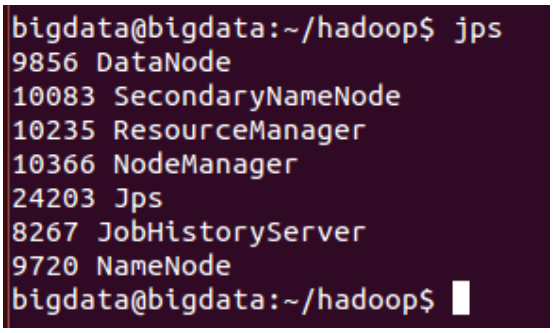
EJERCICIOS DE HIVE

En primer lugar compruebo si están arrancados los demonios:



Veo que están arrancados. Si no estuvieran tendría que arrancarlos con los siguientes comandos:

./sbin/start-dfs.sh

./sbin/start-yarn.sh

./sbin/mr-jobhistory-daemon.sh start historyserver

Estructuras de datos

Partiendo de la discografía de Pink Floyd (año, nombre disco, ranking EEUU, ranking UK)

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

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)

Con el comando: sudo nano /home/bigdata/ejemplosHive/discografia.csv creo un fichero de discografia de PinkFloyd.

Cuando se abra el esdtor pego el contenido y para salir ctrl+x, s + intro.

Con cat compruebo que el fichero se ha creado.

```
bigdata@bigdata:~/ejemplosHive$ sudo nano /home/bigdata/ejemplosHive/discografia.csv
[sudo] password for bigdata:
bigdata@bigdata:~/ejemplosHive$ cat discografia.csv
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
bigdata@bigdata:~/ejemplosHive$
```

2. Acceder a Hive y crear una base de datos llamada ejercicios

Creo una variable de entorno

$HIVE_HOME=/home/bigdata/hive

Entro en la carpeta hive: cd $HIVE_HOME

Arranco la consola hive:

hive

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

Logging initialized using configuration in jar:file:/home/bigdata/hive/lib/hive-common-2.1.0.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez, spark) or using Hive 1.X releases.
hive>
```

Para crear la base de datos utilizo el siguiente comando:

CREATE DATABASE ejercicios COMMENT 'Discocrafias' WITH DBPROPERTIES ('creator' = 'Ani', 'date' = '2016-08-04');

```
hive> CREATE DATABASE ejercicios COMMENT 'Discocrafias' WITH DBPROPERTIES ('creator' = 'Ani', 'date' = '2016-08-04');
OK
Time taken: 1.129 seconds
hive>
```

3. Usar la base de datos anterior

USE ejercicios;

```
hive> USE ejercicios;
OK
Time taken: 0.039 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 …… (.....) row format delimited fields terminated by ',' stored as textfile;)

CREATE TABLE PinkFloyd(
Anio INT,
Titulo STRING,

Ranking_USA INT,

Ranking_UK INT

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\,';

```
hive> CREATE TABLE PinkFloyd(
    > Anio INT,
    > Titulo STRING,
    > Ranking_USA INT,
    > Ranking_UK INT
    > )
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\,';
OK
Time taken: 0.48 seconds
hive>
```

5. Cargar el fichero de texto

LOAD DATA LOCAL INPATH '/home/bigdata/ejemplosHive/discografia.csv' INTO TABLE PinkFloyd;

```
hive> LOAD DATA LOCAL INPATH '/home/bigdata/ejemplosHive/discografia.csv' INTO TABLE PinkFloyd;
Loading data to table ejercicios.pinkfloyd
OK
Time taken: 0.926 seconds
hive>
```

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

select * from PinkFloyd;

```
hive>
    > select * from PinkFloyd;
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     NULL    6
1973    The Dark Side of the Moon      NULL    1
1975    Wish you Were Here     NULL    1
1977    Animals        NULL    2
1979    The Wall       NULL    3
1983    The Final Cut  NULL    1
1987    A Momentary Lapse of Reason    3       3
1994    The Division Bell      NULL    1
2014    The Endless River      NULL    NULL
Time taken: 1.04 seconds, Fetched: 14 row(s)
hive>
```

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

SELECT * FROM PinkFloyd WHERE Ranking_USA <= 5 AND Ranking_UK <= 5;

```
hive> SELECT * FROM PinkFloyd WHERE Ranking_USA <= 5 AND Ranking_UK <= 5;
OK
1987    A Momentary Lapse of Reason    3       3
Time taken: 0.543 seconds, Fetched: 1 row(s)
hive>
```

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)

Mejor ranking en los Estados Unidos:

select min(Ranking_USA) from PinkFloyd;

```
    > select min(Ranking_USA) from PinkFloyd;
FAILED: ParseException line 3:0 missing EOF at 'select' near 'PinkFloyd'
hive> select min(Ranking_USA) from PinkFloyd;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider u
Query ID = bigdata_20160804210608_cfbe6065-40ac-421f-920f-4d772e303236
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1469204298658_0019, Tracking URL = http://bigdata:8088/proxy/application_1469204298
Kill Command = /home/bigdata/hadoop/bin/hadoop job  -kill job_1469204298658_0019
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2016-08-04 21:06:18,088 Stage-1 map = 0%,  reduce = 0%
2016-08-04 21:06:24,730 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.87 sec
2016-08-04 21:06:33,285 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.91 sec
MapReduce Total cumulative CPU time: 1 seconds 910 msec
Ended Job = job_1469204298658_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 1.91 sec   HDFS Read: 8577 HDFS Write: 101 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 910 msec
OK
3
Time taken: 26.28 seconds, Fetched: 1 row(s)
hive>
```

El mejor ranking en los EEUU era 3.

Peor rating en los Estados Unidos:

select max(Ranking_USA) from PinkFloyd;

```
hive> select max(Ranking_USA) from PinkFloyd;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a diffe
Query ID = bigdata_20160804211304_ef19e235-9192-423b-88c1-dd91d58da621
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1469204298658_0020, Tracking URL = http://bigdata:8088/proxy/application_1469204298658_0020/
Kill Command = /home/bigdata/hadoop/bin/hadoop job  -kill job_1469204298658_0020
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2016-08-04 21:13:13,015 Stage-1 map = 0%,  reduce = 0%
2016-08-04 21:13:19,361 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.86 sec
2016-08-04 21:13:26,863 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.91 sec
MapReduce Total cumulative CPU time: 1 seconds 910 msec
Ended Job = job_1469204298658_0020
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 1.91 sec   HDFS Read: 8585 HDFS Write: 103 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 910 msec
OK
999
Time taken: 23.23 seconds, Fetched: 1 row(s)
hive>
```

El peor ranking en EEUU era 999.

Mejor ranking en UK:

select min(Ranking_UK) from PinkFloyd;

```
hive> select min(Ranking_UK) from PinkFloyd;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a diffe
Query ID = bigdata_20160804211553_d24723ce-4133-47e8-b81c-1d9d81330f55
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1469204298658_0021, Tracking URL = http://bigdata:8088/proxy/application_1469204298658_0021/
Kill Command = /home/bigdata/hadoop/bin/hadoop job  -kill job_1469204298658_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2016-08-04 21:16:02,400 Stage-1 map = 0%,  reduce = 0%
2016-08-04 21:16:09,844 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.93 sec
2016-08-04 21:16:18,373 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.06 sec
MapReduce Total cumulative CPU time: 2 seconds 60 msec
Ended Job = job_1469204298658_0021
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.06 sec   HDFS Read: 8577 HDFS Write: 101 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 60 msec
OK
1
Time taken: 26.221 seconds, Fetched: 1 row(s)
hive>
```

El mejor rankig en UK era 1.

Peor ranking en UK:

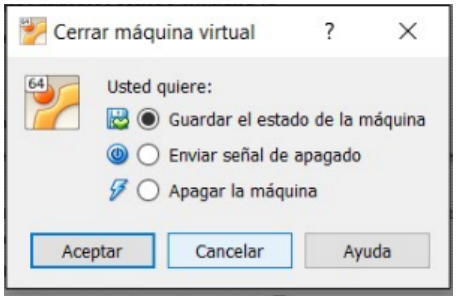
select max(Ranking_UK) from PinkFloyd;

```
hive> select max(Ranking_UK) from PinkFloyd;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a dif
Query ID = bigdata_20160804211919_a4127940-2b78-4387-8114-9dd6298cdcd3
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1469204298658_0022, Tracking URL = http://bigdata:8088/proxy/application_1469204298658_0022/
Kill Command = /home/bigdata/hadoop/bin/hadoop job  -kill job_1469204298658_0022
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2016-08-04 21:19:27,134 Stage-1 map = 0%,  reduce = 0%
2016-08-04 21:19:33,625 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.84 sec
2016-08-04 21:19:42,156 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.04 sec
MapReduce Total cumulative CPU time: 2 seconds 40 msec
Ended Job = job_1469204298658_0022
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.04 sec   HDFS Read: 8577 HDFS Write: 101 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 40 msec
OK
9
Time taken: 24.212 seconds, Fetched: 1 row(s)
hive>
```

El peor ranking en UK era 9.

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

Para finalizar cierro y guardo la máquina Archivo/cerrar y guardar el estado de la máquina.



Si no hago eso tendía que parar los demonios:

cd $HADOOP_HOME

./sbin/stop-dfs.sh
./sbin/stop-yarn.sh
./sbin/mr-jobhistory-daemon.sh stop historyserver