EJERCICIOS DE HIVE

NOMBRE ALUMNO: Tania Batista

Estructuras de datos

Indicar los comandos empleados para resolver las siguientes preguntas:

1. Crear un fichero de texto con la información anterior
(IMPORTANTE: al crear el fichero tener cuidado con los caracteres al final de línea)

Start the demons:
Start-dfs.sh
start-yarn.sh
jps
mr-jobhistory-daemon.sh start historyserver

open another terminal
mapred historyserver
nohup mapred historyserver &
jps

```
bigdata@bigdata:~$ jps
8816 JobHistoryServer
2338 ResourceManager
2484 NodeManager
2153 SecondaryNameNode
8969 Jps
1755 NameNode
1902 DataNode
bigdata@bigdata:~$
```

Create the discography.txt file & remove spaces after comma
nano /home/bigdata/ejemplos/Hive/pinkfloyd.txt

```
1967,The Piper at the Gates of Dawn,131,6
1968,A Saucerful of Secrets,999,9
1969,Music from the Film More,153,9
1969,Ummagumma,74,5
1970,Atom Heart Mother,55,1
1972,Obscured by Clouds,46,6
1973,The Dark Side of the Moon,1,1
1975,Wish you Were Here,1,1
1977,Animals,3,2
1979,The Wall,1,3
1983,The Final Cut,6,1
1987,A Momentary Lapse of Reason,3,3
1994,The Division Bell,1,1
2014,The Endless River,3,1
```

2. Acceder a Hive y crear una base de datos llamada ejercicios (1 punto)

cd $HIVE_HOME

hive

hive> CREATE DATABASE IF NOT EXISTS ejercicios;

hive> SHOW DATABASES;

```
bigdata@bigdata:~$ cd $HIVE_HOME
bigdata@bigdata:~/hive$ hive

Logging initialized using configuration in jar:file:/home/bigdata/hive/lib/hive-common-2.3.0.jar
!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider usi
ng a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> CREATE DATABASE IF NOT EXISTS ejercicios;
OK
Time taken: 9.39 seconds
hive> SHOW DATABASES;
OK
default
ejercicios
Time taken: 0.251 seconds, Fetched: 2 row(s)
hive> bigdata@bigdata:~/hive$
```

3. Usar la base de datos anterior (1 punto)

hive

hive> USE ejercicios;

```
hive> USE ejercicios;
OK
Time taken: 8.691 seconds
hive>
```

4. Crear una tabla en Hive en la base de datos anterior que permita almacenar los datos anteriores indicando que el formato de separación es como el siguiente de tipo tabulación (
create table …… (……) (1 punto)

```
CREATE TABLE discography_tab (
        year INT,
        name STRING,
        eu_ranking INT,
        uk_ranking INT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;
```

```
hive> CREATE TABLE discography_tab (
    > year INT,
    > name STRING,
    > eu_ranking INT,
    > uk_ranking INT)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > STORED AS TEXTFILE;
OK
```

5. Cargar el fichero de texto (1 punto

LOAD DATA LOCAL INPATH '/home/bigdata/ejemplos/Hive/pinkfloyd.txt' INTO TABLE discography_tab;

```
hive> LOAD DATA LOCAL INPATH '/home/bigdata/ejemplos/Hive/pinkfloyd.txt' INTO TABLE discography_tab;
```

6. Acceder a Hive y ejecutar un consulta sencilla (select *) para verificar que hay datos y se han cargado correctamente. En caso contrario, volver a cargar los datos (1 punto)

hive> SELECT * FROM discography_tab;

```
hive> SELECT * FROM discography_tab;
OK
1967    The Piper at the Gates of Dawn  131     6
1968    A Saucerful of Secrets  999     9
1969    Music from the Film More        153     9
1969    Ummagumma       74      5
1970    Atom Heart Mother       55      1
1972    Obscured by Clouds      46      6
1973    The Dark Side of the Moon       1       1
1975    Wish you Were Here      1       1
1977    Animals 3       2
1979    The Wall        1       3
1983    The Final Cut   6       1
1987    A Momentary Lapse of Reason     3       3
1994    The Division Bell       1       1
2014    The Endless River       3       1
Time taken: 2.401 seconds, Fetched: 14 row(s)
hive>
```

I realized too late that EEUU means US, not EU! So here I change the column name:

ALTER TABLE discography_tab CHANGE eu_ranking us_ranking INT;

```
hive> ALTER TABLE discography_tab CHANGE eu_ranking us_ranking INT;
OK
Time taken: 0.971 seconds
hive> SELECT us_ranking FROM discography_tab;
OK
131
```

7. Calcular los discos que estuvieron a la vez entre los 5 primeros lugares en EEUU y UK (1 punto)

hive> SELECT name FROM discography_tab WHERE us_ranking < 6 AND uk_ranking < 6 ORDER BY us_ranking DESC;

```
hive> SELECT name FROM discography_tab WHERE us_ranking  < 6 AND uk_ranking < 6 ORDER BY name DESC;
Wish you Were Here
The Wall
The Endless River
The Division Bell
The Dark Side of the Moon
Animals
A Momentary Lapse of Reason
Time taken: 39.37 seconds, Fetched: 7 row(s)
```

8. (OPCIONAL) Obtener la máxima y mínima posición que ocuparon los discos de Pink Floyd en EEUU y en UK (por ejemplo empleando el comando order y limit en dos sentencias)

9. (OPCIONAL) Repetir todos los ejercicios empleando una tabla con estructuras de datos complejas