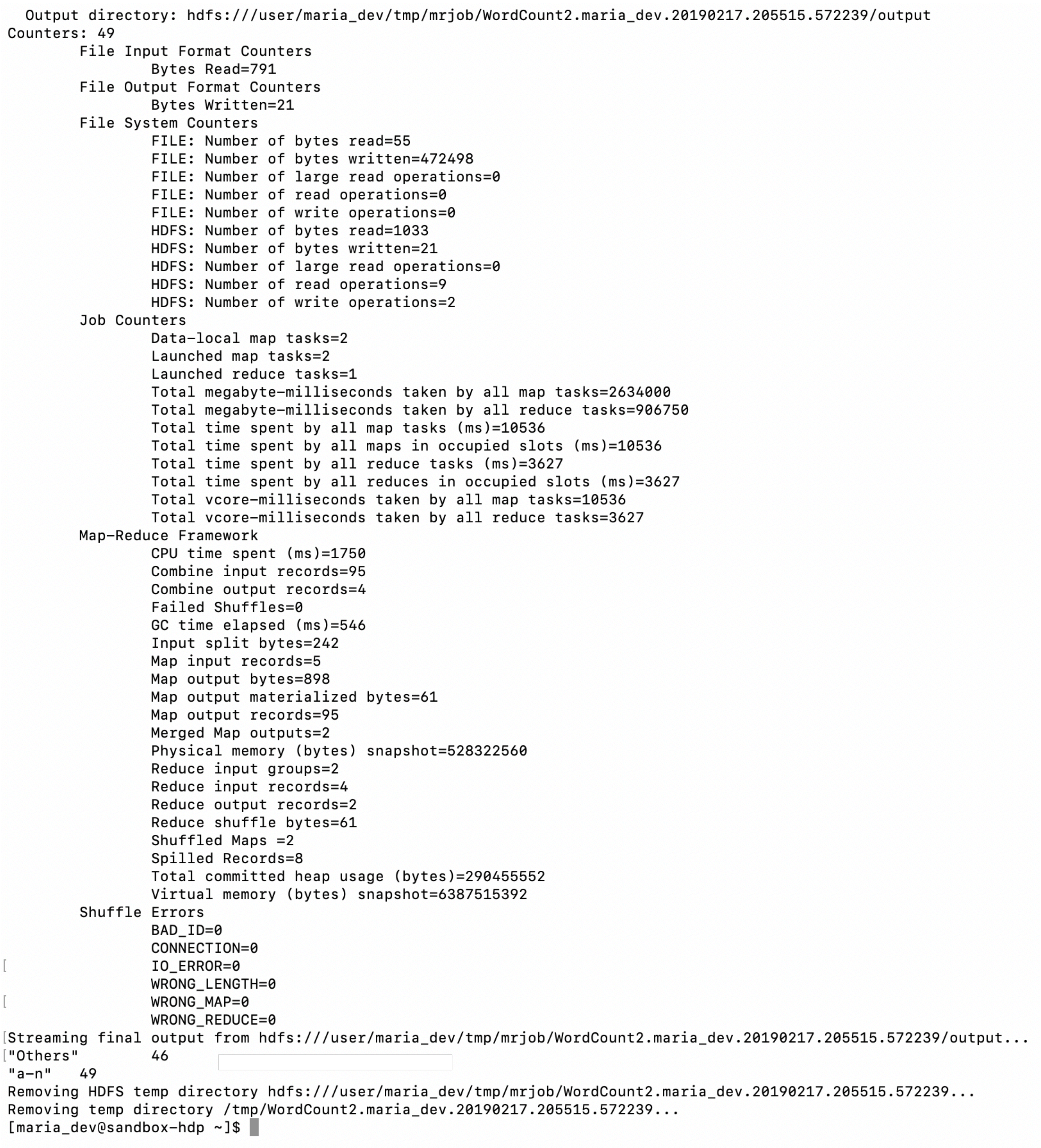
Assignment #3
CSP 554 – Big Data Technologies
Adarsh Mathad Vijayakumar – A20424847

Question 6) Word count program.

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.

Screenshot of the output of the modified program WordCount2.py

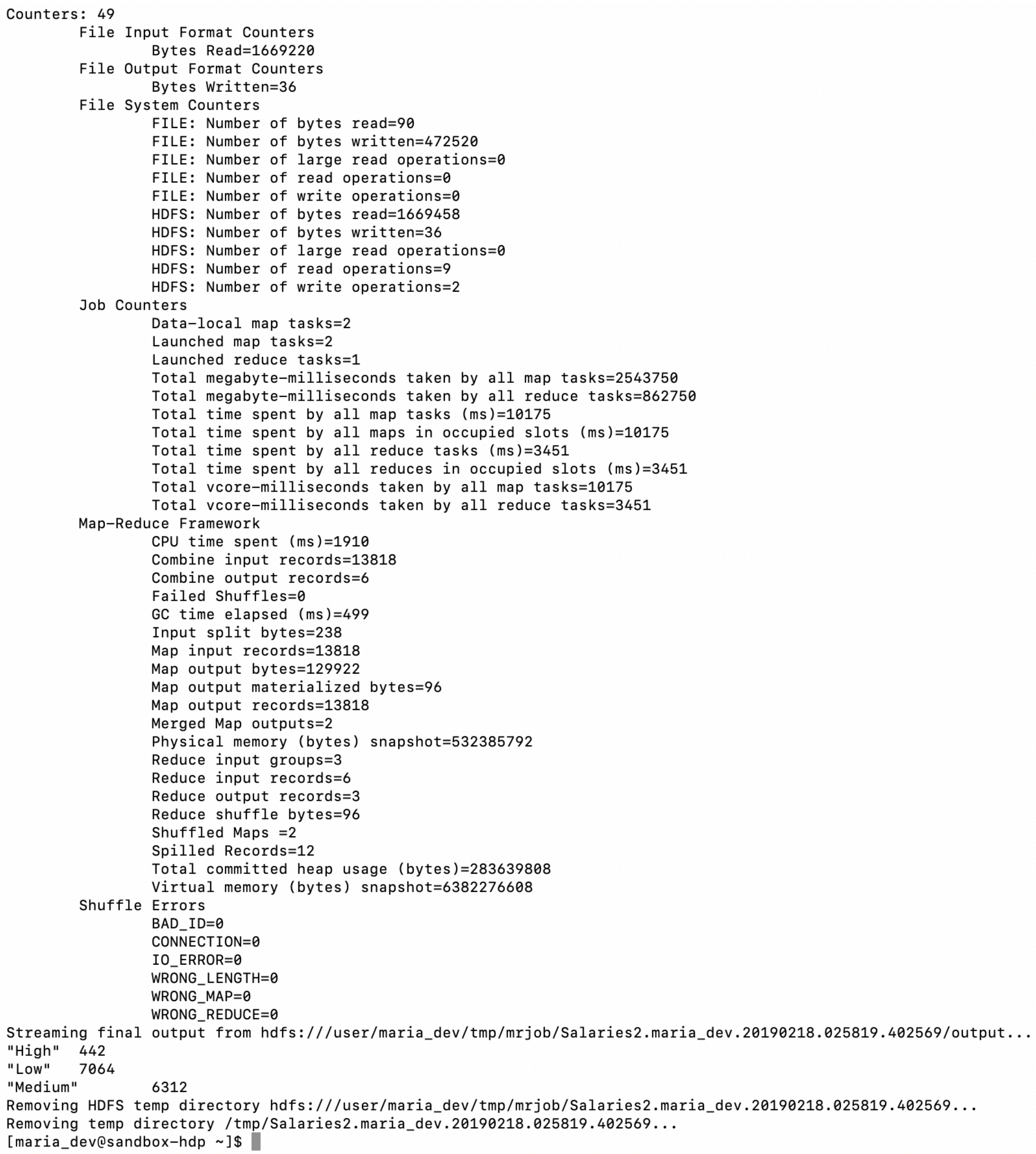


```
  Output directory: hdfs:///user/maria_dev/tmp/mrjob/WordCount2.maria_dev.20190217.205515.572239/output
Counters: 49
        File Input Format Counters
                Bytes Read=791
        File Output Format Counters
                Bytes Written=21
        File System Counters
                FILE: Number of bytes read=55
                FILE: Number of bytes written=472498
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1033
                HDFS: Number of bytes written=21
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Data-local map tasks=2
                Launched map tasks=2
                Launched reduce tasks=1
                Total megabyte-milliseconds taken by all map tasks=2634000
                Total megabyte-milliseconds taken by all reduce tasks=906750
                Total time spent by all map tasks (ms)=10536
                Total time spent by all maps in occupied slots (ms)=10536
                Total time spent by all reduce tasks (ms)=3627
                Total time spent by all reduces in occupied slots (ms)=3627
                Total vcore-milliseconds taken by all map tasks=10536
                Total vcore-milliseconds taken by all reduce tasks=3627
        Map-Reduce Framework
                CPU time spent (ms)=1750
                Combine input records=95
                Combine output records=4
                Failed Shuffles=0
                GC time elapsed (ms)=546
                Input split bytes=242
                Map input records=5
                Map output bytes=898
                Map output materialized bytes=61
                Map output records=95
                Merged Map outputs=2
                Physical memory (bytes) snapshot=528322560
                Reduce input groups=2
                Reduce input records=4
                Reduce output records=2
                Reduce shuffle bytes=61
                Shuffled Maps =2
                Spilled Records=8
                Total committed heap usage (bytes)=290455552
                Virtual memory (bytes) snapshot=6387515392
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
[               IO_ERROR=0
                WRONG_LENGTH=0
[               WRONG_MAP=0
                WRONG_REDUCE=0
[Streaming final output from hdfs:///user/maria_dev/tmp/mrjob/WordCount2.maria_dev.20190217.205515.572239/output...
["Others"       46
"a-n"   49
Removing HDFS temp directory hdfs:///user/maria_dev/tmp/mrjob/WordCount2.maria_dev.20190217.205515.572239...
Removing temp directory /tmp/WordCount2.maria_dev.20190217.205515.572239...
[maria_dev@sandbox-hdp ~]$
```

Question 10) Salaries program.

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.

Screenshot of the output of the modified program Salaries2.py

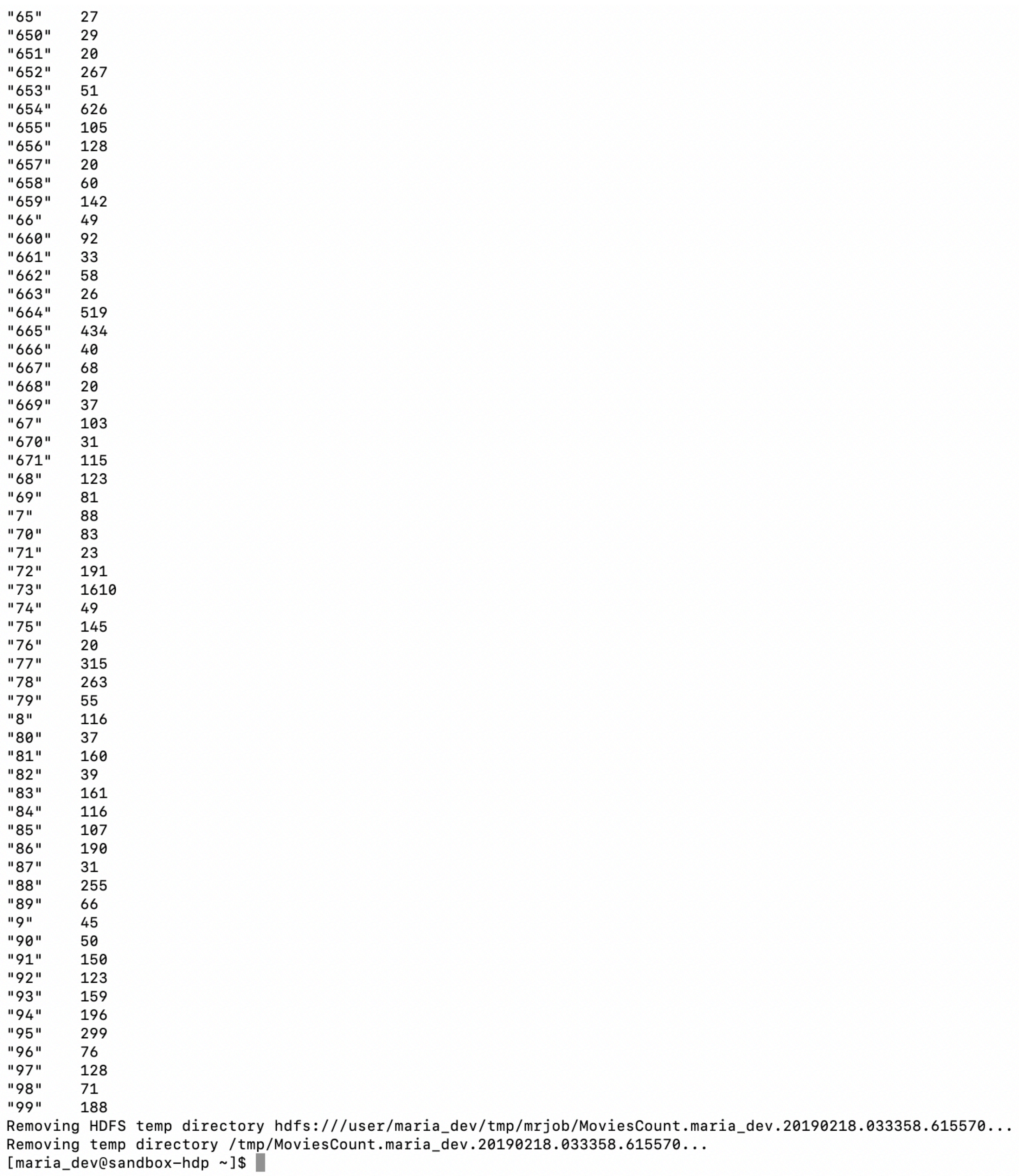


```
Counters: 49
        File Input Format Counters
                Bytes Read=1669220
        File Output Format Counters
                Bytes Written=36
        File System Counters
                FILE: Number of bytes read=90
                FILE: Number of bytes written=472520
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1669458
                HDFS: Number of bytes written=36
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Data-local map tasks=2
                Launched map tasks=2
                Launched reduce tasks=1
                Total megabyte-milliseconds taken by all map tasks=2543750
                Total megabyte-milliseconds taken by all reduce tasks=862750
                Total time spent by all map tasks (ms)=10175
                Total time spent by all maps in occupied slots (ms)=10175
                Total time spent by all reduce tasks (ms)=3451
                Total time spent by all reduces in occupied slots (ms)=3451
                Total vcore-milliseconds taken by all map tasks=10175
                Total vcore-milliseconds taken by all reduce tasks=3451
        Map-Reduce Framework
                CPU time spent (ms)=1910
                Combine input records=13818
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=499
                Input split bytes=238
                Map input records=13818
                Map output bytes=129922
                Map output materialized bytes=96
                Map output records=13818
                Merged Map outputs=2
                Physical memory (bytes) snapshot=532385792
                Reduce input groups=3
                Reduce input records=6
                Reduce output records=3
                Reduce shuffle bytes=96
                Shuffled Maps =2
                Spilled Records=12
                Total committed heap usage (bytes)=283639808
                Virtual memory (bytes) snapshot=6382276608
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/maria_dev/tmp/mrjob/Salaries2.maria_dev.20190218.025819.402569/output...
"High"  442
"Low"   7064
"Medium"        6312
Removing HDFS temp directory hdfs:///user/maria_dev/tmp/mrjob/Salaries2.maria_dev.20190218.025819.402569...
Removing temp directory /tmp/Salaries2.maria_dev.20190218.025819.402569...
[maria_dev@sandbox-hdp ~]$ 
```

Question 12) Write a program to perform the task of outputting a count of the number of movies each user (identified via their user id) reviewed.

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

Screenshot of few lines from the output of the program MoviesCount.py



```
"65"    27
"650"   29
"651"   20
"652"   267
"653"   51
"654"   626
"655"   105
"656"   128
"657"   20
"658"   60
"659"   142
"66"    49
"660"   92
"661"   33
"662"   58
"663"   26
"664"   519
"665"   434
"666"   40
"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"    160
"82"    39
"83"    161
"84"    116
"85"    107
"86"    190
"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/maria_dev/tmp/mrjob/MoviesCount.maria_dev.20190218.033358.615570...
Removing temp directory /tmp/MoviesCount.maria_dev.20190218.033358.615570...
[maria_dev@sandbox-hdp ~]$
```