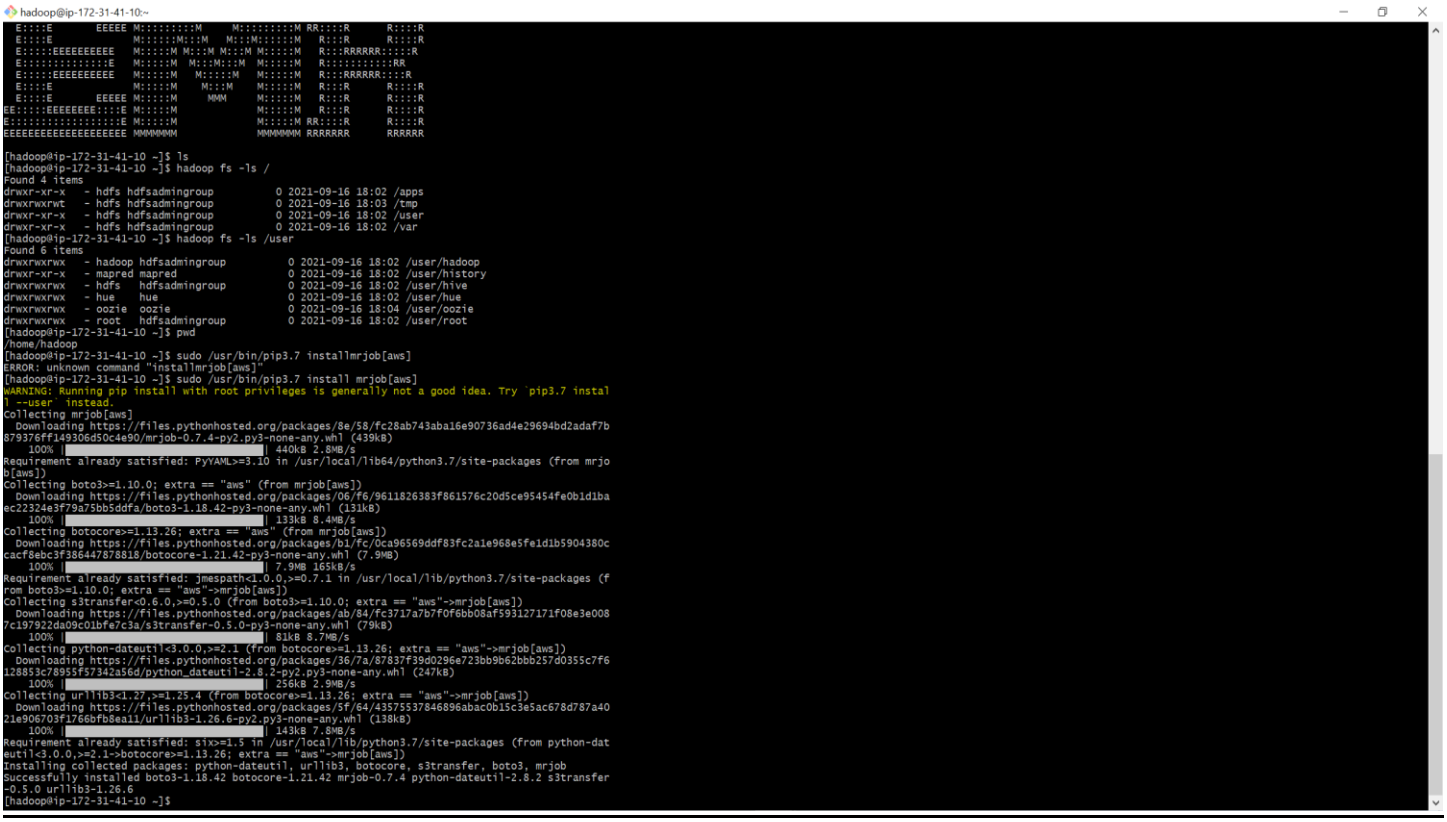
CSP – 554 Big Data Technologies

Assignment #3

Rutul Mehta
A20476293

Installed mrjob Library Successfully



Question 6

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

WORD_RE = re.compile(r"[\w']+")
aton = ('a','b','c','d','e','f','g','h','i','j','k','l','m','n')

class MRWordCount(MRJob):

    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            if(word.startswith(aton)):
                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()
```

```
hadoop@ip-172-31-41-10:~
  map 100% reduce 100%
 Job job_1631815372645_0012 completed successfully
 Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20210916.204710.895692/output
Counters: 50
        File Input Format Counters
                Bytes Read=2376
        File Output Format Counters
                Bytes Written=23
        File System Counters
                FILE: Number of bytes read=118
                FILE: Number of bytes written=2472942
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3376
                HDFS: Number of bytes written=23
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=33
                HDFS: Number of write operations=6
        Job Counters
                Data-local map tasks=8
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=178063872
                Total megabyte-milliseconds taken by all reduce tasks=60186624
                Total time spent by all map tasks (ms)=115927
                Total time spent by all maps in occupied slots (ms)=5564496
                Total time spent by all reduce tasks (ms)=19592
                Total time spent by all reduces in occupied slots (ms)=1880832
                Total vcore-milliseconds taken by all map tasks=115927
                Total vcore-milliseconds taken by all reduce tasks=19592
        Map-Reduce Framework
                CPU time spent (ms)=18500
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=2549
                Input split bytes=1000
                Map input records=6
                Map output bytes=996
                Map output materialized bytes=464
                Map output records=95
                Merged Map outputs=24
                Physical memory (bytes) snapshot=4866211840
                Reduce input groups=2
                Reduce input records=6
                Reduce output records=2
                Reduce shuffle bytes=464
                Shuffled Maps =24
                Spilled Records=12
                Total committed heap usage (bytes)=4226285568
                Virtual memory (bytes) snapshot=40449302528
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20210916.204710.895692/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20210916.204710.895692/output...
"a_to_n"        46
"other" 49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20210916.204710.895692...
Removing temp directory /tmp/WordCount2.hadoop.20210916.204710.895692...
[hadoop@ip-172-31-41-10 ~]$
```

Question 10

```python
from mrjob.job import MRJob

class MRSalaries(MRJob):

    def mapper(self, _, line):
        (name,jobTitle,agencyID,agency,hireDate,annualSalary,grossPay) = line.split('\t')
        f_annualSalary = float(annualSalary)
        if f_annualSalary>0 and f_annualSalary<49999.99:
            yield 'Low', 1
        elif f_annualSalary>50000 and f_annualSalary<99999.99:
            yield 'Medium', 1
        elif f_annualSalary>=100000:
            yield 'High', 1

    def combiner(self, SalaryCategory, counts):
        yield SalaryCategory, sum(counts)

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


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

```
hadoop@ip-172-31-41-10:~
  Job job_1631815372645_0014 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20210916.211948.719587/output
Counters: 50
        File Input Format Counters
                Bytes Read=1567508
        File Output Format Counters
                Bytes Written=36
        File System Counters
                FILE: Number of bytes read=215
                FILE: Number of bytes written=2473198
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1568556
                HDFS: Number of bytes written=36
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=33
                HDFS: Number of write operations=6
        Job Counters
                Data-local map tasks=8
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=188273664
                Total megabyte-milliseconds taken by all reduce tasks=64816128
                Total time spent by all map tasks (ms)=122574
                Total time spent by all maps in occupied slots (ms)=5883552
                Total time spent by all reduce tasks (ms)=21099
                Total time spent by all reduces in occupied slots (ms)=2025504
                Total vcore-milliseconds taken by all map tasks=122574
                Total vcore-milliseconds taken by all reduce tasks=21099
        Map-Reduce Framework
                CPU time spent (ms)=21720
                Combine input records=13775
                Combine output records=24
                Failed Shuffles=0
                GC time elapsed (ms)=2462
                Input split bytes=1048
                Map input records=13818
                Map output bytes=129536
                Map output materialized bytes=696
                Map output records=13775
                Merged Map outputs=24
                Physical memory (bytes) snapshot=4817993728
                Reduce input groups=3
                Reduce input records=24
                Reduce output records=3
                Reduce shuffle bytes=696
                Shuffled Maps =24
                Spilled Records=48
                Total committed heap usage (bytes)=4219994112
                Virtual memory (bytes) snapshot=40480157696
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20210916.211948.719587/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20210916.211948.719587/output...
"High"  442
"Low"   7035
"Medium"        6298
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20210916.211948.719587...
Removing temp directory /tmp/Salaries2.hadoop.20210916.211948.719587...
[hadoop@ip-172-31-41-10 ~]$
```

Question 12

```python
from mrjob.job import MRJob

class MRSalaries(MRJob):

    def mapper(self, _, line):
        (userId,movieId,rating,timeStamp) = line.split(',')
        yield userId, 1

    def combiner(self, userId, counts):
        yield userId, sum(counts)

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


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