CSP554—Big Data Technologies

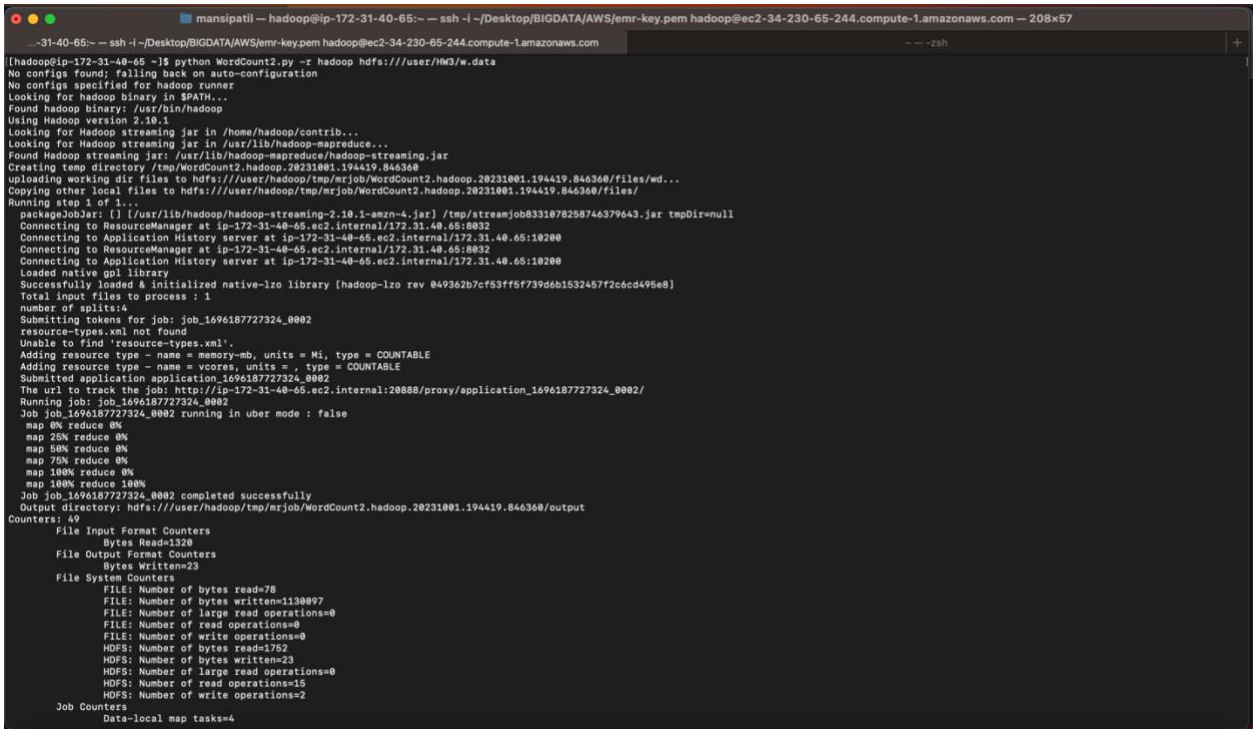
Assignment #3 (Modules 03a & 03b, 15 points)

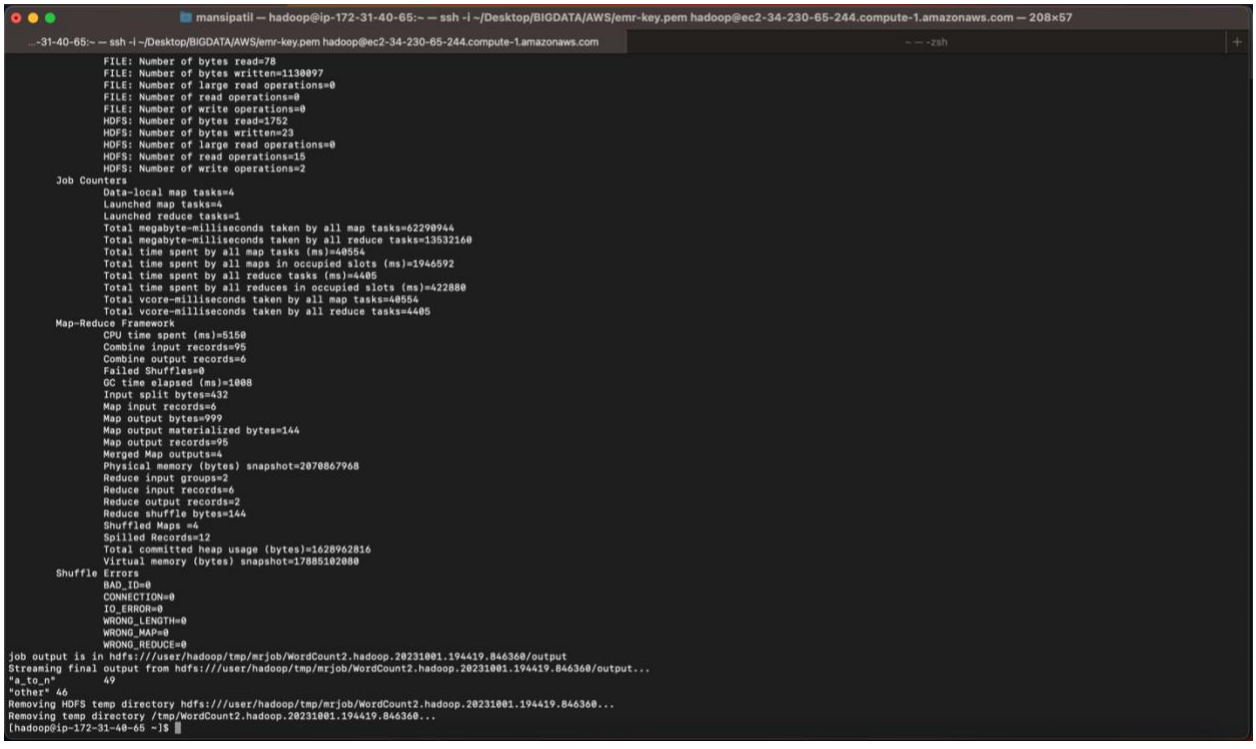
7) (5 points) Submit a copy of this modified program and a screen shot of the results of the program's execution as the output of your assignment.
ANS:

Program File Name – **WordCount2.py**

```
    from mrjob.job import MRJob
import re
WORD_RE = re.compile(r"[\w']+")
class Mansi_Wordcount(MRJob):
def configure_args(self):
    super(Mansi_Wordcount, self).configure_args()
    self.add_passthru_arg('--output-format', type=str, default='wordcount2')
def mapper(self, _, line):
    for word in WORD_RE.findall(line):
      if word.lower()[0] >= 'a' and word.lower()[0] <= 'n':
        yield 'a_to_n', 1
      else:
        yield 'other', 1
  def combiner(self, key, counts):
    yield key, sum(counts)
  def reducer(self, key, counts):
    yield key, sum(counts)
if __name__ == '__main__':
  Mansi_Wordcount.run()
```

OUTPUT:


mansipatil — hadoop@ip-172-31-40-65:~ — ssh -i ~/Desktop/BIGDATA/AWS/emr-key.pem hadoop@ec2-34-230-65-244.compute-1.amazonaws.com — 208×57
...-31-40-65:~ — ssh -i ~/Desktop/BIGDATA/AWS/emr-key.pem hadoop@ec2-34-230-65-244.compute-1.amazonaws.com
~ — -zsh
[hadoop@ip-172-31-40-65 ~]$ python WordCount2.py -r hadoop hdfs:///user/HW3/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.20231001.194419.846360
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20231001.194419.846360/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20231001.194419.846360/files/
Running step 1 of 1...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar] /tmp/streamjob8331078258746379643.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-40-65.ec2.internal/172.31.40.65:8032
Connecting to Application History server at ip-172-31-40-65.ec2.internal/172.31.40.65:10200
Connecting to ResourceManager at ip-172-31-40-65.ec2.internal/172.31.40.65:8032
Connecting to Application History server at ip-172-31-40-65.ec2.internal/172.31.40.65:10200
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
Total input files to process : 1
number of splits:4
Submitting tokens for job: job_1696187727324_0002
resource-types.xml not found
Unable to find 'resource-types.xml'.
Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
Adding resource type - name = vcores, units = , type = COUNTABLE
Submitted application application_1696187727324_0002
The url to track the job: http://ip-172-31-40-65.ec2.internal:20888/proxy/application_1696187727324_0002/
Running job: job_1696187727324_0002
Job job_1696187727324_0002 running in uber mode : false
map 0% reduce 0%
map 25% reduce 0%
map 50% reduce 0%
map 75% reduce 0%
map 100% reduce 0%
map 100% reduce 100%
Job job_1696187727324_0002 completed successfully
Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20231001.194419.846360/output
Counters: 49
File Input Format Counters
Bytes Read=1320
File Output Format Counters
Bytes Written=23
File System Counters
FILE: Number of bytes read=78
FILE: Number of bytes written=1130097
FILE: Number of large read operations=0
FILE: Number of read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1752
HDFS: Number of bytes written=23
HDFS: Number of large read operations=0
HDFS: Number of read operations=15
HDFS: Number of write operations=2
Job Counters
Data-local map tasks=4

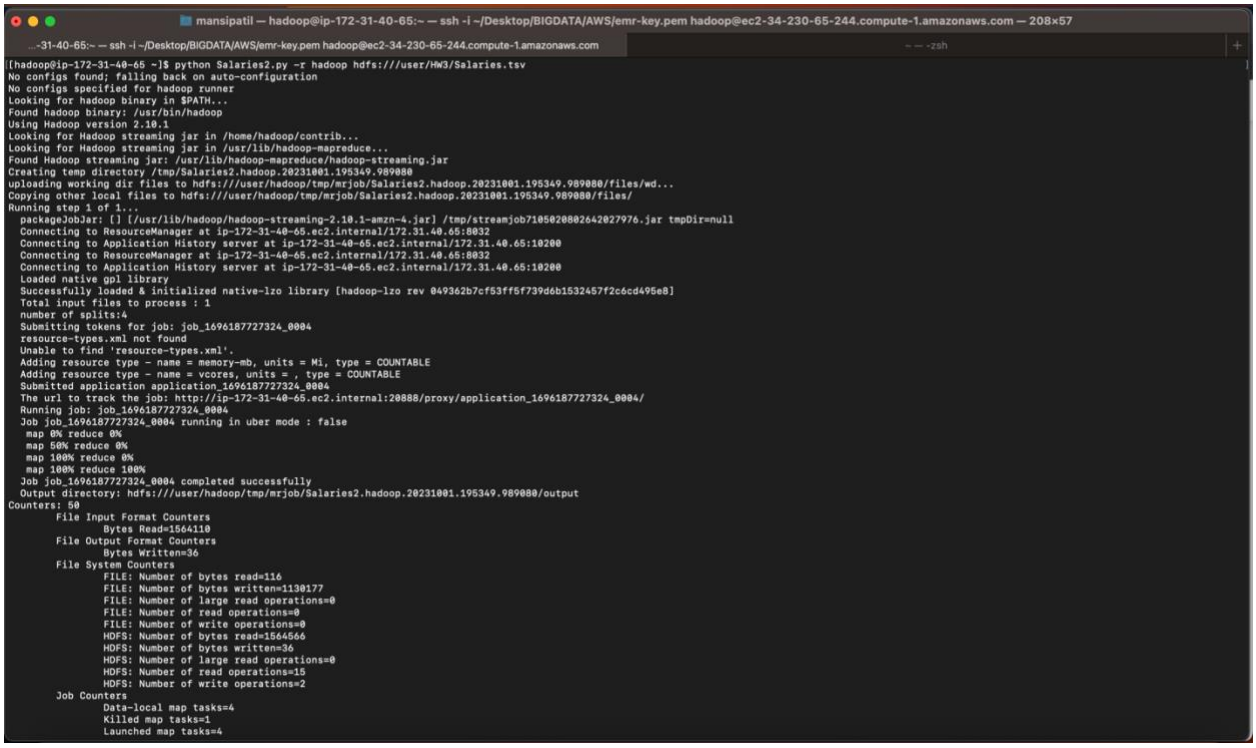

mansipatil — hadoop@ip-172-31-40-65:~ — ssh -i ~/Desktop/BIGDATA/AWS/emr-key.pem hadoop@ec2-34-230-65-244.compute-1.amazonaws.com — 208×57
...-31-40-65:~ — ssh -i ~/Desktop/BIGDATA/AWS/emr-key.pem hadoop@ec2-34-230-65-244.compute-1.amazonaws.com
~ — -zsh
FILE: Number of bytes read=78
FILE: Number of bytes written=1130097
FILE: Number of large read operations=0
FILE: Number of read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1752
HDFS: Number of bytes written=23
HDFS: Number of large read operations=0
HDFS: Number of read operations=15
HDFS: Number of write operations=2
Job Counters
Data-local map tasks=4
Launched map tasks=4
Launched reduce tasks=1
Total megabyte-milliseconds taken by all map tasks=62290944
Total megabyte-milliseconds taken by all reduce tasks=13532160
Total time spent by all map tasks (ms)=40554
Total time spent by all maps in occupied slots (ms)=1946592
Total time spent by all reduce tasks (ms)=4405
Total time spent by all reduces in occupied slots (ms)=422880
Total vcore-milliseconds taken by all map tasks=40554
Total vcore-milliseconds taken by all reduce tasks=4405
Map-Reduce Framework
CPU time spent (ms)=5150
Combine input records=95
Combine output records=6
Failed Shuffles=0
GC time elapsed (ms)=1008
Input split bytes=432
Map input records=6
Map output bytes=999
Map output materialized bytes=144
Map output records=95
Merged Map outputs=4
Physical memory (bytes) snapshot=2070867968
Reduce input groups=2
Reduce input records=6
Reduce output records=2
Reduce shuffle bytes=144
Shuffled Maps =4
Spilled Records=12
Total committed heap usage (bytes)=1628962816
Virtual memory (bytes) snapshot=17885102080
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.20231001.194419.846360/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20231001.194419.846360/output...
"a_to_n" 49
"other" 46
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20231001.194419.846360...
Removing temp directory /tmp/WordCount2.hadoop.20231001.194419.846360...
[hadoop@ip-172-31-40-65 ~]$

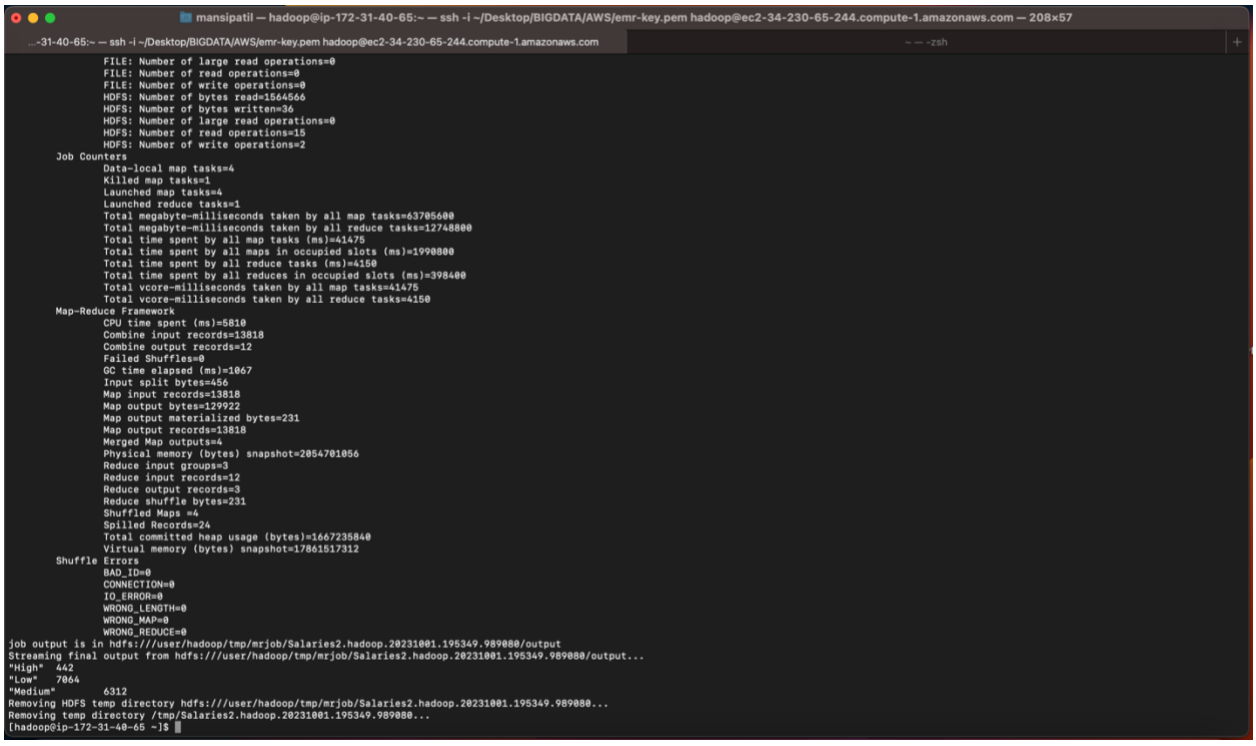
11)(5 points) Submit a copy of this modified program and a screen shot of the results of the program's execution as the output of your assignment.

ANS:

Program File Name – **Salaries2.py**

```
from mrjob.job import MRJob

class Mansi(MRJob):

    def mapper(self, _, line):

        (name, jobTitle, agencyID, agency, hireDate, annualSalary, grossPay) = line.split('\t')

        annual_salary = float(annualSalary.strip('$'))

        if annual_salary >= 100000.00:

            yield 'High', 1

        elif 50000.00 <= annual_salary <= 99999.99:

            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__':

    Mansi.run()
```


mansipatil — hadoop@ip-172-31-40-65:~ — ssh -i ~/Desktop/BIGDATA/AWS/emr-key.pem hadoop@ec2-34-230-65-244.compute-1.amazonaws.com — 208×57
...-31-40-65:~ — ssh -i ~/Desktop/BIGDATA/AWS/emr-key.pem hadoop@ec2-34-230-65-244.compute-1.amazonaws.com
~ — -zsh
[hadoop@ip-172-31-40-65 ~]$ python Salaries2.py -r hadoop hdfs:///user/HW3/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.20231001.195349.989080
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20231001.195349.989080/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20231001.195349.989080/files/
Running step 1 of 1...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar] /tmp/streamjob7105020802642027976.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-40-65.ec2.internal/172.31.40.65:8032
Connecting to Application History server at ip-172-31-40-65.ec2.internal/172.31.40.65:10200
Connecting to ResourceManager at ip-172-31-40-65.ec2.internal/172.31.40.65:8032
Connecting to Application History server at ip-172-31-40-65.ec2.internal/172.31.40.65:10200
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
Total input files to process : 1
number of splits:4
Submitting tokens for job: job_1696187727324_0004
resource-types.xml not found
Unable to find 'resource-types.xml'.
Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
Adding resource type - name = vcores, units = , type = COUNTABLE
Submitted application application_1696187727324_0004
The url to track the job: http://ip-172-31-40-65.ec2.internal:20888/proxy/application_1696187727324_0004/
Running job: job_1696187727324_0004
Job job_1696187727324_0004 running in uber mode : false
map 0% reduce 0%
map 50% reduce 0%
map 100% reduce 0%
map 100% reduce 100%
Job job_1696187727324_0004 completed successfully
Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20231001.195349.989080/output
Counters: 50
File Input Format Counters
Bytes Read=1564110
File Output Format Counters
Bytes Written=36
File System Counters
FILE: Number of bytes read=116
FILE: Number of bytes written=1130177
FILE: Number of large read operations=0
FILE: Number of read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1564566
HDFS: Number of bytes written=36
HDFS: Number of large read operations=0
HDFS: Number of read operations=15
HDFS: Number of write operations=2
Job Counters
Data-local map tasks=4
Killed map tasks=1
Launched map tasks=4


mansipatil — hadoop@ip-172-31-40-65:~ — ssh -i ~/Desktop/BIGDATA/AWS/emr-key.pem hadoop@ec2-34-230-65-244.compute-1.amazonaws.com — 208×57
...-31-40-65:~ — ssh -i ~/Desktop/BIGDATA/AWS/emr-key.pem hadoop@ec2-34-230-65-244.compute-1.amazonaws.com
~ — -zsh
FILE: Number of large read operations=0
FILE: Number of read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1564566
HDFS: Number of bytes written=36
HDFS: Number of large read operations=0
HDFS: Number of read operations=15
HDFS: Number of write operations=2
Job Counters
Data-local map tasks=4
Killed map tasks=1
Launched map tasks=4
Launched reduce tasks=1
Total megabyte-milliseconds taken by all map tasks=63705600
Total megabyte-milliseconds taken by all reduce tasks=12748800
Total time spent by all map tasks (ms)=41475
Total time spent by all maps in occupied slots (ms)=1990800
Total time spent by all reduce tasks (ms)=4150
Total time spent by all reduces in occupied slots (ms)=398400
Total vcore-milliseconds taken by all map tasks=41475
Total vcore-milliseconds taken by all reduce tasks=4150
Map-Reduce Framework
CPU time spent (ms)=5810
Combine input records=13818
Combine output records=12
Failed Shuffles=0
GC time elapsed (ms)=1067
Input split bytes=456
Map input records=13818
Map output bytes=129922
Map output materialized bytes=231
Map output records=13818
Merged Map outputs=4
Physical memory (bytes) snapshot=2054701056
Reduce input groups=3
Reduce input records=12
Reduce output records=3
Reduce shuffle bytes=231
Shuffled Maps =4
Spilled Records=24
Total committed heap usage (bytes)=1667235840
Virtual memory (bytes) snapshot=17861517312
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.20231001.195349.989080/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20231001.195349.989080/output...
"High" 442
"Low" 7064
"Medium" 6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20231001.195349.989080...
Removing temp directory /tmp/Salaries2.hadoop.20231001.195349.989080...
[hadoop@ip-172-31-40-65 ~]$

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.

Output might look something like the following:

186: 2

192: 2

112: 1 etc.

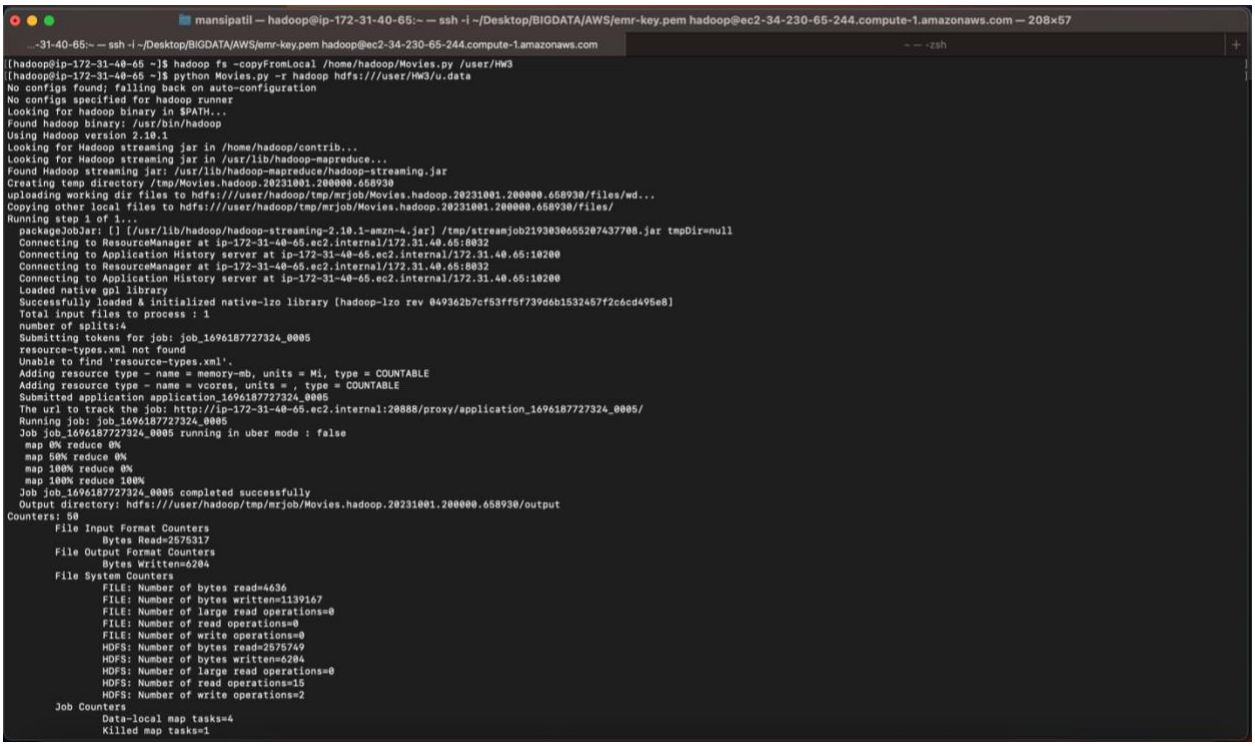
Submit a copy of this program and a screen shot of the results of the program's execution (only 10 lines or so of the result) as the output of your assignment.

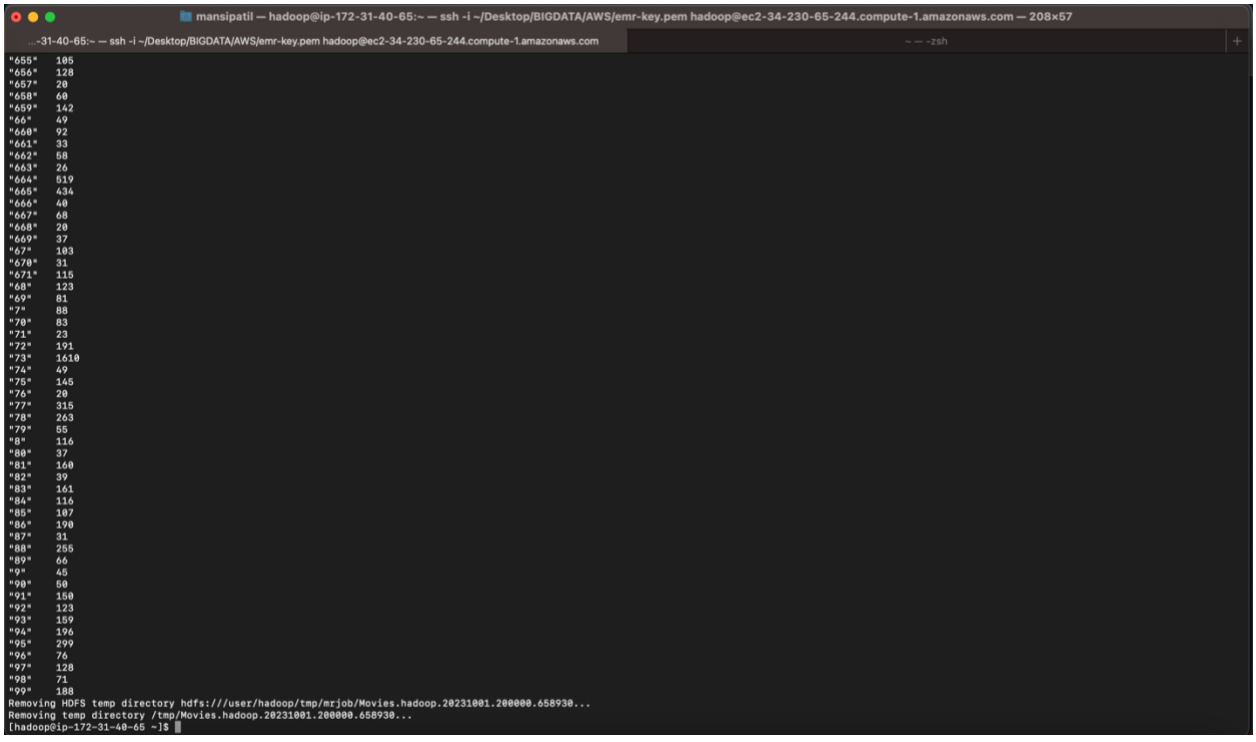
ANS:

Program File Name – Movies.py

```
from mrjob.job import MRJob

class MovieReview_Mansi(MRJob):

    def configure_args(self):

        super(MovieReview_Mansi, self).configure_args()

        self.add_passthru_arg('--output-format', type=str, default='review_count')

    def mapper(self, _, line):

        user_id, movie_id, rating, _ = line.strip().split(',')

        yield user_id, 1

    def combiner(self, user_id, counts):

        yield user_id, sum(counts)

    def reducer(self, user_id, counts):

        yield user_id, sum(counts)

if __name__ == '__main__':

    MovieReview_Mansi.run()
```

OUTPUT:


mansipatil — hadoop@ip-172-31-40-65:~ — ssh -i ~/Desktop/BIGDATA/AWS/emr-key.pem hadoop@ec2-34-230-65-244.compute-1.amazonaws.com — 208×57
[hadoop@ip-172-31-40-65 ~]$ hadoop fs -copyFromLocal /home/hadoop/Movies.py /user/HW3
[hadoop@ip-172-31-40-65 ~]$ python Movies.py -r hadoop hdfs:///user/HW3/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 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/Movies.hadoop.20231001.200000.658930
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20231001.200000.658930/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20231001.200000.658930/files/
Running step 1 of 1...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar] /tmp/streamjob2193030655207437708.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-40-65.ec2.internal/172.31.40.65:8032
Connecting to Application History server at ip-172-31-40-65.ec2.internal/172.31.40.65:10200
Connecting to ResourceManager at ip-172-31-40-65.ec2.internal/172.31.40.65:8032
Connecting to Application History server at ip-172-31-40-65.ec2.internal/172.31.40.65:10200
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
Total input files to process : 1
number of splits:4
Submitting tokens for job: job_1696187727324_0005
resource-types.xml not found
Unable to find 'resource-types.xml'.
Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
Adding resource type - name = vcores, units = , type = COUNTABLE
Submitted application application_1696187727324_0005
The url to track the job: http://ip-172-31-40-65.ec2.internal:20888/proxy/application_1696187727324_0005/
Running job: job_1696187727324_0005
Job job_1696187727324_0005 running in uber mode : false
map 0% reduce 0%
map 50% reduce 0%
map 100% reduce 0%
map 100% reduce 100%
Job job_1696187727324_0005 completed successfully
Output directory: hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20231001.200000.658930/output
Counters: 50
File Input Format Counters
Bytes Read=2675317
File Output Format Counters
Bytes Written=6204
File System Counters
FILE: Number of bytes read=4636
FILE: Number of bytes written=1139167
FILE: Number of large read operations=0
FILE: Number of read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=2575749
HDFS: Number of bytes written=6204
HDFS: Number of large read operations=0
HDFS: Number of read operations=15
HDFS: Number of write operations=2
Job Counters
Data-local map tasks=4
Killed map tasks=1


mansipatil — hadoop@ip-172-31-40-65:~ — ssh -i ~/Desktop/BIGDATA/AWS/emr-key.pem hadoop@ec2-34-230-65-244.compute-1.amazonaws.com — 208×57
"655" 105
"656" 128
"657" 20
"658" 60
"659" 142
"66" 49
"660" 92
"661" 33
"662" 58
"663" 26
"664" 519
"665" 434
"666" 40
"667" 68
"668" 20
"669" 37
"67" 103
"670" 31
"671" 115
"68" 123
"69" 81
"7" 88
"70" 83
"71" 23
"72" 191
"73" 1610
"74" 49
"75" 145
"76" 20
"77" 315
"78" 263
"79" 55
"8" 116
"80" 37
"81" 160
"82" 39
"83" 161
"84" 116
"85" 107
"86" 190
"87" 31
"88" 255
"89" 66
"9" 45
"90" 50
"91" 150
"92" 123
"93" 159
"94" 196
"95" 299
"96" 76
"97" 120
"98" 71
"99" 188
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20231001.200000.658930...
Removing temp directory /tmp/Movies.hadoop.20231001.200000.658930...
[hadoop@ip-172-31-40-65 ~]$