Name: Deep Pawar(A20545137)
Professor: Joseph Rosen
Institute: Illinois Institute of Technology

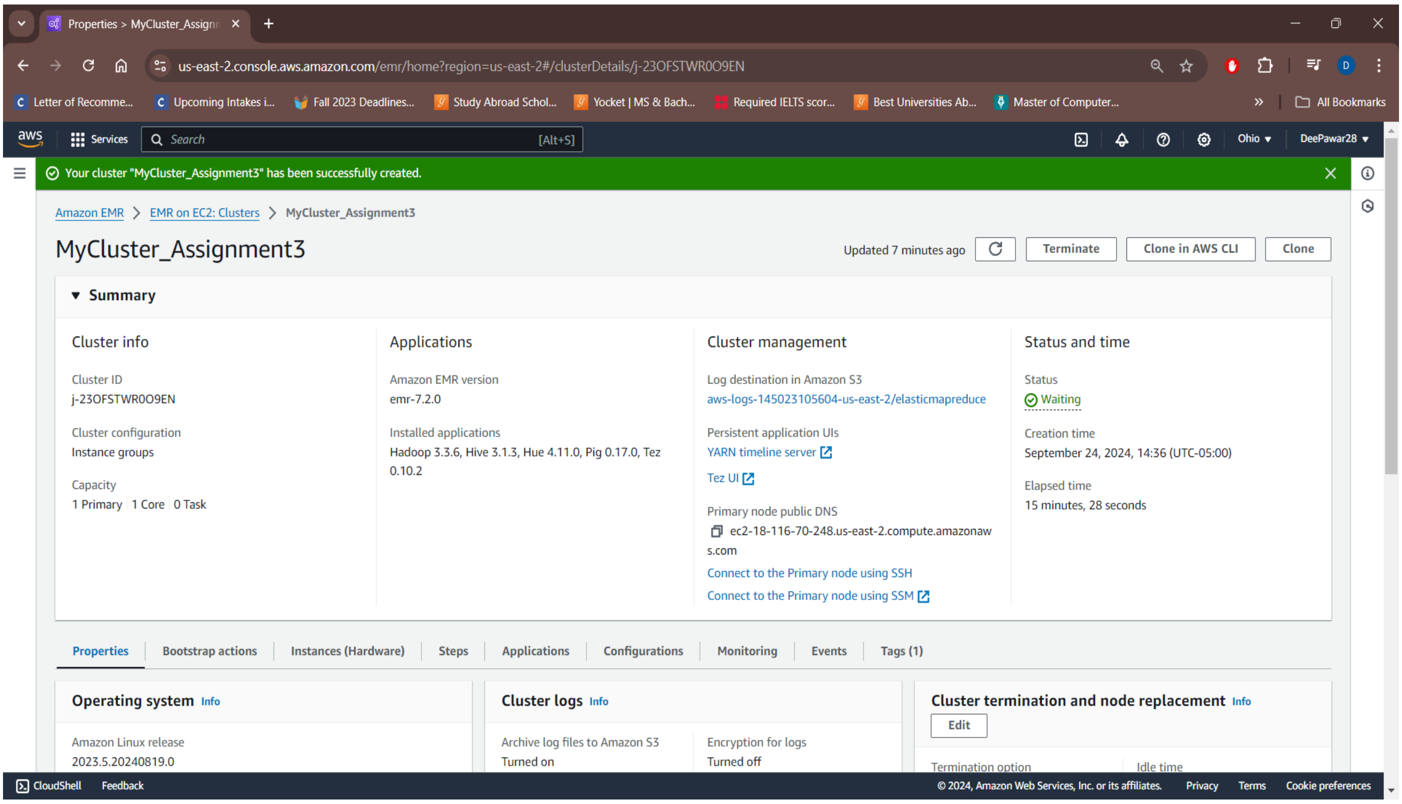
CSP 554: Big Data Technologies

Fall 2024 - Assignment 3

- **Questions and Answers:**

4) Create a new EMR cluster the same as you did previously. Since you already have a security key (“.pem” or “.cer” file) just use that one during cluster creation. Or, if you deleted your security key, just create a new one.

Ans:



Your cluster "MyCluster_Assignment3" has been successfully created.

Amazon EMR > EMR on EC2: Clusters > MyCluster_Assignment3

MyCluster_Assignment3

Summary

Cluster info

Cluster ID
j-23OFSTWR0O9EN

Cluster configuration
Instance groups

Capacity
1 Primary | 1 Core | 0 Task

Applications

Amazon EMR version
emr-7.2.0

Installed applications
Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, Pig 0.17.0, Tez 0.10.2

Cluster management

Log destination in Amazon S3
aws-logs-145023105604-us-east-2/elasticmapreduce

Persistent application UIs
YARN timeline server
Tez UI

Primary node public DNS
ec2-18-116-70-248.us-east-2.compute.amazonaws.com

Connect to the Primary node using SSH
Connect to the Primary node using SSM

Status and time

Status
Waiting

Creation time
September 24, 2024, 14:36 (UTC-05:00)

Elapsed time
15 minutes, 28 seconds

Operating system Info

Amazon Linux release
2023.5.20240819.0

Cluster logs Info

Archive log files to Amazon S3
Turned on

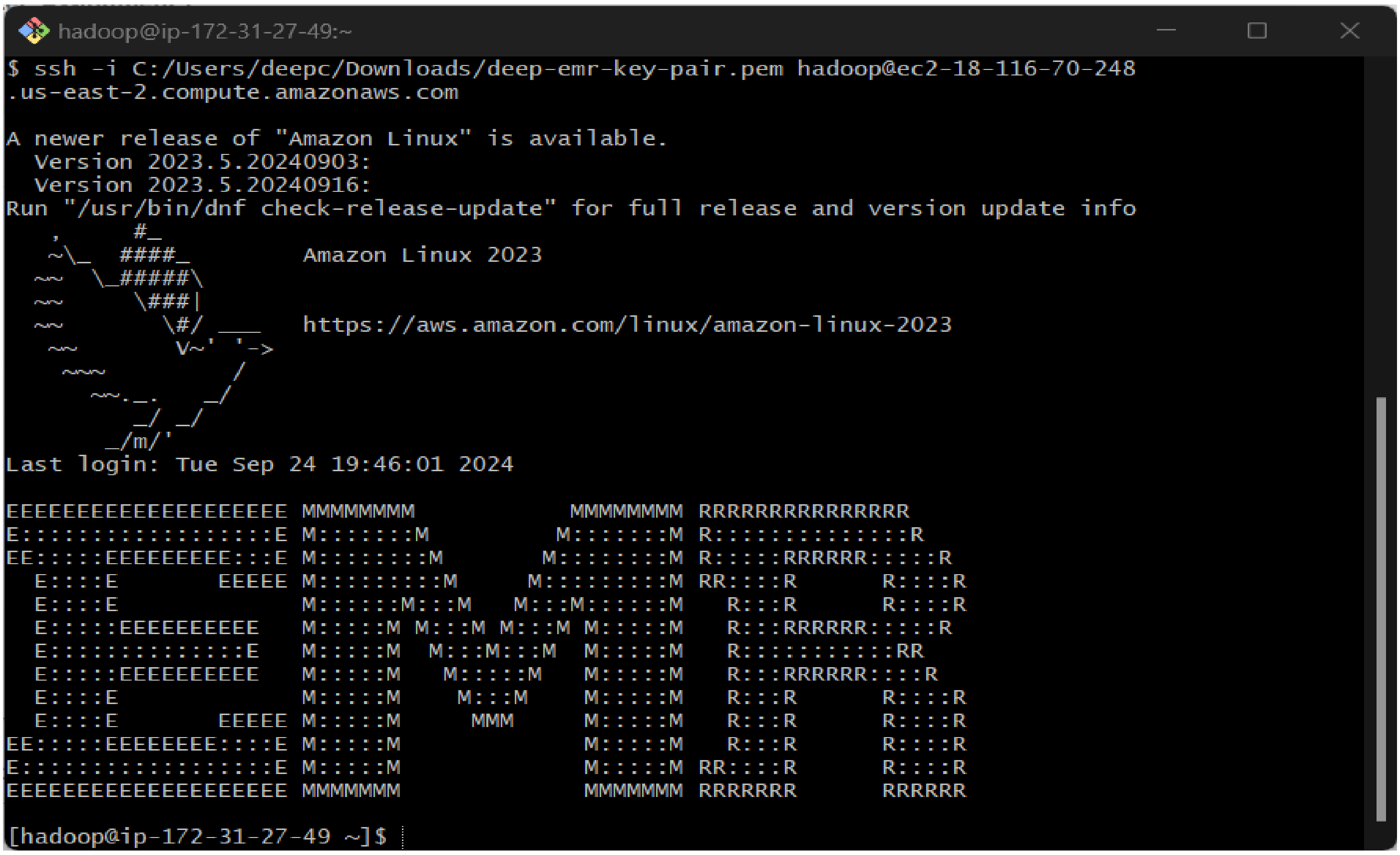
Encryption for logs
Turned off

Cluster termination and node replacement Info

5) Install the mrjob library on your EMR primary node.

a) ssh to the primary node (/home/hadoop) as you did in assignment #2

Ans:



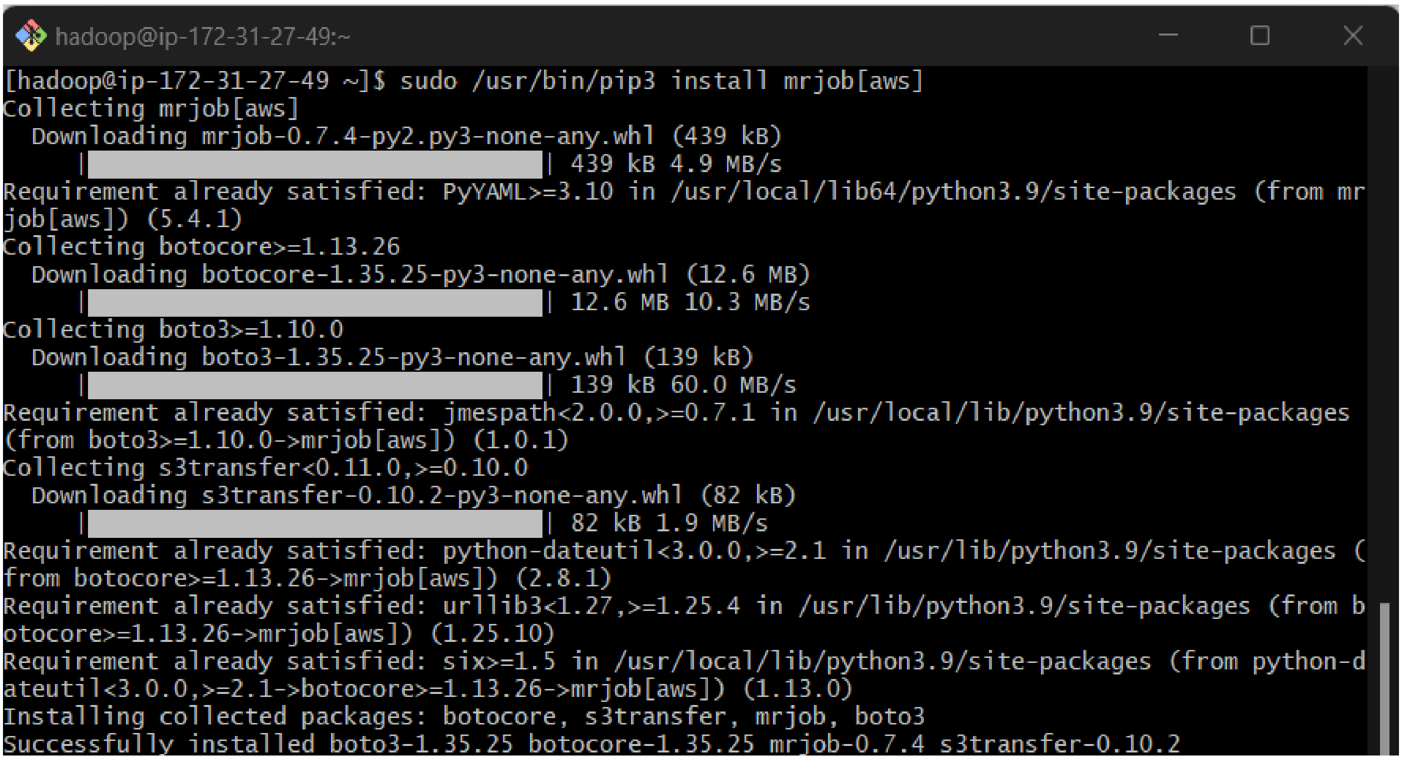
hadoop@ip-172-31-27-49:~

```
$ ssh -i C:/Users/deepc/Downloads/deep-emr-key-pair.pem hadoop@ec2-18-116-70-248
.us-east-2.compute.amazonaws.com

A newer release of "Amazon Linux" is available.
  Version 2023.5.20240903:
  Version 2023.5.20240916:
Run "/usr/bin/dnf check-release-update" for full release and version update info
   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Tue Sep 24 19:46:01 2024

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-172-31-27-49 ~]$
```

b) Enter the following (note if the first command does not work, try the second)

Ans:

- **Command:** sudo /usr/bin/pip3 install mrjob[aws]



hadoop@ip-172-31-27-49:~

```
[hadoop@ip-172-31-27-49 ~]$ sudo /usr/bin/pip3 install mrjob[aws]
Collecting mrjob[aws]
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
     |████████████████████████████████| 439 kB 4.9 MB/s
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.9/site-packages (from mr
job[aws]) (5.4.1)
Collecting botocore>=1.13.26
  Downloading botocore-1.35.25-py3-none-any.whl (12.6 MB)
     |████████████████████████████████| 12.6 MB 10.3 MB/s
Collecting boto3>=1.10.0
  Downloading boto3-1.35.25-py3-none-any.whl (139 kB)
     |████████████████████████████████| 139 kB 60.0 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)
Collecting s3transfer<0.11.0,>=0.10.0
  Downloading s3transfer-0.10.2-py3-none-any.whl (82 kB)
     |████████████████████████████████| 82 kB 1.9 MB/s
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: urllib3<1.27,>=1.25.4 in /usr/lib/python3.9/site-packages (from b
otocore>=1.13.26->mrjob[aws]) (1.25.10)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.9/site-packages (from python-d
ateutil<3.0.0,>=2.1->botocore>=1.13.26->mrjob[aws]) (1.13.0)
Installing collected packages: botocore, s3transfer, mrjob, boto3
Successfully installed boto3-1.35.25 botocore-1.35.25 mrjob-0.7.4 s3transfer-0.10.2
```

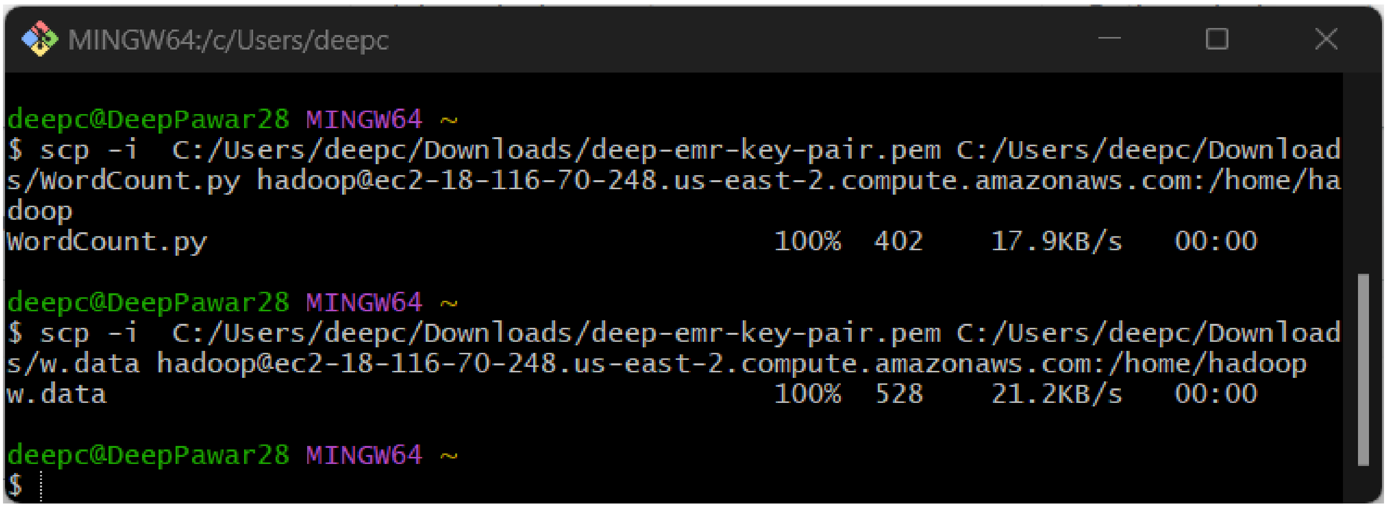
6) Next you will set up to execute the provided WordCount.py map reduce program found in the "Assignments" section of the Blackboard. This is the exact same program we saw in class.

- 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.

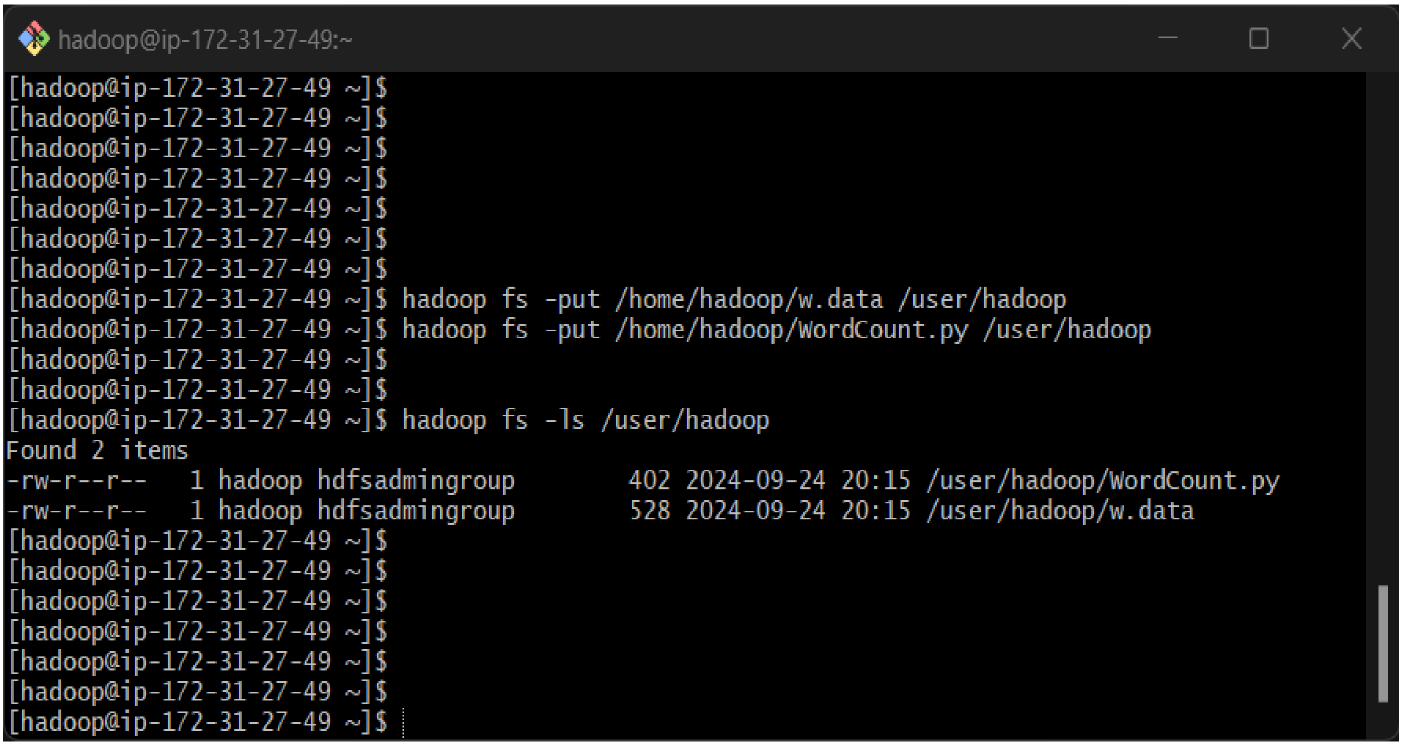
- Step 2:

Ans:



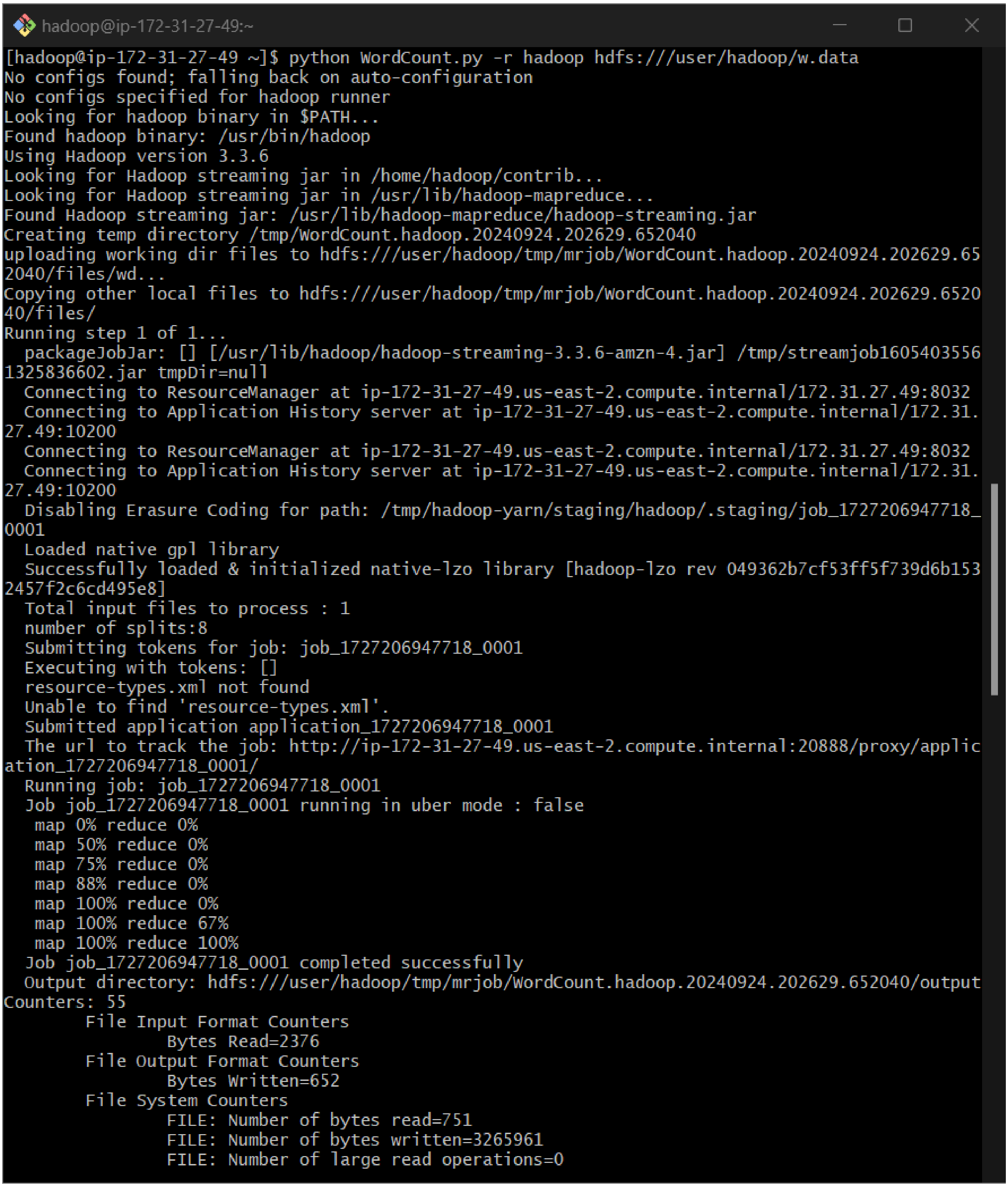
```
deepc@DeepPawar28 MINGW64 ~
$ scp -i  C:/Users/deepc/Downloads/deep-emr-key-pair.pem C:/Users/deepc/Download
s/WordCount.py hadoop@ec2-18-116-70-248.us-east-2.compute.amazonaws.com:/home/ha
doop
WordCount.py                                  100%  402    17.9KB/s   00:00

deepc@DeepPawar28 MINGW64 ~
$ scp -i  C:/Users/deepc/Downloads/deep-emr-key-pair.pem C:/Users/deepc/Download
s/w.data hadoop@ec2-18-116-70-248.us-east-2.compute.amazonaws.com:/home/hadoop
w.data                                        100%  528    21.2KB/s   00:00

deepc@DeepPawar28 MINGW64 ~
$
```

- Step 3:

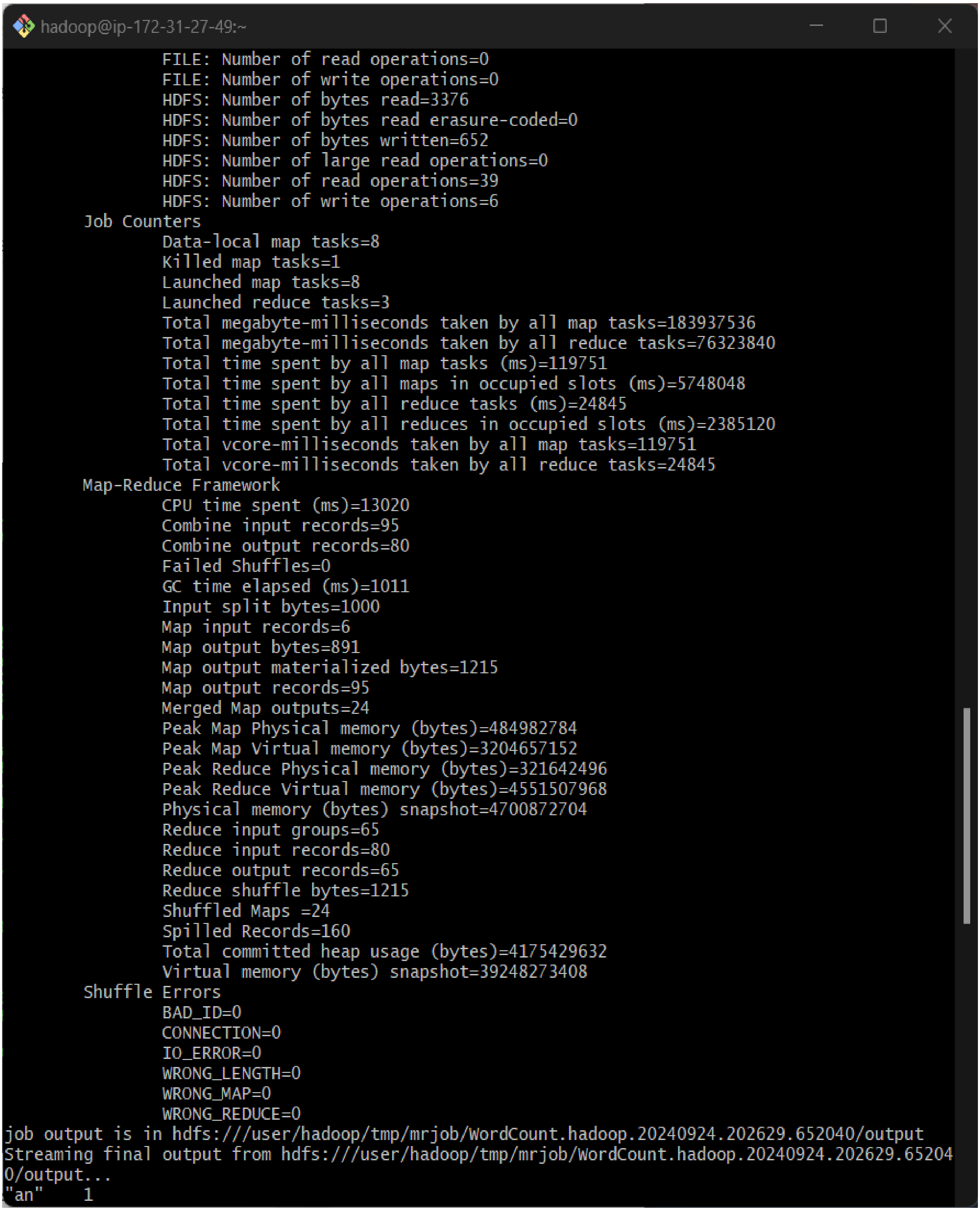


```
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$ hadoop fs -put /home/hadoop/w.data /user/hadoop
[hadoop@ip-172-31-27-49 ~]$ hadoop fs -put /home/hadoop/WordCount.py /user/hadoop
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$ hadoop fs -ls /user/hadoop
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup        402 2024-09-24 20:15 /user/hadoop/WordCount.py
-rw-r--r--   1 hadoop hdfsadmingroup        528 2024-09-24 20:15 /user/hadoop/w.data
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
[hadoop@ip-172-31-27-49 ~]$
```

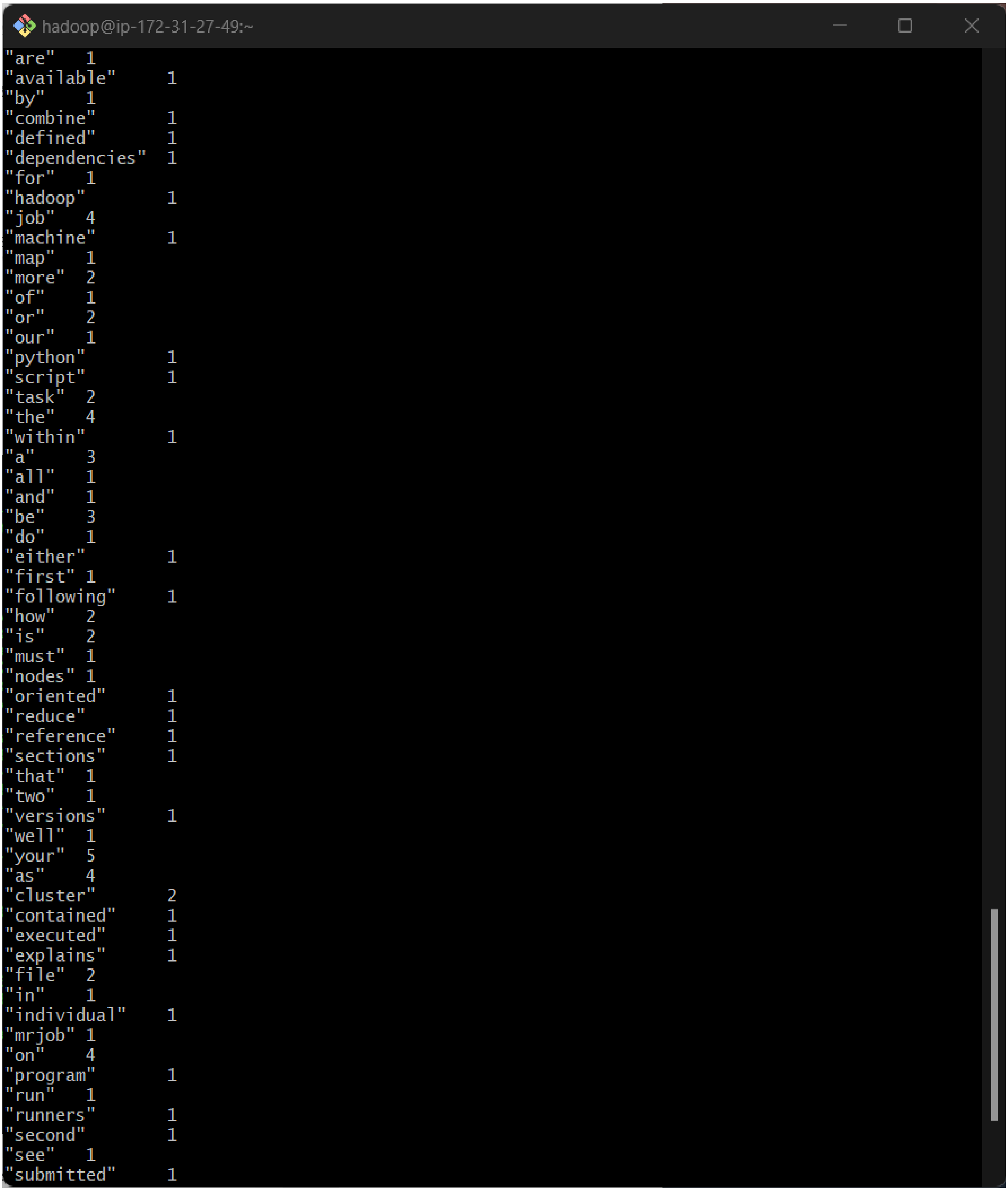
- Step 4:



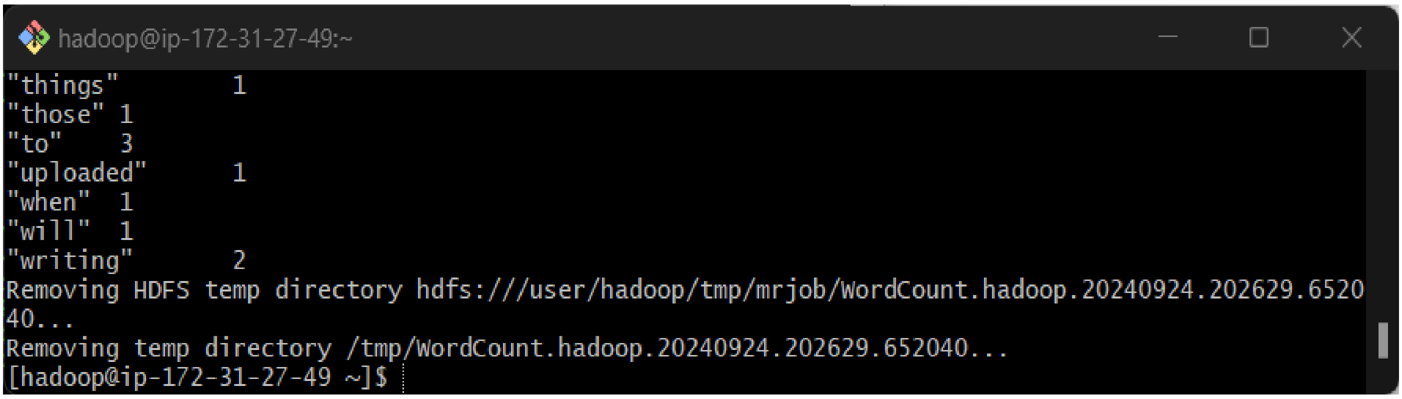
```
[hadoop@ip-172-31-27-49 ~]$ python WordCount.py -r hadoop hdfs:///user/hadoop/w.data
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.3.6
Looking for Hadoop 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.20240924.202629.652040
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20240924.202629.65
2040/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20240924.202629.6520
40/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob1605403556
1325836602.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1727206947718_
0001
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b153
2457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1727206947718_0001
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1727206947718_0001
  The url to track the job: http://ip-172-31-27-49.us-east-2.compute.internal:20888/proxy/applic
ation_1727206947718_0001/
  Running job: job_1727206947718_0001
  Job job_1727206947718_0001 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 75% reduce 0%
   map 88% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1727206947718_0001 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20240924.202629.652040/output
Counters: 55
        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=3265961
                FILE: Number of large read operations=0
```



```
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3376
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=652
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=39
                HDFS: Number of write operations=6
        Job Counters
                Data-local map tasks=8
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=183937536
                Total megabyte-milliseconds taken by all reduce tasks=76323840
                Total time spent by all map tasks (ms)=119751
                Total time spent by all maps in occupied slots (ms)=5748048
                Total time spent by all reduce tasks (ms)=24845
                Total time spent by all reduces in occupied slots (ms)=2385120
                Total vcore-milliseconds taken by all map tasks=119751
                Total vcore-milliseconds taken by all reduce tasks=24845
        Map-Reduce Framework
                CPU time spent (ms)=13020
                Combine input records=95
                Combine output records=80
                Failed Shuffles=0
                GC time elapsed (ms)=1011
                Input split bytes=1000
                Map input records=6
                Map output bytes=891
                Map output materialized bytes=1215
                Map output records=95
                Merged Map outputs=24
                Peak Map Physical memory (bytes)=484982784
                Peak Map Virtual memory (bytes)=3204657152
                Peak Reduce Physical memory (bytes)=321642496
                Peak Reduce Virtual memory (bytes)=4551507968
                Physical memory (bytes) snapshot=4700872704
                Reduce input groups=65
                Reduce input records=80
                Reduce output records=65
                Reduce shuffle bytes=1215
                Shuffled Maps =24
                Spilled Records=160
                Total committed heap usage (bytes)=4175429632
                Virtual memory (bytes) snapshot=39248273408
        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/WordCount.hadoop.20240924.202629.652040/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20240924.202629.65204
0/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
"explains"      1
"file"  2
"in"    1
"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.20240924.202629.6520
40...
Removing temp directory /tmp/WordCount.hadoop.20240924.202629.652040...
[hadoop@ip-172-31-27-49 ~]$ 
```

5) 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 lower-case letters a-n (a through n inclusive) and how many begin with anything else.

The output file should look something like

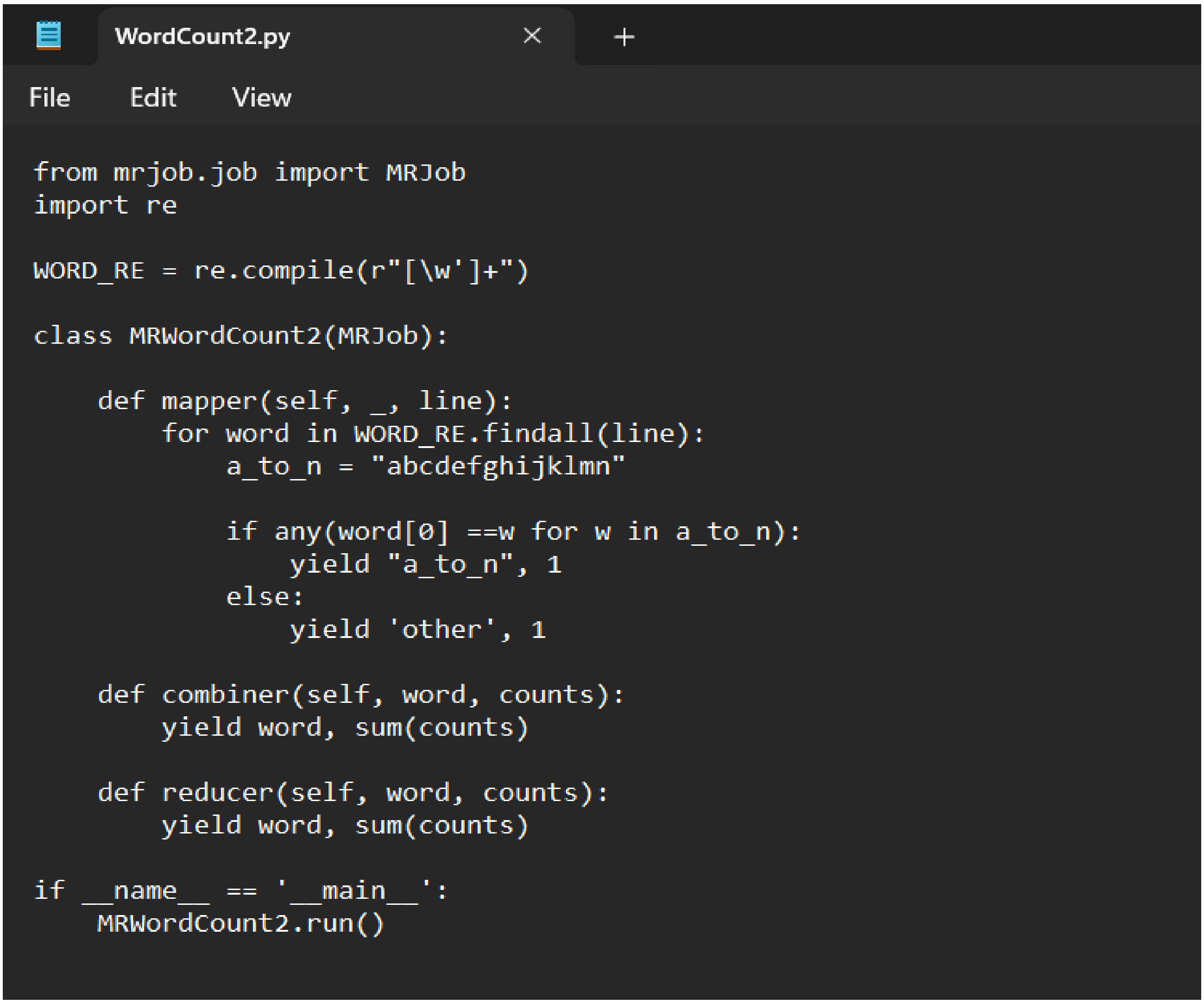
a_to_n, 12

other, 21

Note, do not force words to all lower case. Now execute the program and see what happens.

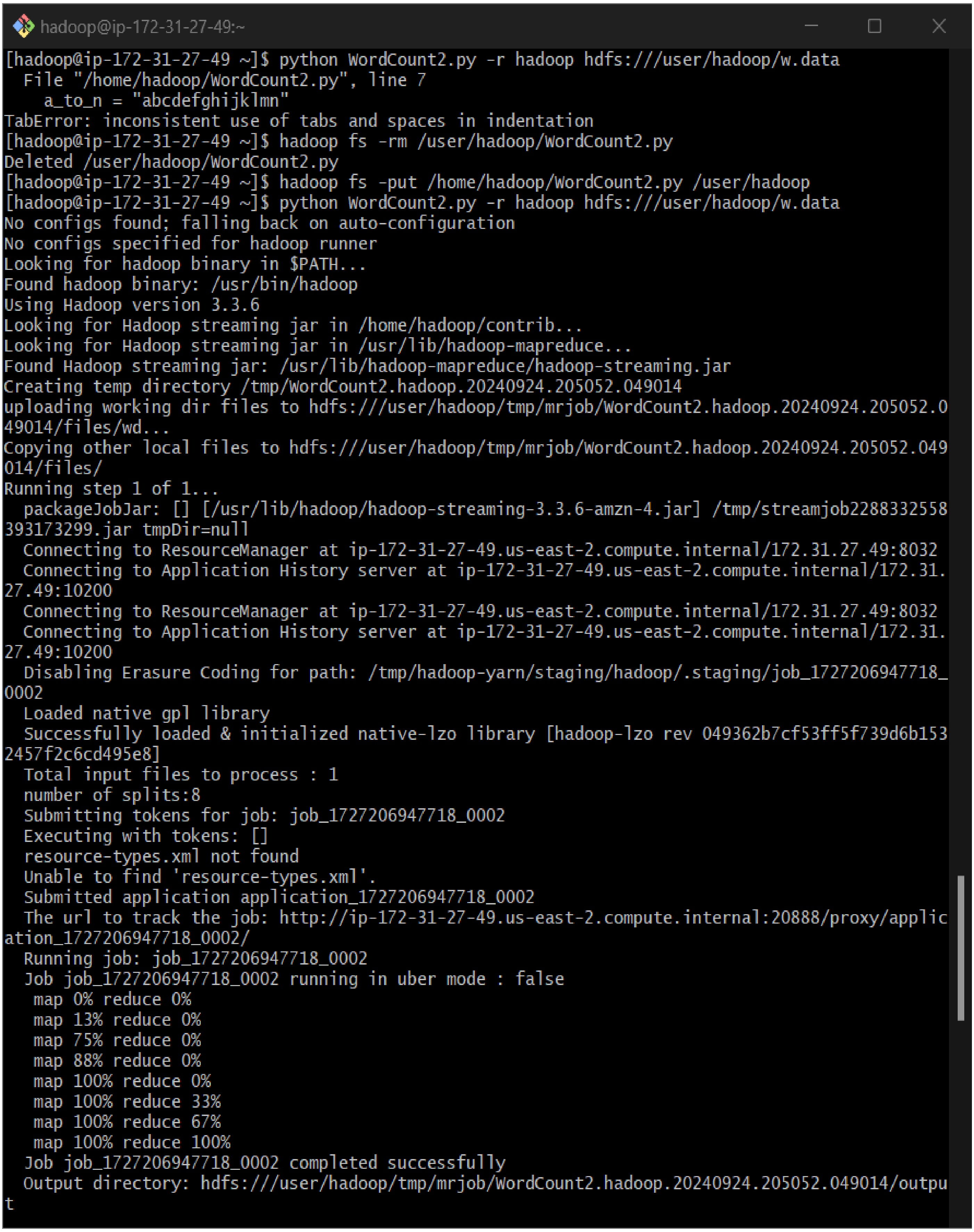
6) (3 points) Submit (1) a copy of this modified program and (2) a screen shot of the results of the program's execution as the output of your assignment.

- **Code: WordCount2.py**

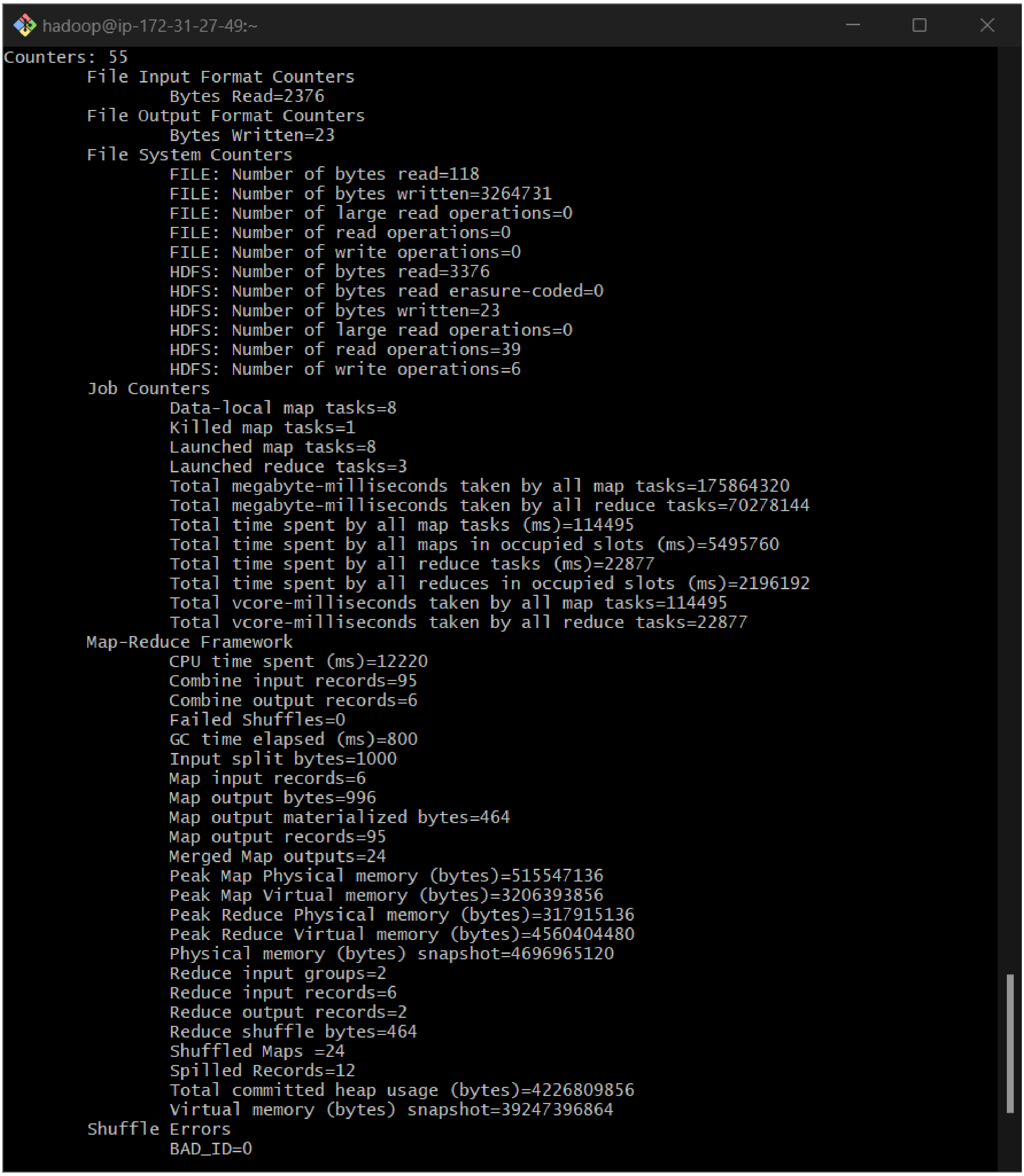


```
from mrjob.job import MRJob
import re

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

class MRWordCount2(MRJob):

    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            a_to_n = "abcdefghijklmn"

            if any(word[0] ==w for w in a_to_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__':
    MRWordCount2.run()
```

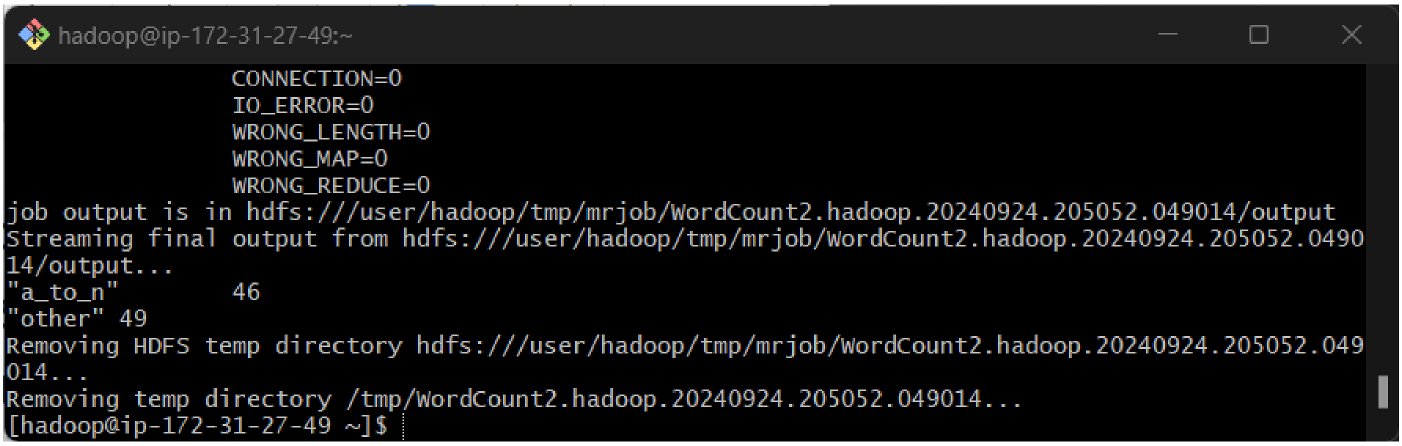
- **Output:**



```
[hadoop@ip-172-31-27-49 ~]$ python WordCount2.py -r hadoop hdfs:///user/hadoop/w.data
  File "/home/hadoop/WordCount2.py", line 7
    a_to_n = "abcdefghijklmn"
TabError: inconsistent use of tabs and spaces in indentation
[hadoop@ip-172-31-27-49 ~]$ hadoop fs -rm /user/hadoop/WordCount2.py
Deleted /user/hadoop/WordCount2.py
[hadoop@ip-172-31-27-49 ~]$ hadoop fs -put /home/hadoop/WordCount2.py /user/hadoop
[hadoop@ip-172-31-27-49 ~]$ 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.20240924.205052.049014
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240924.205052.0
49014/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240924.205052.049
014/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob2288332558
393173299.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1727206947718_
0002
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b153
2457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1727206947718_0002
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1727206947718_0002
  The url to track the job: http://ip-172-31-27-49.us-east-2.compute.internal:20888/proxy/applic
ation_1727206947718_0002/
  Running job: job_1727206947718_0002
  Job job_1727206947718_0002 running in uber mode : false
   map 0% reduce 0%
   map 13% reduce 0%
   map 75% reduce 0%
   map 88% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1727206947718_0002 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240924.205052.049014/outpu
t
```



```
Counters: 55
        File Input Format Counters
                Bytes Read=2376
        File Output Format Counters
                Bytes Written=23
        File System Counters
                FILE: Number of bytes read=118
                FILE: Number of bytes written=3264731
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3376
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=23
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=39
                HDFS: Number of write operations=6
        Job Counters
                Data-local map tasks=8
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=175864320
                Total megabyte-milliseconds taken by all reduce tasks=70278144
                Total time spent by all map tasks (ms)=114495
                Total time spent by all maps in occupied slots (ms)=5495760
                Total time spent by all reduce tasks (ms)=22877
                Total time spent by all reduces in occupied slots (ms)=2196192
                Total vcore-milliseconds taken by all map tasks=114495
                Total vcore-milliseconds taken by all reduce tasks=22877
        Map-Reduce Framework
                CPU time spent (ms)=12220
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=800
                Input split bytes=1000
                Map input records=6
                Map output bytes=996
                Map output materialized bytes=464
                Map output records=95
                Merged Map outputs=24
                Peak Map Physical memory (bytes)=515547136
                Peak Map Virtual memory (bytes)=3206393856
                Peak Reduce Physical memory (bytes)=317915136
                Peak Reduce Virtual memory (bytes)=4560404480
                Physical memory (bytes) snapshot=4696965120
                Reduce input groups=2
                Reduce input records=6
                Reduce output records=2
                Reduce shuffle bytes=464
                Shuffled Maps =24
                Spilled Records=12
                Total committed heap usage (bytes)=4226809856
                Virtual memory (bytes) snapshot=39247396864
        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.20240924.205052.049014/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240924.205052.0490
14/output...
"a_to_n"        46
"other" 49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240924.205052.049
014...
Removing temp directory /tmp/WordCount2.hadoop.20240924.205052.049014...
[hadoop@ip-172-31-27-49 ~]$
```

7) Let’s modify the WordCount.py program again. Call the new program WordCount3.py.

Instead of counting words, calculate the count of words having the same number of letters. For example, if we have a file consisting of one record of the form:

hello there joe

our job should output key value pairs similar to the following:

3, 1

5, 2

Hint, the key in a key-value pair can be an integer just as well as a string.

So, your task is to write a MrJob MapReduce program which again accepts the following file as input

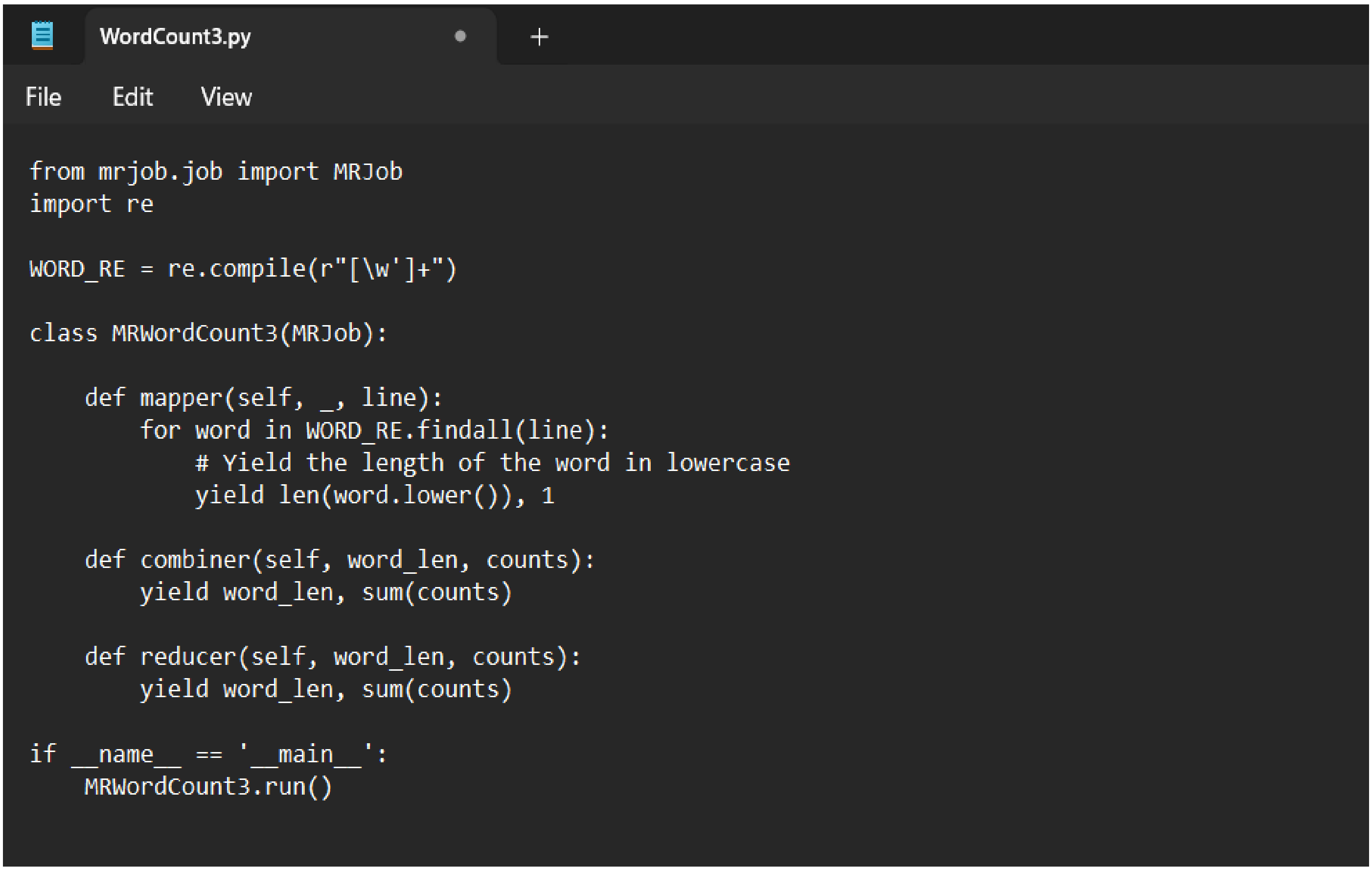
hdfs:///user/hadoop/w.data

and outputs key value pairs where each one has a key with is some number of characters, and the value a count of words having that many characters. Note, please convert all words to lower case on input, so “Hello” and “hello” become the same word.

8) (4 points) When you have accomplished this, please submit the following, (1) a copy of your MRJob code and (2) a copy of the output of the execution of that code.

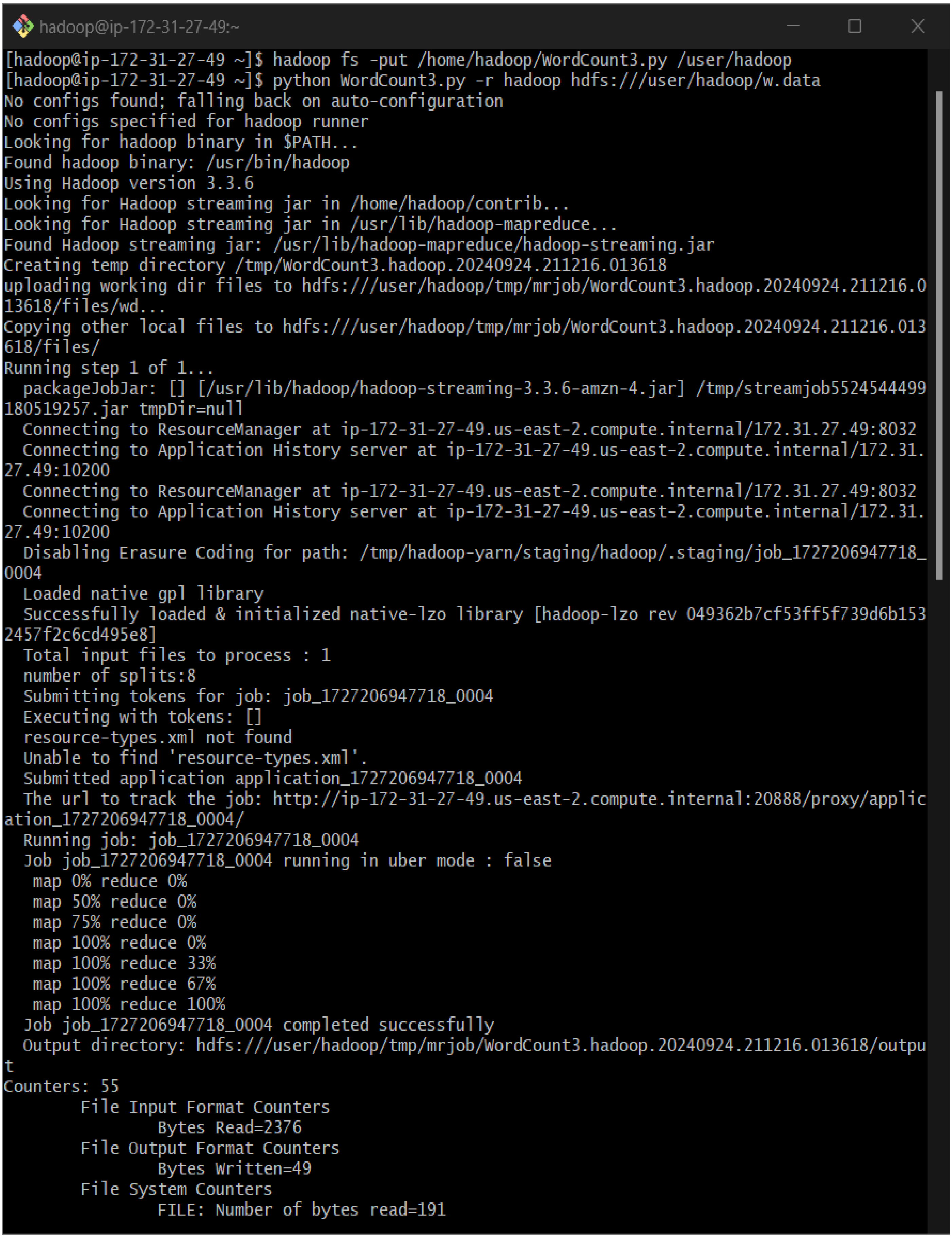
Ans:

- **Code: WordCount3.py**

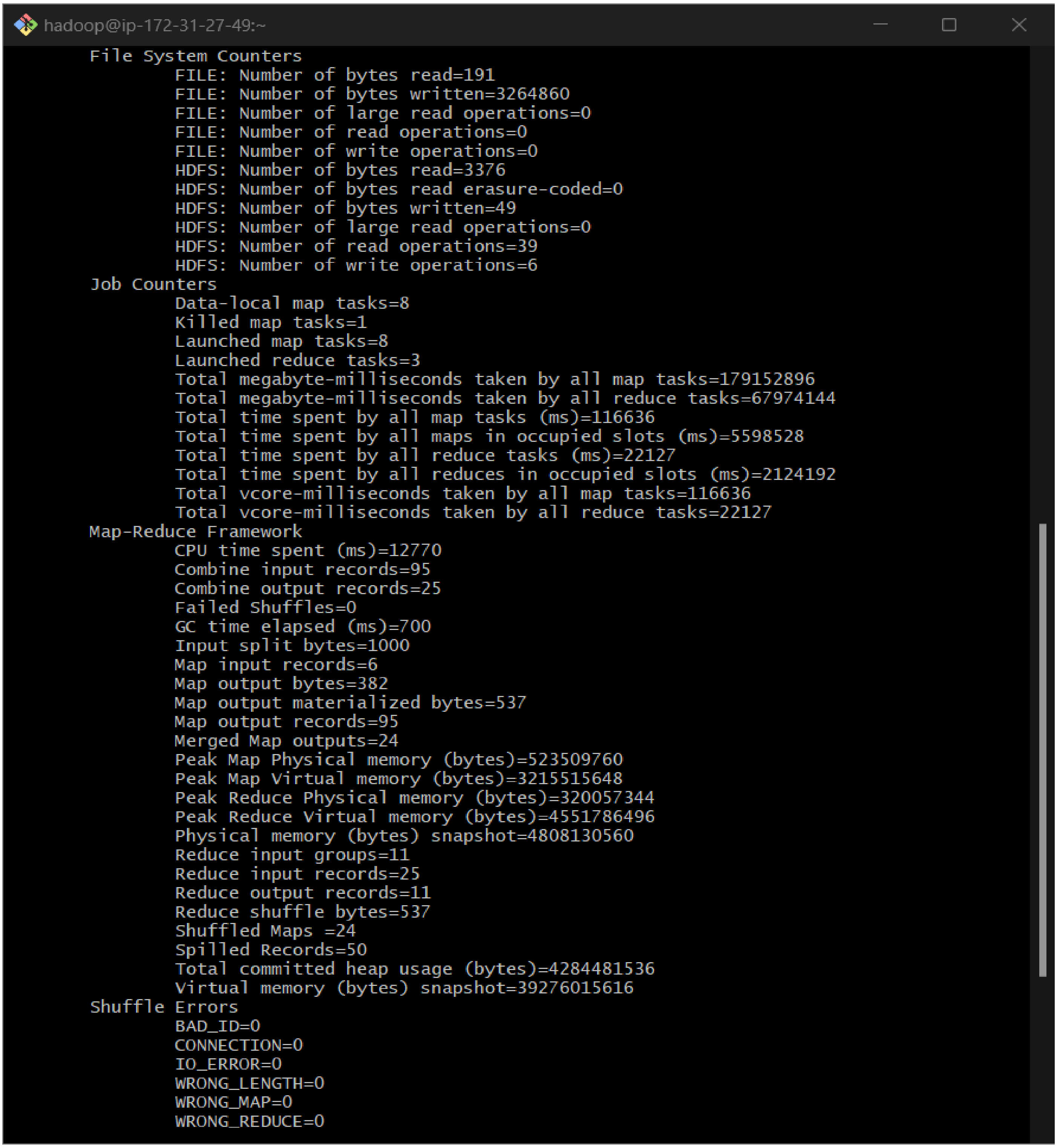


```
from mrjob.job import MRJob
import re

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

class MRWordCount3(MRJob):

    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            # Yield the length of the word in lowercase
            yield len(word.lower()), 1

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

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

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

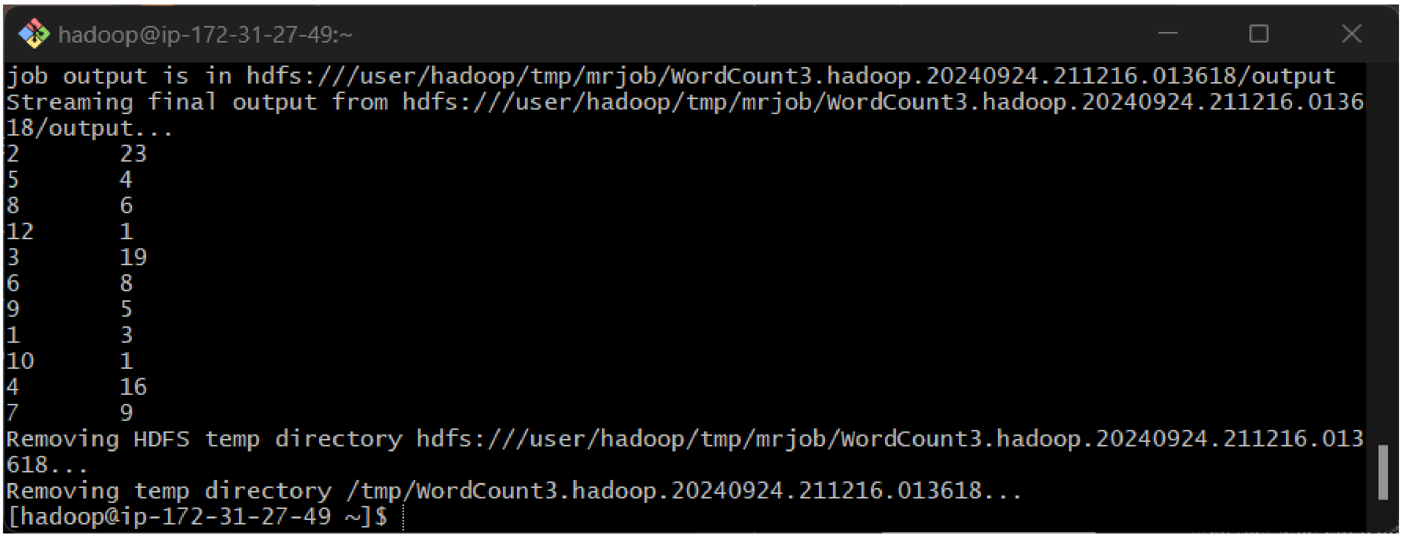
- **Output:**



```
[hadoop@ip-172-31-27-49 ~]$ hadoop fs -put /home/hadoop/WordCount3.py /user/hadoop
[hadoop@ip-172-31-27-49 ~]$ 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.20240924.211216.013618
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240924.211216.0
13618/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240924.211216.013
618/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob5524544499
180519257.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1727206947718_
0004
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b153
2457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1727206947718_0004
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1727206947718_0004
  The url to track the job: http://ip-172-31-27-49.us-east-2.compute.internal:20888/proxy/applic
ation_1727206947718_0004/
  Running job: job_1727206947718_0004
  Job job_1727206947718_0004 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1727206947718_0004 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240924.211216.013618/outpu
t
Counters: 55
        File Input Format Counters
                Bytes Read=2376
        File Output Format Counters
                Bytes Written=49
        File System Counters
                FILE: Number of bytes read=191
```



```
        File System Counters
                FILE: Number of bytes read=191
                FILE: Number of bytes written=3264860
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3376
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=49
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=39
                HDFS: Number of write operations=6
        Job Counters
                Data-local map tasks=8
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=179152896
                Total megabyte-milliseconds taken by all reduce tasks=67974144
                Total time spent by all map tasks (ms)=116636
                Total time spent by all maps in occupied slots (ms)=5598528
                Total time spent by all reduce tasks (ms)=22127
                Total time spent by all reduces in occupied slots (ms)=2124192
                Total vcore-milliseconds taken by all map tasks=116636
                Total vcore-milliseconds taken by all reduce tasks=22127
        Map-Reduce Framework
                CPU time spent (ms)=12770
                Combine input records=95
                Combine output records=25
                Failed Shuffles=0
                GC time elapsed (ms)=700
                Input split bytes=1000
                Map input records=6
                Map output bytes=382
                Map output materialized bytes=537
                Map output records=95
                Merged Map outputs=24
                Peak Map Physical memory (bytes)=523509760
                Peak Map Virtual memory (bytes)=3215515648
                Peak Reduce Physical memory (bytes)=320057344
                Peak Reduce Virtual memory (bytes)=4551786496
                Physical memory (bytes) snapshot=4808130560
                Reduce input groups=11
                Reduce input records=25
                Reduce output records=11
                Reduce shuffle bytes=537
                Shuffled Maps =24
                Spilled Records=50
                Total committed heap usage (bytes)=4284481536
                Virtual memory (bytes) snapshot=39276015616
        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/WordCount3.hadoop.20240924.211216.013618/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240924.211216.0136
18/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.20240924.211216.013
618...
Removing temp directory /tmp/WordCount3.hadoop.20240924.211216.013618...
[hadoop@ip-172-31-27-49 ~]$
```

9) Again, modify the WordCount.py program. Call the new program WordCount4.py.

Now we will write a MapReduce job to calculate the count of unique per record word bigrams. A word bigram is a two word sequence. For example, if we have a file consisting of records of the form:

> hello there joe
> hi there
> there joe go
> joe

Bigrams for these records are create by sliding a two word "window" across the words of the record.

> hello there joe => "hello there", "there joe"
> hi there => "hi there"
> there joe there => "there joe", "joe there"
> joe => *Note, this record has no bigrams*

Notice, that there are 2 instances of the word bigram "there Joe".

So, your task is to write a MrJob MapReduce program which accepts the following file as input

> hdfs:///user/hadoop/w.data

and outputs key value pairs where each one has a key which is some word bigram string, and the value a count of the number of occurrences of that word bigram. Note, please convert all words to lower case on input, so Hello and hello become the same word.

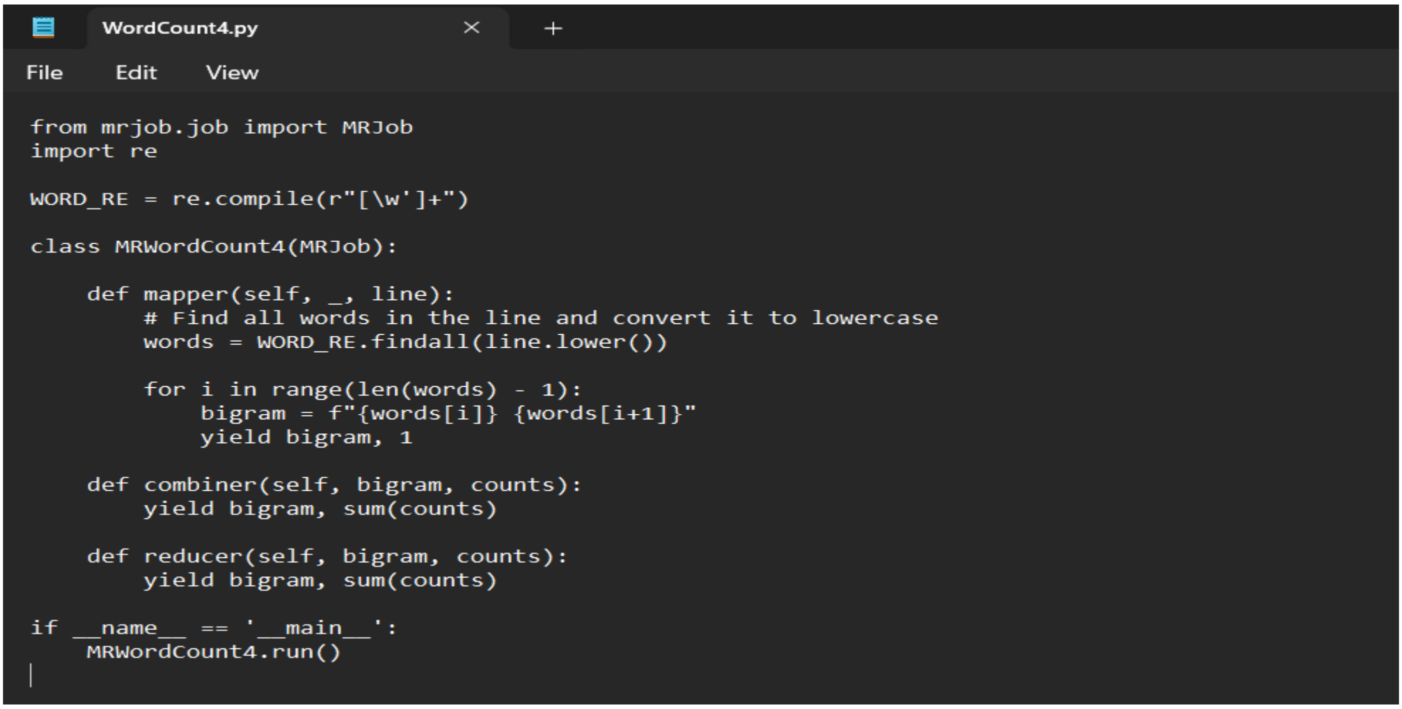
Our job should output key value pairs similar to the following:

> "hello there", 1
> "hi there", 1
> "joe there", 1
> "there joe", 2

10) (5 points) When you have accomplished this, please submit the following, (1) a copy of your MRJob code and (2) a copy of the output of the execution of that code for at least the first 10 bigram key value pairs.

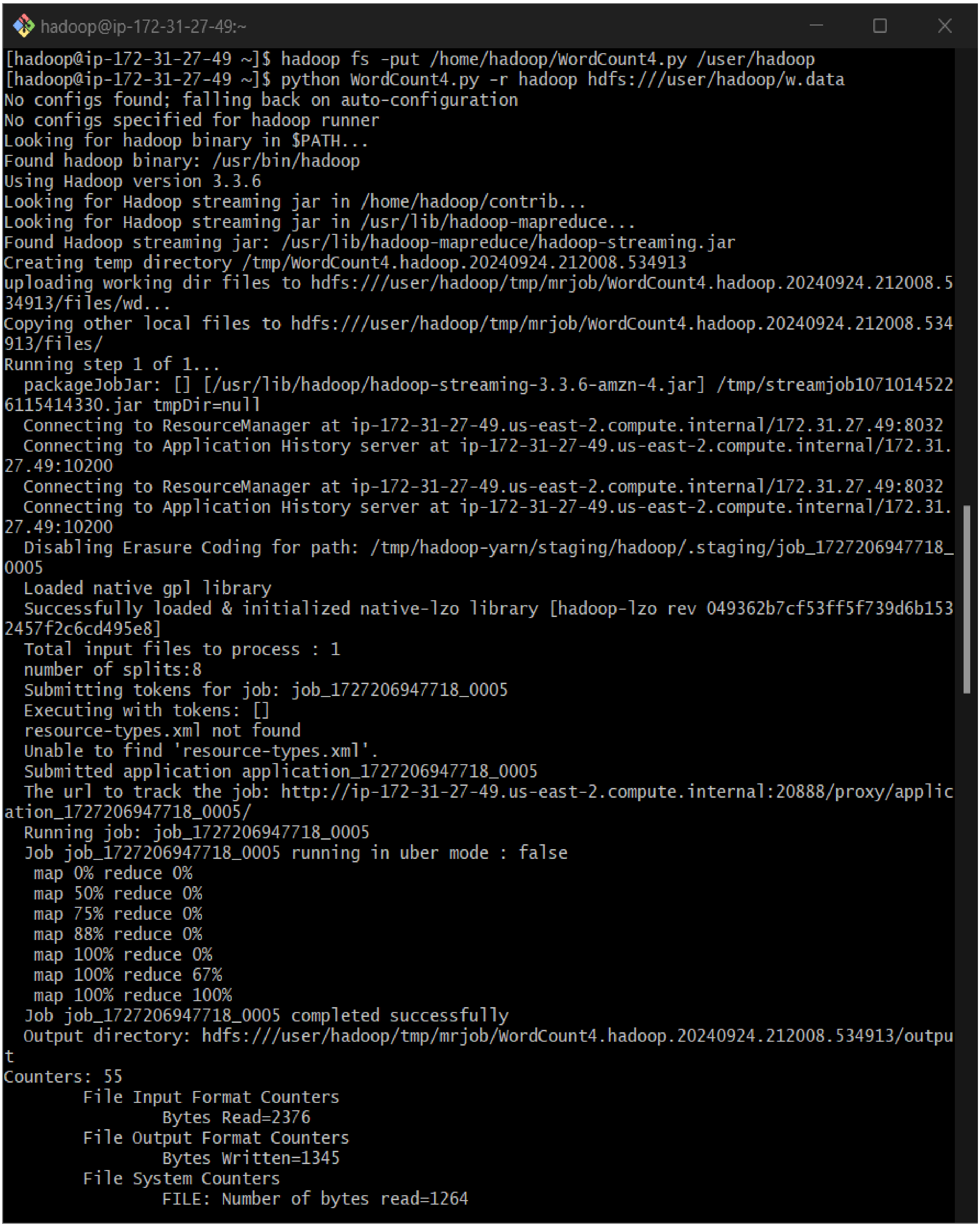
Ans:

- **Code: WordCount4.py**

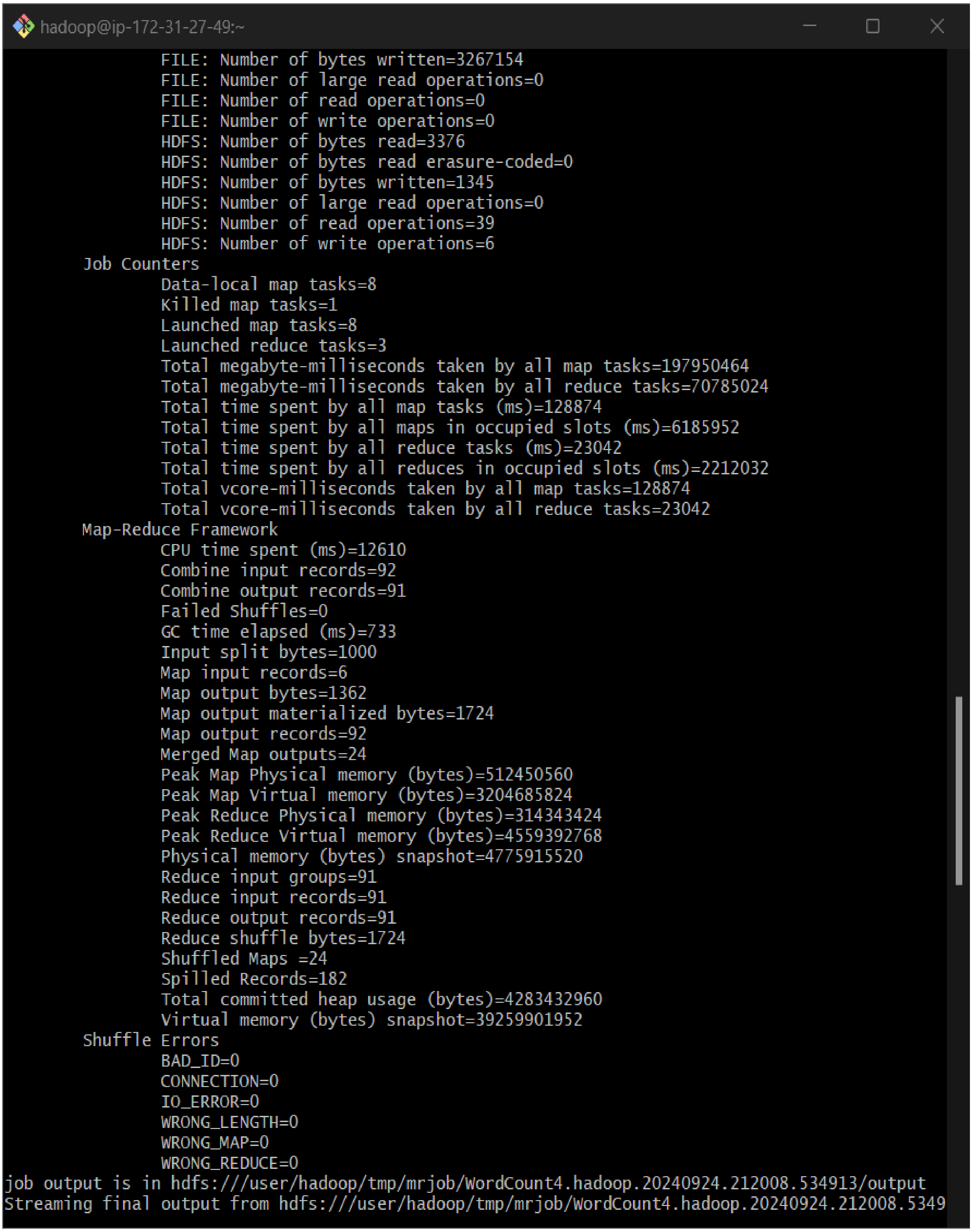


```
from mrjob.job import MRJob
import re

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

class MRWordCount4(MRJob):

    def mapper(self, _, line):
        # Find all words in the line and convert it to lowercase
        words = WORD_RE.findall(line.lower())

        for i in range(len(words) - 1):
            bigram = f"{words[i]} {words[i+1]}"
            yield bigram, 1

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

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

if __name__ == '__main__':
    MRWordCount4.run()
```

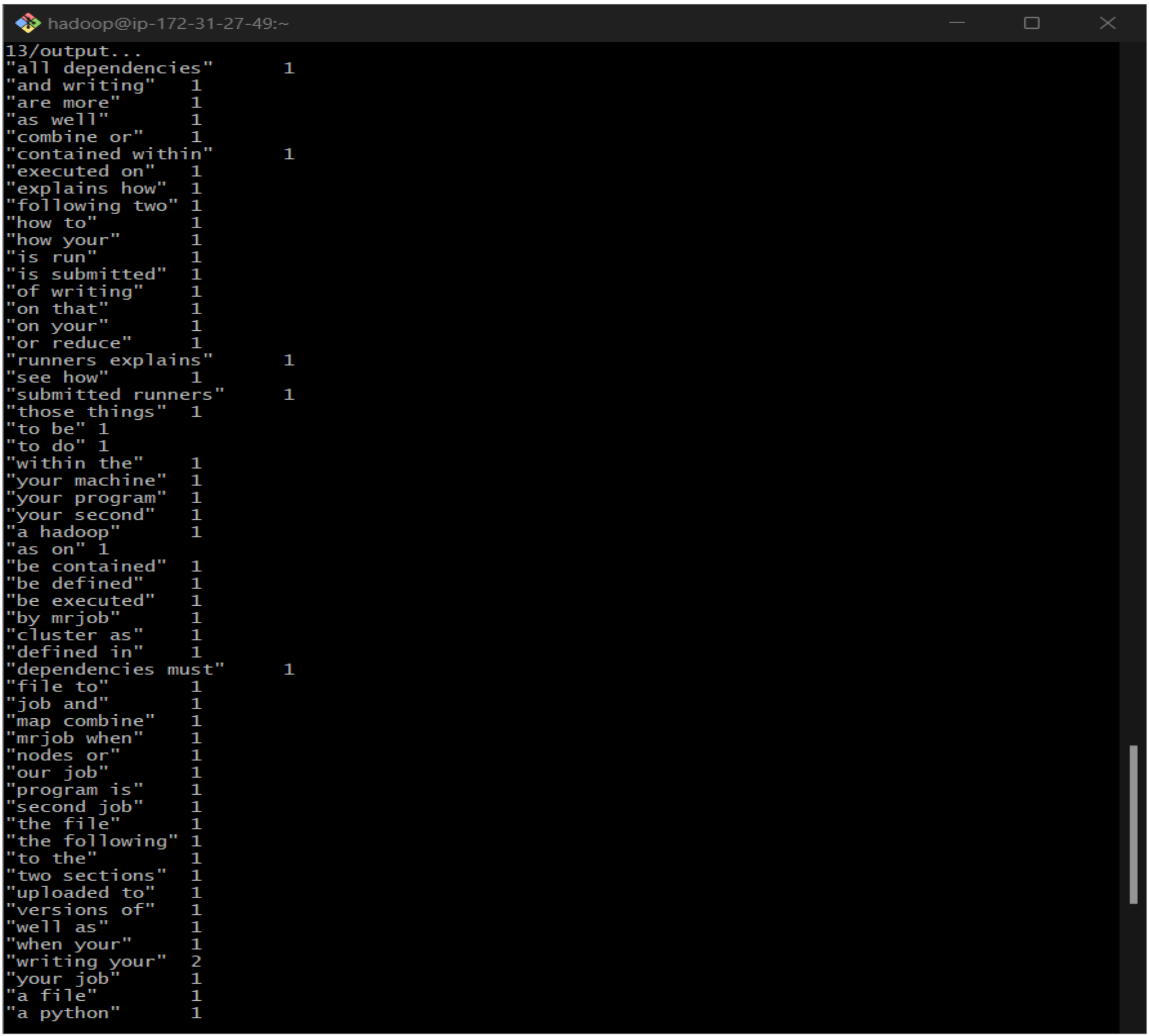
- **Output:**



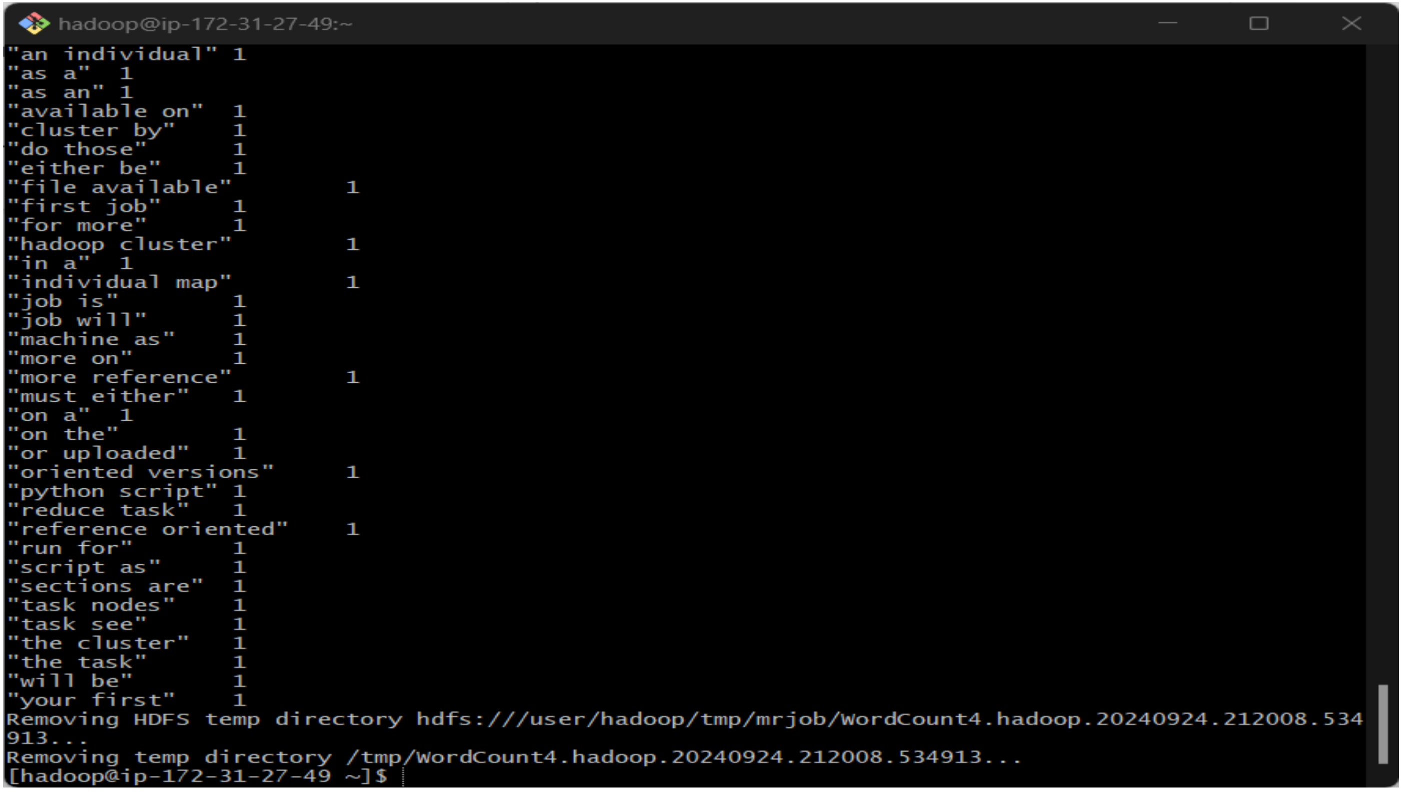
```
[hadoop@ip-172-31-27-49 ~]$ hadoop fs -put /home/hadoop/WordCount4.py /user/hadoop
[hadoop@ip-172-31-27-49 ~]$ python WordCount4.py -r hadoop hdfs:///user/hadoop/w.data
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
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.20240924.212008.534913
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240924.212008.5
34913/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240924.212008.534
913/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob1071014522
6115414330.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1727206947718_
0005
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b153
2457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1727206947718_0005
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1727206947718_0005
  The url to track the job: http://ip-172-31-27-49.us-east-2.compute.internal:20888/proxy/applic
ation_1727206947718_0005/
  Running job: job_1727206947718_0005
  Job job_1727206947718_0005 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 75% reduce 0%
   map 88% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1727206947718_0005 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240924.212008.534913/outpu
t
Counters: 55
        File Input Format Counters
                Bytes Read=2376
        File Output Format Counters
                Bytes Written=1345
        File System Counters
                FILE: Number of bytes read=1264
```



```
                FILE: Number of bytes written=3267154
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3376
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=1345
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=39
                HDFS: Number of write operations=6
        Job Counters
                Data-local map tasks=8
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=197950464
                Total megabyte-milliseconds taken by all reduce tasks=70785024
                Total time spent by all map tasks (ms)=128874
                Total time spent by all maps in occupied slots (ms)=6185952
                Total time spent by all reduce tasks (ms)=23042
                Total time spent by all reduces in occupied slots (ms)=2212032
                Total vcore-milliseconds taken by all map tasks=128874
                Total vcore-milliseconds taken by all reduce tasks=23042
        Map-Reduce Framework
                CPU time spent (ms)=12610
                Combine input records=92
                Combine output records=91
                Failed Shuffles=0
                GC time elapsed (ms)=733
                Input split bytes=1000
                Map input records=6
                Map output bytes=1362
                Map output materialized bytes=1724
                Map output records=92
                Merged Map outputs=24
                Peak Map Physical memory (bytes)=512450560
                Peak Map Virtual memory (bytes)=3204685824
                Peak Reduce Physical memory (bytes)=314343424
                Peak Reduce Virtual memory (bytes)=4559392768
                Physical memory (bytes) snapshot=4775915520
                Reduce input groups=91
                Reduce input records=91
                Reduce output records=91
                Reduce shuffle bytes=1724
                Shuffled Maps =24
                Spilled Records=182
                Total committed heap usage (bytes)=4283432960
                Virtual memory (bytes) snapshot=39259901952
        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/WordCount4.hadoop.20240924.212008.534913/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240924.212008.5349
```



```
13/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
"to the"        1
"two sections"  1
"uploaded to"   1
"versions of"   1
"well as"       1
"when your"     1
"writing your"  2
"your job"      1
"a file"        1
"a python"      1
```



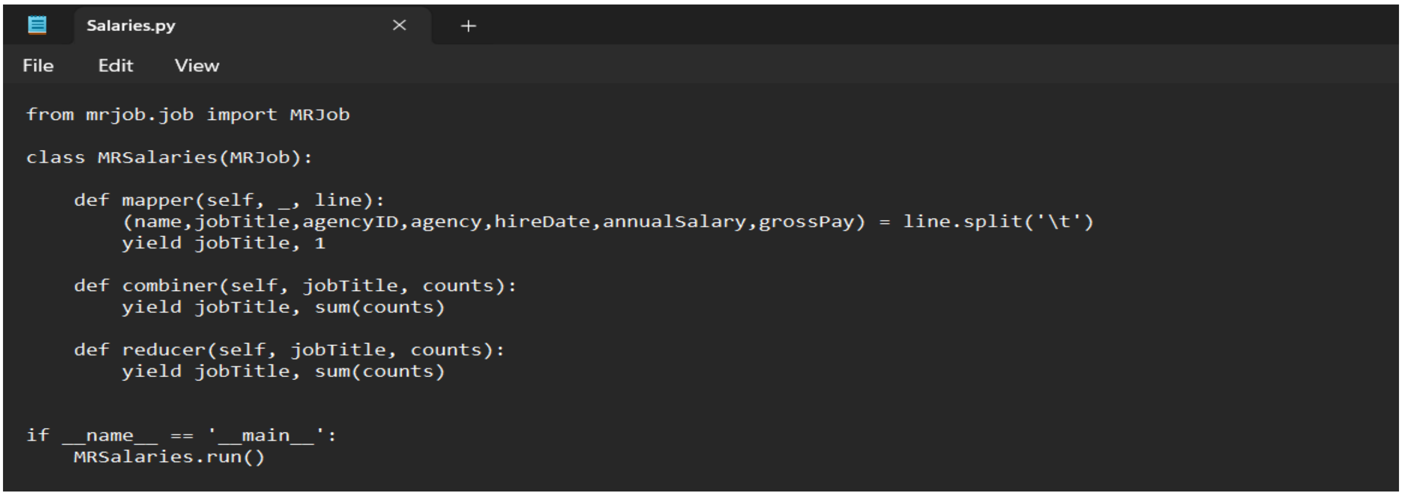
```
"an individual" 1
"as a"  1
"as an" 1
"available on"  1
"cluster by"    1
"do those"      1
"either be"     1
"file available"        1
"first job"     1
"for more"      1
"hadoop cluster"        1
"in a"  1
"individual map"        1
"job is"        1
"job will"      1
"machine as"    1
"more on"       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.20240924.212008.534
913...
Removing temp directory /tmp/WordCount4.hadoop.20240924.212008.534913...
[hadoop@ip-172-31-27-49 ~]$
```

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.

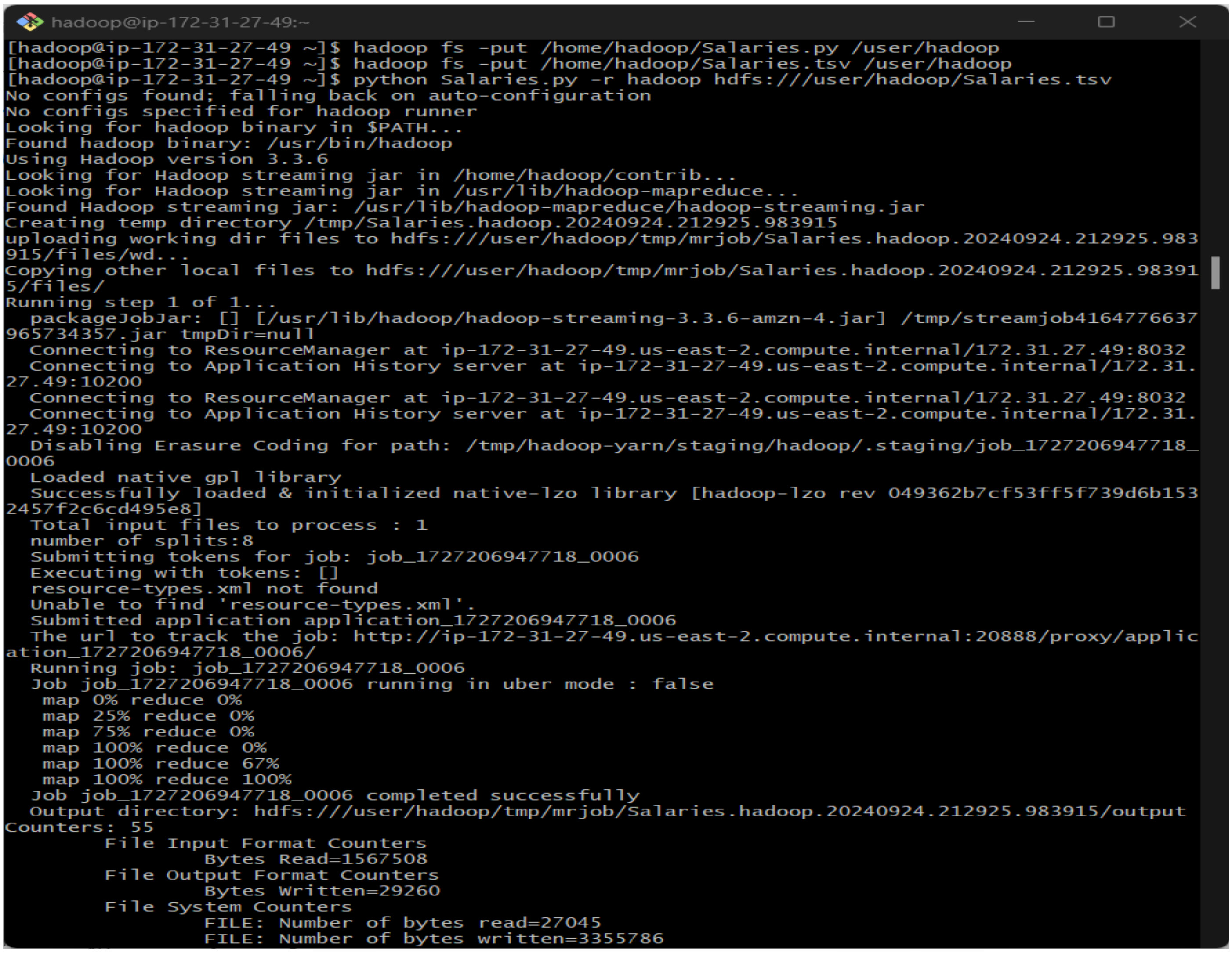
Ans:

- **Code: Salaries.py**

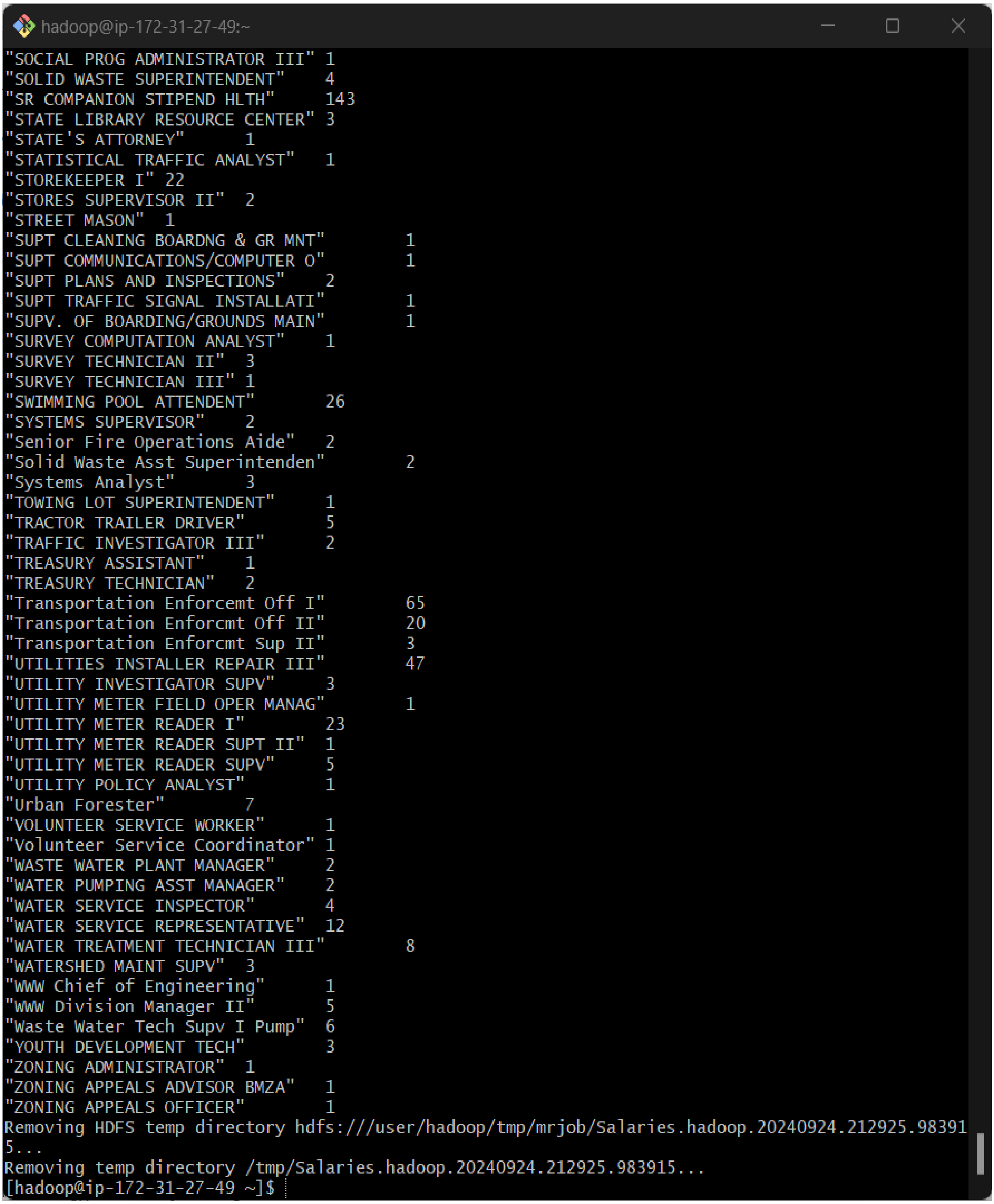


```
from mrjob.job import MRJob

class MRSalaries(MRJob):

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

- **Output:**



```
[hadoop@ip-172-31-27-49 ~]$ hadoop fs -put /home/hadoop/Salaries.py /user/hadoop
[hadoop@ip-172-31-27-49 ~]$ hadoop fs -put /home/hadoop/Salaries.tsv /user/hadoop
[hadoop@ip-172-31-27-49 ~]$ 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.20240924.212925.983915
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20240924.212925.983
915/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20240924.212925.98391
5/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob4164776637
965734357.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1727206947718_
0006
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b153
2457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1727206947718_0006
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1727206947718_0006
  The url to track the job: http://ip-172-31-27-49.us-east-2.compute.internal:20888/proxy/applic
ation_1727206947718_0006/
  Running job: job_1727206947718_0006
  Job job_1727206947718_0006 running in uber mode : false
   map 0% reduce 0%
   map 25% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1727206947718_0006 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20240924.212925.983915/output
Counters: 55
        File Input Format Counters
                Bytes Read=1567508
        File Output Format Counters
                Bytes Written=29260
        File System Counters
                FILE: Number of bytes read=27045
                FILE: Number of bytes written=3355786
```



```
"SOCIAL PROG ADMINISTRATOR III" 1
"SOLID WASTE SUPERINTENDENT"    4
"SR COMPANION STIPEND HLTH"     143
"STATE LIBRARY RESOURCE CENTER" 3
"STATE'S ATTORNEY"      1
"STATISTICAL TRAFFIC ANALYST"   1
"STOREKEEPER I" 22
"STORES SUPERVISOR II"  2
"STREET MASON"  1
"SUPT CLEANING BOARDNG & GR MNT"        1
"SUPT COMMUNICATIONS/COMPUTER O"        1
"SUPT PLANS AND INSPECTIONS"    2
"SUPT TRAFFIC SIGNAL INSTALLATI"        1
"SUPV. OF BOARDING/GROUNDS MAIN"        1
"SURVEY COMPUTATION ANALYST"    1
"SURVEY TECHNICIAN II"  3
"SURVEY TECHNICIAN III" 1
"SWIMMING POOL ATTENDENT"       26
"SYSTEMS SUPERVISOR"    2
"Senior Fire Operations Aide"   2
"Solid Waste Asst Superintenden"        2
"Systems Analyst"       3
"TOWING LOT SUPERINTENDENT"     1
"TRACTOR TRAILER DRIVER"        5
"TRAFFIC INVESTIGATOR III"      2
"TREASURY ASSISTANT"    1
"TREASURY TECHNICIAN"   2
"Transportation Enforcemt Off I"        65
"Transportation Enforcmt Off II"        20
"Transportation Enforcmt Sup II"        3
"UTILITIES INSTALLER REPAIR III"        47
"UTILITY INVESTIGATOR SUPV"     3
"UTILITY METER FIELD OPER MANAG"        1
"UTILITY METER READER I"        23
"UTILITY METER READER SUPT II"  1
"UTILITY METER READER SUPV"     5
"UTILITY POLICY ANALYST"        1
"Urban Forester"        7
"VOLUNTEER SERVICE WORKER"      1
"Volunteer Service Coordinator" 1
"WASTE WATER PLANT MANAGER"     2
"WATER PUMPING ASST MANAGER"    2
"WATER SERVICE INSPECTOR"       4
"WATER SERVICE REPRESENTATIVE"  12
"WATER TREATMENT TECHNICIAN III"        8
"WATERSHED MAINT SUPV"  3
"WWW Chief of Engineering"      1
"WWW Division Manager II"       5
"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.20240924.212925.98391
5...
Removing temp directory /tmp/Salaries.hadoop.20240924.212925.983915...
[hadoop@ip-172-31-27-49 ~]$
```

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. This is defined as follows:

High	100,000.00 and above
Medium	50,000.00 to 99,999.99
Low	0.00 to 49,999.99

The output of the program should be something like the following (in any order):

High 20
Medium 30
Low 10

Some important hints:

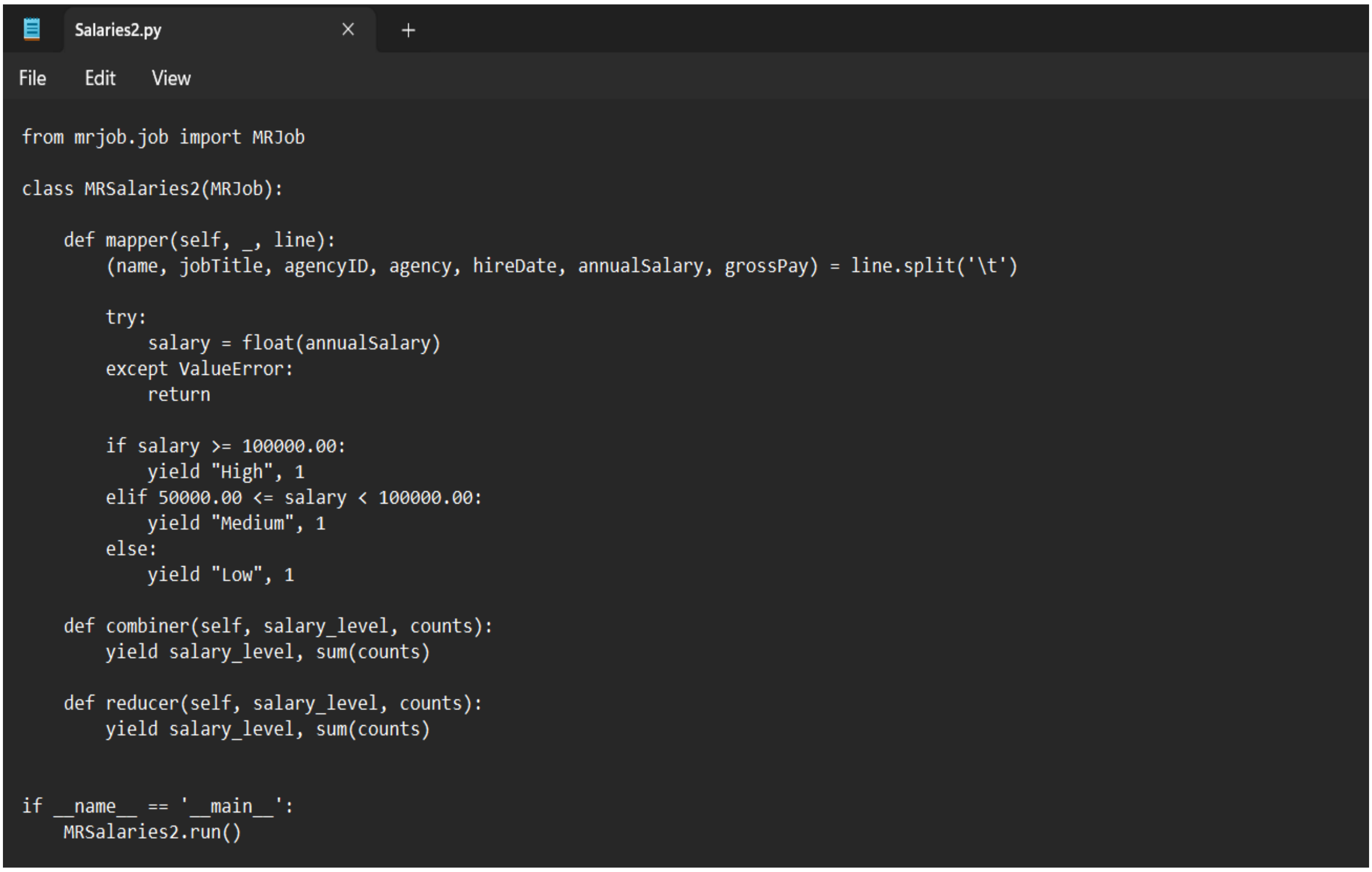
- The annual salary is a string that will need to be converted to a float.
- The mapper should output tuples with one of three keys depending on the annual salary: High, Medium and Low
- The value part of the tuple is not a salary. (What should it be?)

Now execute the program and see what happens.

14) (3 points) Submit (1) a copy of this modified program and (2) a screen shot of the results of the program's execution as the output of your assignment.

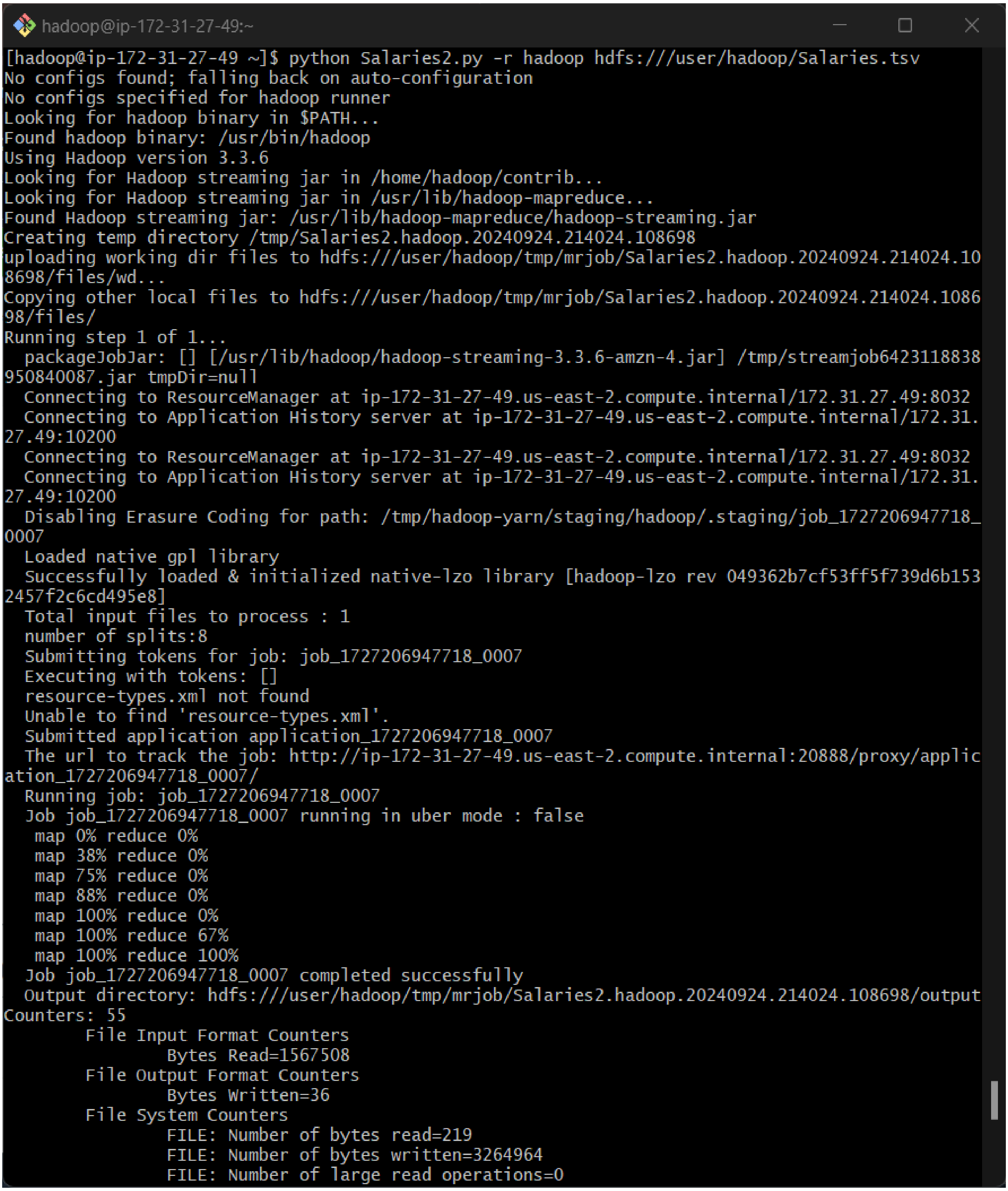
Ans:

- **Code: Salaries2.py**

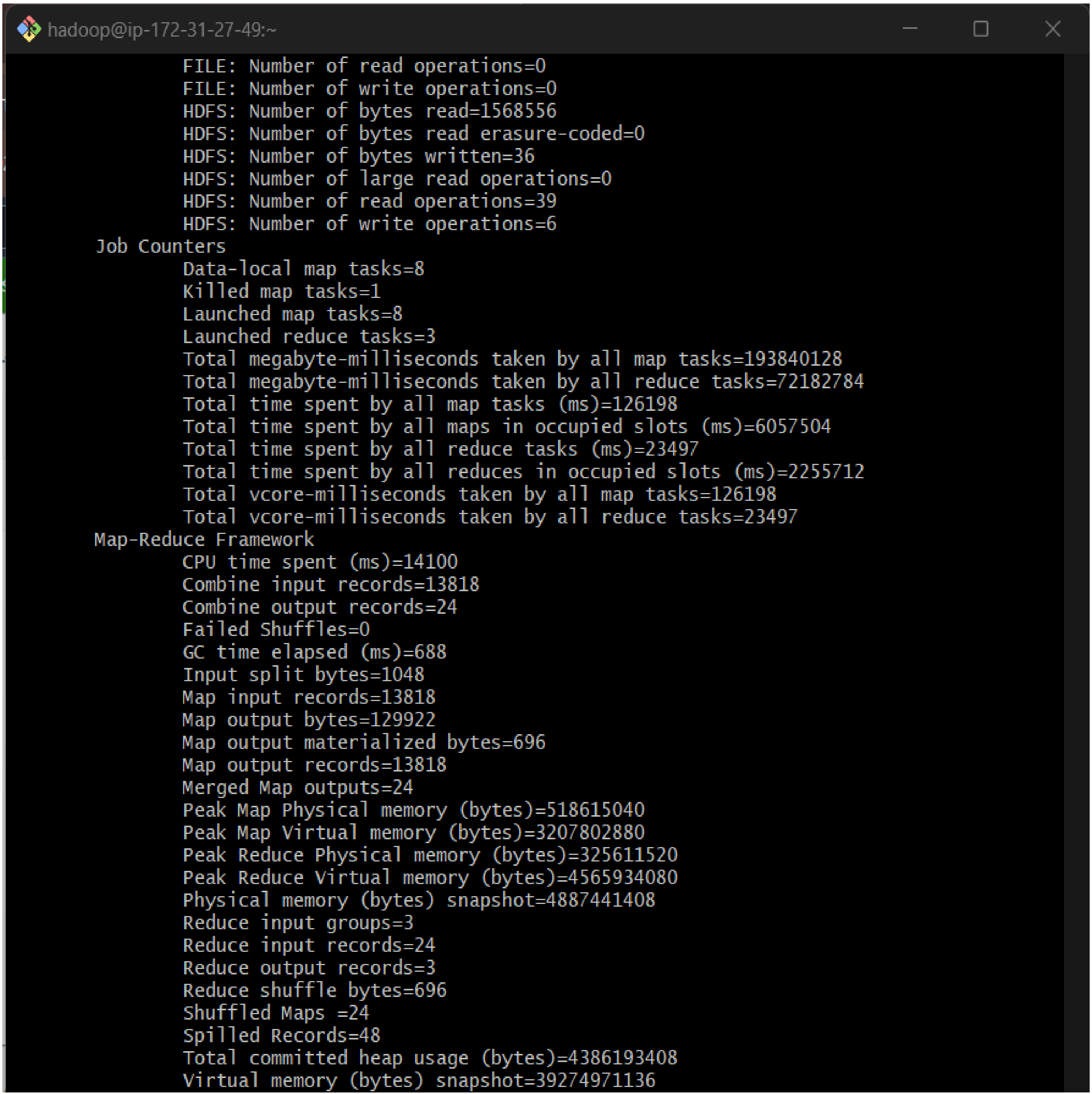


```
from mrjob.job import MRJob

class MRSalaries2(MRJob):

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

        try:
            salary = float(annualSalary)
        except ValueError:
            return

        if salary >= 100000.00:
            yield "High", 1
        elif 50000.00 <= salary < 100000.00:
            yield "Medium", 1
        else:
            yield "Low", 1

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

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


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

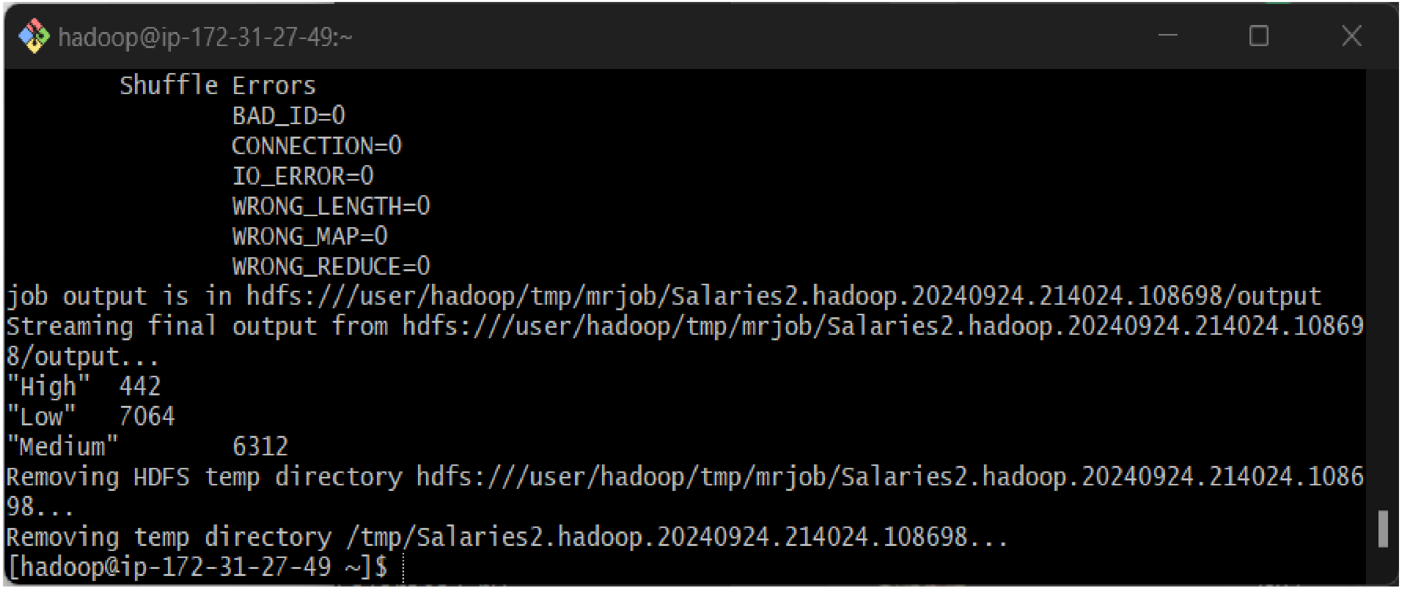
- **Output:**



```
[hadoop@ip-172-31-27-49 ~]$ 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.20240924.214024.108698
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240924.214024.10
8698/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240924.214024.1086
98/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-4.jar] /tmp/streamjob6423118838
950840087.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Connecting to ResourceManager at ip-172-31-27-49.us-east-2.compute.internal/172.31.27.49:8032
  Connecting to Application History server at ip-172-31-27-49.us-east-2.compute.internal/172.31.
27.49:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1727206947718_
0007
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b153
2457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1727206947718_0007
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1727206947718_0007
  The url to track the job: http://ip-172-31-27-49.us-east-2.compute.internal:20888/proxy/applic
ation_1727206947718_0007/
  Running job: job_1727206947718_0007
  Job job_1727206947718_0007 running in uber mode : false
   map 0% reduce 0%
   map 38% reduce 0%
   map 75% reduce 0%
   map 88% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1727206947718_0007 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240924.214024.108698/output
Counters: 55
        File Input Format Counters
                Bytes Read=1567508
        File Output Format Counters
                Bytes Written=36
        File System Counters
                FILE: Number of bytes read=219
                FILE: Number of bytes written=3264964
                FILE: Number of large read operations=0
```



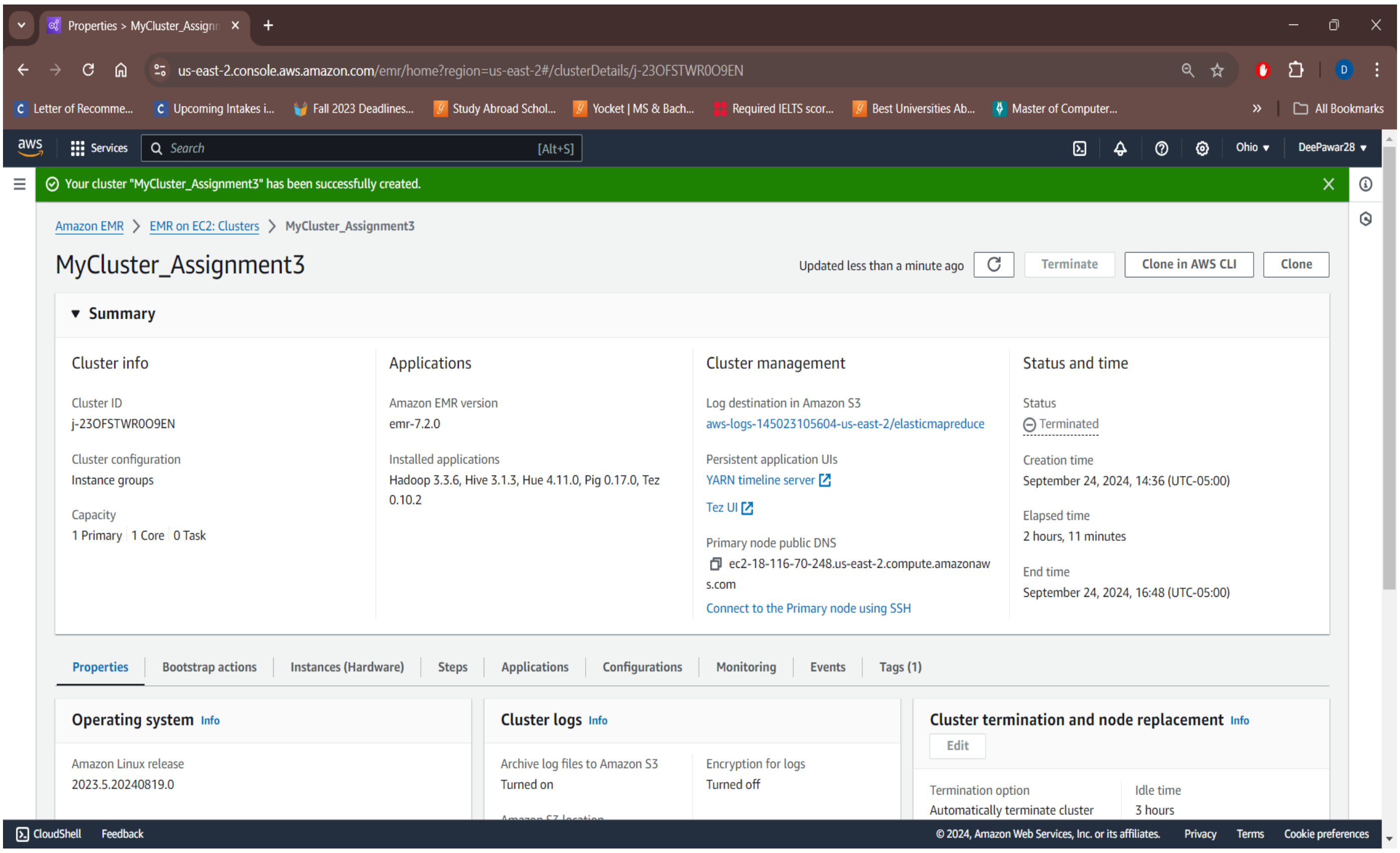
```
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1568556
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=36
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=39
                HDFS: Number of write operations=6
        Job Counters
                Data-local map tasks=8
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=193840128
                Total megabyte-milliseconds taken by all reduce tasks=72182784
                Total time spent by all map tasks (ms)=126198
                Total time spent by all maps in occupied slots (ms)=6057504
                Total time spent by all reduce tasks (ms)=23497
                Total time spent by all reduces in occupied slots (ms)=2255712
                Total vcore-milliseconds taken by all map tasks=126198
                Total vcore-milliseconds taken by all reduce tasks=23497
        Map-Reduce Framework
                CPU time spent (ms)=14100
                Combine input records=13818
                Combine output records=24
                Failed Shuffles=0
                GC time elapsed (ms)=688
                Input split bytes=1048
                Map input records=13818
                Map output bytes=129922
                Map output materialized bytes=696
                Map output records=13818
                Merged Map outputs=24
                Peak Map Physical memory (bytes)=518615040
                Peak Map Virtual memory (bytes)=3207802880
                Peak Reduce Physical memory (bytes)=325611520
                Peak Reduce Virtual memory (bytes)=4565934080
                Physical memory (bytes) snapshot=4887441408
                Reduce input groups=3
                Reduce input records=24
                Reduce output records=3
                Reduce shuffle bytes=696
                Shuffled Maps =24
                Spilled Records=48
                Total committed heap usage (bytes)=4386193408
                Virtual memory (bytes) snapshot=39274971136
```



```
        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.20240924.214024.108698/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240924.214024.10869
8/output...
"High"  442
"Low"   7064
"Medium"        6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240924.214024.1086
98...
Removing temp directory /tmp/Salaries2.hadoop.20240924.214024.108698...
[hadoop@ip-172-31-27-49 ~]$
```

15) Remember to terminate your EMR cluster and remove your S3 bucket!

Ans:



Your cluster "MyCluster_Assignment3" has been successfully created.

Amazon EMR > EMR on EC2: Clusters > MyCluster_Assignment3

MyCluster_Assignment3

Updated less than a minute ago

▼ Summary

Cluster info

Cluster ID
j-23OFSTWR0O9EN

Cluster configuration
Instance groups

Capacity
1 Primary 1 Core 0 Task

Applications

Amazon EMR version
emr-7.2.0

Installed applications
Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, Pig 0.17.0, Tez 0.10.2

Cluster management

Log destination in Amazon S3
aws-logs-145023105604-us-east-2/elasticmapreduce

Persistent application UIs
YARN timeline server
Tez UI

Primary node public DNS
ec2-18-116-70-248.us-east-2.compute.amazonaws.com

Connect to the Primary node using SSH

Status and time

Status
Terminated

Creation time
September 24, 2024, 14:36 (UTC-05:00)

Elapsed time
2 hours, 11 minutes

End time
September 24, 2024, 16:48 (UTC-05:00)

Operating system

Amazon Linux release
2023.5.20240819.0

Cluster logs

Archive log files to Amazon S3
Turned on

Encryption for logs
Turned off

Cluster termination and node replacement

Termination option
Automatically terminate cluster

Idle time
3 hours