

Tutorial Big Data CLEI 2019



Ciudad de Panamá, Panamá

Por: Edwin Montoya
emontoya@eafit.edu.co

Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]

Edwin Montoya



Ingeniero de Sistemas

Doctor Ingeniero en Telecomunicación

Jefe Depto Sistemas

Profesor: Big Data, Cloud Computing, Text Mining

Investigador: Big Data, E-learning con Big Data e IA

Comité de Maestría en Ciencia de Datos y Analítica



* Participó en la formulación y consolidación del Centro de Excelencia en Big Data y Analytics

* Comité técnico y Junta Directiva

* Investigador Adjunto

Experiencia en Big Data

- Varios cursos formales en Pregrado y Maestría en Big Data
- Varios cursos en Educación continua en Tecnologías Big Data.
- 7 cursos de tecnologías Big Data a nivel nacional dentro del programa de Ciudadano de Datos con Caoba, 4 ciudades en Colombia.
- 5 proyectos de maestría recientes relacionados con Big Data
- Asesorías y Consultorías: Modelo de Referencia para Big Data a Terpel (Distribuidor de Petróleo)
- Proyectos de Investigación en Caoba

Laboratorios

- <https://github.com/edwinmontoya/tutorialbigdatalei2019.git>
- Se tendrá acceso a una serie de Recursos en Nube:
 - AWS EMR:
 - <http://emr1.emontoya.ml:8888> (hue)
 - <http://emr2.emontoya.ml:8888> (si necesita)
 - <http://emr3.emontoya.ml:8888> (si necesita)
 - User: admin Password: Clei2019*
 - <http://emr1.emontoya.ml:8890> (zeppelin)
 - <http://emr2.emontoya.ml:8890> (si necesita)
 - <http://emr3.emontoya.ml:8890> (si necesita)
 - Sin autenticación

Recursos Gratuitos para Big Data

- **Cuenta FreeTier o AWS Educate en Amazon.**
 - Servicios RDS/Mysql (base de datos relacional)
 - Servicio EMR (Elastic MapReduce)
 - Servicio S3 (Almacenamiento por objetos)
- **Microsoft Azure FreeTier o Azure for Students:**
 - <https://azure.microsoft.com/en-us/free/students/>
 - <https://visualstudio.microsoft.com/dev-essentials/>
- **Databricks Community**
 - Cluster gratis restringido para Spark + Jupyter Notebooks
- **Máquinas virtuales Sandbox de:**
 - Hortonworks:
 - <https://www.cloudera.com/downloads/hortonworks-sandbox.html>
 - Cloudera:
 - https://www.cloudera.com/downloads/quickstart_vms/5-13.html

Algunas iniciativas en Colombia

Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]



[Inicio](#) | [¿Qué hacemos?](#) | [¿Qué es CAOBA?](#) | [Contacto](#)



Resultados científicos de los proyectos de la ALIANZA CAOBA 2018

Resultados científicos de los proyectos d...
Ver más tarde Compartir

EnVivo Universidad EAFIT

IA
INTELIGENCIA ARTIFICIAL CIENCIA VS. FICCIÓN

¿Qué es CAOBA?

Centro de Excelencia y apropiación en Big Data y Data Analytics

Primer acuerdo público – privado que promueve el Big Data y la analítica en Colombia.

Nace de la convocatoria
Colciencias-MinTIC 2015

En Operación desde 2016



¿Quiénes somos?

Participan:



El futuro
es de todos

DNP
Departamento
Nacional de Planeación



El futuro digital
es de todos

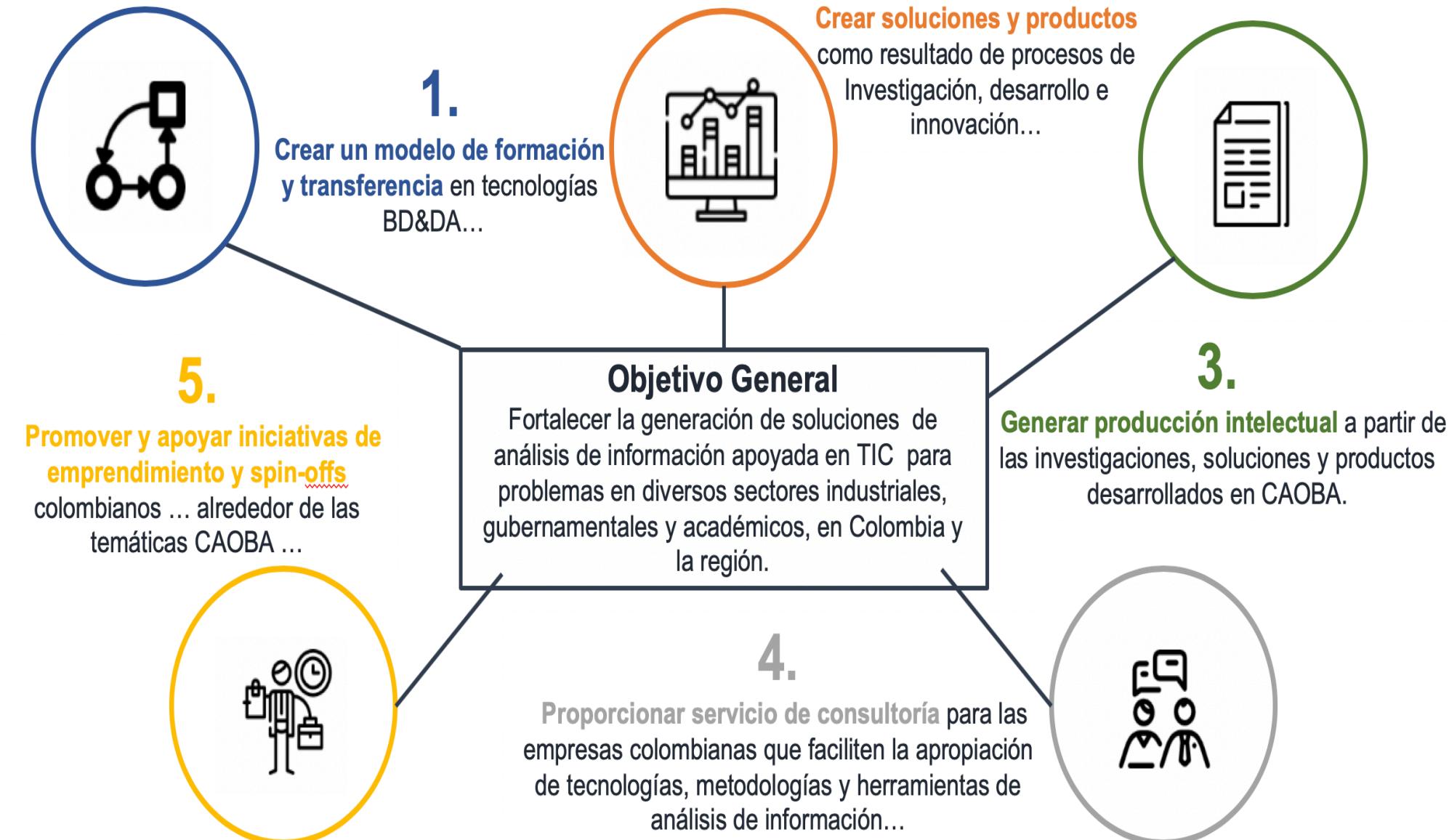
MinTIC



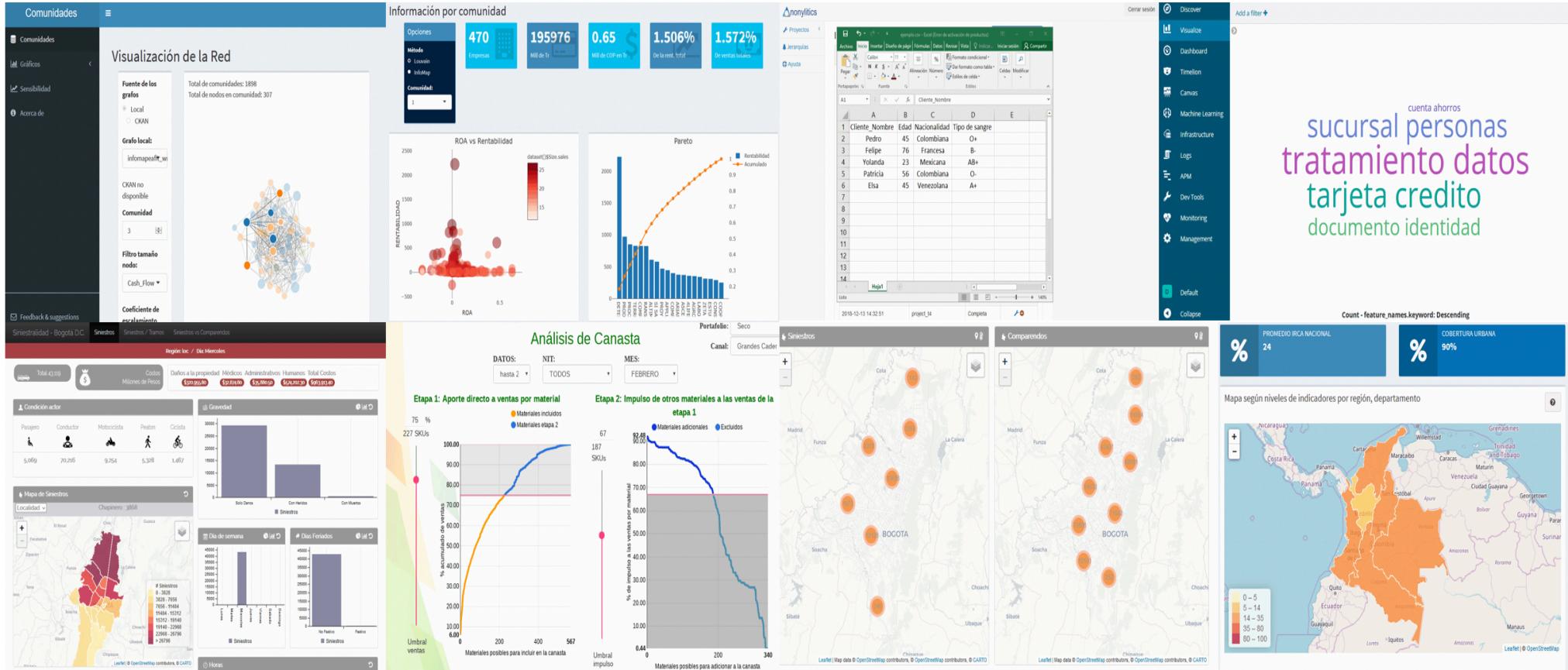
El conocimiento
es de todos

Colciencias

Objetivos CAOBA



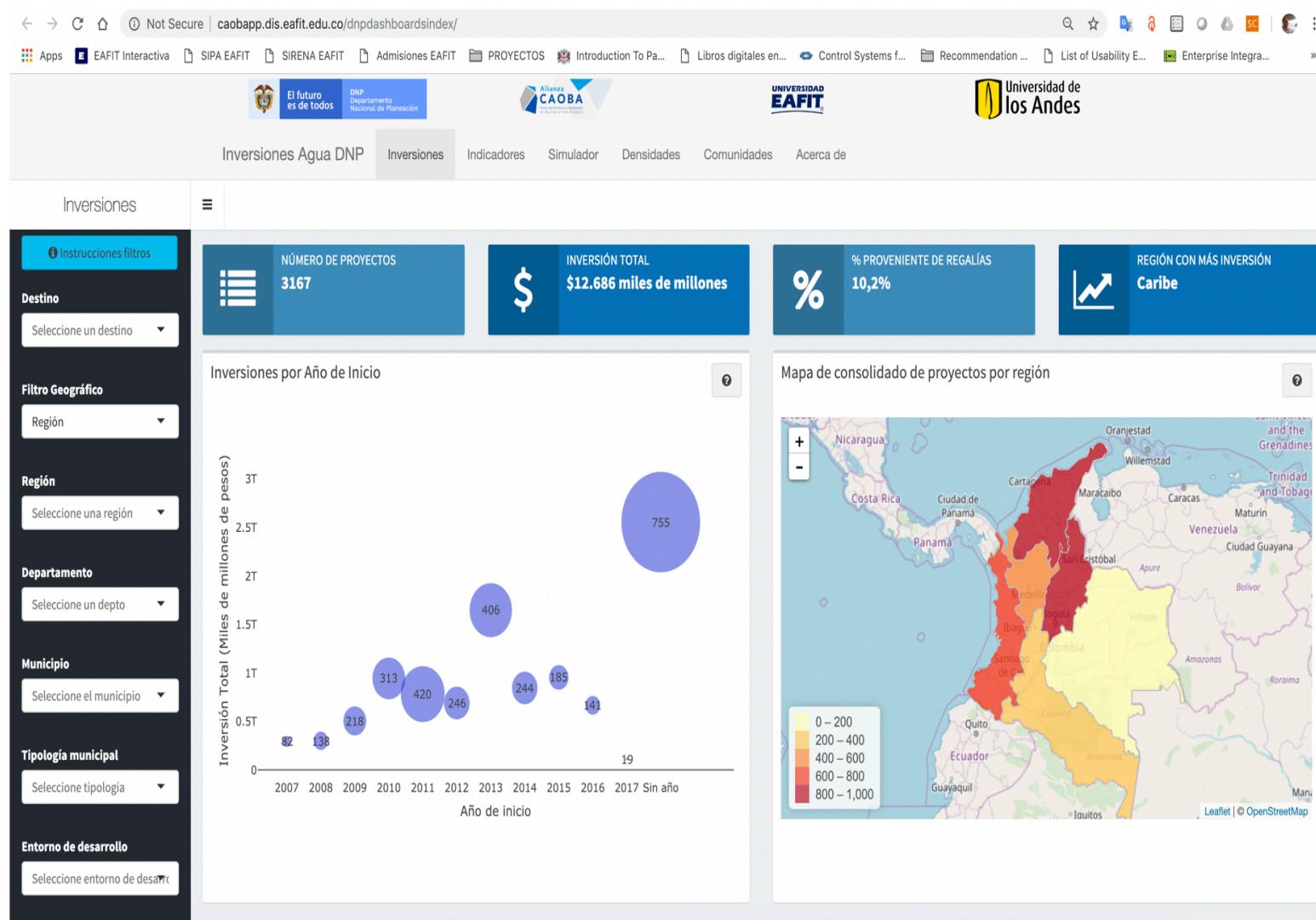
Resultados de la investigación aplicada



cuenta ahorros
sucursal personas
tratamiento datos
tarjeta crédito
documento identidad

INVERSIONES AGUAS DNP

Cuantificación del efecto de las inversiones en el sector de agua potable en la prestación del servicio de agua.



CENTROS PARA LA 4a Revolución Industrial



El Centro es una plataforma para discutir cuestiones éticas, valores y regulación de tecnologías de la Cuarta Revolución Industrial, tales como Internet de las cosas (IoT) y la inteligencia artificial / aprendizaje automático. Los proyectos involucran a múltiples partes interesadas, incluyendo a los reguladores, para desarrollar marcos de políticas que puedan ser aplicados a través de industrias y fronteras nacionales.

<https://es.weforum.org/focus/centro-para-la-cuarta-revolucion-industrial>

13 de diciembre de 2017

200 ciudadanos podrán formarse en analítica de datos y TI con la convocatoria de Científicos de Datos

Hasta el 30 de enero de 2018, profesionales y tecnólogos en ingeniería, administración, economía, matemáticas, estadística o carreras afines, podrán postularse a esta convocatoria, que busca fortalecer la productividad y competitividad de la industria TIC nacional.



La convocatoria de Científicos de Datos del Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC), Colciencias y el Centro de Excelencia y Apropiación en Big Data y Data Analytics (CAOBA), busca la formación y certificación de 200 ciudadanos colombianos como Citizen Data Scientists, en un curso de 70 horas de duración.

Información para

Ciudadano

Industria

Emprendedores

Sector Académico

Sector Gobierno

Categorías VUTIC

Registros

Espectro Radioeléctrico

Radiodifusión Sonora

Servicios Postales

RABCA:
Radioaficionados y
Banda Ciudadana

Verificación IMEI /
Autorización Venta
Terminales Móviles



300
Profesionales

se formarán en una de las habilidades necesarias para afrontar los retos que trae la Cuarta Revolución Industrial.

Formación en
**Ciencia
de Datos**

El futuro digital es de todos MinTIC

Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]

RutaN – Medellín – Centro regional para la 4a Rev Industrial



MinTIC

Ministerio de Tecnologías
de la Información y las Comunicaciones

Grandes esfuerzos en promoción, capacitación, regulación, etc para la industria 4.0



EN MEDELLÍN SE INAUGURÓ EL
**CENTRO PARA LA CUARTA
REVOLUCIÓN INDUSTRIAL**

[CONOCE MÁS](#)

**Medellin, sede del Centro para LA de la
4a Rev industrial**

<https://www.rutanmedellin.org/es/cuarta-revolucion-industrial>

Algunos programas de Especialización y Maestría

Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]



**Especialización en
Inteligencia
Artificial**

COMING SOON

Área Curricular de Ingeniería de Sistemas e Informática
Facultad de Minas

Recuerdame Recuérdame

Universidad Nacional de Colombia

UnalMedellin



Estudiar en EAFIT Academia Investigación Bienestar y cultura International Proyección Acerca de EAFIT

Maestría en Ciencias de los Datos y Analítica

Inicio La maestría ▾ Estructura académica Contacto



<http://www.eafit.edu.co/maestria-ciencias-datos-analitica>



Universidad de los Andes Colombia

Facultad de Ingeniería Departamento de Ingeniería Industrial

Inicio Departamento Programas Académicos Investigación Servicios Nuestro equipo Boletines

Usuario Contraseña Recuérdame Identificarse

Maestría en Inteligencia Analítica para la Toma de Decisiones (Analytics)

Código SNIES: 104198

MAESTRÍA EN INTELIGENCIA ANALÍTICA PARA LA TOMA DE DECISIONES (Analytics)

Home Contenido Objetivo

<https://industrial.uniandes.edu.co/es/programas-academicos/maestria/maestria-en-inteligencia-analitica-para-la-toma-de-decisiones>

Inspira Crea Transforma



<https://www.icesi.edu.co/facultad-ingenieria/maestria-en-ciencia-de-datos>

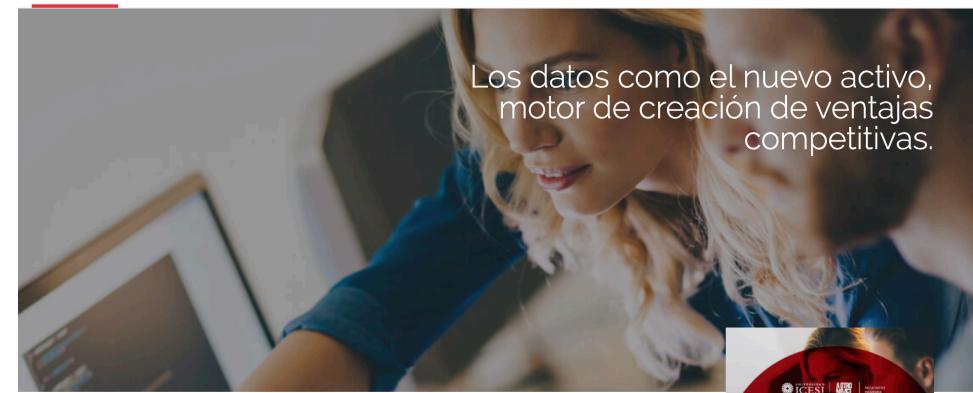
Maestría en Ciencia de Datos

Facultad de Ingeniería

LA MAESTRÍA RAZONES PLAN DE ESTUDIO ADMISIÓN

APLICAR

Los datos como el nuevo activo, motor de creación de ventajas competitivas.



MAESTRÍA EN ANALÍTICA PARA LA INTELIGENCIA DE NEGOCIOS

SNIES 105238



"Nuestra tradición es construir el futuro"

El programa de Maestría en Analítica para Inteligencia de Negocios es una iniciativa interdisciplinaria para formar consultores en análisis de datos para las organizaciones que apoyen la toma de decisiones estratégicas.

El Programa tiene una mezcla única de formación en:
Capacidades analíticas y de ciencia de los datos.
Capacidades de gestión de información.
Capacidades administrativas, organizacionales y de comunicación.

<https://www.javeriana.edu.co/maestria-analitica-para-inteligencia-de-negocios>

Inspira Crea Transforma

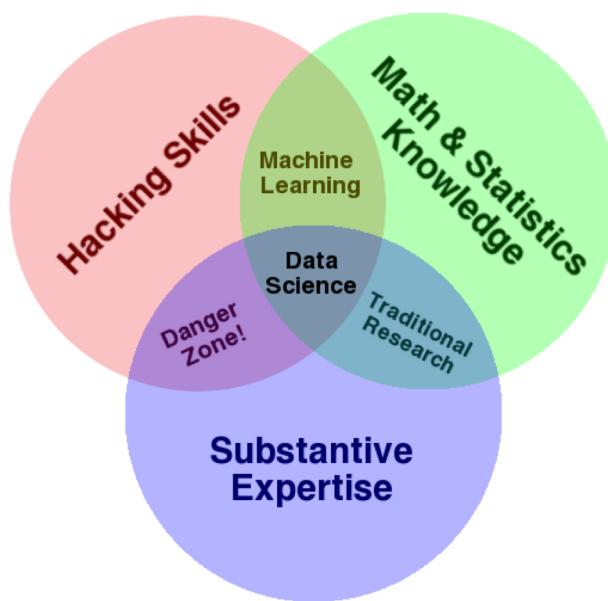
UNIVERSIDAD
EAFIT[®]

Perfiles profesionales en Big Data y Analítica



Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]

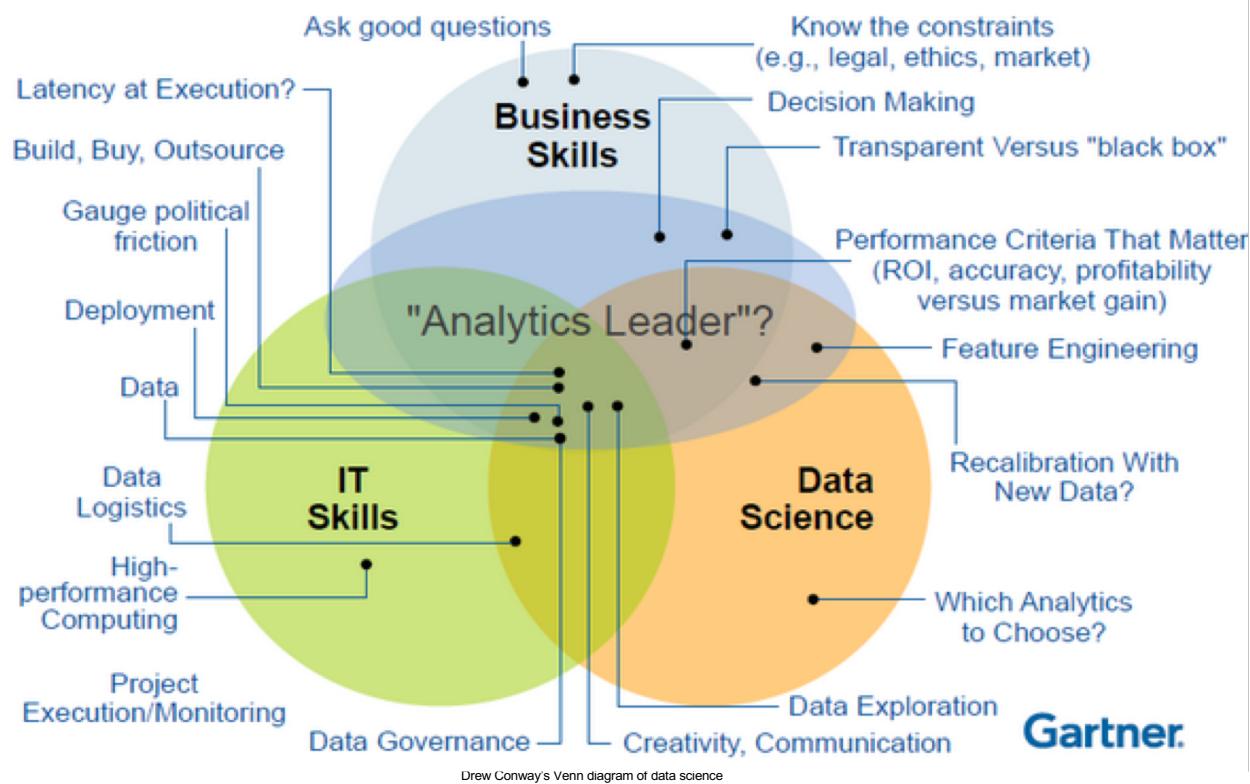


Drew Conway's Venn diagram of data science

Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]

Driving the Success of Data Science Solutions: Skills, Roles and Responsibilities ...



Data Scientist
also known as Data Managers, statisticians.



A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication



Will use programmes such as:
SQL, Python, R

Data Engineers
also known as database administrators and data architects.



They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

Skills: Programming, Mathematics, Big data



Will use programmes such as:
Hadoop, NoSQL, and Python

Data Analysts
also known as business Analysts.



They typically help people from across the company understand specific queries with charts.

Skills: Statistics, Communication, Business knowledge



Will use programmes such as:
Excel, Tableau, SQL

DATA Engineer

Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.

DATA Scientist

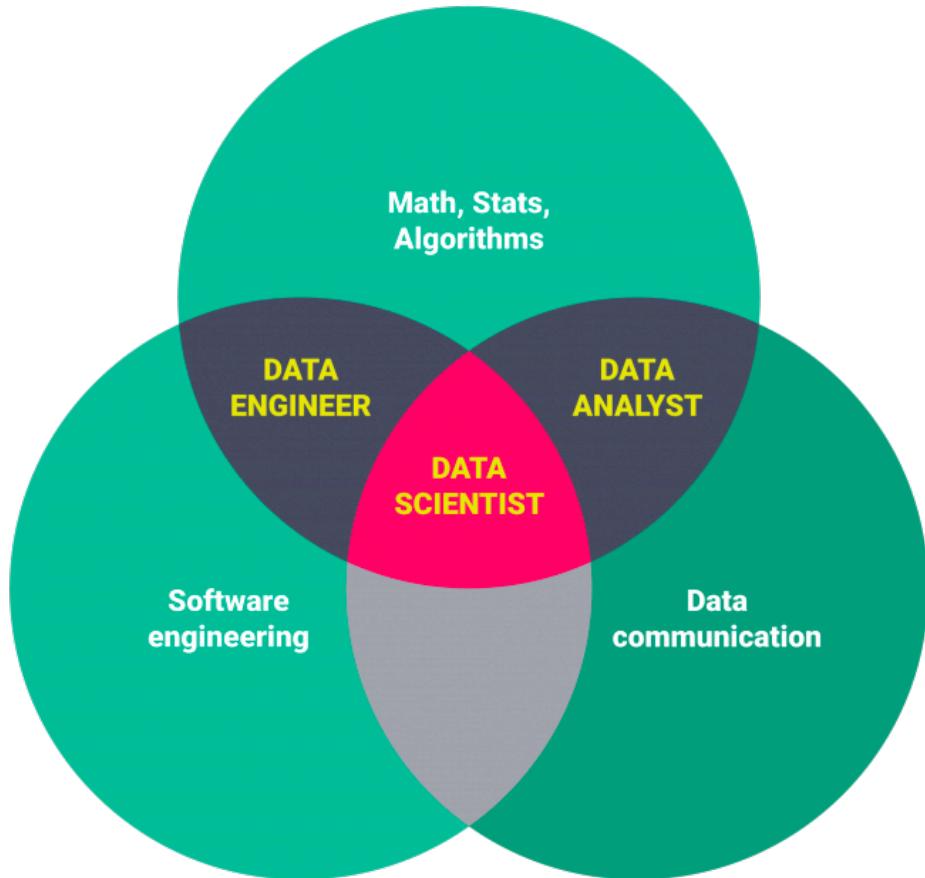
Cleans, massages and organizes (big) data. Performs descriptive statistics and analysis to develop insights, build models and solve a business need.

DataCamp
Learn Data Science By Doing

The graphic illustrates two professionals at their desks. On the left, a man with a beard and headphones is working on a laptop, surrounded by icons representing databases and processing systems. On the right, a woman with glasses is working on a laptop, surrounded by icons representing data cleaning, organization, and analysis. The background features a grid of various data visualization charts like bar graphs, line graphs, and pie charts.

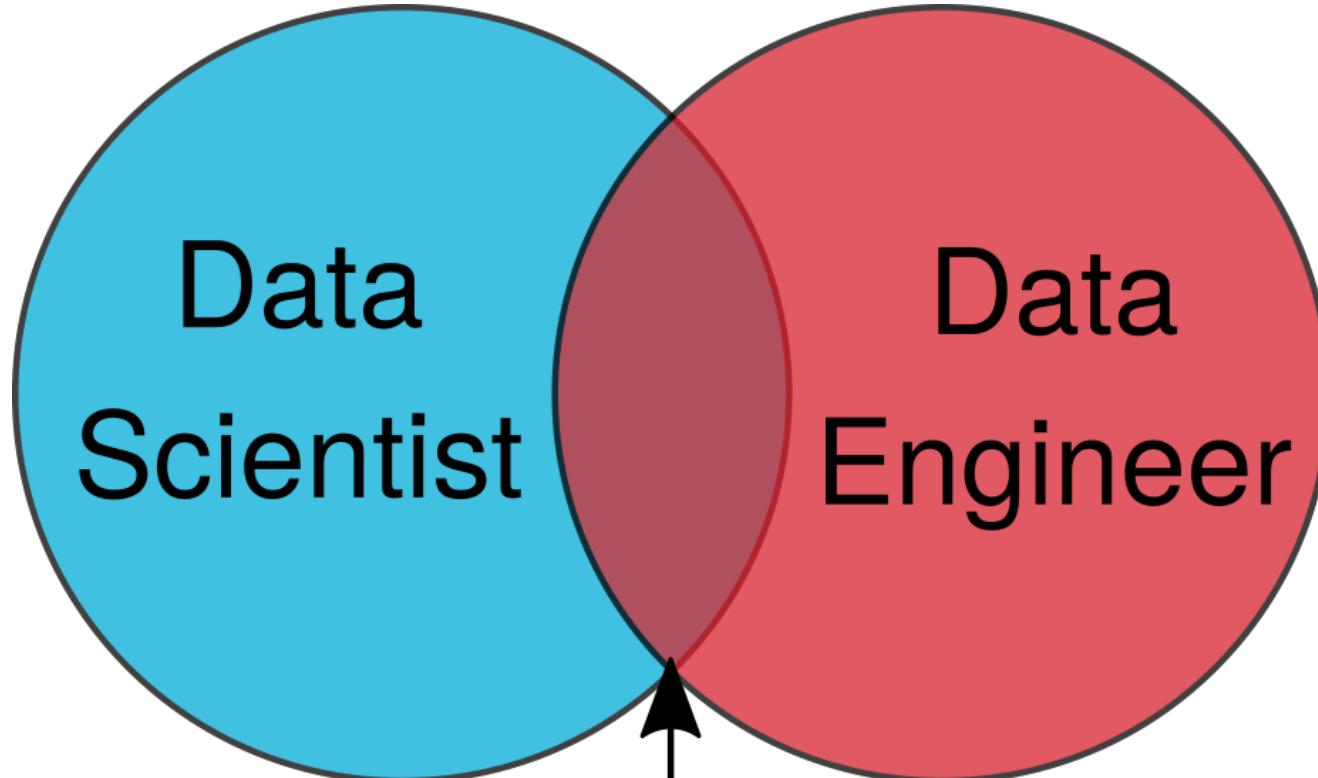
Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]



Inspira Crea Transforma

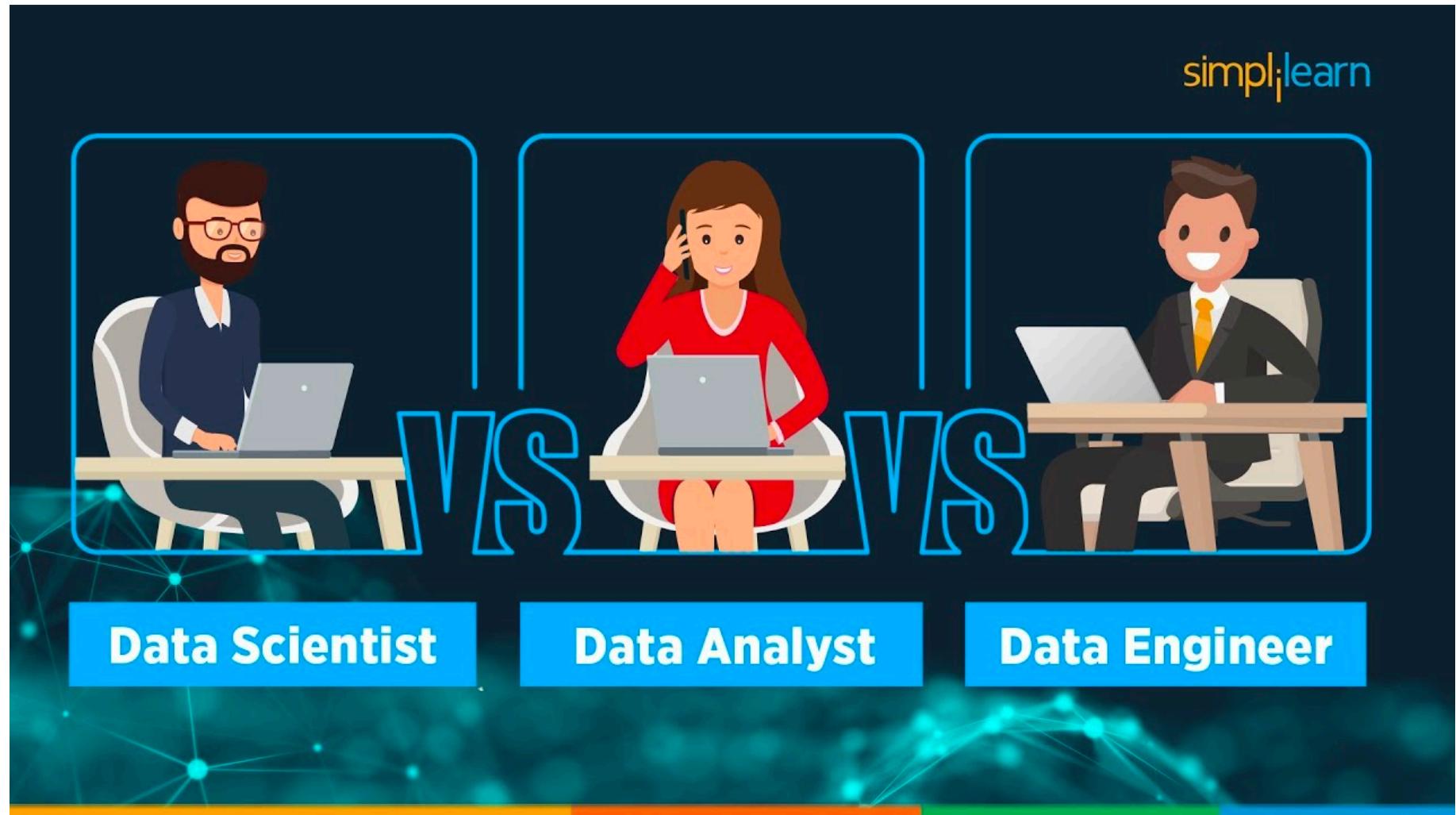
UNIVERSIDAD
EAFIT[®]



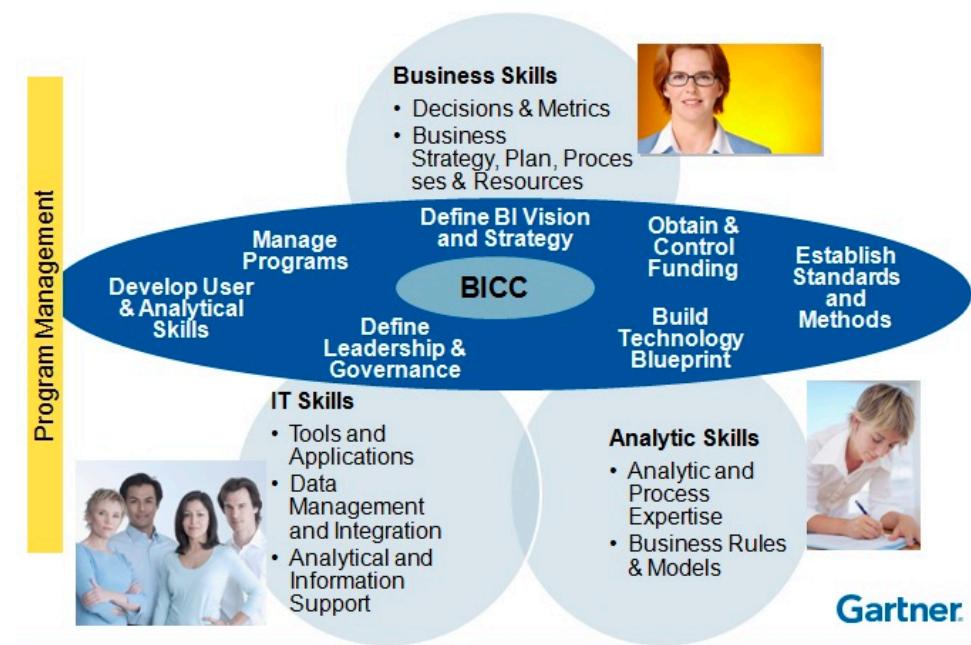
Big Data

Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]



BICC de Gartner



Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]

Conclusión

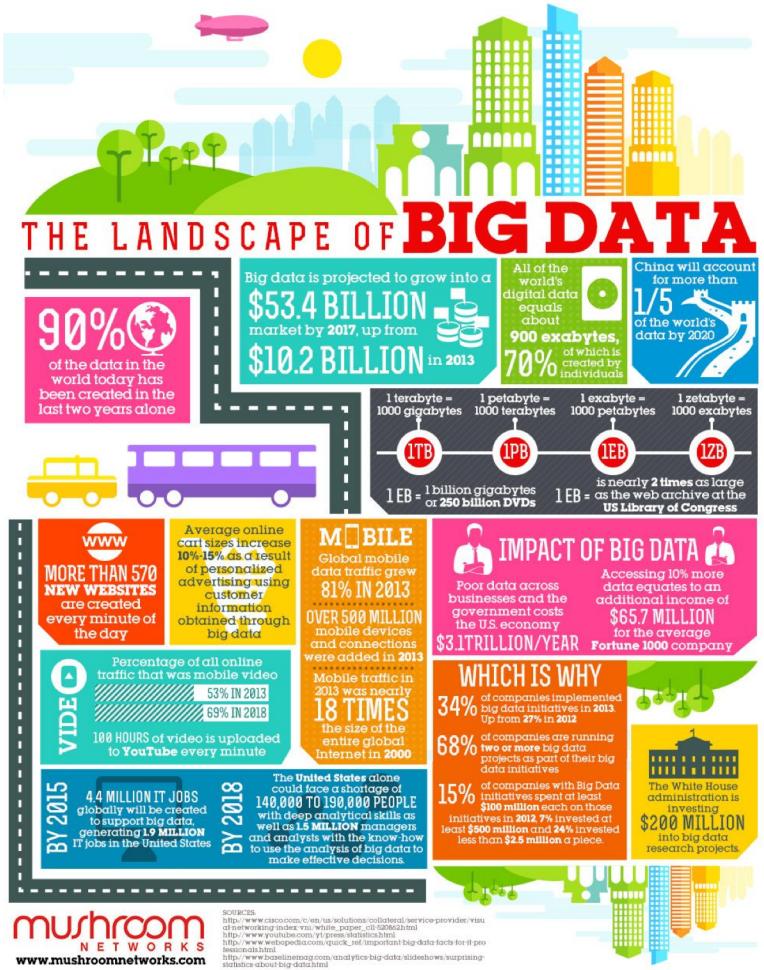
- Una sola persona NO puede manejar todos las habilidades
- Se requiere un trabajo multi e inter disciplinario
- ¿Tú como quieras jugar en este mundo del Big Data?
 - Desde la Tecnología
 - Desde la Ciencia de los datos
 - Desde el Negocio

ESTE TUTORIAL ESTA ENFOCADO A?

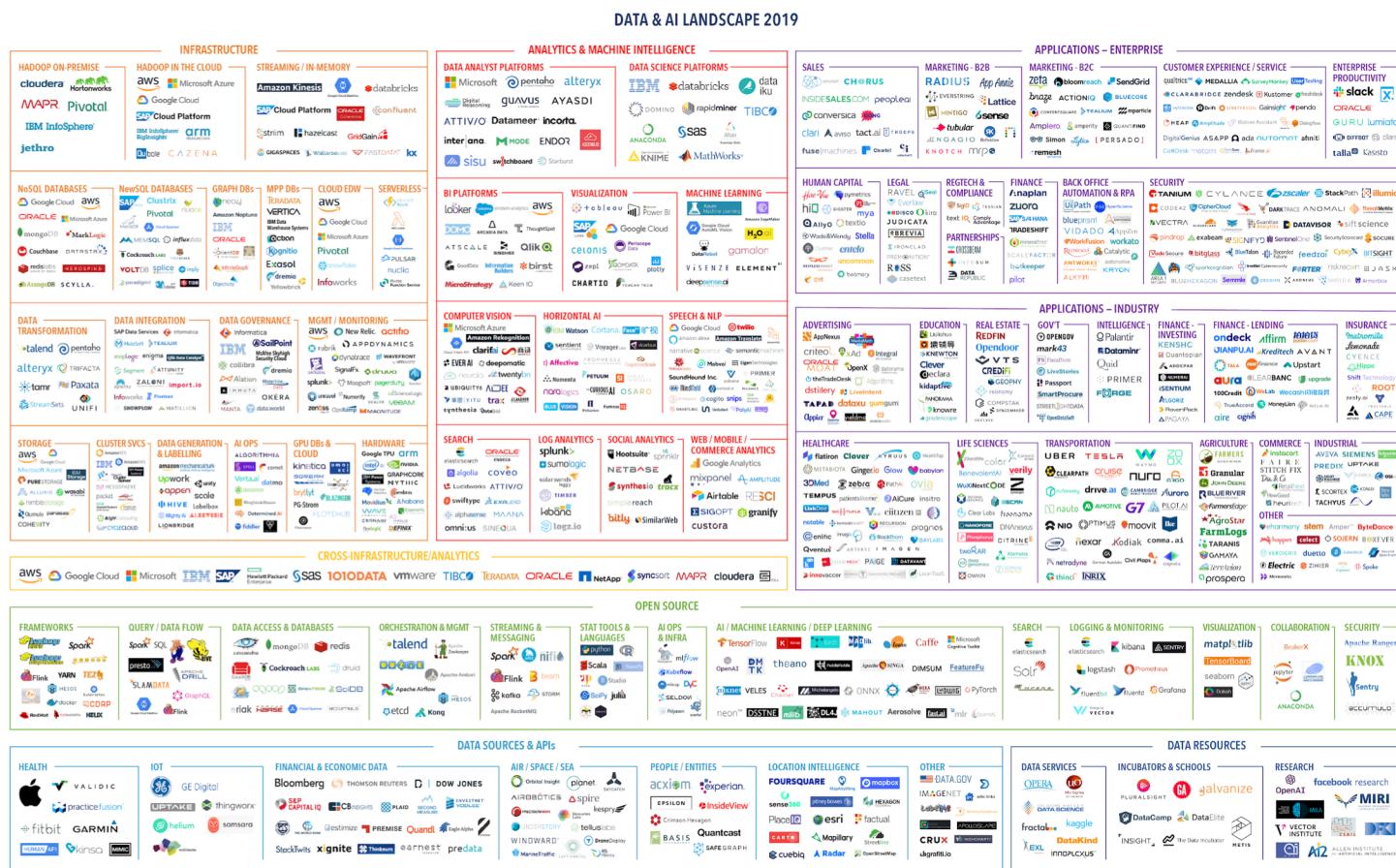
Ingeniería de Datos

Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]



Inspira Crea Transforma



July 16, 2019 - FINAL 2019 VERSION

© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap)

mattturck.com/data2019

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Inspira Crea Transforma

**UNIVERSIDAD
EAFIT**®

1. Introducción a Big Data

Inspira Crea Transforma

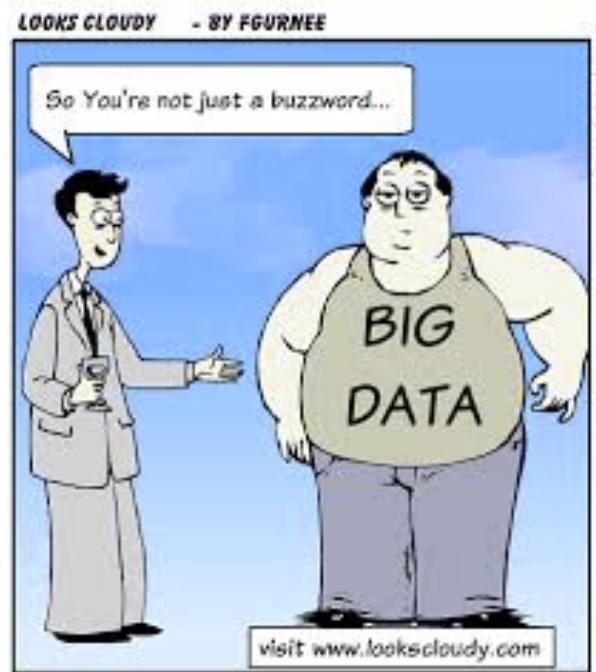


Qué es Big Data

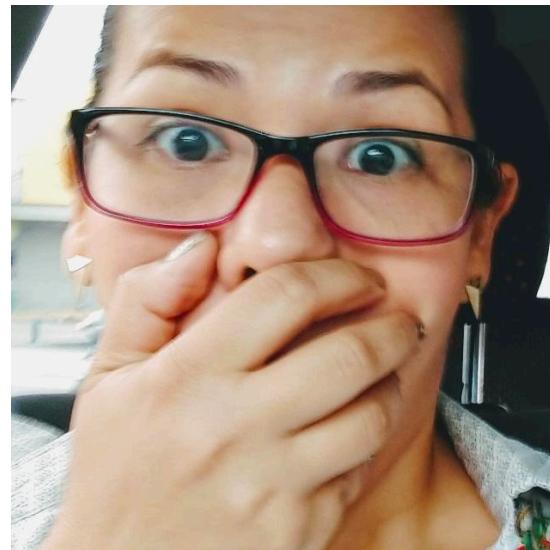
Es el área de conocimiento dentro de la informática que se ocupa de solucionar problemas de transporte, almacenamiento, procesamiento, análisis y aprovechamiento de datos caracterizados como Big Data, esto ocurre especialmente cuando las tecnologías tradicionales no son suficientes.

¿Qué es Big Data?

- Un Buzzword?



¿El tamaño si importa?



Porqué???

Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]

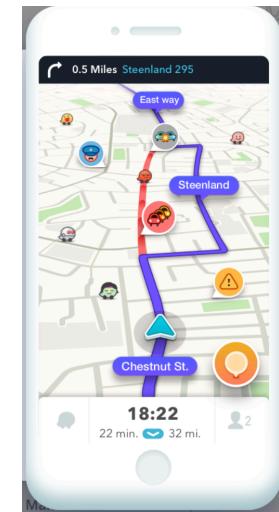
Big (Grandes) o Many (Muchos)

- El colapso viene por la cantidad, NO tanto por el tamaño
- Ej:
 - Waze



Google Maps usa cerca de 0.6MB de data por hora.

Waze usa cerca de 0.23MB por hora



Pero el tamaño si importa

- 151 millones de suscriptores, en más de 200 países.
- Donde los almacena?
 - En la Nube – S3
 - 60+ PB con 1500+ millones de objetos (algo así como películas)
 - Diario procesa más de 400 TB

NETFLIX



Características Big Data

Las 5 Vs pueden ser utilizadas para diferenciar conjuntos de datos categorizados como “Big Data” de los demás conjuntos de datos.

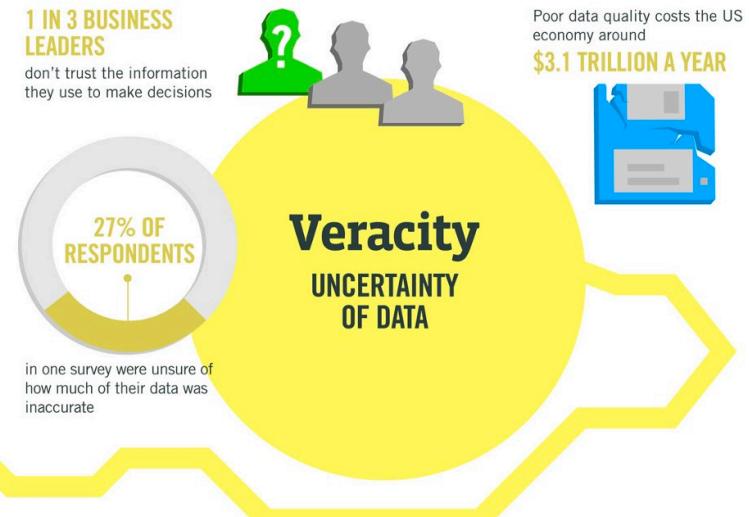
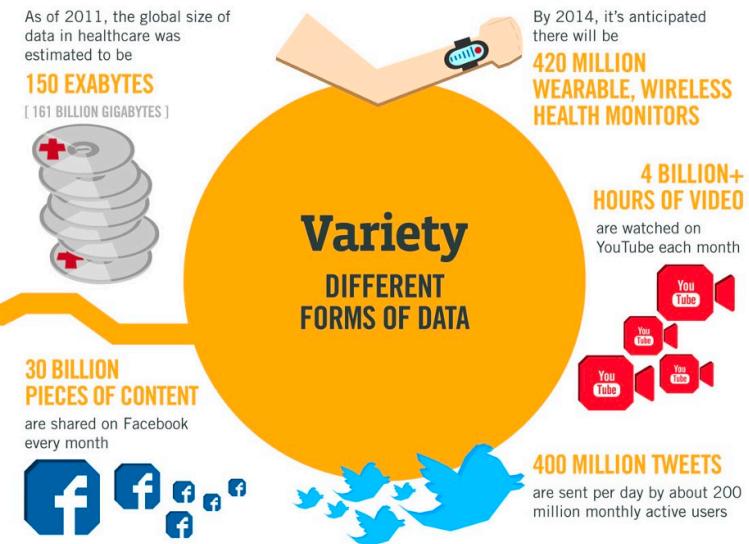
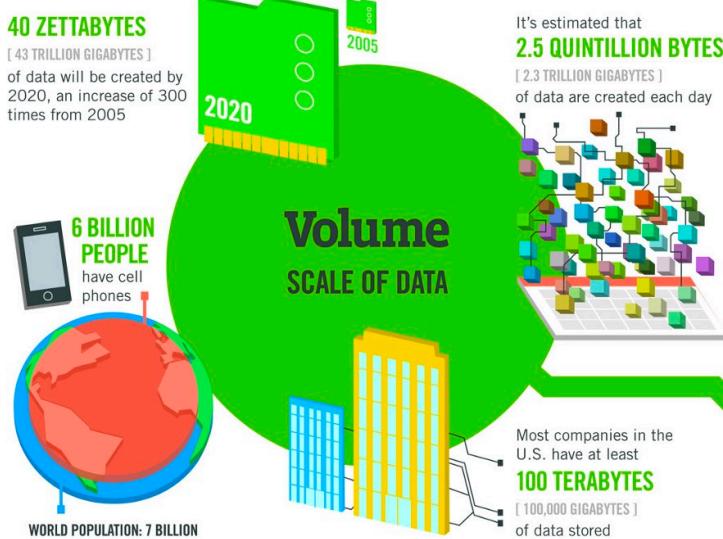
- Volumen
- Velocidad
- Variedad
- Valor
- Veracidad

Tipos de datos Big Data

Los tipos de datos que una solución Big Data puede procesar son:

- Estructurados.
- No estructurados.
- Semi-estructurados.

Big Data - V's



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Inspira Crea Transforma

Cómo reconocer y abordar un caso Big Data

Antes de iniciar la implementación de una solución Big Data debe:

- Cubrir por lo menos 2 de las 3 primeras Vs, las otras 2 Vs no son negociables.
- Proporcionar beneficios claros y de valor para el negocio.
- Implementar mecanismos para la divulgación y entrenamiento Big Data.
- Implementar modelos de gobierno Big Data.
- Implementar mecanismos de integración con otras áreas como: operaciones, seguridad, etc.
- Proveer mecanismos para la gestión, trazabilidad y lineage de los datos.

Big Data: Ciclo de vida Desde la Ingeniería de Datos

- **DATA:** Generación y Adquisición de datos
- **STORAGE:** Sistemas de almacenamiento masivo y distribuido
- **PROCESAMIENTO:** Análisis & Data Analytics
- **APLICACIONES:**
 - Negocios, Científicos, Económicos, ... para casi todos los dominios
 - Apps por: Datos estructurados, texto, web, multimedia, redes, móviles
 - Empresariales, IoT, Social Networks, Health, e-Science, Smart grid, etc.
 - Visualización

Big Data: Ciclo de vida

- **DATA:** Generación y Adquisición de datos (ETL: Extract & Transform & Load)
 - **Generación:** Fuente
 - Datos empresariales operacionales - OLTP
 - Datos IoT, Redes Sociales
 - Datos científicos en muchos otros campos
 - **Adquisición:**
 - Datasets y/o real-time -> Colecciones de datos
 - Transmisión de datos (Inter-DCN & Intra-DCN)
 - Transformación de datos & Pre-procesamiento de datos
 - Integración
 - Limpieza
 - Eliminación de redundancia

Big Data: Ciclo de vida

- **STORAGE:** Sistemas de almacenamiento masivos y distribuidos.
 - Datos transaccionales vs Datos para Analítica
 - CAP (Consistency & Availability & Partition Tolerance)
 - Almacenamiento para datos masivos
 - Sistemas de almacenamiento distribuidos
 - Mecanismos de almacenamiento:
 - Tecnologías tradicionales: DAS, SAN, NAS
 - File System (NFS, GFS, HDFS)
 - Database system (SQL & NoSQL)
 - Tradicionales -> SQL
 - Key-Value, Column-oriented, Document, etc, -> NoSQL
 - Repositorios, ej: tipo CKAN

Big Data: Ciclo de vida

- **PROCESAMIENTO:** Análisis & Data Analytics (big data analytics)
 - Análisis de datos tradicional – OLAP-DWH
 - Métodos de analítica en Big Data
 - Arquitectura para análisis de datos Big Data
 - Análisis Real-Time vs Offline (batch)
 - Análisis a diferentes niveles
 - Data Mining & ML

Big Data: Ciclo de vida

- **ANALÍTICA:** Procesamiento & Data Analytics

- Arquitectura para análisis de datos Big Data
 - Análisis Real-Time
 - Datos que cambian constantemente
 - Arquitecturas:
 - Clúster de procesamiento paralelo usando RDBMS tradicionales
 - Plataformas de computación basadas en memoria (ej: greenplum de EMC, HANA de SAP, Spark)
 - Offline (batch)
 - El tiempo de respuesta no es critico
 - Hadoop, Scribe, Kafka, Chukwa, etc.
 - Hibridas -> lamda

Big Data: Ciclo de vida

- **APLICACIONES:**

- Evolución de las apps:
 - Comerciales
 - Científicas
 - Sociales

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
- HortonWorks
 - www.hortonworks.com
- MapR
 - www.mapr.com

Apache Hadoop
Software



cloudera



LIDERES EN NUBE

- Amazon
- Google
- Microsoft
- IBM



Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]

Zoologico de Hadoop



Apache Ambari



Apache Flink



Apache Mahout



APACHE ZooKeeper™

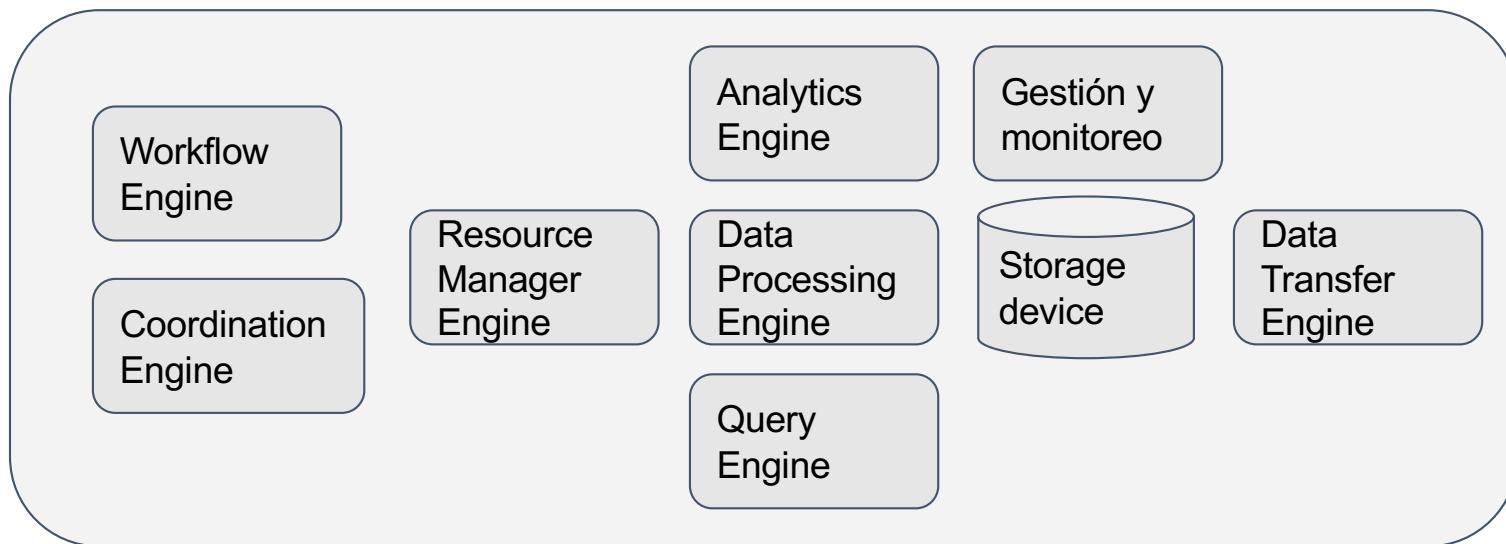
Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]

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

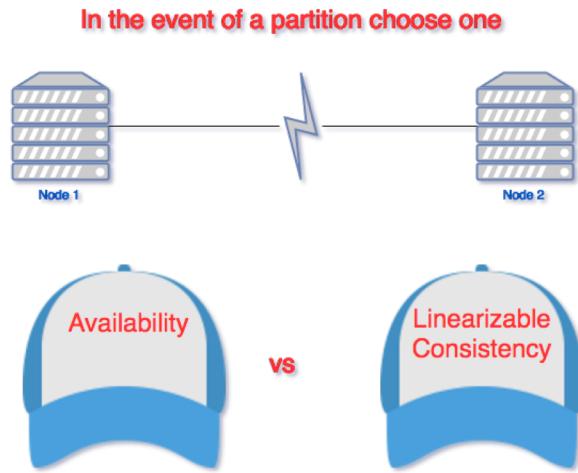
Side-by-side comparison: Object vs. traditional storage			
	OBJECT STORAGE	FILE-BASED STORAGE	BLOCK-BASED STORAGE
Transaction units	Objects, that is, files with custom metadata	Files	Blocks
Supported type of update	No in-place update support; updates create new object versions	Supports in-place updates	Supports in-place updates
Protocols	REST and SOAP over HTTP	CIFS and NFS	SCSI, Fibre Channel, SATA
Metadata support	Support of custom metadata	Fixed file-system attributes	Fixed system attributes
Best suited for	Relatively static file data and as cloud storage	Shared file data	Transactional data and frequently changing data
Biggest strength	Scalability and distributed access	Simplified access and management of shared files	High performance
Limitations	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[®]

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

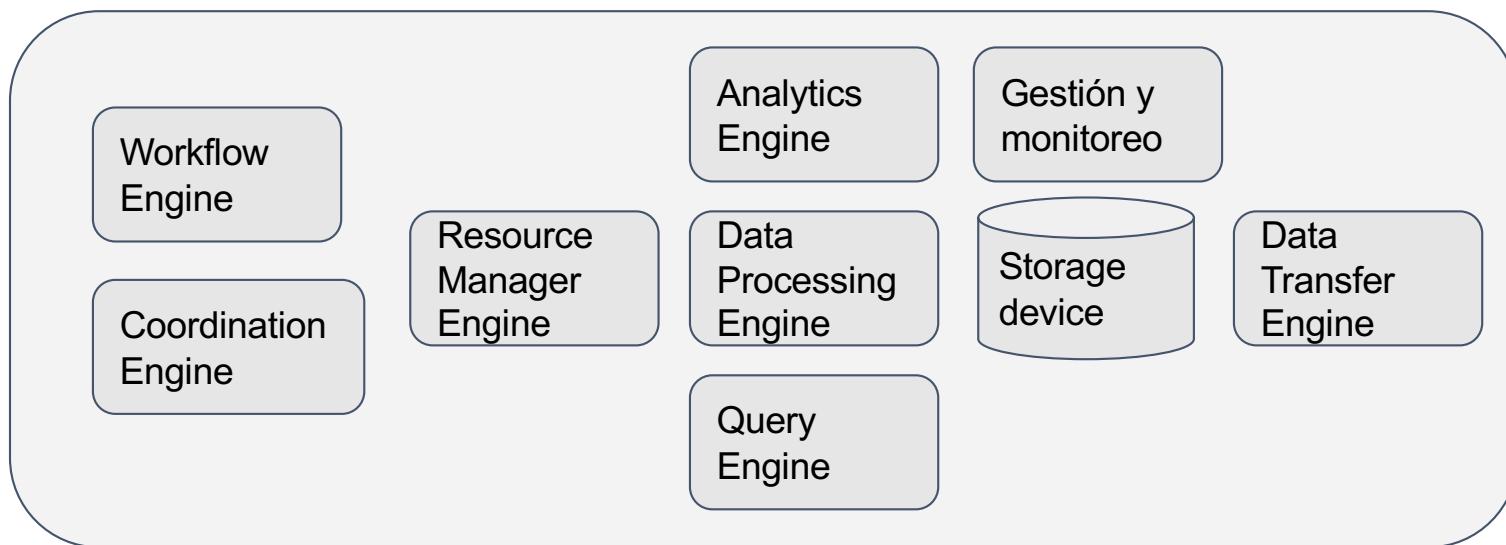
- Este motor es responsable por la monitorización y la coordinación entre las tareas de los distintos componentes de una plataforma Big Data.
- Una solución distribuida de Big Data que deba ejecutarse en varios servidores depende de un motor de coordinación, a fin de garantizar la consistencia operativa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

3.Oferta de soluciones Big Data

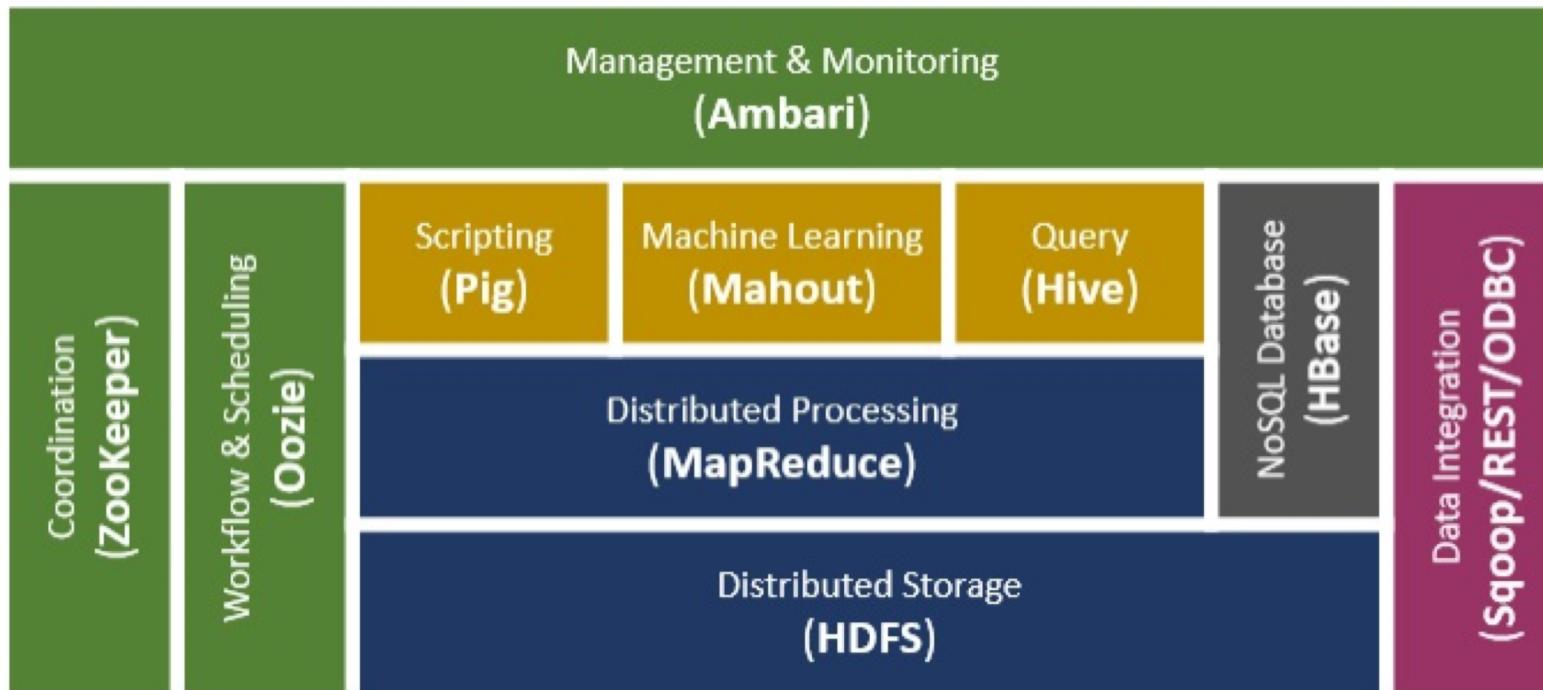
Inspira Crea Transforma



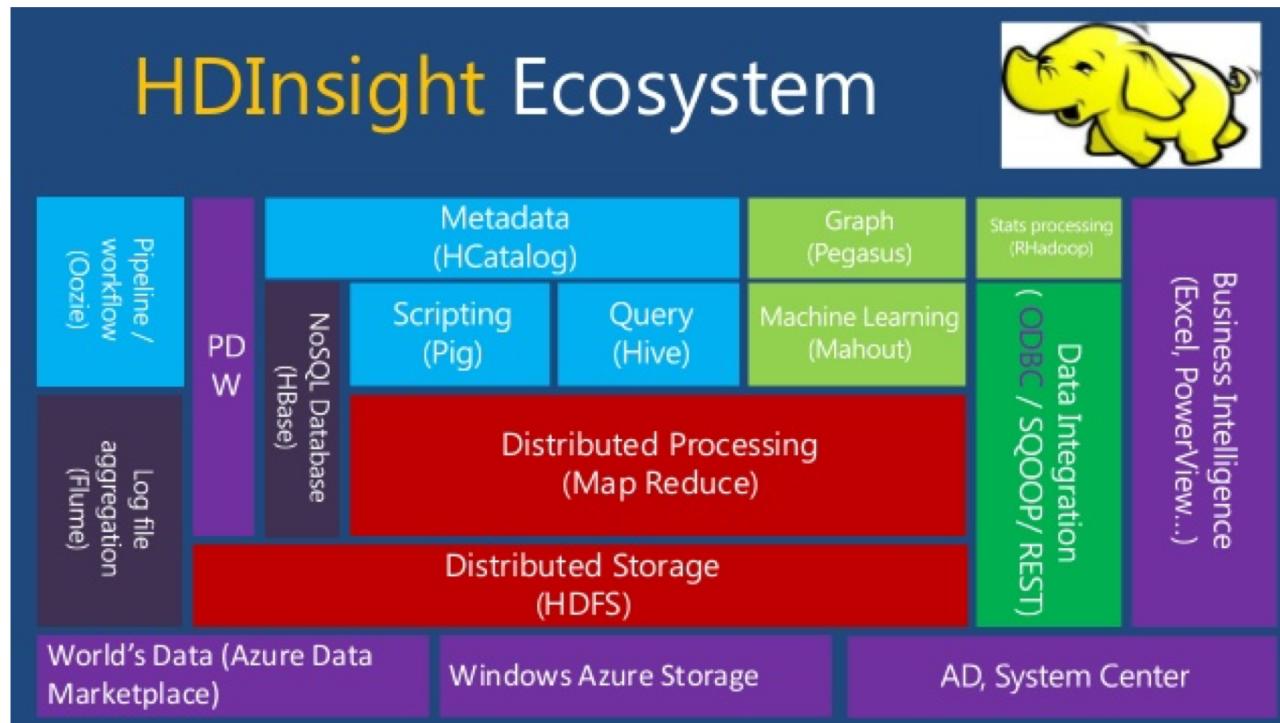
Ecosistema tecnológico de una solución Big Data



Ecosistema Hadoop (estándar)

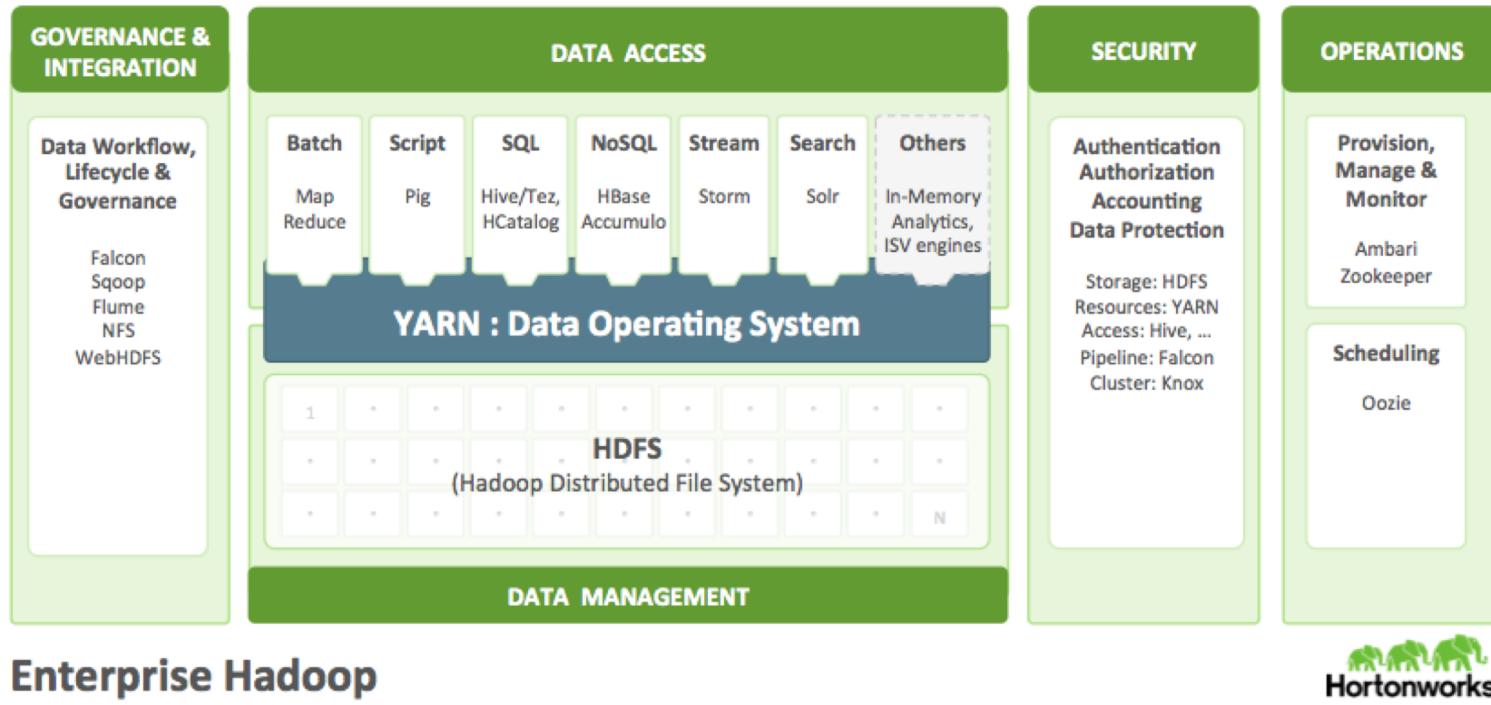


Ecosistema Hadoop: Azure

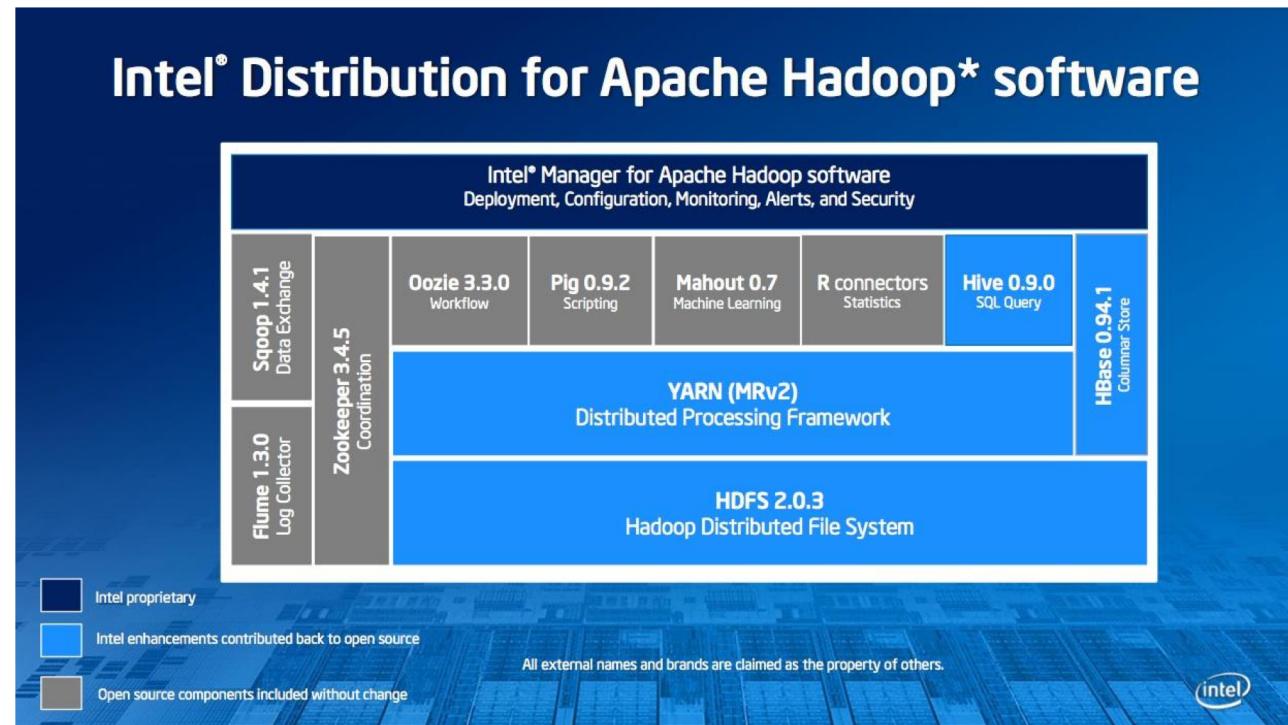


Inspira Crea Transforma

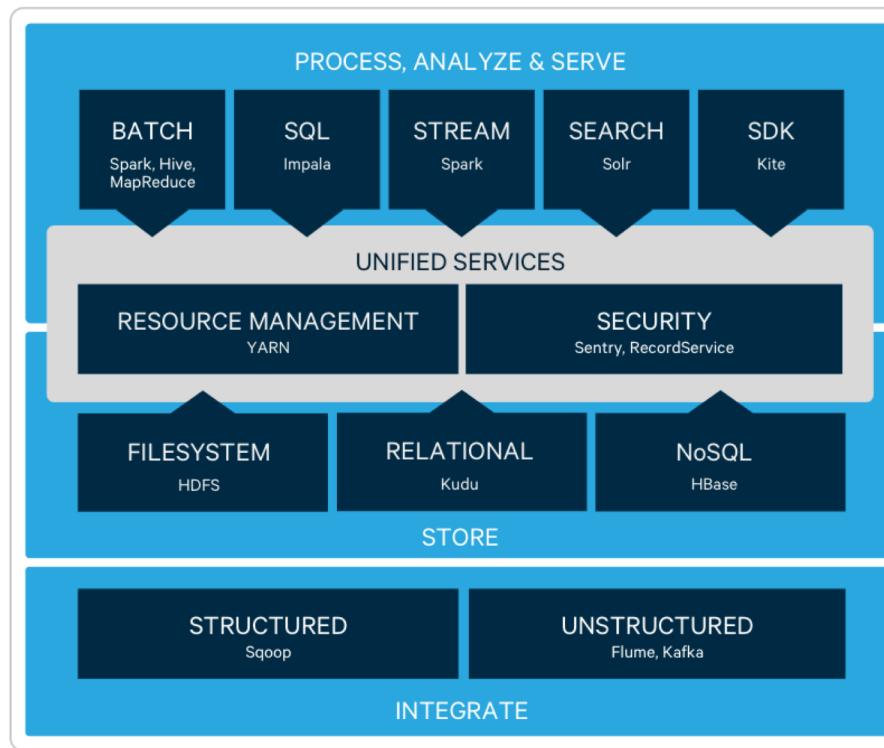
Ecosistema Hadoop: Hortonworks



Ecosistema Hadoop: Intel

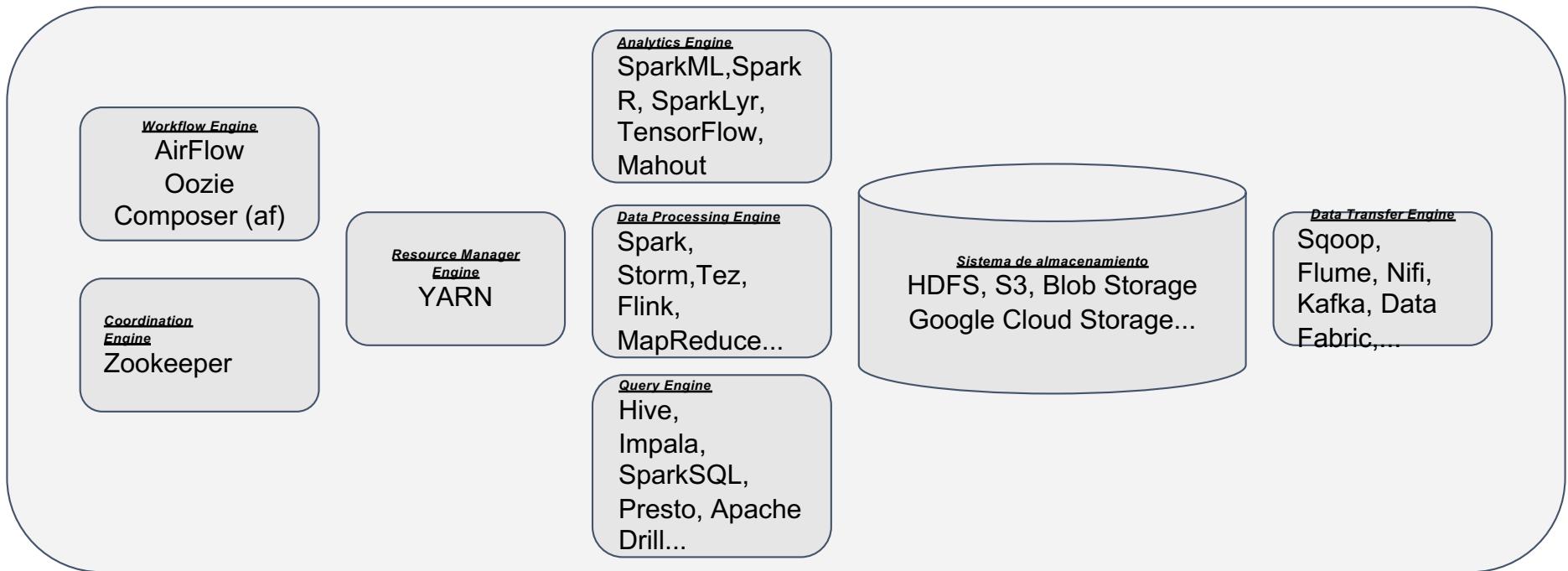


Ecosistema Hadoop: Cloudera



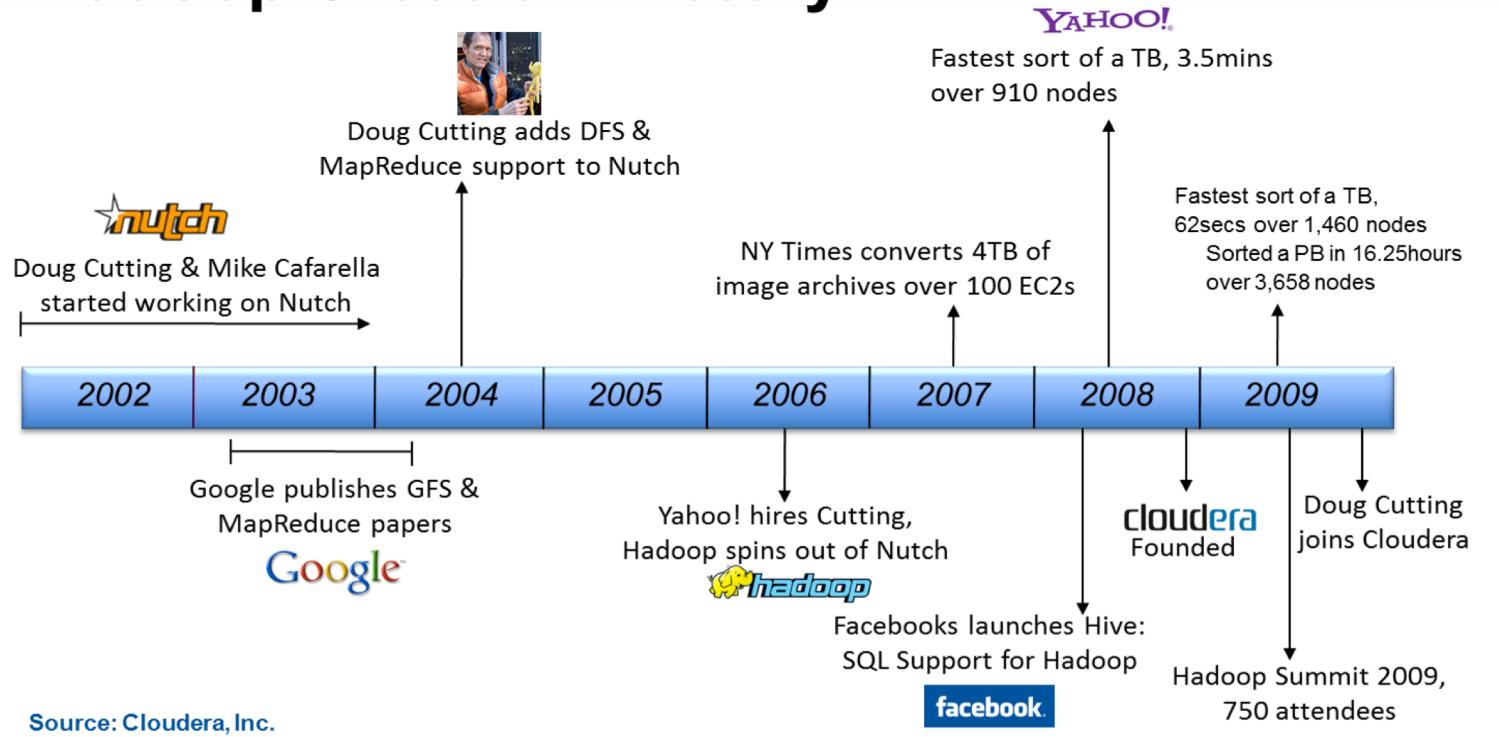
Inspira Crea Transforma

Ecosistema Big Data



6. Ecosistema Hadoop

Hadoop Creation History



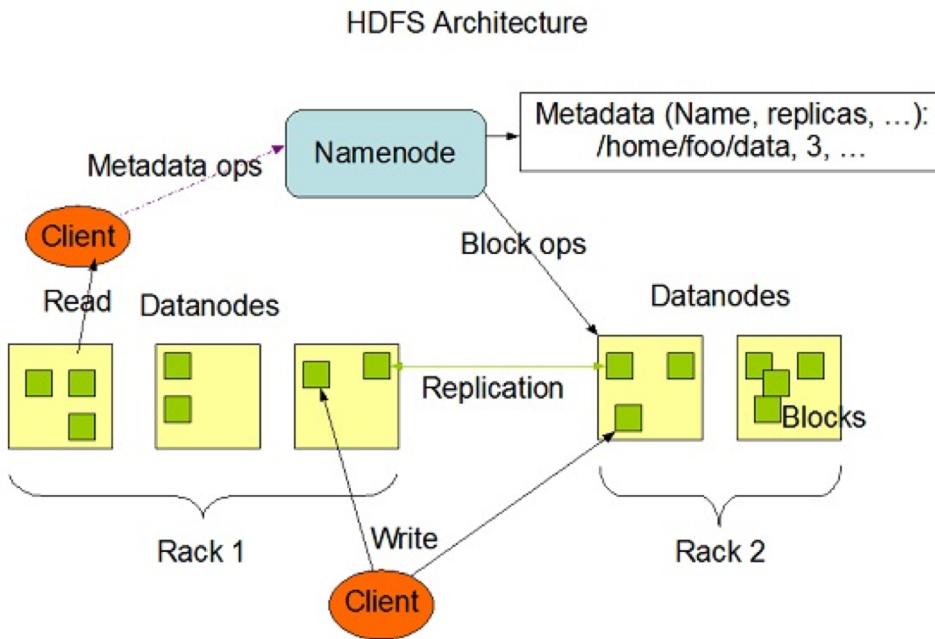
Hadoop

- 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.

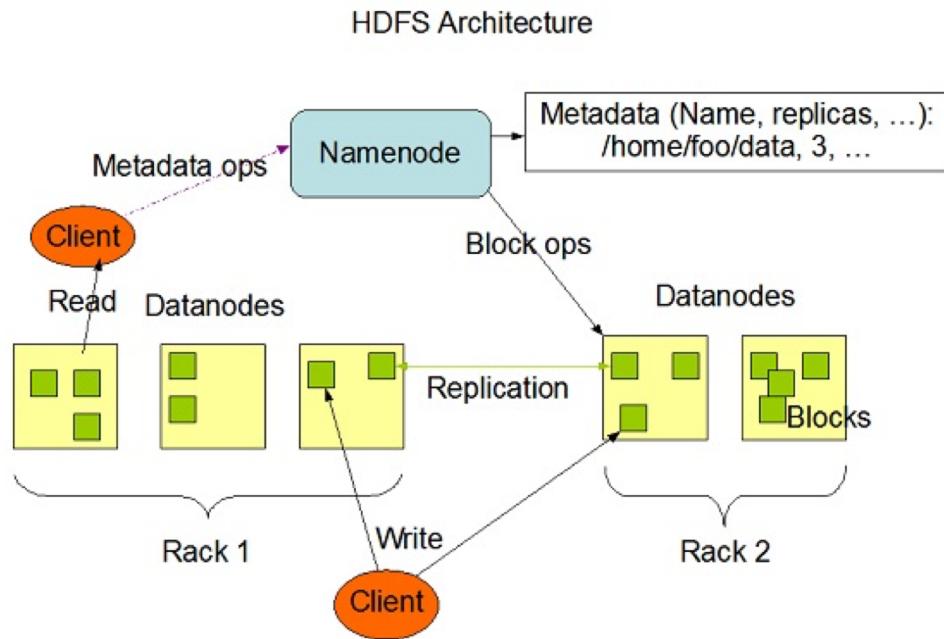
Los datos de la aplicación se almacenan en los servidores DataNode.

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

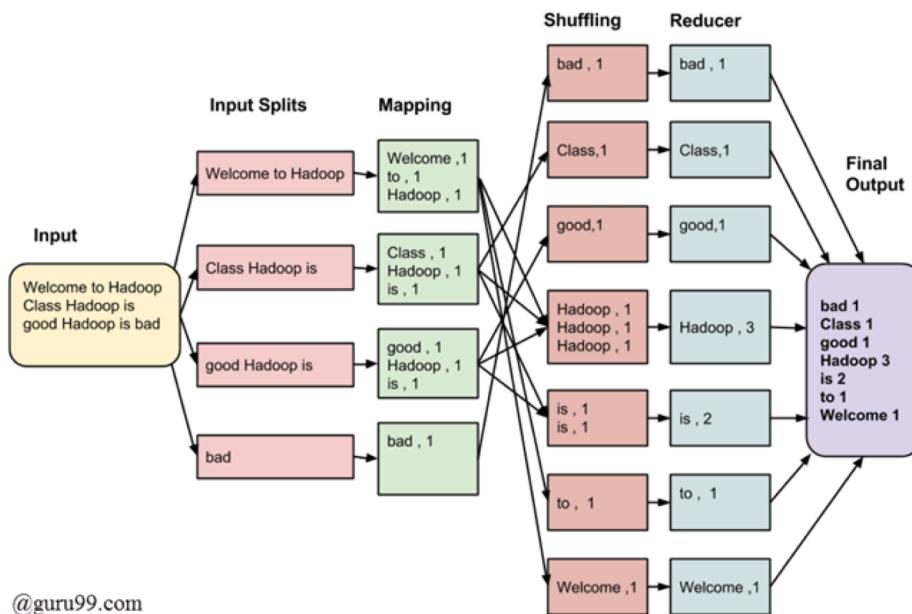


Todos los archivos y directorios en HDFS están representados en el NameNode por Inodes que contienen atributos como: permisos, marca de modificación, cuota, etc.

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

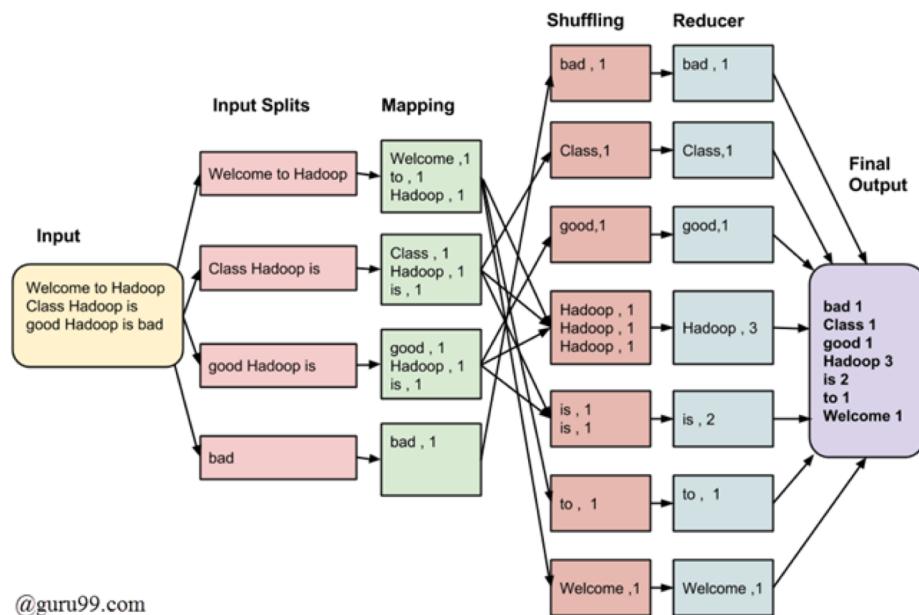
El cliente envía una tarea al JobTracker

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

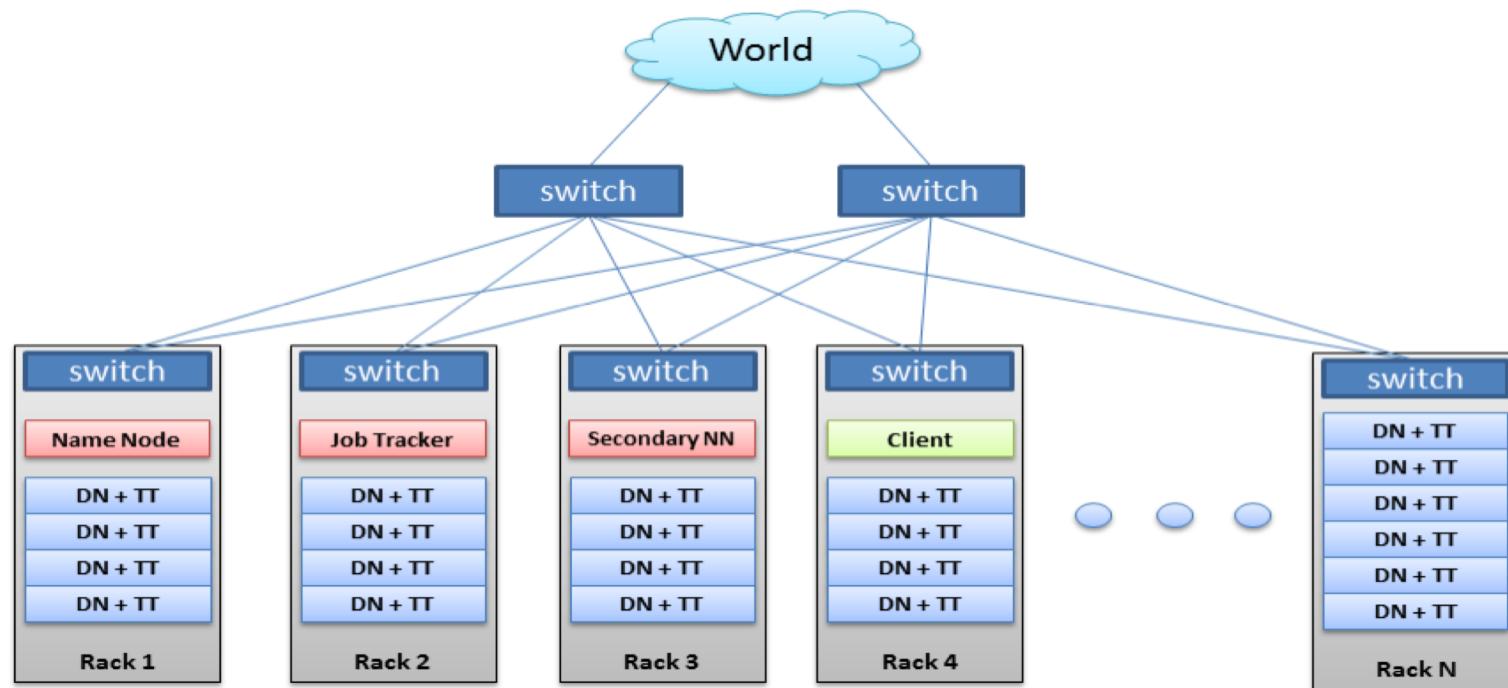


El JobTracker informa al cliente sobre la terminación de la tarea.

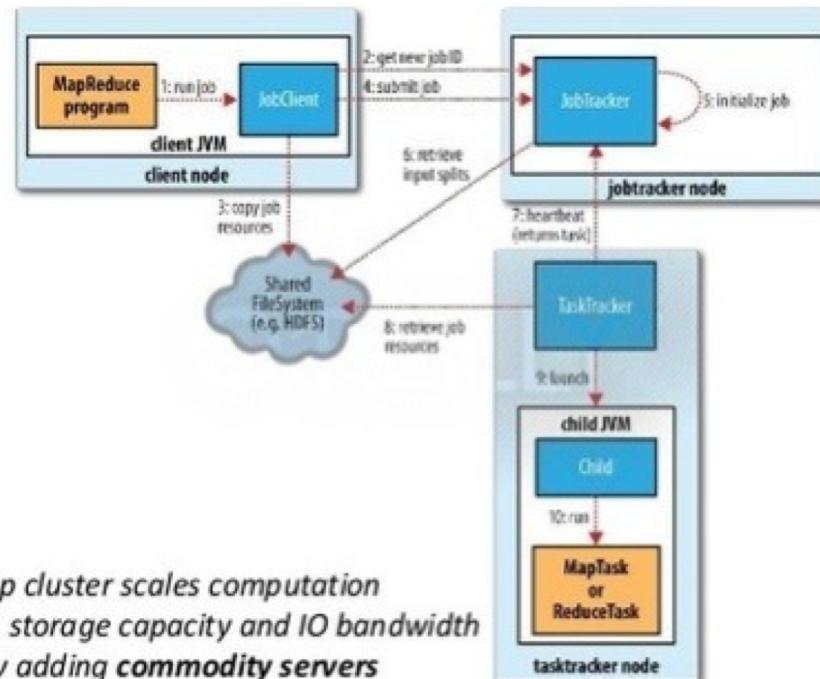
El cliente contacta al NameNode y recibe los resultados.

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

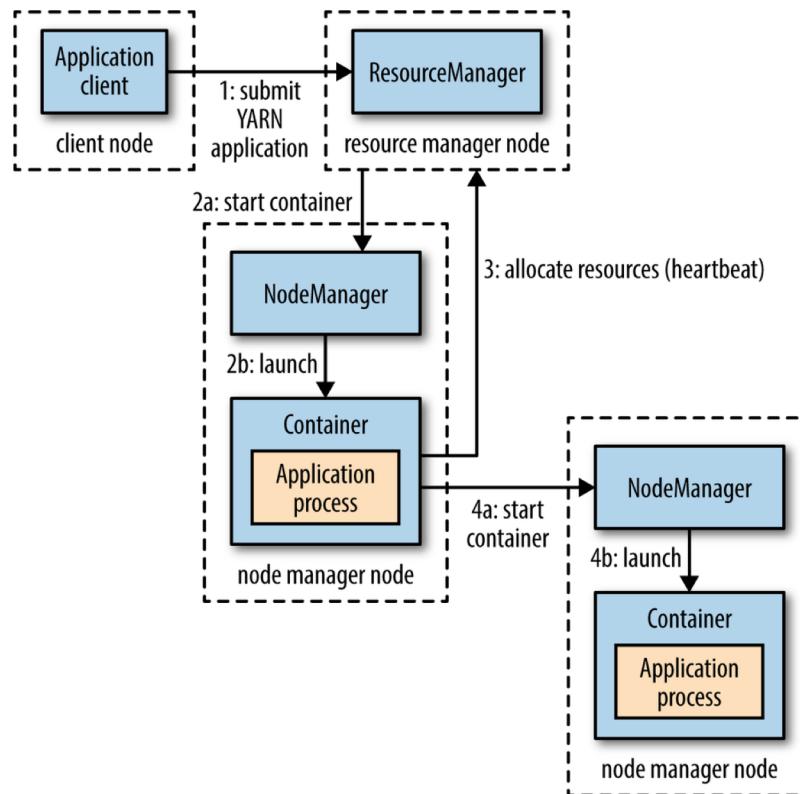
Hadoop Cluster v1



Arquitectura de Hadoop v1



Arquitectura de Hadoop v2 -YARN

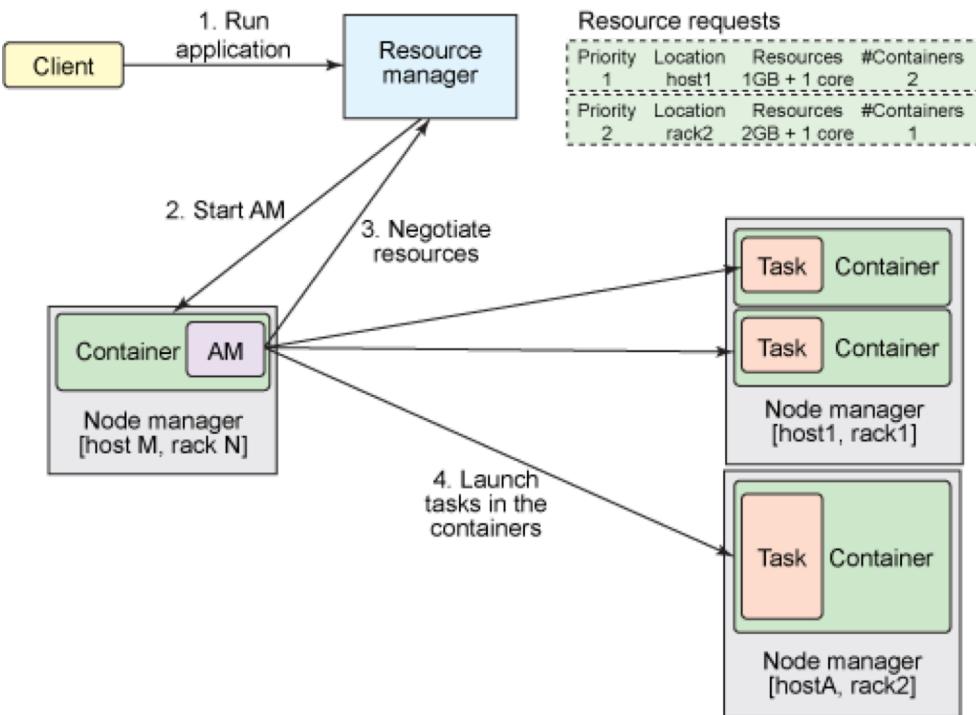


Tiene dos componentes principales:

- 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

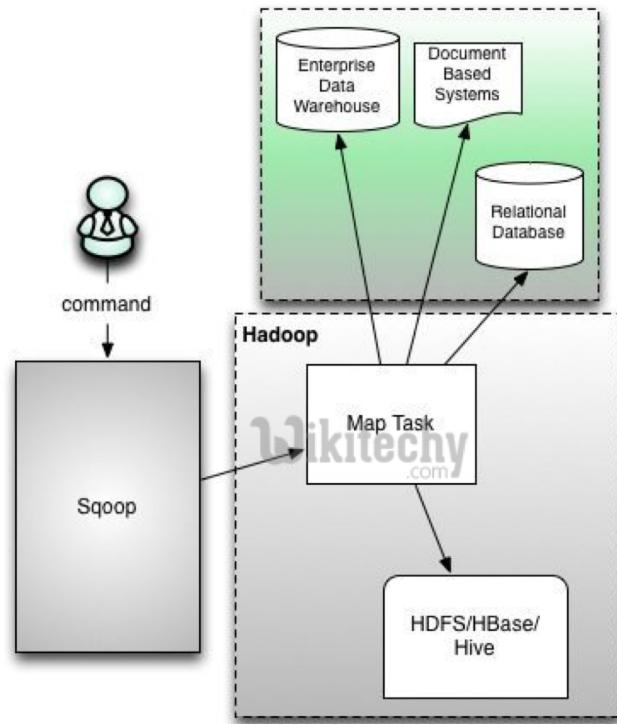


Tiene dos componentes principales:

- 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

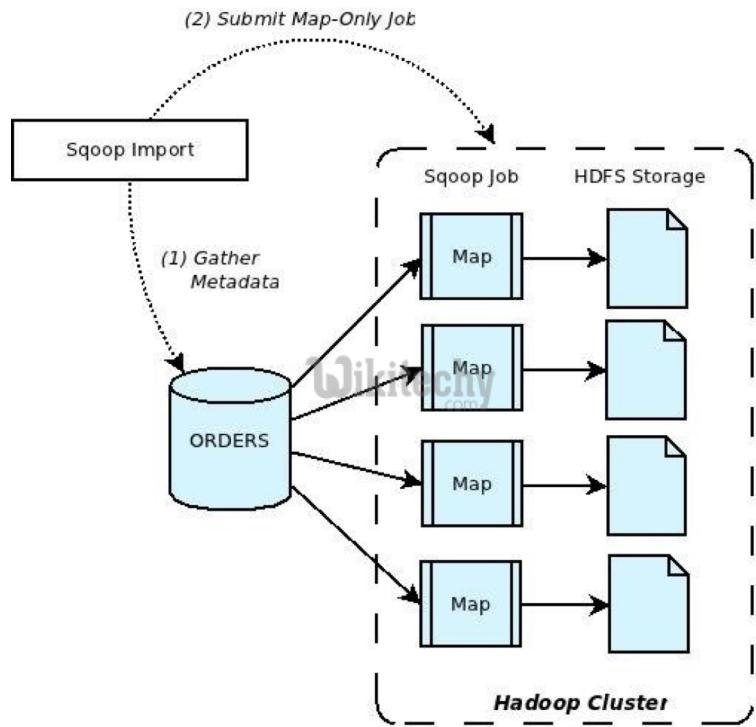


Sqoop solicita información sobre el schema en la BD y captura los metadatos.

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

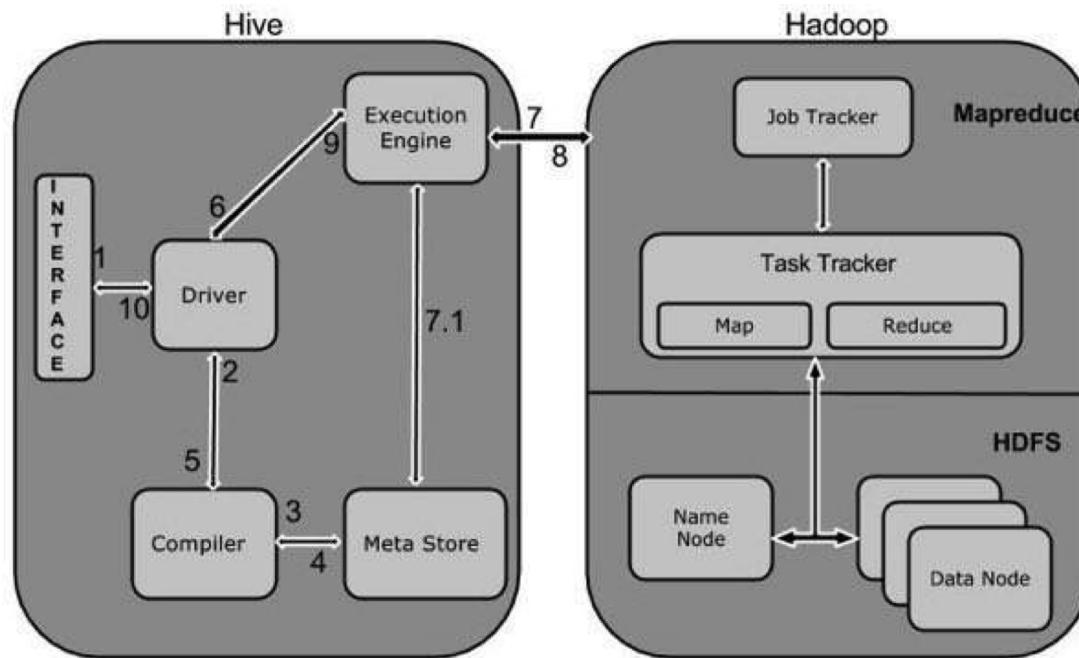


Sqoop analiza la bd y captura los metadatos

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

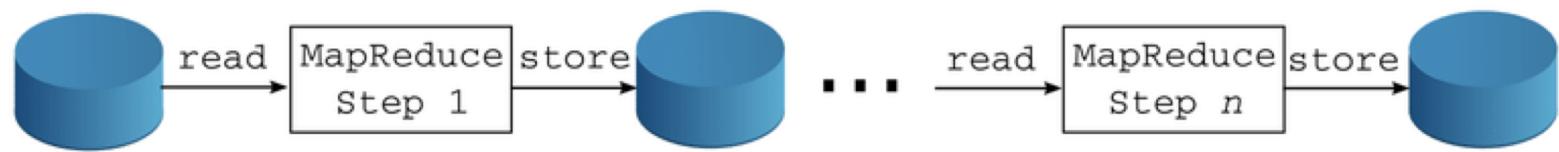


1. Ejecutar la solicitud
2. Obtener plan
3. Obtener metadatos
4. Enviar metadatos
5. Enviar plan
6. Ejecutar plan
7. Ejecutar trabajo
8. Metadata Ops
9. Resultado de búsqueda
10. Enviar resultados

MapReduce vs Spark

 hadoop
MapReduce

a)

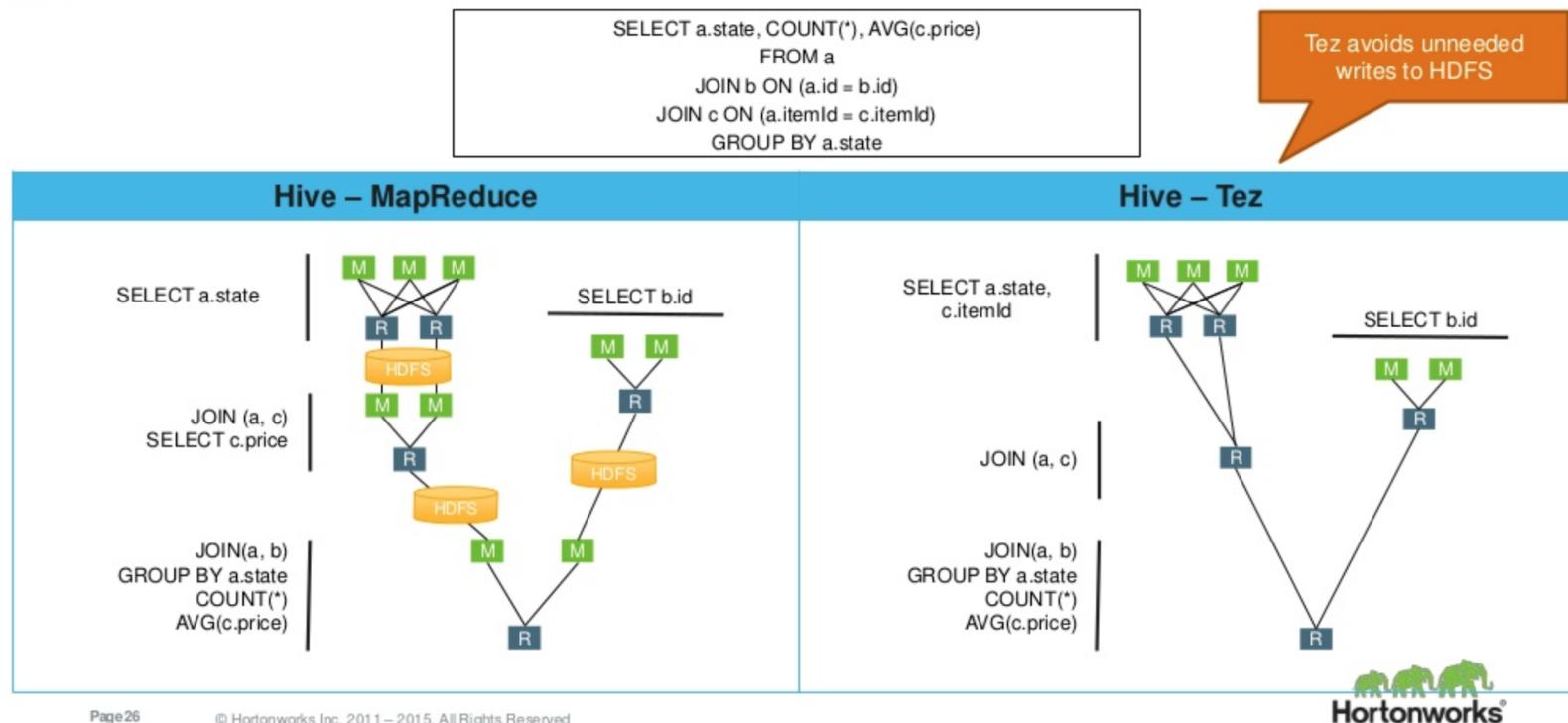


 Spark

b)



MapReduce vs Tez

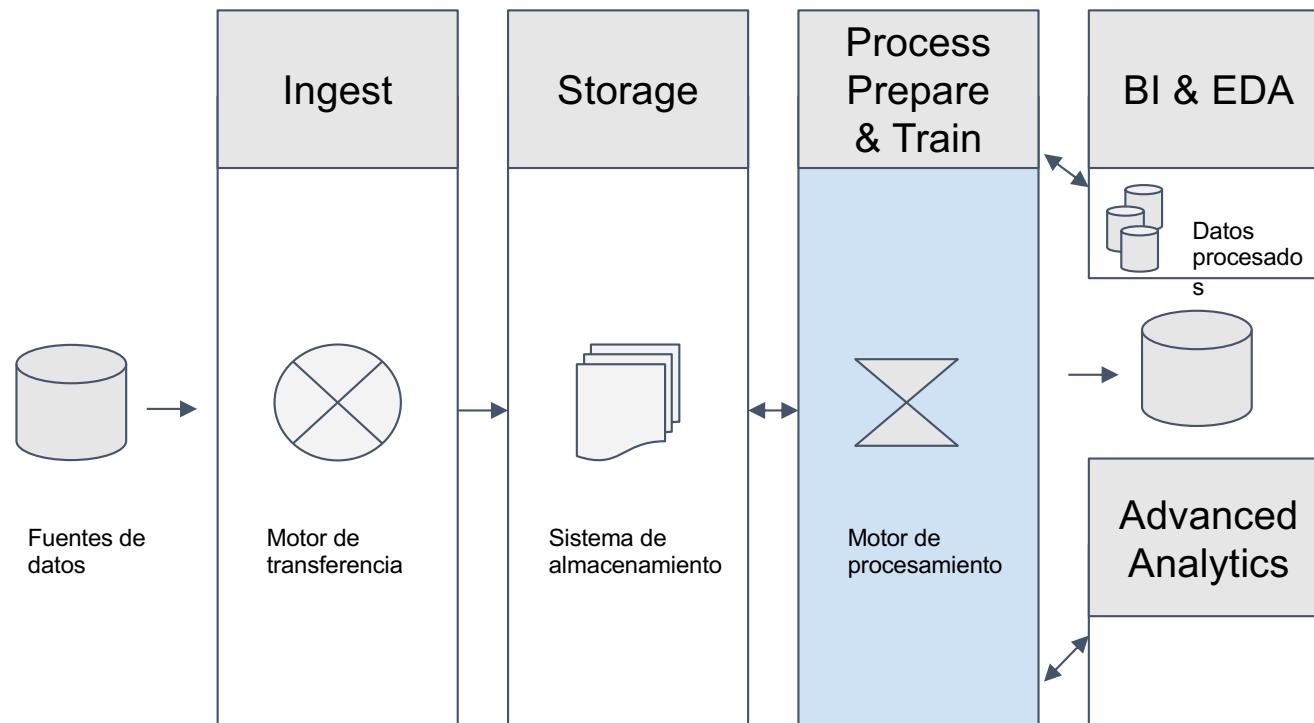


Ecosistemas Big Data y analítica - Nube

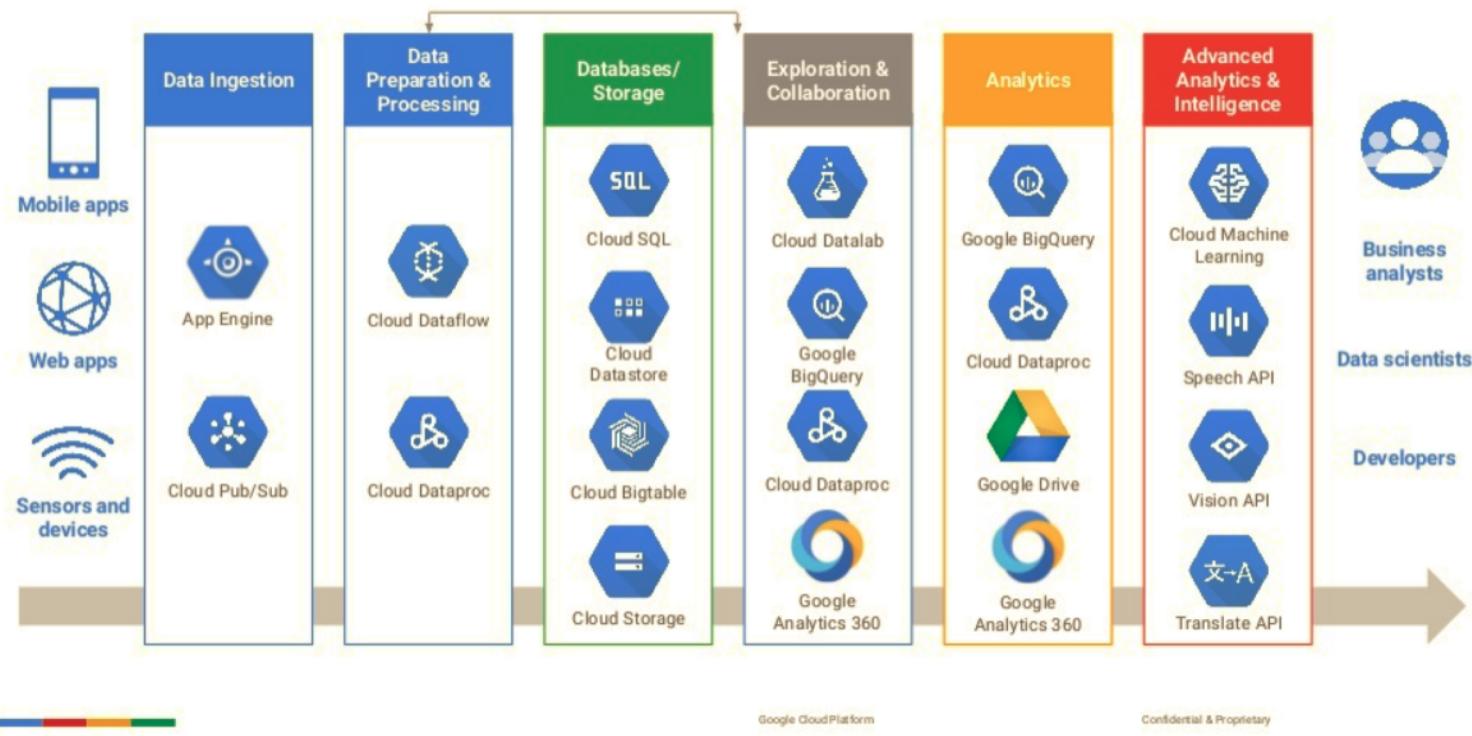
Inspira Crea Transforma



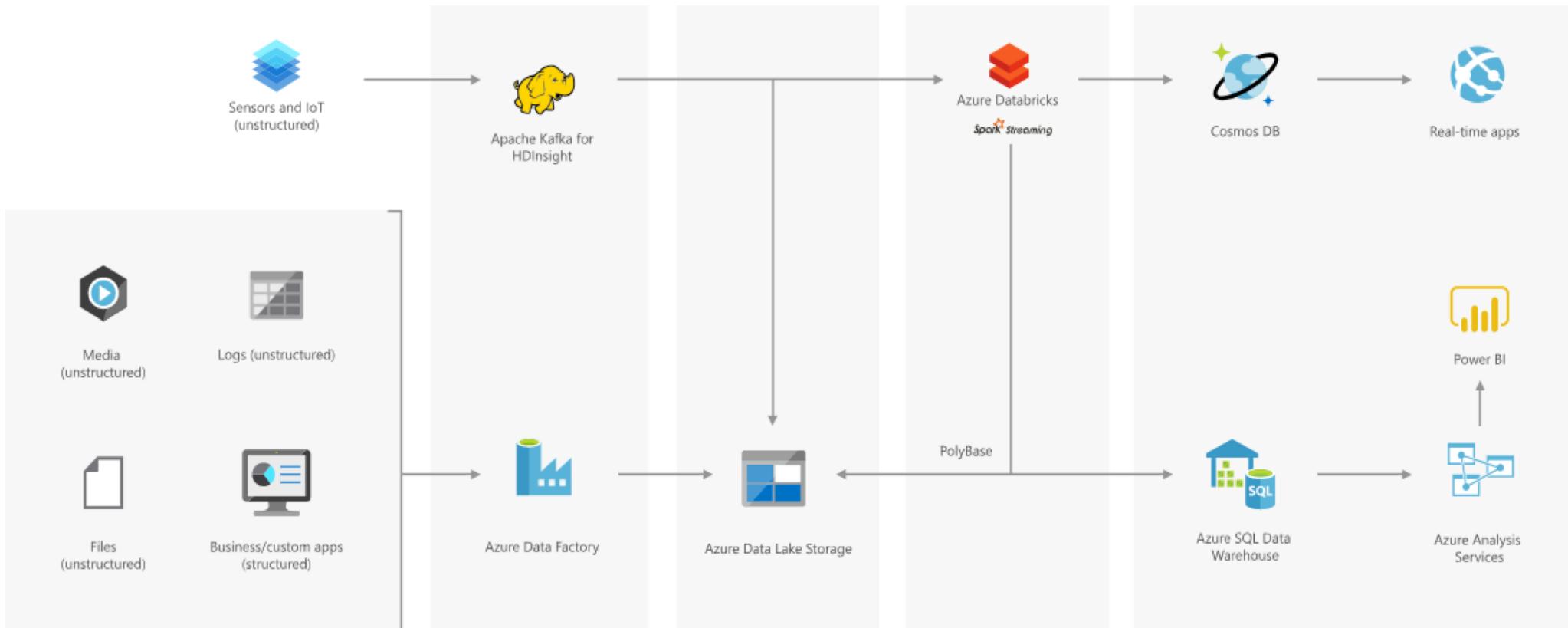
Proceso general de Big Data & Analytics



Google

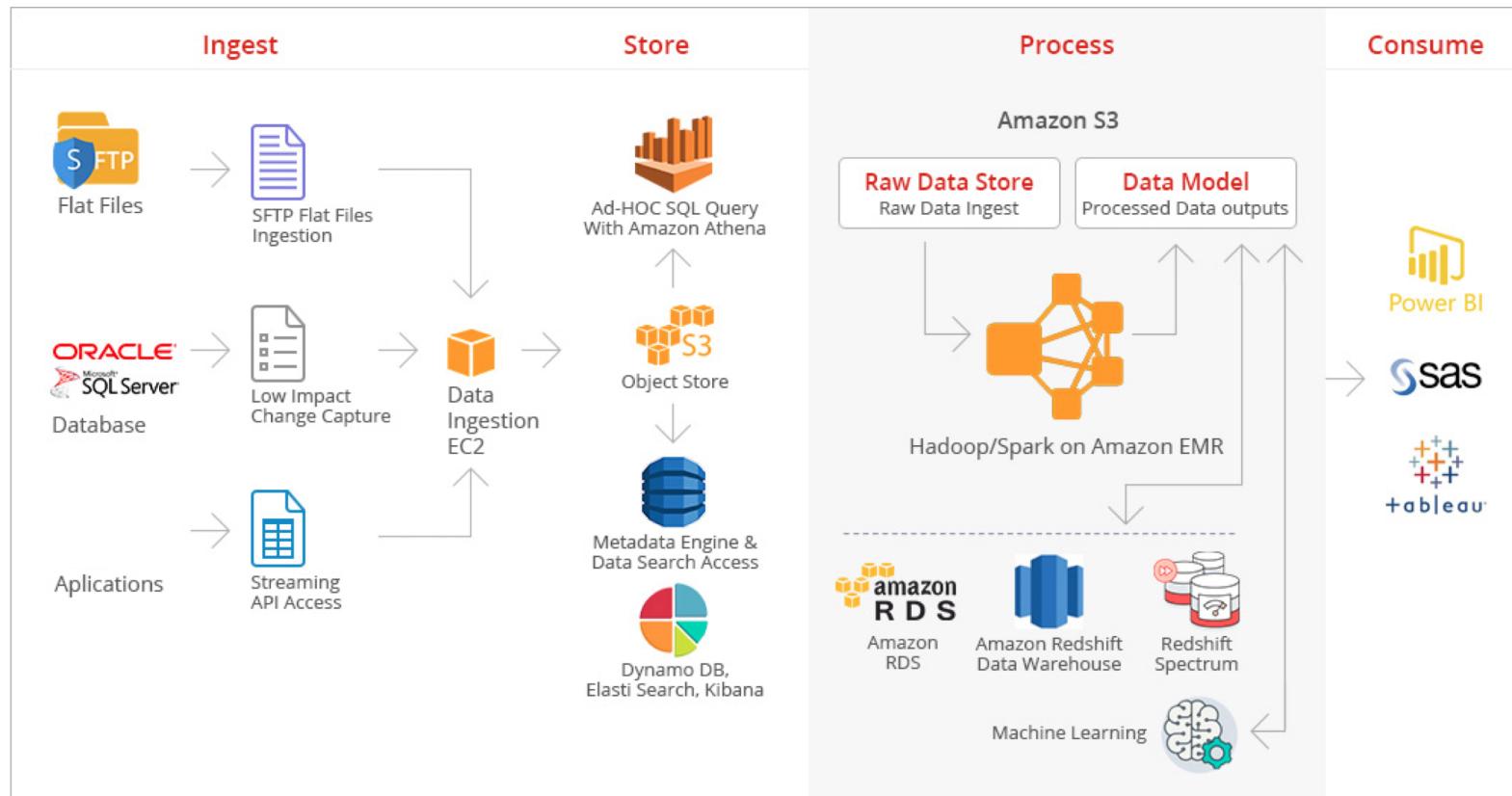


Azure



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



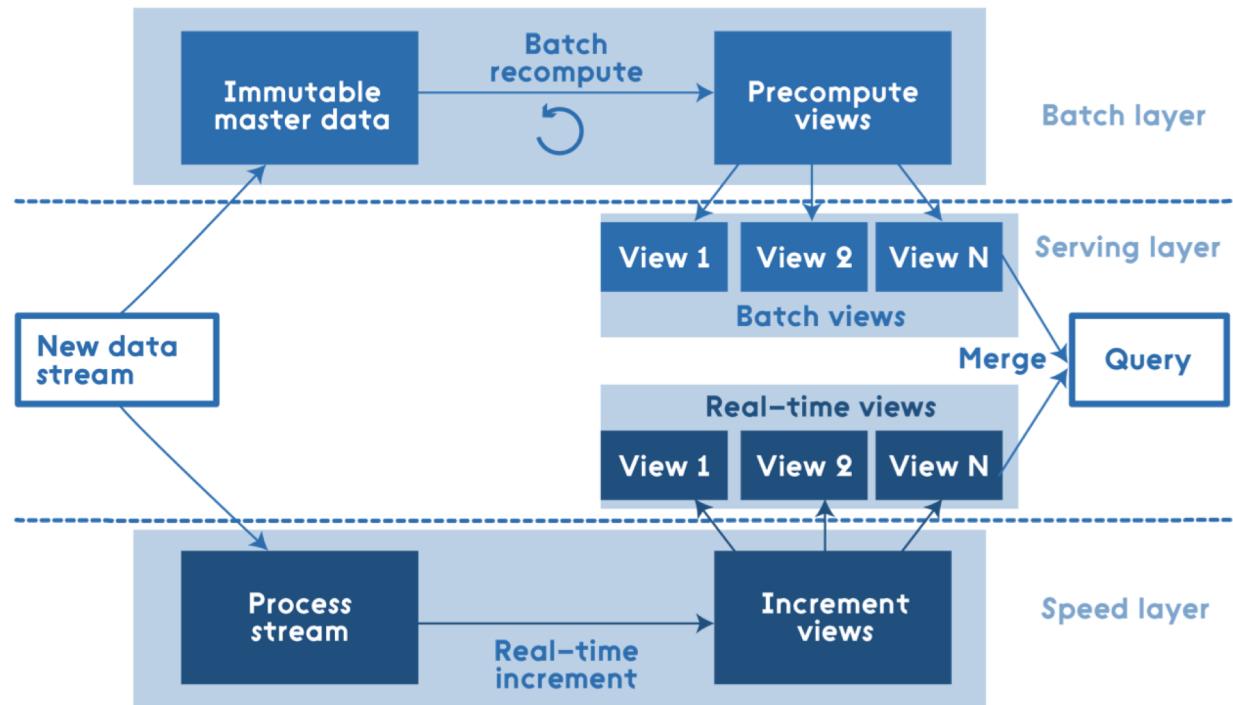
Inspira Crea Transforma

4. Arquitecturas Big Data

Inspira Crea Transforma



Arquitectura Lambda



Arquitectura Lambda

