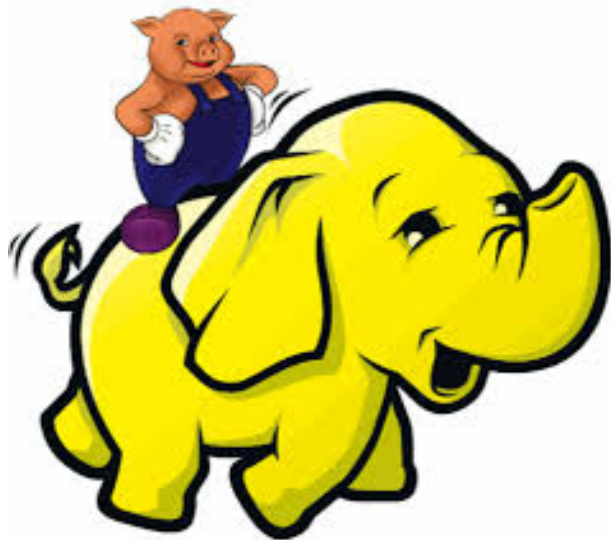


# PRÁCTICA Nº 2.7

## Pig II



- Nombre y apellidos: Alvaro Lucio-Villegas de Cea



## Índice

---

Ejemplo	3
Ejercicios	5
5 escritores más valorados.	5
10 escritores con más páginas.	7

## 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 (rexpoted) 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 svstem at: hdfs://quickstart.clo
```

## Salida del comando

```
at org.apache.hadoop.mapred.JobClient.getMapTaskReports(JobClient.java:667)
at org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher.launchPig(MapReduceLauncher.java:468)
at org.apache.pig.PigServer.launchPlan(PigServer.java:1334)
at org.apache.pig.PigServer.executeCompiledLogicalPlan(PigServer.java:1319)
at org.apache.pig.PigServer.execute(PigServer.java:1309)
at org.apache.pig.PigServer.executeBatch(PigServer.java:387)
at org.apache.pig.PigServer.executeBatch(PigServer.java:365)
at org.apache.pig.tools.grunt.GruntParser.executeBatch(GruntParser.java:148)
at org.apache.pig.tools.grunt.GruntParser.parseStopOnError(GruntParser.java:282)
at org.apache.pig.tools.grunt.GruntParser.parseStopOnError(GruntParser.java:173)
at org.apache.pig.tools.grunt.Grunt.exec(Grunt.java:84)
at org.apache.pig.Main.run(Main.java:484)
at org.apache.pig.Main.main(Main.java:158)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:686)
at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
2023-04-21 07:43:27,830 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2023-04-21 07:43:27,270 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2023-04-21 07:43:27,371 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
[cloudera@quickstart ~]$
```

## Contenido del fichero en HDFS.

File Browser

Search data and saved documents...

Home / user / cloudera / pig / out / part-r-00000

View as binary

Edit file

Download

View file location

Refresh

Last modified: 04/21/2023 2:43 PM

User: cloudera

Group: cloudera

Size: 151.24 KB

Mode: 100644

the	15691
a	8695
and	7963
to	7917
I	7688
is	5286
of	4912
it	4493
you	3532
in	3488
for	3390
camera	3314
that	3264
The	3098
with	2984
on	2611
this	2326
have	2096
my	1870
was	1796
but	1689
can	1648

## 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 x Ejer1.pig x proceso\_opiniones.pig x books.csv x rankingAutores.pig x

```
-- 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  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  2  sorted_authors  SAMPLER
job_1681898044415_0020  1  1  3  3  3  3  2  2  2  2  sorted_authors  ORDER_BY,COMBINER
job_1681898044415_0021  1  1  2  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.

TOPautores

\_SUCCESS

part-r-00000

View as binary

Edit file

Download

View file location

Refresh

Last modified  
04/21/2023 2:53 PM

User  
cloudera

Group  
cloudera

Size  
166 B

Mode  
100644

Home

/ user / cloudera / pig / TOPautores / part-r-00000

John Diamond,5.0

Julie Sylvester/David Sylvester,5.0

Chris Green/Chris Wright/Paul Douglas Gardner,5.0

Middlesex Borough Heritage Committee,5.0

Keith Donohue,5.0

## 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(';');
```

```
-- 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.

TOPPaginas

\_SUCCESS

part-r-00000

View as binary

Edit file

Download

View file location

Refresh

Last modified

04/21/2023 3:13 PM

User

cloudera

Group

cloudera

Size

Home

/ user / cloudera / pig / TOPPaginas / part-r-00000

Stephen King,18219

Orson Scott Card,14066

J.R.R. Tolkien,12537

Dan Simmons,11700

P.G. Wodehouse,11619

Mercedes Lackey,11480

Piers Anthony,10883

Agatha Christie,10556

Laurell K. Hamilton,10219

Sandra Brown,10091