CSP 554 Big Data Technologies

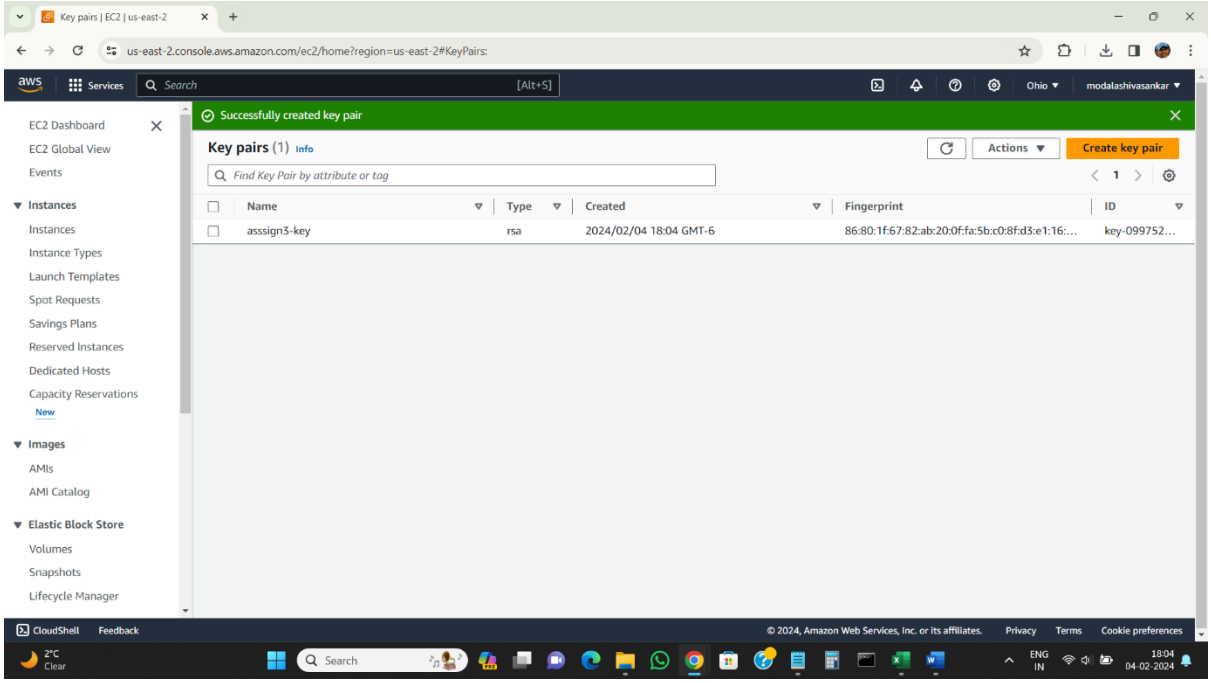
Assignment – #3

Shiva Sankar Modala(A20517528)

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.

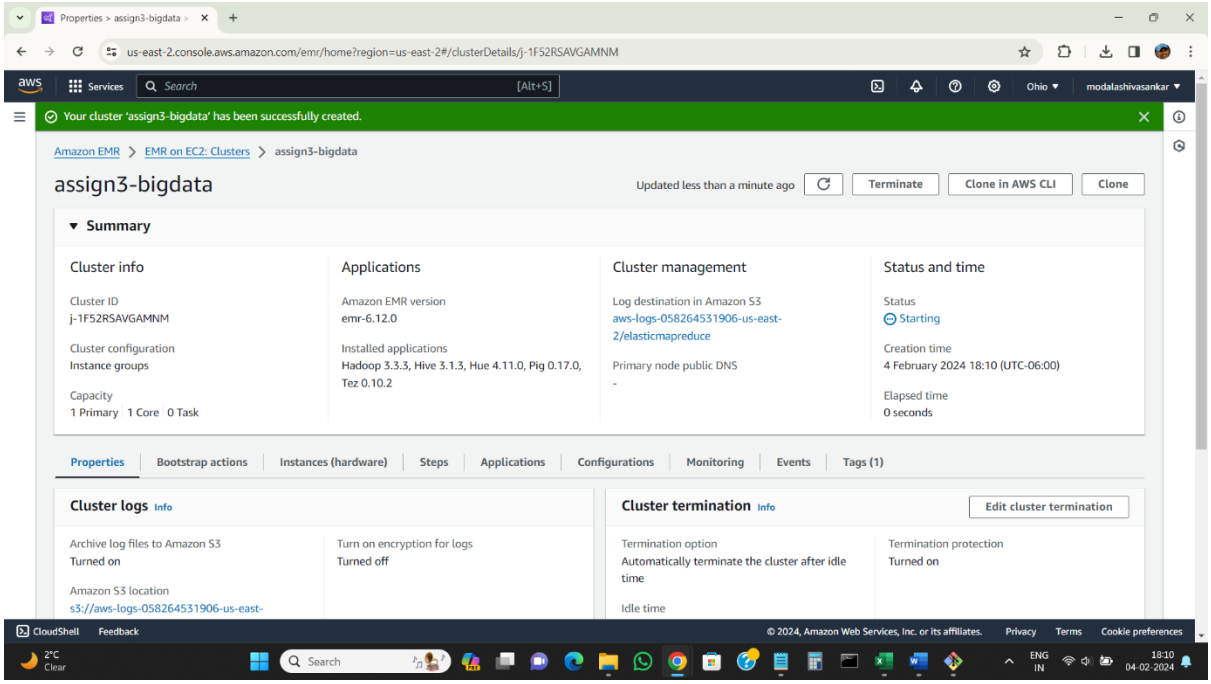
I have created a key pair

Key pair name – asssign3-key



I have created an EMR cluster

Cluster name: assign3-bigdata



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

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

```
hadoop@ip-172-31-4-106:~

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ chmod 400 C:/Users/shiva/Downloads/asssign3-key.pem

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ ssh -i C:/Users/shiva/Downloads/asssign3-key.pem hadoop@ec2-18-189-194-37.us-e
ast-2.compute.amazonaws.com
The authenticity of host 'ec2-18-189-194-37.us-east-2.compute.amazonaws.com (18.
189.194.37)' can't be established.
ED25519 key fingerprint is SHA256:batVXQ7LZPDHcjzjXv1tTpfmubyY73VkxEPaYGdWc2I.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-18-189-194-37.us-east-2.compute.amazonaws.com' (
ED25519) to the list of known hosts.
Last login: Mon Feb  5 00:16:33 2024
       #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/


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-172-31-4-106 ~]$
```

b) Enter the following (note if the first command does not work, try the second)
sudo /usr/bin/pip3.7 install mrjob[aws]

or

try:

sudo /usr/bin/pip3 install mrjob[aws]

```
[hadoop@ip-172-31-4-106 ~]$ pip3.7 install mrjob[aws]
Defaulting to user installation because normal site-packages is not writeable
Collecting mrjob[aws]
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
     |▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓| 439 kB 4.5 MB/s
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.7/site-p
ackages (from mrjob[aws]) (5.4.1)
Collecting botocore>=1.13.26; extra == "aws"
  Downloading botocore-1.33.13-py3-none-any.whl (11.8 MB)
     |▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓| 11.8 MB 54.1 MB/s
Collecting boto3>=1.10.0; extra == "aws"
  Downloading boto3-1.33.13-py3-none-any.whl (139 kB)
     |▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓| 139 kB 51.1 MB/s
Collecting urllib3<1.27,>=1.25.4; python_version < "3.10"
  Downloading urllib3-1.26.18-py2.py3-none-any.whl (143 kB)
     |▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓| 143 kB 53.3 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 50.1 MB/s
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /usr/local/lib/python3.
7/site-packages (from botocore>=1.13.26; extra == "aws"->mrjob[aws]) (1.0.1)
Collecting s3transfer<0.9.0,>=0.8.2
  Downloading s3transfer-0.8.2-py3-none-any.whl (82 kB)
     |▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓| 82 kB 327 kB/s
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-package
s (from python-dateutil<3.0.0,>=2.1->botocore>=1.13.26; extra == "aws"->mrjob[aw
s]) (1.13.0)
Installing collected packages: urllib3, python-dateutil, botocore, s3transfer, b
oto3, mrjob
  WARNING: The scripts mrjob, mrjob-3 and mrjob-3.7 are installed in '/home/hado
op/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warn
ing, use --no-warn-script-location.
Successfully installed boto3-1.33.13 botocore-1.33.13 mrjob-0.7.4 python-dateuti
l-2.8.2 s3transfer-0.8.2 urllib3-1.26.18
[hadoop@ip-172-31-4-106 ~]$ |
```

```
[hadoop@ip-172-31-4-106 ~]$ pip3.7 install mrjob[aws] --no-warn-script-location
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: mrjob[aws] in ./.local/lib/python3.7/site-package
s (0.7.4)
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.7/site-p
ackages (from mrjob[aws]) (5.4.1)
Requirement already satisfied: boto3>=1.10.0; extra == "aws" in ./.local/lib/pyt
hon3.7/site-packages (from mrjob[aws]) (1.33.13)
Requirement already satisfied: botocore>=1.13.26; extra == "aws" in ./.local/lib
/python3.7/site-packages (from mrjob[aws]) (1.33.13)
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)
Requirement already satisfied: s3transfer<0.9.0,>=0.8.2 in ./.local/lib/python3.
7/site-packages (from boto3>=1.10.0; extra == "aws"->mrjob[aws]) (0.8.2)
Requirement already satisfied: urllib3<1.27,>=1.25.4; python_version < "3.10" in
 ./.local/lib/python3.7/site-packages (from botocore>=1.13.26; extra == "aws"->m
rjob[aws]) (1.26.18)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in ./.local/lib/pytho
n3.7/site-packages (from botocore>=1.13.26; extra == "aws"->mrjob[aws]) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-package
s (from python-dateutil<3.0.0,>=2.1->botocore>=1.13.26; extra == "aws"->mrjob[aw
s]) (1.13.0)
[hadoop@ip-172-31-4-106 ~]$ |
```

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.

I have downloaded WordCount.py and w.data from blackboard

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.

```
MINGW64:/c/Users/shiva/Downloads
shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ scp -i C:/Users/shiva/Downloads/asssign3-key.pem C:/Users/shiva/Downloads/WordCount.py hadoop@ec2-18-189-194-37.us-east-2.compute.amazonaws.com:/home/hadoop
WordCount.py                                                                                  100%  402    17.5KB/s   00:00

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ |
```

```
MINGW64:/c/Users/shiva/Downloads
shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ scp -i C:/Users/shiva/Downloads/asssign3-key.pem C:/Users/shiva/Downloads/WordCount.py hadoop@ec2-18-189-194-37.us-east-2.compute.amazonaws.com:/home/hadoop
WordCount.py                                                                                  100%  402    17.5KB/s   00:00

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ scp -i C:/Users/shiva/Downloads/asssign3-key.pem C:/Users/shiva/Downloads/w.data hadoop@ec2-18-189-194-37.us-east-2.compute.amazonaws.com:/home/hadoop
w.data                                                                                        100%  528    19.8KB/s   00:00

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$
```

```
[hadoop@ip-172-31-4-106 ~]$ ls
w.data  WordCount.py
[hadoop@ip-172-31-4-106 ~]$ |
```

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

```
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -put WordCount.py /user/hadoop/
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -put w.data /user/hadoop/
[hadoop@ip-172-31-4-106 ~]$ |
```

```
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -put WordCount.py /user/hadoop/
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -put w.data /user/hadoop/
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -ls /user/hadoop/
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup        402 2024-02-05 00:29 /user/hadoop/Wo
rdCount.py
-rw-r--r--   1 hadoop hdfsadmingroup        528 2024-02-05 00:30 /user/hadoop/w.
data
[hadoop@ip-172-31-4-106 ~]$ |
```

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

Note, the above command will erase all output files in hdfs. If you want to keep the output use the following command instead:

python WordCount.py -r hadoop hdfs:///user/hadoop/w.data - -output-dir /user/hadoop/some-non-existent-directory

hadoop@ip-172-31-4-106:~

```
rdCount.py
-rw-r--r--   1 hadoop hdfsadmingroup        528 2024-02-05 00:30 /user/hadoop/w.
data
[hadoop@ip-172-31-4-106 ~]$ 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.20240205.003323.365313
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20
240205.003323.365313/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.2024
0205.003323.365313/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/str
eamjob2828419677101200443.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/17
2.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.
internal/172.31.4.106:10200
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/17
2.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.
internal/172.31.4.106:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/jo
b_1707092264542_0001
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1707092264542_0001
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1707092264542_0001
  The url to track the job: http://ip-172-31-4-106.us-east-2.compute.internal:20
888/proxy/application_1707092264542_0001/
  Running job: job_1707092264542_0001
  Job job_1707092264542_0001 running in uber mode : false
   map 0% reduce 0%
   map 25% 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_1707092264542_0001 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20240205.0033
23.365313/output
Counters: 55
        File Input Format Counters
```

hadoop@ip-172-31-4-106:~

```
23.365313/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=3258258
                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=198268416
                Total megabyte-milliseconds taken by all reduce tasks=76176384
                Total time spent by all map tasks (ms)=129081
                Total time spent by all maps in occupied slots (ms)=6195888
                Total time spent by all reduce tasks (ms)=24797
                Total time spent by all reduces in occupied slots (ms)=2380512
                Total vcore-milliseconds taken by all map tasks=129081
                Total vcore-milliseconds taken by all reduce tasks=24797
        Map-Reduce Framework
                CPU time spent (ms)=22300
                Combine input records=95
                Combine output records=80
                Failed Shuffles=0
                GC time elapsed (ms)=2921
                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)=581853184
                Peak Map Virtual memory (bytes)=3142402048
                Peak Reduce Physical memory (bytes)=313581568
                Peak Reduce Virtual memory (bytes)=4431839232
                Physical memory (bytes) snapshot=5140561920
                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)=4596432896
                Virtual memory (bytes) snapshot=38012604416
        Shuffle Errors
                BAD_ID=0
```

hadoop@ip-172-31-4-106:~

```
                Spilled Records=160
                Total committed heap usage (bytes)=4596432896
                Virtual memory (bytes) snapshot=38012604416
        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.20240205.003323.
365313/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20240
205.003323.365313/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
```

```
hadoop@ip-172-31-4-106:~
"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.2024
0205.003323.365313...
Removing temp directory /tmp/WordCount.hadoop.20240205.003323.365313...
[hadoop@ip-172-31-4-106 ~]$
```

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 screenshot of the results of the program's execution as the output of your assignment.

a copy of this modified program

WordCount2.py

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

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


class MRWordCount(MRJob):

    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            if(word[0] >= "a" and word[0] <= "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__':
    MRWordCount.run()
```

a screenshot of the results of the program's execution as the output of your assignment

hadoop@ip-172-31-4-106:~

```
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -put WordCount2.py /user/hadoop/
[hadoop@ip-172-31-4-106 ~]$ 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.20240205.004504.282659
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240205.004504.282659/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240205.004504.282659/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob3734706622101458792.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:10200
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1707092264542_0002
  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_1707092264542_0002
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1707092264542_0002
  The url to track the job: http://ip-172-31-4-106.us-east-2.compute.internal:20888/proxy/application_1707092264542_0002/
  Running job: job_1707092264542_0002
  Job job_1707092264542_0002 running in uber mode : false
   map 0% reduce 0%
   map 13% 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_1707092264542_0002 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240205.004504.282659/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=3257039
                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
```

hadoop@ip-172-31-4-106:~

```
        File System Counters
                FILE: Number of bytes read=118
                FILE: Number of bytes written=3257039
                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=184129536
                Total megabyte-milliseconds taken by all reduce tasks=66496512
                Total time spent by all map tasks (ms)=119876
                Total time spent by all maps in occupied slots (ms)=5754048
                Total time spent by all reduce tasks (ms)=21646
                Total time spent by all reduces in occupied slots (ms)=2078016
                Total vcore-milliseconds taken by all map tasks=119876
                Total vcore-milliseconds taken by all reduce tasks=21646
        Map-Reduce Framework
                CPU time spent (ms)=19560
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=2888
                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)=549797888
                Peak Map Virtual memory (bytes)=3139661824
                Peak Reduce Physical memory (bytes)=337686528
                Peak Reduce Virtual memory (bytes)=4437590016
                Physical memory (bytes) snapshot=4958785536
                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)=4487380992
                Virtual memory (bytes) snapshot=37973901312
        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.20240205.004504.282659/output
```

```
hadoop@ip-172-31-4-106:~
                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=184129536
                Total megabyte-milliseconds taken by all reduce tasks=66496512
                Total time spent by all map tasks (ms)=119876
                Total time spent by all maps in occupied slots (ms)=5754048
                Total time spent by all reduce tasks (ms)=21646
                Total time spent by all reduces in occupied slots (ms)=2078016
                Total vcore-milliseconds taken by all map tasks=119876
                Total vcore-milliseconds taken by all reduce tasks=21646
        Map-Reduce Framework
                CPU time spent (ms)=19560
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=2888
                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)=549797888
                Peak Map Virtual memory (bytes)=3139661824
                Peak Reduce Physical memory (bytes)=337686528
                Peak Reduce Virtual memory (bytes)=4437590016
                Physical memory (bytes) snapshot=4958785536
                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)=4487380992
                Virtual memory (bytes) snapshot=37973901312
        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.20240205.004504.282659/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240205.004504.282659/output...
"a_to_n"        46
"other" 49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20240205.004504.282659...
Removing temp directory /tmp/WordCount2.hadoop.20240205.004504.282659...
[hadoop@ip-172-31-4-106 ~]$
```

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.

a copy of your MRJob code

WordCount3.py

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

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


class MRWordCount(MRJob):

    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            word = word.lower()
            yield len(word), 1

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

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


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

```

a copy of the output of the execution of that code

hadoop@ip-172-31-4-106:~

```
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -put WordCount3.py /user/hadoop/
[hadoop@ip-172-31-4-106 ~]$ 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 3.3.3
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/WordCount3.hadoop.20240205.004834.041564
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240205.004834.041564/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240205.004834.041564/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob3136130272702048242.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:10200
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1707092264542_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_1707092264542_0003
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1707092264542_0003
  The url to track the job: http://ip-172-31-4-106.us-east-2.compute.internal:20888/proxy/application_1707092264542_0003/
  Running job: job_1707092264542_0003
  Job job_1707092264542_0003 running in uber mode : false
   map 0% reduce 0%
   map 25% 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_1707092264542_0003 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240205.004834.041564/output
Counters: 56
        File Input Format Counters
                Bytes Read=2376
        File Output Format Counters
                Bytes Written=49
        File System Counters
                FILE: Number of bytes read=191
                FILE: Number of bytes written=3257185
                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
```

hadoop@ip-172-31-4-106:~

```
                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
                Killed reduce tasks=1
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=194456064
                Total megabyte-milliseconds taken by all reduce tasks=74784768
                Total time spent by all map tasks (ms)=126599
                Total time spent by all maps in occupied slots (ms)=6076752
                Total time spent by all reduce tasks (ms)=24344
                Total time spent by all reduces in occupied slots (ms)=2337024
                Total vcore-milliseconds taken by all map tasks=126599
                Total vcore-milliseconds taken by all reduce tasks=24344
        Map-Reduce Framework
                CPU time spent (ms)=21620
                Combine input records=95
                Combine output records=25
                Failed Shuffles=0
                GC time elapsed (ms)=2706
                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)=549089280
                Peak Map Virtual memory (bytes)=3142336512
                Peak Reduce Physical memory (bytes)=279142400
                Peak Reduce Virtual memory (bytes)=4434116608
                Physical memory (bytes) snapshot=4918398976
                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)=4396679168
                Virtual memory (bytes) snapshot=38050115584
        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.20240205.004834.041564/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240205.004834.041564/output...
2       23
5       4
8       6
12      1
3       19
6       8
9       5
```

```
hadoop@ip-172-31-4-106:~
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=194456064
                Total megabyte-milliseconds taken by all reduce tasks=74784768
                Total time spent by all map tasks (ms)=126599
                Total time spent by all maps in occupied slots (ms)=6076752
                Total time spent by all reduce tasks (ms)=24344
                Total time spent by all reduces in occupied slots (ms)=2337024
                Total vcore-milliseconds taken by all map tasks=126599
                Total vcore-milliseconds taken by all reduce tasks=24344
        Map-Reduce Framework
                CPU time spent (ms)=21620
                Combine input records=95
                Combine output records=25
                Failed Shuffles=0
                GC time elapsed (ms)=2706
                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)=549089280
                Peak Map Virtual memory (bytes)=3142336512
                Peak Reduce Physical memory (bytes)=279142400
                Peak Reduce Virtual memory (bytes)=4434116608
                Physical memory (bytes) snapshot=4918398976
                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)=4396679168
                Virtual memory (bytes) snapshot=38050115584
        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.20240205.004834.041564/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20240205.004834.041564/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.20240205.004834.041564...
Removing temp directory /tmp/WordCount3.hadoop.20240205.004834.041564...
[hadoop@ip-172-31-4-106 ~]$
```

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.

a copy of your MRJob code

WordCount4.py

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

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


class MRWordCount(MRJob):

    def mapper(self, _, line):
        words = [word.lower() for word in WORD_RE.findall(line)]
        bigrams = zip(words, words[1:])
        for bigram in bigrams:
            yield bigram, 1

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

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


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

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

hadoop@ip-172-31-4-106:~

```
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -put WordCount4.py /user/hadoop/
[hadoop@ip-172-31-4-106 ~]$ 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 3.3.3
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/WordCount4.hadoop.20240205.005803.577840
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240205.005803.577840/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240205.005803.577840/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob5018870864067614302.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:10200
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1707092264542_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_1707092264542_0005
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1707092264542_0005
  The url to track the job: http://ip-172-31-4-106.us-east-2.compute.internal:20888/proxy/application_1707092264542_0005/
  Running job: job_1707092264542_0005
  Job job_1707092264542_0005 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 63% reduce 0%
   map 75% reduce 0%
   map 88% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1707092264542_0005 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240205.005803.577840/output
Counters: 55
        File Input Format Counters
                Bytes Read=2376
        File Output Format Counters
                Bytes Written=1800
        File System Counters
                FILE: Number of bytes read=1472
                FILE: Number of bytes written=3259909
                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=1800
                HDFS: Number of large read operations=0
```

hadoop@ip-172-31-4-106:~

```
                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=194956800
                Total megabyte-milliseconds taken by all reduce tasks=74870784
                Total time spent by all map tasks (ms)=126925
                Total time spent by all maps in occupied slots (ms)=6092400
                Total time spent by all reduce tasks (ms)=24372
                Total time spent by all reduces in occupied slots (ms)=2339712
                Total vcore-milliseconds taken by all map tasks=126925
                Total vcore-milliseconds taken by all reduce tasks=24372
        Map-Reduce Framework
                CPU time spent (ms)=21190
                Combine input records=92
                Combine output records=91
                Failed Shuffles=0
                GC time elapsed (ms)=2900
                Input split bytes=1000
                Map input records=6
                Map output bytes=1822
                Map output materialized bytes=1988
                Map output records=92
                Merged Map outputs=24
                Peak Map Physical memory (bytes)=580911104
                Peak Map Virtual memory (bytes)=3096424448
                Peak Reduce Physical memory (bytes)=334462976
                Peak Reduce Virtual memory (bytes)=4441595904
                Physical memory (bytes) snapshot=4921057280
                Reduce input groups=91
                Reduce input records=91
                Reduce output records=91
                Reduce shuffle bytes=1988
                Shuffled Maps =24
                Spilled Records=182
                Total committed heap usage (bytes)=4495245312
                Virtual memory (bytes) snapshot=37940473856
        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.20240205.005803.577840/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240205.005803.577840/output...
["a", "python"] 1
["as", "on"]    1
["as", "well"]  1
["by", "mrjob"] 1
["cluster", "as"]       1
["combine", "or"]       1
["do", "those"] 1
["either", "be"]        1
```

hadoop@ip-172-31-4-106:~

```
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240205.005803.577840/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240205.005803.577840/output...
["a", "python"] 1
["as", "on"]    1
["as", "well"]  1
["by", "mrjob"] 1
["cluster", "as"]       1
["combine", "or"]       1
["do", "those"] 1
["either", "be"]        1
["executed", "on"]      1
["explains", "how"]     1
["file", "available"]   1
["first", "job"]        1
["how", "your"] 1
["in", "a"]     1
["job", "and"]  1
["machine", "as"]       1
["more", "on"]  1
["or", "reduce"]        1
["or", "uploaded"]      1
["our", "job"]  1
["python", "script"]    1
["reference", "oriented"]       1
["run", "for"]  1
["runners", "explains"] 1
["task", "see"] 1
["those", "things"]     1
["uploaded", "to"]      1
["when", "your"]        1
["writing", "your"]     2
["your", "program"]     1
["your", "second"]      1
["a", "hadoop"] 1
["all", "dependencies"] 1
["an", "individual"]    1
["and", "writing"]      1
["are", "more"] 1
["as", "a"]     1
["as", "an"]    1
["be", "contained"]     1
["cluster", "by"]       1
["defined", "in"]       1
["following", "two"]    1
["for", "more"] 1
["how", "to"]   1
["is", "run"]   1
["job", "will"] 1
["on", "a"]     1
["on", "that"]  1
["on", "your"]  1
["program", "is"]       1
["script", "as"]        1
["second", "job"]       1
["sections", "are"]     1
["see", "how"]  1
["submitted", "runners"]        1
```

```
hadoop@ip-172-31-4-106:~
["as", "a"]     1
["as", "an"]    1
["be", "contained"]     1
["cluster", "by"]       1
["defined", "in"]       1
["following", "two"]    1
["for", "more"] 1
["how", "to"]   1
["is", "run"]   1
["job", "will"] 1
["on", "a"]     1
["on", "that"]  1
["on", "your"]  1
["program", "is"]       1
["script", "as"]        1
["second", "job"]       1
["sections", "are"]     1
["see", "how"]  1
["submitted", "runners"]        1
["task", "nodes"]       1
["the", "cluster"]      1
["the", "file"] 1
["two", "sections"]     1
["versions", "of"]      1
["well", "as"]  1
["within", "the"]       1
["a", "file"]   1
["available", "on"]     1
["be", "defined"]       1
["be", "executed"]      1
["contained", "within"] 1
["dependencies", "must"]        1
["file", "to"]  1
["hadoop", "cluster"]   1
["individual", "map"]   1
["is", "submitted"]     1
["job", "is"]   1
["map", "combine"]      1
["more", "reference"]   1
["mrjob", "when"]       1
["must", "either"]      1
["nodes", "or"] 1
["of", "writing"]       1
["on", "the"]   1
["oriented", "versions"]        1
["reduce", "task"]      1
["the", "following"]    1
["the", "task"] 1
["to", "be"]    1
["to", "do"]    1
["to", "the"]   1
["will", "be"]  1
["your", "first"]       1
["your", "job"] 1
["your", "machine"]     1
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20240205.005803.577840...
Removing temp directory /tmp/WordCount4.hadoop.20240205.005803.577840...
[hadoop@ip-172-31-4-106 ~]$
```

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.

```
MINGW64:/c/Users/shiva/Downloads

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ scp -i C:/Users/shiva/Downloads/asssign3-key.pem C:/Users/shiva/Downloads/Sala
ries.py hadoop@ec2-18-189-194-37.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.py                                   100%  411    17.0KB/s   00:00

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ scp -i C:/Users/shiva/Downloads/asssign3-key.pem C:/Users/shiva/Downloads/Sala
ries.tsv hadoop@ec2-18-189-194-37.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.tsv                                  100% 1502KB 520.6KB/s   00:02

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ |
```

```
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -put Salaries.py /user/hadoop/
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -put Salaries.tsv /user/hadoop/
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -ls /user/hadoop/
Found 8 items
-rw-r--r--   1 hadoop hdfsadmingroup        411 2024-02-05 01:04 /user/hadoop/Sa
laries.py
-rw-r--r--   1 hadoop hdfsadmingroup    1538148 2024-02-05 01:05 /user/hadoop/Sa
laries.tsv
-rw-r--r--   1 hadoop hdfsadmingroup        402 2024-02-05 00:29 /user/hadoop/Wo
rdCount.py
-rw-r--r--   1 hadoop hdfsadmingroup        504 2024-02-05 00:43 /user/hadoop/Wo
rdCount2.py
-rw-r--r--   1 hadoop hdfsadmingroup        431 2024-02-05 00:48 /user/hadoop/Wo
rdCount3.py
-rw-r--r--   1 hadoop hdfsadmingroup        497 2024-02-05 00:57 /user/hadoop/Wo
rdCount4.py
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-02-05 00:33 /user/hadoop/tm
p
-rw-r--r--   1 hadoop hdfsadmingroup        528 2024-02-05 00:30 /user/hadoop/w.
data
[hadoop@ip-172-31-4-106 ~]$
```

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

hadoop@ip-172-31-4-106:~

```
-rw-r--r--   1 hadoop hdfsadmingroup        528 2024-02-05 00:30 /user/hadoop/w.
data
[hadoop@ip-172-31-4-106 ~]$ python Salaries.py -r hadoop hdfs:///user/hadoop/Sal
aries.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.20240205.010902.545764
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.202
40205.010902.545764/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20240
205.010902.545764/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/str
eamjob7352577277991000547.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/17
2.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.
internal/172.31.4.106:10200
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/17
2.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.
internal/172.31.4.106:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/jo
b_1707092264542_0006
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1707092264542_0006
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1707092264542_0006
  The url to track the job: http://ip-172-31-4-106.us-east-2.compute.internal:20
888/proxy/application_1707092264542_0006/
  Running job: job_1707092264542_0006
  Job job_1707092264542_0006 running in uber mode : false
   map 0% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1707092264542_0006 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20240205.010902.545764/output
Counters: 55
        File Input Format Counters
                Bytes Read=1567508
        File Output Format Counters
                Bytes Written=29260
        File System Counters
```

hadoop@ip-172-31-4-106:~

```
        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=3348104
                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=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=217833984
                Total megabyte-milliseconds taken by all reduce tasks=76978176
                Total time spent by all map tasks (ms)=141819
                Total time spent by all maps in occupied slots (ms)=6807312
                Total time spent by all reduce tasks (ms)=25058
                Total time spent by all reduces in occupied slots (ms)=2405568
                Total vcore-milliseconds taken by all map tasks=141819
                Total vcore-milliseconds taken by all reduce tasks=25058
        Map-Reduce Framework
                CPU time spent (ms)=26110
                Combine input records=13818
                Combine output records=3366
                Failed Shuffles=0
                GC time elapsed (ms)=2618
                Input split bytes=1048
                Map input records=13818
                Map output bytes=356416
                Map output materialized bytes=64873
                Map output records=13818
                Merged Map outputs=24
                Peak Map Physical memory (bytes)=552423424
                Peak Map Virtual memory (bytes)=3145736192
                Peak Reduce Physical memory (bytes)=284209152
                Peak Reduce Virtual memory (bytes)=4452864000
                Physical memory (bytes) snapshot=4885106688
                Reduce input groups=1037
                Reduce input records=3366
                Reduce output records=1037
                Reduce shuffle bytes=64873
                Shuffled Maps =24
                Spilled Records=6732
                Total committed heap usage (bytes)=4417126400
                Virtual memory (bytes) snapshot=38064472064
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
```

hadoop@ip-172-31-4-106:~

```
                Reduce shuffle bytes=64873
                Shuffled Maps =24
                Spilled Records=6732
                Total committed heap usage (bytes)=4417126400
                Virtual memory (bytes) snapshot=38064472064
        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/Salaries.hadoop.20240205.010902.545764/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20240205.010902.545764/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
```

```
hadoop@ip-172-31-4-106:~
"SIGN FABRICATOR I"     2
"SIGN PAINTER II"       4
"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.20240205.010902.545764...
Removing temp directory /tmp/Salaries.hadoop.20240205.010902.545764...
[hadoop@ip-172-31-4-106 ~]$
```

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 screenshot of the results of the program's execution as the output of your assignment.

a copy of this modified program

Salaries2.py

```
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) >= 0.0 and float(annualSalary) <= 49999.99:
            yield "Low", 1
        elif float(annualSalary) >= 50000.0 and float(annualSalary) <= 99999.99:
            yield "Medium", 1
        elif float(annualSalary) >= 100000.0:
            yield "High", 1

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

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


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

a screenshot of the results of the program's execution as the output of your assignment

hadoop@ip-172-31-4-106:~

```
[hadoop@ip-172-31-4-106 ~]$ hadoop fs -put Salaries2.py /user/hadoop/
[hadoop@ip-172-31-4-106 ~]$ 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.20240205.011947.573692
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240205.011947.573692/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240205.011947.573692/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob4060574941583345699.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:10200
  Connecting to ResourceManager at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:8032
  Connecting to Application History server at ip-172-31-4-106.us-east-2.compute.internal/172.31.4.106:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1707092264542_0008
  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_1707092264542_0008
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1707092264542_0008
  The url to track the job: http://ip-172-31-4-106.us-east-2.compute.internal:20888/proxy/application_1707092264542_0008/
  Running job: job_1707092264542_0008
  Job job_1707092264542_0008 running in uber mode : false
   map 0% reduce 0%
   map 63% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1707092264542_0008 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240205.011947.573692/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=216
                FILE: Number of bytes written=3257270
                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
```

```
hadoop@ip-172-31-4-106:~
                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=205827072
                Total megabyte-milliseconds taken by all reduce tasks=70056960
                Total time spent by all map tasks (ms)=134002
                Total time spent by all maps in occupied slots (ms)=6432096
                Total time spent by all reduce tasks (ms)=22805
                Total time spent by all reduces in occupied slots (ms)=2189280
                Total vcore-milliseconds taken by all map tasks=134002
                Total vcore-milliseconds taken by all reduce tasks=22805
        Map-Reduce Framework
                CPU time spent (ms)=24700
                Combine input records=13818
                Combine output records=24
                Failed Shuffles=0
                GC time elapsed (ms)=2907
                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)=536207360
                Peak Map Virtual memory (bytes)=3104477184
                Peak Reduce Physical memory (bytes)=321630208
                Peak Reduce Virtual memory (bytes)=4459819008
                Physical memory (bytes) snapshot=4844777472
                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)=4339531776
                Virtual memory (bytes) snapshot=37964152832
        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.20240205.011947.573692/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240205.011947.573692/output...
"High"  442
"Low"   7064
"Medium"        6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20240205.011947.573692...
Removing temp directory /tmp/Salaries2.hadoop.20240205.011947.573692...
[hadoop@ip-172-31-4-106 ~]$
[hadoop@ip-172-31-4-106 ~]$
```

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

Properties > assign3-bigdata

us-east-2.console.aws.amazon.com/emr/home?region=us-east-2#/clusterDetails/j-1F52RSAVGAMNM

aws | Services | Search [Alt+S] | Ohio | modalashivasankar

Amazon EMR > EMR on EC2: Clusters > assign3-bigdata

assign3-bigdata

Updated less than a minute ago | Terminate | Clone in AWS CLI | Clone

Summary

Cluster info

Cluster ID
j-1F52RSAVGAMNM

Cluster configuration
Instance groups

Capacity
1 Primary | 1 Core | 0 Task

Applications

Amazon EMR version
emr-6.12.0

Installed applications
Hadoop 3.3.3, 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-058264531906-us-east-2/elasticmapreduce

Persistent application UIs
YARN timeline server
Tez UI

Primary node public DNS
ec2-18-189-194-37.us-east-2.compute.amazonaws.com
Connect to the Primary node using SSH
Connect to the Primary node using SSM

Status and time

Status
Terminating

Creation time
4 February 2024 18:10 (UTC-06:00)

Elapsed time
1 hour, 12 minutes

Properties | Bootstrap actions | Instances (hardware) | Steps | Applications | Configurations | Monitoring | Events | Tags (1)

Operating system Info
Amazon Linux release

Cluster logs Info
Archive log files to Amazon S3

Cluster termination Info
Edit cluster termination

CloudShell | Feedback | © 2024, Amazon Web Services, Inc. or its affiliates. | Privacy | Terms | Cookie preferences

2°C Clear | Search | ENG IN | 19:22 04-02-2024

Key pairs | EC2 | us-east-2

us-east-2.console.aws.amazon.com/ec2/home?region=us-east-2#KeyPairs:

aws | Services | Search [Alt+S] | Ohio | modalashivasankar

Spot Requests
Savings Plans
Reserved Instances
Dedicated Hosts
Capacity Reservations
New

Images
AMIs
AMI Catalog

Elastic Block Store
Volumes
Snapshots
Lifecycle Manager

Network & Security
Security Groups
Elastic IPs
Placement Groups
Key Pairs
Network Interfaces

Load Balancing

Successfully deleted 1 key pairs

Key pairs Info

Actions | Create key pair

Find Key Pair by attribute or tag

Name	Type	Created	Fingerprint	ID
No key pairs to display				

CloudShell | Feedback | © 2024, Amazon Web Services, Inc. or its affiliates. | Privacy | Terms | Cookie preferences

2°C Clear | Search | ENG IN | 19:23 04-02-2024