



Universidad **Ricardo Palma**

RECTORADO

PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

BIG DATA APLICADO

SESIÓN 7

Expositores:

David Narváez

Eder Pineda

bigdataplicado@gmail.com

AGENDA

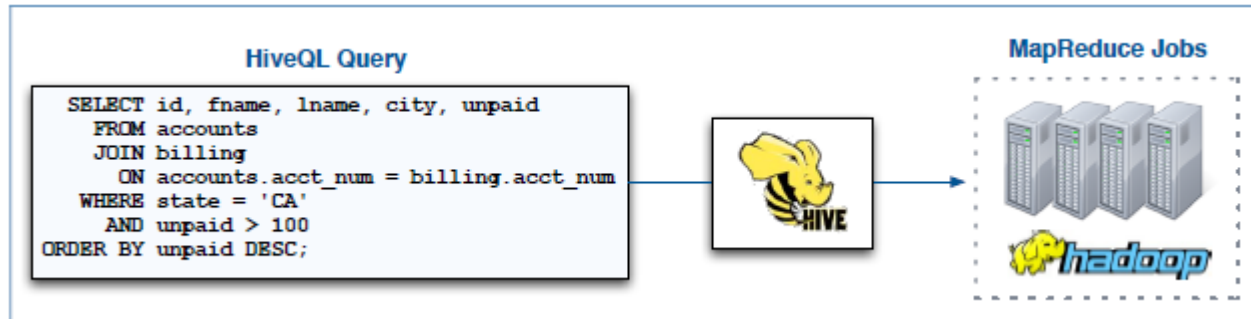
- INTRODUCCIÓN:
 - HIVE
 - QUÉ ES HIVE?
 - CÓMO FUNCIONA?
 - ARQUITECTURA
 - IMPALA
 - QUE ES IMPALA?
 - ARQUITECTURA
 - DATA ANALYSIS CON HIVE/IMPALA/PYSPARK



HIVE – QUÉ ES HIVE?

Hive es una herramienta open source desarrollada inicialmente en Facebook, que utiliza **sintaxis tipo SQL** para consultar (Query) data del HDFS.

Resuelve las consultas ejecutando jobs MapReduce dentro del cluster*

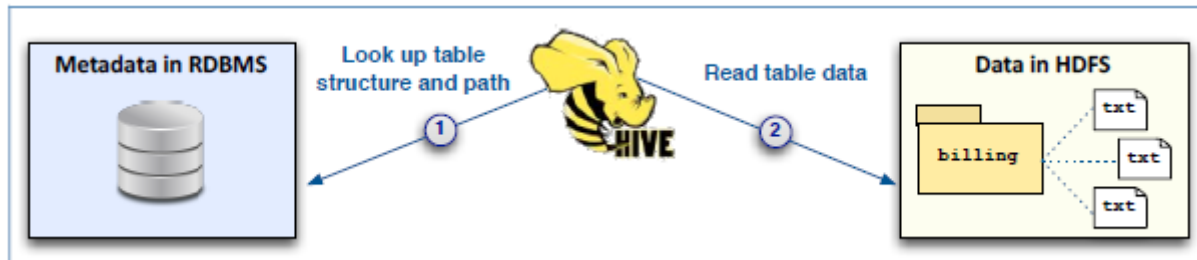


* Aunque se puede configurar Spark como motor de ejecución de procesos. En EMR el motor preferido es TEZ.



HIVE – CÓMO FUNCIONA?

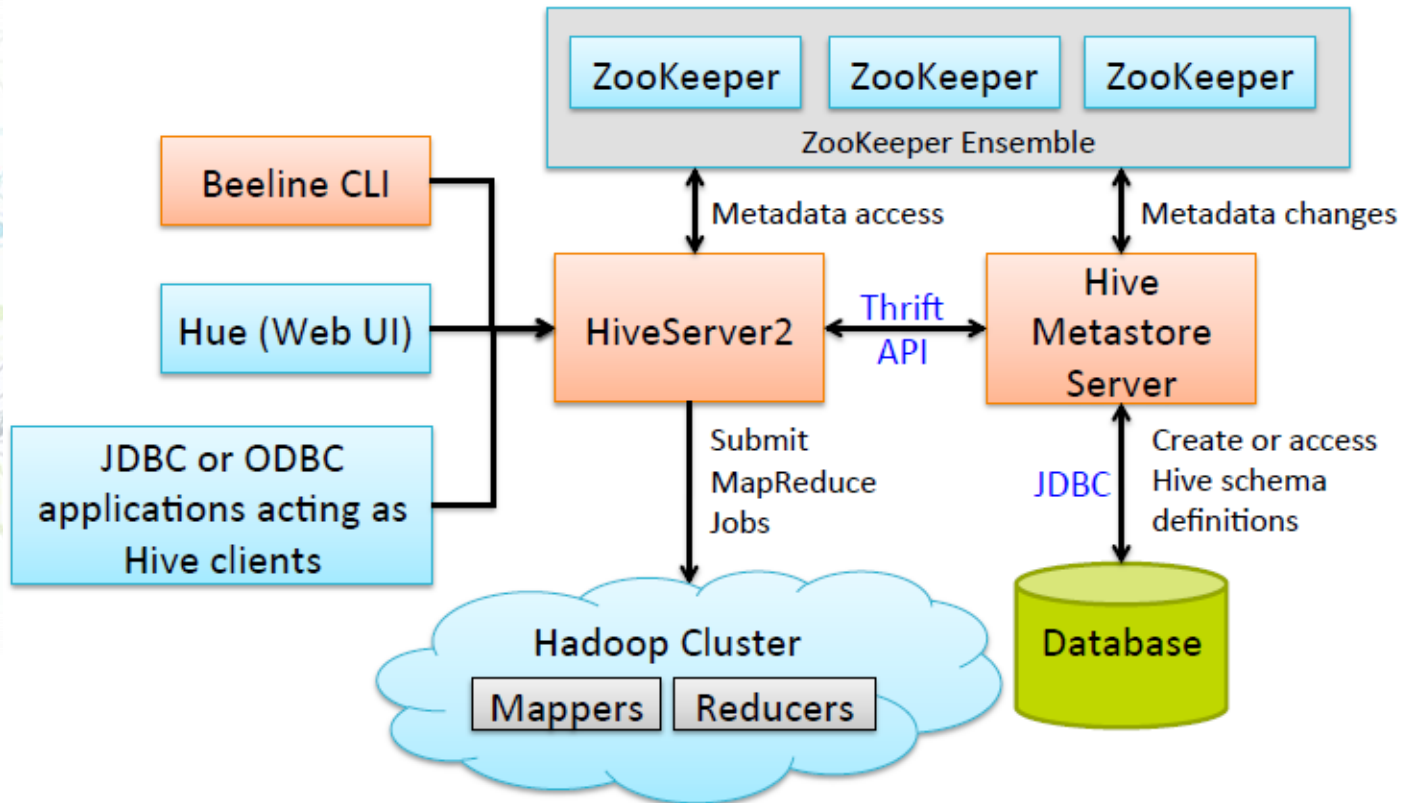
- *Una Tabla Hive es un directorio del HDFS con una metadata asociada.*
 - *Hive interpreta todos los archivos en el directorio como el contexto de la tabla.*
 - *Las tablas son creadas como Administradas(Managed) o Externas(External)*
 - *Managed: Si la tabla es eliminada, la estructura y la data del HDFS son borrados*
 - *External: Si la tabla es eliminada, solamente la estructura es borrada*
- *La Metadata (estructura de la tabla y ruta de la data) es almacenada en un RDBMS*



*** NOTA: Hive no es un RDBMS, No soporta *Update* ni *Delete***



HIVE – ARQUITECTURA





IMPALA – QUÉ ES IMPALA?

Impala es un proyecto 100% open source creado por Cloudera.

Al igual que Hive, Impala permite a los usuarios consultar (Query) data del HDFS utilizando un lenguaje tipo SQL.

A diferencia de Hive, Impala no convierte las consultas en jobs MapReduce; tiene un motor propio

Las consultas en Impala se ejecutan considerablemente mas rápido que las consultas en Hive

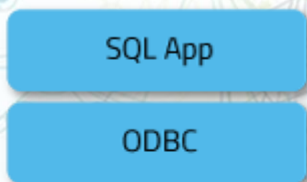
- Algunas pruebas muestran mejoras desde 10x hasta 50x o más

*** NOTA:** Impala utiliza el mismo Metastore que Hive. Las tablas que son creadas en Hive son Visibles en Impala y Viceversa.

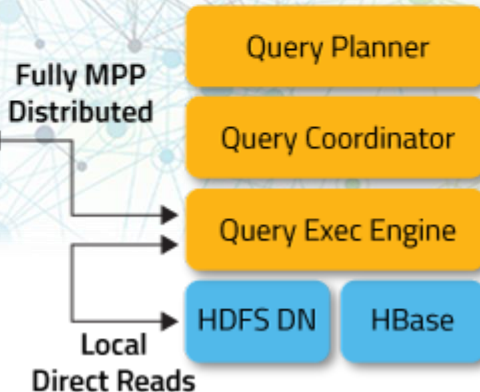
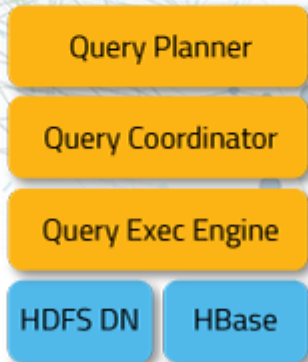
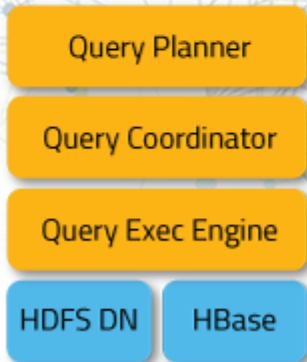


HIVE – ARQUITECTURA

Common Hive SQL and interface



Unified metadata





HIVE – ARQUITECTURA

Impala: Architecture

Common Hive SQL and interface

SQL application
(Beeswax)

ODBC

Unified metadata store

Hive metastore

HDFS Namenode

Impala statestore

impalad's continually talk to
statestore to update their state
and to receive metadata to use
for query planning

impalad

Query planner

Query coordinator

Query exec engine

HDFS DN

HBase RS

impalad

Query planner

Query coordinator

Query exec engine

HDFS DN

HBase RS

impalad

Query planner

Query coordinator

Query exec engine

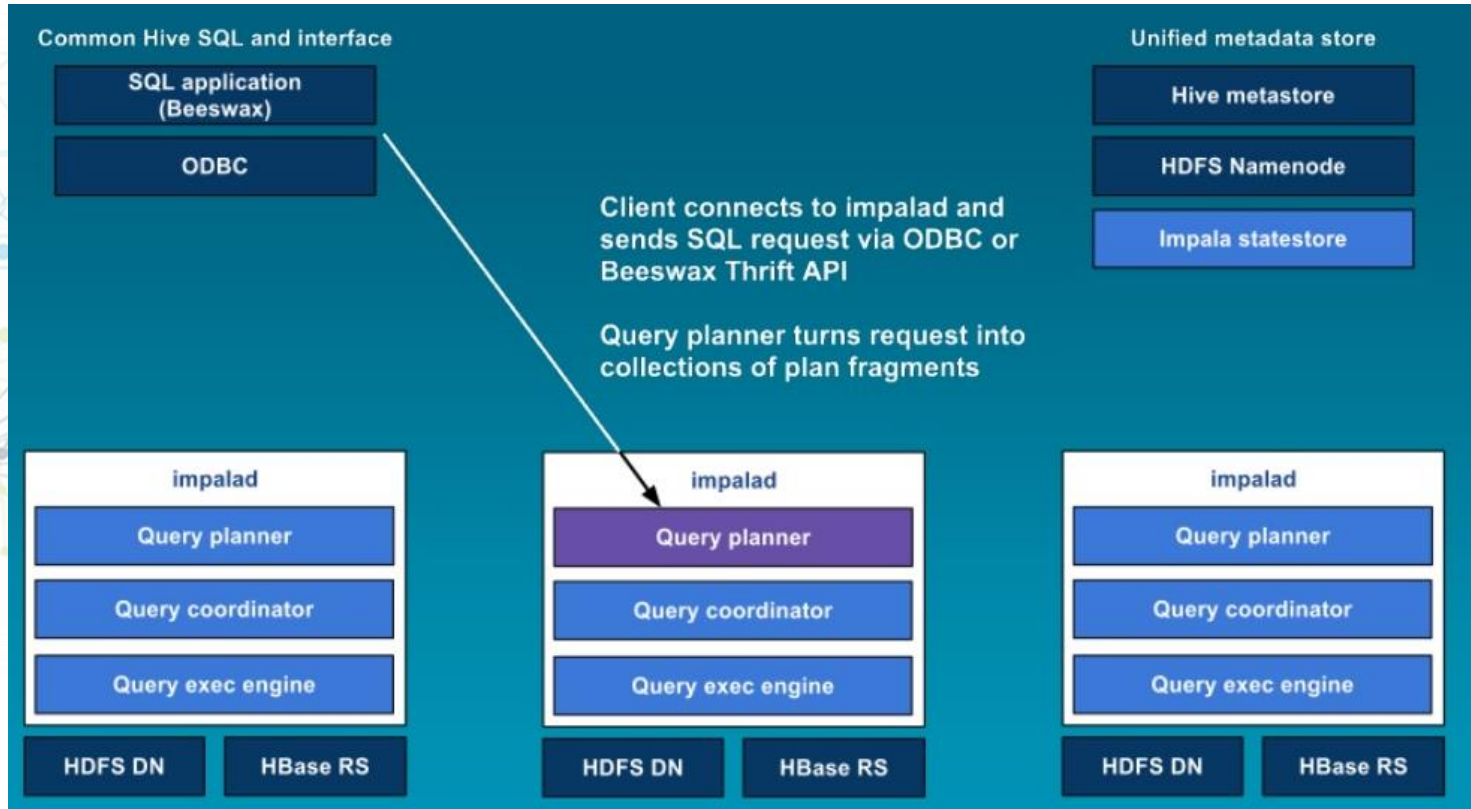
HDFS DN

HBase RS



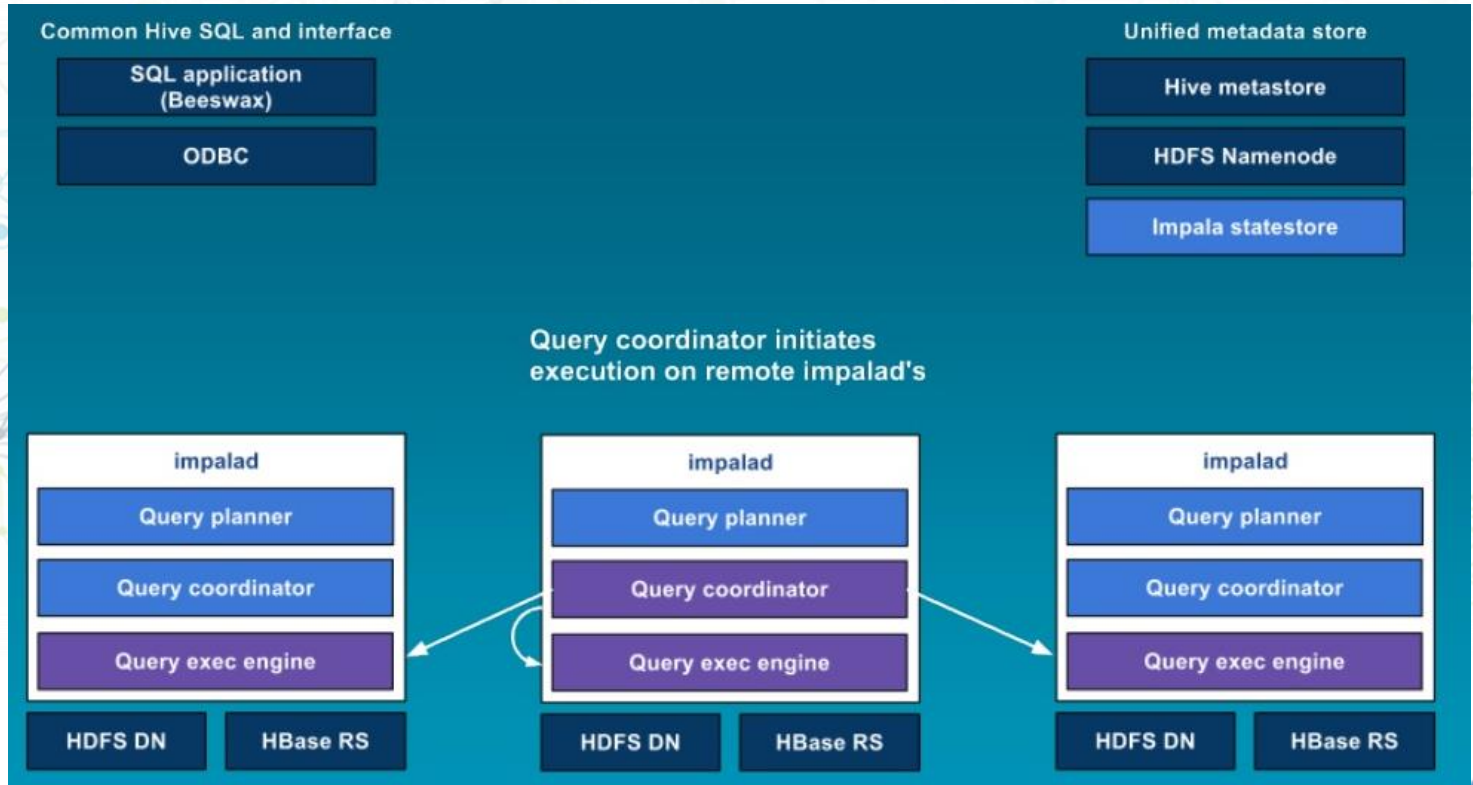


HIVE – ARQUITECTURA



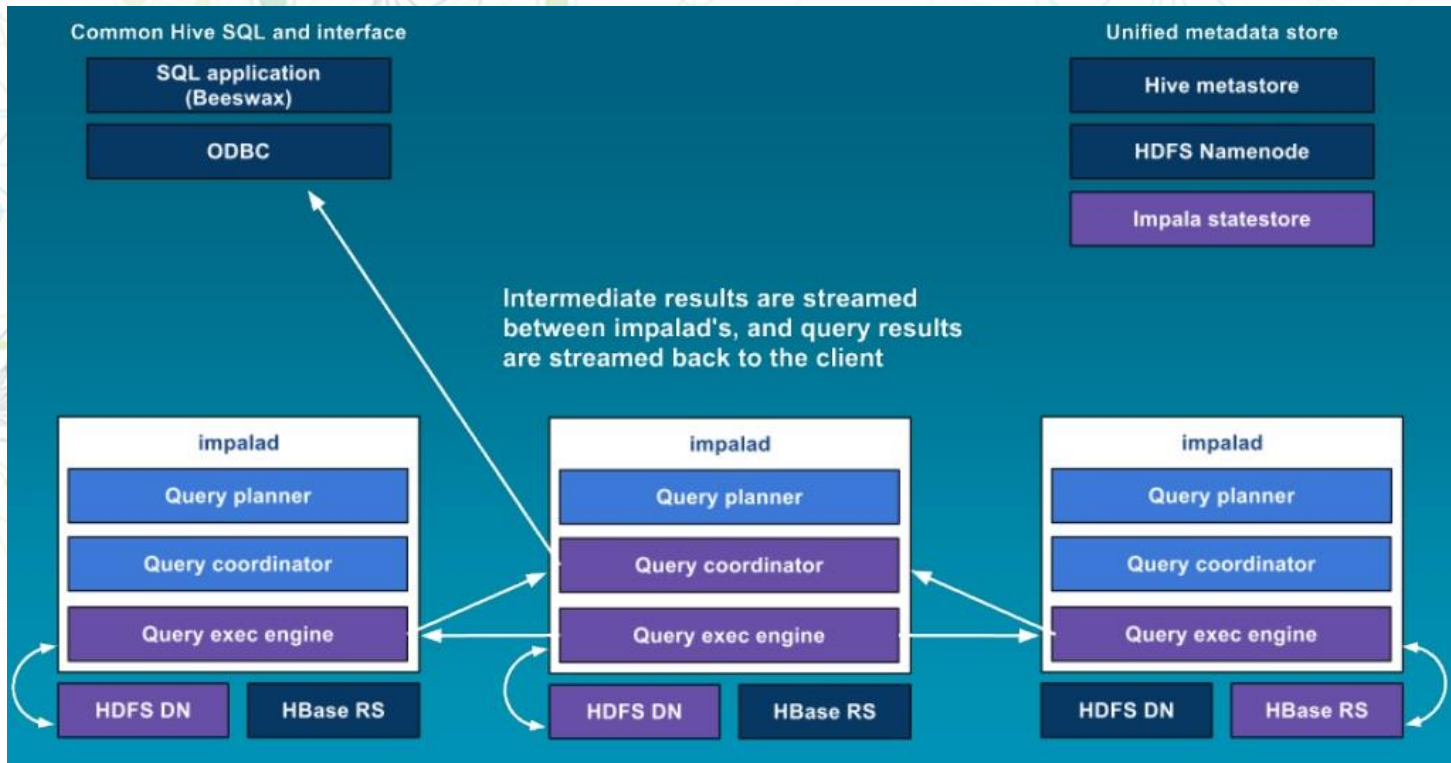


HIVE – ARQUITECTURA



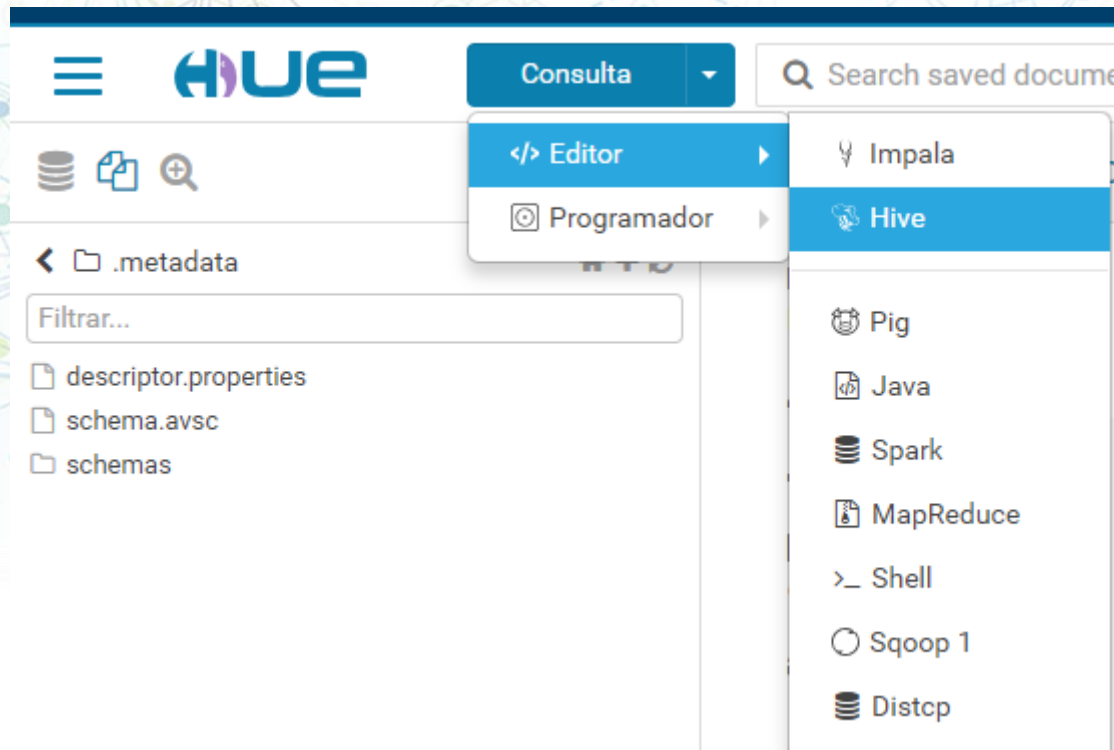


HIVE – ARQUITECTURA



DATA ANALYSIS CON HIVE/IMPALA

Ingresar a Hive/Impala utilizando Hue.



DATA ANALYSIS CON HIVE/IMPALA

Invalidate Metada: ayudará a actualizar el metastore con los últimos cambios realizados.

The screenshot displays the Hue web interface for managing data. The top navigation bar includes the Hue logo, a 'Consulta' dropdown menu, and a search bar for saved documents. The left sidebar shows a file explorer with a tree view containing '.metadata', 'descriptor.properties', 'schema.avsc', and 'schemas'. The main area is the Impala console, where the command 'INVALIDATE METADATA' has been entered and executed. The execution time is 4.59s, and the status is 'Correcto.' (Correct). Below the console, there are sections for 'Historial de consultas' (Query History) and 'Consultas guardadas' (Saved Queries). The query history shows the command 'INVALIDATE METADATA' was executed 'hace unos segundos' (a few seconds ago).

Hue

Consulta

Search saved documents...

Impala

Add a name... Añadir una d...

4.59s default text

1 INVALIDATE METADATA

Correcto.

Historial de consultas Consultas guardadas

hace unos segundos INVALIDATE METADATA

DATA ANALYSIS CON HIVE/IMPALA

Creación de Tablas:

```
CREATE EXTERNAL TABLE DEFAULT.PILOTO_VENTAS
( DESREGION STRING,
  DESZONA STRING,
  MES STRING,
  CODTIENDA STRING,
  NOMBRETIENDA STRING,
  NOMBREEMPLEADO STRING,
  PUESTOEMPLEADO STRING,
  DOCUMENTO STRING,
  NOMBRECLIENTE STRING,
  FECHAVENTA STRING,
  MONEDA STRING,
  PRODUCTO STRING,
  SUBPRODUCTO STRING,
  MONTOVENTA STRING,
  MONTOVENTASOLES STRING,
  MONTOINCENTIVO STRING
) ROW FORMAT DELIMITED
  FIELDS TERMINATED BY ';'
LOCATION '/user/dnarvaez/ventas/piloto_ventas_2018'
```



Hive



Add a name...

Añadir una descripción...

1.2s

```
1 CREATE EXTERNAL TABLE DEFAULT.PILOTO_VENTAS
2 ( DESREGION STRING,
3   DESZONA STRING,
4   MES STRING,
5   CODTIENDA STRING,
6   NOMBRETIENDA STRING,
7   NOMBREEMPLEADO STRING,
8   PUESTOEMPLEADO STRING,
9   DOCUMENTO STRING,
10  NOMBRECLIENTE STRING,
11  FECHAVENTA STRING,
12  MONEDA STRING,
13  PRODUCTO STRING,
14  SUBPRODUCTO STRING,
15  MONTOVENTA STRING,
16  MONTOVENTASOLES STRING,
17  MONTOINCENTIVO STRING
18 ) ROW FORMAT DELIMITED
19   FIELDS TERMINATED BY ';'
20 LOCATION '/user/dnarvaez/ventas/piloto_ventas_2018'
```

DATA ANALYSIS CON HIVE/IMPALA

Creación de Tablas:

```
CREATE EXTERNAL TABLE DEFAULT.PERSONAS_RENIEC  
( NUMERODOCUMENTO STRING,  
  APEPAT STRING,  
  APEMAT STRING,  
  NOMBRES STRING,  
  SEXO STRING,  
  EDAD STRING,  
  DEPARTAMENTO STRING,  
  PROVINCIA STRING,  
  DISTRITO STRING,  
  ESTADOCIVIL INT,  
  CANTHIJOS INT  
) STORED AS PARQUET  
LOCATION '/user/dnarvaez/reniec/personas_reniev'
```

Impala



Add a name...

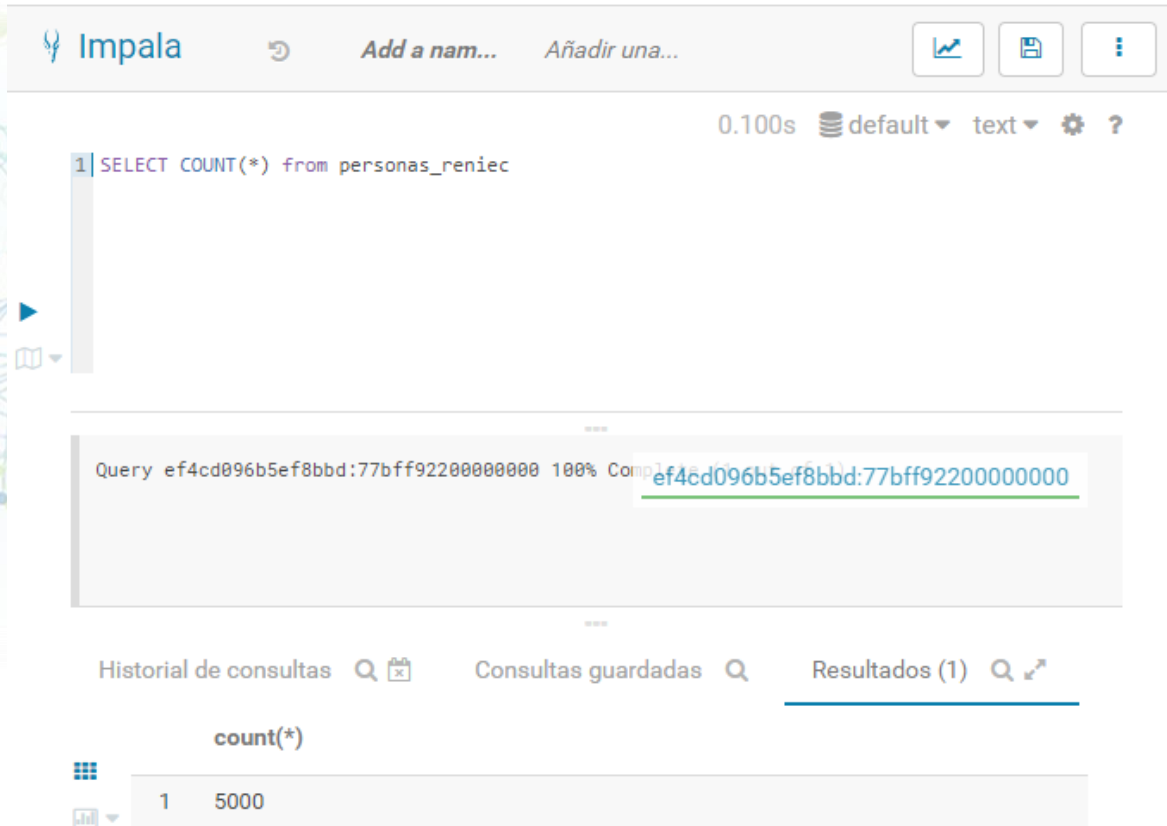
Añadir una descripción...

0.54s

```
1 CREATE EXTERNAL TABLE DEFAULT.PERSONAS_RENIEC  
2 ( NUMERODOCUMENTO STRING,  
3   APEPAT STRING,  
4   APEMAT STRING,  
5   NOMBRES STRING,  
6   SEXO STRING,  
7   EDAD STRING,  
8   DEPARTAMENTO STRING,  
9   PROVINCIA STRING,  
10  DISTRITO STRING,  
11  ESTADOCIVIL INT,  
12  CANTHIJOS INT  
13 ) STORED AS PARQUET  
14 LOCATION '/user/dnarvaez/reniec/personas_reniev'
```



DATA ANALYSIS CON HIVE/IMPALA



The screenshot shows the Impala web interface. At the top, the 'Impala' logo is on the left, and navigation links 'Add a nam...' and 'Añadir una...' are in the center. On the right, there are icons for a chart, a save button, and a menu. Below the header, the query execution time '0.100s' is shown, followed by a dropdown menu set to 'default', a 'text' dropdown, and settings icons. The main area contains a SQL query: `1|SELECT COUNT(*) from personas_reniec`. Below the query, a status bar indicates 'Query ef4cd096b5ef8bbd:77bff92200000000 100% Completed (0 rows)' with the query ID highlighted. At the bottom, there are tabs for 'Historial de consultas', 'Consultas guardadas', and 'Resultados (1)'. The 'Resultados (1)' tab is active, showing a table with one column 'count(*)' and one row with the value '5000'.

Impala

Add a nam... Añadir una...

0.100s default text ?

```
1|SELECT COUNT(*) from personas_reniec
```

Query ef4cd096b5ef8bbd:77bff92200000000 100% Completed (0 rows)

Historial de consultas Consultas guardadas Resultados (1)

	count(*)
1	5000

DATA ANALYSIS CON HIVE/IMPALA

Creación de Tablas Particionadas:

```
CREATE EXTERNAL TABLE DEFAULT.PERSONAS_RENIEC  
( NUMERODOCUMENTO STRING,  
  AEPAT STRING,  
  APEMAT STRING,  
  NOMBRES STRING,  
  SEXO STRING,  
  EDAD STRING,  
  DEPARTAMENTO STRING,  
  PROVINCIA STRING,  
  DISTRITO STRING,  
  ESTADOCIVIL INT,  
  CANTHIJOS INT  
  FECPROCESO INT  
) PARTITIONED BY (FECPROCESO INT)  
STORED AS PARQUET  
LOCATION '/user/dnarvaez/reniec/personas_reniev'
```

Impala



Add a nam...

Añadir una...

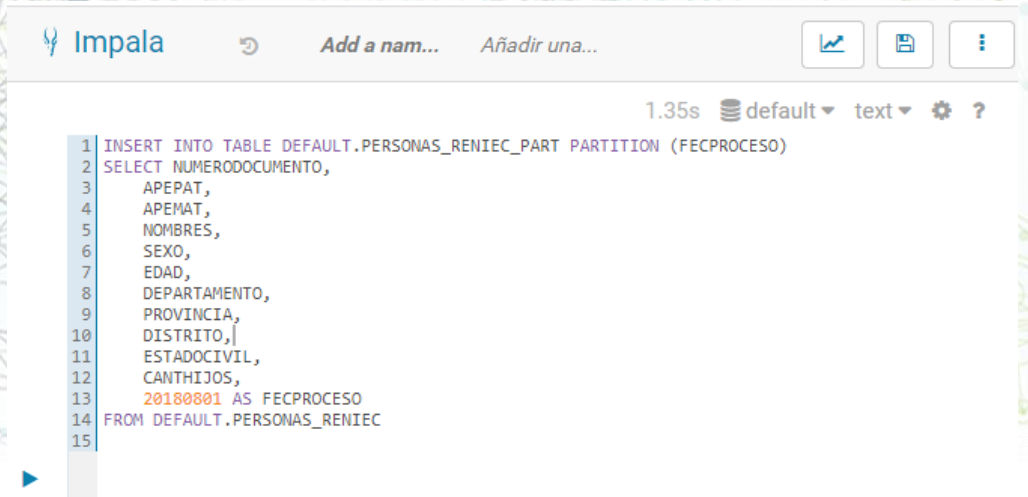
1.36s

```
1 CREATE EXTERNAL TABLE DEFAULT.PERSONAS_RENIEC_PART  
2 ( NUMERODOCUMENTO STRING,  
3   AEPAT STRING,  
4   APEMAT STRING,  
5   NOMBRES STRING,  
6   SEXO STRING,  
7   EDAD STRING,  
8   DEPARTAMENTO STRING,  
9   PROVINCIA STRING,  
10  DISTRITO STRING,  
11  ESTADOCIVIL INT,  
12  CANTHIJOS INT  
13 ) PARTITIONED BY (FECPROCESO INT) STORED AS PARQUET  
14 LOCATION '/user/dnarvaez/reniec/personas_reniev_part'  
15
```

DATA ANALYSIS CON HIVE/IMPALA


Insertar registros en Tablas :

```
INSERT INTO TABLE
DEFAULT.PERSONAS_RENIEC_PART PARTITION
(FECPROCESO)
SELECT NUMERODOCUMENTO,
APEPAT,
APEMAT,
NOMBRES,
SEXO,
EDAD,
DEPARTAMENTO,
PROVINCIA,
DISTRITO,
ESTADOCIVIL,
CANTHIJOS,
20180801 AS FECPROCESO
FROM DEFAULT.PERSONAS_RENIEC
```



DATA ANALYSIS CON HIVE/IMPALA

Verificar directorios creados en HDFS

 root@ip-10-0-0-25:/home/ec2-user

```
[root@ip-10-0-0-25 ec2-user]# hdfs dfs -ls /user/dnarvaez/reniec/personas_reniev_part
Found 3 items
drwxrwxrwx   - impala supergroup          0 2018-08-31 00:22 /user/dnarvaez/reniec/personas_reniev_part/_impala_insert_staging
drwxr-xr-x   - impala supergroup          0 2018-08-31 00:20 /user/dnarvaez/reniec/personas_reniev_part/fecproceso=20180801
drwxr-xr-x   - impala supergroup          0 2018-08-31 00:22 /user/dnarvaez/reniec/personas_reniev_part/fecproceso=20180802
[root@ip-10-0-0-25 ec2-user]#
```



GRACIAS!