EJERCICIOS DE PIG

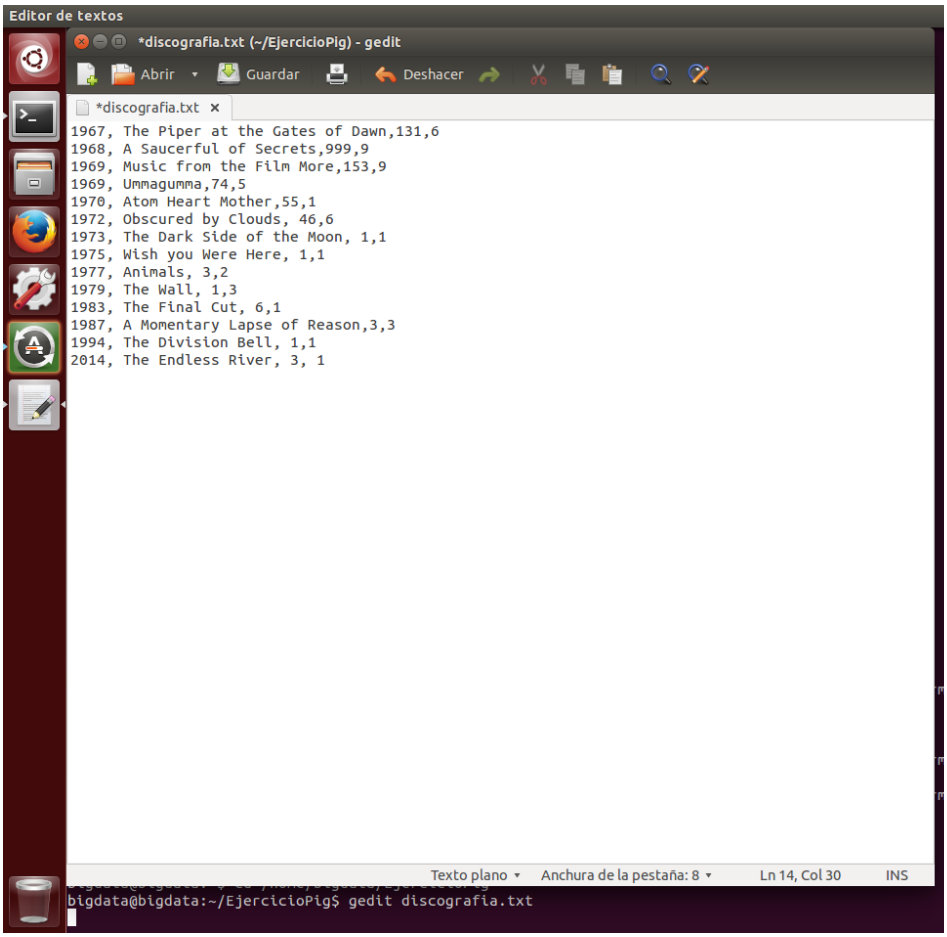
Partiendo de la discografía de Pink Floyd (año, nombre disco, ranking EEUU, ranking UK)

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

1. Crear un fichero llamado discos.txt

mkdir /home/bigdata/EjercicioPig
cd /home/bigdata/EjercicioPig

gedit discografia.txt



Compruebo el contenido del fichero discografia.txt

head discografia.txt

```
bigdata@bigdata:~/EjercicioPig$ head discografia.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
bigdata@bigdata:~/EjercicioPig$
```

2. Arrancar HDFS, Yarn y el job history

Con jps compruebo si tengo arrancados los demonios.

```
bigdata@bigdata:~/EjercicioPig$ jps
4722 Jps
9856 DataNode
10083 SecondaryNameNode
10235 ResourceManager
10366 NodeManager
8267 JobHistoryServer
9720 NameNode
bigdata@bigdata:~/EjercicioPig$
```

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

```
./sbin/start-dfs.sh
./sbin/start-yarn.sh
./sbin/mr-jobhistory-daemon.sh start historyserver
```

3. Subir el fichero a HDFS dentro de la carpeta /ejerciciosPig/discografia.txt

Primero creo una carpeta ejerciciosPig en HDFS:

hdfs dfs -mkdir /ejerciciosPig

Subo el fichero discografia.txt a la carpeta ejerciciosPig en HDFS:

hdfs dfs -put discografia.txt /ejerciciosPig/discografia.txt

Compruebo queel fichero se ha subido:

hdfs dfs -cat /ejerciciosPig/discografia.txt

```
bigdata@bigdata:~/EjercicioPig$ hdfs dfs -cat /ejerciciosPig/discografia.txt
16/08/03 23:51:44 WARN util.NativeCodeLoader: Unable to load native-hadoop li
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:~/EjercicioPig$
```

4. Ejecutar la instrucción ls sobre Hadoop para indicar el tamaño del fichero

hdfs dfs -ls /ejerciciosPig/discografia.txt

```
bigdata@bigdata:~/EjercicioPig$ hdfs dfs -ls /ejerciciosPig/discografia.txt
16/08/04 00:33:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platf
-rw-r--r--   1 bigdata supergroup        424 2016-08-03 23:46 /ejerciciosPig/discografia.txt
bigdata@bigdata:~/EjercicioPig$
```

5. Arrancar pig en modo distribuido (si se desea eliminar trazas de log) y ejecutar el siguiente comando:
cat hdfs://localhost:9000/ejerciciosPig/discografia.txt
para confirmar que los primeros puntos han funcionado correctamente y el fichero está subido a HDFS

Creo una variable de entorno: PIG_HOME=/home/bigdata/pig

Arranco PIG en modo distribuido:
pig -4 $PIG_HOME/conf/nolog.conf -x mapreduce

Ejecuto cat hdfs://localhost:9000/ejerciciosPig/discografia.txt para confirmar que el fichero está subido a HDFS.

```
bigdata@bigdata:~/EjercicioPig$ PIG_HOME=/home/bigdata/pig
bigdata@bigdata:~/EjercicioPig$ pig -4 $PIG_HOME/conf/nolog.conf -x mapreduce
16/08/04 00:43:36 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/08/04 00:43:36 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/08/04 00:43:36 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
16/08/04 00:43:36 INFO pig.Main: Loaded log4j properties from file: /home/bigdata/pig/conf/nolog.conf
grunt> REGISTER '/home/bigdata/ejemplosMapReduce/wc.jar';
grunt> DEFINE MY_UDF com.myudfs.UPPER();
grunt> SET default_parallel 10;
grunt>
grunt> cat hdfs://localhost:9000/ejerciciosPig/discografia.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
grunt>
```

6. Cargar el fichero de hdfs en una variable llamada discos

discos = LOAD 'hdfs://localhost:9000/ejerciciosPig/discografia.txt' using PigStorage (',') AS (annio: int, nombredisco: chararray, rankingEEUU: int, rankingUK: int);
dump discos;

```
grunt> discos = LOAD 'hdfs://localhost:9000/ejerciciosPig/discografia.txt' using PigStorage (',') AS (annio: int, nombredisco: chararray, rankingEEUU: int, rankingUK: int);
grunt> dump discos;
(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)
grunt>
```

7. Calcular los discos que estuvieron a la vez en el top 5 de EEUU y de UK (indicar también el resultado)

top5 = filter discos by rankingEEUU <= 5 and rankingUK <= 5;
dump top5;

```
grunt> top5 = filter discos by rankingEEUU <= 5 and rankingUK <= 5;
grunt> dump top5;
(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)
(1987, A Momentary Lapse of Reason,3,3)
(1994, The Division Bell,1,1)
(2014, The Endless River,3,1)
grunt>
```

8. Obtener la máxima y mínima posición que ocuparon los discos de Pink Floyd en EEUU y en UK (indicar también el resultado) empleando los comandos de LATIN PIG

Máxima y mínima posiciòn en los Estados Unidos:
USAmaxmin = foreach (group discos all) generate MAX(discos.rankingEEUU) as maxEEUU, MIN(discos.rankingEEUU) as minEEUU;
dump USAmaxmin;

```
grunt> USAmaxmin = foreach (group discos all) generate MAX(discos.rankingEEUU) as maxEEUU, MIN(discos.rankingEEUU) as minEEUU;
grunt> dump USAmaxmin;
(999,1)
```

Máxima y mínima posiciòn en UK:
UKmaxmin = foreach (group discos all) generate MAX(discos.rankingUK) as maxUK, MIN(discos.rankingUK) as minUK;
dump Ukmaxmin;

```
grunt> UKmaxmin = foreach (group discos all) generate MAX(discos.rankingUK) as maxUK, MIN(discos.rankingUK) as minUK;
grunt> dump UKmaxmin;
(9,1)
```

9. Explica con tus propias palabras lo que se desea obtener con los siguientes comandos e indica el resultado obtenido.

a = foreach discos generate anio;

b = distinct a;

dump b;

Se define como a una lista se los años por cada registro de disco.
Se cuarda en la variable b una selección de a, en la que solo aparecen elementos distintos, es decir, se elimina años que se repiten.
Finalmente se muestra los años que se han guardado en la variable b: distintos años de los elementos guardados en variable discos.

10. (opcional) Empleando UDFs extrae información útil de la discografía.