EJERCICIOS DE PIG

NOMBRE ALUMNO: Tania Batista

1. Crear un fichero llamado discos.txt (1 punto)

mkdir -p /home/bigdata/ejemplos/Pig/hdfs/upload/

```
bigdata@bigdata:~/ejemplos/Pig$ mkdir -p /home/bigdata/ejemplos/Pig/hdfs/upload/
```

nano /home/bigdata/ejemplos/Pig/hdfs/upload/discos.txt

```
bigdata@bigdata:~/ejemplos/Pig$ nano /home/bigdata/ejemplos/Pig/hdfs/upload/discos.txt
  GNU nano 2.8.6     Archivo: /home/bigdata/ejemplos/Pig/hdfs/upload/discos.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. Arrancar HDFS, Yarn y el job history (1 punto)

start-dfs.sh
start-yarn.sh
mr-jobhistory-daemon.sh start historyserver
mapred historyserver
nohup mapred historyserver &
jps

```
bigdata@bigdata:~/ejemplos/Pig$ jps
2498 NodeManager
2355 ResourceManager
7571 Jps
7397 JobHistoryServer
1895 DataNode
1755 NameNode
2094 SecondaryNameNode
```

3. Subir el fichero a HDFS dentro de la carpeta /ejerciciosPig/discografia.txt (1 punto)

nano /home/bigdata/ejemplos/Pig/discografia.txt
hdfs dfs -put /home/bigdata/ejemplos/Pig/discografia.txt /exercisesPig/

```
bigdata@bigdata:~$ hdfs dfs -put /home/bigdata/ejemplos/Pig/discografia.txt /exercisesPig/
```

I didn't really understand this question.
Were we supposed to take the previous "discos.txt" file, and have it uploaded to /ejerciciosPig/ with the new name of "discografia.txt"? In that case, managed it this way, but then I couldn't get it to run on the localhost link below.

Create the new HDFS folder
hdfs dfs -mkdir /user/bigdata/testingPig

```
bigdata@bigdata:~$ hdfs dfs -mkdir  /user/bigdata/testingPig
```

Load discos.txt to HDFS
hdfs dfs -put /home/bigdata/ejemplos/Pig/hdfs/upload/discos.txt /user/bigdata/testingPig

```
bigdata@bigdata:~$ hdfs dfs -put /home/bigdata/ejemplos/Pig/hdfs/upload/discos.txt  /user/bigdata/testingPig
```

Create a new folder, where discos.txt will be transferred and renamed
hdfs dfs -get /user/bigdata/testingPig/discos.txt ejemplos/Hadoop/hdfs/downloadTesting/discografia.txt
mkdir /home/bigdata/ejemplos/Hadoop/hdfs/downloadTesting

```
bigdata@bigdata:~$ mkdir /home/bigdata/ejemplos/Hadoop/hdfs/downloadTesting/
bigdata@bigdata:~$ hdfs dfs -get /user/bigdata/testingPig/discos.txt ejemplos/Hadoop/hdfs/downloadTesting/discografia.txt
```

Here is discografia.txt with the same content… except no access here from the Grunt Shell!

```
bigdata@bigdata:~$ cat ejemplos/Hadoop/hdfs/downloadTesting/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
bigdata@bigdata:~$
```

4. Ejecutar la instrucción ls sobre Hadoop para indicar el tamaño del fichero (1 punto)

hdfs dfs -ls /exercisesPig/
hdfs dfs -du /exercisesPig/

```
bigdata@bigdata:~$ hdfs dfs -ls /exercisesPig/

Found 1 items
-rw-r--r--   1 bigdata supergroup        401 2018-02-25 13:41 /exercisesPig/discografia.txt
bigdata@bigdata:~$
bigdata@bigdata:~$ hdfs dfs -du /exercisesPig/
401  /exercisesPig/discografia.txt
```

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 (1 punto)

pig -x mapreduce

```
bigdata@bigdata:~/ejemplos/Pig$ pig -x mapreduce
```

cat hdfs://localhost:9000/exercisesPig/discografia.txt

```
grunt> cat hdfs://localhost:9000/exercisesPig/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 (1 punto)

discos = load 'hdfs://localhost:9000/exercisesPig/discografia.txt' using PigStorage(',') as (year: int, disk: chararray, usa: int, uk: int);

```
grunt> discos = load 'hdfs://localhost:9000/exercisesPig/discografia.txt' using PigStorage(',') as (year: int, disk: chararray, usa: int, uk: int);
```

7. Calcular los discos que estuvieron a la vez en el top 5 de EEUU y de UK (indicar también el resultado) (1 punto)

top5 = filter discos by (usa > 6) and (uk > 6)

```
grunt> top5 = filter discos by (usa < 6) and (uk < 6);
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)
```

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 (1 punto)

discos_grouped = GROUP discos ALL;
topUSAnumber = FOREACH discos_grouped GENERATE MIN(discos.usa) AS USA_hits;
top_USA_discs = FILTER discos BY usa == (int)topUSAnumber.USA_hits;
dump top_USA_discs;

```
grunt> discos_grouped = GROUP discos ALL;
grunt> topUSAnumber = FOREACH discos_grouped GENERATE MIN(discos.usa) AS USA_hits;
grunt> top_USA_discs = FILTER discos BY usa == (int)topUSAnumber.USA_hits;
grunt> dump top_USA_discs;
```

```
(1973,The Dark Side of the Moon,1,1)
(1975,Wish you Were Here,1,1)
(1979,The Wall,1,3)
(1994,The Division Bell,1,1)
```

discos_grouped = GROUP discos ALL;
topUKnumber = FOREACH discos_grouped GENERATE MIN(discos.uk) AS UK_hits;
top_uk_discs = FILTER discos BY uk == (int)topUKnumber.UK_hits;
dump top_uk_discs;

```
grunt> discos_grouped = GROUP discos ALL;
grunt> topUKnumber = FOREACH discos_grouped GENERATE MIN(discos.uk) AS UK_hits;
grunt> top_uk_discs = FILTER discos BY uk == (int)topUKnumber.UK_hits;
grunt> dump top_uk_discs;
(1970,Atom Heart Mother,55,1)
(1973,The Dark Side of the Moon,1,1)
(1975,Wish you Were Here,1,1)
(1983,The Final Cut,6,1)
(1994,The Division Bell,1,1)
(2014,The Endless River,3,1)
```

9. Explica con tus propias palabras lo que se desea obtener con los siguientes comandos e indica el resultado obtenido (2 puntos)

a = foreach discos generate year;
b = distinct a;
dump b;

Create a variable that contains only the years associated with each disc. Then eliminate any duplicates! So you only get 1 disc per year. Then print all the years that Pink Floyd had new music available.