Girish Rajani
A20503736
CSP-554 Big Data Technologies
Homework 3

Figure 1 - Showing successful connection to Master Node using SSH and installing mrjob

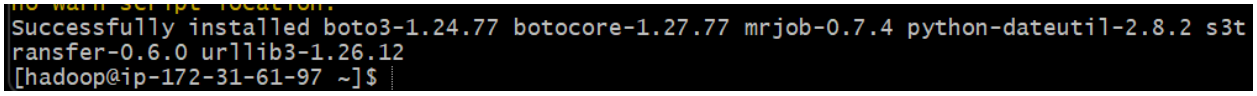

Figure 2 - Showing successful installation of mrjob

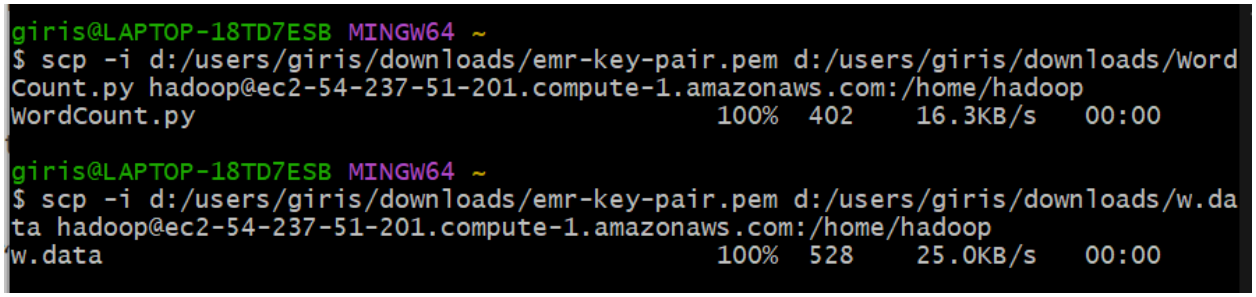

Figure 3 - Moved both WordCount.py and w.data files to the /home/hadoop directory using scp

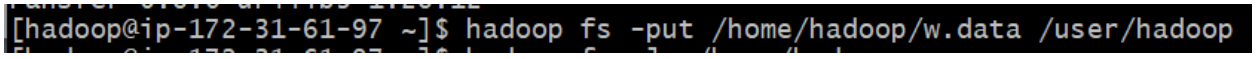

Figure 4 - Using -put command to copy w.data file from the /home/hadoop directory to HDFS /user/hadoop directory

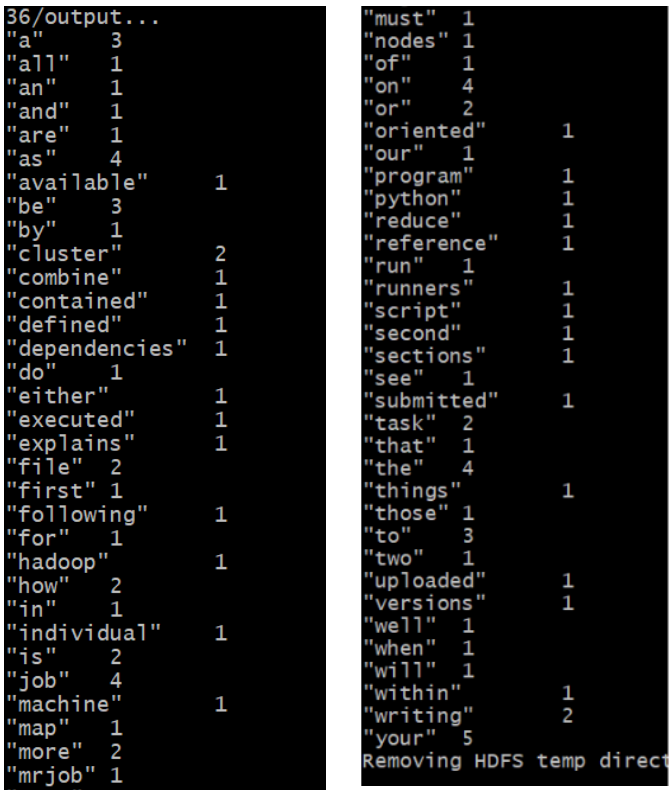

Figure 5 - Showing output after reading the WordCount.py file using the following command:
python WordCount.py -r hadoop hdfs:///user/hadoop/w.data

7) (5 points) Submit a copy of this modified program and a screen shot of the results of the program's execution as the output of your assignment.

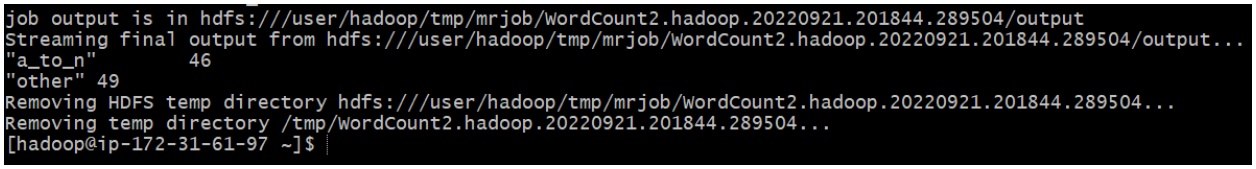

Figure 6 - Showing screenshot of the results after reading the WordCount2.py file using the following command:
python WordCount2.py -r hadoop hdfs:///user/hadoop/w.data

The program counted 46 words beginning with small letters a to n and 49 words beginning with anything else.

```
[hadoop@ip-172-31-61-97 ~]$ python Salaries.py -r hadoop hdfs:///user/hadoop/Salaries.tsv
```

Figure 7 - Uploaded Salaries.py and Salaries.tvs to HDFS using scp and then read the python file using the following command:
python Salaries.py -r hadoop hdfs:///user/hadoop/Salaries.tsv

```
"Waste Water Opns 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.20220921.211610.724953...
Removing temp directory /tmp/Salaries.hadoop.20220921.211610.724953...
[hadoop@ip-172-31-61-97 ~]$
```

Figure 8 - Illustrating the output, which shows the number of workers who share the same job

11) (5 points) Submit a copy of this modified program and a screen shot of the results of the program's execution as the output of your assignment.

```
job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220921.222046.731251/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220921.222046.731251/output...
"High"  442
"Low"   7064
"Medium"        6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220921.222046.731251...
Removing temp directory /tmp/Salaries2.hadoop.20220921.222046.731251...
[hadoop@ip-172-31-61-97 ~]$ client_loop: send disconnect: Connection reset by peer
```

Figure 9 - Showing screenshot of the results after reading the Salaries2.py file using the following command:
python Salaries2.py -r hadoop hdfs:///user/hadoop/Salaries.tsv

The program counted 442 workers who have 'High' annual salaries, 7064 workers who have 'Low' annual salaries, and 6312 employees who have 'Medium' annual salaries.

Note: From this question onwards, the cluster had been terminated and then cloned at a later date. That is why the Hadoop IP address will be different in the new screenshots below

```
[hadoop@ip-172-31-63-51 ~]$ hadoop fs -put /home/hadoop/u.data /user/hadoop
```

Figure 10 - Using -put command to copy u.data file from the /home/hadoop directory to HDFS /user/hadoop directory

13) (5 points) Write a program to perform the task of outputting a count of the number of movies each user (identified via their user id) reviewed.

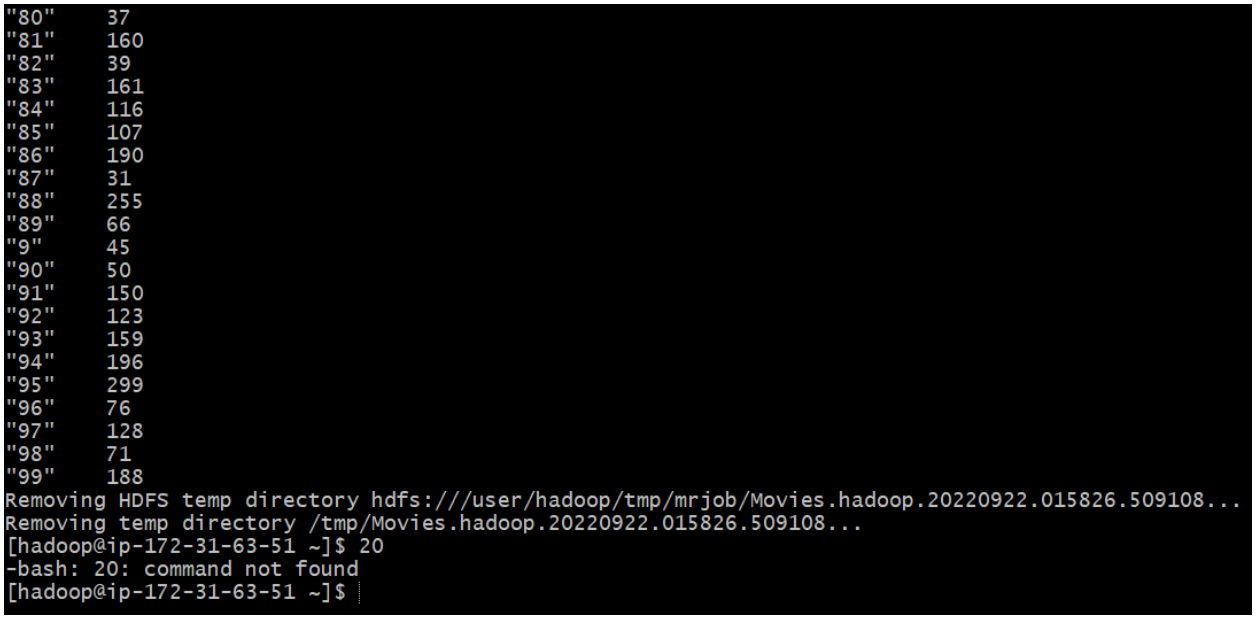

Figure 11 - Showing screenshot of the results after reading the Movies.py file using the following command:
python Movies.py -r hadoop hdfs:///user/hadoop/u.data

The program shows the number of movies each user (identified via their user id) reviewed. So for example, from figure 11 above, we can see that user 9 reviewed 45 movies.

In total, this movie dataset contains over 10,000 review samples, and to be able to count the number of movies each user reviewed (identified via their user id) within a few lines of code is astounding.