

**CSP – 554 Big Data Technologies**

**Assignment #3**

Rutul Mehta

A20476293

## Installed mrjob Library Successfully

```
E:::E EEEEE M:::M:::M:::M:::M M:::M:::M RR:::R R:::R
E:::E M:::M:::M M:::M:::M M:::M:::M R:::R R:::R
E:::E EEEEEEE M:::M:::M M:::M:::M M:::M:::M R:::RRRRR R:::R
E:::E EEEEEEE M:::M:::M M:::M:::M M:::M:::M R:::R
E:::E EEEEEEE M:::M:::M M:::M:::M M:::M:::M R:::RRRRR R:::R
E:::E EEEEE EEEE M:::M:::M M:::M R:::R R:::R
E:::E EEEEE EEEE M:::M:::M M:::M R:::R R:::R
E:::E EEEEEEE:::E M:::M:::M M:::M:::M R:::R R:::R
E:::E EEEEE EEEE M:::M:::M M:::M RR:::R R:::R
EEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRR RRRRRR

[hadoop@pip-172-31-41-10 ~]$ ls
[hadoop@pip-172-31-41-10 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x - hdfs hdfsadmingroup 0 2021-09-16 18:02 /apps
drwxr-xrwT - hdfs hdfsadmingroup 0 2021-09-16 18:03 /tmp
drwxr-xr-x - hdfs hdfsadmingroup 0 2021-09-16 18:02 /user
drwxr-xr-x - hdfs hdfsadmingroup 0 2021-09-16 18:02 /var
[hadoop@pip-172-31-41-10 ~]$ hadoop fs -ls /user
Found 6 items
drwxr-xrwx - hadoop hdfsadmingroup 0 2021-09-16 18:02 /user/hadoop
drwxr-xr-x - mapped mapped 0 2021-09-16 18:02 /user/history
drwxr-xrwx - hdfs hdfsadmingroup 0 2021-09-16 18:02 /user/hive
drwxr-xrwx - hue hue 0 2021-09-16 18:02 /user/hue
drwxr-xrwx - locale locale 0 2021-09-16 18:04 /user/oozie
drwxr-xrwx - root hdfsadmingroup 0 2021-09-16 18:02 /user/root
[hadoop@pip-172-31-41-10 ~]$ pwd
/home/hadoop
[hadoop@pip-172-31-41-10 ~]$ sudo /usr/bin/pip3.7 install mrjob[aaws]
ERROR: unknown command "installmrjob[aaws]"
[hadoop@pip-172-31-41-10 ~]$ sudo /usr/bin/pip3.7 install mrjob[aaws]
WARNING: Running pip install with root privileges is generally not a good idea. Try 'pip3.7 install --user' instead.
Collecting mrjob[aaws]
  Downloading https://files.pythonhosted.org/packages/8e/5f/fc28ab743aba1e90736ada29694bd2adafb78336fd9d30650c4e80/mrjob-0.7.4-py2.py3-none-any.whl (439K)
    100% |#####| 440k 2.8MB/s
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.7/site-packages (from mrjob[aaws])
Collecting boto3<=1.0.0, extra== "aws" (from mrjob[aaws])
  Downloading https://files.pythonhosted.org/packages/06/f6/8181e2638f8fe157c20dc9e545afe0dbidaec2232463f79a75bb5dfda/boto3-1.18.42-py3-none-any.whl (136kB)
    100% |#####| 136k 8.8MB/s
Collecting botocore<=1.13.26, extra== "aws" (from mrjob[aaws])
  Downloading https://files.pythonhosted.org/packages/1b/fc/0a9656dd8df3ca2ae6865fe1db5904380ccac8ebc3f38644787881s/botocore-1.21.42-py3-none-any.whl (7.9Mb)
    100% |#####| 7.9Mb
Requirement already satisfied: jmespath<1.0.0, >=0.7.1 in /usr/local/lib/python3.7/site-packages (from boto3<=1.0.0, extra== "aws">->mrjob[aaws])
Collecting s3transfer<0.6.0, >=0.5.0 (from boto3<=1.0.0, extra== "aws">->mrjob[aaws])
  Downloading https://files.pythonhosted.org/packages/a0/84/fc371a7b70f6bb0eaf532171f08e9c0087c197922da09c01bf6c3a/s3transfer-0.5.0-py3-none-any.whl (79K)
    100% |#####| 81k 8.7MB/s
Requirement already satisfied: python-dateutil<3.0.0, >=2.8.1 in /usr/local/lib/python3.7/site-packages (from s3transfer<0.6.0, >=0.5.0->mrjob[aaws])
Collecting https://files.pythonhosted.org/packages/36/74/87837f309d96273bb962ebb257d0355cf76128853c78955f57342a56d/python-dateutil-2.8.2-py3-none-any.whl (247K)
    100% |#####| 256k 2.9MB/s
Collecting urllib3<1.27, >=1.25.4 (from boto3<=1.13.26, extra== "aws">->mrjob[aaws])
  Downloading https://files.pythonhosted.org/packages/5f/64/43575337646896abac0b35ace3678d87a4021e906703f166bf8eaall/urllib3-1.26.6-py2.py3-none-any.whl (138K)
    100% |#####| 143k 7.8MB/s
Requirement already satisfied: six<1.16.0, >=1.10 in /usr/local/lib/python3.7/site-packages (from python-dateutil<3.0.0, >=2.8.2->boto3<=1.13.26, extra== "aws">->mrjob[aaws])
Installing collected packages: python-dateutil, urllib3, botocore, s3transfer, boto3, mrjob
  Attempting uninstall: boto3-1.18.42
    Found existing installation: boto3-1.18.42
    Uninstalling boto3-1.18.42:
      Successfully uninstalled boto3-1.18.42
  Attempting uninstall: python-dateutil-2.8.2
    Found existing installation: python-dateutil-2.8.2
    Uninstalling python-dateutil-2.8.2:
      Successfully uninstalled python-dateutil-2.8.2
  Successfully installed boto3-1.18.42 botocore-1.21.42 mrjob-0.7.4 python-dateutil-2.8.2 s3transfer-0.5.0 urllib3-1.26.6
```

## Question 6

```
1 from mrjob.job import MRJob
2 import re
3
4 WORD_RE = re.compile(r"[\w']+")
5 aton = ('a','b','c','d','e','f','g','h','i','j','k','l','m','n')
6
7 class MRWordCount(MRJob):
8
9     def mapper(self, _, line):
10         for word in WORD_RE.findall(line):
11             if word.startswith(aton):
12                 yield 'a_to_n', 1
13             else:
14                 yield 'other', 1
15
16     def combiner(self, word, counts):
17         yield word, sum(counts)
18
19     def reducer(self, word, counts):
20         yield word, sum(counts)
21
22
23 if __name__ == '__main__':
24     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=376
  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=1
  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

```
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         f_annualSalary = float(annualSalary)
8         if f_annualSalary>0 and f_annualSalary<49999.99:
9             yield 'Low', 1
10        elif f_annualSalary>50000 and f_annualSalary<99999.99:
11            yield 'Medium', 1
12        elif f_annualSalary>=100000:
13            yield 'High', 1
14
15    def combiner(self, SalaryCategory, counts):
16        yield SalaryCategory, sum(counts)
17
18    def reducer(self, SalaryCategory, counts):
19        yield SalaryCategory, sum(counts)
20
21
22 if __name__ == '__main__':
23     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
  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

```
1 from mrjob.job import MRJob
2
3 class MRSalaries(MRJob):
4
5     def mapper(self, _, line):
6         (userId,movieId,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     MRSalaries.run()
```

hadoop@ip-172-31-41-10~

```
"530" 78
"533" 163
"536" 109
"539" 26
"542" 63
"545" 92
"548" 138
"551" 85
"554" 64
"557" 66
"56" 522
"560" 100
"563" 158
"566" 22
"569" 85
"572" 106
"575" 547
"578" 34
"581" 49
"584" 193
"587" 504
"59" 78
"590" 89
"593" 70
"596" 487
"599" 192
"602" 129
"605" 437
"608" 296
"611" 35
"614" 99
"617" 75
"62" 53
"620" 172
"623" 103
"626" 150
"629" 34
"632" 39
"635" 22
"638" 20
"641" 140
"644" 39
"647" 150
"65" 27
"650" 29
"653" 51
"656" 128
"659" 142
"662" 58
"665" 434
"668" 20
"671" 115
"68" 123
"71" 23
"74" 49
"77" 315
"8" 116
"80" 37
"83" 161
"86" 190
"89" 66
"92" 123
"95" 299
"98" 71
Removing HDFS temp directory hdfs://user/hadoop/tmp/mrjob/MovieReviewed.hadoop.20210916.215715.123242...
Removing temp directory /tmp/MovieReviewed.hadoop.20210916.215715.123242...
[hadoop@ip-172-31-41-10 ~]$
```