Instalación de Pig

wget http://www-us.apache.org/dist/pig/pig-0.16.0/pig-0.16.0.tar.gz

tar -xzf pig-0.16.0.tar.gz

mkdir /opt/hadoop/pig

[hadoop@nodo1 Descargas]$ mv pig-0.16.0/* /opt/hadoop/pig/

Editamos: ".bashrc" de directorio del usuario /home/hadoop

gedit /home/hadoop/.bashrc

Copiamos y pegamos:

 # Set PIG_HOME

export PIG_HOME=/opt/hadoop/pig

export PATH=$PATH:/opt/hadoop/pig/bin

export PIG_CLASSPATH=$HADOOP_CONF_DIR

Ejecutamos: source /home/hadoop/.bashrc

pig -version

Primer programa en PIG

Crear un fichero llamado empleados.txt con lo campos:

Nombre, apellidos, número de teléfono, ciudad y profesión, separados los campos por tabulador

Copiar el fichero a HDFS

hadoop fs –copyFromLocal /home/hadoop/empleados.txt /input/empleados

o

hdfs dfs –copyFromLocal /home/hadoop/empleados.txt /input/empleados

Escribir el programa PIG

gedit /home/hadoop/output.pig si disponemos de interface gráfica en Linux.

En caso contrario, es decir, tan solo disponemos de acceso a la línea de comando de linux se debería utilizar

nano /home/hadoop/output.pig

vi /home/hadoop/output.pig

con el siguiente código

```
A = LOAD '/home/hadoop/empleados.txt' using PigStorage ('\t') as (FName:
chararray, LName: chararray, MobileNo: chararray, City: chararray,
Profession: chararray);

B = FOREACH A generate FName, MobileNo, Profession;

DUMP B;
```

Ejecutar

pig /home/hadoop/output.pig

Acceso a HDFS

$ pig

```
grunt> cd hdfs:///
grunt> ls
```

Copiar ficheros desde local a HDFS

```
grunt> mkdir test
grunt> cd test
grunt> copyFromLocal /etc/passwd passwd
grunt> ls
```

Leer archivo y realizar dump

```
grunt> passwd = LOAD '/etc/passwd' USING PigStorage(':') AS (user:chararray,
\
passwd:chararray, uid:int, gid:int, userinfo:chararray, home:chararray, \
shell:chararray);
grunt> DUMP passwd;
(root,x,0,0,root,/root,/bin/bash)
```

Tratar con un dataset con la herramienta PIG

Dataset con atletas y medallas

¿Qué país ha ganado más medallas?

```
atletas = LOAD 'atletas.csv' USING

org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE
','SKIP_INPUT_HEADER') as

(atleta:chararray,pais:chararray,deporte:chararray,oro:int,plata:int,bronce:int,total:
int);

DUMP atletas;

atletas_limit = LIMIT atletas 10;
DUMP atletas_limit;

atletas_group_pais = GROUP atletas BY pais;

DESCRIBE atletas_group_pais;
```

```
-- Ejecución paso a paso
ILLUSTRATE atletas_group_pais;


-- foreach
atletas = LOAD 'atletas.csv' USING

org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE
','SKIP_INPUT_HEADER') as

(atleta:chararray,pais:chararray,deporte:chararray,oro:int,plata:int,bronce:int,total:
int);
atletas_group_pais = GROUP atletas BY pais;
suma_medallas=FOREACH atletas_group_pais GENERATE group AS
pais,SUM(atletas.total) as cuenta_medallas;
DUMP suma_medallas



-- ORDER
atletas = LOAD 'atletas.csv' USING

org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE
','SKIP_INPUT_HEADER') as

(atleta:chararray,pais:chararray,deporte:chararray,oro:int,plata:int,bronce:int,total:
int);
atletas_group_pais = GROUP atletas BY pais;
suma_medallas=FOREACH atletas_group_pais GENERATE group AS
pais,SUM(atletas.total) as cuenta_medallas;
ordena_medallas = ORDER suma_medallas BY cuenta_medallas DESC;
ordena_medallas_primera = LIMIT ordena_medallas 1;
DUMP ordena_medallas_primera;
```

-- País con más medallas excluyendo swimming

```
atletas = LOAD 'atletas.csv' USING

org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE
','SKIP_INPUT_HEADER') as

(atleta:chararray,pais:chararray,deporte:chararray,oro:int,plata:int,bronce:int,total:
int);
atletas_filtrado = FILTER atletas by deporte!='Swimming';
suma_medallas = FOREACH (GROUP atletas_filtrado BY pais) GENERATE gruop as
pais, SUM(atletas_filtrado.total) as cuenta_medallas;
ordena_medallas = ODER suma_medallas BY cuenta_medallas DESC;
ordena_medallas_limit = LIMIT ordena_medallas 1;
```

DUMP ordena_medallas_limit;

-- Distintos países

```
atletas = LOAD 'atletas.csv' USING

org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE
','SKIP_INPUT_HEADER') as

(atleta:chararray,pais:chararray,deporte:chararray,oro:int,plata:int,bronce:int,total:
int);
distintos_paises = DISTINCT (FOREACH atletas GENERATE pais);
DUMP distintos_paises;
```