



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Tecnologías para el Big Data Apache Hive

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

Historia de Hive



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

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?



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

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

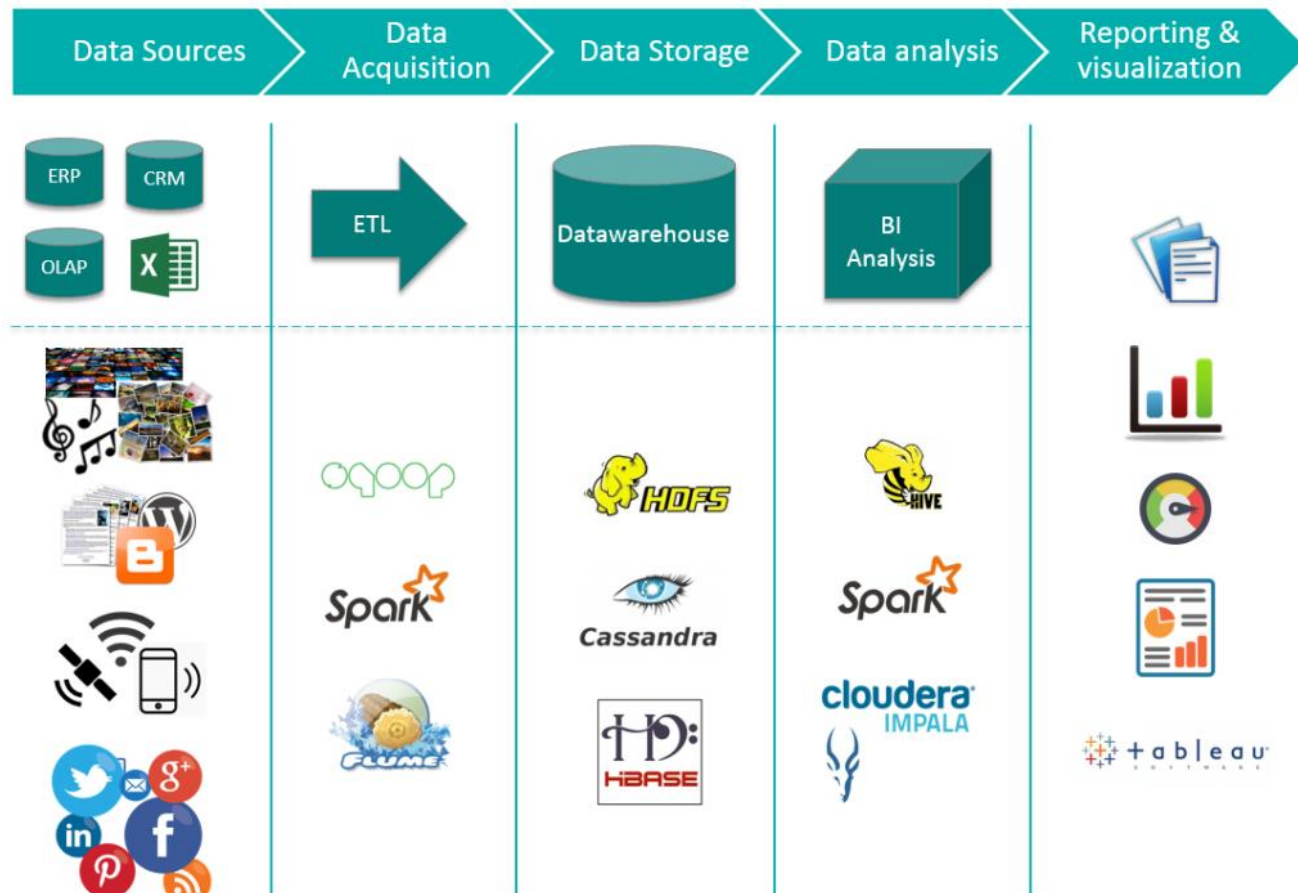


Map Reduce

¿Qué es Hive?



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



Compartamos



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

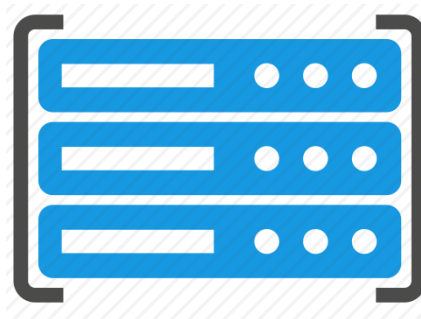
En equipos respondan y debatan las preguntas:

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?

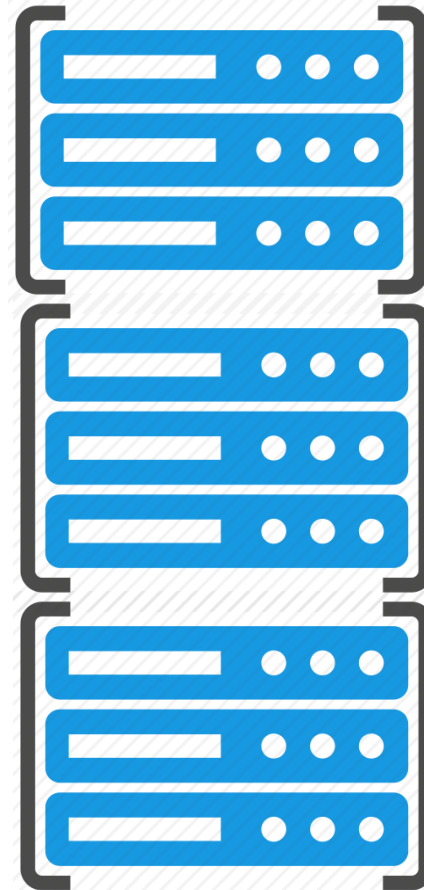
¿Qué hace Hive?



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?



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

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.



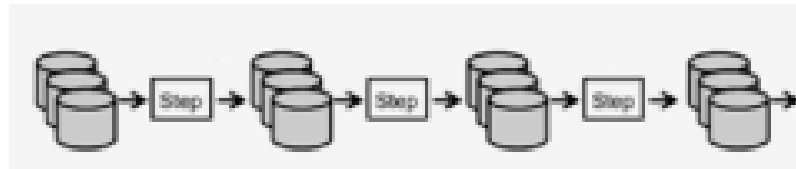
¿Por qué es tan usado?



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



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Why is Impala faster than Hive?

Answer

Request ▾

Follow

14

Comment

Downvote



4 Answers



Chris Schrader, Business Intelligence Consultant

Answered Jun 30

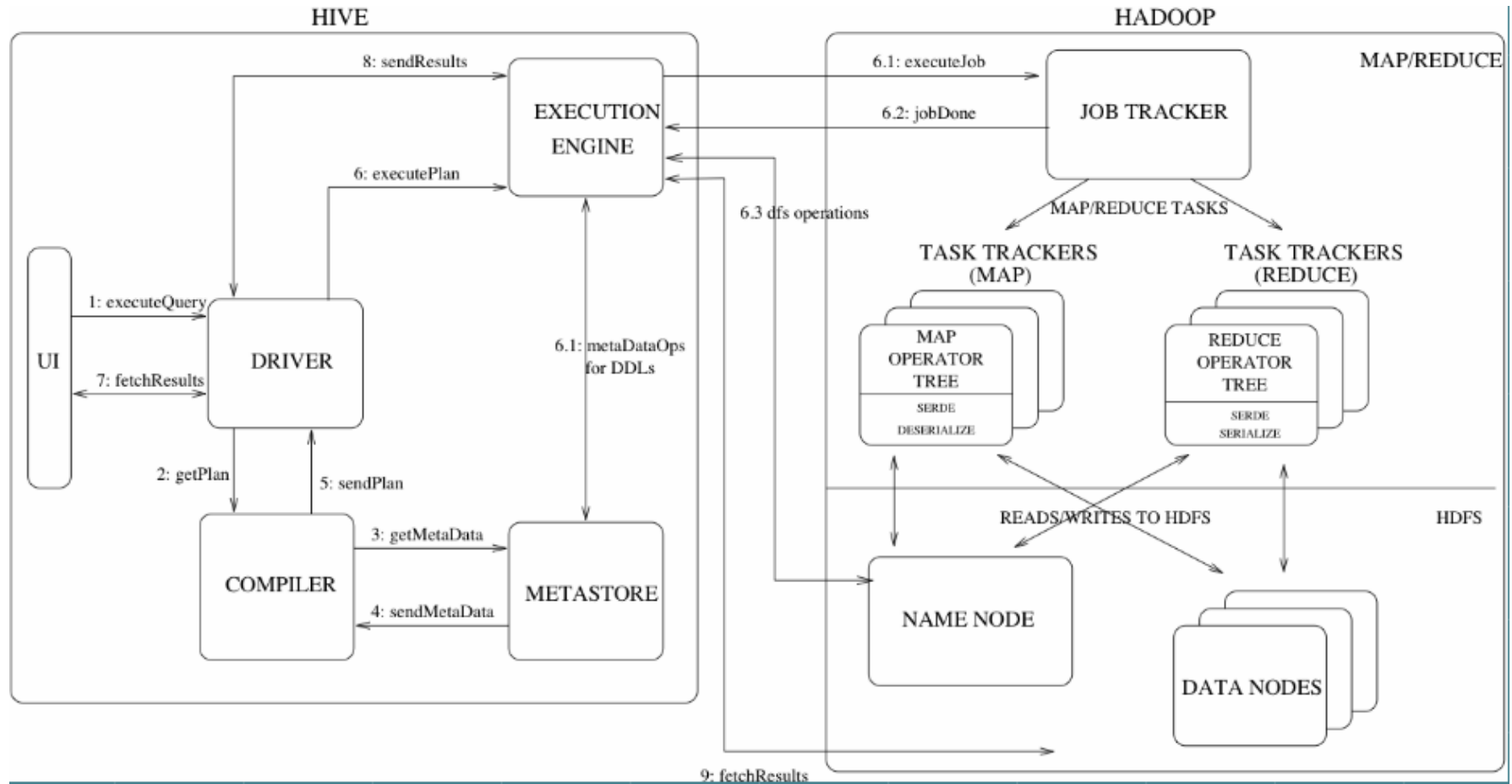


A2A.

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

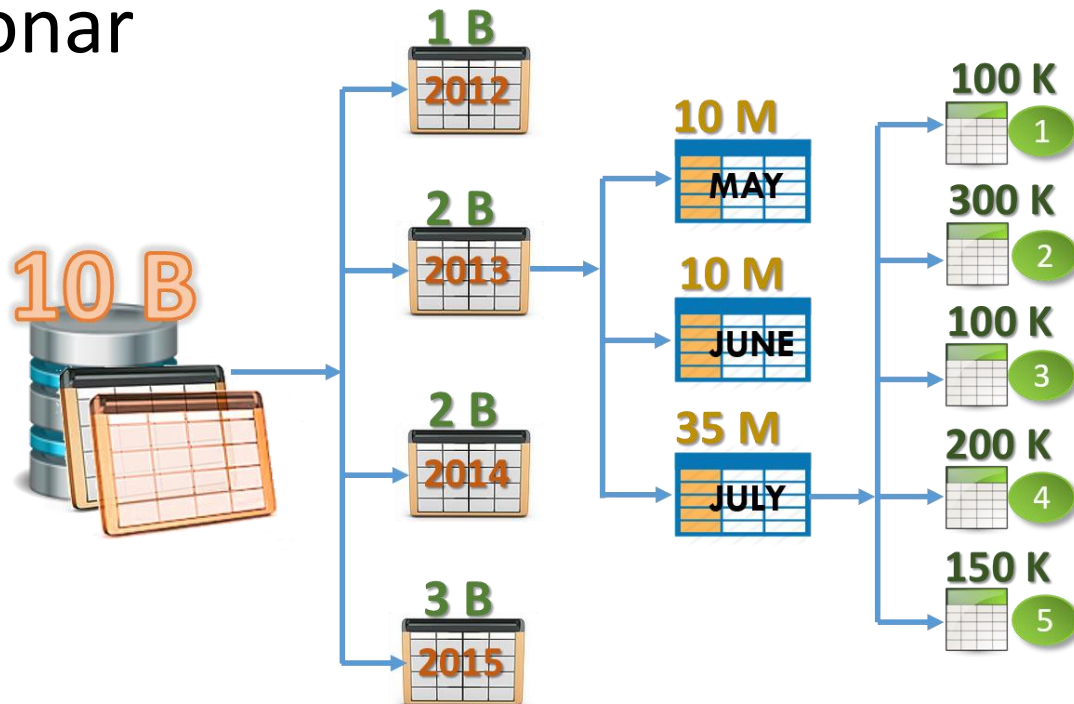


Particiones



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>)
```

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

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

MAP KEYS ITEMS TERMINATED BY ':' ;

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, direccion.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



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Preguntas

Práctica



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Azure Demo Hue

Bibliografía



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

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>



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Tecnologías para el Big Data Apache Hive