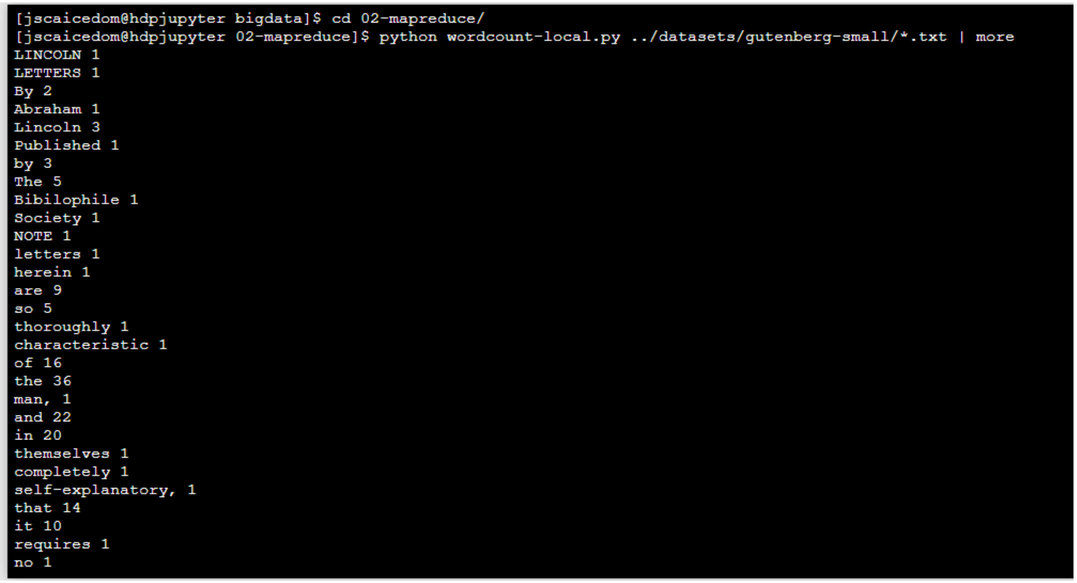
2. MapReduce

miércoles, 29 de abril de 2020 15:34

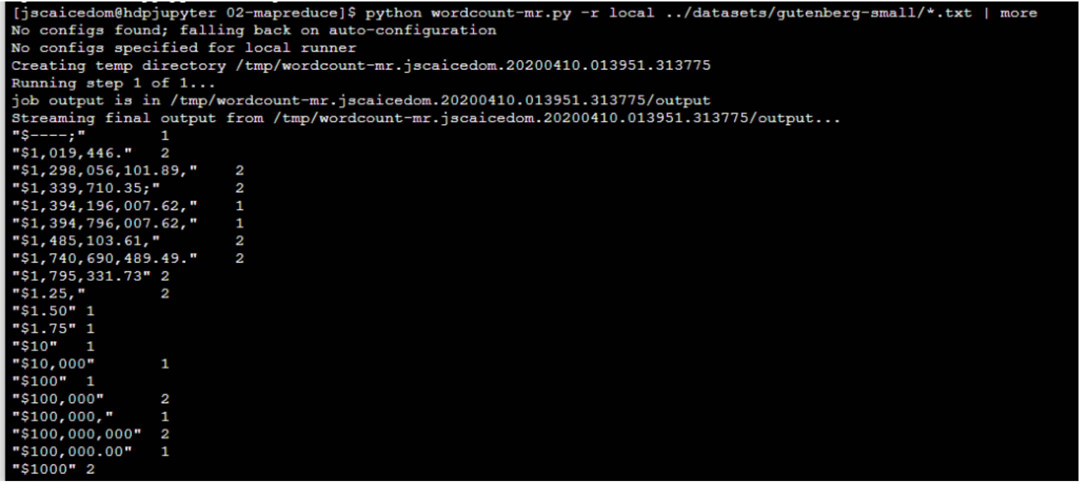
- Ejecutar y registrar en bitacora ejemplo de wordcount-local.py y en MapReduce con mrjob de wordcount-mr.py en versión LOCAL (Jupyterhub):

    - Local:



```
[jscaicedom@hdpjupyter bigdata]$ cd 02-mapreduce/
[jscaicedom@hdpjupyter 02-mapreduce]$ python wordcount-local.py ../datasets/gutenberg-small/*.txt | more
LINCOLN 1
LETTERS 1
By 2
Abraham 1
Lincoln 3
Published 1
by 3
The 5
Bibilophile 1
Society 1
NOTE 1
letters 1
herein 1
are 9
so 5
thoroughly 1
characteristic 1
of 16
the 36
man, 1
and 22
in 20
themselves 1
completely 1
self-explanatory, 1
that 14
it 10
requires 1
no 1
```

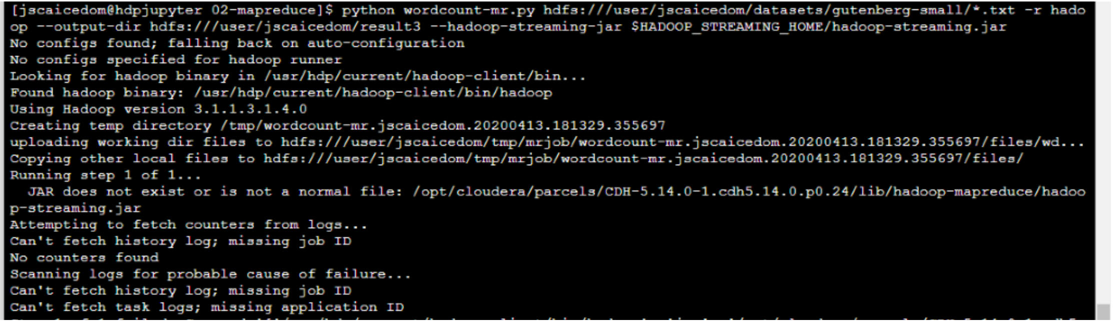
- Mrjob:



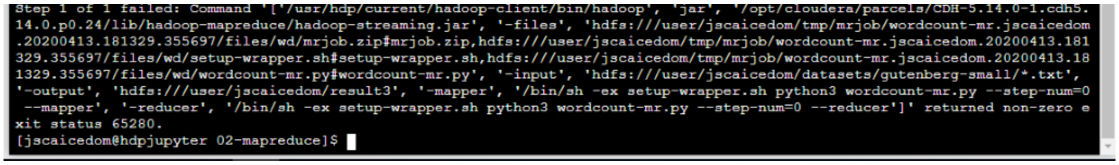
```
[jscaicedom@hdpjupyter 02-mapreduce]$ python wordcount-mr.py -r local ../datasets/gutenberg-small/*.txt | more
No configs found; falling back on auto-configuration
No configs specified for local runner
Creating temp directory /tmp/wordcount-mr.jscaicedom.20200410.013951.313775
Running step 1 of 1...
job output is in /tmp/wordcount-mr.jscaicedom.20200410.013951.313775/output
Streaming final output from /tmp/wordcount-mr.jscaicedom.20200410.013951.313775/output...
"$----;"        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
```

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

    - Ejecuté los comandos del github y creé las variables de ambiente, pero me salen errores y no crea la carpeta result3.

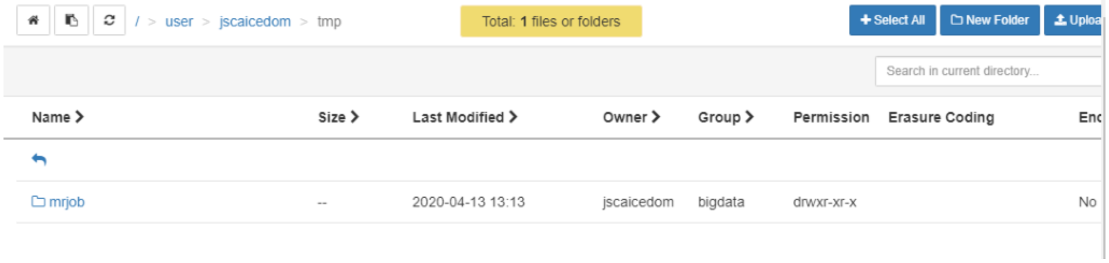


```
[jscaicedom@hdpjupyter 02-mapreduce]$ python wordcount-mr.py hdfs:///user/jscaicedom/datasets/gutenberg-small/*.txt -r hado
op --output-dir hdfs:///user/jscaicedom/result3 --hadoop-streaming-jar $HADOOP_STREAMING_HOME/hadoop-streaming.jar
No configs found; falling back on auto-configuration
No configs 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.jscaicedom.20200413.181329.355697
uploading working dir files to hdfs:///user/jscaicedom/tmp/mrjob/wordcount-mr.jscaicedom.20200413.181329.355697/files/wd...
Copying other local files to hdfs:///user/jscaicedom/tmp/mrjob/wordcount-mr.jscaicedom.20200413.181329.355697/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/hadoo
p-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/jscaicedom/tmp/mrjob/wordcount-mr.jscaicedom
.20200413.181329.355697/files/wd/mrjob.zip#mrjob.zip,hdfs:///user/jscaicedom/tmp/mrjob/wordcount-mr.jscaicedom.20200413.181
329.355697/files/wd/setup-wrapper.sh#setup-wrapper.sh,hdfs:///user/jscaicedom/tmp/mrjob/wordcount-mr.jscaicedom.20200413.18
1329.355697/files/wd/wordcount-mr.py#wordcount-mr.py', '-input', 'hdfs:///user/jscaicedom/datasets/gutenberg-small/*.txt',
'-output', 'hdfs:///user/jscaicedom/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 e
xit status 65280.
[jscaicedom@hdpjupyter 02-mapreduce]$ 
```

- Y me crea una carpeta tmp:

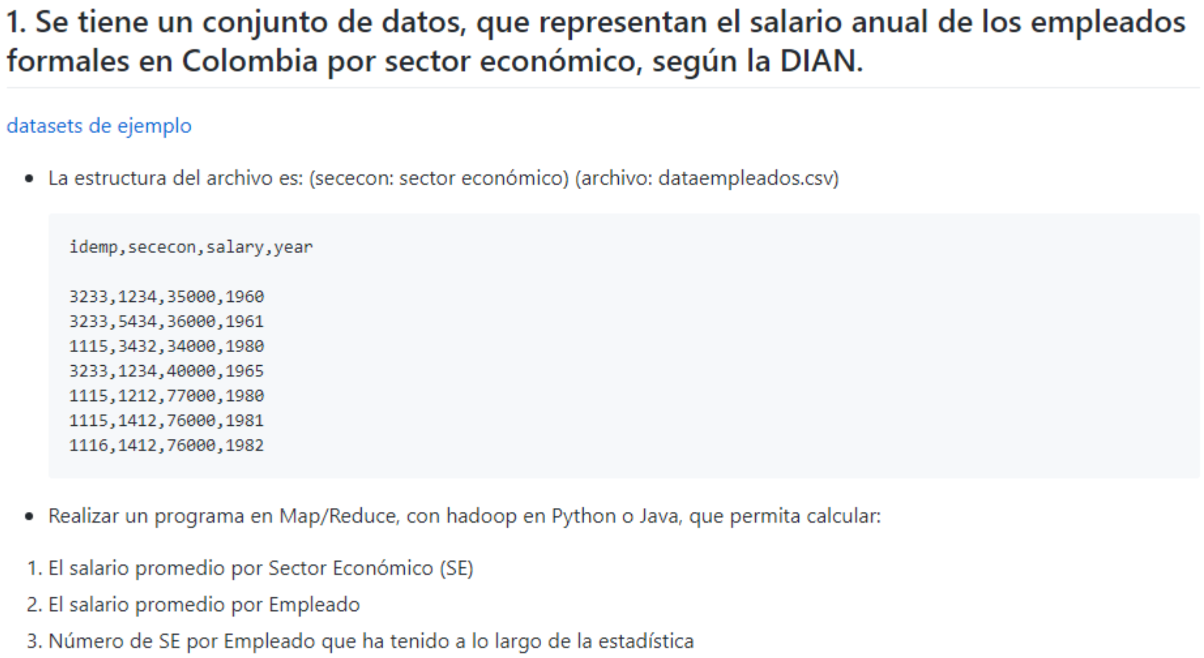


/ > user > jscaicedom > tmp
Total: 1 files or folders
+ Select All
New Folder
Uploa
Search in current directory...

Name	Size	Last Modified	Owner	Group	Permission	Erasure Coding	Enc
mrjob	--	2020-04-13 13:13	jscaicedom	bigdata	drwxr-xr-x		No

- Realizar uno de los 3 ejercicios de mrjob dejados en el github, entregar fuentes e instrucciones de ejecucion en el github de su propio lab y en las bitacoras que considere.

  - Escogí el primer ejercicio:



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

datasets de ejemplo

- La estructura del archivo es: (sececon: sector económico) (archivo: dataempleados.csv)

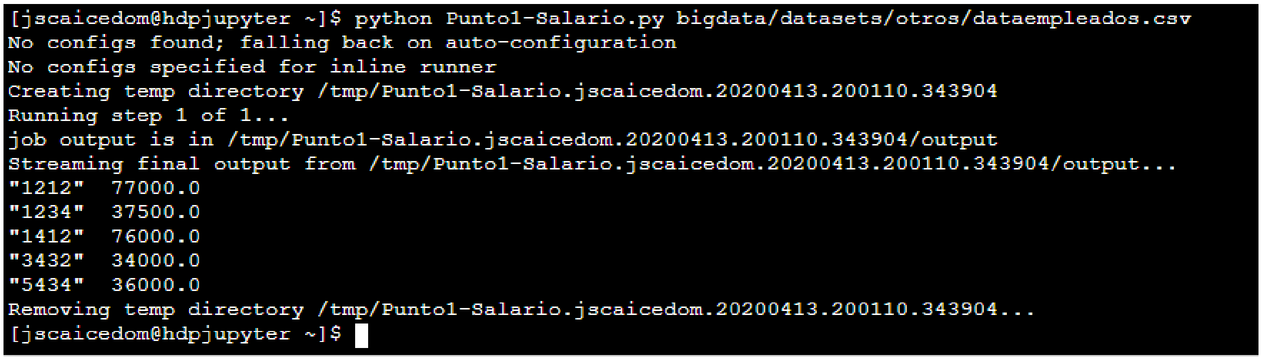
```
idemp,sececon,salary,year

3233,1234,35000,1960
3233,5434,36000,1961
1115,3432,34000,1980
3233,1234,40000,1965
1115,1212,77000,1980
1115,1412,76000,1981
1116,1412,76000,1982
```

- Realizar un programa en Map/Reduce, con hadoop en Python o Java, que permita calcular:

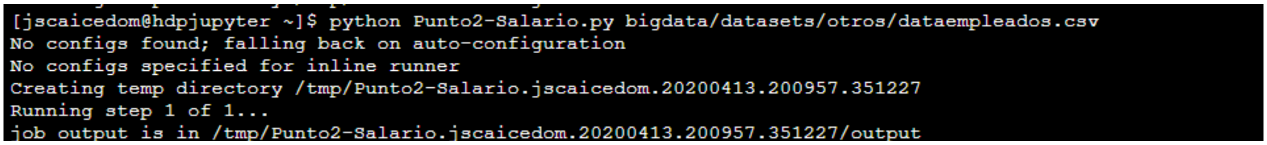
1. El salario promedio por Sector Económico (SE)
2. El salario promedio por Empleado
3. Número de SE por Empleado que ha tenido a lo largo de la estadística

1.

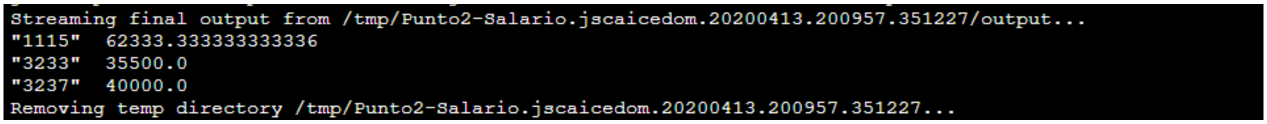


```
[jscaicedom@hdpjupyter ~]$ python Punto1-Salario.py bigdata/datasets/otros/dataempleados.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/Punto1-Salario.jscaicedom.20200413.200110.343904
Running step 1 of 1...
job output is in /tmp/Punto1-Salario.jscaicedom.20200413.200110.343904/output
Streaming final output from /tmp/Punto1-Salario.jscaicedom.20200413.200110.343904/output...
"1212"  77000.0
"1234"  37500.0
"1412"  76000.0
"3432"  34000.0
"5434"  36000.0
Removing temp directory /tmp/Punto1-Salario.jscaicedom.20200413.200110.343904...
[jscaicedom@hdpjupyter ~]$ 
```

2.

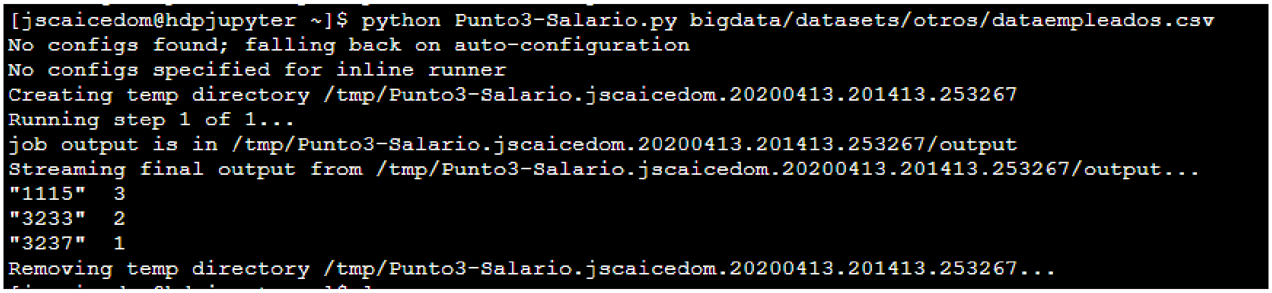


```
[jscaicedom@hdpjupyter ~]$ python Punto2-Salario.py bigdata/datasets/otros/dataempleados.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/Punto2-Salario.jscaicedom.20200413.200957.351227
Running step 1 of 1...
job output is in /tmp/Punto2-Salario.jscaicedom.20200413.200957.351227/output
```



```
Streaming final output from /tmp/Punto2-Salario.jscaicedom.20200413.200957.351227/output...
"1115"  62333.333333333336
"3233"  35500.0
"3237"  40000.0
Removing temp directory /tmp/Punto2-Salario.jscaicedom.20200413.200957.351227...
```

3.



```
[jscaicedom@hdpjupyter ~]$ python Punto3-Salario.py bigdata/datasets/otros/dataempleados.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/Punto3-Salario.jscaicedom.20200413.201413.253267
Running step 1 of 1...
job output is in /tmp/Punto3-Salario.jscaicedom.20200413.201413.253267/output
Streaming final output from /tmp/Punto3-Salario.jscaicedom.20200413.201413.253267/output...
"1115"  3
"3233"  2
"3237"  1
Removing temp directory /tmp/Punto3-Salario.jscaicedom.20200413.201413.253267...
```