Tecnologías para el Big Data II

Agenda

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

- Introducción
- Tecnologías batch procesamiento - Hive
- Tecnologías batch analítica - Hive
- Ejercicios Prácticos

https://www.youtube.com/watch?v=C8WKk6IJIso

Compartamos

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Coméntanos sobre el articulo que leíste

https://www.youtube.com/watch?v=U0r9s4iXwo0

Arquitectura Big Data

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Data Sources

Arquitectura Big Data

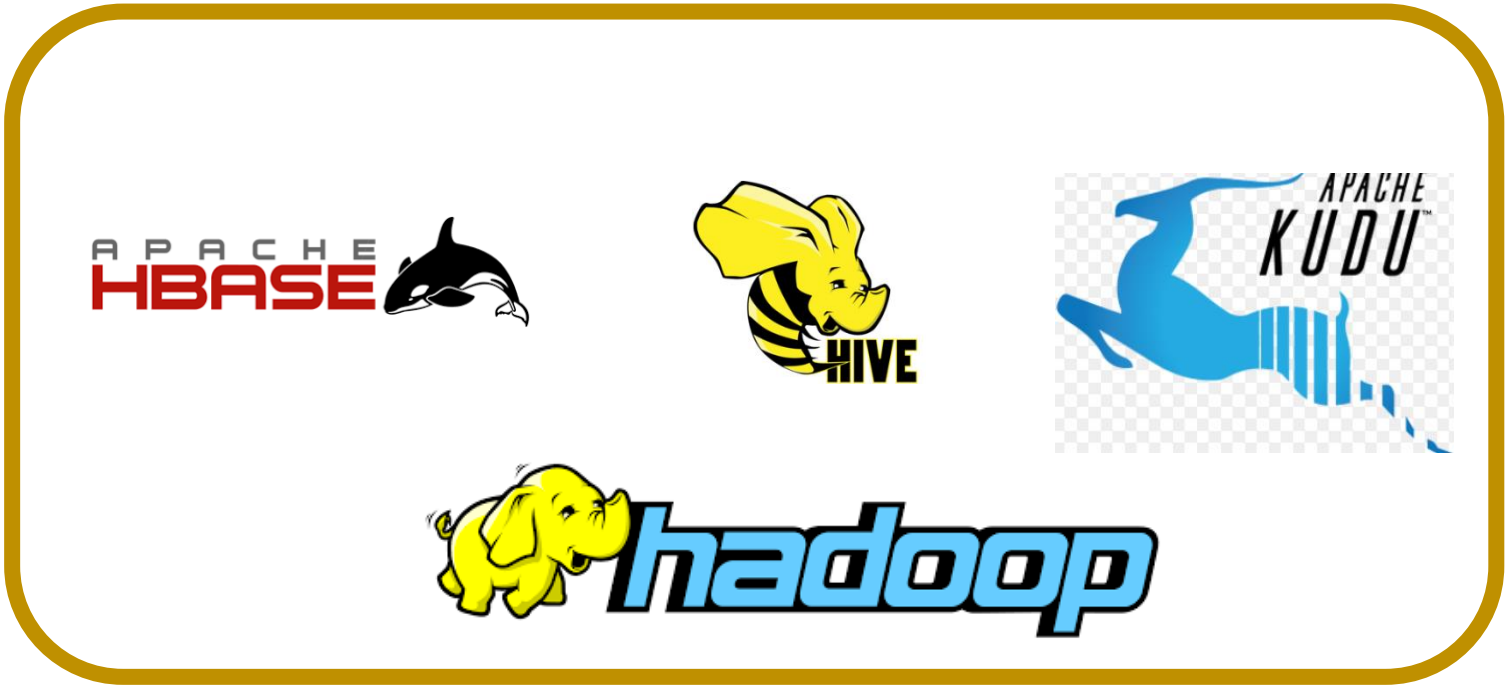
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Data Ingestion/Acquisition

Arquitectura Big Data

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Data Storage



Arquitectura Big Data

Data Analysis

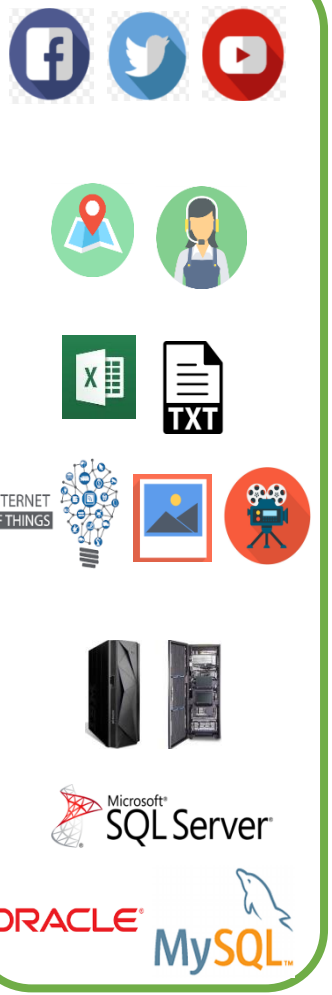
Arquitectura Big Data

Data Reporting & Visualization

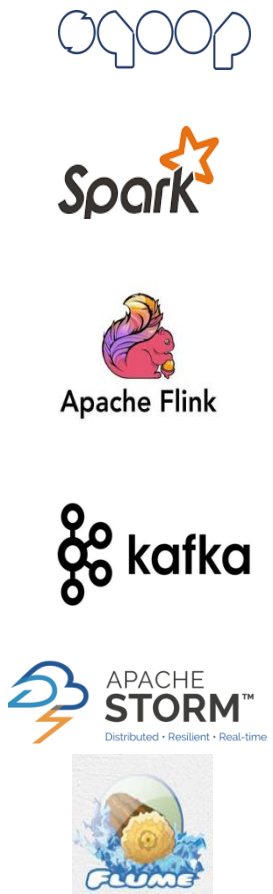
Big Picture

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Sources

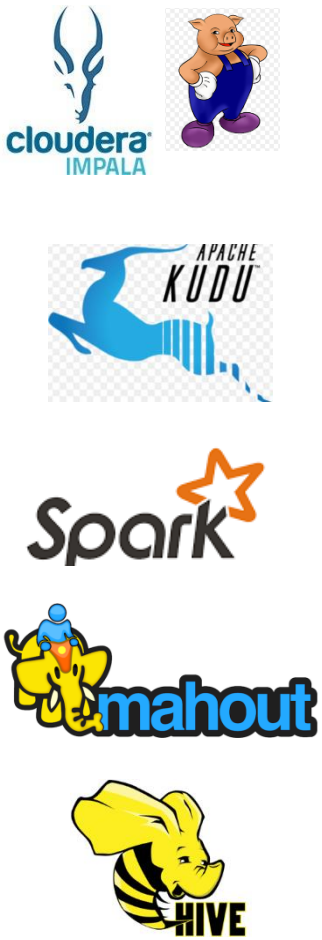


Ingestion



Storage

Analysis



Visualization

Compartamos

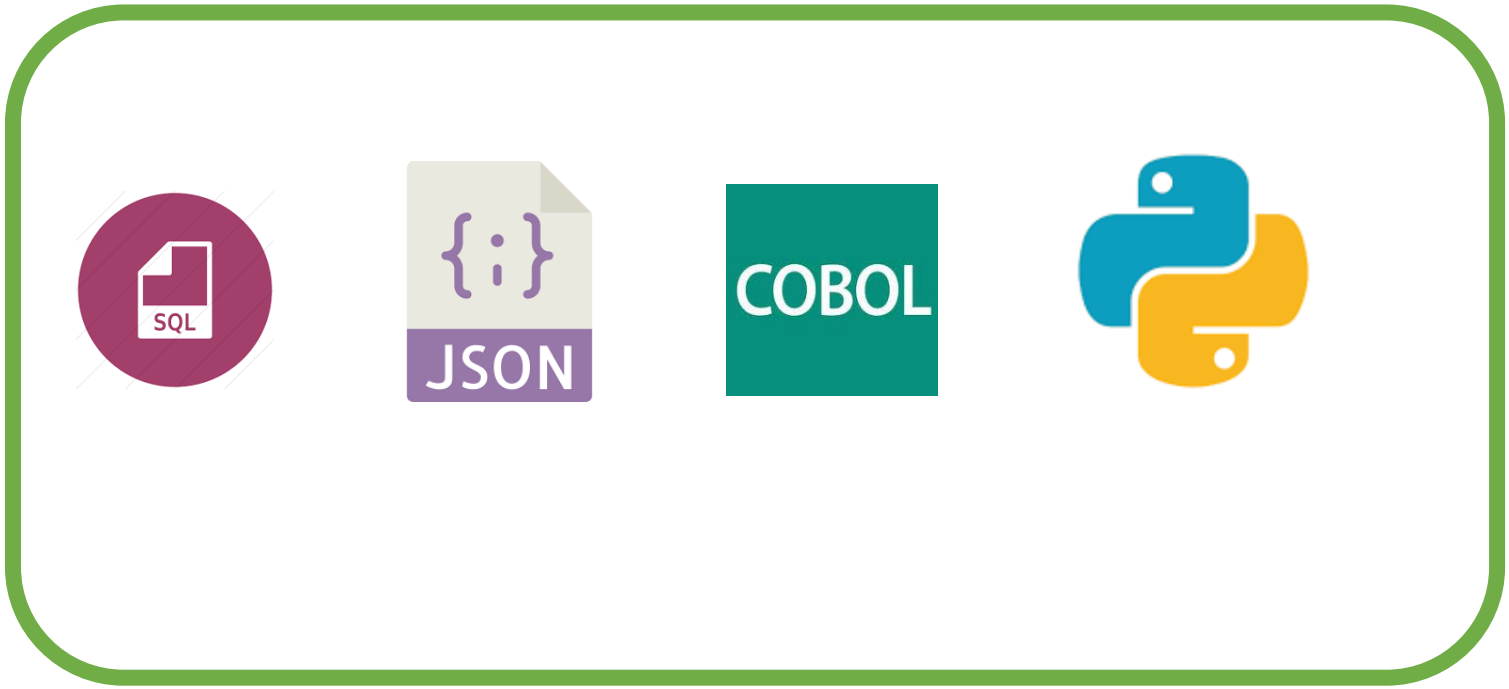
En equipos conversemos sobre lo visto en clase

1. ¿Qué tecnologías consideras más interesante de aprender a corto plazo?
2. ¿Lo podrías aplicar en tu día a día? ¿Cómo? O ¿Qué te faltaría?

Lenguajes en Big Data

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Data Sources



Lenguajes en Big Data

Data Ingestion/Acquisition

Lenguajes en Big Data

Data Storage

Lenguajes en Big Data

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Data Analysis

Lenguajes en Big Data

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

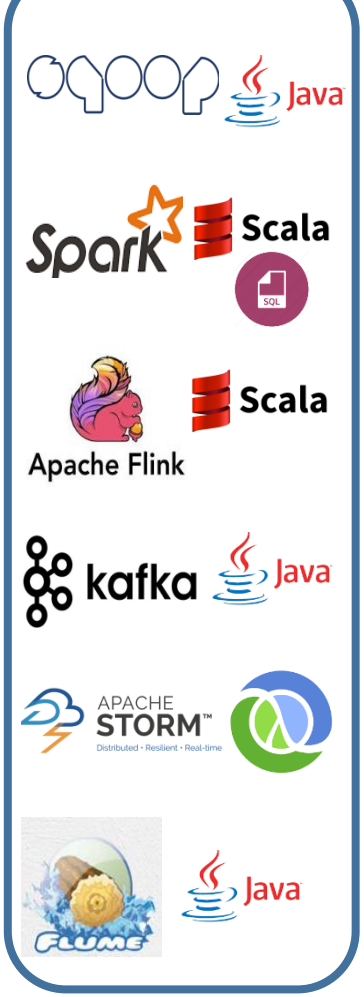
Data Visualization & Reporting

Bigger picture

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Sources

Ingestion



Storage

Analysis

Visualization

Perfiles en Big Data

Data Sources

Data Engineer

Data Scientist

Business and Data Analysts

Data Expert

Perfiles en Big Data

Data Ingestion/Acquisition

Data Engineer

Data Architect

DataOps

Perfiles en Big Data

Data Storage

Hadoop Admin

Data Engineer

Data Architect/ Big Data Specialist

DataOps

Data Governance

Data Quality

Data Modeler

Perfiles en Big Data

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Data Analysis

Data Scientist

Business and Data Analysts

Data Engineer

Data Quality

Perfiles en Big Data

Data Visualization

Data Scientist

Business and Data Analysts

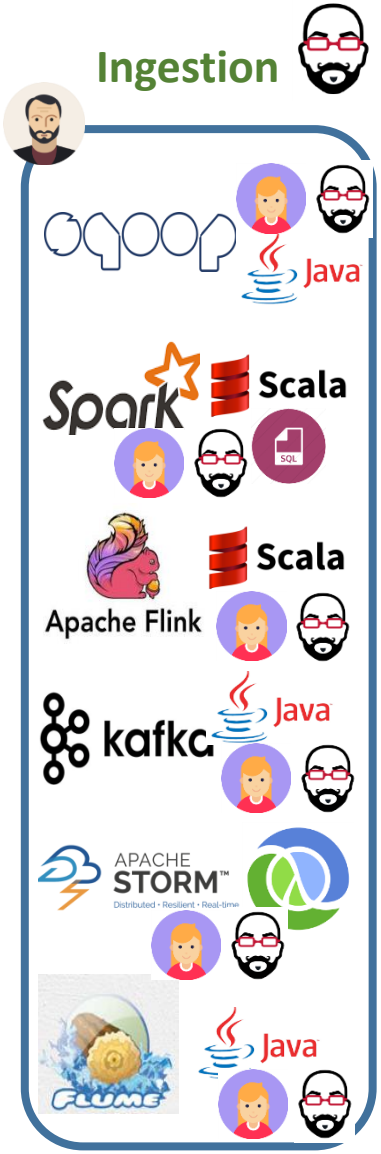
Data Engineer

Biggest picture

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

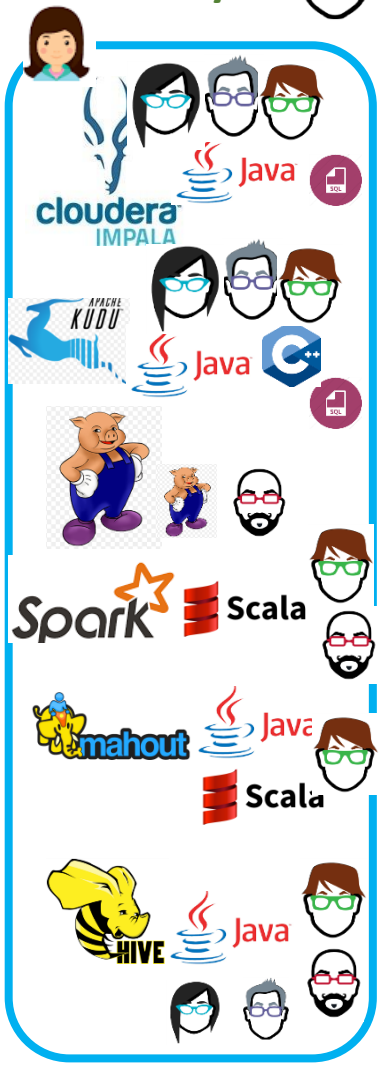
Sources

Ingestion

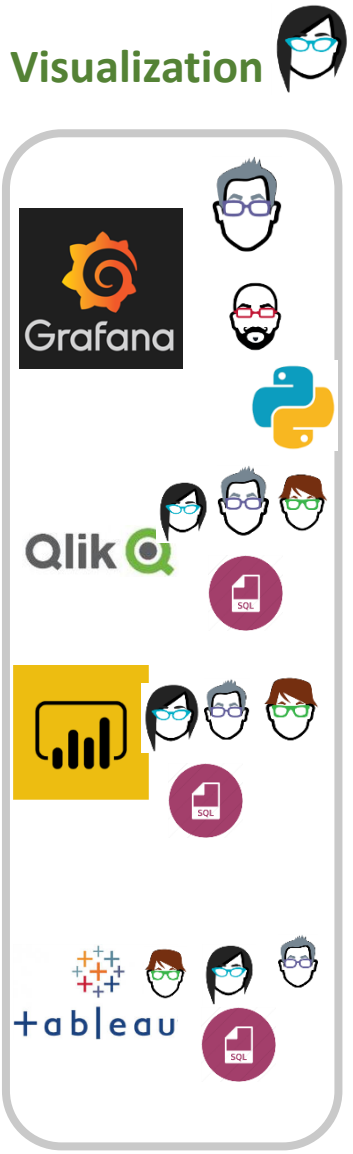


Storage

Analysis



Visualization



Biggest?

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Existen 41 proyectos Apache relacionados a Big Data

Apache Airavata
Project Apache Ambari
Project Apache Apex
Project Apache Avro
Project Apache Beam
Project Apache Bigtop
Project Apache BookKeeper
Project Apache Calcite
Project Apache CarbonData
Project Apache CouchDB
Project Apache Tajo
Project Apache Tez
Project Apache Trafodion
Project Apache VXQuery
Project Apache Zeppelin

Project Apache Crunch
Project Apache DataFu
Project Apache DirectMemory
Project Apache Drill
Project Apache Edgent (Incubating)
Project Apache Falcon
Project Apache Flink
Project Apache Flume
Project Apache Giraph
Project Apache Hama
Project Apache Helix
Project Apache Ignite
Project Apache Kafka
Project Apache Knox
Project Apache Lens

Project Apache MetaModel
Project Apache Oozie
Project Apache ORC
Project Apache Parquet
Project Apache Phoenix
Project Apache PredictionIO
Project Apache REEF
Project Apache Samza
Project Apache Spark
Project Apache Sqoop
Project Apache Storm

Compartamos

En equipos conversemos sobre lo visto en clase

1. Con lo visto hasta el momento puedes diseñar una solución a nivel de arquitectura

Con componentes tecnológicos, lenguaje y perfiles de Big Data a cualquiera de estos casos:

- Análisis para detectar Fraude informático de tarjeta de crédito
- Aplicación para medir el clima en ciertas zonas de la ciudad
- Análisis del sentimiento de los peruanos por temas de corrupción
- Reporte de ventas de una gran cadena de supermercado
- Reporte para el ente regulador en temas de caída de servicio y sus motivos

Deuda técnica

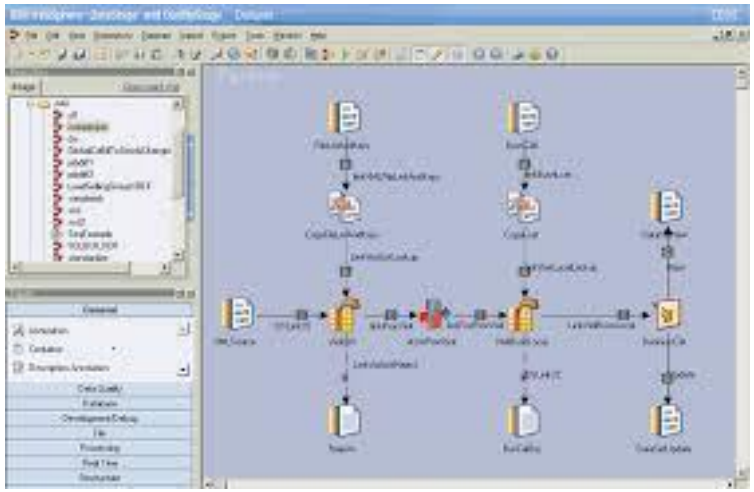
¿Todavía puede ser más grande la imagen?

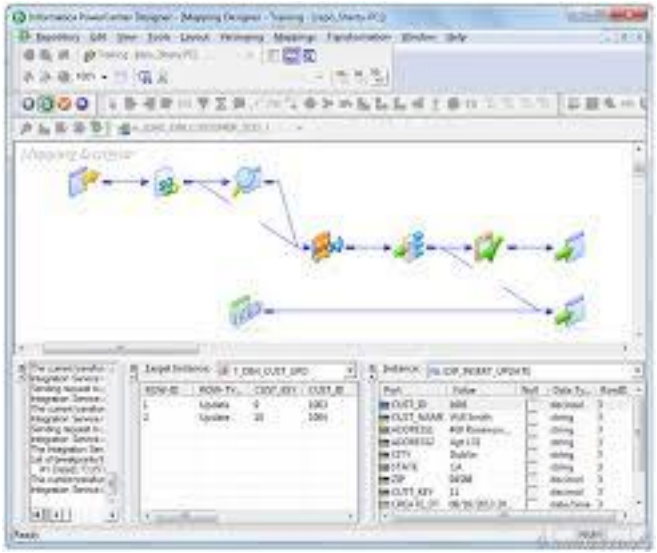
Sí, nos falta hablar de automatización, gobierno, calidad, proveedores,
versionamiento de código, seguridad, ciclo de un proyecto Big Data…

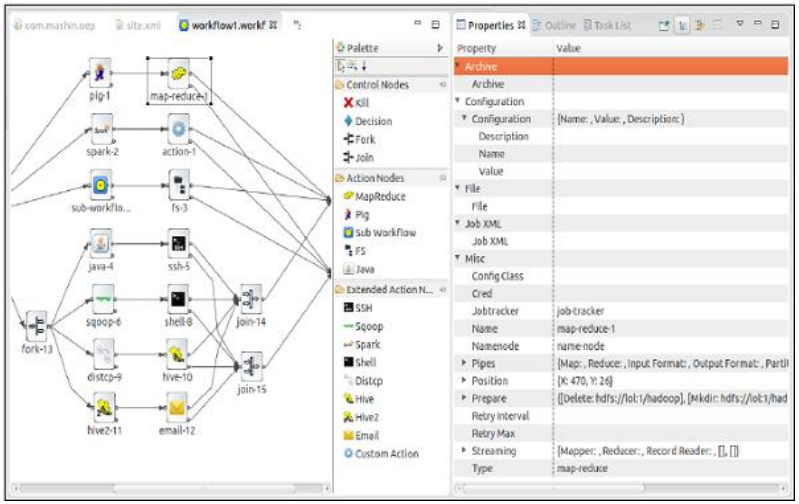
Deuda técnica

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Automatización



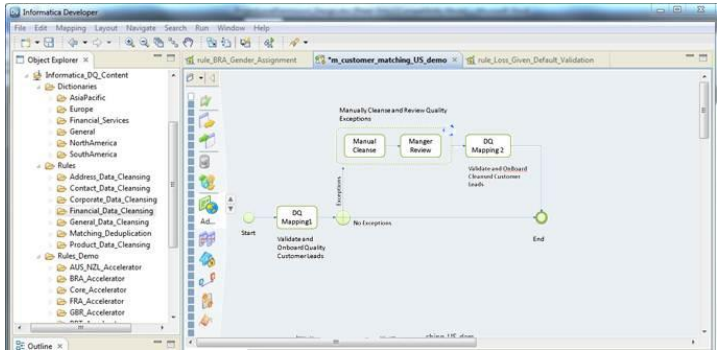


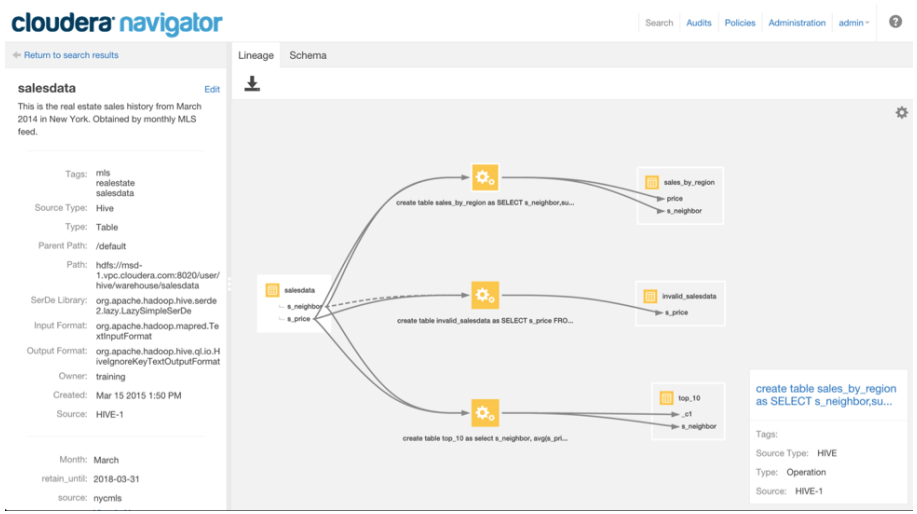


Deuda técnica

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Gobierno y calidad (linaje y trazabilidad)





Deuda técnica

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Versionamiento de código

Deuda técnica

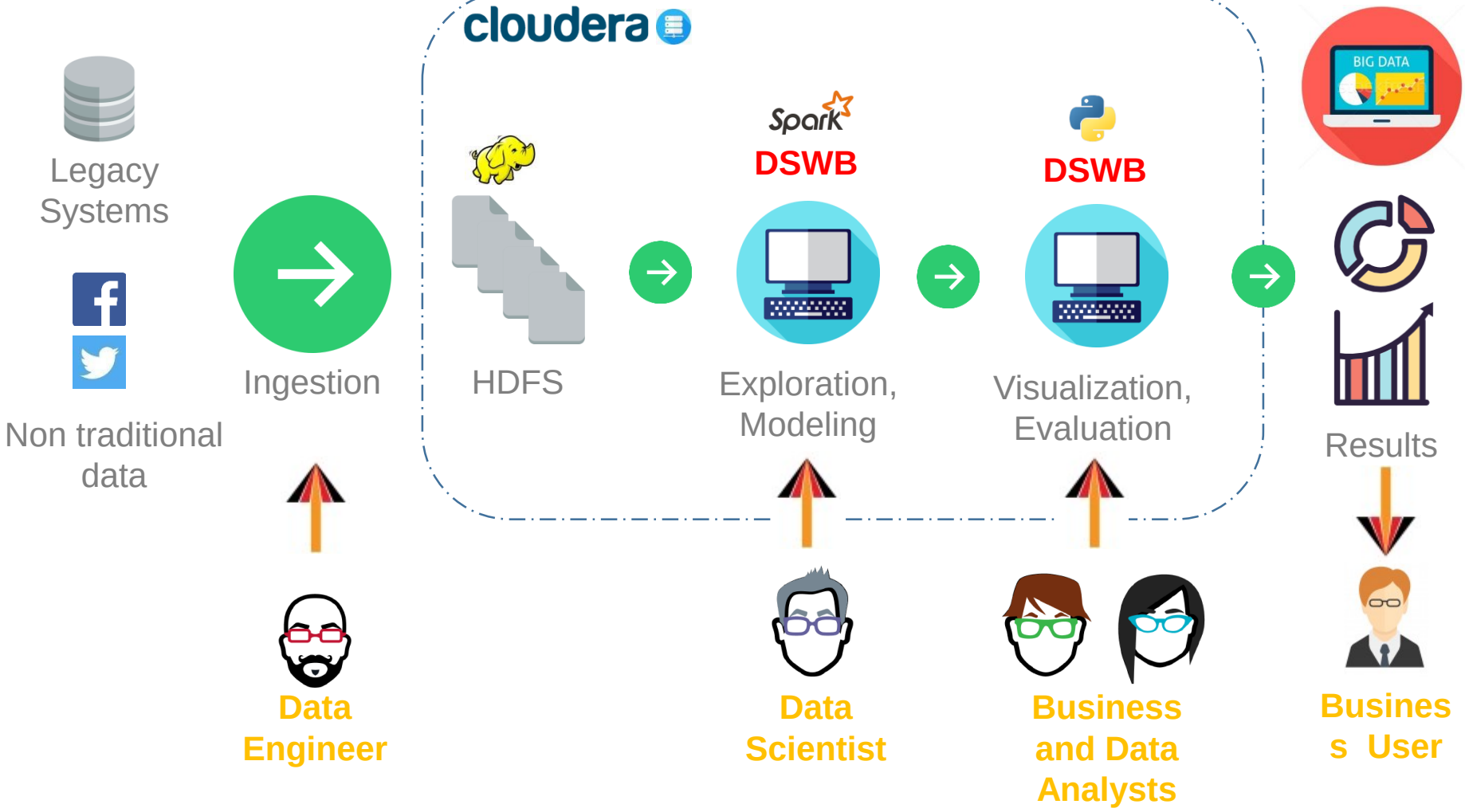
Seguridad

Seguridad

Kerberos y Sentry

Deuda técnica

Ciclo de un proyecto Big Data



Compartamos

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

En equipos conversemos sobre lo visto en clase

1. ¿Qué lenguaje y tecnología consideras relevante para ti?
2. Estos nuevos lenguajes y perfiles los has visto recientemente en alguna convocatoria o dentro de tu trabajo/escuela?
3. ¿Cómo ves ahora Big Data?¿Te interesa más? ¿Existe alguna tecnología no mencionada?

Práctica

https://qwiklabs.com/

Crearse un usario y buscar el laboratorio
Analyze Big Data with Hadoop

Práctica II

https://azure.microsoft.com/en-us/

Crearse una cuenta y luego ir a

https://portal.azure.com/

Historia de Hive

2009 Inició como un proyecto del equipo de infraestructura (Joydeep Sen Sarma and Ashish Thusoo)

2010 Publica el paper de Hive

2014 Pasa a ser un proyecto incubado en Apache

https://www.qubole.com/blog/founders-transformation-hadoop/
http://spark.apache.org/committers.html

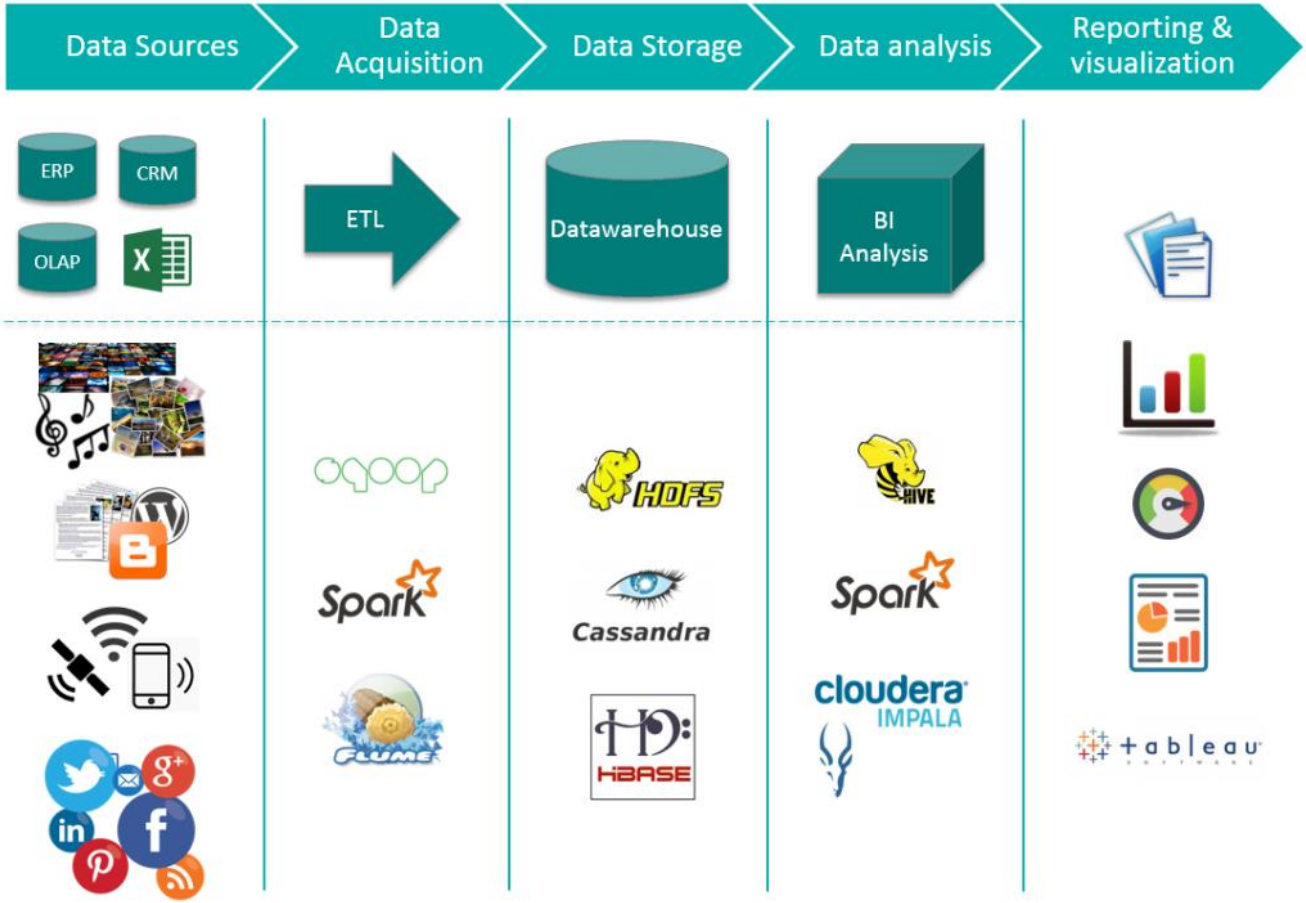
¿Qué es Hive?

Data warehouse software que facilita leer, escribir y manejar largos conjunto de datos, que residen en un almacenamiento distribuido, usando SQL.

Map Reduce

¿Qué es Hive?

Es una capa de abstracción en la parte superior de Hadoop.



http://www.clearpeaks.com/blog/big-data/big-data-ecosystem-spark-and-tableau

Compartamos

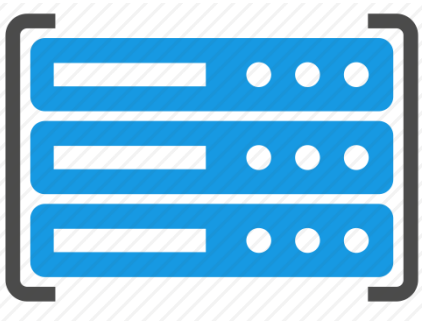
En equipos respondan y debatan las preguntas:

1. ¿Qué tecnologías usan para cargar la extraer y cargar data?
2. En el estado actual, ¿Necesitan el procesamiento y/o almacenamiento en tiempo real? ¿Por qué?

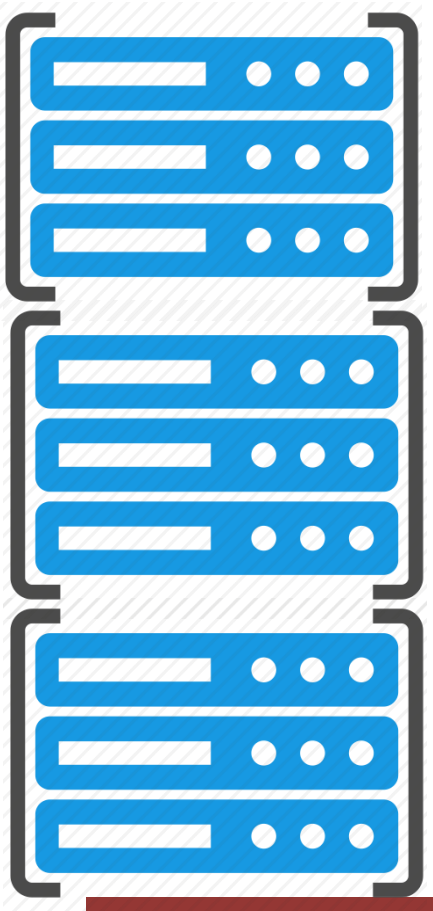
¿Qué hace Hive?

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Hive permite manejar hasta varios petabytes de datos a la vez distribuido a través de un cluster de miles de servidores virtuales o físicos a través de HiveQL.



Driver



Worker

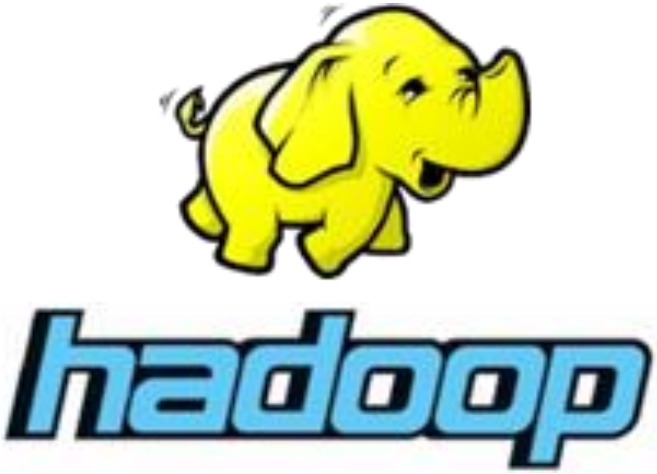
Worker

Worker

¿Qué hace Hive?

El interprete de Hive utiliza Map Reduce o Spark para procesar los datos.

Existe conectores ODBC y JDBC lo que lo hace fácil de integrarse con herramientas de BI.

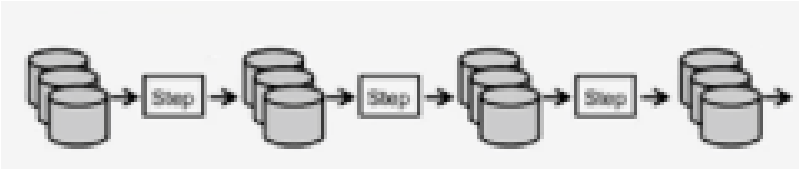


¿Por qué es tan usado?

Los datos pueden ser cargados en HDFS antes de definir una tabla.

Hive al tener un lenguaje parecido al SQL no necesita de lenguajes como Python, Java o Scala.

Schema on read



¿Por qué es tan usado?

Es bueno para datos estructurados, como para datos semi estructurados.

Schema on read

Características	BD	Hive
Lenguaje	SQL	SQL
Update, Delete	Y	N
Procedimiento	Y	N
Índices	Y	Limitado
Formato de archivos	N	Y(Avro, ORC, Parquet)

Mitos en Hive

¿Cómo está relacionado Hive con Apache Hadoop?

Hive es una capa en el top de Hadoop y gestionado por YARN.

¿Hive tiene el mismo comportamiento que BD regular?

No, tiene características similares acercándose más o no a una BD regular.

¿Existen realmente tablas Hive?

No, las tablas son realmente archivos en HDFS

Mitos en Hive

Why is Impala faster than Hive?

Answer | Request | Follow 14 | Comment | Downvote

4 Answers

Chris Schrader, Business Intelligence Consultant
Answered Jun 30

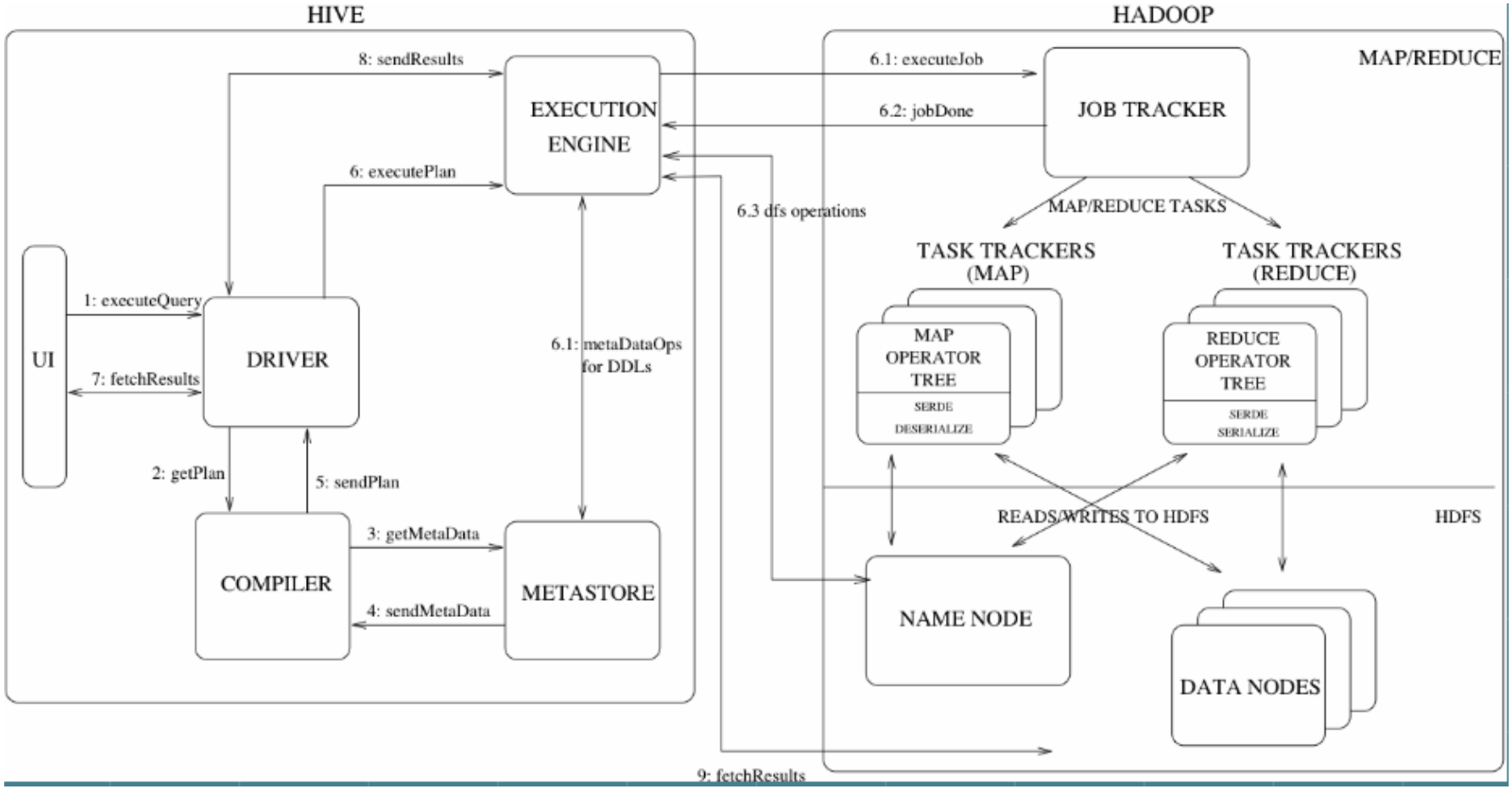
A2A.

Well, it's not always. And Hive itself is a hodgepodge of multiple processing engines and storage types. Hive can be run with MapReduce, Tez, or Spark as its engine. It also supports many data formats including plain text, avro, Parquet, Orc, RCFile, and probably a lot more I'm not thinking about. Hive most recently also added Live Long and Process (LLAP) to it's architecture which holds a lot of pre-computed vectors in memory.

As Shahzad Aslam mentions in his answer, Impala is an MPP style processing architecture and doesn't have many of the startup overheads of Hive (since Hive effectively submits a batch job to it's underlying processing engine vs running in "Always On").

Arquitectura Hive

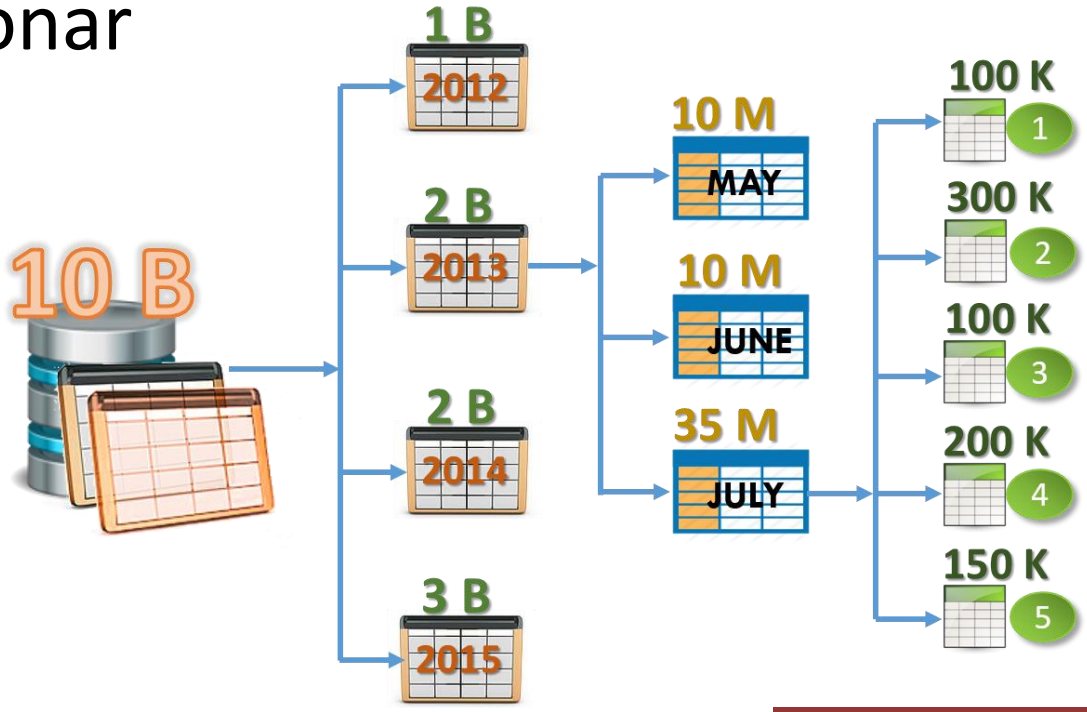
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería



Particiones

Se utiliza particiones cuando la lectura del dataset es muy larga.

Las consultas son en su mayoría por el campo a particionar



Formato Archivos - AVRO

Requiere definir un esquema previamente .avsc

Fácil cambio de esquema o campos.

```
CREATE TABLE order_details_avro ()
STORED AS AVRO
TBLPROPERTIES ('avro.schema.literal'=
'{"name": "order",
"type": "record",
"fields": [
{"name":"order_id", "type":"int"},
{"name":"cust_id", "type":"int"},
{"name":"order_date", "type":"string"}
]}');
```

Formato Archivos - Parquet

Formato columnar de archivos open source (Soportado por Cloudera).

Incrementa el performance.

```
CREATE TABLE order_details_parquet (
order_id INT,
prod_id INT)
STORED AS PARQUET;
```

Estructuras Complejas

Array

Cada elemento del arreglo debe ser del mismo tipo de dato.

Se puede especificar el tipo delimitado

```
Create table clientes(
nombre String,
telefono array<String>)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY  ','
COLLECTION ITEMS TERMINATED BY  '|' ;
```

```
Select nombre,
telefono[0],
telefono[1] from
clientes;
```

Estructuras Complejas

MAP

Son los tipo clave valor. Se puede especificar el tipo delimitador. Deben tener el mismo tipo las claves y valores.

```
Create table clientes(
nombre String,
telefono map<String, String>)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY  ','
MAP KEYS ITEMS TERMINATED BY  ':' ;
```

```
Select nombre,
telefono['casa'],
telefono['trabajo']
from clientes;
```

Estructuras Complejas

Structs

Cada elemento tiene un tipo de dato propio. Se define un tipo de dato propio.

```
Create table clientes(
nombre String,
direccion struct<calle: String,
ciudad: String>)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY  ','
MAP KEYS ITEMS TERMINATED BY  ':' ;
```

```
Select
nombre,  direccio
n.calle from
clientes;
```

Ejercicio

En equipo contesta las siguientes preguntas:

1. ¿En tu empresa/organización manejan datos complejos?
2. Piensas que es necesario para alguna familia de datos que manejas
3. Comparte 6 ejemplos reales (Map, Struct, Array)

Cómo se trabaja con Hive

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Data Engineer

ETL → Producción

Data Scientist/Business User/Data Analyst

Query → Exploración, cálculos

Preguntas

Práctica

CTIC UNI
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Azure

Demo Hue

Bibliografía

1. Paper inicial Hive

http://www.vldb.org/pvldb/2/vldb09-938.pdf

2. MySQL vs Hive

https://2xbbhjxc6wk3v21p62t8n4d4-wpengine.netdna-ssl.com/wp-content/uploads/2016/05/Hortonworks.CheatSheet.SQLtoHive.pdf

Tecnologías para el Big Data II