Assignment 3

Name: Akshay Jain || CWID: A20592846 || Course: CSP-554

Question - 4: Steps to copy files from 'Local' -> '/home/hadoop' -> '/user/hadoop'

```
(base) akshayjain@dhcp156 Assignment3 % scp -i emr-key-pair.pem WordCount.py hadoop@ec2-3-238-242-207.compute-1.amazonaws.com:/home/hadoop
WordCount.py                                  100%  402    10.0KB/s   00:00

(base) akshayjain@dhcp156 Assignment3 % scp -i emr-key-pair.pem w.data hadoop@ec2-3-238-242-207.compute-1.amazonaws.com:/home/hadoop
w.data                                        100%  528    17.4KB/s   00:00
```

```
[hadoop@ip-172-31-68-171 ~]$ hadoop fs -copyFromLocal /home/hadoopWordCount.py /user/hadoop
copyFromLocal: `/home/hadoopWordCount.py': No such file or directory
[hadoop@ip-172-31-68-171 ~]$ hadoop fs -copyFromLocal /home/hadoop/WordCount.py /user/hadoop
[hadoop@ip-172-31-68-171 ~]$ hadoop fs -copyFromLocal /home/hadoop/w.data /user/hadoop
[hadoop@ip-172-31-68-171 ~]$ hadoop fs -ls /user/hadoop
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup        402 2022-09-22 17:12 /user/hadoop/WordCount.py
-rw-r--r--   1 hadoop hdfsadmingroup        528 2022-09-22 17:13 /user/hadoop/w.data
[hadoop@ip-172-31-68-171 ~]$
```

Question -5: InstalledmrjobLibrarySuccessfully

```
[hadoop@ip-172-31-68-171 ~]$ sudo /usr/bin/pip3.7 install mrjob[aws]
WARNING: Running pip install with root privileges is generally not a good idea. Try `pip3.7 install --user` instead.
Collecting mrjob[aws]
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
     |████████████████████████████████| 439 kB 18.1 MB/s
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.7/site-packages (from mrjob[aws]) (5.4.1)
Collecting boto3>=1.10.0; extra == "aws"
  Downloading boto3-1.24.78-py3-none-any.whl (132 kB)
     |████████████████████████████████| 132 kB 76 kB/s
Collecting botocore>=1.13.26; extra == "aws"
  Downloading botocore-1.27.78-py3-none-any.whl (9.1 MB)
     |████████████████████████████████| 9.1 MB 39.3 MB/s
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /usr/local/lib/python3.7/site-packages (from boto3>=1.10.0; extra == "aws"->mrjob[aws]) (1.0.0)
Collecting s3transfer<0.7.0,>=0.6.0
  Downloading s3transfer-0.6.0-py3-none-any.whl (79 kB)
     |████████████████████████████████| 79 kB 107 kB/s
Collecting urllib3<1.27,>=1.25.4
  Downloading urllib3-1.26.12-py2.py3-none-any.whl (140 kB)
     |████████████████████████████████| 140 kB 45.4 MB/s
Collecting python-dateutil<3.0.0,>=2.1
  Downloading python_dateutil-2.8.2-py2.py3-none-any.whl (247 kB)
     |████████████████████████████████| 247 kB 44.2 MB/s
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (from python-dateutil<3.0.0,>=2.1->botocore>=1.13.26; extra == "aws"->mrjob[aws]) (1.13.0)
Installing collected packages: urllib3, python-dateutil, botocore, s3transfer, boto3, mrjob
  WARNING: The scripts mrjob, mrjob-3 and mrjob-3.7 are installed in '/usr/local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed boto3-1.24.78 botocore-1.27.78 mrjob-0.7.4 python-dateutil-2.8.2 s3transfer-0.6.0 urllib3-1.26.12
```

Question -6: Modified Wordcount.py and its output

WordCount2 〉No Selection

```python
from mrjob.job import MRJob
import re

WORD_RE = re.compile(r"[\w']+")


class MRWordCount(MRJob):

    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            if word[0] >= 'a' and word[0] <='n':
                yield "a_to_n", 1
            else:
                yield "other", 1
    def combiner(self, word, counts):
        yield word, sum(counts)

    def reducer(self, word, counts):
        yield word, sum(counts)


if __name__ == '__main__':
    MRWordCount.run()
```

Run:

```
[hadoop@ip-172-31-68-171 ~]$ python WordCount2.py -r hadoop hdfs:///user/hadoop/w.data
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/WordCount2.hadoop.20220922.172907.184095
```

Output:

```
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220922.172907.184095/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220922.172907.184095/output...
"a_to_n"        46
"other" 49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220922.172907.184095...
Removing temp directory /tmp/WordCount2.hadoop.20220922.172907.184095...
```

Question 10: Modified Salaries2.py and its output

Salaries2 〉 No Selection

```
1  from mrjob.job import MRJob
2
3  class MRSalaries(MRJob):
4
5      def mapper(self, _, line):
6          (name,jobTitle,agencyID,agency,hireDate,annualSalary,grossPay) =
               line.split('\t')
7          if float(annualSalary) >= 100,000.00:
8              yield "High", 1
9          if(float(annualSalary) >= 50000.00 and float(annualSalary) <= 99999.99 :
10             yield "Medium", 1
11         else:
12             yield "Low", 1
13
14     def combiner(self, annualSalary, counts):
15         yield annualSalary, sum(counts)
16
17     def reducer(self, annualSalary, counts):
18         yield annualSalary, sum(counts)
19
20
21 if __name__ == '__main__':
22     MRSalaries.run()
23
```

Run:

```
[hadoop@ip-172-31-68-171 ~]$ python Salaries2.py -r hadoop hdfs:///user/hadoop/Salaries.tsv
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/Salaries2.hadoop.20220922.181729.227572
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220922.181729.227572/files/wd..
.
```

Output:

```
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220922.181729.227572/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220922.181729.227572/output...
"High"  442
"Low"   7506
"Medium"        6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220922.181729.227572...
Removing temp directory /tmp/Salaries2.hadoop.20220922.181729.227572...
```

Question 12: Moviesperuser.py and its output

Moviesperuser > No Selection

```
1  from mrjob.job import MRJob
2
3  class MRWovieUserRating(MRJob):
4
5      def mapper(self, _, line):
6          (userId,mvId,rating,timestamp) = line.split(',')
7          yield userId, 1
8
9      def combiner(self, userId, counts):
10         yield userId, sum(counts)
11
12     def reducer(self, userId, counts):
13         yield userId, sum(counts)
14
15
16 if __name__ == '__main__':
17     MRMovieUserRating.run()
18
```

Output:

```
job output is in hdfs:///user/hadoop/tmp/mrjob/Moviesperuser.hadoop.20220922.185745.137962/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Moviesperuser.hadoop.20220922.185745.137962/output...
"1"     20
"10"    46
"100"   25
"101"   55
"102"   678
"103"   94
"104"   76
"105"   525
"106"   45
"107"   32
"108"   31
"109"   23
"11"    38
"110"   120
"111"   341
"112"   21
"113"   27
"114"   25
"115"   41
"116"   25
"117"   55
"118"   189
"119"   641
"12"    61
"120"   138
"121"   80
"122"   40
"123"   33
"124"   85
"125"   210
"126"   64
"127"   21
"128"   323
"129"   26
"13"    53
"130"   375
"131"   44
"132"   94
"133"   178
"134"   311
"135"   22
"136"   50
"137"   80
"138"   81
```