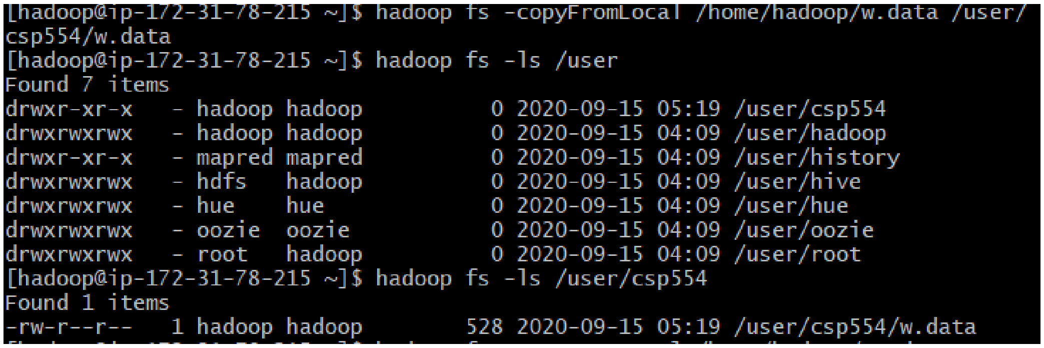
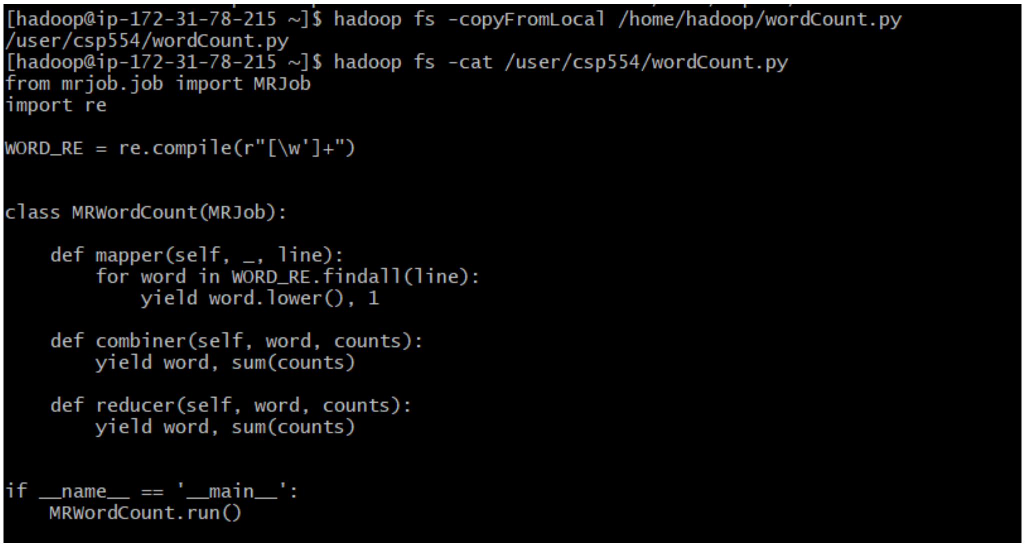
Step2:

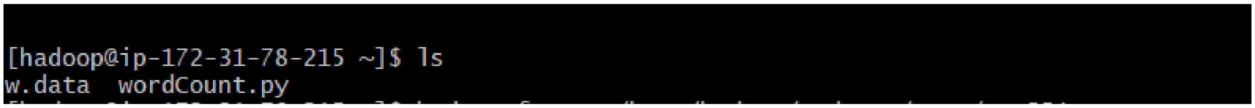


```
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -copyFromLocal /home/hadoop/w.data /user/
csp554/w.data
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -ls /user
Found 7 items
drwxr-xr-x   - hadoop hadoop          0 2020-09-15 05:19 /user/csp554
drwxrwxrwx   - hadoop hadoop          0 2020-09-15 04:09 /user/hadoop
drwxr-xr-x   - mapred mapred          0 2020-09-15 04:09 /user/history
drwxrwxrwx   - hdfs   hadoop          0 2020-09-15 04:09 /user/hive
drwxrwxrwx   - hue    hue             0 2020-09-15 04:09 /user/hue
drwxrwxrwx   - oozie  oozie           0 2020-09-15 04:09 /user/oozie
drwxrwxrwx   - root   hadoop          0 2020-09-15 04:09 /user/root
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -ls /user/csp554
Found 1 items
-rw-r--r--   1 hadoop hadoop        528 2020-09-15 05:19 /user/csp554/w.data
```

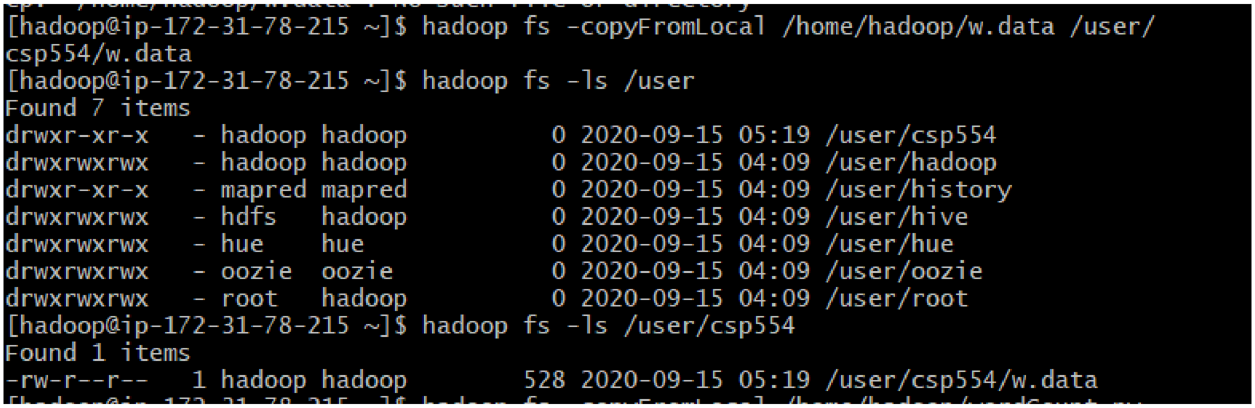


```
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -copyFromLocal /home/hadoop/wordCount.py
/user/csp554/wordCount.py
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -cat /user/csp554/wordCount.py
from mrjob.job import MRJob
import re

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


class MRWordCount(MRJob):

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

Step3:

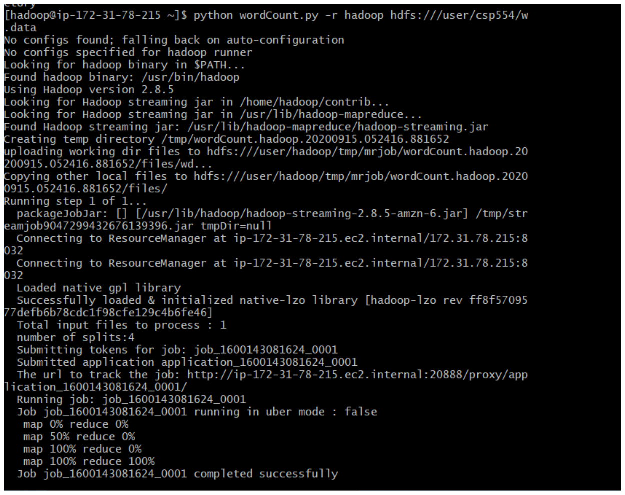


```
[hadoop@ip-172-31-78-215 ~]$ ls
w.data  wordCount.py
```

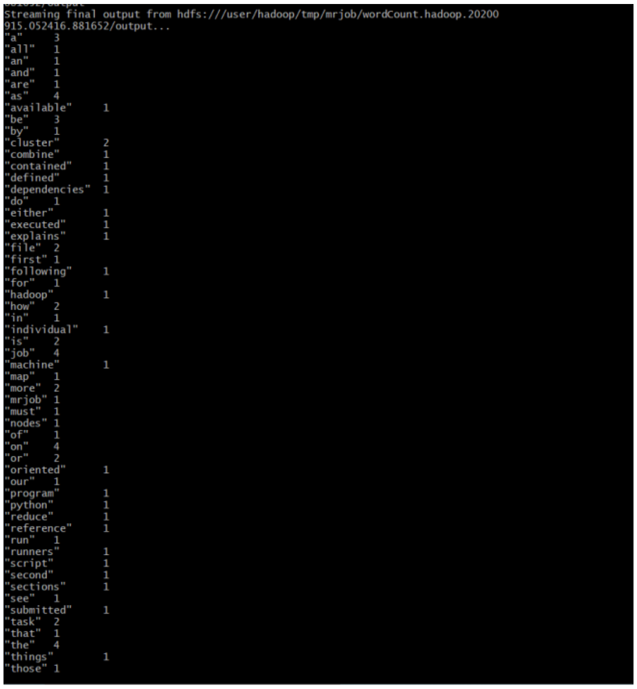


```
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -copyFromLocal /home/hadoop/w.data /user/
csp554/w.data
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -ls /user
Found 7 items
drwxr-xr-x   - hadoop hadoop          0 2020-09-15 05:19 /user/csp554
drwxrwxrwx   - hadoop hadoop          0 2020-09-15 04:09 /user/hadoop
drwxr-xr-x   - mapred mapred          0 2020-09-15 04:09 /user/history
drwxrwxrwx   - hdfs   hadoop          0 2020-09-15 04:09 /user/hive
drwxrwxrwx   - hue    hue             0 2020-09-15 04:09 /user/hue
drwxrwxrwx   - oozie  oozie           0 2020-09-15 04:09 /user/oozie
drwxrwxrwx   - root   hadoop          0 2020-09-15 04:09 /user/root
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -ls /user/csp554
Found 1 items
-rw-r--r--   1 hadoop hadoop        528 2020-09-15 05:19 /user/csp554/w.data
```

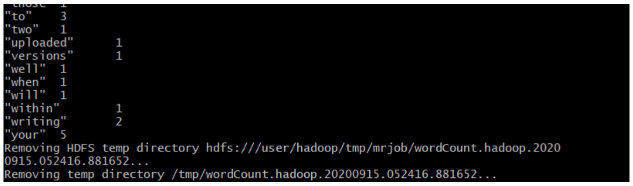
Step4:



```
[hadoop@ip-172-31-78-215 ~]$ python wordCount.py -r hadoop hdfs:///user/csp554/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 2.8.5
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.20200915.052416.881652
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/wordCount.hadoop.20
200915.052416.881652/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/wordCount.hadoop.2020
0915.052416.881652/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-6.jar] /tmp/str
eamjob9047299432676139396.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-78-215.ec2.internal/172.31.78.215:8
032
  Connecting to ResourceManager at ip-172-31-78-215.ec2.internal/172.31.78.215:8
032
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev ff8f57095
77defb6b78cdc1f98cfe129c4b6fe46]
  Total input files to process : 1
  number of splits:4
  Submitting tokens for job: job_1600143081624_0001
  Submitted application application_1600143081624_0001
  The url to track the job: http://ip-172-31-78-215.ec2.internal:20888/proxy/app
lication_1600143081624_0001/
  Running job: job_1600143081624_0001
  Job job_1600143081624_0001 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1600143081624_0001 completed successfully
```



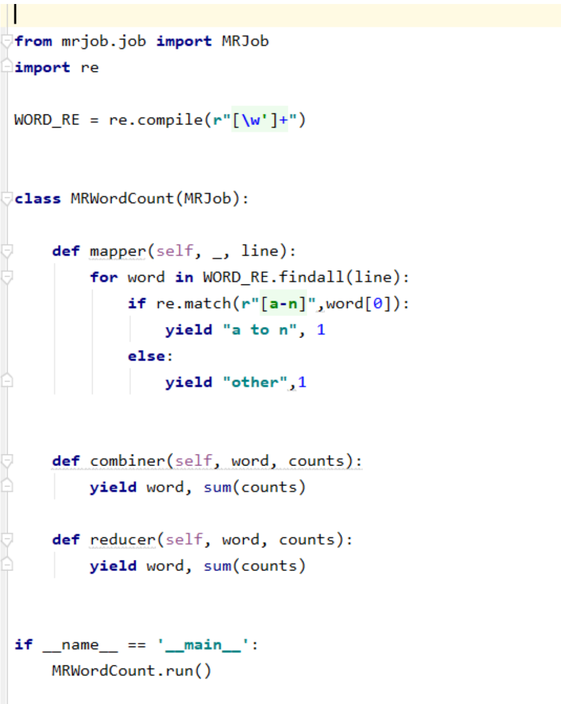
```
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/wordCount.hadoop.20200
915.052416.881652/output...
"a"     3
"all"   1
"an"    1
"and"   1
"are"   1
"as"    4
"available"     1
"be"    3
"by"    1
"cluster"       2
"combine"       1
"contained"     1
"defined"       1
"dependencies"  1
"do"    1
"either"        1
"executed"      1
"explains"      1
"file"  2
"first" 1
"following"     1
"for"   1
"hadoop"        1
"how"   2
"in"    1
"individual"    1
"is"    2
"job"   4
"machine"       1
"map"   1
"more"  2
"mrjob" 1
"must"  1
"nodes" 1
"of"    1
"on"    4
"or"    2
"oriented"      1
"our"   1
"program"       1
"python"        1
"reduce"        1
"reference"     1
"run"   1
"runners"       1
"script"        1
"second"        1
"sections"      1
"see"   1
"submitted"     1
"task"  2
"that"  1
"the"   4
"things"        1
"those" 1
```



```
"to"    3
"two"   1
"uploaded"      1
"versions"      1
"well"  1
"when"  1
"will"  1
"within"        1
"writing"       2
"your"  5
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/wordCount.hadoop.2020
0915.052416.881652...
Removing temp directory /tmp/wordCount.hadoop.20200915.052416.881652...
```

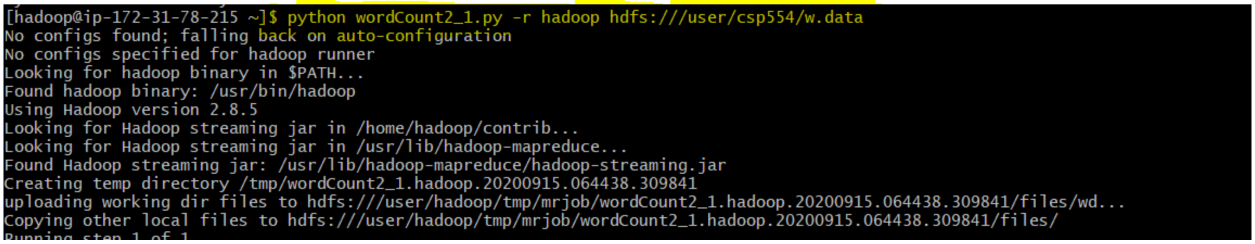
## Step 5:

WordCount2_1.py

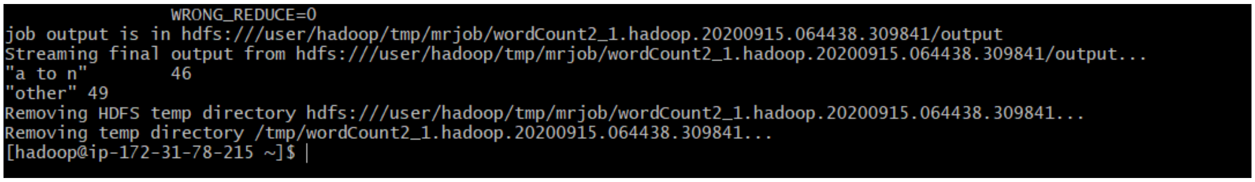


```python
from mrjob.job import MRJob
import re

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


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

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


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

## Step 6:

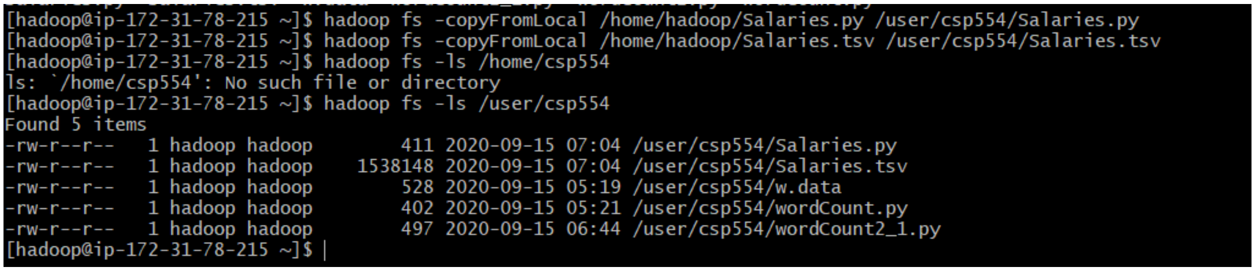


```
[hadoop@ip-172-31-78-215 ~]$ python wordCount2_1.py -r hadoop hdfs:///user/csp554/w.data
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.8.5
Looking for Hadoop 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_1.hadoop.20200915.064438.309841
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/wordCount2_1.hadoop.20200915.064438.309841/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/wordCount2_1.hadoop.20200915.064438.309841/files/
Running step 1 of 1
```



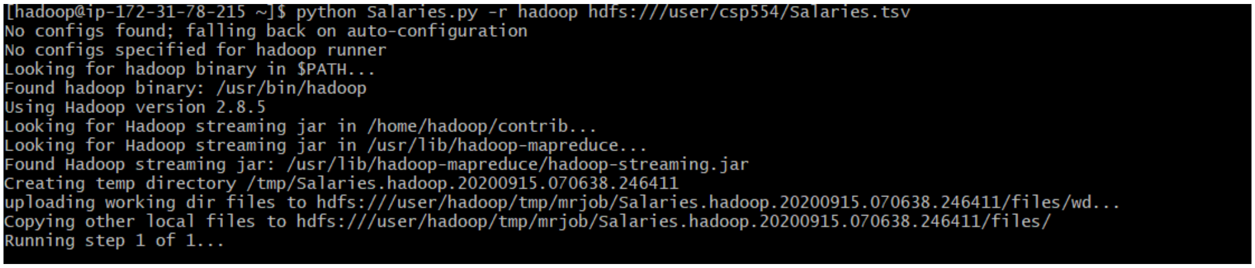
```
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/wordCount2_1.hadoop.20200915.064438.309841/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/wordCount2_1.hadoop.20200915.064438.309841/output...
"a to n"        46
"other" 49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/wordCount2_1.hadoop.20200915.064438.309841...
Removing temp directory /tmp/wordCount2_1.hadoop.20200915.064438.309841...
[hadoop@ip-172-31-78-215 ~]$ |
```

Step 7:

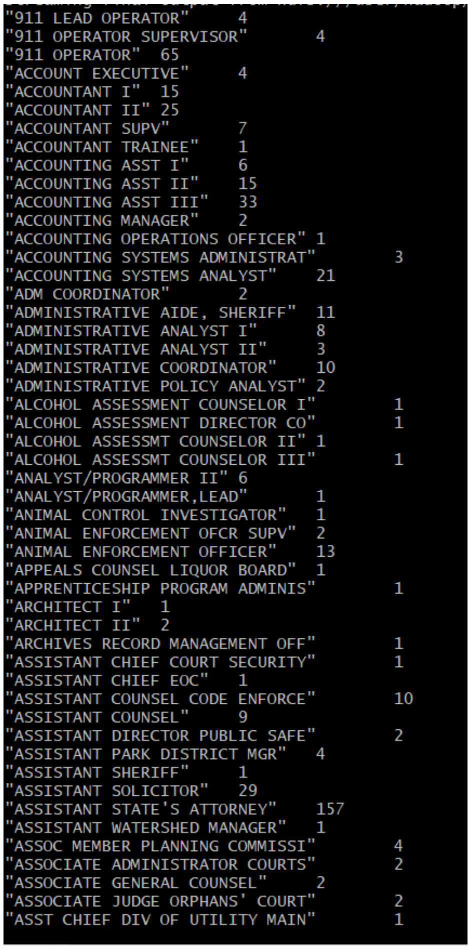


```
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -copyFromLocal /home/hadoop/Salaries.py /user/csp554/Salaries.py
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -copyFromLocal /home/hadoop/Salaries.tsv /user/csp554/Salaries.tsv
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -ls /home/csp554
ls: `/home/csp554': No such file or directory
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -ls /user/csp554
Found 5 items
-rw-r--r--   1 hadoop hadoop        411 2020-09-15 07:04 /user/csp554/Salaries.py
-rw-r--r--   1 hadoop hadoop    1538148 2020-09-15 07:04 /user/csp554/Salaries.tsv
-rw-r--r--   1 hadoop hadoop        528 2020-09-15 05:19 /user/csp554/w.data
-rw-r--r--   1 hadoop hadoop        402 2020-09-15 05:21 /user/csp554/wordCount.py
-rw-r--r--   1 hadoop hadoop        497 2020-09-15 06:44 /user/csp554/wordCount2_1.py
[hadoop@ip-172-31-78-215 ~]$ |
```

Step 8:



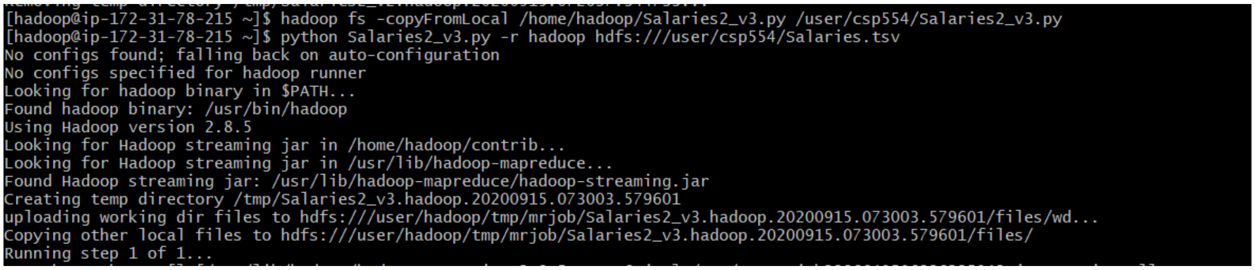
```
[hadoop@ip-172-31-78-215 ~]$ python Salaries.py -r hadoop hdfs:///user/csp554/Salaries.tsv
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.8.5
Looking for 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/Salaries.hadoop.20200915.070638.246411
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20200915.070638.246411/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20200915.070638.246411/files/
Running step 1 of 1...
```



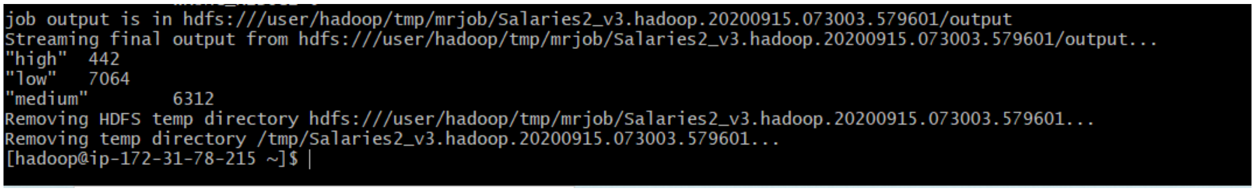
```
"911 LEAD OPERATOR"     4
"911 OPERATOR SUPERVISOR"       4
"911 OPERATOR"  65
"ACCOUNT EXECUTIVE"     4
"ACCOUNTANT I"  15
"ACCOUNTANT II" 25
"ACCOUNTANT SUPV"       7
"ACCOUNTANT TRAINEE"    1
"ACCOUNTING ASST I"     6
"ACCOUNTING ASST II"    15
"ACCOUNTING ASST III"   33
"ACCOUNTING MANAGER"    2
"ACCOUNTING OPERATIONS OFFICER" 1
"ACCOUNTING SYSTEMS ADMINISTRAT"        3
"ACCOUNTING SYSTEMS ANALYST"    21
"ADM COORDINATOR"       2
"ADMINISTRATIVE AIDE, SHERIFF"  11
"ADMINISTRATIVE ANALYST I"      8
"ADMINISTRATIVE ANALYST II"     3
"ADMINISTRATIVE COORDINATOR"    10
"ADMINISTRATIVE POLICY ANALYST" 2
"ALCOHOL ASSESSMENT COUNSELOR I"        1
"ALCOHOL ASSESSMENT DIRECTOR CO"        1
"ALCOHOL ASSESSMT COUNSELOR II" 1
"ALCOHOL ASSESSMT COUNSELOR III"        1
"ANALYST/PROGRAMMER II" 6
"ANALYST/PROGRAMMER,LEAD"       1
"ANIMAL CONTROL INVESTIGATOR"   1
"ANIMAL ENFORCEMENT OFCR SUPV"  2
"ANIMAL ENFORCEMENT OFFICER"    13
"APPEALS COUNSEL LIQUOR BOARD"  1
"APPRENTICESHIP PROGRAM ADMINIS"        1
"ARCHITECT I"   1
"ARCHITECT II"  2
"ARCHIVES RECORD MANAGEMENT OFF"        1
"ASSISTANT CHIEF COURT SECURITY"        1
"ASSISTANT CHIEF EOC"   1
"ASSISTANT COUNSEL CODE ENFORCE"        10
"ASSISTANT COUNSEL"     9
"ASSISTANT DIRECTOR PUBLIC SAFE"        2
"ASSISTANT PARK DISTRICT MGR"   4
"ASSISTANT SHERIFF"     1
"ASSISTANT SOLICITOR"   29
"ASSISTANT STATE'S ATTORNEY"    157
"ASSISTANT WATERSHED MANAGER"   1
"ASSOC MEMBER PLANNING COMMISSI"        4
"ASSOCIATE ADMINISTRATOR COURTS"        2
"ASSOCIATE GENERAL COUNSEL"     2
"ASSOCIATE JUDGE ORPHANS' COURT"        2
"ASST CHIEF DIV OF UTILITY MAIN"        1
```

Step 10:

Salaries2_v3.py

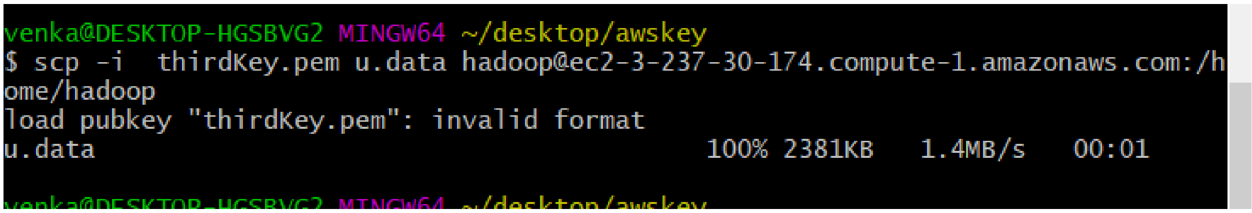


```
[hadoop@ip-172-31-78-215 ~]$ hadoop fs -copyFromLocal /home/hadoop/Salaries2_v3.py /user/csp554/Salaries2_v3.py
[hadoop@ip-172-31-78-215 ~]$ python Salaries2_v3.py -r hadoop hdfs:///user/csp554/Salaries.tsv
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.8.5
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_v3.hadoop.20200915.073003.579601
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2_v3.hadoop.20200915.073003.579601/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2_v3.hadoop.20200915.073003.579601/files/
Running step 1 of 1...
```



```
job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries2_v3.hadoop.20200915.073003.579601/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2_v3.hadoop.20200915.073003.579601/output...
"high"  442
"low"   7064
"medium"        6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2_v3.hadoop.20200915.073003.579601...
Removing temp directory /tmp/Salaries2_v3.hadoop.20200915.073003.579601...
[hadoop@ip-172-31-78-215 ~]$ |
```

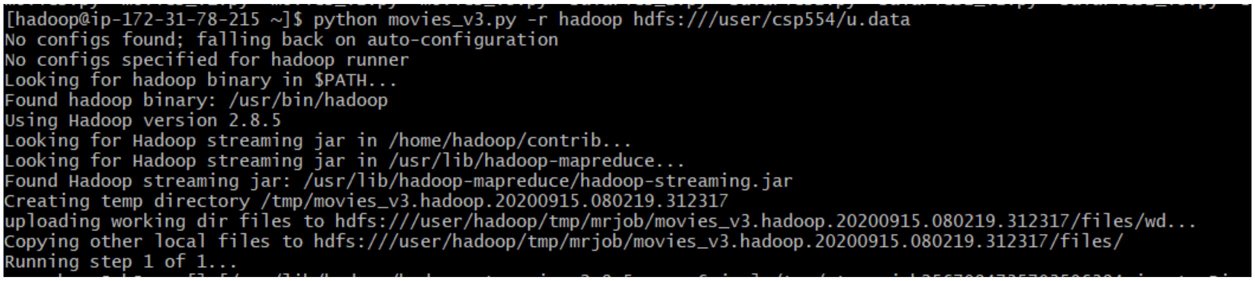
Step 11:



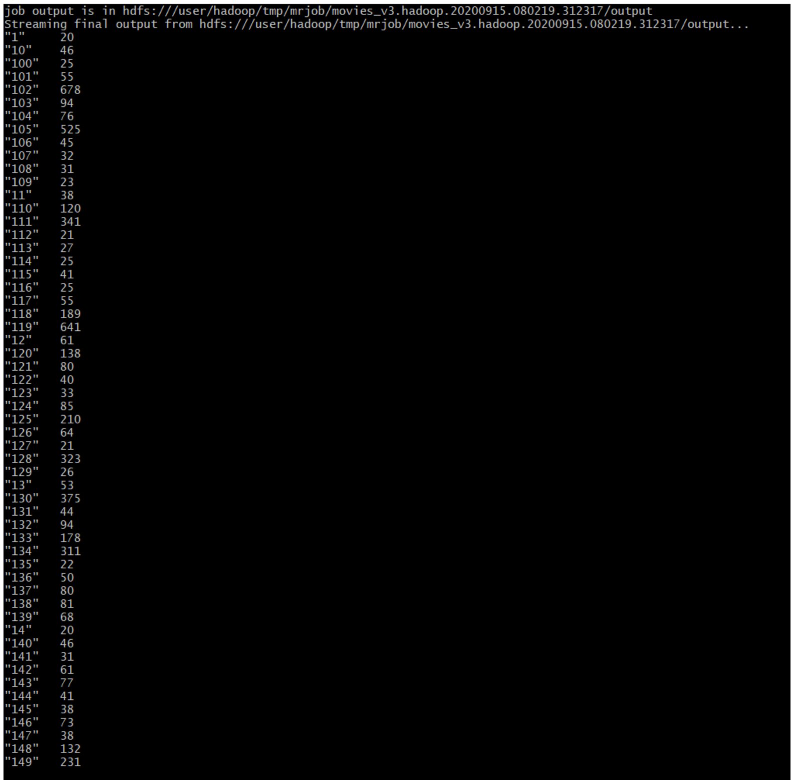
```
venka@DESKTOP-HGSBVG2 MINGW64 ~/desktop/awskey
$ scp -i  thirdKey.pem u.data hadoop@ec2-3-237-30-174.compute-1.amazonaws.com:/h
ome/hadoop
load pubkey "thirdKey.pem": invalid format
u.data                                        100% 2381KB   1.4MB/s   00:01

venka@DESKTOP-HGSBVG2 MINGW64 ~/desktop/awskey
```

Step12:

movies_v3.py



```
[hadoop@ip-172-31-78-215 ~]$ python movies_v3.py -r hadoop hdfs:///user/csp554/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 2.8.5
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_v3.hadoop.20200915.080219.312317
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/movies_v3.hadoop.20200915.080219.312317/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/movies_v3.hadoop.20200915.080219.312317/files/
Running step 1 of 1...
```



```
job output is in hdfs:///user/hadoop/tmp/mrjob/movies_v3.hadoop.20200915.080219.312317/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/movies_v3.hadoop.20200915.080219.312317/output...
"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
"149"   231
```