La finalidad de esta práctica es familiarizarse con el entorno HIVE.

Vete haciendo capturas de pantalla de todos los pasos que vayas dando, acompañándolas de comentarios descriptivos de los mismos.

INTRODUCCIÓN

A.- Continuaremos trabajando con el conjunto de datos **Movielens** de **Kaggle:**

https://www.kaggle.com/datasets/prajitdatta/movielens-100k-dataset

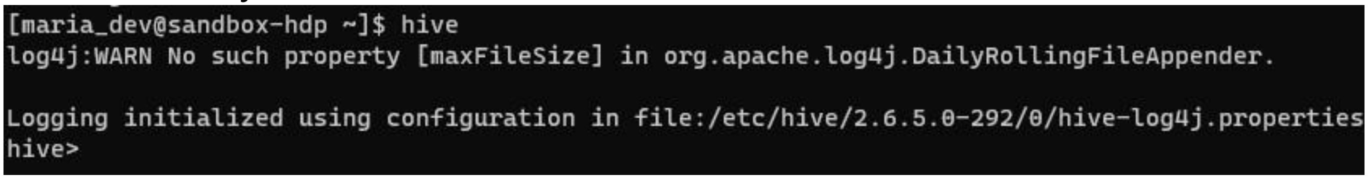
CONTENIDO

APARTADO A

Práctica con HIVE

De modo similar a como hicimos con PIG, por cada uno de los *scripts* siguientes:

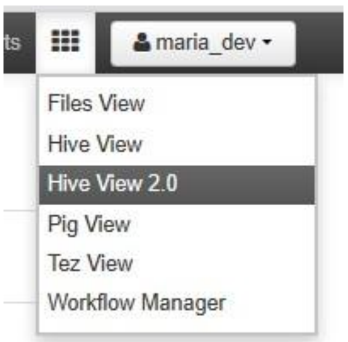
Prueba su ejecución tanto desde la consola de Hive



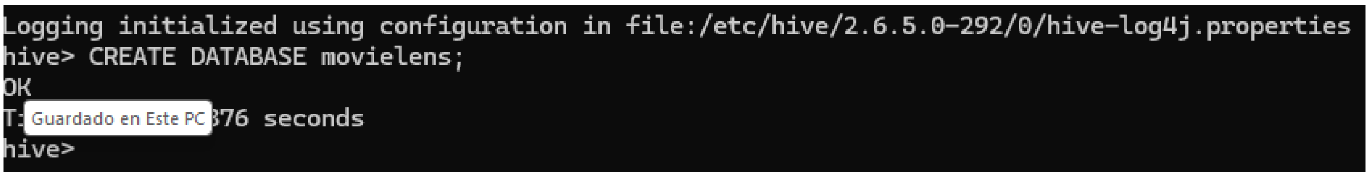
```
[maria_dev@sandbox-hdp ~]$ hive
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
hive>
```

como desde Ambari Hive View 2.0.

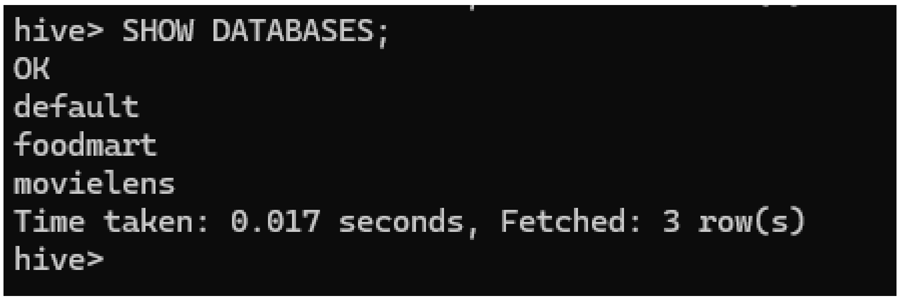
- Muestra el contenido de los *scripts.*
- Muestra un ejemplo de su ejecución con la salida de los datos por pantalla tanto en la consola como desde el entorno de Ambari.
- Posteriormente guarda las consultas en un archivo y ejecútalas todas juntas desde la línea de comandos.

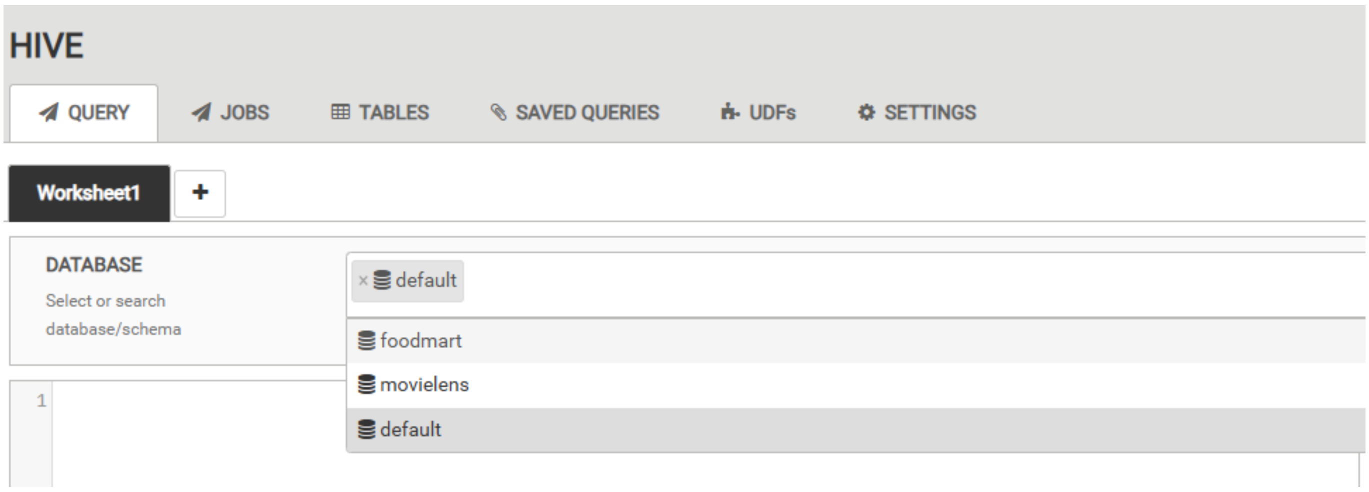


1.- Crea una base de datos que llamaremos movielens para almacenar las tablas necesarias. Para cada una de las consultas deberás crear previamente las tablas y cargar los datos necesarios para poder realizarlas.



```
Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
hive> CREATE DATABASE movielens;
OK
T: Guardado en Este PC 376 seconds
hive>
```


```
hive> SHOW DATABASES;
OK
default
foodmart
movielens
Time taken: 0.017 seconds, Fetched: 3 row(s)
hive>
```



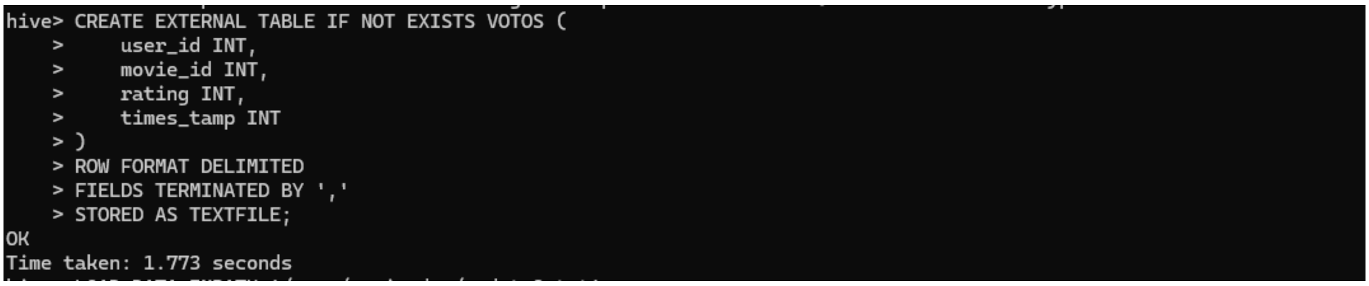
Tablas necesarias PELICULAS, USUARIOS, VOTACIONES

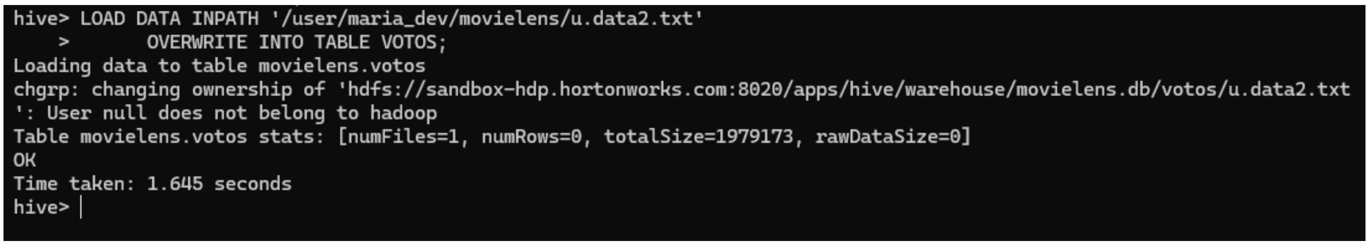
VOTOS: u.data2.txt : user id , movie id , rating , timestamp. (AHORA SEPARADO POR COMAS)

USUARIOS: - u.user2.txt: user id , age , gender , occupation | (AHORA SEPARADO POR COMAS)

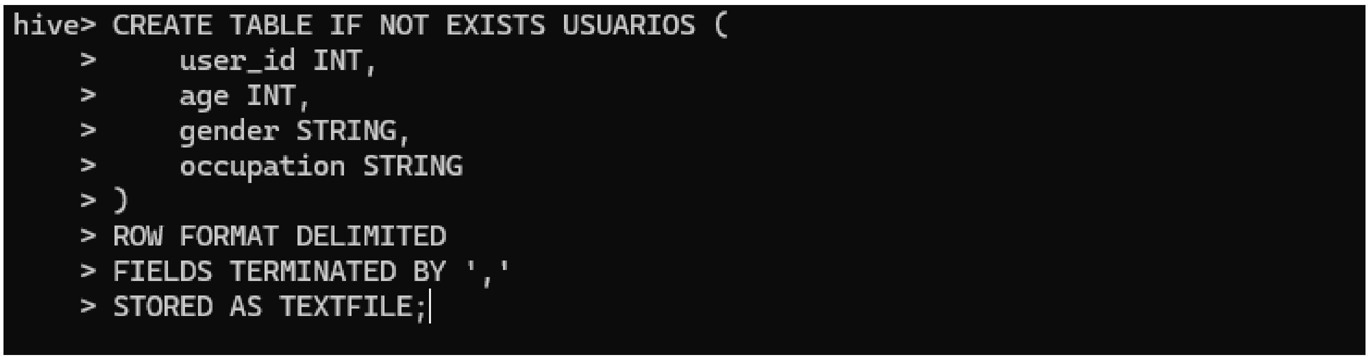
PELICULAS: - u.item2.txt: movie id | movie title | release date (SEPARADAO POR TAB)

CARGAMOS VOTOS

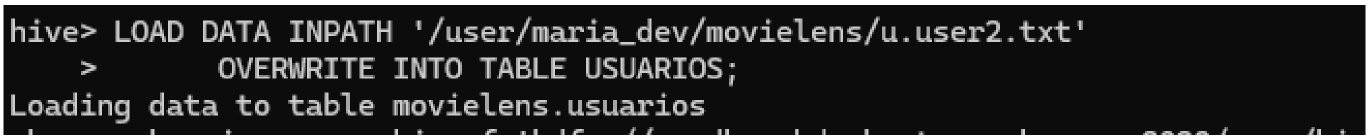

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS VOTOS (
    >     user_id INT,
    >     movie_id INT,
    >     rating INT,
    >     times_tamp INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 1.773 seconds
```


```
hive> LOAD DATA INPATH '/user/maria_dev/movielens/u.data2.txt'
    >       OVERWRITE INTO TABLE VOTOS;
Loading data to table movielens.votos
chgrp: changing ownership of 'hdfs://sandbox-hdp.hortonworks.com:8020/apps/hive/warehouse/movielens.db/votos/u.data2.txt
': User null does not belong to hadoop
Table movielens.votos stats: [numFiles=1, numRows=0, totalSize=1979173, rawDataSize=0]
OK
Time taken: 1.645 seconds
hive> |
```

CARGAMOS LOS USUARIOS

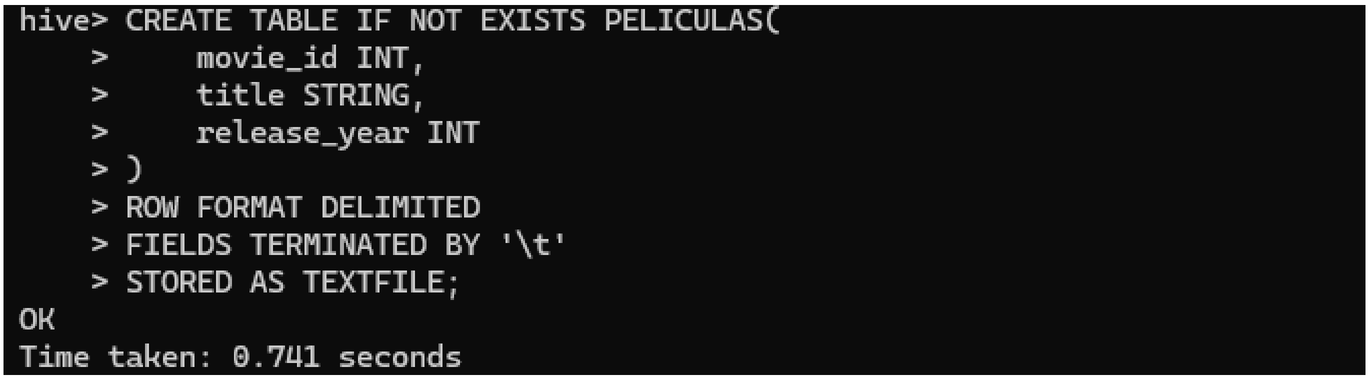


```
hive> CREATE TABLE IF NOT EXISTS USUARIOS (
    >     user_id INT,
    >     age INT,
    >     gender STRING,
    >     occupation STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
```

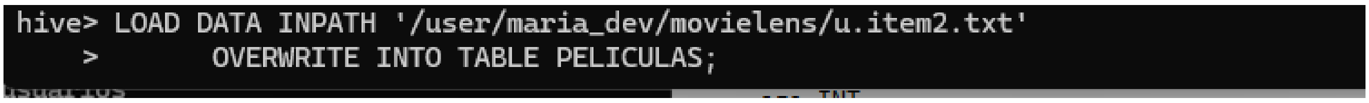


```
hive> LOAD DATA INPATH '/user/maria_dev/movielens/u.user2.txt'
    >       OVERWRITE INTO TABLE USUARIOS;
Loading data to table movielens.usuarios
```

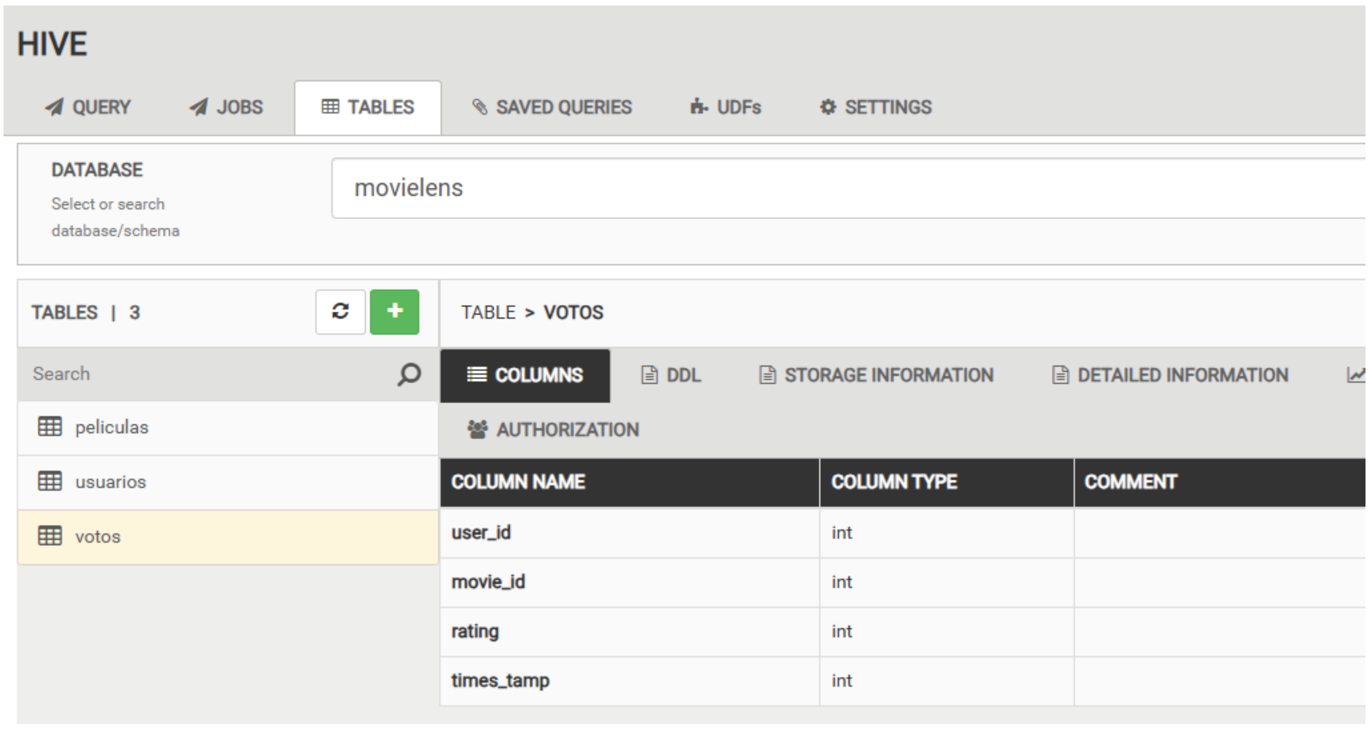
CARGAMOS LAS PELICULAS



```
hive> CREATE TABLE IF NOT EXISTS PELICULAS(
    >     movie_id INT,
    >     title STRING,
    >     release_year INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t'
    > STORED AS TEXTFILE;
OK
Time taken: 0.741 seconds
```



```
hive> LOAD DATA INPATH '/user/maria_dev/movielens/u.item2.txt'
    >       OVERWRITE INTO TABLE PELICULAS;
```



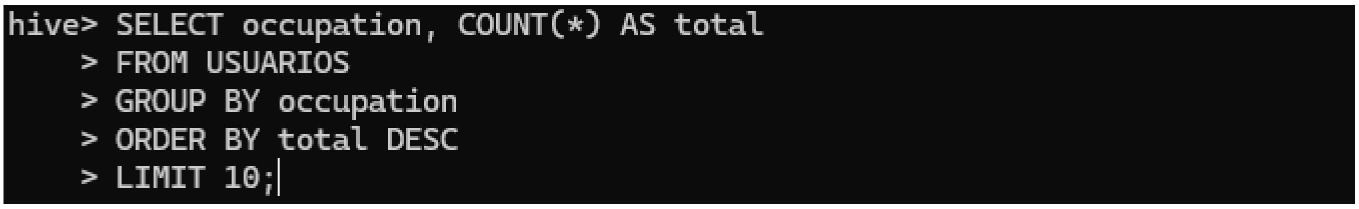
movielens

peliculas
usuarios
votos

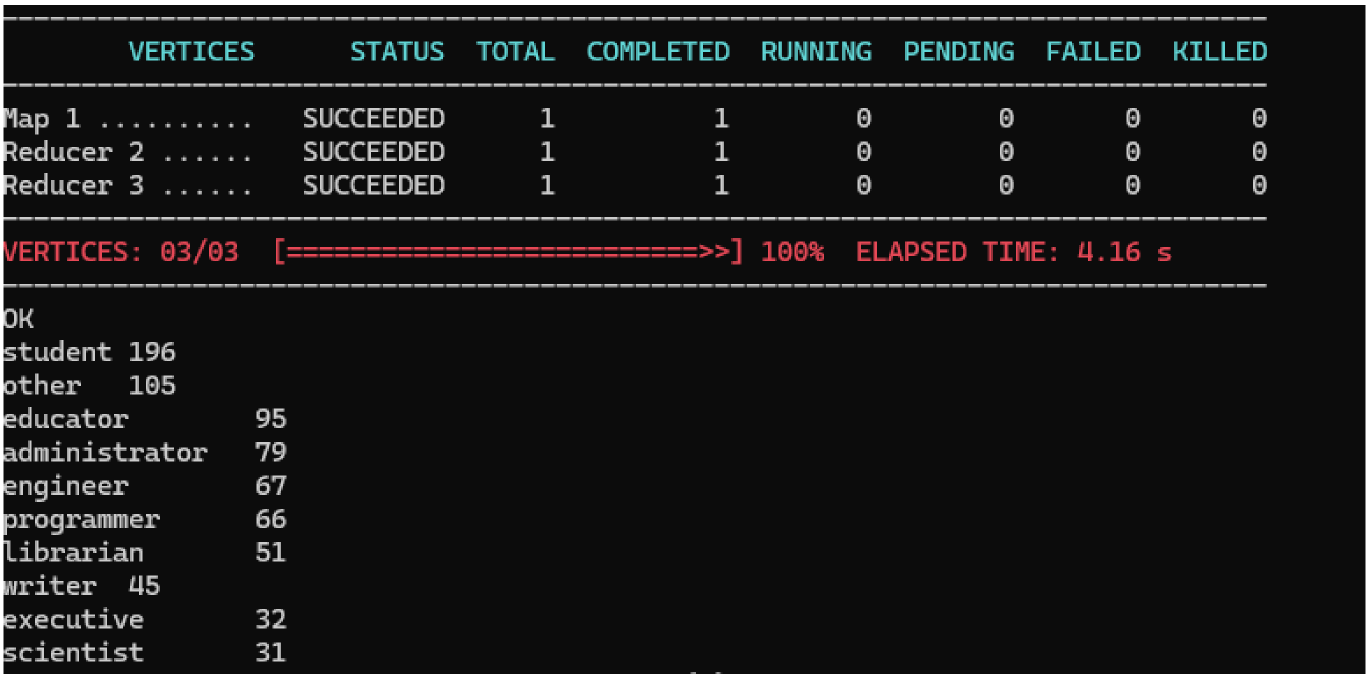
TABLE > VOTOS

COLUMN NAME	COLUMN TYPE	COMMENT
user_id	int	
movie_id	int	
rating	int	
times_tamp	int	

2.- Encontrar las 10 ocupaciones más frecuentes entre los votantes

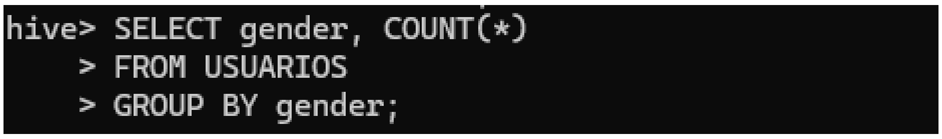


```
hive> SELECT occupation, COUNT(*) AS total
    > FROM USUARIOS
    > GROUP BY occupation
    > ORDER BY total DESC
    > LIMIT 10;
```

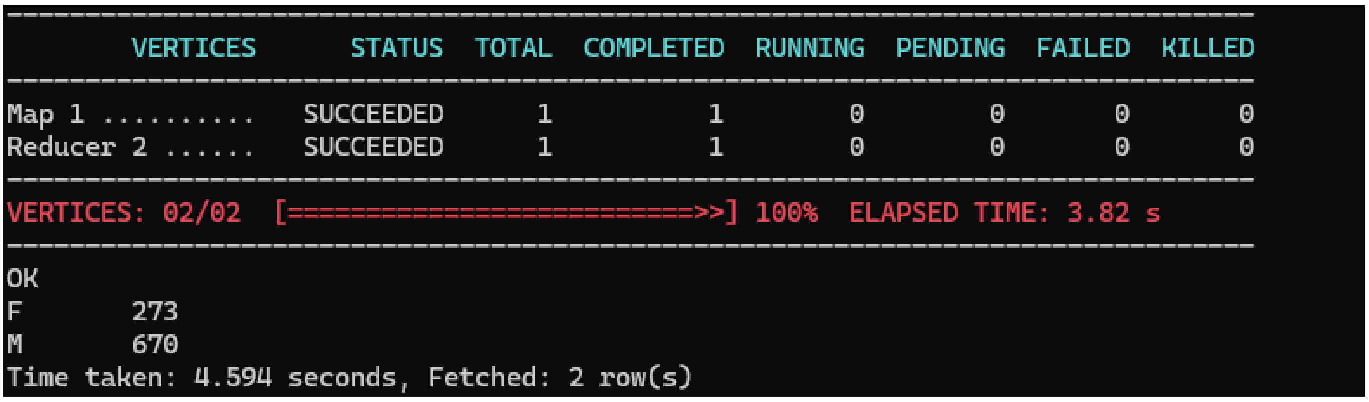


```
----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.16 s
----------------------------------------------------------------------------------
OK
student 196
other   105
educator        95
administrator   79
engineer        67
programmer      66
librarian       51
writer  45
executive       32
scientist       31
```

3.- Y luego el número de hombres y mujeres

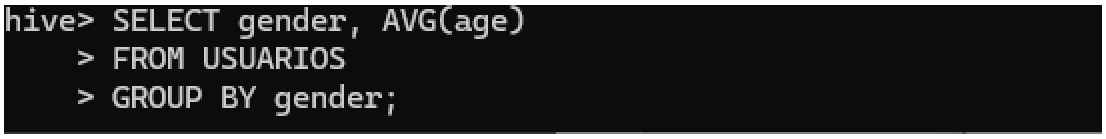


```
hive> SELECT gender, COUNT(*)
    > FROM USUARIOS
    > GROUP BY gender;
```

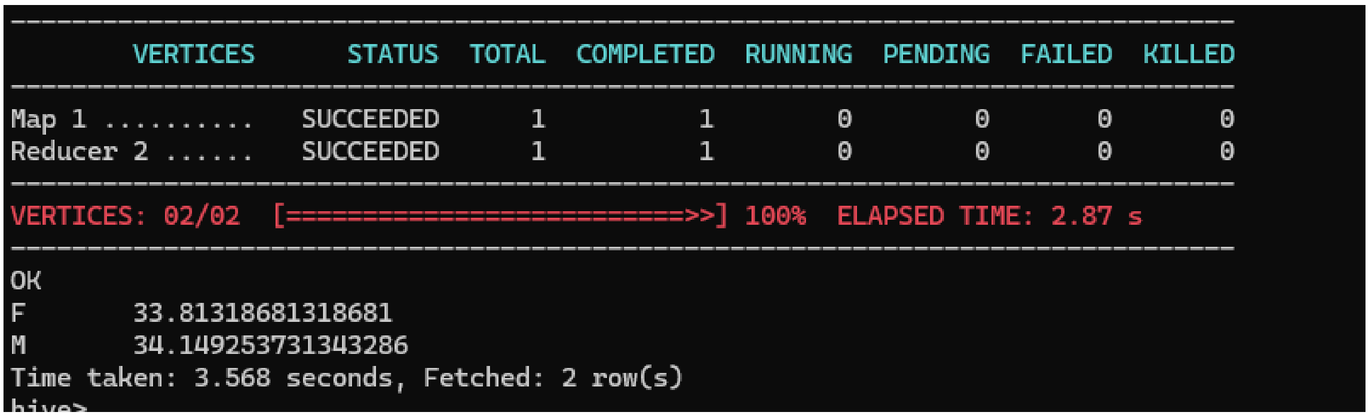


```
----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 3.82 s
----------------------------------------------------------------------------------
OK
F       273
M       670
Time taken: 4.594 seconds, Fetched: 2 row(s)
```

4.- Muestra la edad media por géneros.

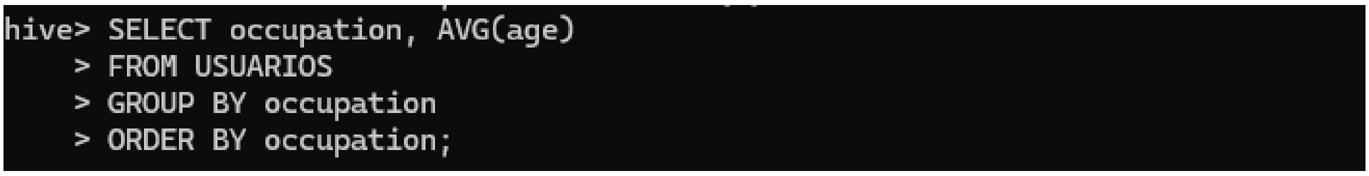


```
hive> SELECT gender, AVG(age)
    > FROM USUARIOS
    > GROUP BY gender;
```

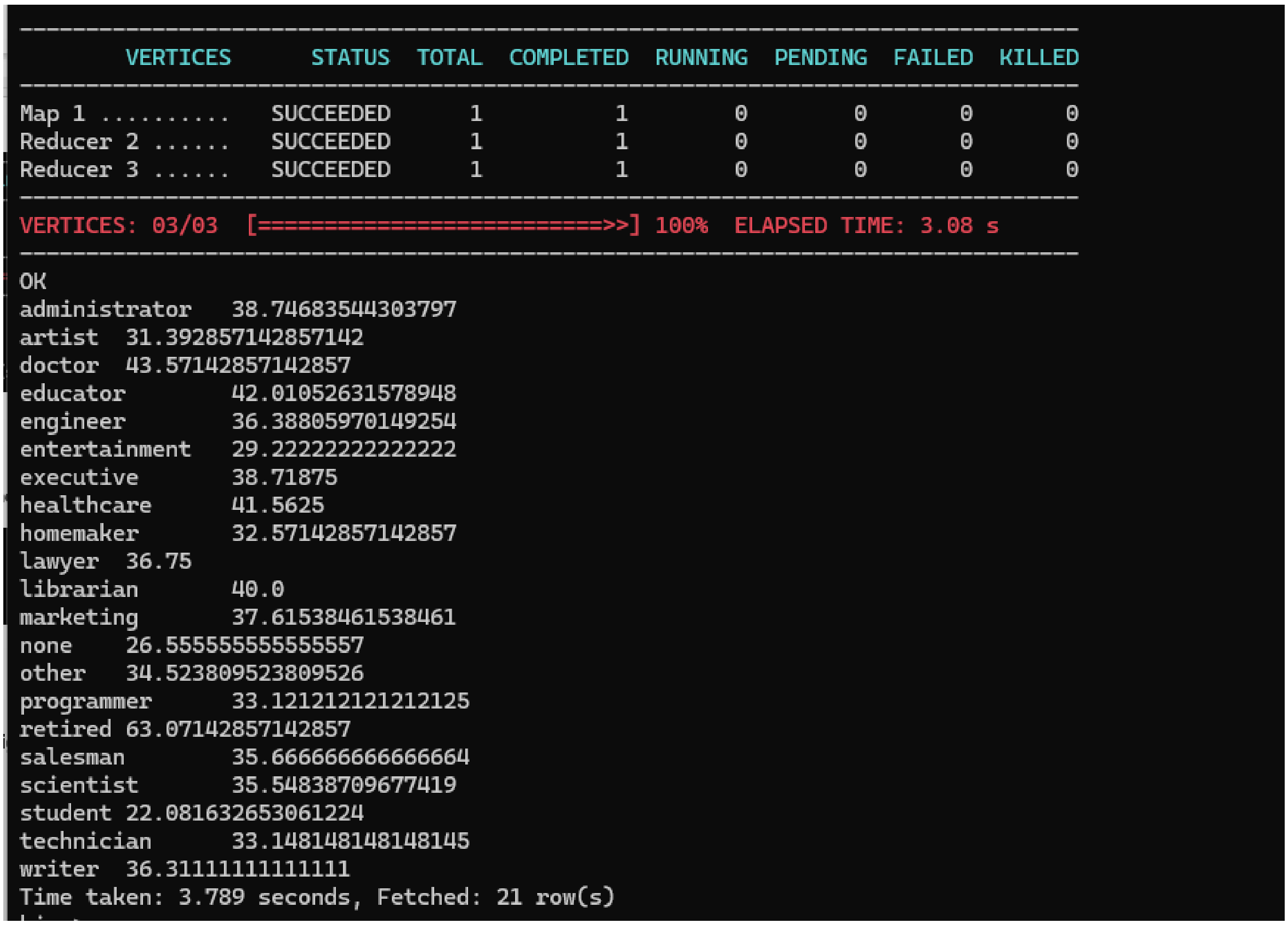


```
----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 2.87 s
----------------------------------------------------------------------------------------------
OK
F       33.81318681318681
M       34.149253731343286
Time taken: 3.568 seconds, Fetched: 2 row(s)
```

5.- Muestra la edad media por ocupaciones.

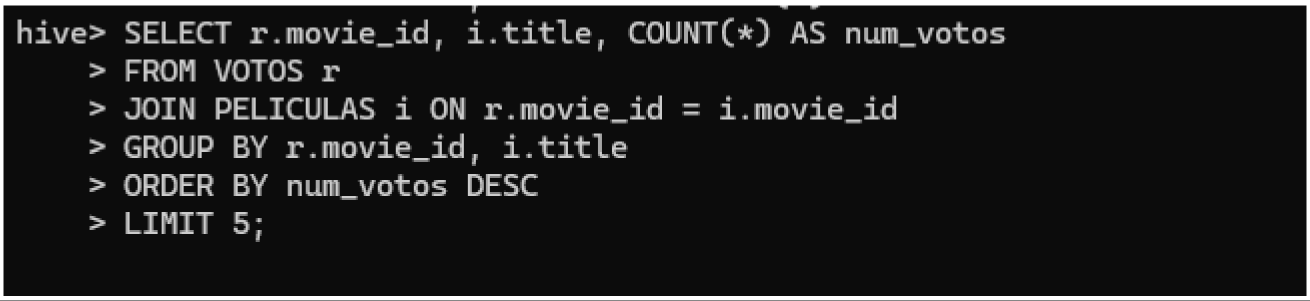


```
hive> SELECT occupation, AVG(age)
    > FROM USUARIOS
    > GROUP BY occupation
    > ORDER BY occupation;
```

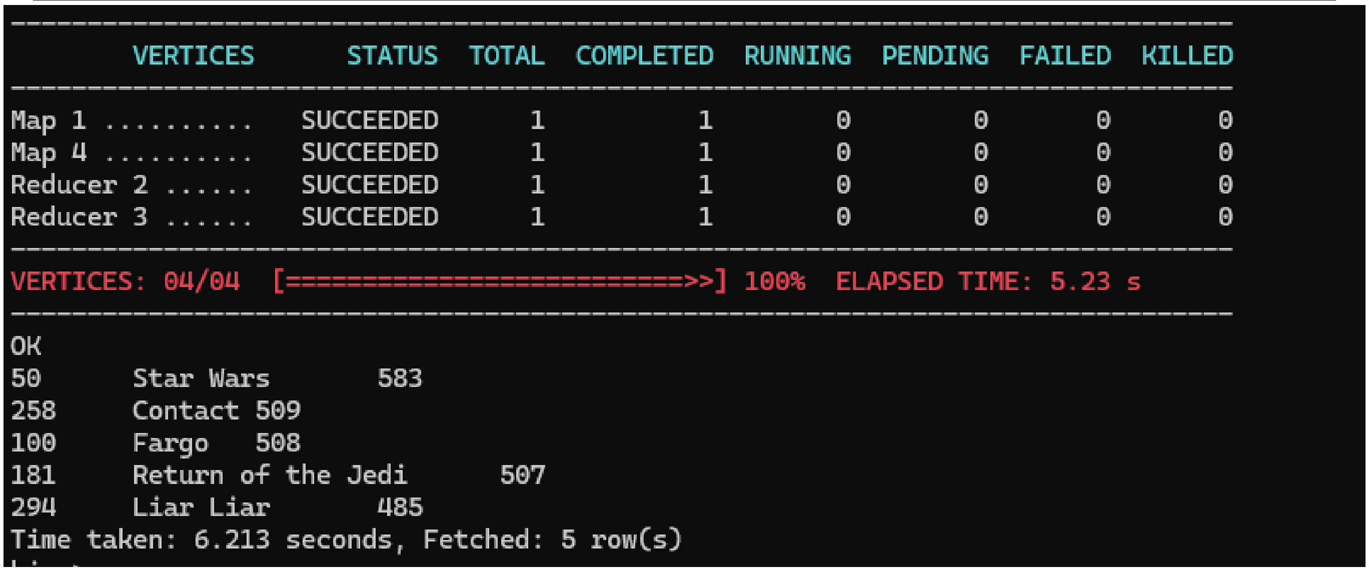


```
----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 3.08 s
----------------------------------------------------------------------------------------------
OK
administrator   38.74683544303797
artist  31.392857142857142
doctor  43.57142857142857
educator        42.01052631578948
engineer        36.38805970149254
entertainment   29.22222222222222
executive       38.71875
healthcare      41.5625
homemaker       32.57142857142857
lawyer  36.75
librarian       40.0
marketing       37.61538461538461
none    26.555555555555557
other   34.523809523809526
programmer      33.121212121212125
retired 63.07142857142857
salesman        35.666666666666664
scientist       35.54838709677419
student 22.081632653061224
technician      33.148148148148145
writer  36.31111111111111
Time taken: 3.789 seconds, Fetched: 21 row(s)
```

6.- Encontrar las cinco películas (código, título y número de votos) más votadas (recuento de votos, no media).



```
hive> SELECT r.movie_id, i.title, COUNT(*) AS num_votos
    > FROM VOTOS r
    > JOIN PELICULAS i ON r.movie_id = i.movie_id
    > GROUP BY r.movie_id, i.title
    > ORDER BY num_votos DESC
    > LIMIT 5;
```



```
--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 4 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 5.23 s
--------------------------------------------------------------------------------
OK
50      Star Wars       583
258     Contact 509
100     Fargo   508
181     Return of the Jedi      507
294     Liar Liar       485
Time taken: 6.213 seconds, Fetched: 5 row(s)
```