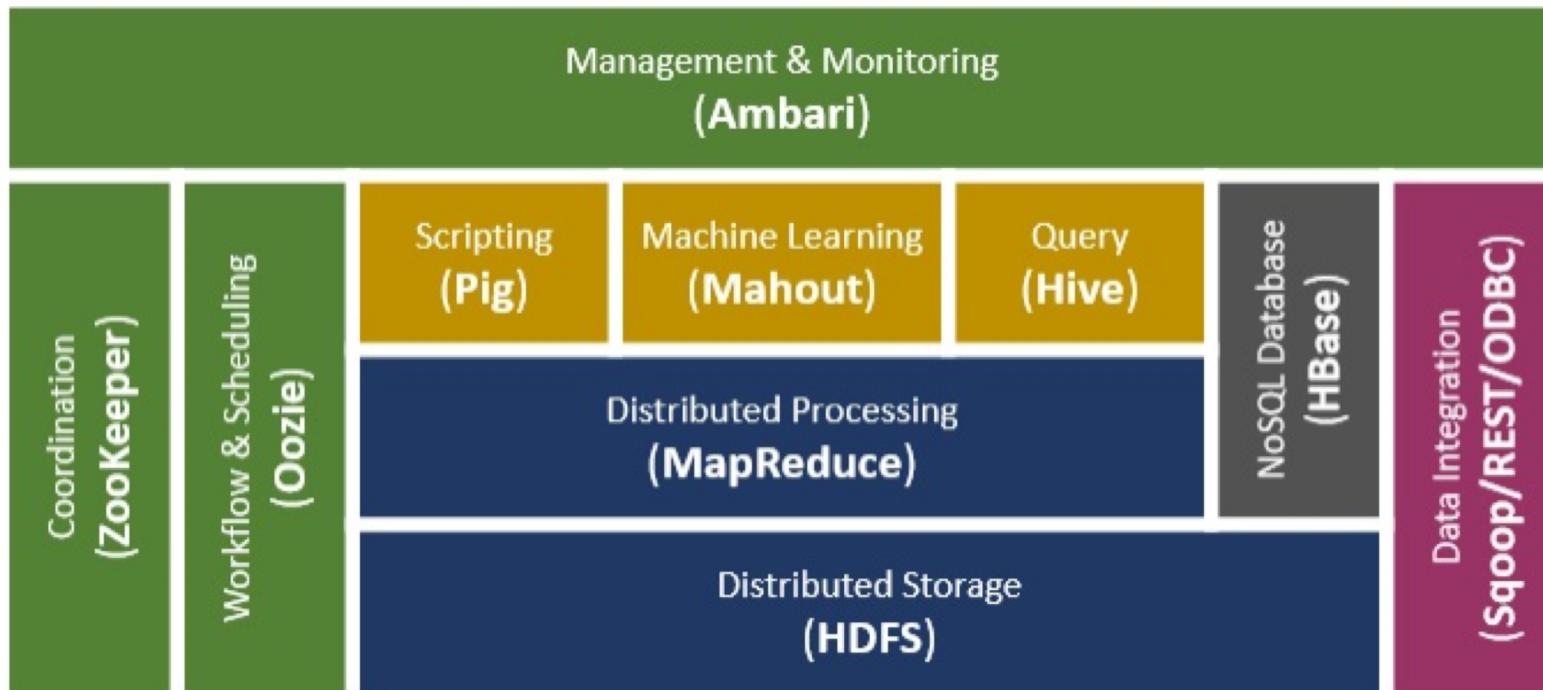
Inspira Crea Transforma



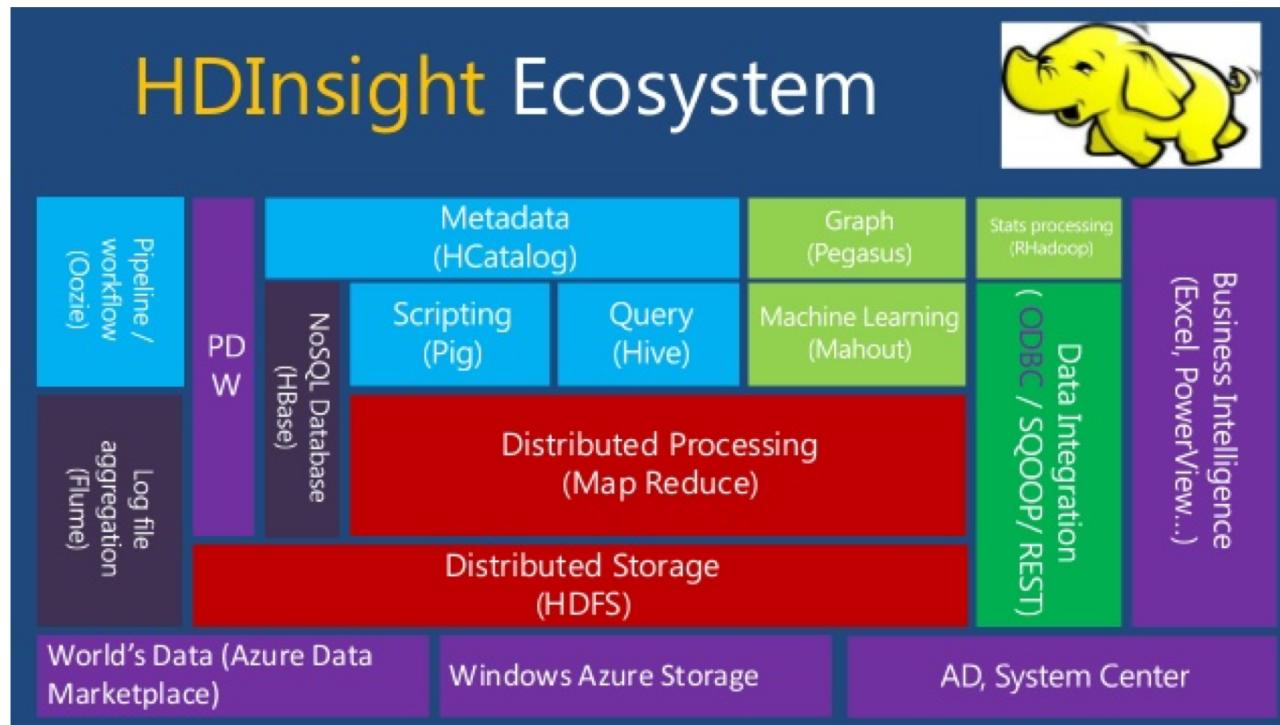
# Ecosistema tecnológico de una solución Big Data



# Ecosistema Hadoop (estándar)

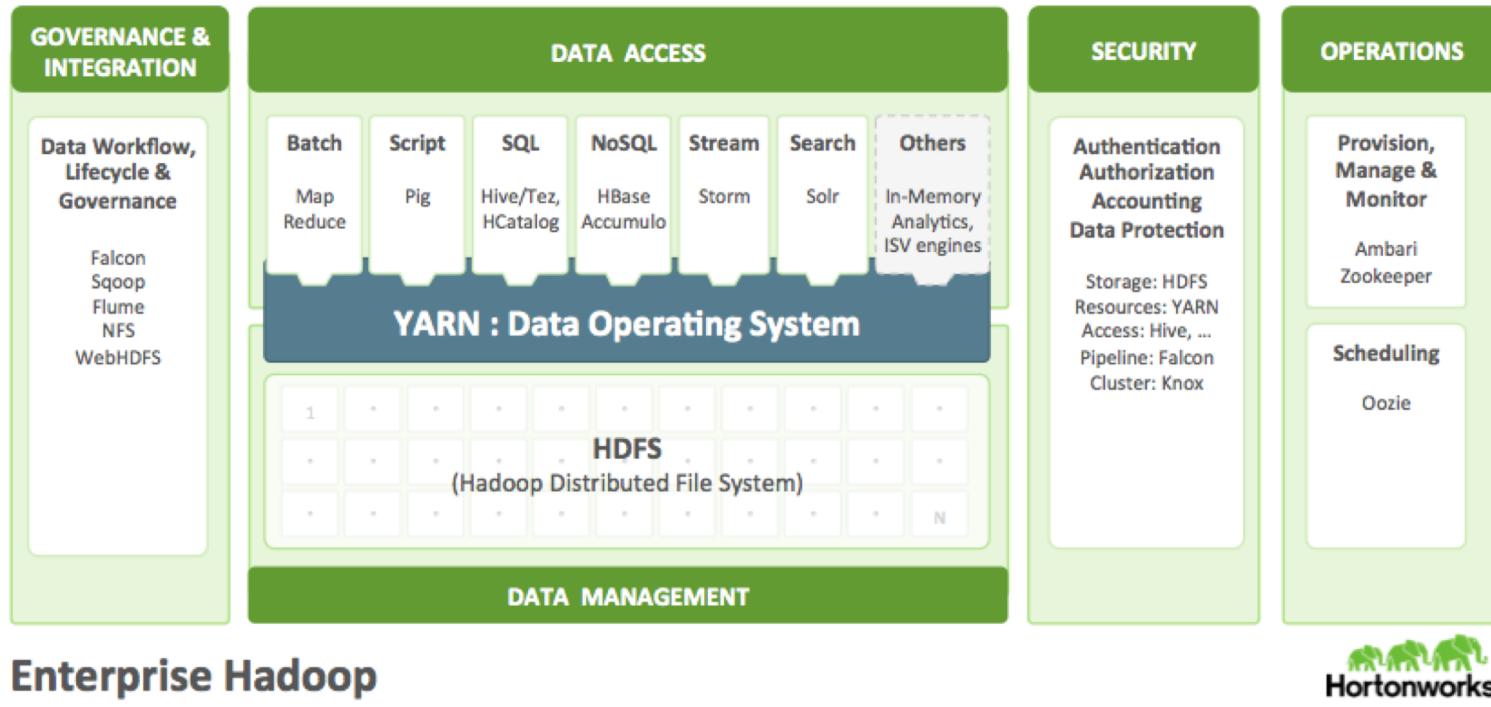


# Ecosistema Hadoop: Azure

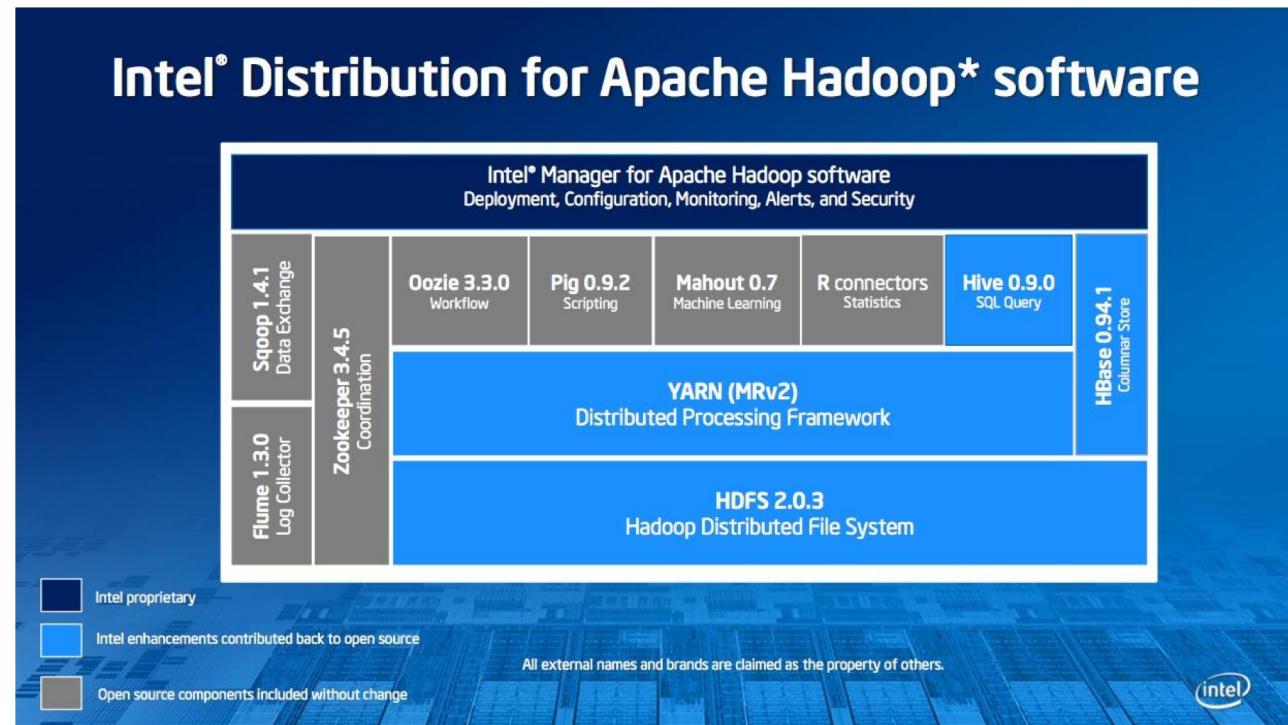


Inspira Crea Transforma

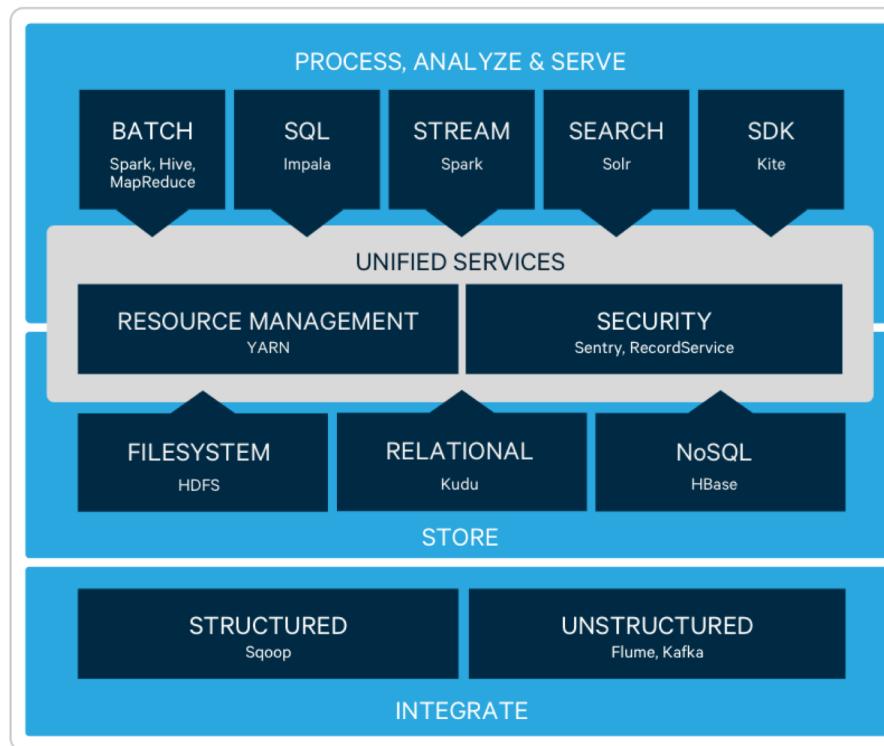
# Ecosistema Hadoop: Hortonworks



# Ecosistema Hadoop: Intel

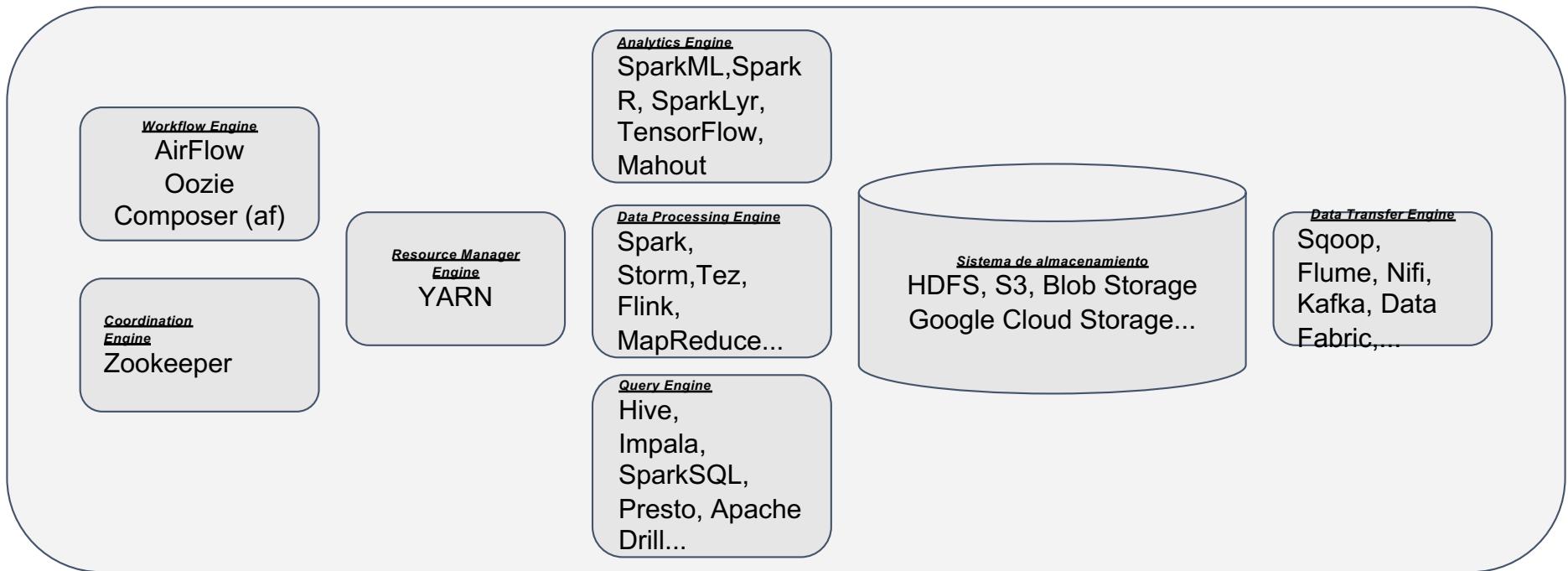


# Ecosistema Hadoop: Cloudera



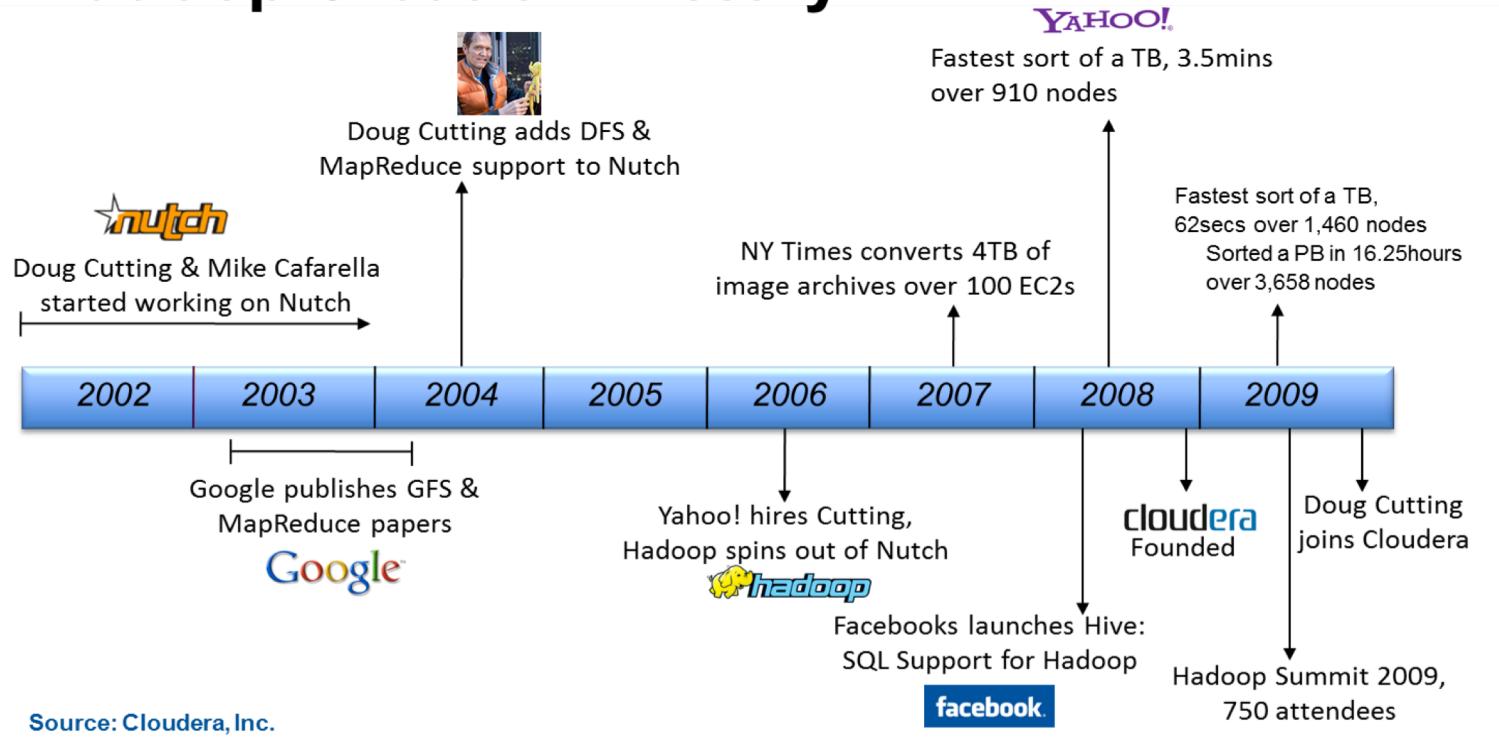
Inspira Crea Transforma

# Ecosistema Big Data



## 6. Ecosistema Hadoop

# Hadoop Creation History



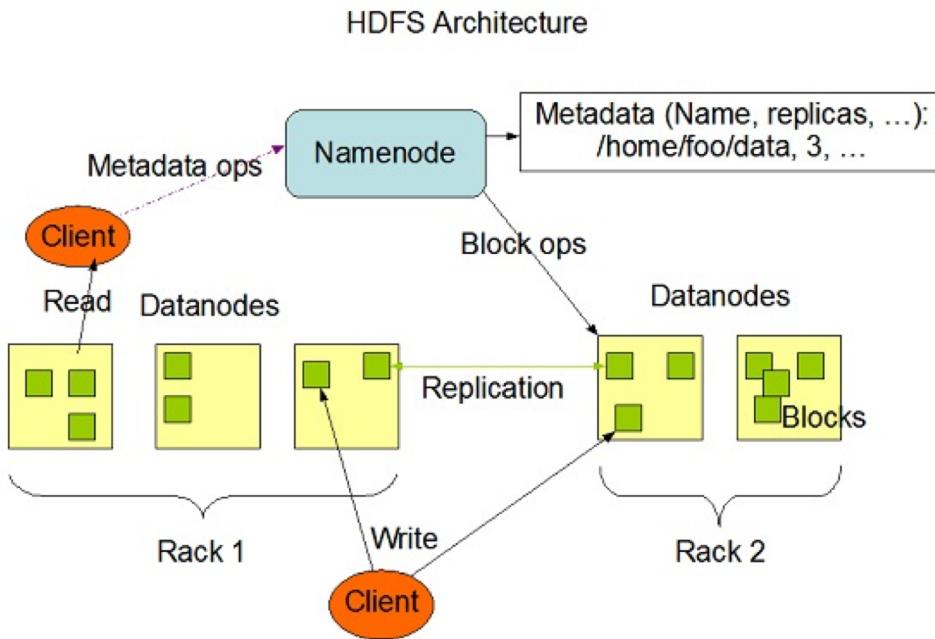
# Hadoop

- Apache Hadoop es un Framework de Big Data, que provee un modelo de computación distribuida que usa “commodity hardware” de forma flexible y que puede ser escalable.
- Es una arquitectura Master / Slave
- Sigue el paradigma Hadoop MapReduce

# Hadoop

- Emplea 3 componentes básicos dentro de su arquitectura:
  - Un sistema de archivos distribuido (HDFS)
  - Un motor para el procesamiento de grandes volúmenes de datos (MapReduce)
  - Un gestor de recursos (YARN) ??

# Arquitectura de Hadoop v1 - HDFS



Almacena los datos y los metadatos por separado en el sistema de archivos.

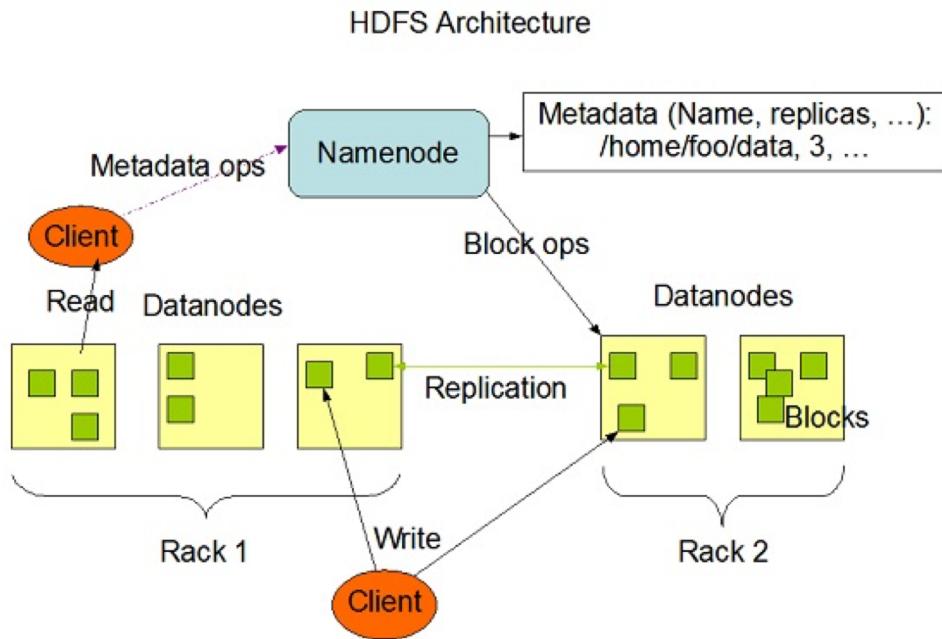
Los datos de la aplicación se almacenan en los servidores DataNode.

Los metadatos del sistema de archivos se almacenan en los servidores NameNode

HDFS replica todo el contenido de los DataNodes para garantizar la confiabilidad de los datos.

NameNode y DataNode usan protocolos basados TCP para su comunicación.

# Arquitectura de Hadoop v1 - HDFS

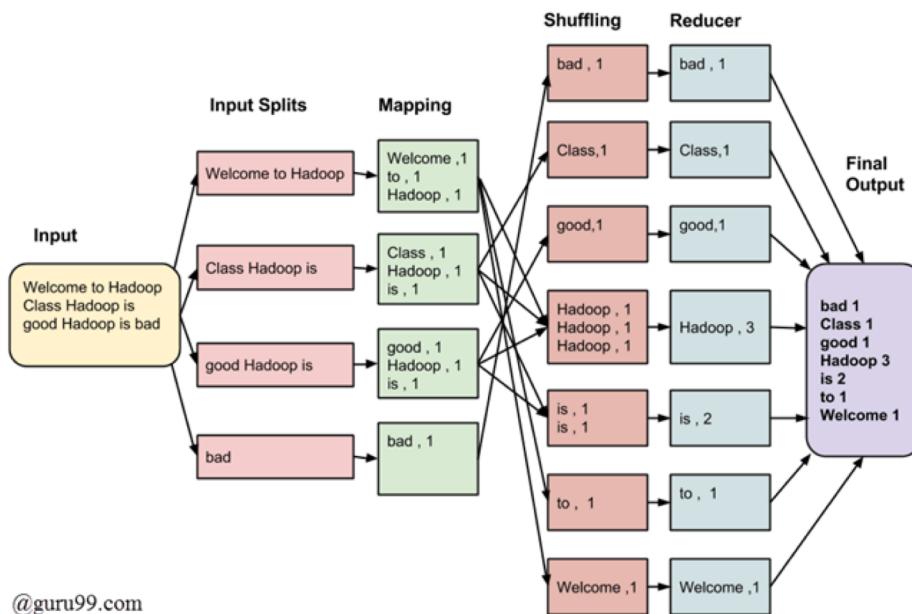


Todos los archivos y directorios en HDFS están representados en el NameNode por Inodes que contienen atributos como: permisos, marca de modificación, cuota, etc.

Un NameNode mapea toda la estructura del sistema de archivos en memoria.

Un DataNode administra el estado de un nodo HDFS, es quien realiza tareas sobre los datos (E/S, cálculos estadísticos, ML,...)

# Arquitectura de Hadoop - MapReduce



@guru99.com

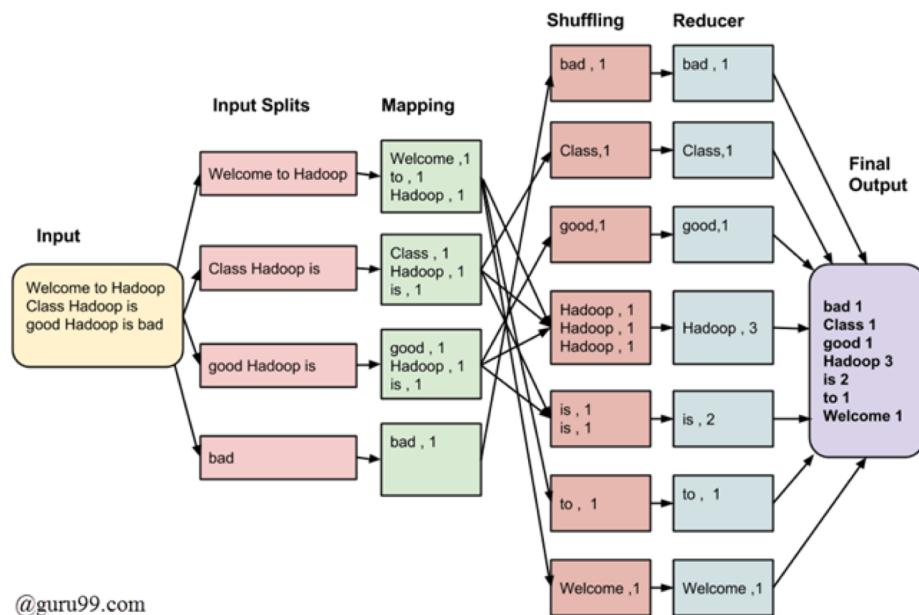
El cliente envía una tarea al JobTracker

El JobTracker interroga al NameNode sobre la ubicación de los datos.

Según la respuesta del NameNode el JobTracker solicita a los respectivos Task Trackers para que ejecuten una tarea en sus datos.

Los resultados son guardados en los DataNode y el NameNode es informado.

# Arquitectura de Hadoop - MapReduce

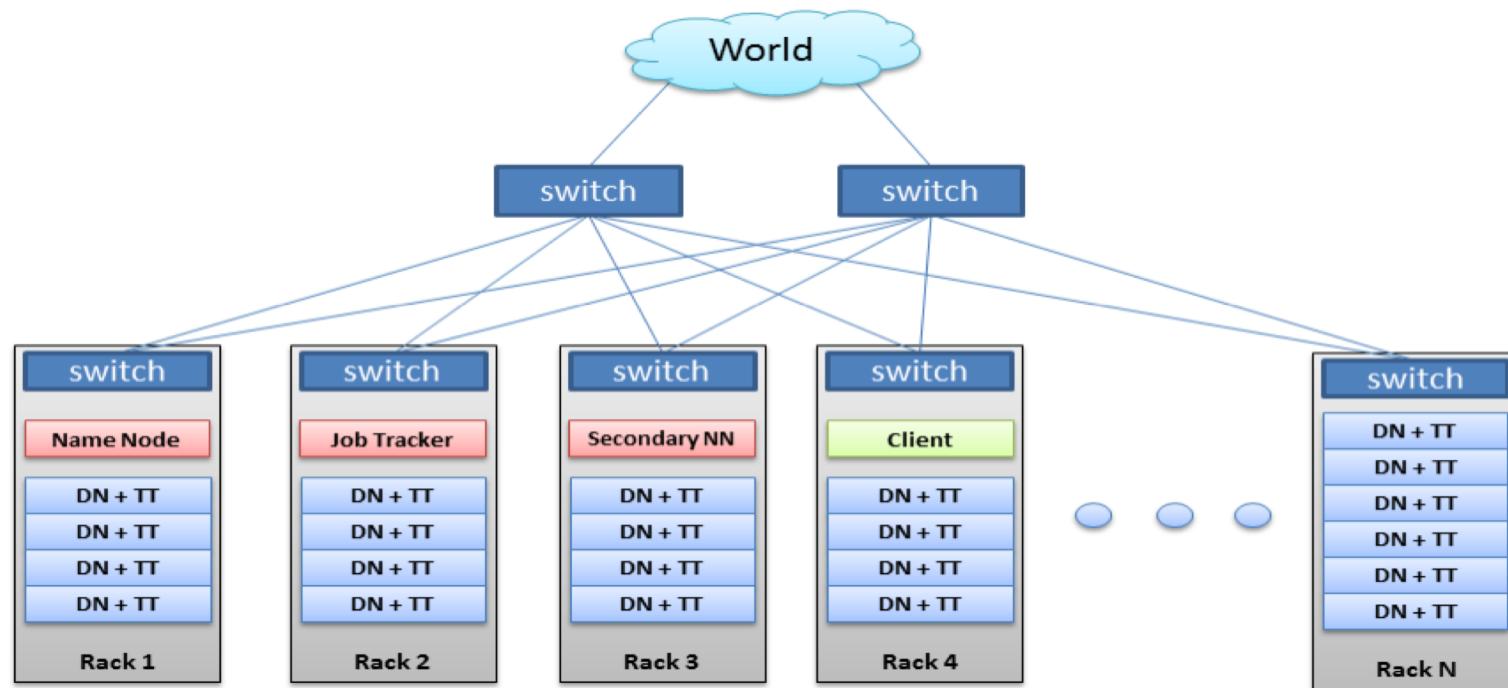


El JobTracker informa al cliente sobre la terminación de la tarea.

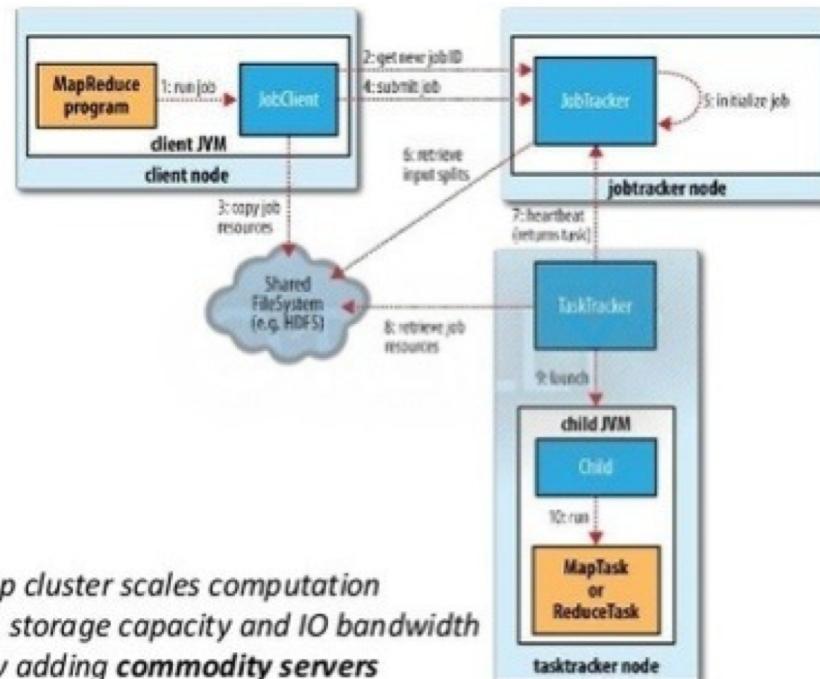
El cliente contacta al NameNode y recibe los resultados.

<https://www.youtube.com/watch?v=lgWy7BwIKKQ>

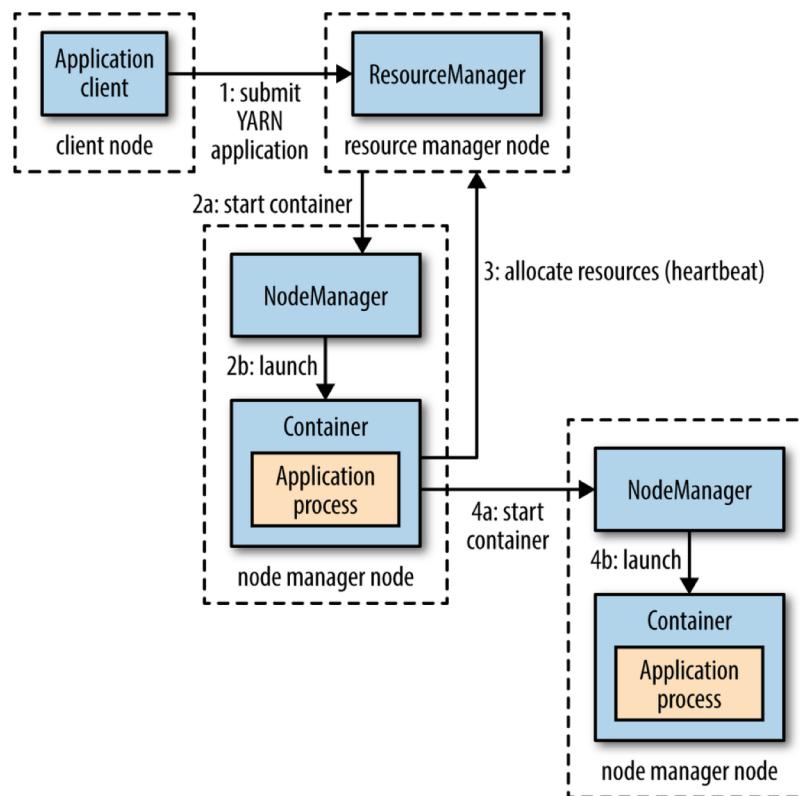
# Hadoop Cluster v1



# Arquitectura de Hadoop v1



# Arquitectura de Hadoop v2 -YARN

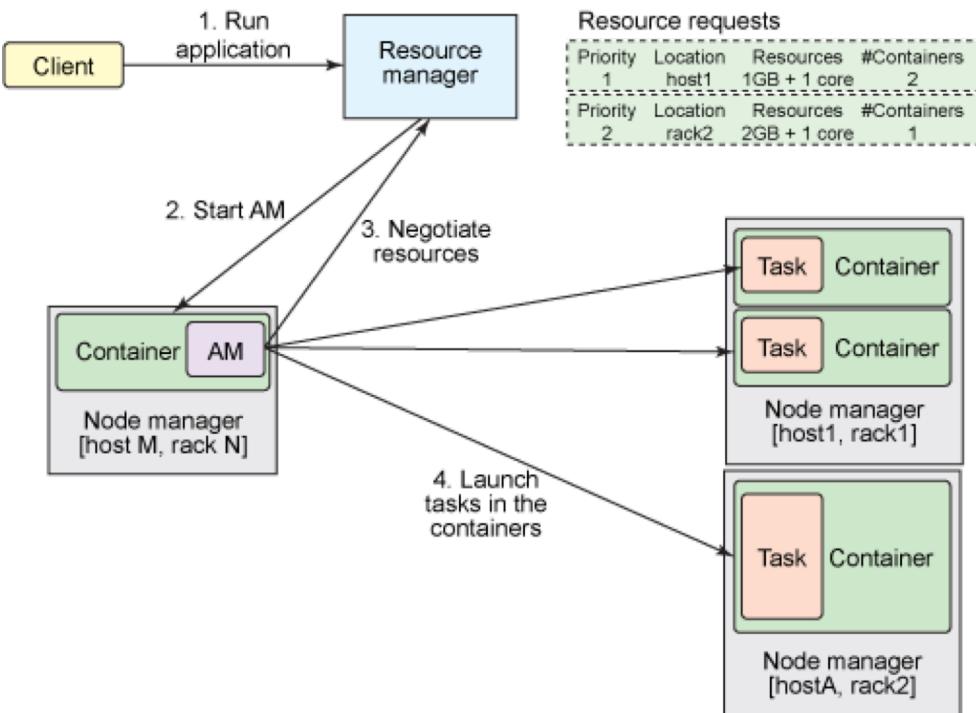


Tiene dos componentes principales:

- JobScheduler: asigna recursos (CPU,memoria...)
- Application Manager: supervisar el progreso de la aplicación enviada (MapR)

Node Manager: cada nodo tiene un node NM que mantiene los recursos disponibles, administra el ciclo de vida del contenedor.

# Arquitectura de Hadoop v2 -YARN

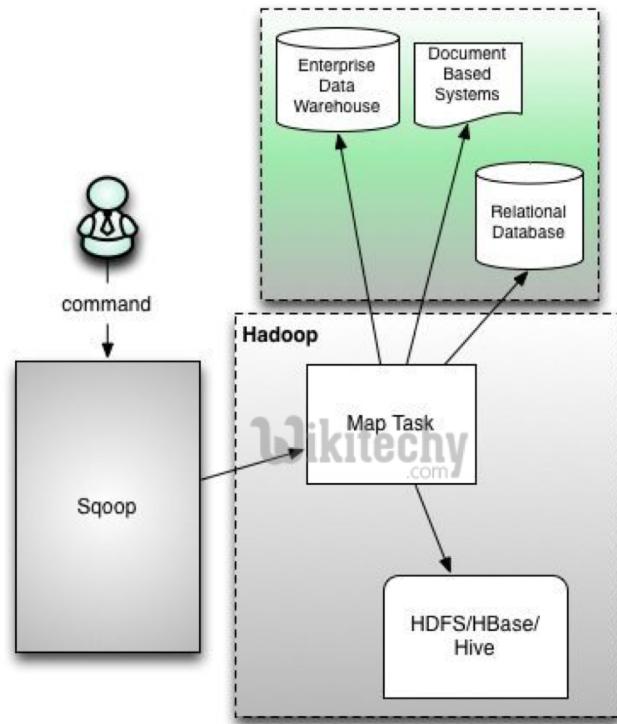


Tiene dos componentes principales:

- JobScheduler: asigna recursos (CPU,memoria...)
- Application Manager: supervisar el progreso de la aplicación enviada (MapR)

Node Manager: cada nodo tiene un node NM que mantiene los recursos disponibles, administra el ciclo de vida del contenedor.

# Arquitectura de Hadoop - Sqoop

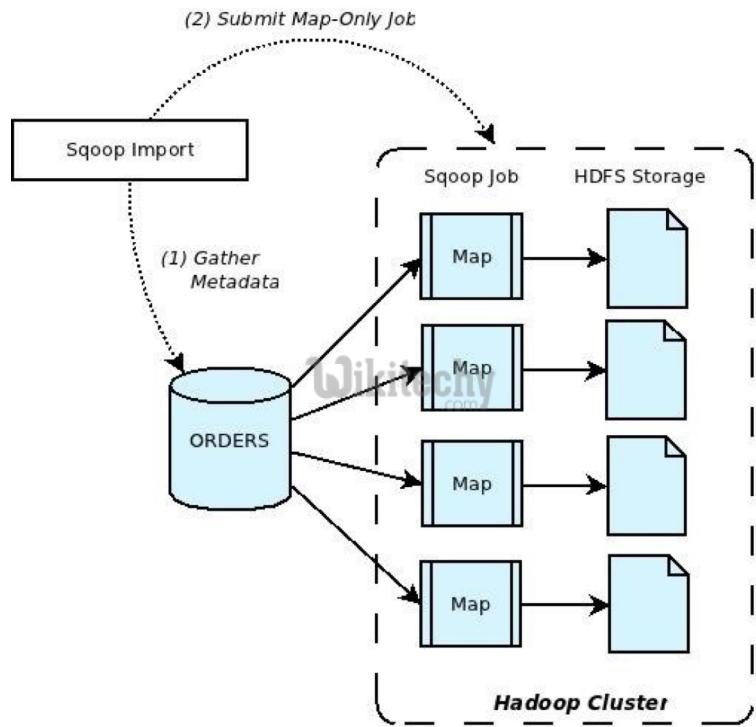


Sqoop solicita información sobre el schema en la BD y captura los metadatos.

Crea un Map Job que es enviado al cluster de Hadoop.

Cada job captura la información de la fuente de datos y la almacena en HDFS.

# Arquitectura de Hadoop - Sqoop

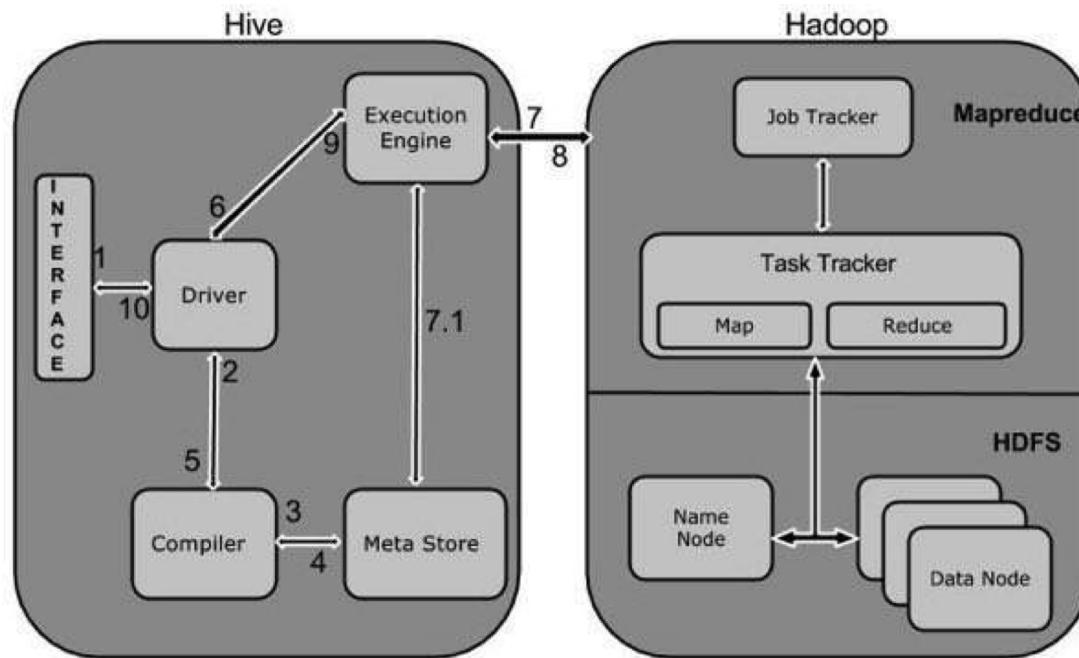


Sqoop analiza la bd y captura los metadatos

Crea un Map Job que es enviado el cluster de Hadoop.

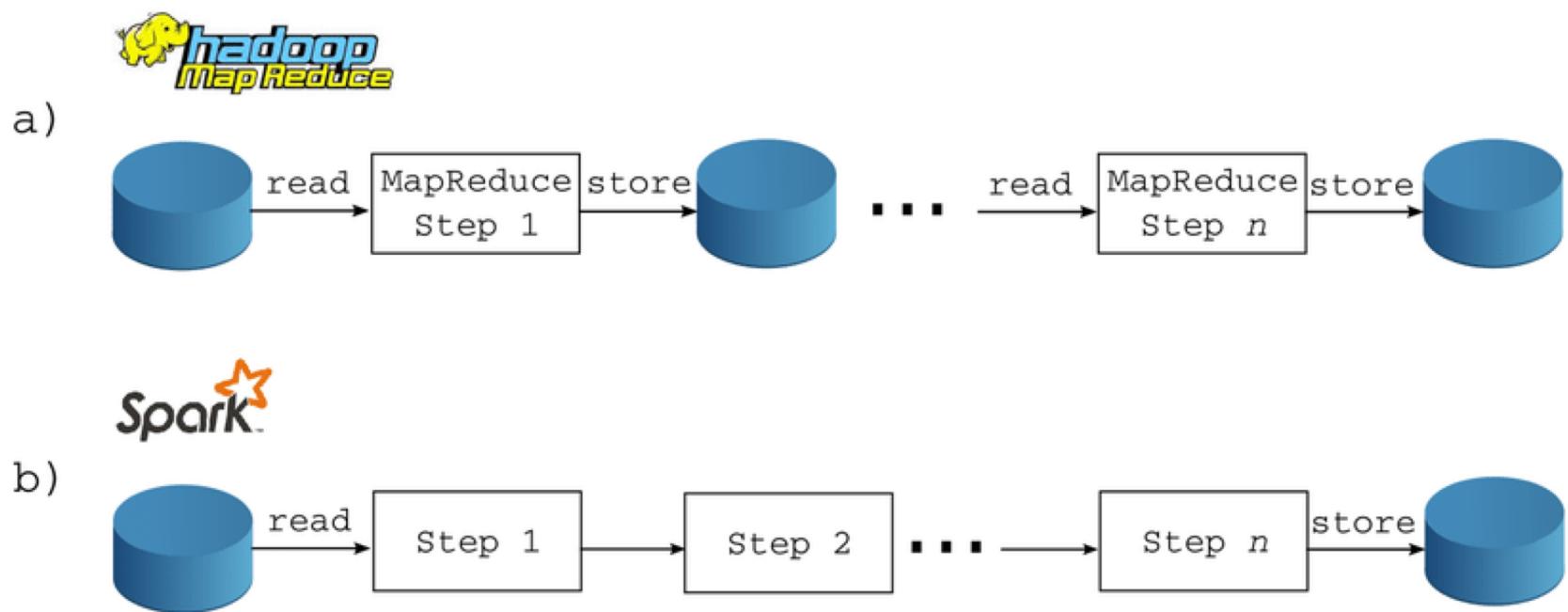
Cada job captura la información de la fuente de datos y la almacena en HDFS.

# Arquitectura de Hadoop - Hive

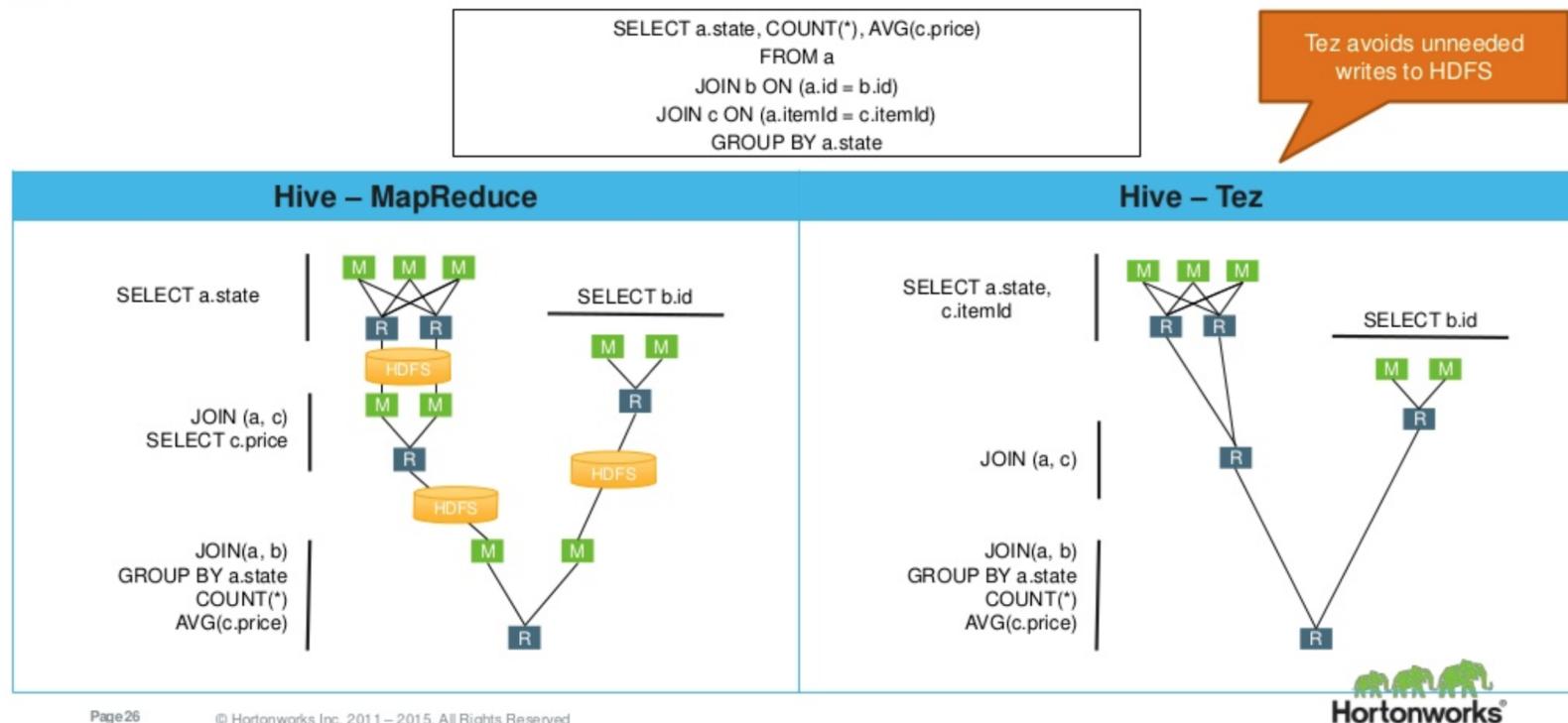


1. Ejecutar la solicitud
2. Obtener plan
3. Obtener metadatos
4. Enviar metadatos
5. Enviar plan
6. Ejecutar plan
7. Ejecutar trabajo
8. Metadata Ops
9. Resultado de búsqueda
10. Enviar resultados

# MapReduce vs Spark



# MapReduce vs Tez

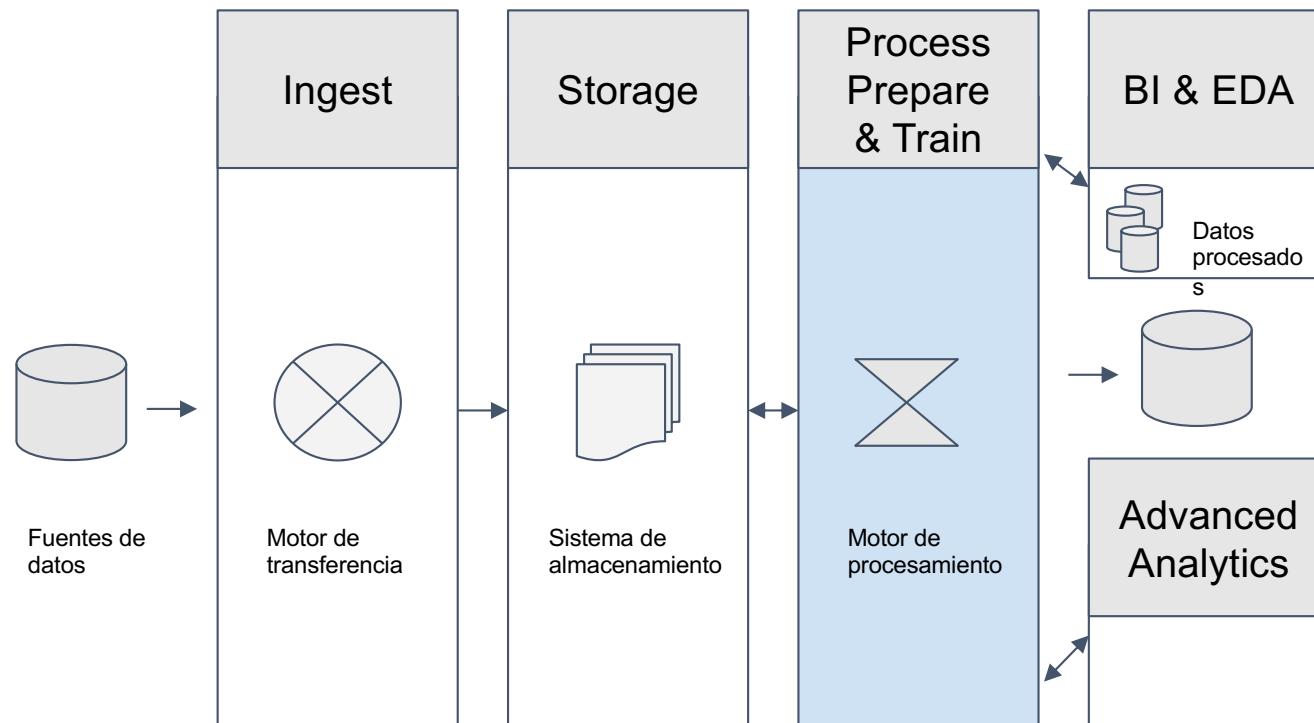


# Ecosistemas Big Data y analítica - Nube

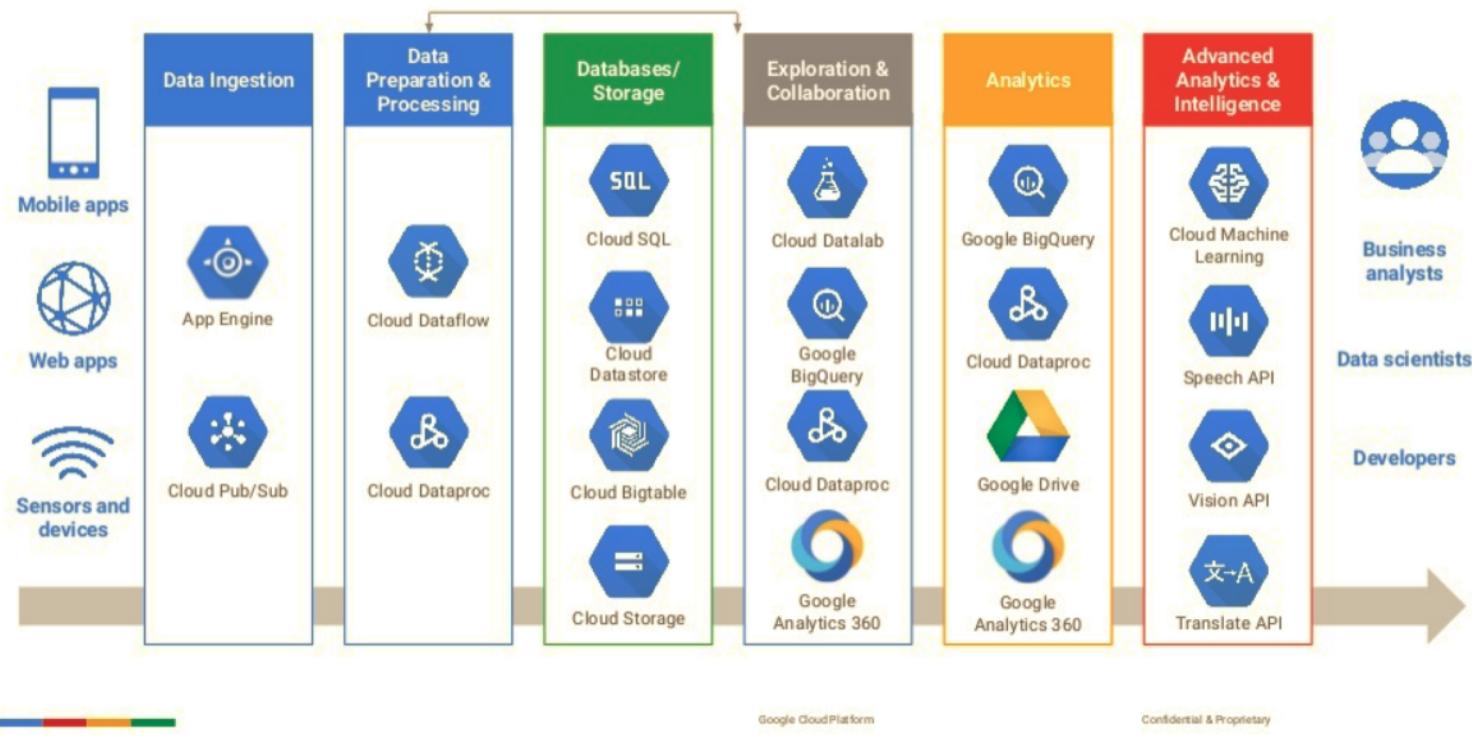
Inspira Crea Transforma



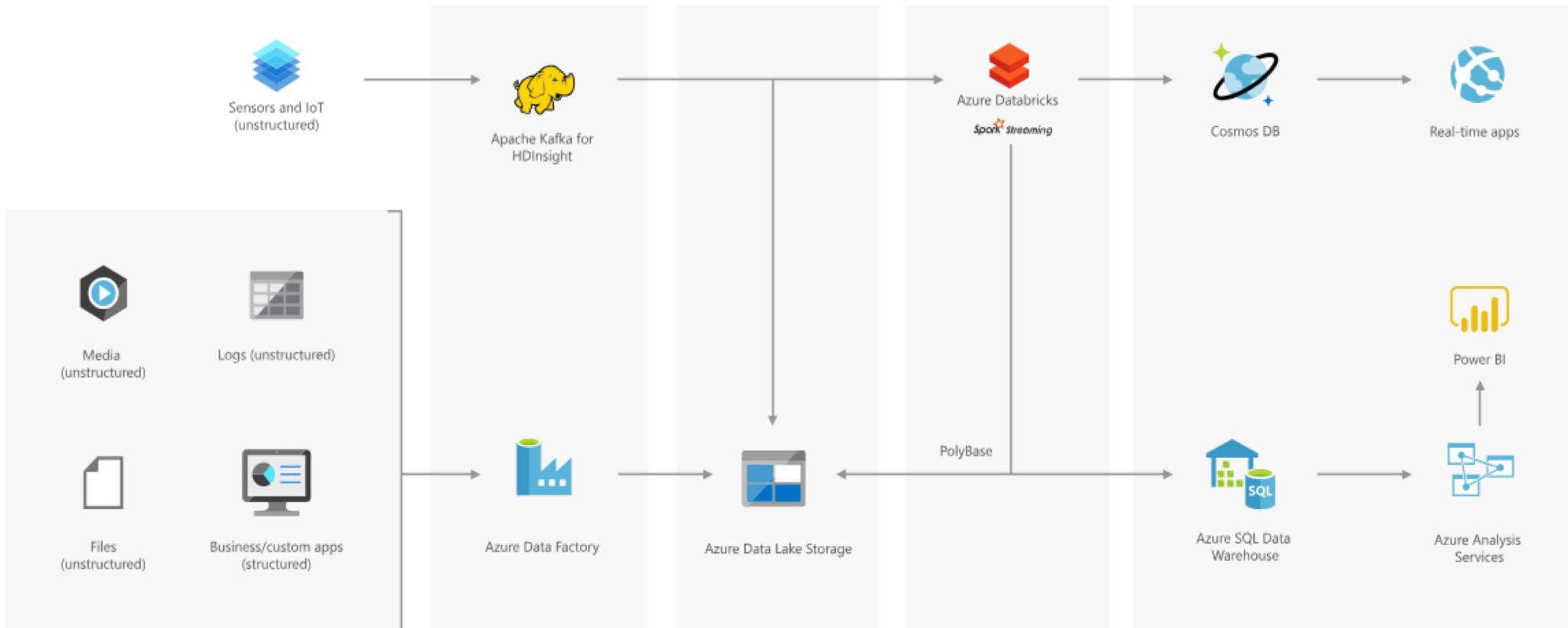
# Proceso general de Big Data & Analytics



# Google

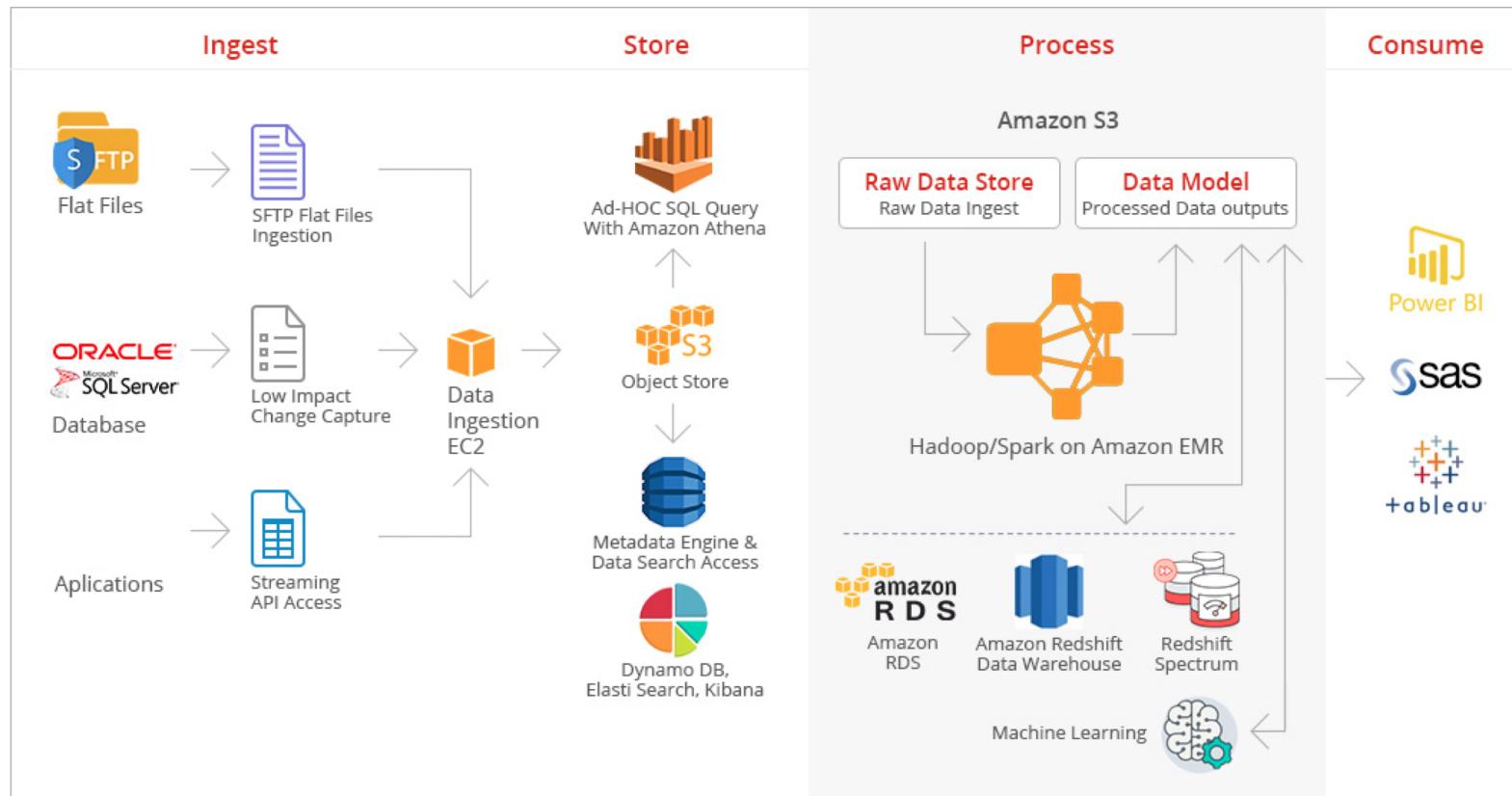


# Azure



Microsoft Azure also supports other Big Data services like Azure IoT Hub, Azure Event Hubs, Azure Machine Learning and Azure Data Lake to allow customers to tailor the above architecture to meet their unique needs.

# AWS



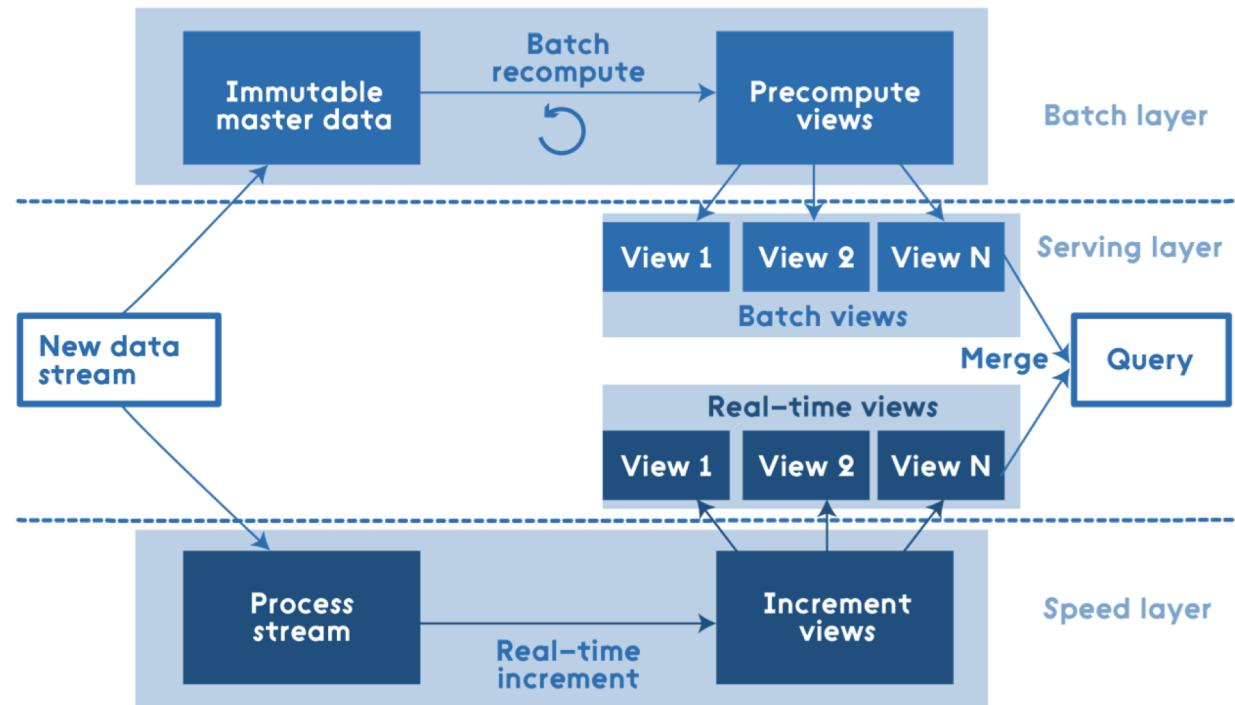
Inspira Crea Transforma

# 4. Arquitecturas Big Data

Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>

# Arquitectura Lambda



# Arquitectura Lambda

