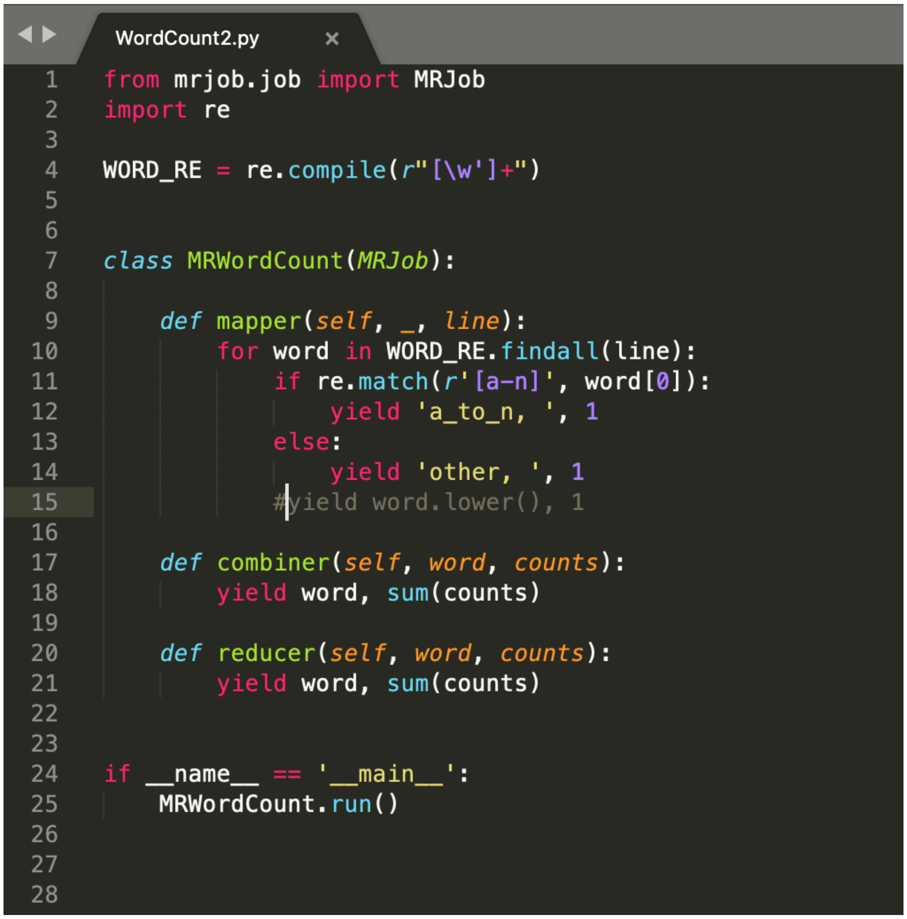
Assignment 3

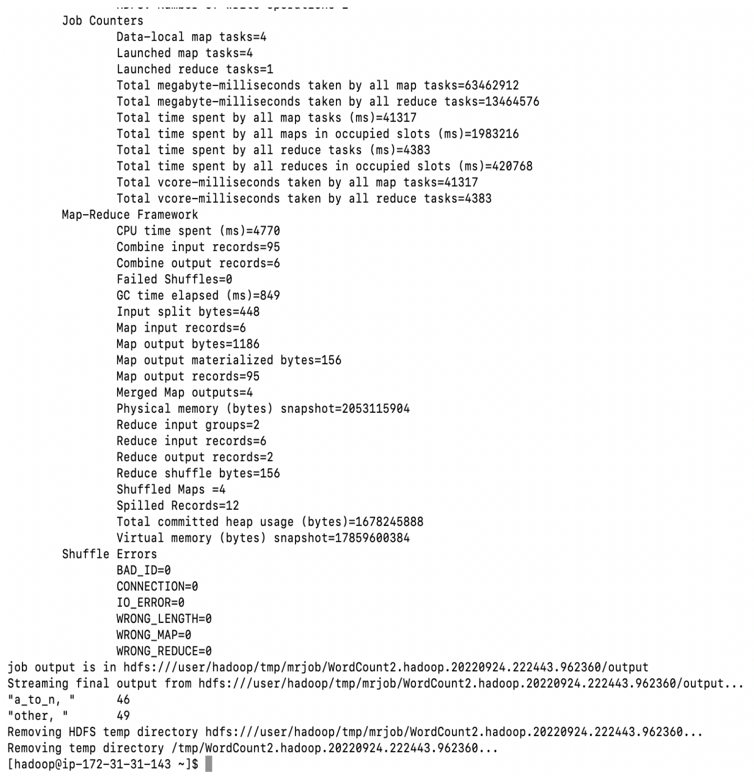
Atharva Tanaji Kadam A20467229

Question 6:

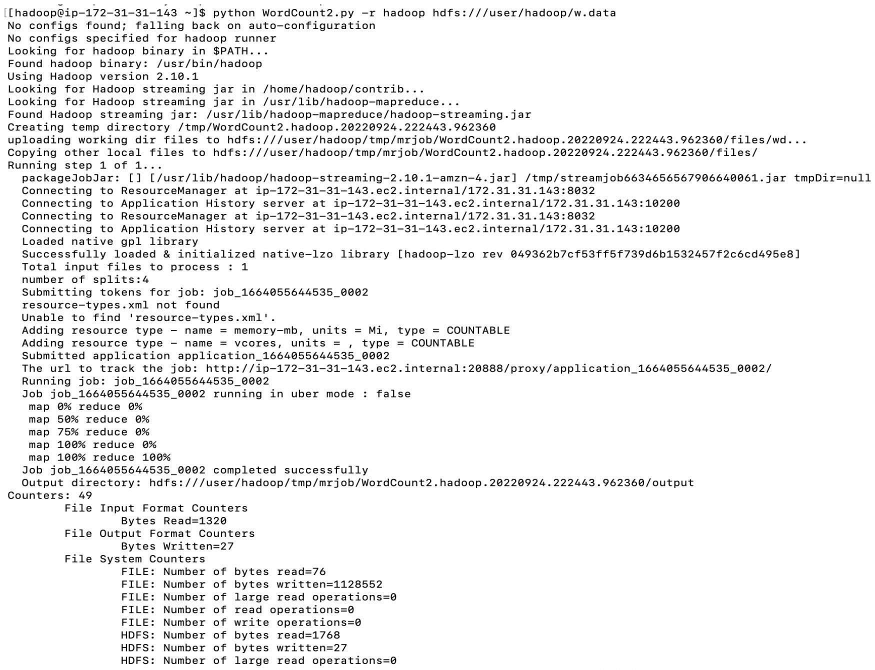


```python
from mrjob.job import MRJob
import re

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


class MRWordCount(MRJob):

    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            if re.match(r'[a-n]', word[0]):
                yield 'a_to_n, ', 1
            else:
                yield 'other, ', 1
            #yield word.lower(), 1

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

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


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

```
"a_to_n, " 46
"other, "    49
```

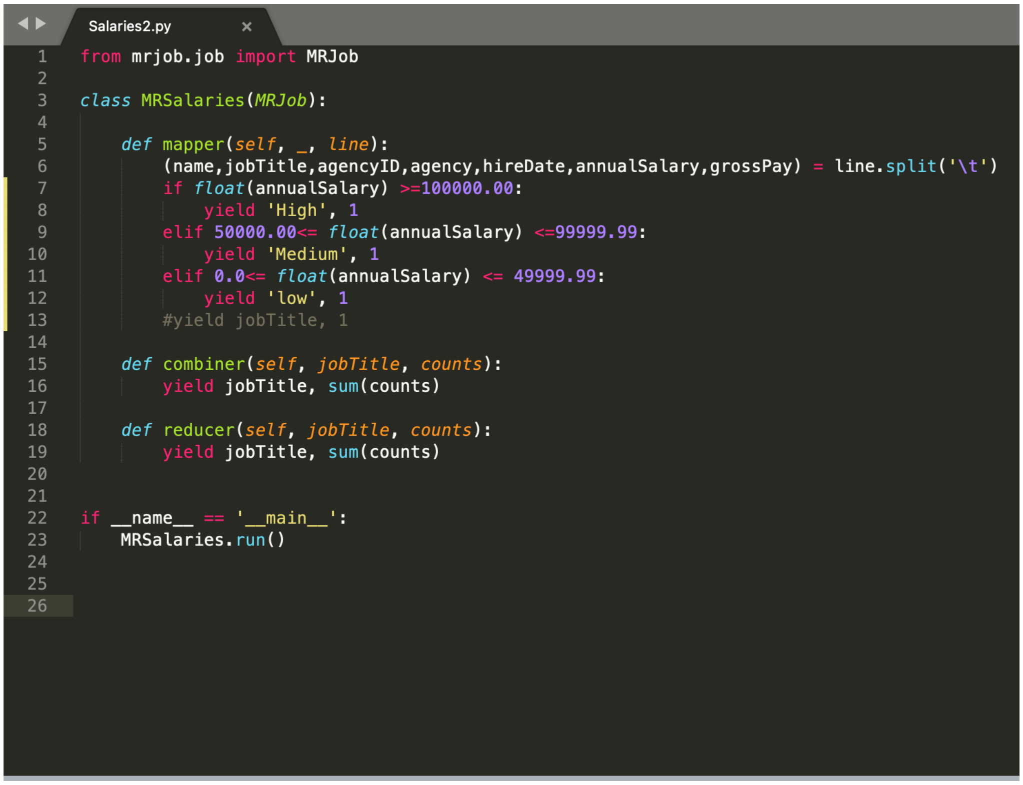


```
        Job Counters
                Data-local map tasks=4
                Launched map tasks=4
                Launched reduce tasks=1
                Total megabyte-milliseconds taken by all map tasks=63462912
                Total megabyte-milliseconds taken by all reduce tasks=13464576
                Total time spent by all map tasks (ms)=41317
                Total time spent by all maps in occupied slots (ms)=1983216
                Total time spent by all reduce tasks (ms)=4383
                Total time spent by all reduces in occupied slots (ms)=420768
                Total vcore-milliseconds taken by all map tasks=41317
                Total vcore-milliseconds taken by all reduce tasks=4383
        Map-Reduce Framework
                CPU time spent (ms)=4770
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=849
                Input split bytes=448
                Map input records=6
                Map output bytes=1186
                Map output materialized bytes=156
                Map output records=95
                Merged Map outputs=4
                Physical memory (bytes) snapshot=2053115904
                Reduce input groups=2
                Reduce input records=6
                Reduce output records=2
                Reduce shuffle bytes=156
                Shuffled Maps =4
                Spilled Records=12
                Total committed heap usage (bytes)=1678245888
                Virtual memory (bytes) snapshot=17859600384
        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.20220924.222443.962360/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220924.222443.962360/output...
"a_to_n, "      46
"other, "       49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220924.222443.962360...
Removing temp directory /tmp/WordCount2.hadoop.20220924.222443.962360...
[hadoop@ip-172-31-31-143 ~]$ 
```



```
[hadoop@ip-172-31-31-143 ~]$ 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.20220924.222443.962360
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220924.222443.962360/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220924.222443.962360/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar] /tmp/streamjob6634656567906640061.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-31-143.ec2.internal/172.31.31.143:8032
  Connecting to Application History server at ip-172-31-31-143.ec2.internal/172.31.31.143:10200
  Connecting to ResourceManager at ip-172-31-31-143.ec2.internal/172.31.31.143:8032
  Connecting to Application History server at ip-172-31-31-143.ec2.internal/172.31.31.143:10200
  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_1664055644535_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_1664055644535_0002
  The url to track the job: http://ip-172-31-31-143.ec2.internal:20888/proxy/application_1664055644535_0002/
  Running job: job_1664055644535_0002
  Job job_1664055644535_0002 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1664055644535_0002 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220924.222443.962360/output
Counters: 49
        File Input Format Counters
                Bytes Read=1320
        File Output Format Counters
                Bytes Written=27
        File System Counters
                FILE: Number of bytes read=76
                FILE: Number of bytes written=1128552
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1768
                HDFS: Number of bytes written=27
                HDFS: Number of large read operations=0
```

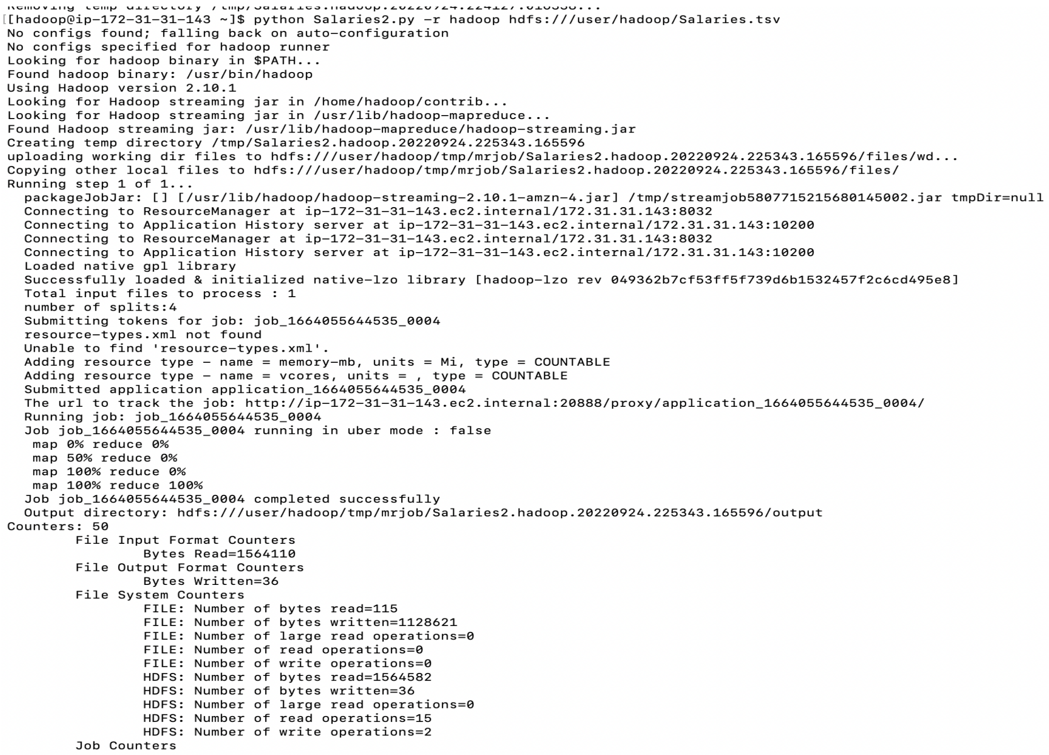
Question 10:



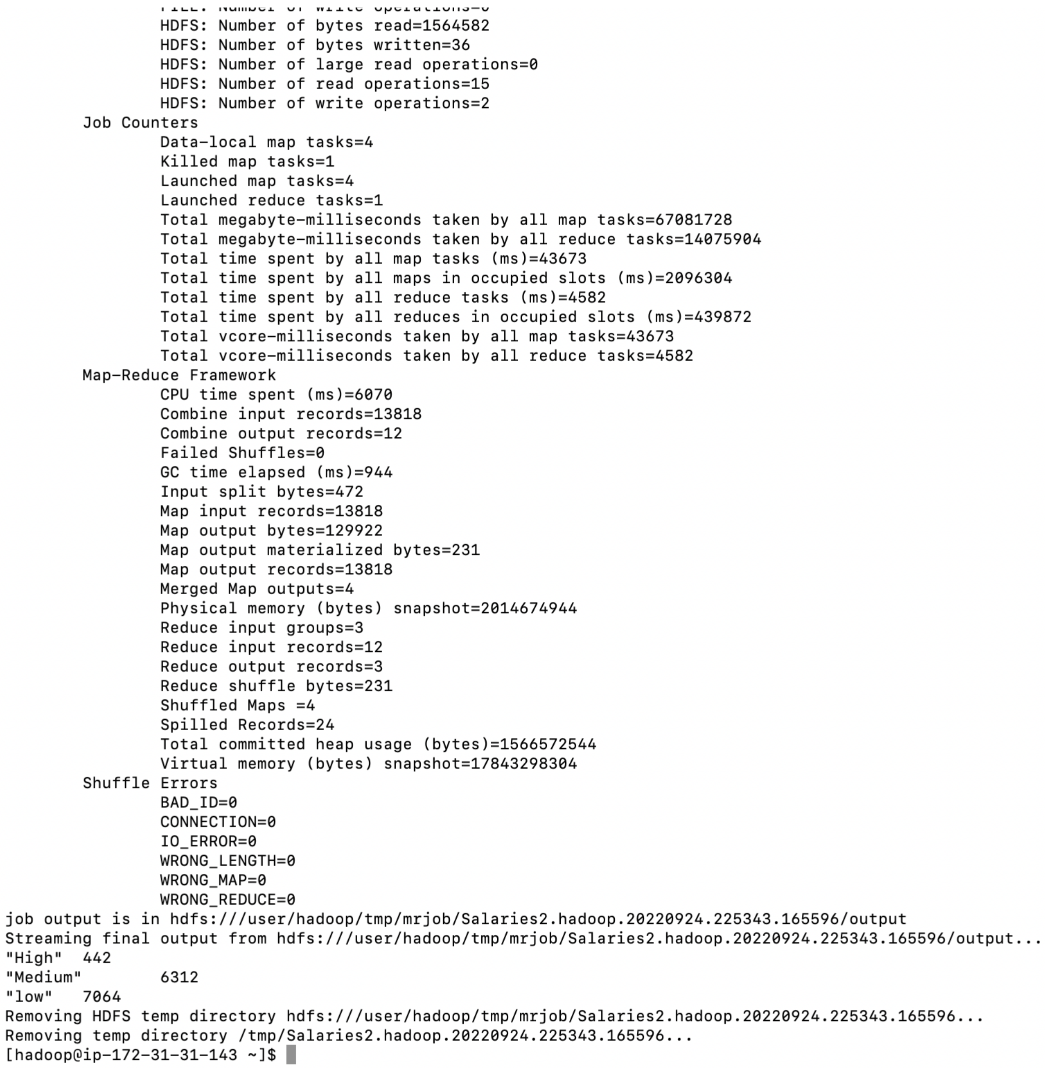
```python
from mrjob.job import MRJob

class MRSalaries(MRJob):

    def mapper(self, _, line):
        (name,jobTitle,agencyID,agency,hireDate,annualSalary,grossPay) = line.split('\t')
        if float(annualSalary) >=100000.00:
            yield 'High', 1
        elif 50000.00<= float(annualSalary) <=99999.99:
            yield 'Medium', 1
        elif 0.0<= float(annualSalary) <= 49999.99:
            yield 'low', 1
        #yield jobTitle, 1

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

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


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

Result -

```
"High"     442
"Medium”     6312
"low"      7064
```

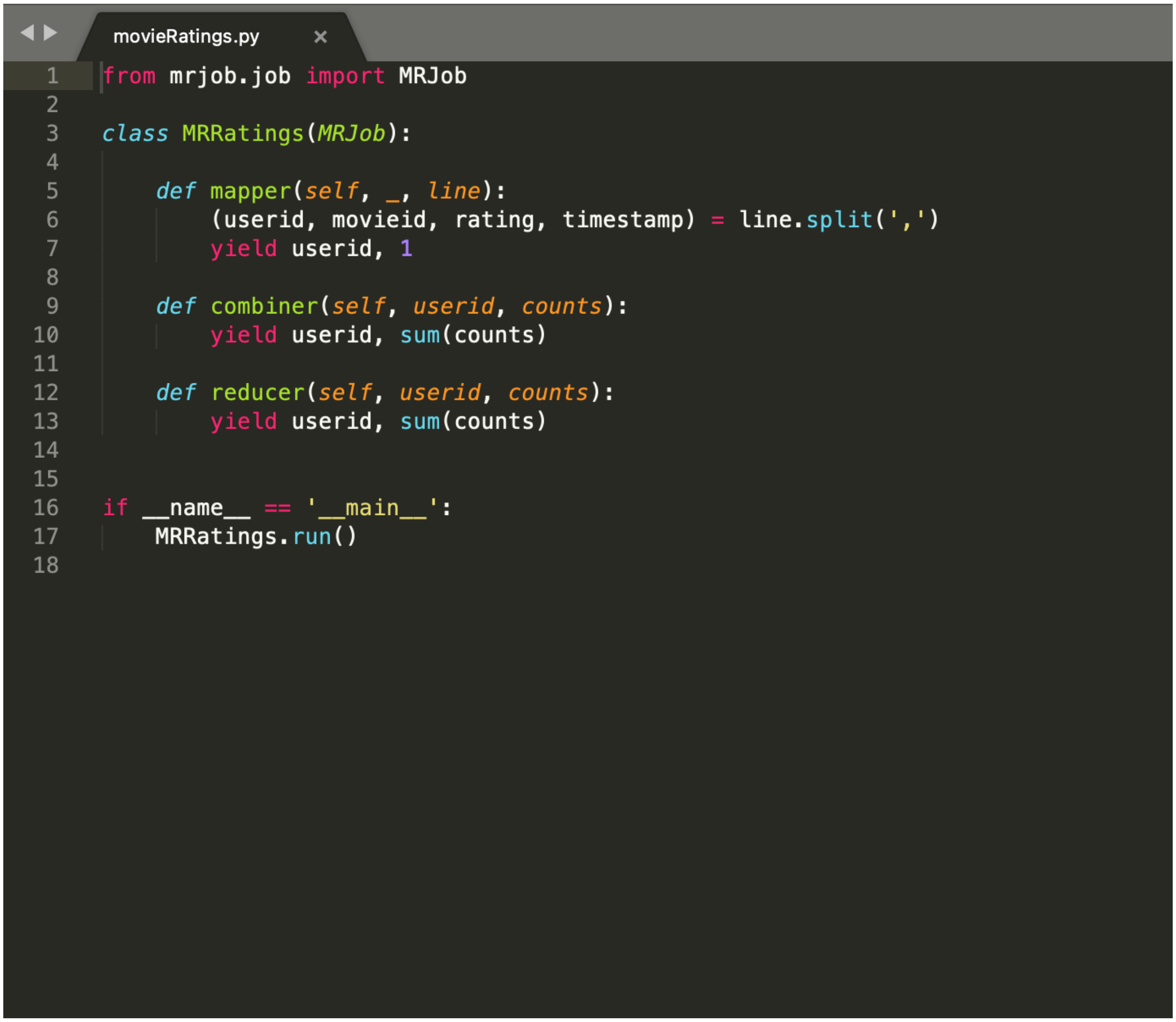


```
Removing temp directory /tmp/Salaries.hadoop.20220924.224127.610338...
[[hadoop@ip-172-31-31-143 ~]$ python Salaries2.py -r hadoop hdfs:///user/hadoop/Salaries.tsv
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found 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.20220924.225343.165596
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220924.225343.165596/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220924.225343.165596/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar] /tmp/streamjob5807715215680145002.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-31-143.ec2.internal/172.31.31.143:8032
  Connecting to Application History server at ip-172-31-31-143.ec2.internal/172.31.31.143:10200
  Connecting to ResourceManager at ip-172-31-31-143.ec2.internal/172.31.31.143:8032
  Connecting to Application History server at ip-172-31-31-143.ec2.internal/172.31.31.143:10200
  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_1664055644535_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_1664055644535_0004
  The url to track the job: http://ip-172-31-31-143.ec2.internal:20888/proxy/application_1664055644535_0004/
  Running job: job_1664055644535_0004
  Job job_1664055644535_0004 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1664055644535_0004 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220924.225343.165596/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=115
                FILE: Number of bytes written=1128621
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1564582
                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
```



```
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1564582
                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=67081728
                Total megabyte-milliseconds taken by all reduce tasks=14075904
                Total time spent by all map tasks (ms)=43673
                Total time spent by all maps in occupied slots (ms)=2096304
                Total time spent by all reduce tasks (ms)=4582
                Total time spent by all reduces in occupied slots (ms)=439872
                Total vcore-milliseconds taken by all map tasks=43673
                Total vcore-milliseconds taken by all reduce tasks=4582
        Map-Reduce Framework
                CPU time spent (ms)=6070
                Combine input records=13818
                Combine output records=12
                Failed Shuffles=0
                GC time elapsed (ms)=944
                Input split bytes=472
                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=2014674944
                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)=1566572544
                Virtual memory (bytes) snapshot=17843298304
        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.20220924.225343.165596/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220924.225343.165596/output...
"High"  442
"Medium"        6312
"low"   7064
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220924.225343.165596...
Removing temp directory /tmp/Salaries2.hadoop.20220924.225343.165596...
[hadoop@ip-172-31-31-143 ~]$ 
```

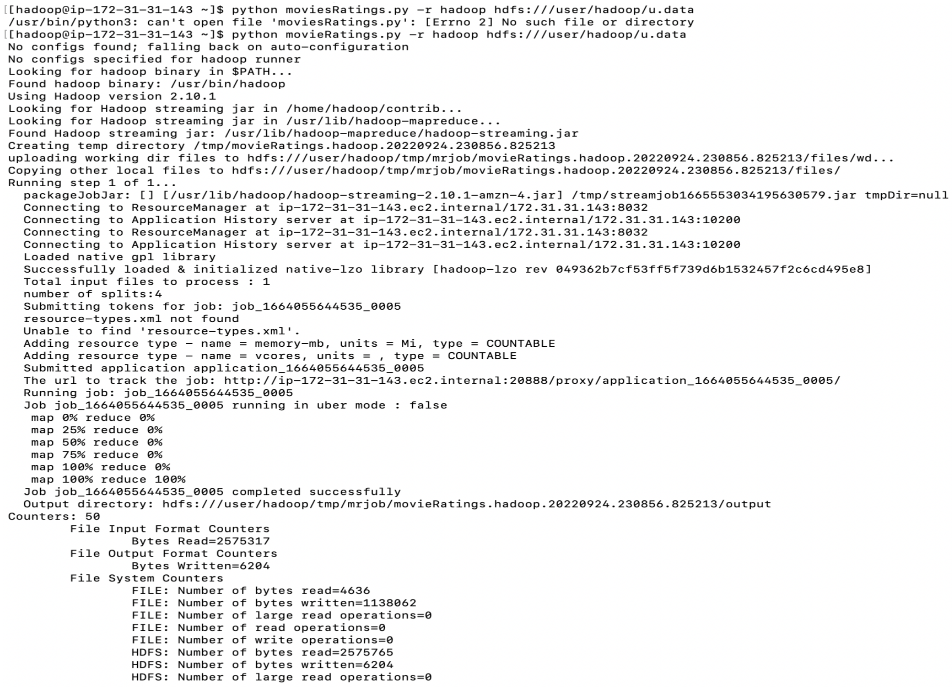
Question 12:



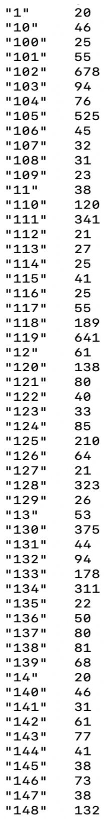
```python
from mrjob.job import MRJob

class MRRatings(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__':
    MRRatings.run()

```

Result -

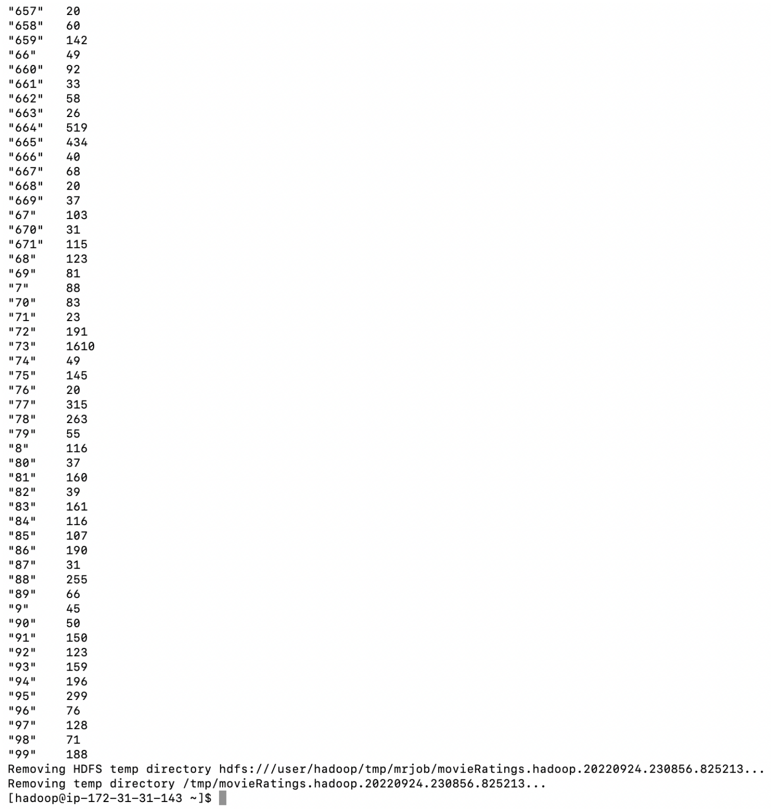
```
"97"    128
"98"    71
"99"    188
```



```
[[hadoop@ip-172-31-31-143 ~]$ python moviesRatings.py -r hadoop hdfs:///user/hadoop/u.data
 /usr/bin/python3: can't open file 'moviesRatings.py': [Errno 2] No such file or directory
[[hadoop@ip-172-31-31-143 ~]$ python movieRatings.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 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/movieRatings.hadoop.20220924.230856.825213
 uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/movieRatings.hadoop.20220924.230856.825213/files/wd...
 Copying other local files to hdfs:///user/hadoop/tmp/mrjob/movieRatings.hadoop.20220924.230856.825213/files/
 Running step 1 of 1...
   packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar] /tmp/streamjob1665553034195630579.jar tmpDir=null
   Connecting to ResourceManager at ip-172-31-31-143.ec2.internal/172.31.31.143:8032
   Connecting to Application History server at ip-172-31-31-143.ec2.internal/172.31.31.143:10200
   Connecting to ResourceManager at ip-172-31-31-143.ec2.internal/172.31.31.143:8032
   Connecting to Application History server at ip-172-31-31-143.ec2.internal/172.31.31.143:10200
   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_1664055644535_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_1664055644535_0005
   The url to track the job: http://ip-172-31-31-143.ec2.internal:20888/proxy/application_1664055644535_0005/
   Running job: job_1664055644535_0005
   Job job_1664055644535_0005 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_1664055644535_0005 completed successfully
   Output directory: hdfs:///user/hadoop/tmp/mrjob/movieRatings.hadoop.20220924.230856.825213/output
 Counters: 50
         File Input Format Counters
                 Bytes Read=2575317
         File Output Format Counters
                 Bytes Written=6204
         File System Counters
                 FILE: Number of bytes read=4636
                 FILE: Number of bytes written=1138062
                 FILE: Number of large read operations=0
                 FILE: Number of read operations=0
                 FILE: Number of write operations=0
                 HDFS: Number of bytes read=2575765
                 HDFS: Number of bytes written=6204
                 HDFS: Number of large read operations=0
```



```
"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
"139"   68
"14"    20
"140"   46
"141"   31
"142"   61
"143"   77
"144"   41
"145"   38
"146"   73
"147"   38
"148"   132
```



```
"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"	128
"98"	71
"99"	188
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/movieRatings.hadoop.20220924.230856.825213...
Removing temp directory /tmp/movieRatings.hadoop.20220924.230856.825213...
[hadoop@ip-172-31-31-143 ~]$ 
```