Ejercicios de Impala

- **Creación de una BBDD**

1. Conéctate a `impala-shell`:

```
[cloudera@quickstart ~]$ impala-shell
Starting Impala Shell without Kerberos authentication
Connected to quickstart.cloudera:21000
Server version: impalad version 2.10.0-cdh5.13.0 RELEASE (build 2511805f1eaa991d
f1460276c7e9f19d819cd4e4)
***************************************************************************
***
Welcome to the Impala shell.
(Impala Shell v2.10.0-cdh5.13.0 (2511805) built on Wed Oct  4 10:55:37 PDT 2017)

To see live updates on a query's progress, run 'set LIVE_SUMMARY=1;'.
***************************************************************************
***
[quickstart.cloudera:21000] >
```

2. Crea una base de datos llamada `practica_impala`:

```
[quickstart.cloudera:21000] > CREATE DATABASE IF NOT EXISTS practica_impala;
Query: create DATABASE IF NOT EXISTS practica_impala
Fetched 0 row(s) in 1.39s
```

3. Usa la base de datos que acabas de crear:

```
[quickstart.cloudera:21000] > USE practica_impala;
Query: use practica impala
```

- **Trabajando con Tablas…**

1. Crea una tabla llamada `empleados` con las siguientes columnas: `id` (INT), `nombre` (STRING), `apellido` (STRING), `edad` (INT), `salario` (DOUBLE):

```
[quickstart.cloudera:21000] > CREATE TABLE IF NOT EXISTS empleados (
                            > id INT,
                            > nombre STRING,
                            > apellido STRING,
                            > edad INT,
                            > salario DOUBLE
                            > );
Query: create TABLE IF NOT EXISTS empleados (
id INT,
nombre STRING,
apellido STRING,
edad INT,
salario DOUBLE
)
Fetched 0 row(s) in 0.46s
```

2. Inserta algunos registros en la tabla `empleados`:

```
[quickstart.cloudera:21000] > INSERT INTO empleados VALUES (1, 'Juan', 'Perez',
30, 35000.50);
Query: insert INTO empleados VALUES (1, 'Juan', 'Perez', 30, 35000.50)
Query submitted at: 2024-07-17 09:49:04 (Coordinator: http://quickstart.cloudera
:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?
query_id=1e40dd87e409d3bb:fc44fdde00000000
Modified 1 row(s) in 5.39s
[quickstart.cloudera:21000] > INSERT INTO empleados VALUES (2, 'Maria', 'Gomez',
 25, 42000.00);
Query: insert INTO empleados VALUES (2, 'Maria', 'Gomez', 25, 42000.00)
Query submitted at: 2024-07-17 09:49:09 (Coordinator: http://quickstart.cloudera
:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?
query_id=9543be1cca7a277b:56c2898300000000
Modified 1 row(s) in 0.11s
[quickstart.cloudera:21000] > INSERT INTO empleados VALUES (3, 'Pedro', 'Lopez',
 45, 55000.75);
Query: insert INTO empleados VALUES (3, 'Pedro', 'Lopez', 45, 55000.75)
Query submitted at: 2024-07-17 09:49:10 (Coordinator: http://quickstart.cloudera
:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?
query_id=de4b858306cc35cf:61c6c31b00000000
Modified 1 row(s) in 0.11s
```

3. Consultar datos

a) **Selecciona todos los registros de la tabla `empleados`:**

```
[quickstart.cloudera:21000] > SELECT * FROM empleados;
Query: select * FROM empleados
Query submitted at: 2024-07-17 09:51:18 (Coordinator: http://quickstart.cloudera
:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?
query_id=2847f4feb7fcc5f5:2814b87f00000000
+----+--------+----------+------+----------+
| id | nombre | apellido | edad | salario  |
+----+--------+----------+------+----------+
| 1  | Juan   | Perez    | 30   | 35000.5  |
| 2  | Maria  | Gomez    | 25   | 42000    |
| 3  | Pedro  | Lopez    | 45   | 55000.75 |
+----+--------+----------+------+----------+
Fetched 3 row(s) in 0.41s
```

b) **Selecciona los nombres y apellidos de los empleados que ganan más de 40000:**

```
[quickstart.cloudera:21000] > SELECT nombre, apellido
                            > FROM empleados
                            > WHERE salario > 40000;
Query: select nombre, apellido
FROM empleados
WHERE salario > 40000
Query submitted at: 2024-07-17 09:52:35 (Coordinator: http://quickstart.cloudera
:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?
query_id=7a4a86a39a6e4de3:b9c66ecd00000000
+--------+----------+
| nombre | apellido |
+--------+----------+
| Maria  | Gomez    |
| Pedro  | Lopez    |
+--------+----------+
Fetched 2 row(s) in 0.16s
```

4. Actualiza el salario de `Juan Perez` a 37000.00:

Lo meto en una tabla nueva

```
[quickstart.cloudera:21000] > CREATE TABLE IF NOT EXISTS empleados_temp AS
                            > SELECT id, nombre, apellido, edad,
                            > CASE WHEN nombre='Juan' AND apellido='Perez' THEN 37000.00 ELSE salario END AS salario
                            > FROM empleados;
Query: create TABLE IF NOT EXISTS empleados_temp AS
SELECT id, nombre, apellido, edad,
CASE WHEN nombre='Juan' AND apellido='Perez' THEN 37000.00 ELSE salario END AS salario
FROM empleados
+-------------------+
| summary           |
+-------------------+
| Inserted 3 row(s) |
+-------------------+
Fetched 1 row(s) in 0.57s
```

Compruebo que esta bien:

```
[quickstart.cloudera:21000] > SELECT * FROM empleados_temp;
Query: select * FROM empleados_temp
Query submitted at: 2024-07-17 10:08:01 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=cc444060858e3e41:6903535f00000000
+----+--------+----------+------+----------+
| id | nombre | apellido | edad | salario  |
+----+--------+----------+------+----------+
| 3  | Pedro  | Lopez    | 45   | 55000.75 |
| 2  | Maria  | Gomez    | 25   | 42000    |
| 1  | Juan   | Perez    | 30   | 37000    |
+----+--------+----------+------+----------+
Fetched 3 row(s) in 0.14s
```

Elimino la tabla original:

```
[quickstart.cloudera:21000] > DROP TABLE empleados;
Query: drop TABLE empleados
```

Y renombro la creada para que se llame igual que la original:

```
[quickstart.cloudera:21000] > ALTER TABLE empleados_temp RENAME TO emplados;
Query: alter TABLE empleados_temp RENAME TO emplados
Fetched 0 row(s) in 0.27s
```

Compruebo que ha ido bien:

```
[quickstart.cloudera:21000] > select * from empleados;
Query: select * from empleados
Query submitted at: 2024-07-17 10:13:44 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=af48d87a8639de7f:1783e99d00000000
+----+--------+----------+------+----------+
| id | nombre | apellido | edad | salario  |
+----+--------+----------+------+----------+
| 2  | Maria  | Gomez    | 25   | 42000    |
| 3  | Pedro  | Lopez    | 45   | 55000.75 |
| 1  | Juan   | Perez    | 30   | 37000    |
+----+--------+----------+------+----------+
Fetched 3 row(s) in 0.14s
```

5. Elimina el registro del empleado con `id` 3:

Creo tabla temporal para poder meter los datos menos los de id = 3.

```
[quickstart.cloudera:21000] > CREATE TABLE empleados_temp AS
                            > SELECT * FROM empleados WHERE id != 3;
Query: create TABLE empleados_temp AS
SELECT * FROM empleados WHERE id != 3
+-------------------+
| summary           |
+-------------------+
| Inserted 2 row(s) |
+-------------------+
Fetched 1 row(s) in 0.31s
```

Verifico que la tabla tiene los datos correctos

```
[quickstart.cloudera:21000] > select * from empleados_temp;
Query: select * from empleados_temp
Query submitted at: 2024-07-17 10:20:24 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=6e4f07fd8a64e648:5c73d07500000000
+----+--------+----------+------+---------+
| id | nombre | apellido | edad | salario |
+----+--------+----------+------+---------+
| 2  | Maria  | Gomez    | 25   | 42000   |
| 1  | Juan   | Perez    | 30   | 37000   |
+----+--------+----------+------+---------+
Fetched 2 row(s) in 0.14s
```

Elimino la tabla original

```
[quickstart.cloudera:21000] > DROP TABLE empleados;
Query: drop TABLE empleados
```

Renombro la tabla temporal a la original

```
[quickstart.cloudera:21000] > ALTER TABLE empleados_temp RENAME TO empleados;
Query: alter TABLE empleados_temp RENAME TO empleados
Fetched 0 row(s) in 0.15s
```

Verifico la nueva tabla

```
[quickstart.cloudera:21000] > ALTER TABLE empleados_temp RENAME TO empleados;
Query: alter TABLE empleados_temp RENAME TO empleados
Fetched 0 row(s) in 0.15s
[quickstart.cloudera:21000] > select * from empleados;
Query: select * from empleados
Query submitted at: 2024-07-17 10:21:39 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=3e411c739634c531:9f913f9500000000
+----+--------+----------+------+---------+
| id | nombre | apellido | edad | salario |
+----+--------+----------+------+---------+
| 2  | Maria  | Gomez    | 25   | 42000   |
| 1  | Juan   | Perez    | 30   | 37000   |
+----+--------+----------+------+---------+
Fetched 2 row(s) in 4.44s
```

6. Describe la estructura de la tabla `empleados`

```
[quickstart.cloudera:21000] > DESCRIBE empleados;
Query: describe empleados
+----------+--------+---------+
| name     | type   | comment |
+----------+--------+---------+
| id       | int    |         |
| nombre   | string |         |
| apellido | string |         |
| edad     | int    |         |
| salario  | double |         |
+----------+--------+---------+
Fetched 5 row(s) in 0.18s
```

7. Cargar Datos desde un Archivo

a) **Crea un archivo `empleados.csv` con los siguientes datos y cárgalo en HDFS:**

He creado un archivo csv y lo he cargado hdfs.

```
[cloudera@quickstart Desktop]$  hdfs dfs -put ~/Desktop/empleados.csv /user/clou
dera/empleados.csv;
```

b) **Carga los datos del archivo `empleados.csv` en la tabla `empleados`:**

Me daba fallos de permisos

```
[quickstart.cloudera:21000] > LOAD DATA INPATH '/user/cloudera/empleados.csv' INTO TABLE empleados;
Query: load DATA INPATH '/user/cloudera/empleados.csv' INTO TABLE empleados
ERROR: AnalysisException: Unable to LOAD DATA from hdfs://quickstart.cloudera:8020/user/cloudera/empleados.csv because Impala
 does not have WRITE permissions on its parent directory hdfs://quickstart.cloudera:8020/user/cloudera
```

Le he cambiado los permisos:

```
[cloudera@quickstart Desktop]$ sudo -u hdfs hdfs dfs -chmod -R 777 /user/cloudera
```

y lo he cargado con éxito:

```
[quickstart.cloudera:21000] > LOAD DATA INPATH '/user/cloudera/empleados.csv' INTO TABLE empleados;
Query: load DATA INPATH '/user/cloudera/empleados.csv' INTO TABLE empleados
+----------------------------------------------------------+
| summary                                                  |
+----------------------------------------------------------+
| Loaded 1 file(s). Total files in destination location: 2 |
+----------------------------------------------------------+
Fetched 1 row(s) in 0.31s
```

8. Crear una Tabla Particionada

La partición en Impala es una técnica que divide físicamente los datos durante la carga, basándose en valores de una o más columnas. Esto acelera las consultas que prueban esas columnas.

a) Crea una tabla llamada `ventas` particionada por `año` y `mes`:

```
[quickstart.cloudera:21000] > CREATE TABLE IF NOT EXISTS ventas (
                            >     id INT,
                            >     producto STRING,
                            >     cantidad INT,
                            >     precio DOUBLE
                            > )
                            > PARTITIONED BY (anio INT, mes INT);
Query: create TABLE IF NOT EXISTS ventas (
    id INT,
    producto STRING,
    cantidad INT,
    precio DOUBLE
)
PARTITIONED BY (anio INT, mes INT)
Fetched 0 row(s) in 0.18s
```

b) Inserta datos en la tabla `ventas` especificando las particiones:

```
[quickstart.cloudera:21000] > INSERT INTO ventas PARTITION (anio=2023, mes=7) VALUES (1, 'Producto A', 100, 9.99);
Query: insert INTO ventas PARTITION (anio=2023, mes=7) VALUES (1, 'Producto A', 100, 9.99)
Query submitted at: 2024-07-17 11:28:58 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=224341cab801ff7d:b5ad8b5800000000
Modified 1 row(s) in 3.19s
[quickstart.cloudera:21000] > INSERT INTO ventas PARTITION (anio=2023, mes=7) VALUES (2, 'Producto B', 200, 19.99);
Query: insert INTO ventas PARTITION (anio=2023, mes=7) VALUES (2, 'Producto B', 200, 19.99)
Query submitted at: 2024-07-17 11:29:01 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=10416f88b1d07466:7639229500000000
Modified 1 row(s) in 0.11s
[quickstart.cloudera:21000] > INSERT INTO ventas PARTITION (anio=2023, mes=8) VALUES (3, 'Producto C', 150, 29.99);
Query: insert INTO ventas PARTITION (anio=2023, mes=8) VALUES (3, 'Producto C', 150, 29.99)
Query submitted at: 2024-07-17 11:29:01 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=c74497f03e670e47:7b5f01a300000000
Modified 1 row(s) in 0.31s
```

c) Muestra todas las particiones de la tabla `ventas`:

```
[quickstart.cloudera:21000] > SHOW PARTITIONS ventas;
Query: show PARTITIONS ventas
+-------+-----+-------+--------+------+--------------+-------------------+--------+-------------------+----------------------
-------------------------------------------------------------------------+
| anio  | mes | #Rows | #Files | Size | Bytes Cached | Cache Replication | Format | Incremental stats | Location
                                                                        |
+-------+-----+-------+--------+------+--------------+-------------------+--------+-------------------+----------------------
-------------------------------------------------------------------------+
| 2023  | 7   | -1    | 2      | 45B  | NOT CACHED   | NOT CACHED        | TEXT   | false             | hdfs://quickstart.clo
udera:8020/user/hive/warehouse/practica_impala.db/ventas/anio=2023/mes=7 |
| 2023  | 8   | -1    | 1      | 23B  | NOT CACHED   | NOT CACHED        | TEXT   | false             | hdfs://quickstart.clo
udera:8020/user/hive/warehouse/practica_impala.db/ventas/anio=2023/mes=8 |
| Total |     | -1    | 3      | 68B  | 0B           |                   |        |                   |
                                                                        |
+-------+-----+-------+--------+------+--------------+-------------------+--------+-------------------+----------------------
-------------------------------------------------------------------------+
Fetched 3 row(s) in 0.12s
```

9. Optimización y Metadatos

a) Actualiza los metadatos de la tabla `empleados`;

```
[quickstart.cloudera:21000] > INVALIDATE METADATA empleados;
Query: invalidate METADATA empleados
Query submitted at: 2024-07-17 11:33:53 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=e6440bc7c8c4705c:b79e903b00000000
Fetched 0 row(s) in 0.12s
```

b) Refresca los metadatos de la tabla `ventas`:

```
[quickstart.cloudera:21000] > REFRESH ventas
                            > ;
Query: refresh ventas
Query submitted at: 2024-07-17 11:35:18 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=a64af089f5725052:d2ccefa300000000
Fetched 0 row(s) in 0.14s
```

10. Uso de Vistas

Una vista en Impala es una consulta almacenada que se comporta como una tabla virtual. A diferencia de las tablas físicas, las vistas no almacenan datos, sino que proporcionan una representación lógica de los datos existentes en otras tablas.

a) Crea una vista llamada `vista_empleados` que muestra solo los nombres y salarios de los empleados

```
[quickstart.cloudera:21000] > CREATE VIEW IF NOT EXISTS vista_empleados AS SELECT nombre, salario FROM empleados;
Query: create VIEW IF NOT EXISTS vista_empleados AS SELECT nombre, salario FROM empleados
Fetched 0 row(s) in 3.35s
```

b) Consulta la vista `vista_empleados`:

```
[quickstart.cloudera:21000] > SELECT * FROM vista_empleados;
Query: select * FROM vista_empleados
Query submitted at: 2024-07-17 11:38:26 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=aa4b42fff169b076:b17ad55800000000
+--------+---------+
| nombre | salario |
+--------+---------+
| Maria  | 42000   |
| Juan   | 37000   |
| NULL   | NULL    |
| NULL   | NULL    |
+--------+---------+
Fetched 4 row(s) in 3.30s
```

11. Calcula el salario promedio por edad en la tabla `empleados`.

```
[quickstart.cloudera:21000] > SELECT edad, AVG(salario) OVER (PARTITION BY edad) AS salario_promedio
                            > FROM empleados;
Query: select edad, AVG(salario) OVER (PARTITION BY edad) AS salario_promedio
FROM empleados
Query submitted at: 2024-07-17 11:42:56 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=4244ddc3266dc24a:cc6b3d0400000000
+------+------------------+
| edad | salario_promedio |
+------+------------------+
| NULL | NULL             |
| NULL | NULL             |
| 25   | 42000            |
| 30   | 37000            |
+------+------------------+
Fetched 4 row(s) in 0.21s
```