

# Práctica con HIVE

## 1) Apartado A

- 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.

Script creación de la base de datos, es el mismo que la práctica anterior pero adaptado a Hive:

```
CREATE DATABASE IF NOT EXISTS movielens;

USE movielens;

-- Tabla usuario (u.user2)
CREATE TABLE usuario (
    id INT,
    edad INT,
    genero STRING,
    ocupacion STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t';

-- Tabla pelicula (u.item2)
CREATE TABLE pelicula (
    id INT,
    titulo STRING,
    anyo INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|';

-- Tabla voto (u.data)
CREATE TABLE voto (
    usuario_id INT,
    pelicula_id INT,
    valoracion INT,
    fecha BIGINT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t';
```

# PRÁCTICA CON HIVE

The screenshot shows the Apache Hive Workbench interface. In the top left, it says "Worksheet1 \*". Below that, under "DATABASE", it shows "Select or search database/schema" and "default". The main area contains the following SQL code:

```
1 CREATE DATABASE IF NOT EXISTS movielens;
2
3 USE movielens;
4
5 -- Tabla usuario (u.user2)
6 CREATE TABLE usuario (
7     id INT,
8     edad INT,
9     genero STRING,
10    ocupacion STRING
11 )
12 ROW FORMAT DELIMITED
13 FIELDS TERMINATED BY '|';
14
```

Below the code, there are buttons for "Execute", "Save As", "Insert UDF", and "Visual Explain". The "RESULTS" tab is selected, showing the output of the EXPLAIN command:

```
{"STAGE DEPENDENCIES": "Stage-0: {\"ROOT STAGE\": \"TRUE\"}"; "STAGE PLANS": "[Stage-0: {\"Create Table Operator: {\"Create Table\" : {\"serde name\": \"org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe\", \"field delimiter\": \":\", \"name\": \"movielens.voto\", \"input format\": \"org.apache.hadoop.mapred.TextInputFormat\", \"output format\": \"org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat\", \"columns\": [\"usuario_id int\", \"pelicula_id int\", \"valoracion int\", \"fecha bigint\"]}}}}]"}
```

At the bottom, there are buttons for "Execute", "Save As", "Insert UDF", and "Visual Explain", along with tabs for "RESULTS", "LOG", "VISUAL EXPLAIN", and "TEZ UI".

Insertamos los datos a la base:

The screenshot shows the Apache Hive Workbench interface. In the top left, it says "Worksheet1 \*". Below that, under "DATABASE", it shows "Select or search database/schema" and "default". The main area contains the following SQL code:

```
1 LOAD DATA INPATH '/user/maria_dev/movielens/u.user2'
2 INTO TABLE usuario;
3
4 LOAD DATA INPATH '/user/maria_dev/movielens/u.item2'
5 INTO TABLE pelicula;
6
7 LOAD DATA INPATH '/user/maria_dev/movielens/u.data'
8 INTO TABLE voto;
```

Below the code, there are buttons for "Execute", "Save As", "Insert UDF", and "Visual Explain". The "RESULTS" tab is selected.

Contenido de película:

```
hive> SELECT * FROM pelicula LIMIT 10;
OK
1      Toy Story      1995
2      GoldenEye      1995
3      Four Rooms     1995
4      Get Shorty     1995
5      Copycat        1995
6      Shanghai Triad (Yao a yao yao dao waipo qiao)   1995
7      Twelve Monkeys  1995
8      Babe            1995
9      Dead Man Walking      1995
10     Richard III     1995
Time taken: 1.266 seconds, Fetched: 10 row(s)
```

Contenido de usuario:

```
hive> SELECT * FROM usuario LIMIT 10;
OK
1      24      M      technician
2      53      F      other
3      23      M      writer
4      24      M      technician
5      33      F      other
6      42      M      executive
7      57      M      administrator
8      36      M      administrator
9      29      M      student
10     53      M      lawyer
Time taken: 0.172 seconds, Fetched: 10 row(s)
```

Contenido de voto:

```
hive> SELECT * FROM voto LIMIT 10;
OK
196    242    3    881250949
186    302    3    891717742
22     377    1    878887116
244    51     2    880606923
166    346    1    886397596
298    474    4    884182806
115    265    2    881171488
253    465    5    891628467
305    451    3    886324817
6      86     3    883603013
```

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

```
1 SELECT u.ocupacion, COUNT(u.id) as total FROM usuario AS u
2 GROUP BY u.ocupacion
3   ORDER BY total DESC
4 LIMIT 10;
```

u.ocupacion	total
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

```
SELECT u.genero, COUNT(u.id) as total FROM usuario AS u
GROUP BY u.genero
ORDER BY total DESC;
```

u.genero	total
M	670
F	273

```
hive> SELECT u.genero, COUNT(u.id) as total FROM usuario AS u GROUP BY u.genero ORDER BY total DESC;
Query ID = maria_dev_20251202103629_2926c427-b884-4153-9cab-6bd264124e25
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1764663236812_0016)

-----
      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: 4.81 s
-----
OK
M      670
F      273
Time taken: 8.803 seconds, Fetched: 2 row(s)
```

- 4) Muestra la edad media por géneros.

```
SELECT u.genero, avg(u.edad) as total FROM usuario AS u
GROUP BY u.genero;
```

u.genero	total
F	33.81318681318681
M	34.149253731343286

```
hive> SELECT u.genero, avg(u.edad) as total FROM usuario AS u GROUP BY u.genero;
Query ID = maria_dev_20251202103837_3f31ad0c-423b-42c3-815f-3ac1d729db5e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1764663236812_0016)
```

```
-----  
          VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  
-----  
Map 1 ..... SUCCEEDED      1        1        0        0        0        0        0  
Reducer 2 ..... SUCCEEDED      1        1        0        0        0        0        0  
-----  
VERTICES: 02/02  [=====>>>] 100%  ELAPSED TIME: 3.17 s  
-----  
OK  
F      33.81318681318681  
M      34.149253731343286  
Time taken: 4.296 seconds, Fetched: 2 row(s)
```

- 5) Muestra la edad media por ocupaciones.

```
SELECT u.ocupacion, avg(u.edad) as total FROM usuario AS u GROUP BY u.ocupacion;
```

u.ocupacion	total
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	36.311111111111111

## PRÁCTICA CON HIVE

```
hive> SELECT u.ocupacion, avg(u.edad) as total FROM usuario AS u GROUP BY u.ocupacion;
Query ID = maria_dev_20251202103957_a872f1c3-e1fb-4a62-9fcf-e1b0ea561a4c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1764663236812_0016)

-----
      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: 3.69 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 36.311111111111111
Time taken: 5.151 seconds, Fetched: 21 row(s)
```

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

```
SELECT p.id, p.titulo, count(v.valoracion) AS votos FROM voto AS v
LEFT JOIN pelicula AS p ON v.pelicula_id = p.id
GROUP BY p.id, p.titulo
ORDER BY votos DESC
LIMIT 5;
```

## PRÁCTICA CON HIVE

p.id	p.titulo	votos
50	Star Wars	583
258	Contact	509
100	Fargo	508
181	Return of the Jedi	507
294	Liar Liar	485

```
hive> SELECT p.id, p.titulo, count(v.valoracion) AS votos FROM voto AS v LEFT JOIN pelicula AS p ON v.pelicula_id = p.id GROUP BY p.id, p.titulo ORDER BY votos DESC LIMIT 5;
Query ID = maria_dev_20251202104601_9c78b150-d48b-4625-98b5-d9b29796d498
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1764663236812_0016)

-----  

      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: 6.80 s  

-----  

OK  

50      Star Wars      583  

258     Contact       509  

100     Fargo          508  

181     Return of the Jedi   507  

294     Liar Liar      485  

Time taken: 13.37 seconds, Fetched: 5 row(s)
```