

# MapReduce

miércoles, 29 de abril de 2020 15:36

- Ejecutar y registrar en bitacora ejemplo de wordcount-local.py y en MapReduce con mrjob de wordcount-mr.py en versión LOCAL

- Local

```
Windows PowerShell
PS C:\Users\Usuario\Documents\XVII\TET\bigdata-master> cd .\02-mapreduce\
PS C:\Users\Usuario\Documents\XVII\TET\bigdata-master\02-mapreduce> python wordcount-local.py ../datasets/gutenberg-small/1/*.*txt | more
LINCOLN 108
LETTERS 9
By 225
Abraham 120
Lincoln 376
Published 3
By 4571
The 2460
Bibliophile 1
Society 7
NOTE 9
Letters 49
herein 14
are 2467
so 1917
thoroughly 15
Characteristic 18
of 23820
```

- Mrjob

```
[amartinezv@hdpjupyter 02-mapreduce]$ python wordcount-mr.py -r local ../datasets/gutenberg-small/1/*.*txt | more
No configs found; falling back on auto-configuration
No configs specified for local runner
Creating temp directory /tmp/wordcount-mr.amartinezv.20200410.025856.169324
Running step 1 of 1...
job output is in /tmp/wordcount-mr.amartinezv.20200410.025856.169324/output
streaming final output from /tmp/wordcount-mr.amartinezv.20200410.025856.169324/output...
"225", 1
"$1,019,446.", 2
"$1,298,056,101.89," 2
"$1,339,710.35," 2
"$1,394,196,007.62," 1
"$1,394,796,007.62," 1
"$1,485,103.61," 2
"$1,740,690,489.49," 2
"$1,795,331.73" 2
"$1,25," 2
"$1.50" 1
"$1.75" 1
"$10" 1
"$10,000" 1
"$100" 1
"$100,000" 2
"$100,000," 1
"$100,000,000" 2
"$100,000.00" 1
"$1000" 2
"$102,316,155.99," 2
"$102,322,509.27," 2
"$1025," 1
"$109,741,134.10," 2
"$11,125,364.13," 2
"$11,125,789.59," 2
"$11,318,206.84" 2
"$12,438,253.78" 1
```

- Ejecutar y registrar en bitacora ejemplo de wordcount-mr.py en HADOOP (datos de entrada y salida en HDFS en el DCA y en EMR)

Al crear las variables de ambiente y ejecutar el comando del Github se me creo la carpeta temp pero no la carpeta result

```
[amartinezv@hdpjupyter 02-mapreduce]$ python wordcount-mr.py hdfs://user/amartinezv/datasets/gutenberg-small/1/*.*txt -r hadoop --output-dir hdfs://user/amartinezv/result3 --hadoop-streaming-jar $HADOOP_STREAMING_HOME/hadoop-streaming.jar
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /usr/hdp/current/hadoop-client/bin...
Found hadoop binary: /usr/hdp/current/hadoop-client/bin/hadoop
Using Hadoop version 3.1.1-3.1.4.0
Creating temp directory /tmp/wordcount-mr.amartinezv.20200413.181723.896350
uploading working dir files to hdfs://user/amartinezv/tmp/mrjob/wordcount-mr.amartinezv.20200413.181723.896350/files/wd...
Copying other local files to hdfs://user/amartinezv/tmp/mrjob/wordcount-mr.amartinezv.20200413.181723.896350/files/
Running step 1 of 1...
JAR does not exist or is not a normal file: /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop-mapreduce/hadoop-streaming.jar
Attempting to fetch counters from logs...
Can't fetch history log; missing job ID
No counters found
Scanning logs for probable cause of failure...
Can't fetch history log; missing job ID
Can't fetch task logs; missing application ID
Step 1 of 1 failed: Command '"/usr/hdp/current/hadoop-client/bin/hadoop", "jar", "/opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop-mapreduce/hadoop-streaming.jar", "-files", "hdfs://user/amartinezv/tmp/mrjob/wordcount-mr.amartinezv.20200413.181723.896350/files/wd/mrjob.zip#mrjob.zip,hdfs://user/amartinezv/tmp/mrjob/wordcount-mr.amartinezv.20200413.181723.896350/files/wd/setup-wrapper.sh#setup-wrapper.sh,hdfs://user/amartinezv/tmp/mrjob/wordcount-mr.amartinezv.20200413.181723.896350/files/wd/wordcount-mr.py#wordcount-mr.py", "-input", "hdfs://user/amartinezv/datasets/gutenberg-small/1/*.*txt", "-output", "hdfs://user/amartinezv/result3", "-mapper", "/bin/sh -ex setup-wrapper.sh python3 wordcount-mr.py --step-num=0 --mapper", "-reducer", "/bin/sh -ex setup-wrapper.sh python3 wordcount-mr.py --step-num=0 --reducer"]' returned non-zero exit status 65280
[amartinezv@hdpjupyter 02-mapreduce]$
```

Name >	Size >	Last Modified >	Owner >	Group >	Permission	Erasure Coding
datasets	--	2020-03-27 22:16	amartinezv	bigdata	drwxr-xr-x	
tmp	--	2020-04-13 13:17	amartinezv	bigdata	drwxr-xr-x	

- Ejercicios de mrjob dejados en el github

1. Se tiene un conjunto de datos, que representan el salario anual de los empleados formales en Colombia por sector económico, según la DIAN.

- El salario promedio por Sector Económico (SE)

**PromedioSalarioSE.py** hace unos segundos

```
File Edit View Language
1 from mrjob.job import MRJob
```

```

2
3 class PromedioSalarioSE(MRJob):
4
5     def mapper(self, _, line):
6         idemp, sececon, salary, year = line.split(',')
7         try:
8             salary = float(salary)
9         except ValueError:
10             pass
11         else:
12             yield sececon, salary
13
14
15     def reducer (self, sececon, salario):
16         sumSalario=0
17         cont =0
18         for s in salario:
19             sumSalario += s
20             cont += 1
21
22         yield sececon, sumSalario/cont
23
24 if __name__ == '__main__':
25     PromedioSalarioSE.run()

```

```

[amartinezv@hdpjupyter ~]$ python PromedioSalarioSE.py bigdata/datasets/otros/dataempleados.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/PromedioSalarioSE.amartinezv.20200413.201225.733947
Running step 1 of 1...
job output is in /tmp/PromedioSalarioSE.amartinezv.20200413.201225.733947/output
Streaming final output from /tmp/PromedioSalarioSE.amartinezv.20200413.201225.733947/output...
"1212" 77000.0
"1234" 37500.0
"1412" 76000.0
"3432" 34000.0
"5434" 36000.0
Removing temp directory /tmp/PromedioSalarioSE.amartinezv.20200413.201225.733947...
[amartinezv@hdpjupyter ~]$

```

## 2. El salario promedio por Empleado

 **jupyter** PromedioSalarioEmpleados.py ✓ hace unos segundos

```

File Edit View Language

1 from mrjob.job import MRJob
2
3 class PromedioSalarioEmpleados(MRJob):
4
5     def mapper(self, _, line):
6         idemp, sececon, salary, year = line.split(',')
7         try:
8             salary = float(salary)
9         except ValueError:
10             pass
11         else:
12             yield idemp, salary
13
14
15     def reducer (self, idemp, salario):
16         sumSalario=0
17         cont =0
18         for s in salario:
19             sumSalario += s
20             cont += 1
21
22         yield idemp, sumSalario/cont
23
24 if __name__ == '__main__':
25     PromedioSalarioEmpleados.run()

```

```

[amartinezv@hdpjupyter ~]$ python PromedioSalarioEmpleados.py bigdata/datasets/otros/dataempleados.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/PromedioSalarioEmpleados.amartinezv.20200413.201858.308225
Running step 1 of 1...
job output is in /tmp/PromedioSalarioEmpleados.amartinezv.20200413.201858.308225/output
Streaming final output from /tmp/PromedioSalarioEmpleados.amartinezv.20200413.201858.308225/output...
"1115" 62333.333333333336
"3233" 35500.0
"3237" 40000.0
Removing temp directory /tmp/PromedioSalarioEmpleados.amartinezv.20200413.201858.308225...
[amartinezv@hdpjupyter ~]$

```

## 3. Número de SE por Empleado que ha tenido a lo largo de la estadística



The image shows a Jupyter Notebook interface with a menu bar (File, Edit, View, Language) and a title bar that reads "NumeroSEporEmpleados.py" with a checkmark and the text "hace un minuto". The code is as follows:

```
1 from mrjob.job import MRJob
2
3 class NumeroSEporEmpleados(MRJob):
4
5     def mapper(self, _, line):
6         idemp, sececon, salary, year = line.split(',')
7         try:
8             salary = float(salary)
9         except ValueError:
10             pass
11         else:
12             yield idemp, sececon
13
14
15     def reducer(self, idemp, sector):
16         cont = 0
17         for s in sector:
18             cont += 1
19
20         yield idemp, cont
21
22 if __name__ == '__main__':
23     NumeroSEporEmpleados.run()
```

```
[amartinezv@hdpjupyter ~]$ python NumeroSEporEmpleados.py bigdata/datasets/otros/dataempleados.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/NumeroSEporEmpleados.amartinezv.20200413.202551.910763
Running step 1 of 1...
job output is in /tmp/NumeroSEporEmpleados.amartinezv.20200413.202551.910763/output
Streaming final output from /tmp/NumeroSEporEmpleados.amartinezv.20200413.202551.910763/output...
"1115" 3
"3233" 2
"3237" 1
Removing temp directory /tmp/NumeroSEporEmpleados.amartinezv.20200413.202551.910763...
[amartinezv@hdpjupyter ~]$
```