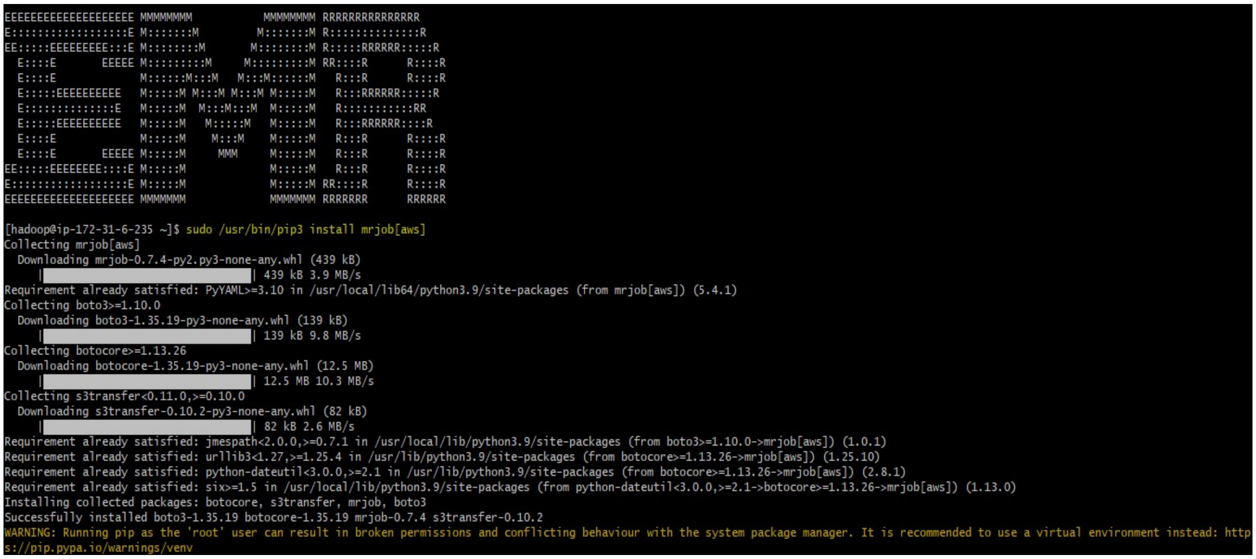
ASSIGNMENT - 3

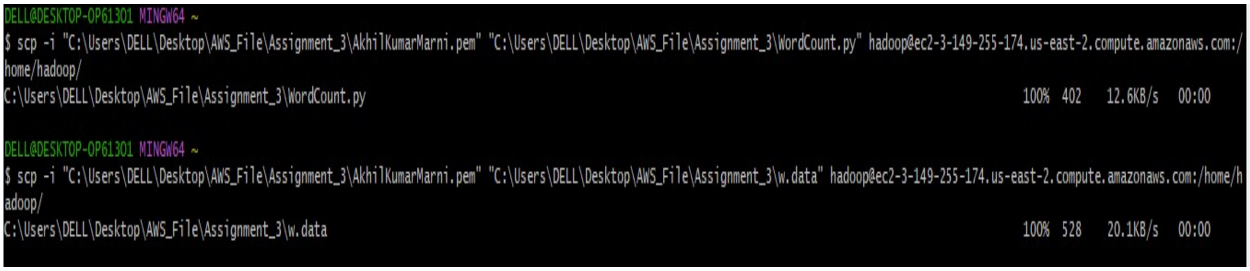
CSP - 554 BIG DATA TECHNOLOGIES

ssh to the master node



```
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-172-31-6-235 ~]$ sudo /usr/bin/pip3 install mrjob[aws]
Collecting mrjob[aws]
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
     |████████████████████████████████| 439 kB 3.9 MB/s
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.9/site-packages (from mrjob[aws]) (5.4.1)
Collecting boto3>=1.10.0
  Downloading boto3-1.35.19-py3-none-any.whl (139 kB)
     |████████████████████████████████| 139 kB 9.8 MB/s
Collecting botocore>=1.13.26
  Downloading botocore-1.35.19-py3-none-any.whl (12.5 MB)
     |████████████████████████████████| 12.5 MB 10.3 MB/s
Collecting s3transfer<0.11.0,>=0.10.0
  Downloading s3transfer-0.10.2-py3-none-any.whl (82 kB)
     |████████████████████████████████| 82 kB 2.6 MB/s
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /usr/local/lib/python3.9/site-packages (from boto3>=1.10.0->mrjob[aws]) (1.0.1)
Requirement already satisfied: urllib3<1.27,>=1.25.4 in /usr/lib/python3.9/site-packages (from botocore>=1.13.26->mrjob[aws]) (1.25.10)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in /usr/lib/python3.9/site-packages (from botocore>=1.13.26->mrjob[aws]) (2.8.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.9/site-packages (from python-dateutil<3.0.0,>=2.1->botocore>=1.13.26->mrjob[aws]) (1.13.0)
Installing collected packages: botocore, s3transfer, mrjob, boto3
Successfully installed boto3-1.35.19 botocore-1.35.19 mrjob-0.7.4 s3transfer-0.10.2
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: http
s://pip.pypa.io/warnings/venv
```

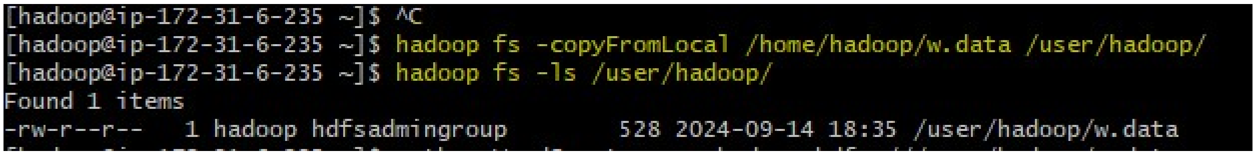
Moving WordCount.py and w.data to /home/Hadoop
Moving W.data to /user/Hadoop



```
DELL@DESKTOP-OP61301 MINGW64 ~
$ scp -i "C:\Users\DELL\Desktop\AWS_File\Assignment_3\AkhilKumarMarni.pem" "C:\Users\DELL\Desktop\AWS_File\Assignment_3\WordCount.py" hadoop@ec2-3-149-255-174.us-east-2.compute.amazonaws.com:/
home/hadoop/
C:\Users\DELL\Desktop\AWS_File\Assignment_3\WordCount.py                                                    100%  402    12.6KB/s   00:00

DELL@DESKTOP-OP61301 MINGW64 ~
$ scp -i "C:\Users\DELL\Desktop\AWS_File\Assignment_3\AkhilKumarMarni.pem" "C:\Users\DELL\Desktop\AWS_File\Assignment_3\w.data" hadoop@ec2-3-149-255-174.us-east-2.compute.amazonaws.com:/home/h
adoop/
C:\Users\DELL\Desktop\AWS_File\Assignment_3\w.data                                                          100%  528    20.1KB/s   00:00
```

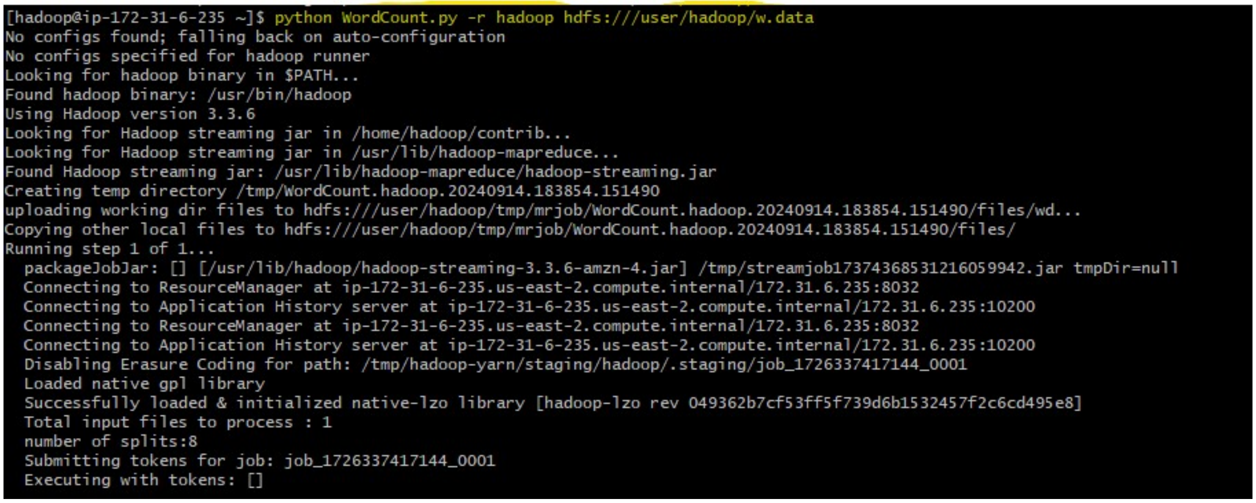
A20554334
AKHIL KUMAR MARNI

Moving W.data to /user/Hadoop



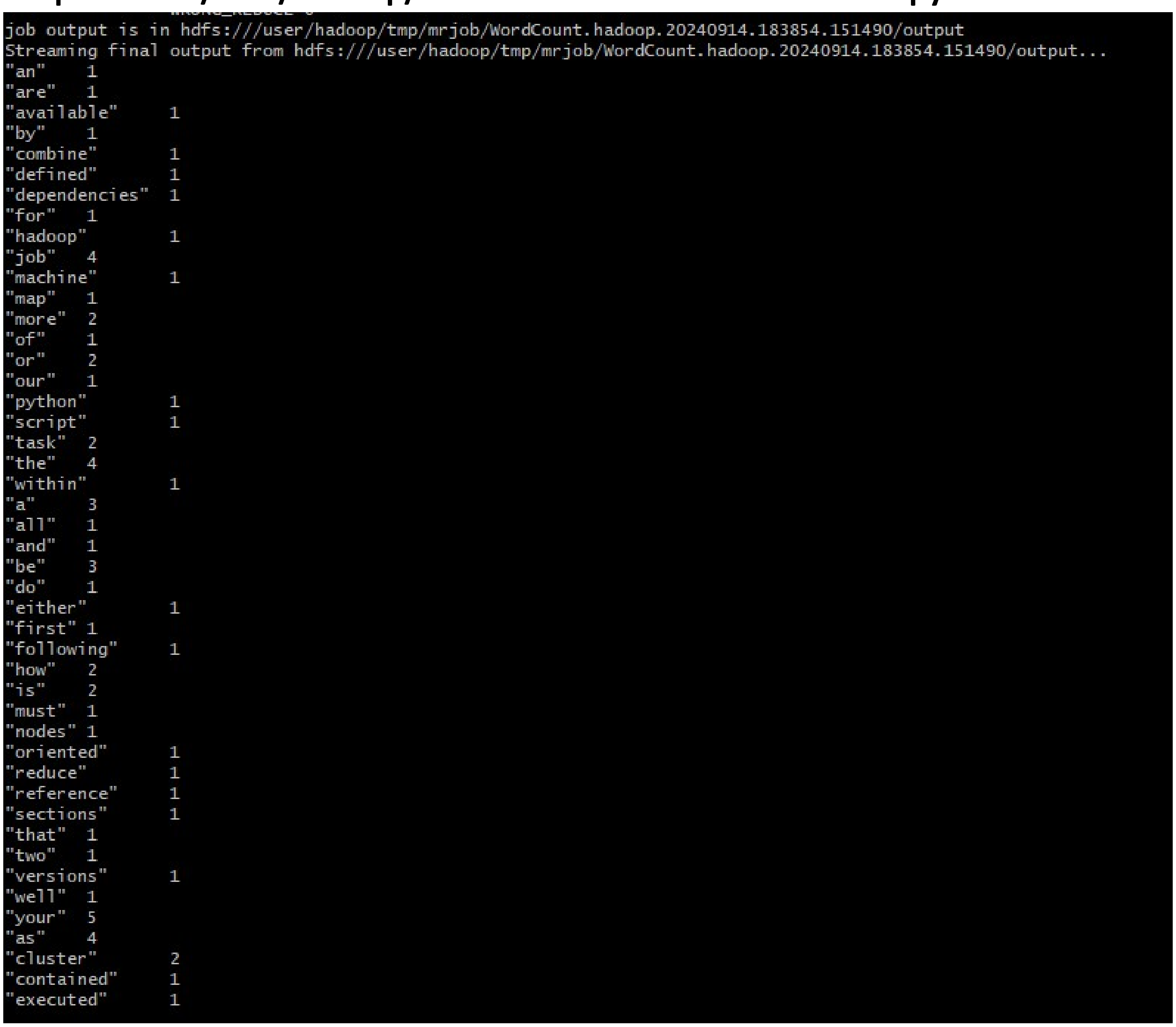
```
[hadoop@ip-172-31-6-235 ~]$ ^C
[hadoop@ip-172-31-6-235 ~]$ hadoop fs -copyFromLocal /home/hadoop/w.data /user/hadoop/
[hadoop@ip-172-31-6-235 ~]$ hadoop fs -ls /user/hadoop/
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup        528 2024-09-14 18:35 /user/hadoop/w.data
```

Executing the first Job from unmodified WordCount.py code



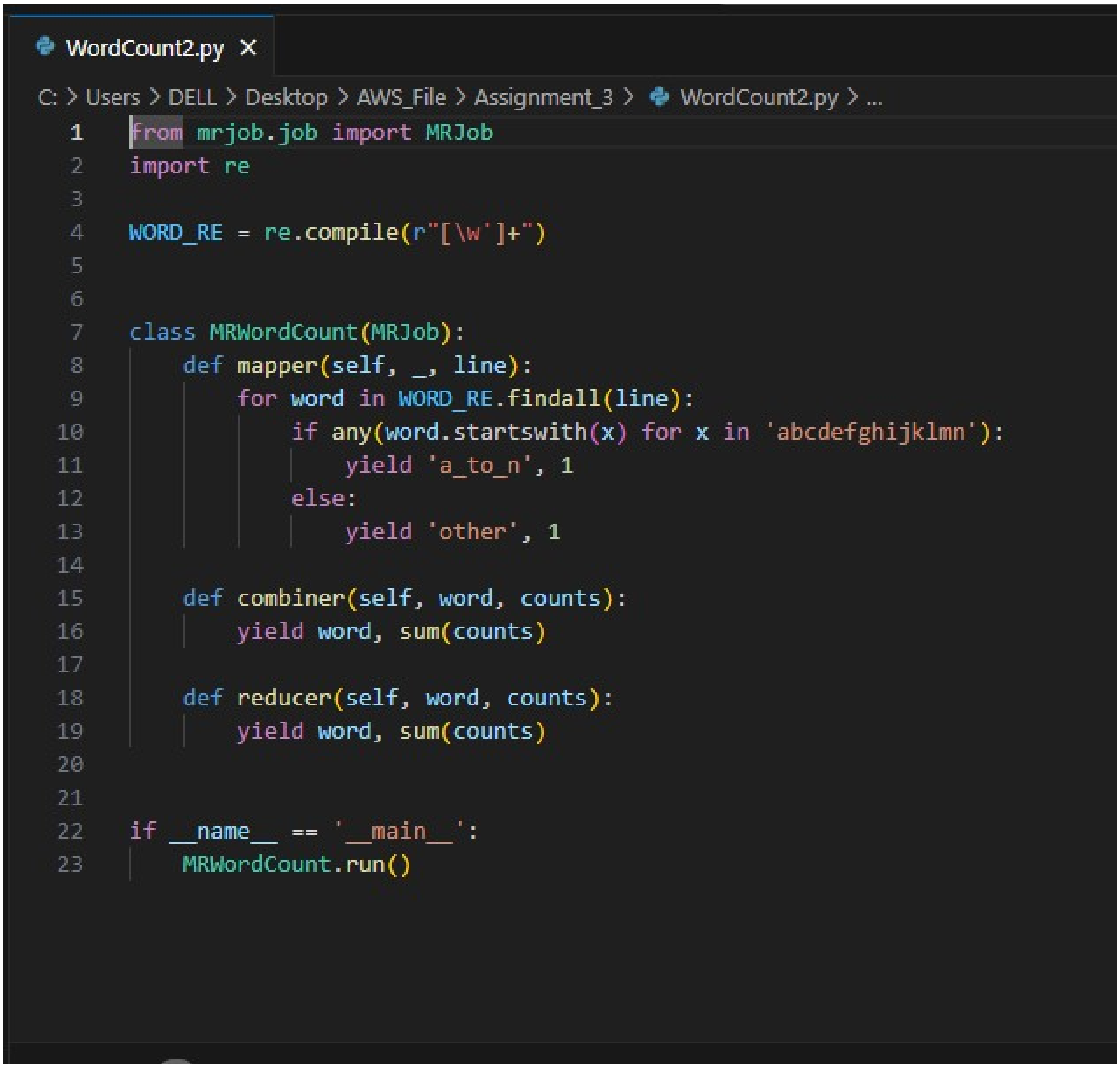
```
[hadoop@ip-172-31-6-235 ~]$ python WordCount.py -r hadoop hdfs:///user/hadoop/w.data
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.3.6
Looking for Hadoop 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.20240914.183854.151490
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20240914.183854.151490/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20240914.183854.151490/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob17374368531216059942.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-6-235.us-east-2.compute.internal/172.31.6.235:8032
  Connecting to Application History server at ip-172-31-6-235.us-east-2.compute.internal/172.31.6.235:10200
  Connecting to ResourceManager at ip-172-31-6-235.us-east-2.compute.internal/172.31.6.235:8032
  Connecting to Application History server at ip-172-31-6-235.us-east-2.compute.internal/172.31.6.235:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1726337417144_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_1726337417144_0001
  Executing with tokens: []
```

Output file in /user/hadoop/out from unmodified WordCount.py code



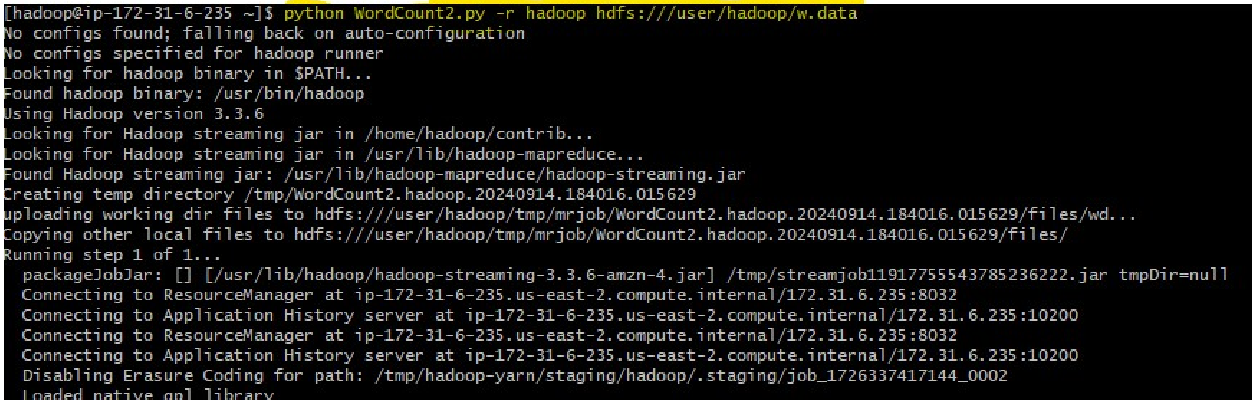
```
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20240914.183854.151490/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20240914.183854.151490/output...
"an"    1
"are"   1
"available"     1
"by"    1
"combine"       1
"defined"       1
"dependencies"  1
"for"   1
"hadoop"        1
"job"   4
"machine"       1
"map"   1
"more"  2
"of"    1
"or"    2
"our"   1
"python"        1
"script"        1
"task"  2
"the"   4
"within"        1
"a"     3
"all"   1
"and"   1
"be"    3
"do"    1
"either"        1
"first" 1
"following"     1
"how"   2
"is"    2
"must"  1
"nodes" 1
"oriented"      1
"reduce"        1
"reference"     1
"sections"      1
"that"  1
"two"   1
"versions"      1
"well"  1
"your"  5
"as"    4
"cluster"       2
"contained"     1
"executed"      1
```

(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. Screenshot of modified WordCount.py code further saved to a new file named "WordCount2.py" for finding words starting from a to n and other words.


WordCount2.py ×
C: > Users > DELL > Desktop > AWS_File > Assignment_3 > WordCount2.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 any(word.startswith(x) for x in 'abcdefghijklmn'):
                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()
```

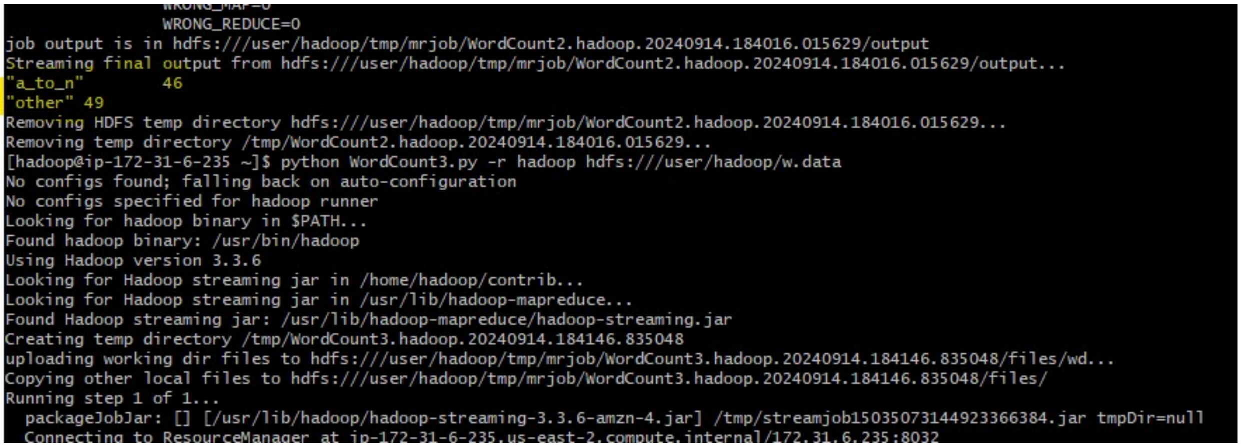
Executing the new modified code “WordCount2.py”



```
[hadoop@ip-172-31-6-235 ~]$ 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 3.3.6
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.20240914.184016.015629
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240914.184016.015629/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240914.184016.015629/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob11917755543785236222.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-6-235.us-east-2.compute.internal/172.31.6.235:8032
  Connecting to Application History server at ip-172-31-6-235.us-east-2.compute.internal/172.31.6.235:10200
  Connecting to ResourceManager at ip-172-31-6-235.us-east-2.compute.internal/172.31.6.235:8032
  Connecting to Application History server at ip-172-31-6-235.us-east-2.compute.internal/172.31.6.235:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1726337417144_0002
  Loaded native gpl library
```

Output

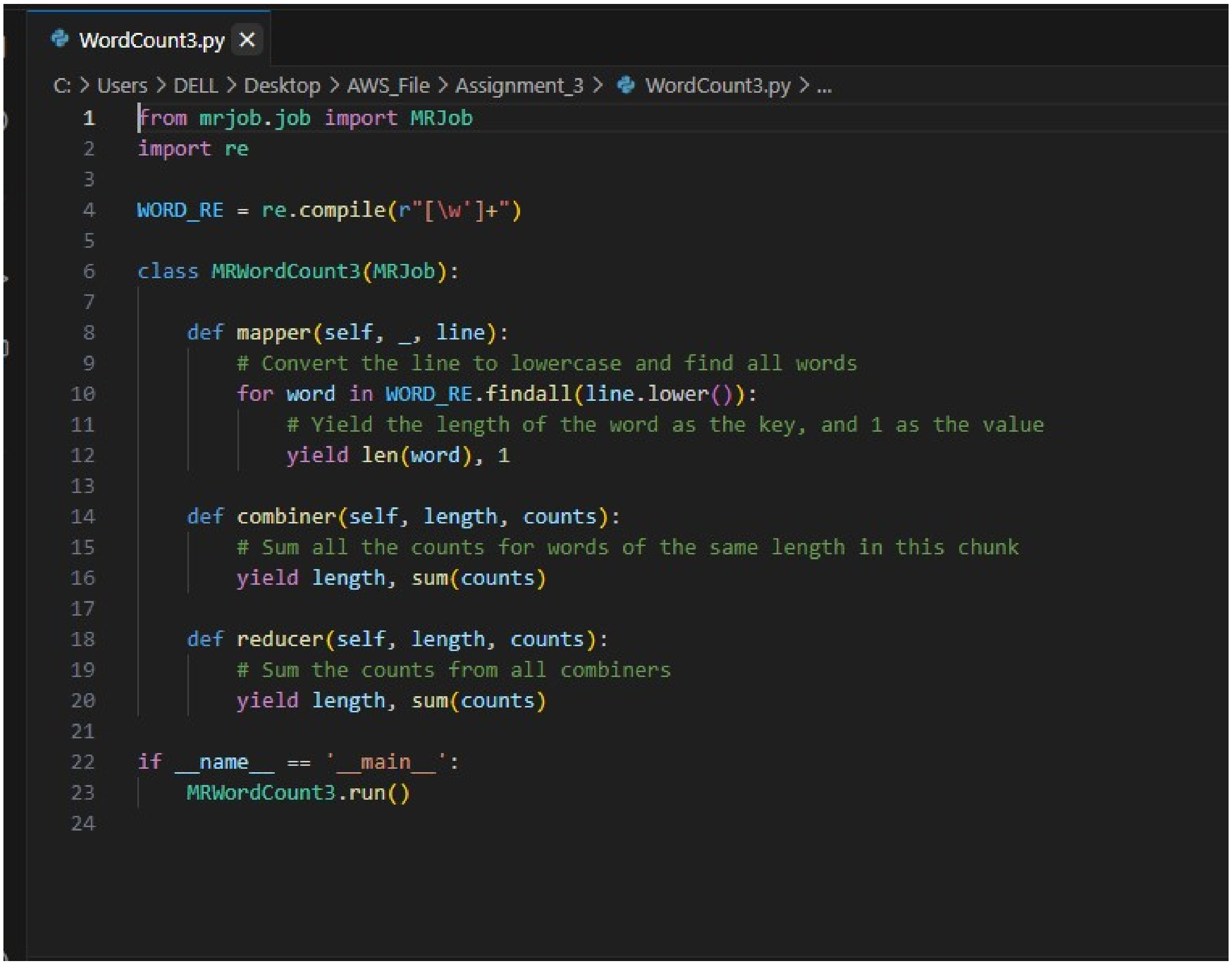
Output [a_to_n = 46 and other = 49]



```
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240914.184016.015629/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240914.184016.015629/output...
"a_to_n"        46
"other" 49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240914.184016.015629...
Removing temp directory /tmp/WordCount2.hadoop.20240914.184016.015629...
[hadoop@ip-172-31-6-235 ~]$ python WordCount3.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 3.3.6
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/WordCount3.hadoop.20240914.184146.835048
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240914.184146.835048/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240914.184146.835048/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob15035073144923366384.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-6-235.us-east-2.compute.internal/172.31.6.235:8032
```

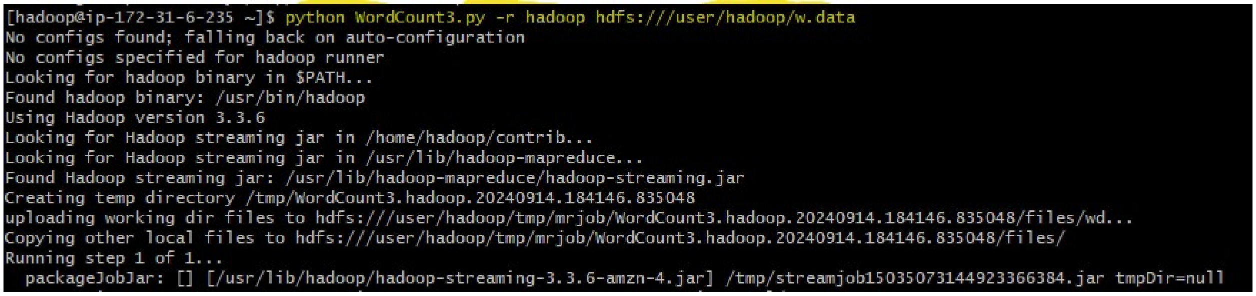
(4 points) When you have accomplished this, please submit the following,

(1) a copy of your MRJob code.

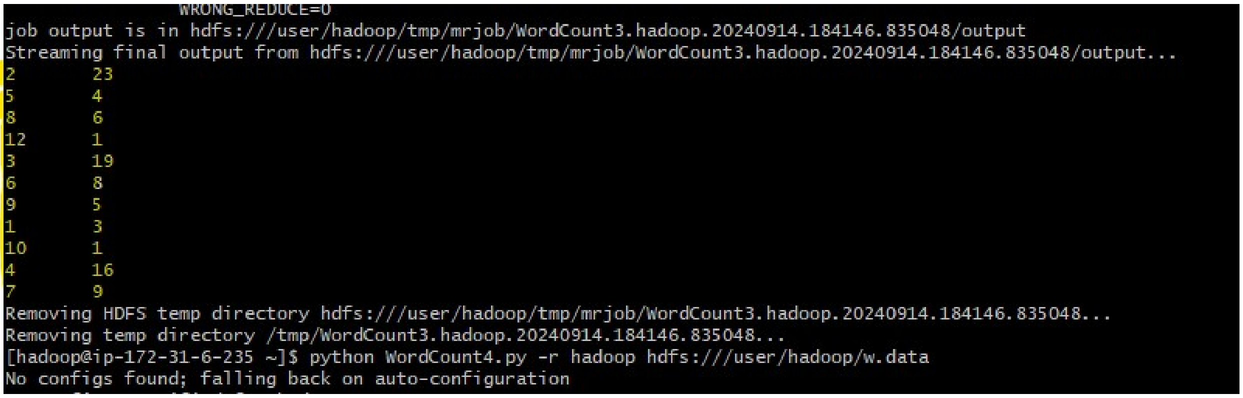

WordCount3.py ✕
C: > Users > DELL > Desktop > AWS_File > Assignment_3 > WordCount3.py > ...

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

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

class MRWordCount3(MRJob):

    def mapper(self, _, line):
        # Convert the line to lowercase and find all words
        for word in WORD_RE.findall(line.lower()):
            # Yield the length of the word as the key, and 1 as the value
            yield len(word), 1

    def combiner(self, length, counts):
        # Sum all the counts for words of the same length in this chunk
        yield length, sum(counts)

    def reducer(self, length, counts):
        # Sum the counts from all combiners
        yield length, sum(counts)

if __name__ == '__main__':
    MRWordCount3.run()

```

(2) a copy of the output of the execution of that code.



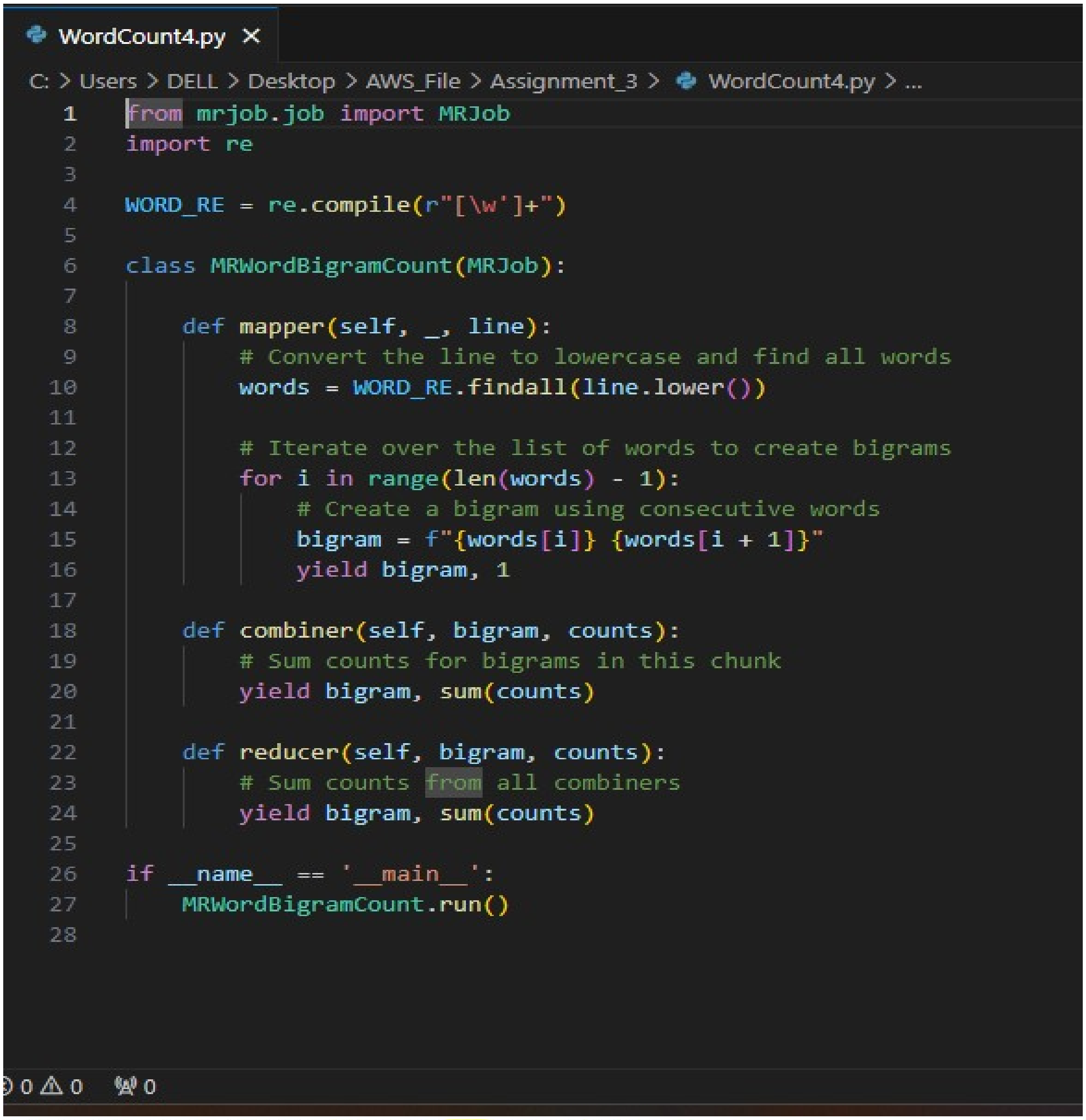
```
[hadoop@ip-172-31-6-235 ~]$ python WordCount3.py -r hadoop hdfs:///user/hadoop/w.data
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.3.6
Looking for Hadoop 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/WordCount3.hadoop.20240914.184146.835048
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240914.184146.835048/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240914.184146.835048/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob15035073144923366384.jar tmpDir=null
```



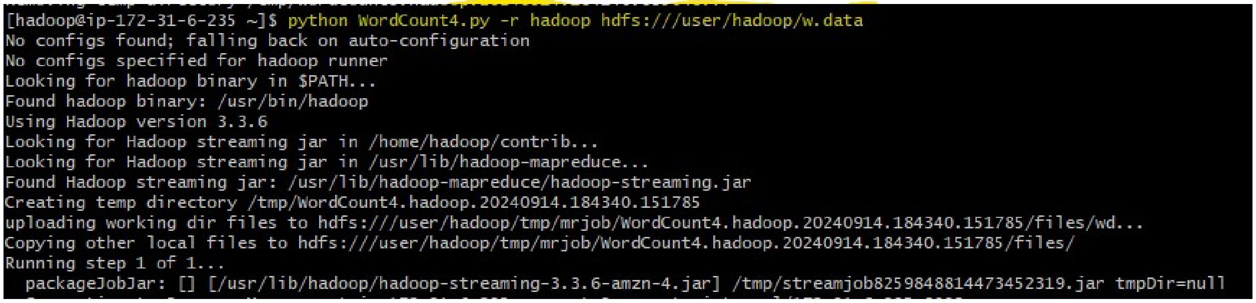
```
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240914.184146.835048/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240914.184146.835048/output...
2       23
5       4
8       6
12      1
3       19
6       8
9       5
1       3
10      1
4       16
7       9
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240914.184146.835048...
Removing temp directory /tmp/WordCount3.hadoop.20240914.184146.835048...
[hadoop@ip-172-31-6-235 ~]$ python WordCount4.py -r hadoop hdfs:///user/hadoop/w.data
No configs found; falling back on auto-configuration
```

(5 points) When you have accomplished this, please submit the following,

(1) a copy of your MRJob code.

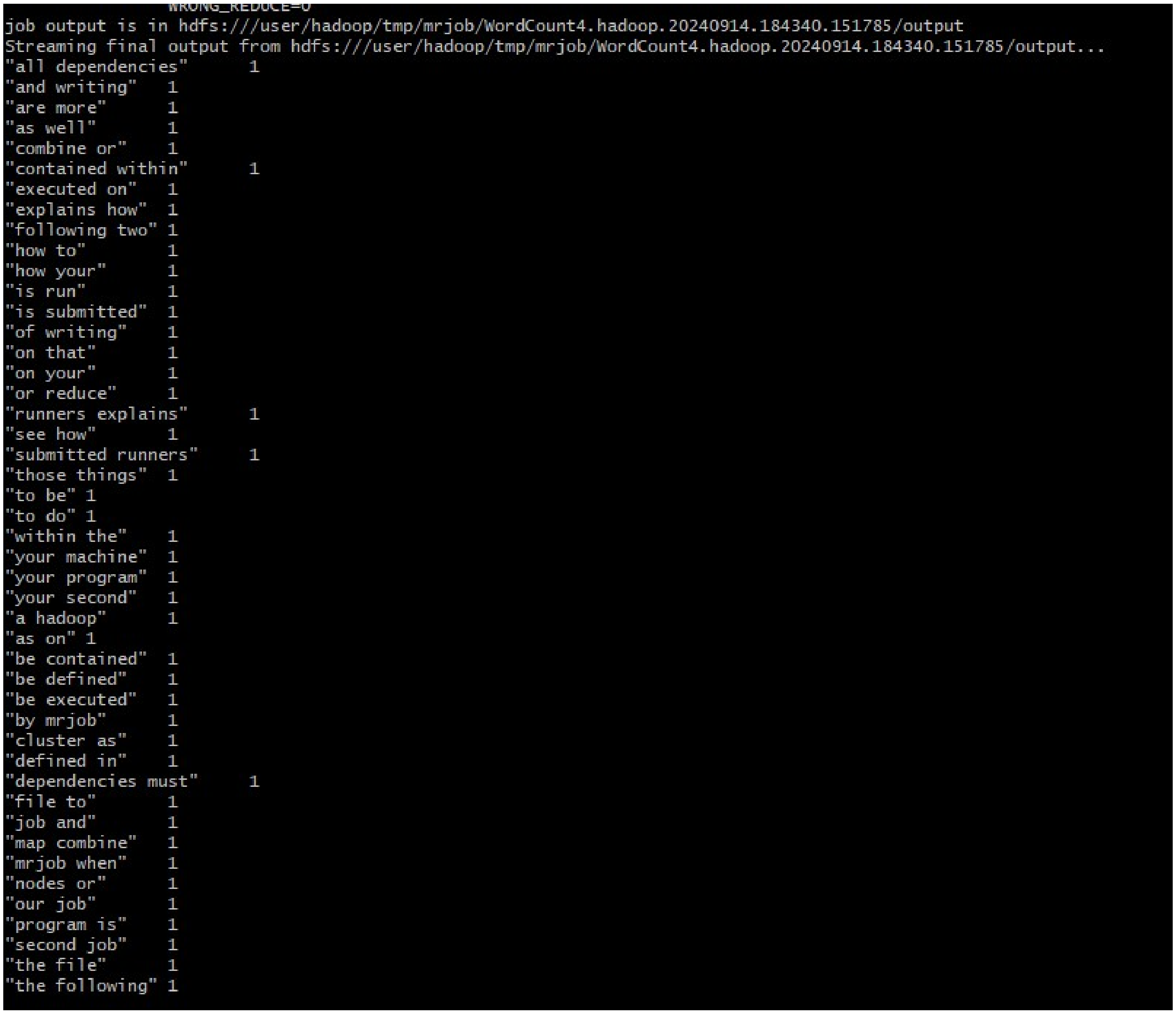

WordCount4.py
C: > Users > DELL > Desktop > AWS_File > Assignment_3 > WordCount4.py > ...

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

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

class MRWordBigramCount(MRJob):

    def mapper(self, _, line):
        # Convert the line to lowercase and find all words
        words = WORD_RE.findall(line.lower())

        # Iterate over the list of words to create bigrams
        for i in range(len(words) - 1):
            # Create a bigram using consecutive words
            bigram = f"{words[i]} {words[i + 1]}"
            yield bigram, 1

    def combiner(self, bigram, counts):
        # Sum counts for bigrams in this chunk
        yield bigram, sum(counts)

    def reducer(self, bigram, counts):
        # Sum counts from all combiners
        yield bigram, sum(counts)

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

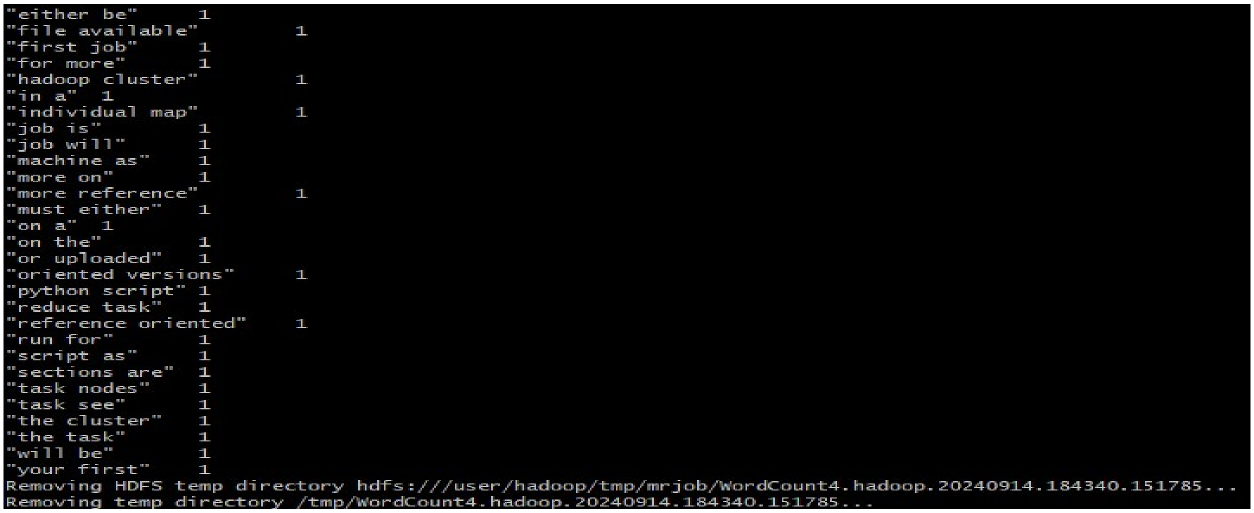


```
[hadoop@ip-172-31-6-235 ~]$ python WordCount4.py -r hadoop hdfs:///user/hadoop/w.data
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.3.6
Looking for Hadoop 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/WordCount4.hadoop.20240914.184340.151785
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240914.184340.151785/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240914.184340.151785/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob8259848814473452319.jar tmpDir=null
```

(2) a copy of the output of the execution of that code for at least the first 10 bigram key value pairs.



```
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240914.184340.151785/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240914.184340.151785/output...
"all dependencies"      1
"and writing"   1
"are more"      1
"as well"       1
"combine or"    1
"contained within"      1
"executed on"   1
"explains how"  1
"following two" 1
"how to"        1
"how your"      1
"is run"        1
"is submitted"  1
"of writing"    1
"on that"       1
"on your"       1
"or reduce"     1
"runners explains"      1
"see how"       1
"submitted runners"     1
"those things"  1
"to be" 1
"to do" 1
"within the"    1
"your machine"  1
"your program"  1
"your second"   1
"a hadoop"      1
"as on" 1
"be contained"  1
"be defined"    1
"be executed"   1
"by mrjob"      1
"cluster as"    1
"defined in"    1
"dependencies must"     1
"file to"       1
"job and"       1
"map combine"   1
"mrjob when"    1
"nodes or"      1
"our job"       1
"program is"    1
"second job"    1
"the file"      1
"the following" 1
```

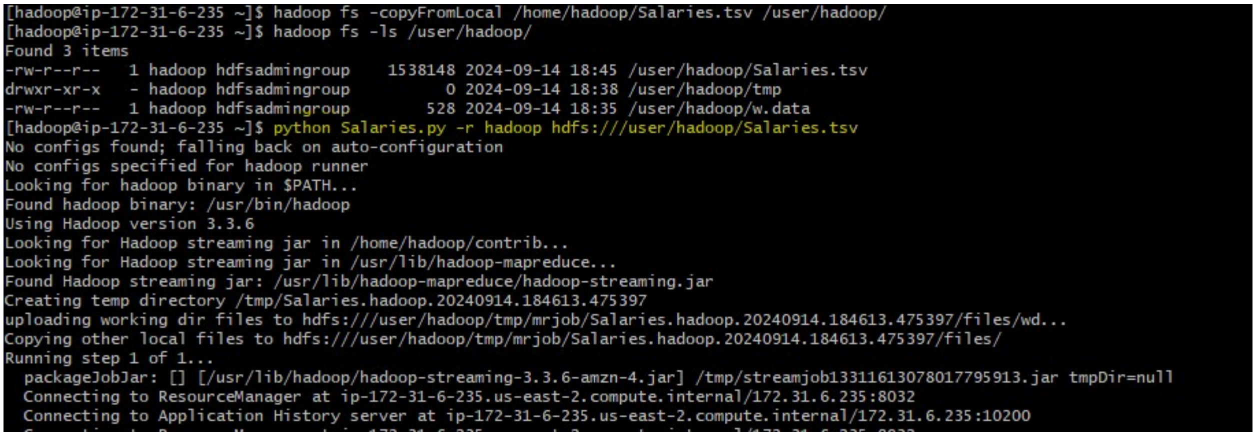


```
"either be"     1
"file available"        1
"first job"     1
"for more"      1
"hadoop cluster"        1
"in a"  1
"individual map"        1
"job is"        1
"job will"      1
"machine as"    1
"more on"       1
"more reference"        1
"must either"   1
"on a"  1
"on the"        1
"or uploaded"   1
"oriented versions"     1
"python script" 1
"reduce task"   1
"reference oriented"    1
"run for"       1
"script as"     1
"sections are"  1
"task nodes"    1
"task see"      1
"the cluster"   1
"the task"      1
"will be"       1
"your first"    1
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240914.184340.151785...
Removing temp directory /tmp/WordCount4.hadoop.20240914.184340.151785...
```

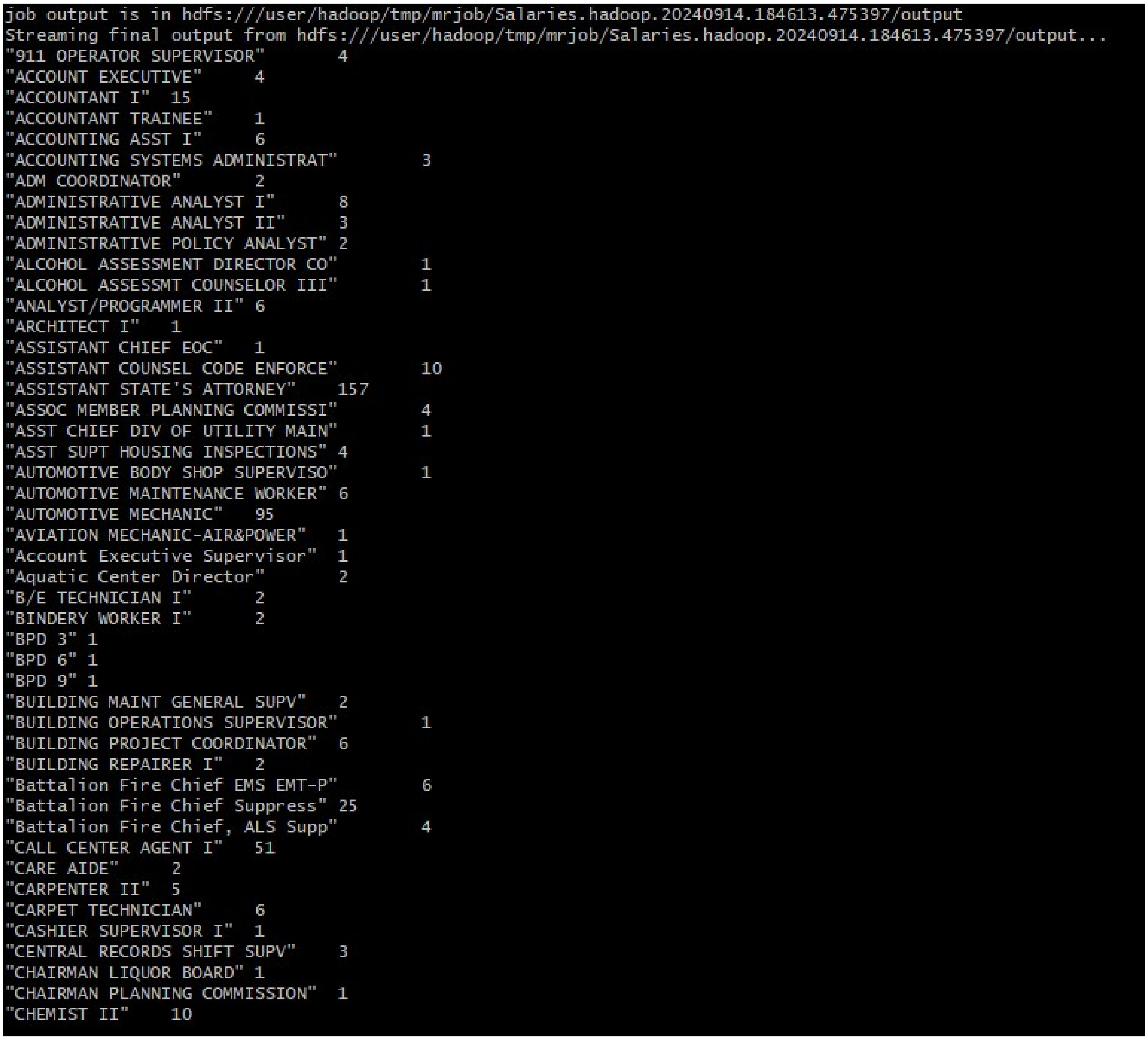
A20554334
AKHIL KUMAR MARNI

11) 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.

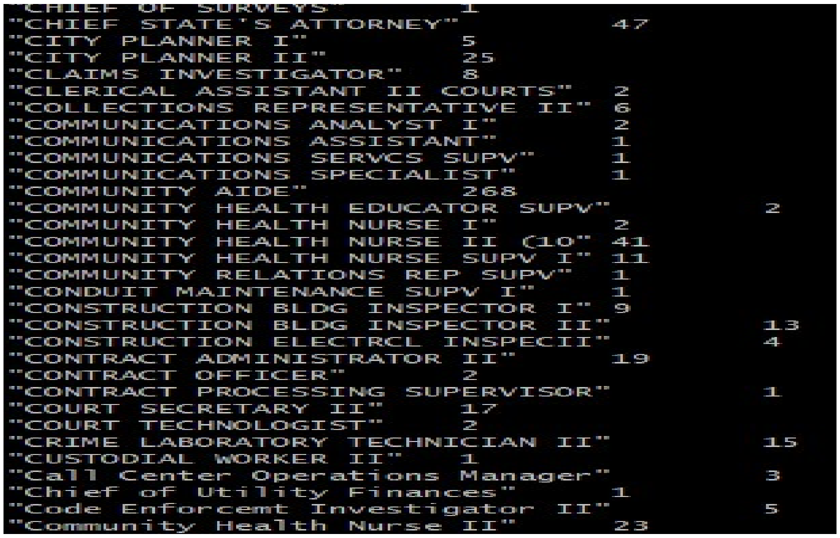
12) Execute the Salaries.py program to make sure it works. It should print out how many workers share each job title.



```
[hadoop@ip-172-31-6-235 ~]$ hadoop fs -copyFromLocal /home/hadoop/Salaries.tsv /user/hadoop/
[hadoop@ip-172-31-6-235 ~]$ hadoop fs -ls /user/hadoop/
Found 3 items
-rw-r--r--   1 hadoop hdfsadmingroup    1538148 2024-09-14 18:45 /user/hadoop/Salaries.tsv
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-09-14 18:38 /user/hadoop/tmp
-rw-r--r--   1 hadoop hdfsadmingroup        528 2024-09-14 18:35 /user/hadoop/w.data
[hadoop@ip-172-31-6-235 ~]$ 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 3.3.6
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.20240914.184613.475397
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20240914.184613.475397/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20240914.184613.475397/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob13311613078017795913.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-6-235.us-east-2.compute.internal/172.31.6.235:8032
  Connecting to Application History server at ip-172-31-6-235.us-east-2.compute.internal/172.31.6.235:10200
```

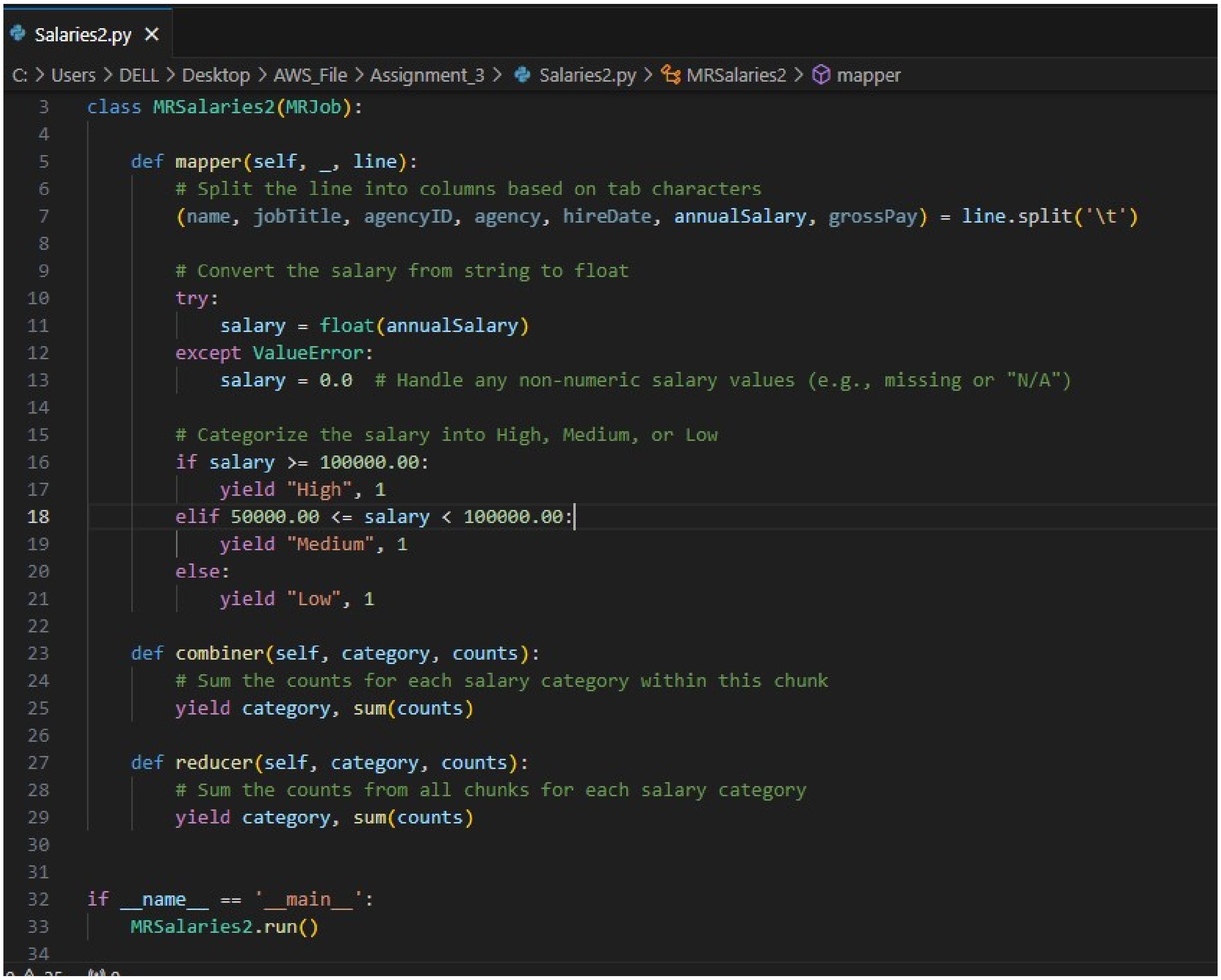


```
job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20240914.184613.475397/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20240914.184613.475397/output...
"911 OPERATOR SUPERVISOR"       4
"ACCOUNT EXECUTIVE"     4
"ACCOUNTANT I"  15
"ACCOUNTANT TRAINEE"    1
"ACCOUNTING ASST I"     6
"ACCOUNTING SYSTEMS ADMINISTRAT"        3
"ADM COORDINATOR"       2
"ADMINISTRATIVE ANALYST I"      8
"ADMINISTRATIVE ANALYST II"     3
"ADMINISTRATIVE POLICY ANALYST" 2
"ALCOHOL ASSESSMENT DIRECTOR CO"        1
"ALCOHOL ASSESSMT COUNSELOR III"        1
"ANALYST/PROGRAMMER II" 6
"ARCHITECT I"   1
"ASSISTANT CHIEF EOC"   1
"ASSISTANT COUNSEL CODE ENFORCE"        10
"ASSISTANT STATE'S ATTORNEY"    157
"ASSOC MEMBER PLANNING COMMISSI"        4
"ASST CHIEF DIV OF UTILITY MAIN"        1
"ASST SUPT HOUSING INSPECTIONS" 4
"AUTOMOTIVE BODY SHOP SUPERVISO"        1
"AUTOMOTIVE MAINTENANCE WORKER" 6
"AUTOMOTIVE MECHANIC"   95
"AVIATION MECHANIC-AIR&POWER"   1
"Account Executive Supervisor"  1
"Aquatic Center Director"       2
"B/E TECHNICIAN I"      2
"BINDERY WORKER I"      2
"BPD 3" 1
"BPD 6" 1
"BPD 9" 1
"BUILDING MAINT GENERAL SUPV"   2
"BUILDING OPERATIONS SUPERVISOR"        1
"BUILDING PROJECT COORDINATOR"  6
"BUILDING REPAIRER I"   2
"Battalion Fire Chief EMS EMT-P"        6
"Battalion Fire Chief Suppress" 25
"Battalion Fire Chief, ALS Supp"        4
"CALL CENTER AGENT I"   51
"CARE AIDE"     2
"CARPENTER II"  5
"CARPET TECHNICIAN"     6
"CASHIER SUPERVISOR I"  1
"CENTRAL RECORDS SHIFT SUPV"    3
"CHAIRMAN LIQUOR BOARD" 1
"CHAIRMAN PLANNING COMMISSION"  1
"CHEMIST II"    10
```



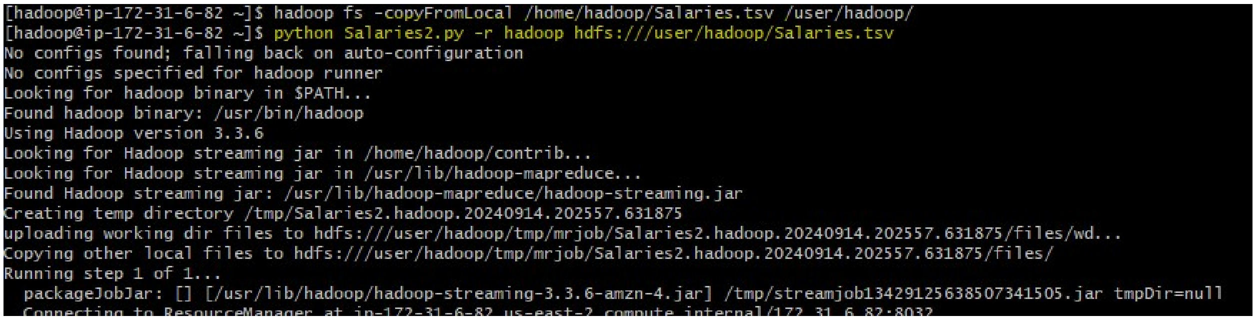
```
"CHIEF OF SURVEYS"      1
"CHIEF STATE'S ATTORNEY"        47
"CITY PLANNER I"        5
"CITY PLANNER II"       25
"CLAIMS INVESTIGATOR"   8
"CLERICAL ASSISTANT II COURTS"  2
"COLLECTIONS REPRESENTATIVE II" 6
"COMMUNICATIONS ANALYST I"      2
"COMMUNICATIONS ASSISTANT"      1
"COMMUNICATIONS SERVCS SUPV"    1
"COMMUNICATIONS SPECIALIST"     1
"COMMUNITY AIDE"        268
"COMMUNITY HEALTH EDUCATOR SUPV"        2
"COMMUNITY HEALTH NURSE I"      2
"COMMUNITY HEALTH NURSE II (10" 41
"COMMUNITY HEALTH NURSE SUPV I" 11
"COMMUNITY RELATIONS REP SUPV"  1
"CONDUIT MAINTENANCE SUPV I"    1
"CONSTRUCTION BLDG INSPECTOR I" 9
"CONSTRUCTION BLDG INSPECTOR II"        13
"CONSTRUCTION ELECTRCL INSPECII"        4
"CONTRACT ADMINISTRATOR II"     19
"CONTRACT OFFICER"      2
"CONTRACT PROCESSING SUPERVISOR"        1
"COURT SECRETARY II"    17
"COURT TECHNOLOGIST"    2
"CRIME LABORATORY TECHNICIAN II"        15
"CUSTODIAL WORKER II"   1
"Call Center Operations Manager"        3
"Chief of Utility Finances"     1
"Code Enforcemt Investigator II"        5
"Community Health Nurse II"     23
```

13. 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. The output of the program should be something like the following (in any order): High 20 Medium 30 Low 10 14) (3 points) Submit (1) a copy of this modified program

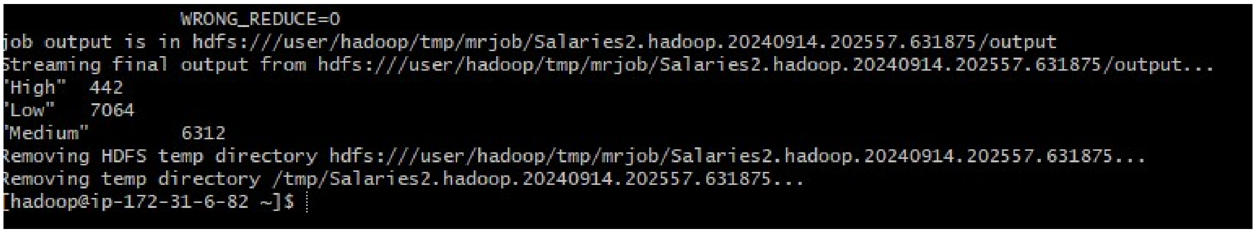

Salaries2.py ×
C: > Users > DELL > Desktop > AWS_File > Assignment_3 > Salaries2.py > MRSalaries2 > mapper

```python
class MRSalaries2(MRJob):

    def mapper(self, _, line):
        # Split the line into columns based on tab characters
        (name, jobTitle, agencyID, agency, hireDate, annualSalary, grossPay) = line.split('\t')

        # Convert the salary from string to float
        try:
            salary = float(annualSalary)
        except ValueError:
            salary = 0.0  # Handle any non-numeric salary values (e.g., missing or "N/A")

        # Categorize the salary into High, Medium, or Low
        if salary >= 100000.00:
            yield "High", 1
        elif 50000.00 <= salary < 100000.00:
            yield "Medium", 1
        else:
            yield "Low", 1

    def combiner(self, category, counts):
        # Sum the counts for each salary category within this chunk
        yield category, sum(counts)

    def reducer(self, category, counts):
        # Sum the counts from all chunks for each salary category
        yield category, sum(counts)


if __name__ == '__main__':
    MRSalaries2.run()
```

(2) a screenshot of the results of the program's execution as the output of your assignment.



```
[hadoop@ip-172-31-6-82 ~]$ hadoop fs -copyFromLocal /home/hadoop/Salaries.tsv /user/hadoop/
[hadoop@ip-172-31-6-82 ~]$ 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 3.3.6
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.20240914.202557.631875
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240914.202557.631875/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240914.202557.631875/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob13429125638507341505.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-6-82.us-east-2.compute.internal/172.31.6.82:8032
```



```
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240914.202557.631875/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240914.202557.631875/output...
"High"  442
"Low"   7064
"Medium"        6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240914.202557.631875...
Removing temp directory /tmp/Salaries2.hadoop.20240914.202557.631875...
[hadoop@ip-172-31-6-82 ~]$
```

-----------------------------THE END-----------------------------