

# Big Data Technologies [CSP 554 - 01]

## Assignment-3 [15 points]

Anusha Sonte Parameshwar [A20518694]

### Step 1: Creating the Hadoop cluster:

The screenshot shows the AWS EMR Cluster Overview page for the 'Assignment2' cluster. The cluster is in a 'Waiting' state, ready to run steps. The cluster ID is j-39J487ZYLMJR3. It was created on 2023-09-13 at 21:16 UTC-5, with an elapsed time of 13 minutes. After the last step completes, the cluster will wait. Termination protection is off. The master public DNS is ec2-3-144-191-147.us-east-2.compute.amazonaws.com. A link to connect to the Master Node Using SSH is provided. The cluster has 1 core and 1 m4.large instance. The Amazon Linux Release is 2.0.20230822.0. The configuration details include release label emr-5.36.1, Hadoop distribution Amazon 2.10.1, applications Hive 2.3.9, Hue 4.10.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2, and log URI s3://aws-logs-991952665029-us-east-2/elasticmapreduce/. The network and hardware section shows the cluster is in the us-east-2c availability zone, with a subnet ID of subnet-0a3028e9ab5898bfd, one master instance running, and one core instance running. Task and cluster scaling are disabled. Auto-termination is also disabled.

### Step 2: Installing mrjob library on EMR primary node:

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

```
[hadoop@ip-172-31-46-234 ~]$ sudo /usr/bin/pip3 install mrjob[aws]
WARNING: Running pip install with root privileges is generally not a good idea. Try `pip3 install --user` instead.
Collecting mrjob[aws]
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
    ██████████ | 439 kB 17.0 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 boto3-1.28.47-py3-none-any.whl (135 kB)
    ██████████ | 135 kB 27.4 MB/s
Collecting botocore>=1.13.26; extra == "aws"
  Downloading botocore-1.31.47-py3-none-any.whl (11.2 MB)
    ██████████ | 11.2 MB 44.9 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]) (1.0.0)
Collecting s3transfer<0.7.0,>=0.6.0
  Downloading s3transfer-0.6.2-py3-none-any.whl (79 kB)
    ██████████ | 79 kB 14.0 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 26.6 MB/s
Collecting urllib3<1.27,>=1.25.4
  Downloading urllib3-1.26.16-py2.py3-none-any.whl (143 kB)
    ██████████ | 143 kB 28.6 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: python-dateutil, urllib3, 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 boto3-1.28.47 botocore-1.31.47 mrjob-0.7.4 python-dateutil-2.8.2 s3transfer-0.6.2 urllib3-1.26.16
[hadoop@ip-172-31-46-234 ~]$
```

### **Step 3: Copying Wordcount.py and w.data from local machine to hadoop primary node:**

Commands used:

[on local terminal]: scp -i emr-key-pair.pem Wordcount.py

[hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop](https://ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop)

[on local terminal]: scp -i emr-key-pair.pem w.data

[hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop](https://ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop)

[on main terminal]: hadoop fs -put ~/Wordcount.py /user/hadoop

[on main terminal]: hadoop fs -put ~/w.data /user/hadoop

```
anushasp@Anushas-MacBook-Air BigData % scp -i emr-key-pair.pem Wordcount.py hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop
Wordcount.py                                100%   402    21.2KB/s  00:00
anushasp@Anushas-MacBook-Air BigData % scp -i emr-key-pair.pem w.data hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop
w.data                                     100%   528    31.2KB/s  00:00
```

### **Step 4: Executing Python program for WordCount:**

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

```
job output is in hdfs:///user/hadoop/tmp/mrjob/Wordcount.hadoop.20230914.024824.447071/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Wordcount.hadoop.20230914.024824.447071/output...
"a"      3
"all"    1
"an"     1
"and"    1
"are"    1
"as"     4
"available" 1
"be"     3
"by"     1
"cluster" 2
"combine" 1
"contained" 1
"defined" 1
"dependencies" 1
"do"     1
"either" 1
"executed" 1
"explains" 1
"file"   2
"first"  1
"following" 1
"for"    1
"hadoop" 1
"how"   2
"in"    1
"individual" 1
"is"    2
"job"   4
"machine" 1
"map"   1
"more"  2
"mrjob" 1
"must"  1
"nodes" 1
"of"    1
"on"    4
```

[5 points]

### Step 5: Modified Python Program: Wordcount2.py

Commands used:

[on local terminal]: scp -i emr-key-pair.pem Wordcount2.py

[hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop](https://ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop)

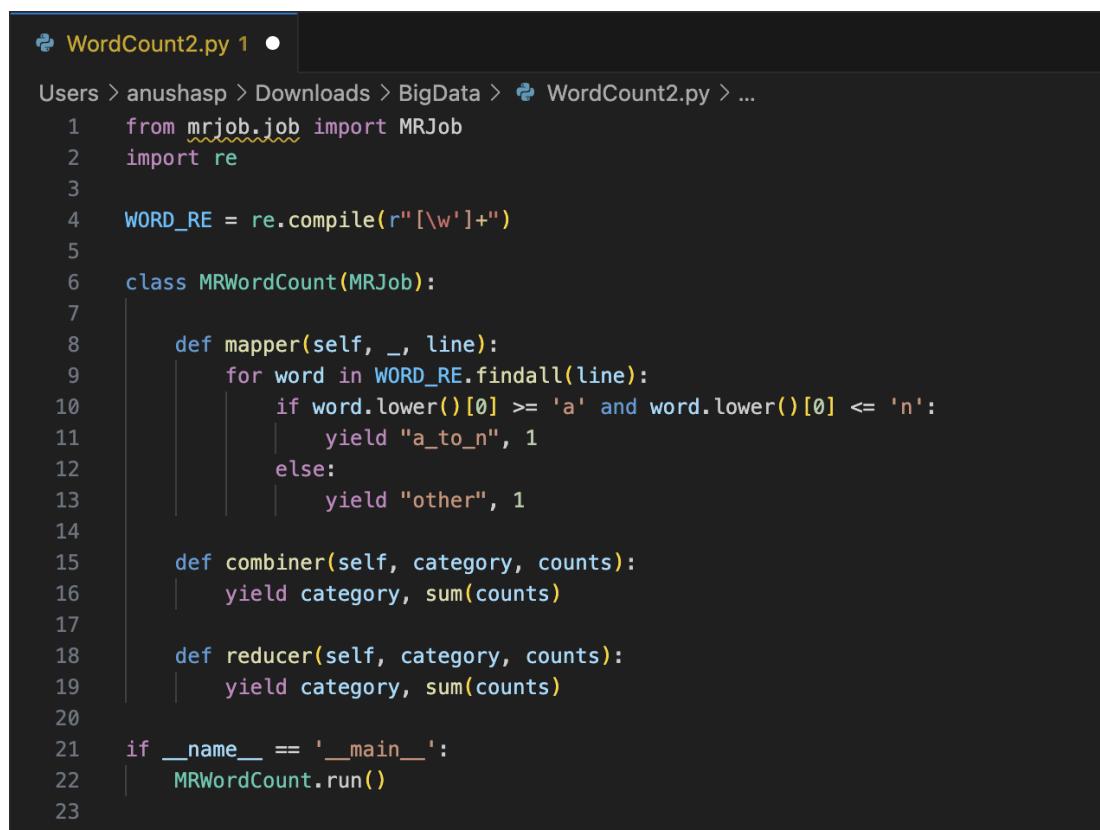
[on main terminal]: hadoop fs -put ~/Wordcount2.py /user/hadoop

```
anushasp@Anushas-MacBook-Air BigData % scp -i emr-key-pair.pem WordCount2.py hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop
WordCount2.py
100% 531    30.9KB/s   00:00
```

### Wordcount2.py:

#### Git link:

<https://github.com/anushasonte/BigData/blob/main/Assignment3/WordCount2.py>



The screenshot shows a code editor window with the file 'WordCount2.py' open. The code is written in Python and uses the mrjob library to process text files. It defines a class MRWordCount that inherits from MRJob. The mapper function processes each line, finds words using a regular expression, and yields them with a value of 1 if they start with 'a' and end with 'n', or with a value of 1 otherwise. The combiner and reducer functions both yield the category and the sum of counts. The script ends with a main block that runs the MRWordCount class.

```
WordCount2.py 1 ●
Users > anushasp > Downloads > BigData > WordCount2.py > ...
1  from mrjob.job import MRJob
2  import re
3
4  WORD_RE = re.compile(r"[\w']+")
5
6  class MRWordCount(MRJob):
7
8      def mapper(self, _, line):
9          for word in WORD_RE.findall(line):
10              if word.lower()[0] >= 'a' and word.lower()[0] <= 'n':
11                  yield "a_to_n", 1
12              else:
13                  yield "other", 1
14
15      def combiner(self, category, counts):
16          yield category, sum(counts)
17
18      def reducer(self, category, counts):
19          yield category, sum(counts)
20
21  if __name__ == '__main__':
22      MRWordCount.run()
23
```

## Output:

```
python WordCount2.py -r hadoop hdfs:///user/hadoop/w.data
```

```
[hadoop@ip-172-31-46-234 ~]$ 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 2.10.1
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.20230914.025114.965568
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230914.025114.965568/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230914.025114.965568/files/
Running step 1 of 1...
  packageJobJar: [] [/tmp/streamjob7058962125034010225.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:8032
Connecting to Application History server at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:10200
Connecting to ResourceManager at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:8032
Connecting to Application History server at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:10200
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff739d6b1532457f2c6cd495e8]
Total input files to process : 1
number of splits:4
Submitting tokens for job: job_1694658178709_0002
resource-types.xml not found
Unable to find 'resource-types.xml'.
Adding resource type - name = memory_mb, units = Mi, type = COUNTABLE
Adding resource type - name = vcores, units = , type = COUNTABLE
Submitted application application_1694658178709_0002
The url to track the job: http://ip-172-31-46-234.us-east-2.compute.internal:20888/proxy/application_1694658178709_0002/
Running job: job_1694658178709_0002
Job job_1694658178709_0002 running in uber mode : false
  map 0% reduce 0%
  map 25% reduce 0%
  map 50% reduce 0%
  map 100% reduce 0%
  map 100% reduce 100%
Job job_1694658178709_0002 completed successfully
Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230914.025114.965568/output
Counters: 50
  File Input Format Counters
    Bytes Read=1320
  File Output Format Counters
    Bytes Written=23
  File System Counters
```

```
Launched reduce tasks=1
Total megabyte-milliseconds taken by all map tasks=63438336
Total megabyte-milliseconds taken by all reduce tasks=12564480
Total time spent by all map tasks (ms)=41301
Total time spent by all maps in occupied slots (ms)=1982448
Total time spent by all reduce tasks (ms)=4090
Total time spent by all reduces in occupied slots (ms)=392640
Total vcore-milliseconds taken by all map tasks=41301
Total vcore-milliseconds taken by all reduce tasks=4090
Map-Reduce Framework
  CPU time spent (ms)=5700
  Combine input records=95
  Combine output records=6
  Failed Shuffles=0
  GC time elapsed (ms)=1039
  Input split bytes=504
  Map input records=6
  Map output bytes=999
  Map output materialized bytes=144
  Map output records=95
  Merged Map outputs=4
  Physical memory (bytes) snapshot=2044612608
  Reduce input groups=2
  Reduce input records=6
  Reduce output records=2
  Reduce shuffle bytes=144
  Shuffled Maps =4
  Spilled Records=12
  Total committed heap usage (bytes)=1668808704
  Virtual memory (bytes) snapshot=17848913920
  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.20230914.025114.965568/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230914.025114.965568/output...
"a_to_n"        49
"other" 46
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230914.025114.965568...
Removing temp directory /tmp/WordCount2.hadoop.20230914.025114.965568...
[hadoop@ip-172-31-46-234 ~]$
```

## Step 6: Executing Python program for Salaries:

Commands used:

[on local terminal]: scp -i emr-key-pair.pem Salaries.py

[hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop](https://ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop)

[on local terminal]: scp -i emr-key-pair.pem Salaries.tsv

[hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop](https://ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop)

[on main terminal]: hadoop fs -put ~/Salaries.py /user/hadoop

[on main terminal]: hadoop fs -put ~/Salaries.tsv /user/hadoop

```
anushasp@Anushas-MacBook-Air BigData % scp -i emr-key-pair.pem Salaries.py hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.py
anushasp@Anushas-MacBook-Air BigData % scp -i emr-key-pair.pem Salaries2.py hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries2.py
anushasp@Anushas-MacBook-Air BigData % scp -i emr-key-pair.pem Salaries.tsv hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.tsv
anushasp@Anushas-MacBook-Air BigData % █
```

python Salaries.py -r hadoop hdfs:///user/hadoop/Salaries.tsv

```
"WASTE WATER PLANT COORDINATOR" 2
"WASTE WATER PLANT MANAGER" 2
"WASTE WATER PLANT OPSN SUPV" 2
"WATER PUMPING ASST MANAGER" 2
"WATER SERVICE INSPECTOR" 4
"WATER SERVICE REPRESENTATIVE" 12
"WATER TREATMENT ASST MANAGER" 2
"WATER TREATMENT TECHNICIAN IT" 17
"WATER TREATMENT TECHNICIAN III" 8
"WATER TREATMENT TECHNICIAN SUP" 6
"WATERSHED MAINT SUPV" 3
"WATERSHED MANAGER" 1
"WATERSHED RANGER II" 5
"WATERSHED RANGER III" 3
"WATERSHED RANGER SUPERVISOR" 1
"WEB DEVELOPER" 1
"WELDER" 8
"WHITEPRINT MACHINE OPR" 1
"WORK STUDY STUDENT" 18
"WORKER'S COMPENSATION CONTRACT" 1
"WW Chief of Engineering" 1
"WW Division Manager I" 1
"WW Division Manager II" 5
"Waste Water Maint Mgr Instrum" 1
"Waste Water Maintenance Mgr Me" 1
"Waste Water Ops Tech II Pump" 10
"Waste Water Opsn Tech II Sanit" 81
"Waste Water Tech Supv I Pump" 6
"Waste Water Tech Supv II Pump" 1
"Waste Water Tech Supv II Sanit" 10
"Waste Water Techn Supv I Sanit" 19
"Water Systems Pumping Supv" 1
"Water Systems Treatment Manage" 1
"Water Systems Treatment Supv" 2
"YOUTH DEVELOPMENT TECH" 3
"ZONING ADMINISTRATOR" 1
"ZONING APPEALS ADVISOR BMZA" 1
"ZONING APPEALS OFFICER" 1
"ZONING ENFORCEMENT OFFICER" 1
"ZONING EXAMINER I" 2
"ZONING EXAMINER II" 1
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230914.025500.423427...
Removing temp directory /tmp/Salaries.hadoop.20230914.025500.423427...
[hadoop@ip-172-31-46-234 ~]$ █
```

[5 points]

### Step 7: Modified Python Program: Salaries2.py

Commands used:

[on local terminal]: scp -i /emr-key-pair.pem /Salaries2.py

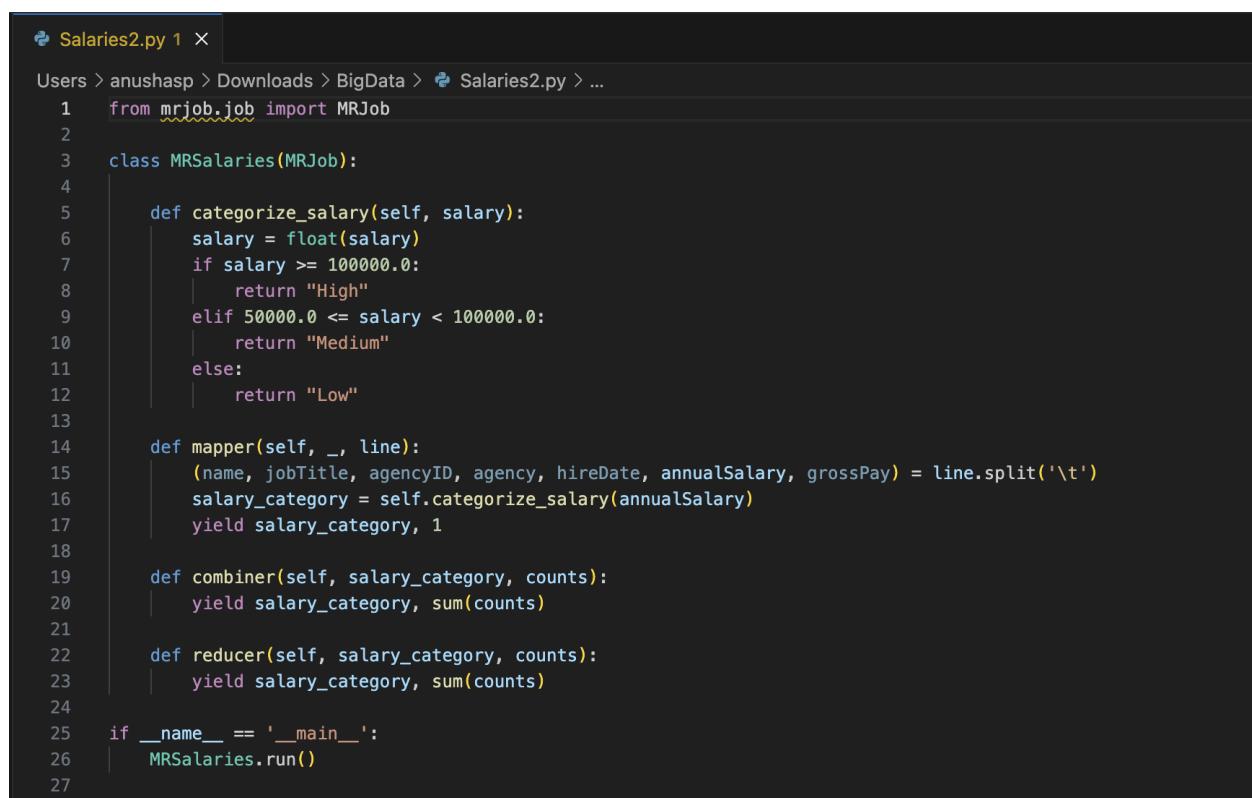
[hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop](https://ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop)

[on main terminal]: hadoop fs -put ~/Salaries2.py /user/hadoop

```
anushasp@Anushas-MacBook-Air BigData % scp -i emr-key-pair.pem Salaries.py hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.co  
m:/home/hadoop  
Salaries.py  
anushasp@Anushas-MacBook-Air BigData % scp -i emr-key-pair.pem Salaries2.py hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.c  
om:/home/hadoop  
Salaries2.py  
anushasp@Anushas-MacBook-Air BigData % scp -i emr-key-pair.pem Salaries.tsv hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.c  
om:/home/hadoop  
Salaries.tsv  
anushasp@Anushas-MacBook-Air BigData % █
```

### Salaries2.py:

Git link: <https://github.com/anushasonte/BigData/blob/main/Assignment3/Salaries2.py>



The screenshot shows a code editor window with the file "Salaries2.py" open. The code is written in Python and uses the mrjob library to process salary data. It defines a class MRSalaries with methods for categorizing salaries, mapping salary categories to counts, and reducing those counts. The code includes imports, a main function, and a run method.

```
Salaries2.py 1 ×  
Users > anushasp > Downloads > BigData > Salaries2.py > ...  
1  from mrjob.job import MRJob  
2  
3  class MRSalaries(MRJob):  
4  
5      def categorize_salary(self, salary):  
6          salary = float(salary)  
7          if salary >= 100000.0:  
8              return "High"  
9          elif 50000.0 <= salary < 100000.0:  
10             return "Medium"  
11         else:  
12             return "Low"  
13  
14      def mapper(self, _, line):  
15          (name, jobTitle, agencyID, agency, hireDate, annualSalary, grossPay) = line.split('\t')  
16          salary_category = self.categorize_salary(annualSalary)  
17          yield salary_category, 1  
18  
19      def combiner(self, salary_category, counts):  
20          yield salary_category, sum(counts)  
21  
22      def reducer(self, salary_category, counts):  
23          yield salary_category, sum(counts)  
24  
25  if __name__ == '__main__':  
26      MRSalaries.run()  
27
```

## Output:

```
python Salaries2.py -r hadoop hdfs:///user/hadoop/Salaries.tsv
```

```
[hadoop@ip-172-31-46-234 ~]$ hadoop fs -put ~/Salaries2.py /user/hadoop
[hadoop@ip-172-31-46-234 ~]$ clear

[hadoop@ip-172-31-46-234 ~]$ hadoop fs -ls /user/hadoop
Found 7 items
-rw-r--r-- 1 hadoop hdfsadmingroup 411 2023-09-14 02:54 /user/hadoop/Salaries.py
-rw-r--r-- 1 hadoop hdfsadmingroup 1538148 2023-09-14 02:54 /user/hadoop/Salaries.tsv
-rw-r--r-- 1 hadoop hdfsadmingroup 752 2023-09-14 03:00 /user/hadoop/Salaries2.py
-rw-r--r-- 1 hadoop hdfsadmingroup 531 2023-09-14 02:50 /user/hadoop/WordCount2.py
-rw-r--r-- 1 hadoop hdfsadmingroup 402 2023-09-14 02:46 /user/hadoop/Wordcount.py
drwxr-xr-x - hadoop hdfsadmingroup 0 2023-09-14 02:48 /user/hadoop/tmp
-rw-r--r-- 1 hadoop hdfsadmingroup 528 2023-09-14 02:45 /user/hadoop/w.data
[hadoop@ip-172-31-46-234 ~]$ 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 2.10.1
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.20230914.030108.534007
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230914.030108.534007/files/...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230914.030108.534007/files/
Running step 1 of 1...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar] /tmp/streamjob4670546331418000762.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:8032
Connecting to Application History server at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:10200
Connecting to ResourceManager at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:8032
Connecting to Application History server at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:10200
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
Total input files to process : 1
number of splits:4
Submitting tokens for job: job_1694658178709_0004
resource-types.xml not found
Unable to find 'resource-types.xml'.
Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
Adding resource type - name = vcores, units = , type = COUNTABLE
Submitted application application_1694658178709_0004
The url to track the job: http://ip-172-31-46-234.us-east-2.compute.internal:20888/proxy/application_1694658178709_0004/
Running job: job_1694658178709_0004
Job job_1694658178709_0004 running in uber mode : false
map 0% reduce 0%
```

```
Total megabyte-milliseconds taken by all map tasks=67236964
Total megabyte-milliseconds taken by all reduce tasks=13314048
Total time spent by all map tasks (ms)=43774
Total time spent by all maps in occupied slots (ms)=2101152
Total time spent by all reduce tasks (ms)=4334
Total time spent by all reduces in occupied slots (ms)=416064
Total vcore-milliseconds taken by all map tasks=43774
Total vcore-milliseconds taken by all reduce tasks=4334
Map-Reduce Framework
  CPU time spent (ms)=6940
  Combine input records=13818
  Combine output records=12
  Failed Shuffles=0
  GC time elapsed (ms)=981
  Input split bytes=528
  Map input records=13818
  Map output bytes=129922
  Map output materialized bytes=231
  Map output records=13818
  Merged Map outputs=4
  Physical memory (bytes) snapshot=2031996928
  Reduce input groups=3
  Reduce input records=12
  Reduce output records=3
  Reduce shuffle bytes=231
  Shuffled Maps =4
  Spilled Records=24
  Total committed heap usage (bytes)=1632632832
  Virtual memory (bytes) snapshot=17853067264
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.20230914.030108.534007/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230914.030108.534007/output...
"High" 442
"Low" 7064
"Medium" 6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230914.030108.534007...
Removing temp directory /tmp/Salaries2.hadoop.20230914.030108.534007...
[hadoop@ip-172-31-46-234 ~]$ x]
```

[5 points]

### Step 8: Executing Python program for Moviecount:

Commands used:

[on local terminal]: scp -i emr-key-pair.pem u.data

[hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop](https://ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop)

[on local terminal]: scp -i emr-key-pair.pem MovieCount.py

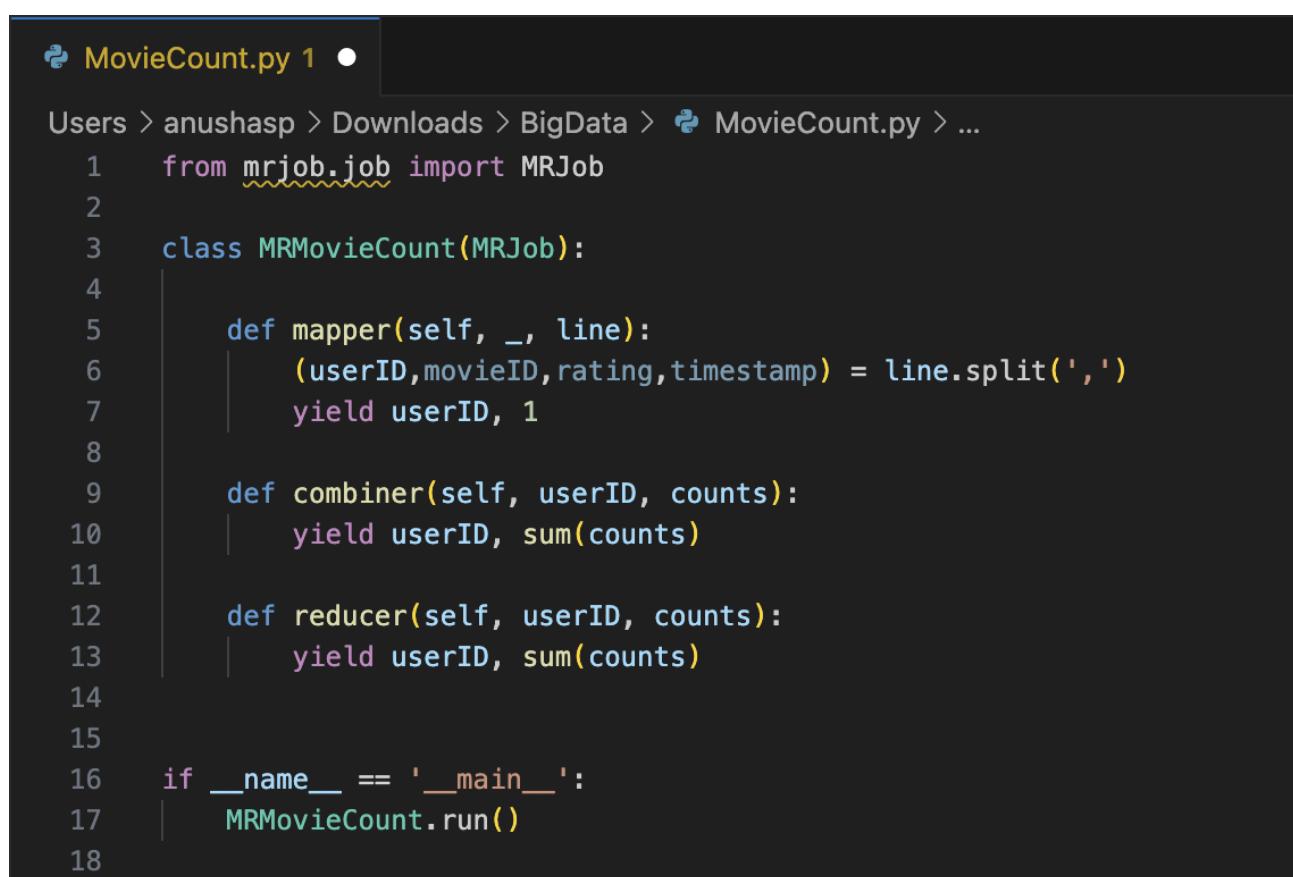
[hadoop@ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop](https://ec2-3-144-191-147.us-east-2.compute.amazonaws.com:/home/hadoop)

[on main terminal]: hadoop fs -put ~/MovieCount.py /user/hadoop

[on main terminal]: hadoop fs -put ~/u.data /user/hadoop

**Moviecount.py:**

Gitlink:<https://github.com/anushasonte/BigData/blob/main/Assignment3/MovieCount.py>



The screenshot shows a code editor window with the file 'MovieCount.py' open. The code is a Python script using the mrjob library to process movie data. It defines a class 'MRMovieCount' that inherits from 'MRJob'. The 'mapper' function splits each line into userID, movieID, rating, and timestamp, and yields the userID with a count of 1. The 'combiner' and 'reducer' functions both yield the userID with the sum of counts. The script ends with a check if the file is run directly, in which case it calls the 'run()' method of the 'MRMovieCount' class.

```
MovieCount.py 1 ●
Users > anushasp > Downloads > BigData > MovieCount.py > ...
1  from mrjob.job import MRJob
2
3  class MRMovieCount(MRJob):
4
5      def mapper(self, _, line):
6          userID,movieID,rating,timestamp = line.split(',')
7          yield userID, 1
8
9      def combiner(self, userID, counts):
10         yield userID, sum(counts)
11
12     def reducer(self, userID, counts):
13         yield userID, sum(counts)
14
15
16     if __name__ == '__main__':
17         MRMovieCount.run()
18
```

## Output:

```
python Moviecount.py -r hadoop hdfs:///user/hadoop/u.data
```

```
[hadoop@ip-172-31-46-234 ~]$ hadoop fs -put ~/MovieCount.py /user/hadoop
[hadoop@ip-172-31-46-234 ~]$ hadoop fs -put ~/u.data /user/hadoop
[hadoop@ip-172-31-46-234 ~]$ hadoop fs -ls /user/hadoop
Found 9 items
-rw-r--r-- 1 hadoop hdfsadmingroup 372 2023-09-14 03:05 /user/hadoop/MovieCount.py
-rw-r--r-- 1 hadoop hdfsadmingroup 411 2023-09-14 02:54 /user/hadoop/Salaries.py
-rw-r--r-- 1 hadoop hdfsadmingroup 1538148 2023-09-14 02:54 /user/hadoop/Salaries.tsv
-rw-r--r-- 1 hadoop hdfsadmingroup 752 2023-09-14 03:00 /user/hadoop/Salaries2.py
-rw-r--r-- 1 hadoop hdfsadmingroup 531 2023-09-14 02:50 /user/hadoop/WordCount2.py
-rw-r--r-- 1 hadoop hdfsadmingroup 402 2023-09-14 02:46 /user/hadoop/Wordcount.py
drwxr-xr-x - hadoop hdfsadmingroup 0 2023-09-14 02:48 /user/hadoop/tmp
-rw-r--r-- 1 hadoop hdfsadmingroup 2438233 2023-09-14 03:05 /user/hadoop/u.data
-rw-r--r-- 1 hadoop hdfsadmingroup 528 2023-09-14 02:45 /user/hadoop/w.data
[hadoop@ip-172-31-46-234 ~]$ python MovieCount.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 2.10.1
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/MovieCount.hadoop.20230914.030614.283848
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/MovieCount.hadoop.20230914.030614.283848/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/MovieCount.hadoop.20230914.030614.283848/files/
Running step 1 of 1...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar! /tmp/streamjob684268210676853853.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:8032
Connecting to Application History server at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:10200
Connecting to ResourceManager at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:8032
Connecting to Application History server at ip-172-31-46-234.us-east-2.compute.internal/172.31.46.234:10200
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf5f739d6b1532457f2c6cd495e8]
Total input files to process : 1
number of splits:4
Submitting tokens for job: job_1694658178709_0005
resource-types.xml not found
Unable to find 'resource-types.xml'.
Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
Adding resource type - name = vcores, units = , type = COUNTABLE
Submitted application application_1694658178709_0005
The url to track the job: http://ip-172-31-46-234.us-east-2.compute.internal:20888/proxy/application_1694658178709_0005/
Running job: job_1694658178709_0005
Job job_1694658178709_0005 running in uber mode : false
```

```
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/MovieCount.hadoop.20230914.030614.283848/output...
```

```
"1" 20
"10" 46
"100" 25
"101" 55
"102" 678
"103" 94
"104" 76
"105" 525
"106" 45
"107" 32
"108" 31
"109" 23
"11" 38
"110" 120
"111" 341
"112" 21
"113" 27
"114" 25
"115" 41
"116" 25
"117" 55
"118" 189
"119" 641
"12" 61
"120" 138
"121" 80
"122" 40
"123" 33
"124" 85
"125" 210
"126" 64
"127" 21
"128" 323
"129" 26
"13" 53
"130" 375
"131" 44
"132" 94
"133" 178
"134" 311
"135" 22
"136" 58
"137" 80
```

```

"667" 68
"668" 20
"669" 37
"67" 103
"670" 31
"671" 115
"68" 123
"69" 81
"7" 88
"70" 83
"71" 23
"72" 191
"73" 1610
"74" 49
"75" 145
"76" 20
"77" 315
"78" 263
"79" 55
"8" 116
"80" 37
"81" 168
"82" 39
"83" 161
"84" 116
"85" 107
"86" 198
"87" 31
"88" 255
"89" 66
"9" 45
"90" 50
"91" 150
"92" 123
"93" 159
"94" 196
"95" 299
"96" 76
"97" 128
"98" 71
"99" 188
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/MovieCount.hadoop.20230914.030614.283848...
Removing temp directory /tmp/MovieCount.hadoop.20230914.030614.283848...
[hadoop@ip-172-31-46-234 ~]$ █

```

## Step 9: Removing all the files from /user/hadoop

```

|[hadoop@ip-172-31-46-234 ~]$ hadoop fs -ls /user/hadoop
Found 9 items
-rw-r--r-- 1 hadoop hdfsadmingroup 372 2023-09-14 03:05 /user/hadoop/MovieCount.py
-rw-r--r-- 1 hadoop hdfsadmingroup 411 2023-09-14 02:54 /user/hadoop/Salaries.py
-rw-r--r-- 1 hadoop hdfsadmingroup 1538148 2023-09-14 02:54 /user/hadoop/Salaries.tsv
-rw-r--r-- 1 hadoop hdfsadmingroup 752 2023-09-14 03:00 /user/hadoop/Salaries2.py
-rw-r--r-- 1 hadoop hdfsadmingroup 531 2023-09-14 02:50 /user/hadoop/WordCount2.py
-rw-r--r-- 1 hadoop hdfsadmingroup 402 2023-09-14 02:46 /user/hadoop/Wordcount.py
drwxr-xr-x - hadoop hdfsadmingroup 0 2023-09-14 02:48 /user/hadoop/tmp
-rw-r--r-- 1 hadoop hdfsadmingroup 2438233 2023-09-14 03:05 /user/hadoop/u.data
-rw-r--r-- 1 hadoop hdfsadmingroup 528 2023-09-14 02:45 /user/hadoop/w.data
[hadoop@ip-172-31-46-234 ~]$
|[hadoop@ip-172-31-46-234 ~]$ hadoop fs -rm /user/hadoop/MovieCount.py
Deleted /user/hadoop/MovieCount.py
|[hadoop@ip-172-31-46-234 ~]$ hadoop fs -rm /user/hadoop/Salaries.py
Deleted /user/hadoop/Salaries.py
|[hadoop@ip-172-31-46-234 ~]$ hadoop fs -rm /user/hadoop/Salaries2.py
Deleted /user/hadoop/Salaries2.py
|[hadoop@ip-172-31-46-234 ~]$ hadoop fs -rm /user/hadoop/Salaries.tsv
Deleted /user/hadoop/Salaries.tsv
|[hadoop@ip-172-31-46-234 ~]$ hadoop fs -rm /user/hadoop/u.data
Deleted /user/hadoop/u.data
|[hadoop@ip-172-31-46-234 ~]$ hadoop fs -rm /user/hadoop/w.data
Deleted /user/hadoop/w.data
|[hadoop@ip-172-31-46-234 ~]$ hadoop fs -rm /user/hadoop/Wordcount.py
Deleted /user/hadoop/Wordcount.py
|[hadoop@ip-172-31-46-234 ~]$ hadoop fs -rm /user/hadoop/WordCount2.py
Deleted /user/hadoop/WordCount2.py

```

## Step 10: Terminating the cluster:

The screenshot shows the AWS Elastic MapReduce (EMR) console interface. The top navigation bar includes the AWS logo, services menu, search bar, and user information (Ohio, Anusha Sonte Parameshwar). A prominent message box at the top left informs users about the transition to the new EMR console.

**Amazon EMR** sidebar:

- EMR Studio
- EMR Serverless New
- EMR on EC2** (selected):
  - Clusters
  - Notebooks
  - Git repositories
  - Security configurations
  - Block public access
  - VPC subnets
  - Events
- EMR on EKS
- Virtual clusters

**Help** and **What's new** links are also present.

**Cluster: Assignment2 cluster** details:

- Status:** Terminating (Terminated by user request)
- Summary** tab selected.
- Configuration details:**
  - ID: j-39J487ZYLMJR3
  - Release label: emr-5.36.1
  - Creation date: 2023-09-13 21:16 (UTC-5)
  - Hadoop distribution: Amazon 2.10.1
  - Elapsed time: 58 minutes
  - Applications: Hive 2.3.9, Hue 4.10.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
  - Log URI: s3://aws-logs-991952665029-us-east-2/elasticmapreduce/
  - EMRFS consistent view: Disabled
  - Custom AMI ID: --
  - Amazon Linux Release: 2.0.20230822.0 [Learn more](#)
- Network and hardware:**
  - Availability zone: us-east-2c
  - Subnet ID: [subnet-0a3028e9ab5898bdf](#)
  - Master: Terminating 1 m4.large
  - Core: Terminating 1 m4.large
  - Task: --
  - Cluster scaling: Not enabled
  - Auto-termination: Not enabled
- Application user interfaces:**
  - Persistent user interfaces: [YARN timeline server](#), [Tez UI](#)
  - On-cluster user interfaces: --
- Security and access:**

Bottom navigation bar includes CloudShell, Feedback, Language, © 2023, Amazon Web Services, Inc. or its affiliates., Privacy, Terms, and Cookie preferences.