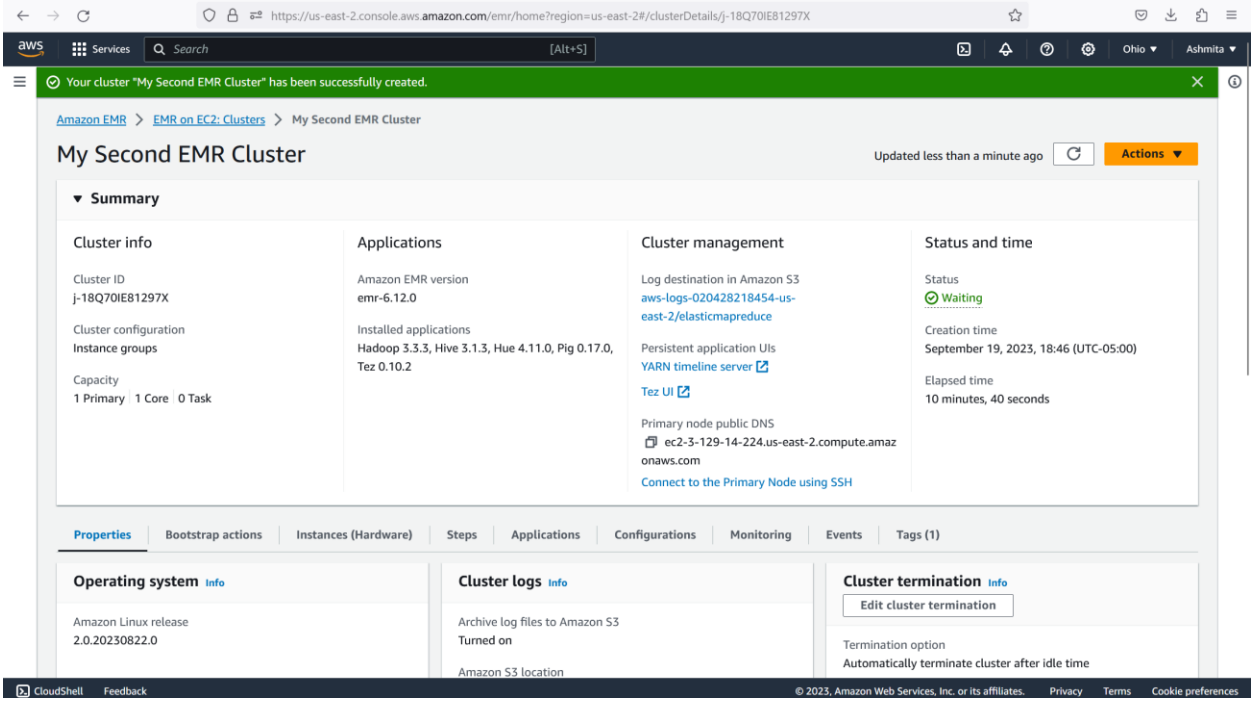
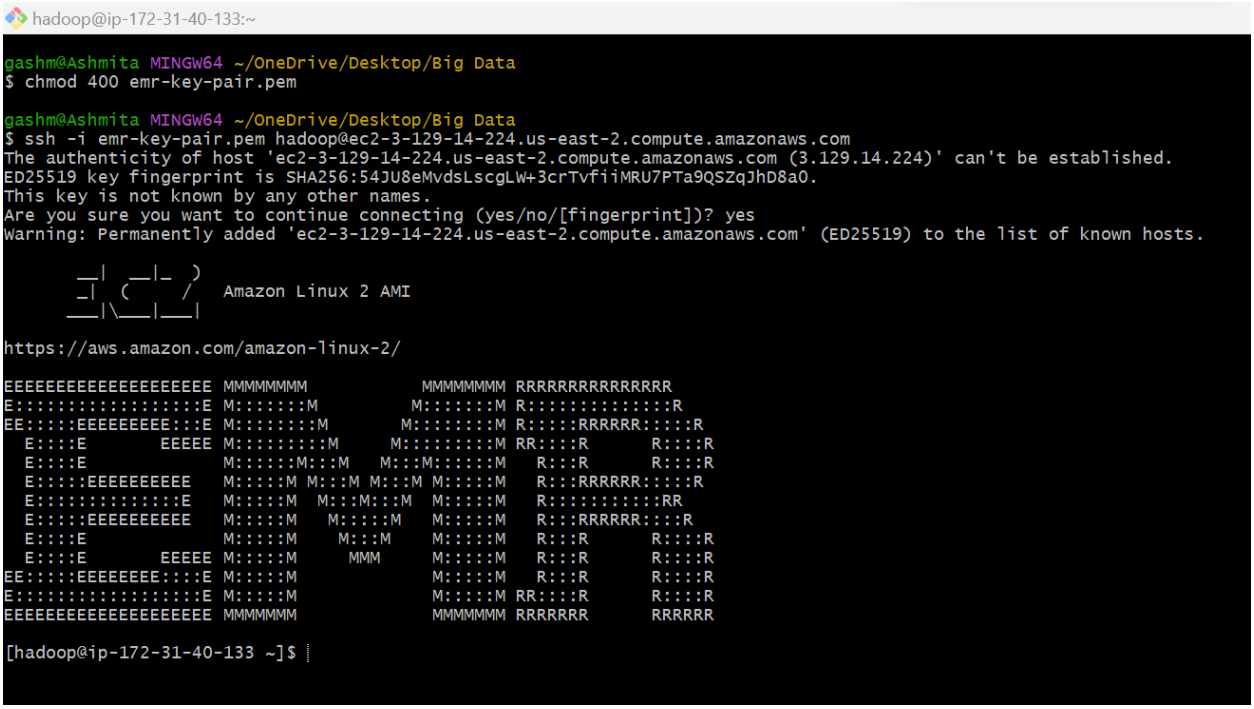
CSP 554

HOMEWORK 2

ASHMITA GUPTA (A20512498)

New EMR cluster:





Installing the mrjob library on EMR primary node:

```
hadoop@ip-172-31-40-133:~
[hadoop@ip-172-31-40-133 ~]$ sudo /usr/bin/pip3.7 install mrjob[aws]
WARNING: Running pip install with root privileges is generally not a good idea. Try `pip3.7 install --user` instead.
Collecting mrjob[aws]
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
     |████████████████████████████████| 439 kB 35.7 MB/s
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.7/site-packages (from mrjob[aws]) (5.4.1)
Collecting boto3>=1.10.0; extra == "aws"
  Downloading boto3-1.28.51-py3-none-any.whl (135 kB)
     |████████████████████████████████| 135 kB 53.5 MB/s
Collecting botocore>=1.13.26; extra == "aws"
  Downloading botocore-1.31.51-py3-none-any.whl (11.2 MB)
     |████████████████████████████████| 11.2 MB 49.7 MB/s
Collecting s3transfer<0.7.0,>=0.6.0
  Downloading s3transfer-0.6.2-py3-none-any.whl (79 kB)
     |████████████████████████████████| 79 kB 15.3 MB/s
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /usr/local/lib/python3.7/site-packages (from boto3>=1.10.0; extra == "aws"->mrjob[aws]) (1.0.1)
Collecting urllib3<1.27,>=1.25.4
  Downloading urllib3-1.26.16-py2.py3-none-any.whl (143 kB)
     |████████████████████████████████| 143 kB 45.1 MB/s
Collecting python-dateutil<3.0.0,>=2.1
  Downloading python_dateutil-2.8.2-py2.py3-none-any.whl (247 kB)
     |████████████████████████████████| 247 kB 49.9 MB/s
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (from python-dateutil<3.0.0,>=2.1->botocore>=1.13.26; extra == "aws"->mrjob[aws]) (1.13.0)
Installing collected packages: urllib3, python-dateutil, botocore, s3transfer, boto3, mrjob
  WARNING: The scripts mrjob, mrjob-3 and mrjob-3.7 are installed in '/usr/local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed boto3-1.28.51 botocore-1.31.51 mrjob-0.7.4 python-dateutil-2.8.2 s3transfer-0.6.2 urllib3-1.26.16
[hadoop@ip-172-31-40-133 ~]$
```

6) Step 1 & 2:

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: Note to prevent confusion: the default directory of your Linux account on the Hadoop primary node is “/home/hadoop.” But when we want to copy something to HDFS we will sometimes copy it to an HDFS directory beginning with “/user/hadoop.” Be aware, the Linux and HDFS file system path names have nothing to do with one another. Any similarity in naming (such as the use of the directory name “hadoop”) is just coincidental.

Now open another terminal window (but don’t use it to ssh to the primary node). This will allow you to access files on your PC or MAC to upload them to the Hadoop primary node.

From this terminal window use the secure copy (scp) program to move the WordCount.py file to the /home/hadoop directory of the primary node.

Moving WordCount.py to home/hadoop:

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair.pem WordCount.py hadoop@ec2-3-129-14-224.us-east-2.compute.amazonaws.com:/home/hadoop
WordCount.py                                                                    100%  402    13.5KB/s   00:00

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$
```

Moving w.data to home/hadoop:

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data
gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair.pem WordCount.py hadoop@ec2-3-129-14-224.us-east-2.compute.amazonaws.com:/home/hadoop
WordCount.py                                                                    100%  402    13.5KB/s   00:00

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair.pem w.data hadoop@ec2-3-129-14-224.us-east-2.compute.amazonaws.com:/home/hadoop
w.data                                                                          100%  528    22.8KB/s   00:00

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ |
```

Step 3:

Do the same for the assignment file w.data. That is move it to the directory /home/hadoop on the Hadoop primary node Linux file system.

In this case copy the file from the Linux "/home/hadoop" directory to the Hadoop file system (HDFS), say to the directory "/user/hadoop"

To check make sure the file w.data is where you think it is in HDFS by executing:

hadoop fs -ls /user/hadoop

Moving w.data to user/hadoop:

```
[hadoop@ip-172-31-40-133 ~]$ hadoop fs -copyFromLocal /home/hadoop/w.data /user/hadoop/w.data
[hadoop@ip-172-31-40-133 ~]$ hadoop fs -ls /user/hadoop
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup        528 2023-09-20 00:15 /user/hadoop/w.data
```

Step 4:

Now execute the following

python WordCount.py -r hadoop hdfs:///user/hadoop/w.data

Note there must be three slashes in "hdfs:///" as "hdfs://" indicates that the file you are reading from is in the hadoop file system and the "/user" is the first part of the path to that file. Also note that sometimes copying and pasting this command from the assignment document does not work and it needs to be entered manually.

Check that it produces some reasonable output. If all is well you should see information in the output similar to this when the program finishes correctly:

```
"well"   1
"when" 1
"will"    1
"within"          1
"writing"         2
"your"   5
```

Executing WordCount.py

```
[hadoop@ip-172-31-40-133 ~]$ 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 3.3.3
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.20230920.001703.496380
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230920.001703.496380/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230920.001703.496380/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob5580029164415459838.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695167659138_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_1695167659138_0001
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1695167659138_0001
  The url to track the job: http://ip-172-31-40-133.us-east-2.compute.internal:20888/proxy/application_1695167659138_0001/
  Running job: job_1695167659138_0001
  Job job_1695167659138_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 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1695167659138_0001 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230920.001703.496380/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=3256636
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3384
                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
```

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data
        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=203533824
                Total megabyte-milliseconds taken by all reduce tasks=79872000
                Total time spent by all map tasks (ms)=132509
                Total time spent by all maps in occupied slots (ms)=6360432
                Total time spent by all reduce tasks (ms)=26000
                Total time spent by all reduces in occupied slots (ms)=2496000
                Total vcore-milliseconds taken by all map tasks=132509
                Total vcore-milliseconds taken by all reduce tasks=26000
        Map-Reduce Framework
                CPU time spent (ms)=23090
                Combine input records=95
                Combine output records=80
                Failed Shuffles=0
                GC time elapsed (ms)=3143
                Input split bytes=1008
                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)=518365184
                Peak Map Virtual memory (bytes)=3145314304
                Peak Reduce Physical memory (bytes)=321626112
                Peak Reduce Virtual memory (bytes)=4464263168
                Physical memory (bytes) snapshot=4849762304
                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)=4318560256
                Virtual memory (bytes) snapshot=38010691584
        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.20230920.001703.496380/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230920.001703.496380/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
```

Output of WordCount.py

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data
job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230920.001703.496380/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230920.001703.496380/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
```

```
"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
```

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 small letters a-n and how many begin with anything else.

The output file should look something like

a_to_n, 12

other, 21

Now execute the program and see what happens.

6) (5 points) Submit a copy of this modified program and a screen shot of the results of the program's execution as the output of your assignment.

WordCount2.py Code

hadoop@ip-172-31-40-133:~

```
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()


~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
-- INSERT --
```

Execution of WordCount2.py:

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data
[hadoop@ip-172-31-40-133 ~]$ python WordCount2.py -r hadoop hdfs:///user/hadoop/
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 3.3.3
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.20230920.003104.202639
Uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/wordCount2.hadoop.20230920.003104.202639/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/wordCount2.hadoop.20230920.003104.202639/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob6954742103275437344.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695167659138_0002
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 2
  Cleaning up the staging area /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695167659138_0002
  Error Launching job : Not a file: hdfs://ip-172-31-40-133.us-east-2.compute.internal:8020/user/hadoop/tmp
  Streaming Command Failed!
Attempting to fetch counters from logs...
Can't fetch history log; missing job ID
No counters found
Scanning logs for probable cause of failure...
Can't fetch history log; missing job ID
Can't fetch task logs; missing application ID
Step 1 of 1 failed: Command '['/usr/bin/hadoop', 'jar', '/usr/lib/hadoop-mapreduce/hadoop-streaming.jar', '-files', 'hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230920.003104.202639/files/wd/Word
Count2.py#WordCount2.py,hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230920.003104.202639/files/wd/mrjob.zip#mrjob.zip,hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230920.003104.202639/files/
wd/setup-wrapper.sh#setup-wrapper.sh', '-input', 'hdfs:///user/hadoop/', '-output', 'hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230920.003104.202639/output', '-mapper', '/bin/sh -ex setup-wrappe
r.sh python3 WordCount2.py --step-num=0 --mapper', '-combiner', '/bin/sh -ex setup-wrapper.sh python3 WordCount2.py --step-num=0 --combiner', '-reducer', '/bin/sh -ex setup-wrapper.sh python3 WordCount
2.py --step-num=0 --reducer']' returned non-zero exit status 1280.
[hadoop@ip-172-31-40-133 ~]$ 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 3.3.3
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.20230920.003155.774466
Uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/wordCount2.hadoop.20230920.003155.774466/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/wordCount2.hadoop.20230920.003155.774466/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob8975782987326048413.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695167659138_0003
  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_1695167659138_0003
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1695167659138_0003
```

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data
 Running job: job_1695167659138_0003
 Job job_1695167659138_0003 running in uber mode : false
  map 0% reduce 0%
  map 13% reduce 0%
  map 50% reduce 0%
  map 75% reduce 0%
  map 100% reduce 0%
  map 100% reduce 67%
  map 100% reduce 100%
 Job job_1695167659138_0003 completed successfully
 Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230920.003155.774466/output
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=3255406
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3384
                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=186746880
                Total megabyte-milliseconds taken by all reduce tasks=77356032
                Total time spent by all map tasks (ms)=121580
                Total time spent by all maps in occupied slots (ms)=5835840
                Total time spent by all reduce tasks (ms)=25181
                Total time spent by all reduces in occupied slots (ms)=2417376
                Total vcore-milliseconds taken by all map tasks=121580
                Total vcore-milliseconds taken by all reduce tasks=25181
        Map-Reduce Framework
                CPU time spent (ms)=19020
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=2651
                Input split bytes=1008
                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)=584384512
                Peak Map Virtual memory (bytes)=3105132544
                Peak Reduce Physical memory (bytes)=274206720
                Peak Reduce Virtual memory (bytes)=4433301504
                Physical memory (bytes) snapshot=4883771392
                Reduce input groups=2
                Reduce input records=6
                Reduce output records=2
                Reduce shuffle bytes=464
                Shuffled Maps =24
                Spilled Records=12
```

Output of WordCount2.py:

```
job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230920.003155.774466/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230920.003155.774466/output...
"a_to_n"        46
"other" 49
```

7) 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.

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

Moving files to /home/hodoop:

```
gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair.pem Salaries.tsv hadoop@ec2-3-129-14-224.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.tsv

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair.pem Salaries.py hadoop@ec2-3-129-14-224.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.py
```

Moving Salaries.tsv to /user/hadoop:

```
[hadoop@ip-172-31-40-133 ~]$ hadoop fs -copyFromLocal /home/hadoop/Salaries.tsv /user/hadoop/Salaries.tsv
[hadoop@ip-172-31-40-133 ~]$ hadoop fs -ls /user/hadoop
Found 3 items
-rw-r--r--   1 hadoop hdfsadmingroup    1538148 2023-09-20 00:37 /user/hadoop/Salaries.tsv
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-20 00:17 /user/hadoop/tmp
-rw-r--r--   1 hadoop hdfsadmingroup        528 2023-09-20 00:15 /user/hadoop/w.data
```

Executing Salaries.py:

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data
[hadoop@ip-172-31-40-133 ~]$ 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 3.3.3
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.20230920.003825.325904
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230920.003825.325904/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230920.003825.325904/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob3261779097100612113.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695167659138_0004
  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_1695167659138_0004
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1695167659138_0004
  The url to track the job: http://ip-172-31-40-133.us-east-2.compute.internal:20888/proxy/application_1695167659138_0004/
  Running job: job_1695167659138_0004
  Job job_1695167659138_0004 running in uber mode : false
   map 0% reduce 0%
   map 38% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1695167659138_0004 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230920.003825.325904/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=3346472
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1568564
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=29260
                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=220609536
                Total megabyte-milliseconds taken by all reduce tasks=77331456
```

Output of Salaries.py:

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data
job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230920.003825.325904/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230920.003825.325904/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 CONTRACT OFFICER"        1
"CHIEF JUDGE ORPHANS' COURT"    1
"CHIEF OF FISCAL SERVICES I"    4
"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
```

Salaries2.py Code

hadoop@ip-172-31-40-133:~

```
from mrjob.job import MRJob

class MRSalaries(MRJob):

    def mapper(self, _, line):
        (name,jobTitle,agencyID,agency,hireDate,annualSalary,grossPay) = line.split('\t')
        if (float(annualSalary) >= 100000.00):
            yield 'High', 1
        if (float(annualSalary) >= 50000.00 and float(annualSalary) <= 99999.99):
            yield 'Medium', 1
        if (float(annualSalary) >= 0.00 and float(annualSalary) <= 49999.99):
            yield 'Low', 1


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

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


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


~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
:wq! Salaries2.py
```

Executing Salaries2.py:

MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data

```
[hadoop@ip-172-31-40-133 ~]$ 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 3.3.3
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.20230920.004731.461881
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230920.004731.461881/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230920.004731.461881/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob7238712571300615916.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695167659138_0005
  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_1695167659138_0005
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1695167659138_0005
  The url to track the job: http://ip-172-31-40-133.us-east-2.compute.internal:20888/proxy/application_1695167659138_0005/
  Running job: job_1695167659138_0005
  Job job_1695167659138_0005 running in uber mode : false
   map 0% reduce 0%
   map 13% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1695167659138_0005 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230920.004731.461881/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=210
                FILE: Number of bytes written=3255633
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1568564
                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=204460032
                Total megabyte-milliseconds taken by all reduce tasks=85149696
```

Output of Salaries2.py:

```
job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230920.004731.461881/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230920.004731.461881/output...
"High"  442
"Low"   7064
"Medium"        6312
```

11) Now copy the file u.data from the assignment to /user/hadoop. This is similar to the file used for some examples in Module 03b. NOTE: unlike the slide deck examples, this version of u.data has fields separated by commas and not tabs.

12) (5 points) Review the slides 52-62 in lecture notes Module 3b. Now write a program to perform the task of outputting a count of the number of movies each user (identified via their user id) reviewed.

Output might look something like the following:

186 2
192 2
112 1
etc.

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

Copying u.data:

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair.pem WordCount.py hadoop@ec2-3-129-14-224.us-east-2.compute.amazonaws.com:/home/hadoop
WordCount.py                                                                                    100%  402    13.5KB/s   00:00

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair.pem w.data hadoop@ec2-3-129-14-224.us-east-2.compute.amazonaws.com:/home/hadoop
w.data                                                                                          100%  528    22.8KB/s   00:00

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair.pem Salaries.tsv hadoop@ec2-3-129-14-224.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.tsv                                                                                    100% 1502KB   1.4MB/s   00:01

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair.pem Salaries.py hadoop@ec2-3-129-14-224.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.py                                                                                     100%  411    17.5KB/s   00:00

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair.pem u.data hadoop@ec2-3-129-14-224.us-east-2.compute.amazonaws.com:/home/hadoop
u.data                                                                                          100% 2381KB   1.4MB/s   00:01
```

```
[hadoop@ip-172-31-40-133 ~]$ hadoop fs -copyFromLocal /home/hadoop/u.data /user/hadoop/u.data
[hadoop@ip-172-31-40-133 ~]$ hadoop fs -ls /user/hadoop
Found 4 items
-rw-r--r--   1 hadoop hdfsadmingroup    1538148 2023-09-20 00:37 /user/hadoop/Salaries.tsv
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-20 00:17 /user/hadoop/tmp
-rw-r--r--   1 hadoop hdfsadmingroup    2438233 2023-09-20 00:51 /user/hadoop/u.data
-rw-r--r--   1 hadoop hdfsadmingroup        528 2023-09-20 00:15 /user/hadoop/w.data
```

Program to perform the task of outputting a count of the number of movies each user (identified via their user id) reviewed:

hadoop@ip-172-31-40-133:~

```
class MRMovies(MRJob):

    def mapper(self, _, line):
        (user_id, movie_id, rating, timeStamp) = line.split(',')
        yield user_id, 1

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

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


if __name__ == '__main__':
    MRMovies.run()
~
~
~
```

Executing MovieReviews.py

MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data

```
[hadoop@ip-172-31-40-133 ~]$ python MovieReviews.py -r hadoop hdfs:///user/hadoop/u.data
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.3.3
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/MovieReviews.hadoop.20230920.005754.096153
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/MovieReviews.hadoop.20230920.005754.096153/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/MovieReviews.hadoop.20230920.005754.096153/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob5591239622663387633.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Connecting to ResourceManager at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:8032
  Connecting to Application History server at ip-172-31-40-133.us-east-2.compute.internal/172.31.40.133:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695167659138_0006
  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_1695167659138_0006
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1695167659138_0006
  The url to track the job: http://ip-172-31-40-133.us-east-2.compute.internal:20888/proxy/application_1695167659138_0006/
  Running job: job_1695167659138_0006
  Job job_1695167659138_0006 running in uber mode : false
   map 0% reduce 0%
   map 13% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1695167659138_0006 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/MovieReviews.hadoop.20230920.005754.096153/output
Counters: 55
        File Input Format Counters
                Bytes Read=2597157
        File Output Format Counters
                Bytes Written=6204
        File System Counters
                FILE: Number of bytes read=5193
                FILE: Number of bytes written=3266743
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2598165
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=6204
                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=226417152
                Total megabyte-milliseconds taken by all reduce tasks=84562944
```

MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data

```
        Map-Reduce Framework
                CPU time spent (ms)=30200
                Combine input records=100004
                Combine output records=678
                Failed Shuffles=0
                GC time elapsed (ms)=2818
                Input split bytes=1008
                Map input records=100004
                Map output bytes=784015
                Map output materialized bytes=6405
                Map output records=100004
                Merged Map outputs=24
                Peak Map Physical memory (bytes)=558329856
                Peak Map Virtual memory (bytes)=3098558464
                Peak Reduce Physical memory (bytes)=307871744
                Peak Reduce Virtual memory (bytes)=4454789120
                Physical memory (bytes) snapshot=4905308160
                Reduce input groups=671
                Reduce input records=678
                Reduce output records=671
                Reduce shuffle bytes=6405
                Shuffled Maps =24
                Spilled Records=1356
                Total committed heap usage (bytes)=4421320704
                Virtual memory (bytes) snapshot=37992497152
        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/MovieReviews.hadoop.20230920.005754.096153/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/MovieReviews.hadoop.20230920.005754.096153/output...
"102"   678
"105"   525
"108"   31
"111"   341
"114"   25
"117"   55
"12"    61
"120"   138
"123"   33
"126"   64
"129"   26
"132"   94
"135"   22
"138"   81
"141"   31
"144"   41
"147"   38
"15"    1700
"150"   413
"153"   51
"156"   45
"159"   148
"162"   30
"165"   487
"168"   116
"171"   48
"174"   21
"177"   224
"18"    51
"180"   24
```

Output:

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data
job output is in hdfs:///user/hadoop/tmp/mrjob/MovieReviews.hadoop.20230920.005754.096153/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/MovieReviews.hadoop.20230920.005754.096153/output...
"102"   678
"105"   525
"108"   31
"111"   341
"114"   25
"117"   55
"12"    61
"120"   138
"123"   33
"126"   64
"129"   26
"132"   94
"135"   22
"138"   81
"141"   31
"144"   41
"147"   38
"15"    1700
"150"   413
"153"   51
"156"   45
"159"   148
"162"   30
"165"   487
"168"   116
"171"   48
"174"   21
"177"   224
"18"    51
"180"   24
"183"   41
"186"   42
"189"   176
"192"   55
"195"   485
"198"   75
"201"   122
"204"   31
"207"   46
"21"    162
"210"   32
"213"   910
"216"   82
"219"   138
"222"   88
"225"   28
"228"   60
"231"   32
"234"   115
"237"   44
"24"    21
"240"   230
"243"   307
"246"   26
"249"   20
"252"   38
"255"   145
"258"   32
"261"   50
"264"   33
"267"   41
"27"    23
```

MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data

```
"551"   85
"554"   64
"557"   66
"56"    522
"560"   100
"563"   158
"566"   22
"569"   85
"572"   106
"575"   547
"578"   34
"581"   49
"584"   193
"587"   504
"59"    78
"590"   89
"593"   70
"596"   487
"599"   192
"602"   129
"605"   437
"608"   296
"611"   35
"614"   99
"617"   75
"62"    53
"620"   172
"623"   103
"626"   150
"629"   34
"632"   39
"635"   22
"638"   20
"641"   140
"644"   39
"647"   150
"65"    27
"650"   29
"653"   51
"656"   128
"659"   142
"662"   58
"665"   434
"668"   20
"671"   115
"68"    123
"71"    23
"74"    49
"77"    315
"8"     116
"80"    37
"83"    161
"86"    190
"89"    66
"92"    123
"95"    299
"98"    71
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/MovieReviews.hadoop.20230920.005754.096153...
Removing temp directory /tmp/MovieReviews.hadoop.20230920.005754.096153...
[hadoop@ip-172-31-40-133 ~]$ Connection to ec2-3-129-14-224.us-east-2.compute.amazonaws.com closed by remote host.
Connection to ec2-3-129-14-224.us-east-2.compute.amazonaws.com closed.
```