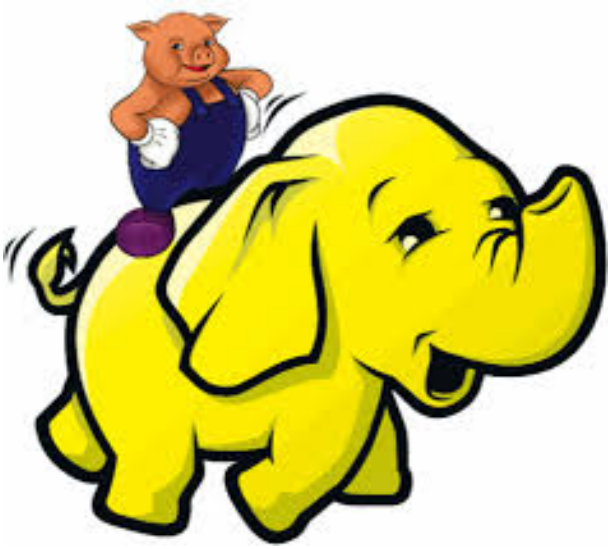
PRÁCTICA Nº 2.7
Pig II



- Nombre y apellidos: Alvaro Lucio-Villegas de Cea

Índice

Ejemplo

- Descarga el archivo Eopinions.csv y proceso_opiniones.pig.
- Ejecuta el script pig sobre el conjunto de datos.

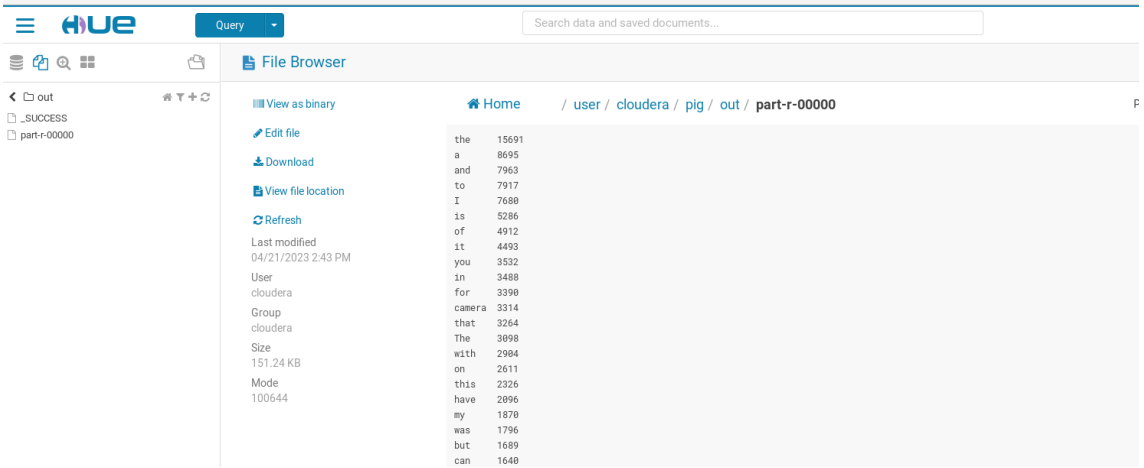
```
-- 1.1 Fase de Extracción LOAD
csv_data = LOAD '/user/cloudera/pig/Eopinions.csv' USING PigStorage (',') as
(class:chararray, opinion:chararray);

-- DUMP csv_data; -- para comprobaciones

-- 1.2 Generar colección de comentarios (campo 1)
comentarios = FOREACH csv_data GENERATE $1;

-- DUMP csv_data; -- para comprobaciones

-- 2.1 Procesar cada comentario y trozearlo en palabras
-- TOKENIZE: cadena -> bolsa de palabras (bag of words)
-- FLATTEN: desanida/aplana tuplas bosas de palabras
wordfile_flat = FOREACH comentarios GENERATE FLATTEN (TOKENIZE($0)) as
wordin;

-- 2.2 Agrupación por palabras (GROUP BY)
wordfile_grpd = GROUP wordfile_flat by wordin;

-- 2.3 Calculo frecuencia de cada palabra
word_counts = FOREACH wordfile_grpd GENERATE group,
COUNT(wordfile_flat.wordin) as cnt;

-- 2.4 Ordenación/Ranking de palabras por frecuencia
word_count_des = ORDER word_counts BY cnt DESC;

-- 3. Carga/almacenamiento
STORE word_count_des into '/user/cloudera/pig/out';
```

Ejecución del script.

```
[cloudera@quickstart ~]$ pig -f proceso_opiniones.pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
2023-04-21 07:41:59,961 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
12.0-cdh5.13.0 (rexported) compiled Oct 04 2017, 11:09:03
2023-04-21 07:41:59,961 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/cloudera/pig_1682088119951.log
2023-04-21 07:42:00,658 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/cloudera/.pigbootup not found
2023-04-21 07:42:00,712 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2023-04-21 07:42:00,712 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-04-21 07:42:00,712 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.clo
```

Salida del comando

Contenido del fichero en HDFS.



Ejercicios

- Descarga el archivo books.csv

5 escritores más valorados.

- Almacena en un archivo el ranking con los 5 escritores mejor valorados

```
-- Carga el archivo de datos en una relación llamada 'books'
books = LOAD '/user/cloudera/pig/books.csv' USING PigStorage(';')
        AS (bookID:int, title:chararray, authors:chararray,
        average_rating:float, isbn:chararray, isbn13:chararray,
        language_code:chararray, num_pages:int, ratings_count:int,
        text_reviews_count:int, publication_date:chararray,
publisher:chararray);

-- Agrupa los autores y calcula el promedio de sus valoraciones,
-- luego ordena los resultados de manera descendente según la valoración
promedio
author_ratings = GROUP books BY authors;
top_authors = FOREACH author_ratings GENERATE group AS
author,AVG(books.average_rating) AS avg_rating;

sorted_authors = ORDER top_authors BY avg_rating DESC;

-- Almacena los 5 mejores autores en un archivo
top_five = LIMIT sorted_authors 5;
STORE top_five INTO '/user/cloudera/pig/TOPautores' USING PigStorage(',');
```

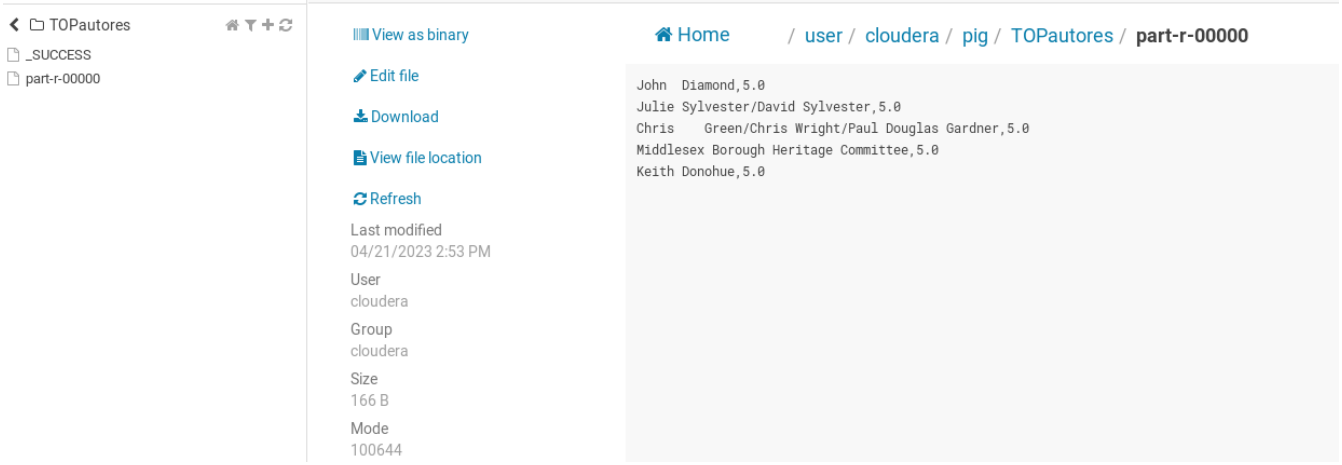
*Unsaved Document 1 | Ejer1.pig | proceso_opiniones.pig | books.csv | rankingAutores.pig

```
-- Carga el archivo de datos en una relación llamada 'books'
books = LOAD '/user/cloudera/pig/books.csv' USING PigStorage(';')
        AS (bookID:int, title:chararray, authors:chararray,
            average_rating:float, isbn:chararray, isbn13:chararray,
            language_code:chararray, num_pages:int, ratings_count:int,
            text_reviews_count:int, publication_date:chararray, publisher:chararray);

-- Agrupa los autores y calcula el promedio de sus valoraciones,
-- luego ordena los resultados de manera descendente según la valoración promedio
author_ratings = GROUP books BY authors;
top_authors = FOREACH author_ratings GENERATE group AS author,AVG(books.average_rating) AS avg_rating;

sorted_authors = ORDER top_authors BY avg_rating DESC;

-- Almacena los 5 mejores autores en un archivo
top_five = LIMIT sorted_authors 5;
STORE top_five INTO '/user/cloudera/pig/TOPautores' USING PigStorage(',');
```

Salida del script y contenido del fichero generado.

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.13.0 0.12.0-cdh5.13.0        cloudera        2023-04-21 08:21:05     2023-04-21 08:22:31     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1681898044415_0018  1       1       3       3       3       2       2       2       2       author_ratings,books,top_authors        GROUP_BY,COMBINER
job_1681898044415_0019  1       1       2       2       2       2       2       2       2       sorted_authors  SAMPLER
job_1681898044415_0020  1       1       3       3       3       2       2       2       2       sorted_authors  ORDER_BY,COMBINER
job_1681898044415_0021  1       1       2       2       2       2       2       2       2       sorted_authors          /user/cloudera/pig/TOPautores,

Input(s):
Successfully read 11128 records (1551118 bytes) from: "/user/cloudera/pig/books.csv"

Output(s):
Successfully stored 5 records (166 bytes) in: "/user/cloudera/pig/TOPautores"

Counters:
Total records written : 5
Total bytes written : 166
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
```

Contenido del fichero en HDFS.



10 escritores con más páginas.

- Almacena en un archivo el ranking con los 10 escritores con más páginas escritas.

```
-- Carga el archivo de datos en una relación llamada 'books'
books = LOAD '/user/cloudera/pig/books.csv' USING PigStorage(';')
        AS (bookID:int, title:chararray, authors:chararray,
        average_rating:float, isbn:chararray, isbn13:chararray,
        language_code:chararray, num_pages:int, ratings_count:int,
        text_reviews_count:int, publication_date:chararray, publisher:chararray);

-- Agrupa los autores y suma el número de páginas de todos sus libros
author_pages = GROUP books BY authors;
total_pages = FOREACH author_pages GENERATE group AS
author,SUM(books.num_pages) AS total_pages;

-- Ordena los resultados de manera descendente según el número de páginas
sorted_authors = ORDER total_pages BY total_pages DESC;

-- Almacena los 10 autores con más páginas escritas en un archivo
top_ten = LIMIT sorted_authors 10;
STORE top_ten INTO '/user/cloudera/pig/TOPPaginas' USING PigStorage(',');
```

Salida del comando

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.13.0 0.12.0-cdh5.13.0        cloudera        2023-04-21 08:24:32     2023-04-21 08:26:09     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1681898044415_0022  1       1       3       3       3       3       2       2       2       2       author_pages,books,total_pages  GROUP_BY,COMBINER
job_1681898044415_0023  1       1       2       2       2       2       2       2       2       2       sorted_authors  SAMPLER
job_1681898044415_0024  1       1       2       2       2       2       3       3       3       3       sorted_authors  ORDER_BY,COMBINER
job_1681898044415_0025  1       1       2       2       2       2       2       2       2       2       sorted_authors          /user/cloudera/pig/TOPPaginas,

Input(s):
Successfully read 11128 records (1551118 bytes) from: "/user/cloudera/pig/books.csv"

Output(s):
Successfully stored 10 records (211 bytes) in: "/user/cloudera/pig/TOPPaginas"

Counters:
Total records written : 10
Total bytes written : 211
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1681898044415_0022  ->      job_1681898044415_0023,
job_1681898044415_0023  ->      job_1681898044415_0024,
job_1681898044415_0024  ->      job_1681898044415_0025,
job_1681898044415_0025
```

Contenido del fichero en HDFS.