4) Install the mrjob library on your EMR master node. a) ssh to the master node (/home/hadoop) as you did in assignment #2 b) Enter the following (note if the first command does not work, try the second) sudo /usr/bin/pip3.7 install mrjob[aws] or try: sudo /usr/bin/pip3 install mrjob[aws]

Command executed-
$ ssh -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem
hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com

Screenshot-

```
Ankan Mazumdar@DESKTOP-CMULEBA MINGW64 /
$ ssh -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com
The authenticity of host 'ec2-18-191-187-150.us-east-2.compute.amazonaws.com (18.191.187.150)' can't be established.
ED25519 key fingerprint is SHA256:2WXi4FOIx2j8aMrJvUaj8MPsj+qfHWasEE8/cRC1rLw.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-18-191-187-150.us-east-2.compute.amazonaws.com' (ED25519) to the list of known hosts.

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
29 package(s) needed for security, out of 61 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R::::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE:::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-35-66 ~]$ pwd
/home/hadoop
```

Command executed-
[hadoop@ip-172-31-35-66 ~]$ python3 -m venv myenv
[hadoop@ip-172-31-35-66 ~]$ source myenv/bin/activate
(myenv) [hadoop@ip-172-31-35-66 ~]$ pip install mrjob
(myenv) [hadoop@ip-172-31-35-66 ~]$ sudo /usr/bin/pip3.7 install mrjob[aws]

Screenshot-

```
[hadoop@ip-172-31-35-66 ~]$ python3 -m venv myenv
[hadoop@ip-172-31-35-66 ~]$ source myenv/bin/activate
(myenv) [hadoop@ip-172-31-35-66 ~]$ pip install mrjob
Defaulting to user installation because normal site-packages is not writeable
Collecting mrjob
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
     |████████████████████████████████| 439 kB 37.0 MB/s
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.7/site-packages (from mrjob) (5.4.1)
Installing collected packages: mrjob
  WARNING: The scripts mrjob, mrjob-3 and mrjob-3.7 are installed in '/home/hadoop/.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 mrjob-0.7.4
(myenv) [hadoop@ip-172-31-35-66 ~]$ mrjob --version
-bash: mrjob: command not found
(myenv) [hadoop@ip-172-31-35-66 ~]$ mrjob
-bash: mrjob: command not found
(myenv) [hadoop@ip-172-31-35-66 ~]$ 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 14.1 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 boto3-1.28.57-py3-none-any.whl (135 kB)
     |████████████████████████████████| 135 kB 43.4 MB/s
Collecting botocore>=1.13.26; extra == "aws"
  Downloading botocore-1.31.57-py3-none-any.whl (11.2 MB)
     |████████████████████████████████| 11.2 MB 35.7 MB/s
Collecting s3transfer<0.8.0,>=0.7.0
  Downloading s3transfer-0.7.0-py3-none-any.whl (79 kB)
     |████████████████████████████████| 79 kB 14.2 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]) (1.0.1)
Collecting urllib3<1.27,>=1.25.4
  Downloading urllib3-1.26.16-py2.py3-none-any.whl (143 kB)
     |████████████████████████████████| 143 kB 36.2 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 33.9 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 boto3-1.28.57 botocore-1.31.57 mrjob-0.7.4 python-dateutil-2.8.2 s3transfer-0.7.0 urllib3-1.26.16
```

5) Next you will set up to execute the provided WordCount.py map reduce program found in the “Assignments” section of the Blackboard. Step 1: Download the two files “w.data” and “WordCount.py” to your PC or Mac. They are part of the documents included with the assignment. Step 2: Note to prevent confusion: the default directory of your Linux account on the Hadoop master node is “/home/hadoop.” But when we want to copy something to HDFS we will sometimes copy it to an HDFS directory beginning with “/user/hadoop.” Be aware, the Linux and HDFS file system path names have nothing to do with one another. Any similarity in naming (such as the use of the directory name “hadoop”) is just coincidental. Now open another terminal window (but don’t use it to ssh to the master node). This will allow you to access files on your PC or MAC to upload them to the Hadoop master node. From this terminal window use the secure copy (scp) program to move the WordCount.py file to the /home/hadoop directory of the master node. Step 3: Do the same for the assignment file w.data. That is move it to the directory /home/Hadoop on the Hadoop master node Linux file system. In this case copy the file from the Linux “/home/hadoop” directory to the Hadoop file system (HDFS), say to the directory “/user/hadoop”

Command executed-

$ scp -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem WordCount.py hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop WordCount.py

$ scp -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem w.data hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop w.data

(myenv) [hadoop@ip-172-31-35-66 ~]$ hdfs dfs -copyFromLocal -f /home/hadoop/w.data /user/hadoop/
(myenv) [hadoop@ip-172-31-35-66 ~]$ hdfs dfs -ls /user/hadoop/
Found 1 items
-rw-r--r-- 1 hadoop hdfsadmingroup 528 2023-09-30 15:31 /user/hadoop/w.data

```
Ankan Mazumdar@DESKTOP-CMULEBA MINGW64 ~/Downloads/Big data CSP 554 HW3
$ scp -i emr_key_pair.pem WordCount.py hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop
Warning: Identity file emr_key_pair.pem not accessible: No such file or directory.
hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com: Permission denied (publickey,gssapi-keyex,gssapi-with-mic).
scp: Connection closed

Ankan Mazumdar@DESKTOP-CMULEBA MINGW64 ~/Downloads/Big data CSP 554 HW3
$ scp -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem WordCount.py hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop
WordCount.py                                                                                          100%  402     7.9KB/s   00:00

Ankan Mazumdar@DESKTOP-CMULEBA MINGW64 ~/Downloads/Big data CSP 554 HW3
$ scp -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem w.data hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop
w.data                                                                                                100%  528    13.9KB/s   00:00
```

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ hdfs dfs -copyFromLocal -f /home/hadoop/w.data /user/hadoop/
(myenv) [hadoop@ip-172-31-35-66 ~]$ hdfs dfs -ls  /user/hadoop/
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup        528 2023-09-30 15:31 /user/hadoop/w.data
```

Step 4: Now execute the following python WordCount.py -r hadoop hdfs:///user/hadoop/w.data Note there must be three slashes in “hdfs:///” as “hdfs://” indicates that the file you are reading from is in the hadoop file system and the “/user” is the first part of the path to that file. Also note that sometimes copying and pasting this command from the assignment document does not work and it needs to be entered manually. Check that it produces some reasonable output. Note, the above command will erase all output files in hdfs. If you want to keep the output use the following command instead: python WordCount.py -r hadoop hdfs:///user/hadoop/w.data --output-dir /user/hadoop/some-nonexistent-directory.

Command executed-
(myenv) [hadoop@ip-172-31-35-66 ~]$ python WordCount.py -r hadoop hdfs:///user/hadoop/w.data

Screenshot-

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ python WordCount.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 3.3.3
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/WordCount.hadoop.20230930.153216.559646
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230930.153216.559646/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230930.153216.559646/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-5.jar] /tmp/streamjob4194686514048133592.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-35-66.us-east-2.compute.internal/172.31.35.66:8032
  Connecting to Application History server at ip-172-31-35-66.us-east-2.compute.internal/172.31.35.66:10200
  Connecting to ResourceManager at ip-172-31-35-66.us-east-2.compute.internal/172.31.35.66:8032
  Connecting to Application History server at ip-172-31-35-66.us-east-2.compute.internal/172.31.35.66:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1696086659197_0001
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1696086659197_0001
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1696086659197_0001
  The url to track the job: http://ip-172-31-35-66.us-east-2.compute.internal:20888/proxy/application_1696086659197_0001/
  Running job: job_1696086659197_0001
  Job job_1696086659197_0001 running in uber mode : false
   map 0% reduce 0%
   map 13% reduce 0%
   map 25% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1696086659197_0001 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230930.153216.559646/output
Counters: 56
        File Input Format Counters
                Bytes Read=2376
        File Output Format Counters
                Bytes Written=652
        File System Counters
                FILE: Number of bytes read=751
                FILE: Number of bytes written=3255301
```

```
"individual"    1
"mrjob" 1
"on"    4
"program"       1
"run"   1
"runners"       1
"second"        1
"see"   1
"submitted"     1
"things"        1
"those" 1
"to"    3
"uploaded"      1
"when"  1
"will"  1
"writing"       2
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230930.153216.559646...
Removing temp directory /tmp/WordCount.hadoop.20230930.153216.559646...
```

6) Now slightly modify the WordCount.py program. Call the new program WordCount2.py. Instead of counting how many words there are in the input documents (w.data), modify the program to count how many words begin with the small letters a-n and how many begin with anything else.

Command executed-
Cat > WordCount2.py

screenshot-

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ cat > WordCount2.py
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].lower() >= 'a' and word[0].lower() <= '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()
^C
(myenv) [hadoop@ip-172-31-35-66 ~]$ cat  WordCount2.py
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].lower() >= 'a' and word[0].lower() <= '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()
```

7) (5 points) Submit a copy of this modified program and a screen shot of the results of the program's execution as the output of your assignment.

Command executed-

(myenv) [hadoop@ip-172-31-35-66 ~]$ python WordCount2.py -r hadoop hdfs:///user/hadoop/w.data

Screenshot-

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ 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
```

Output-

"a_to_n" 49
"other" 46

```
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230930.155116.197106/output...
"a_to_n"        49
"other" 46
```

8) Now do the same as the above for the files Salaries.py and Salaries.tsv. The ".tsv" file holds department and salary information for Baltimore municipal workers. Have a look at Salaries.py for the layout of the ".tsv" file and how to read it in to our map reduce program.

Command executed-
scp -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem Salaries.py hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop Salaries.py

$ scp -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem Salaries.tsv hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop Salaries.tsv

Screenshot-

```
Ankan Mazumdar@DESKTOP-CMULEBA MINGW64 ~/Downloads/Big data CSP 554 HW3
$ scp -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem Salaries.tsv Salaries.py u.data hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop

Ankan Mazumdar@DESKTOP-CMULEBA MINGW64 ~/Downloads/Big data CSP 554 HW3
$ scp -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem Salaries.py  hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.py                                                                                                   100%  411     1.9KB/s   00:00

Ankan Mazumdar@DESKTOP-CMULEBA MINGW64 ~/Downloads/Big data CSP 554 HW3
$ scp -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem Salaries.tsv  hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.tsv                                                                                                  100% 1502KB  36.0KB/s   00:41
```

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ hdfs dfs -copyFromLocal -f /home/hadoop/Salaries.tsv /user/hadoop/
```

9) Execute the Salaries.py program to make sure it works. It should print out how many workers share each job title.

Command executed-

(myenv) [hadoop@ip-172-31-35-66 ~]$ hdfs dfs -copyFromLocal -f /home/hadoop/Salaries.tsv /user/hadoop/
screenshot-

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ python Salaries.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 3.3.3
Looking for Hadoop streaming jar in /home/hadoop/contrib
```

output-

```
job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230930.161355.969103/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230930.161355.969103/output...
"911 OPERATOR SUPERVISOR"       4
"ACCOUNT EXECUTIVE"     4
"ACCOUNTANT I"  15
"ACCOUNTANT TRAINEE"    1
"ACCOUNTING ASST I"     6
"ACCOUNTING SYSTEMS ADMINISTRAT"        3
"ADM COORDINATOR"       2
```

```
"Waste Water Tech Supv I Pump"  6
"YOUTH DEVELOPMENT TECH"        3
"ZONING ADMINISTRATOR"  1
"ZONING APPEALS ADVISOR BMZA"   1
"ZONING APPEALS OFFICER"        1
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230930.161355.969103...
Removing temp directory /tmp/Salaries.hadoop.20230930.161355.969103...
```

10) Now modify the Salaries.py program. Call it Salaries2.py Instead of counting the number of workers per department, change the program to provide the number of workers having High, Medium or Low annual salaries. execute the program and see what happens.

Command executed-

(myenv) [hadoop@ip-172-31-35-66 ~]$ cat > Salaries2.py

```
from mrjob.job import MRJob

class MRWorkerSalaries(MRJob):

    def configure_args(self):
        super(MRWorkerSalaries, self).configure_args()
        self.add_passthru_arg('--salary-thresholds', default='100000,50000',
                              help='Comma-separated salary thresholds for High, Medium, and Low')

    def mapper(self, _, line):
        (name, jobTitle, agencyID, agency, hireDate, annualSalary, grossPay) = line.split('\t')
        annual_salary = float(annualSalary)
```

```
        salary_thresholds = [float(threshold) for threshold in self.options.salary_thresholds.split(',')]
        if annual_salary >= salary_thresholds[0]:
            yield 'High', 1
        elif annual_salary >= salary_thresholds[1]:
            yield 'Medium', 1
        else:
            yield 'Low', 1

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

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


if __name__ == '__main__':
    MRWorkerSalaries.run()
^C
(myenv) [hadoop@ip-172-31-35-66 ~]$ cat Salaries2.py
from mrjob.job import MRJob

class MRWorkerSalaries(MRJob):

    def configure_args(self):
        super(MRWorkerSalaries, self).configure_args()
        self.add_passthru_arg('--salary-thresholds', default='100000,50000',
                              help='Comma-separated salary thresholds for High, Medium, and Low')

    def mapper(self, _, line):
        (name, jobTitle, agencyID, agency, hireDate, annualSalary, grossPay) = line.split('\t')
        annual_salary = float(annualSalary)

        salary_thresholds = [float(threshold) for threshold in self.options.salary_thresholds.split(',')]
        if annual_salary >= salary_thresholds[0]:
            yield 'High', 1
        elif annual_salary >= salary_thresholds[1]:
            yield 'Medium', 1
        else:
            yield 'Low', 1

    def combiner(self, salary_category, counts):
        yield salary_category, sum(counts)
```

```
    def reducer(self, salary_category, counts):
        yield salary_category, sum(counts)


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

Screenshot-

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ cat > Salaries2.py
from mrjob.job import MRJob

class MRWorkerSalaries(MRJob):

    def configure_args(self):
        super(MRWorkerSalaries, self).configure_args()
        self.add_passthru_arg('--salary-thresholds', default='100000,50000',
                              help='Comma-separated salary thresholds for High, Medium, and Low')

    def mapper(self, _, line):
        (name, jobTitle, agencyID, agency, hireDate, annualSalary, grossPay) = line.split('\t')
        annual_salary = float(annualSalary)

        salary_thresholds = [float(threshold) for threshold in self.options.salary_thresholds.split(',')]
        if annual_salary >= salary_thresholds[0]:
            yield 'High', 1
        elif annual_salary >= salary_thresholds[1]:
            yield 'Medium', 1
        else:
            yield 'Low', 1

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

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


if __name__ == '__main__':
    MRWorkerSalaries.run()
^C
(myenv) [hadoop@ip-172-31-35-66 ~]$ cat Salaries2.py
from mrjob.job import MRJob

class MRWorkerSalaries(MRJob):

    def configure_args(self):
        super(MRWorkerSalaries, self).configure_args()
        self.add_passthru_arg('--salary-thresholds', default='100000,50000',
                              help='Comma-separated salary thresholds for High, Medium, and Low')

    def mapper(self, _, line):
        (name, jobTitle, agencyID, agency, hireDate, annualSalary, grossPay) = line.split('\t')
        annual_salary = float(annualSalary)

        salary_thresholds = [float(threshold) for threshold in self.options.salary_thresholds.split(',')]
        if annual_salary >= salary_thresholds[0]:
            yield 'High', 1
        elif annual_salary >= salary_thresholds[1]:
            yield 'Medium', 1
        else:
```

```
            yield 'Low', 1

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

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


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

11) (5 points) Submit a copy of this modified program and a screenshot of the results of the program's execution as the output of your assignment.

Command executed-
(myenv) [hadoop@ip-172-31-35-66 ~]$ python Salaries2.py -r hadoop hdfs:///user/hadoop/Salaries.tsv

screenshot-

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ 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 3.3.3
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.20230930.163208.658414
```

Output-

"High" 442
"Low" 7064
"Medium" 6312

```
        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.20230930.163208.658414/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230930.163208.658414/output...
"High"  442
"Low"   7064
"Medium"        6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230930.163208.658414...
Removing temp directory /tmp/Salaries2.hadoop.20230930.163208.658414...
```

12) Now copy the file u.data from the assignment to /user/hadoop. NOTE: this version of u.data has fields separated by commas and not tabs.
$ scp -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem u.data hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop u.data

Command executed-
(myenv) [hadoop@ip-172-31-35-66 ~]$ hdfs dfs -copyFromLocal -f /home/hadoop/u.data /user/hadoop/

screenshot-

```
Ankan Mazumdar@DESKTOP-CMULEBA MINGW64 ~/Downloads/Big data CSP 554 HW3
$ scp -i /c/Users/Ankan\ Mazumdar/Downloads/emr_key_pair.pem u.data  hadoop@ec2-18-191-187-150.us-east-2.compute.amazonaws.com:/home/hadoop
u.data                                                                                                    100% 2381KB 169.4KB/s   00:14

Ankan Mazumdar@DESKTOP-CMULEBA MINGW64 ~/Downloads/Big data CSP 554 HW3
$
```

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ hdfs dfs -copyFromLocal -f /home/hadoop/u.data /user/hadoop/
```

13) (5 points) Write a program to perform the task of outputting a count of the number of movies each user (identified via their user id) reviewed.

Command executed-

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ cat > MovieReviewCount.py
from mrjob.job import MRJob

class MRUserMovieCount(MRJob):

    def configure_args(self):
        super(MRUserMovieCount, self).configure_args()
        self.add_passthru_arg('--input-file', default='u.data',
                              help='Input CSV file with user reviews')

    def mapper(self, _, line):
        # Split the CSV line into user_id and movie_id
        user_id, movie_id, rating, timestamp = line.split(',')

        # Emit the user_id as the key and a count of 1 as the value
        yield user_id, 1

    def reducer(self, user_id, review_counts):
        # Sum the counts to get the total number of movies reviewed by the user
        total_reviews = sum(review_counts)

        # Emit the user_id and the total number of reviews
        yield user_id, total_reviews

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

screenshot-

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ cat > MovieReviewCount.py
from mrjob.job import MRJob

class MRUserMovieCount(MRJob):

    def configure_args(self):
        super(MRUserMovieCount, self).configure_args()
        self.add_passthru_arg('--input-file', default='u.data',
                              help='Input CSV file with user reviews')

    def mapper(self, _, line):
        # Split the CSV line into user_id and movie_id
        user_id, movie_id, rating, timestamp = line.split(',')

        # Emit the user_id as the key and a count of 1 as the value
        yield user_id, 1

    def reducer(self, user_id, review_counts):
        # Sum the counts to get the total number of movies reviewed by the user
        total_reviews = sum(review_counts)

        # Emit the user_id and the total number of reviews
        yield user_id, total_reviews

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

Command executed-

Screenshot-
(myenv) [hadoop@ip-172-31-35-66 ~]$ hdfs dfs -copyFromLocal -f /home/hadoop/u.data /user/hadoop/
(myenv) [hadoop@ip-172-31-35-66 ~]$ python MovieReviewCount.py -r hadoop hdfs:///user/hadoop/u.data

```
(myenv) [hadoop@ip-172-31-35-66 ~]$ hdfs dfs -copyFromLocal -f /home/hadoop/u.data /user/hadoop/
(myenv) [hadoop@ip-172-31-35-66 ~]$ python MovieReviewCount.py -r hadoop hdfs:///user/hadoop/u.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 3.3.3
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/MovieReviewCount.hadoop.20230930.170527.631909
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/MovieReviewCount.hadoop.20230930.170527.631909/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/MovieReviewCount.hadoop.20230930.170527.631909/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-5.jar] /tmp/streamjob7654214972817850174.jar tmpDir=null
```

Output-

```
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/MovieReviewCount.hadoop.20230930.170527.631909/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/MovieReviewCount.hadoop.20230930.170527.631909/output...
"102"   678
"105"   525
"108"   31
"111"   341
"114"   25
"117"   55
"12"    61
"120"   138
"123"   33
"126"   64
"129"   26
"132"   94
"135"   22
"138"   81
"141"   31
"144"   41
"147"   38
"15"    1700
"150"   413
"153"   51
"156"   45
"159"   148
"162"   30
"165"   487
"168"   116
"171"   48
"174"   21
"177"   224
"18"    51
"180"   24
"183"   41
"186"   42
```

```
"80"    37
"83"    161
"86"    190
"89"    66
"92"    123
"95"    299
"98"    71
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/MovieReviewCount.hadoop.20230930.170527.631909...
Removing temp directory /tmp/MovieReviewCount.hadoop.20230930.170527.631909...
(myenv) [hadoop@ip-172-31-35-66 ~]$
```