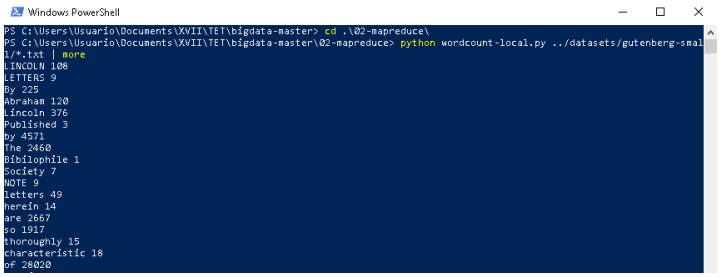
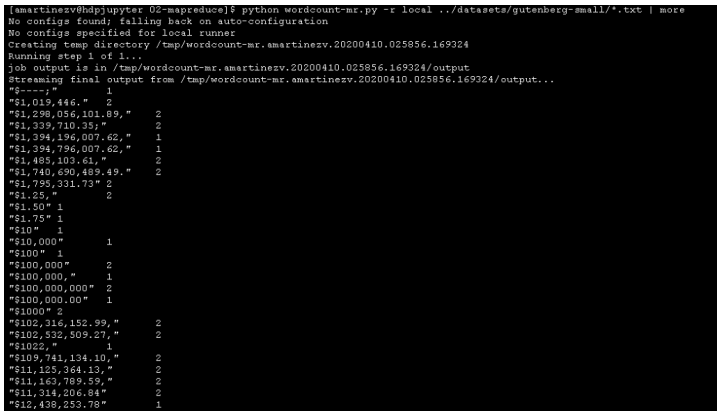
MapReduce

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

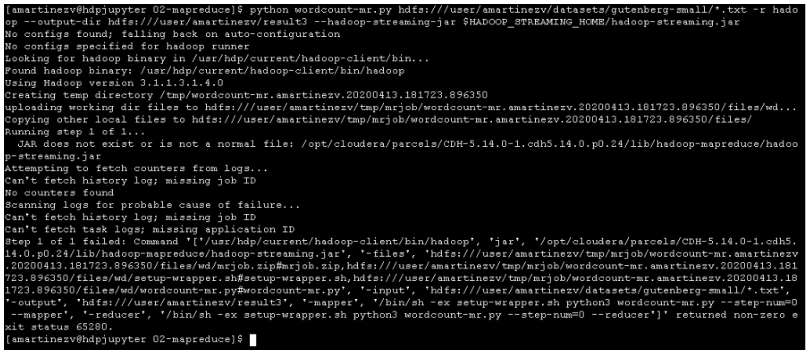


  - Mrjob



- 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



- 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.

  1. El salario promedio por Sector Económico (SE)

```
class PromedioSalarioSE(MRJob):

    def mapper(self, _, line):
        idemp, sececon, salary, year = line.split(',')
        try:
            salary = float(salary)
        except ValueError:
            pass
        else:
            yield sececon, salary


    def reducer (self, sececon, salario):
        sumSalario=0
        cont =0
        for s in salario:
            sumSalario += s
            cont += 1

        yield sececon, sumSalario/cont

if __name__ == '__main__':
    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

```
from mrjob.job import MRJob

class PromedioSalarioEmpleados(MRJob):

    def mapper(self, _, line):
        idemp, sececon, salary, year = line.split(',')
        try:
            salary = float(salary)
        except ValueError:
            pass
        else:
            yield idemp, salary


    def reducer (self, idemp, salario):
        sumSalario=0
        cont =0
        for s in salario:
            sumSalario += s
            cont += 1

        yield idemp, sumSalario/cont

if __name__ == '__main__':
    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

jupyter NumeroSEporEmpleados.py hace un minuto

File Edit View Language

```python
from mrjob.job import MRJob

class NumeroSEporEmpleados(MRJob):

    def mapper(self, _, line):
        idemp, sececon, salary, year = line.split(',')
        try:
            salary = float(salary)
        except ValueError:
            pass
        else:
            yield idemp, sececon


    def reducer (self, idemp, sector):
        cont =0
        for s in sector:
            cont += 1

        yield idemp, cont

if __name__ == '__main__':
    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 ~]$
```