Akash Tanwani **A20448831**

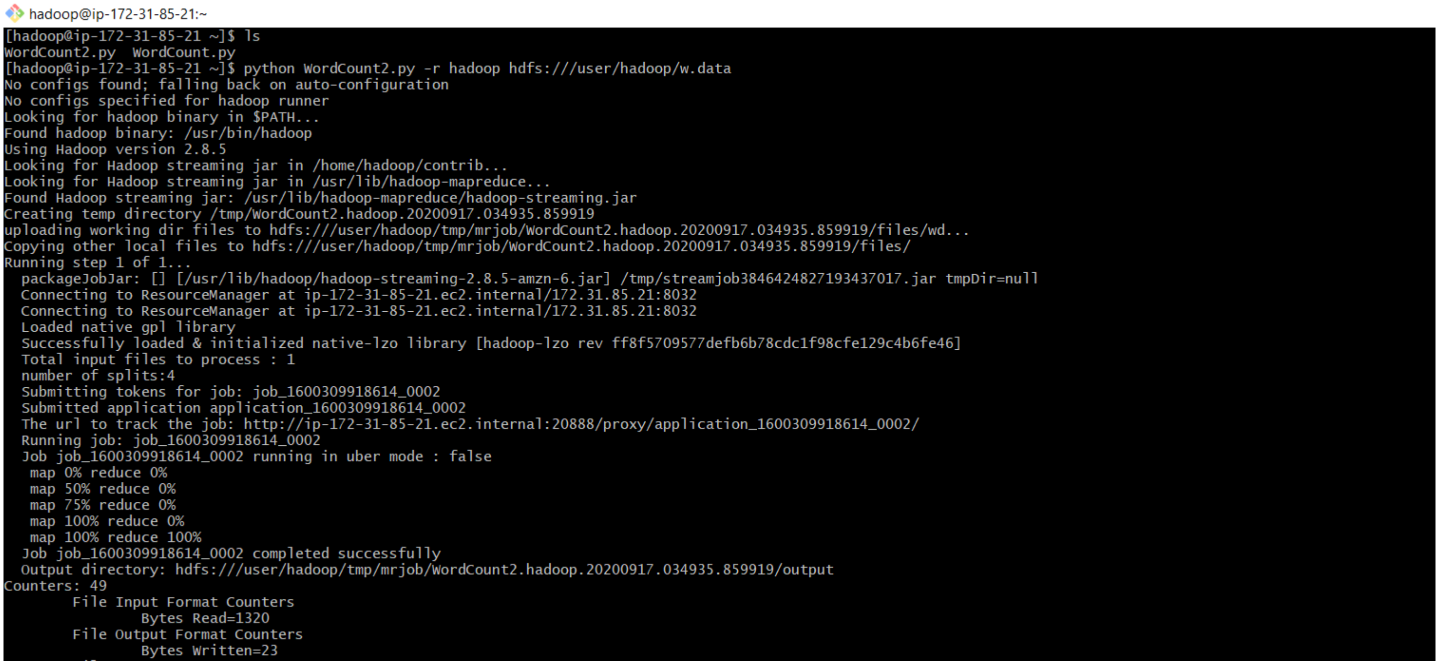
CSP554—Big Data Technologies

Assignment #3 (Modules 03a & 03b, 15 points)

Que 6)

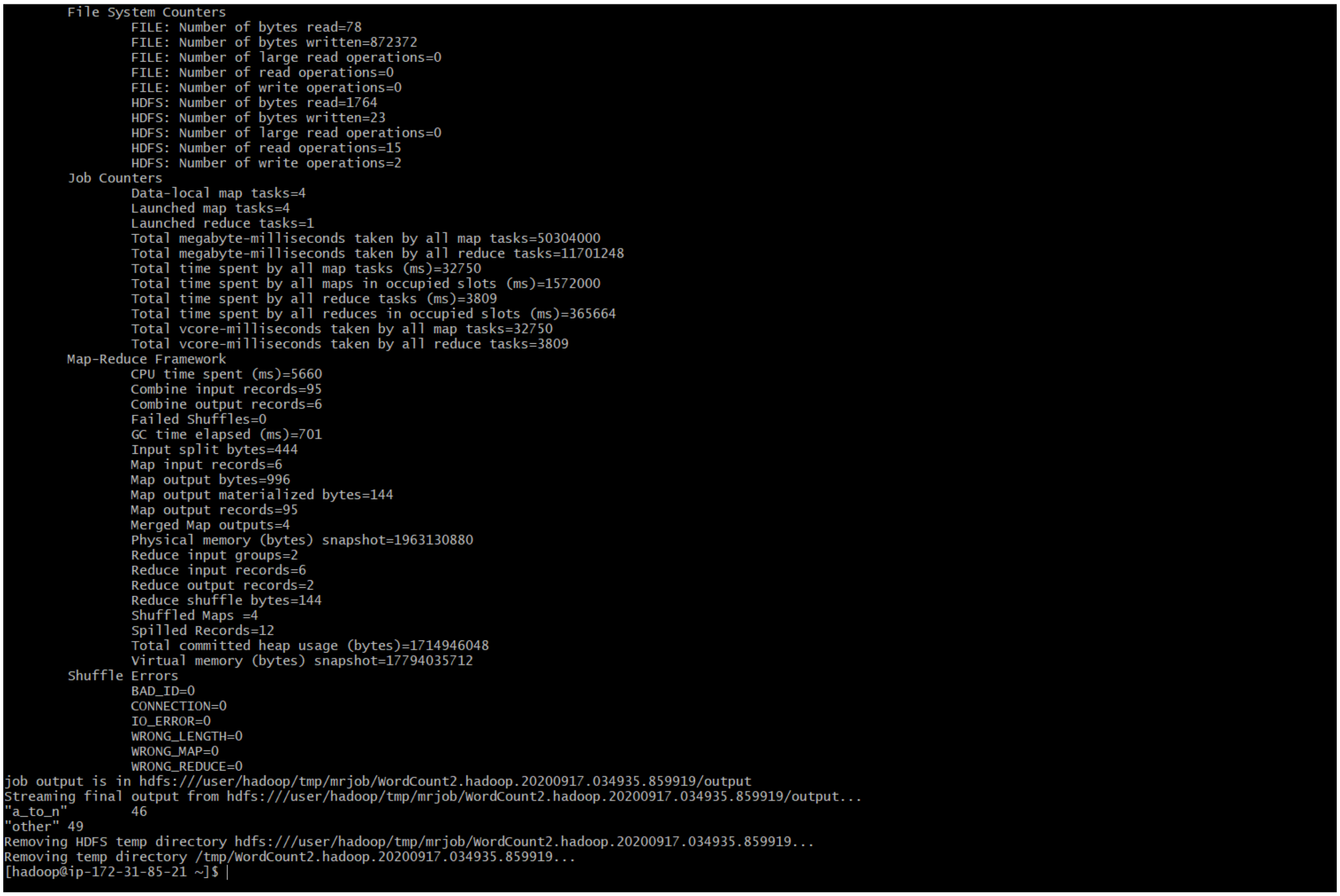
Ans: a_to_n=46

other=49



hadoop@ip-172-31-85-21:~

```
[hadoop@ip-172-31-85-21 ~]$ ls
WordCount2.py  WordCount.py
[hadoop@ip-172-31-85-21 ~]$ 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 2.8.5
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.20200917.034935.859919
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20200917.034935.859919/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20200917.034935.859919/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-6.jar] /tmp/streamjob3846424827193437017.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-85-21.ec2.internal/172.31.85.21:8032
  Connecting to ResourceManager at ip-172-31-85-21.ec2.internal/172.31.85.21:8032
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev ff8f5709577defb6b78cdc1f98cfe129c4b6fe46]
  Total input files to process : 1
  number of splits:4
  Submitting tokens for job: job_1600309918614_0002
  Submitted application application_1600309918614_0002
  The url to track the job: http://ip-172-31-85-21.ec2.internal:20888/proxy/application_1600309918614_0002/
  Running job: job_1600309918614_0002
  Job job_1600309918614_0002 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1600309918614_0002 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20200917.034935.859919/output
Counters: 49
        File Input Format Counters
                Bytes Read=1320
        File Output Format Counters
                Bytes Written=23
```



```
        File System Counters
                FILE: Number of bytes read=78
                FILE: Number of bytes written=872372
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1764
                HDFS: Number of bytes written=23
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=15
                HDFS: Number of write operations=2
        Job Counters
                Data-local map tasks=4
                Launched map tasks=4
                Launched reduce tasks=1
                Total megabyte-milliseconds taken by all map tasks=50304000
                Total megabyte-milliseconds taken by all reduce tasks=11701248
                Total time spent by all map tasks (ms)=32750
                Total time spent by all maps in occupied slots (ms)=1572000
                Total time spent by all reduce tasks (ms)=3809
                Total time spent by all reduces in occupied slots (ms)=365664
                Total vcore-milliseconds taken by all map tasks=32750
                Total vcore-milliseconds taken by all reduce tasks=3809
        Map-Reduce Framework
                CPU time spent (ms)=5660
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=701
                Input split bytes=444
                Map input records=6
                Map output bytes=996
                Map output materialized bytes=144
                Map output records=95
                Merged Map outputs=4
                Physical memory (bytes) snapshot=1963130880
                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)=1714946048
                Virtual memory (bytes) snapshot=17794035712
        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.20200917.034935.859919/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20200917.034935.859919/output...
"a_to_n"        46
"other" 49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20200917.034935.859919...
Removing temp directory /tmp/WordCount2.hadoop.20200917.034935.859919...
[hadoop@ip-172-31-85-21 ~]$ |
```

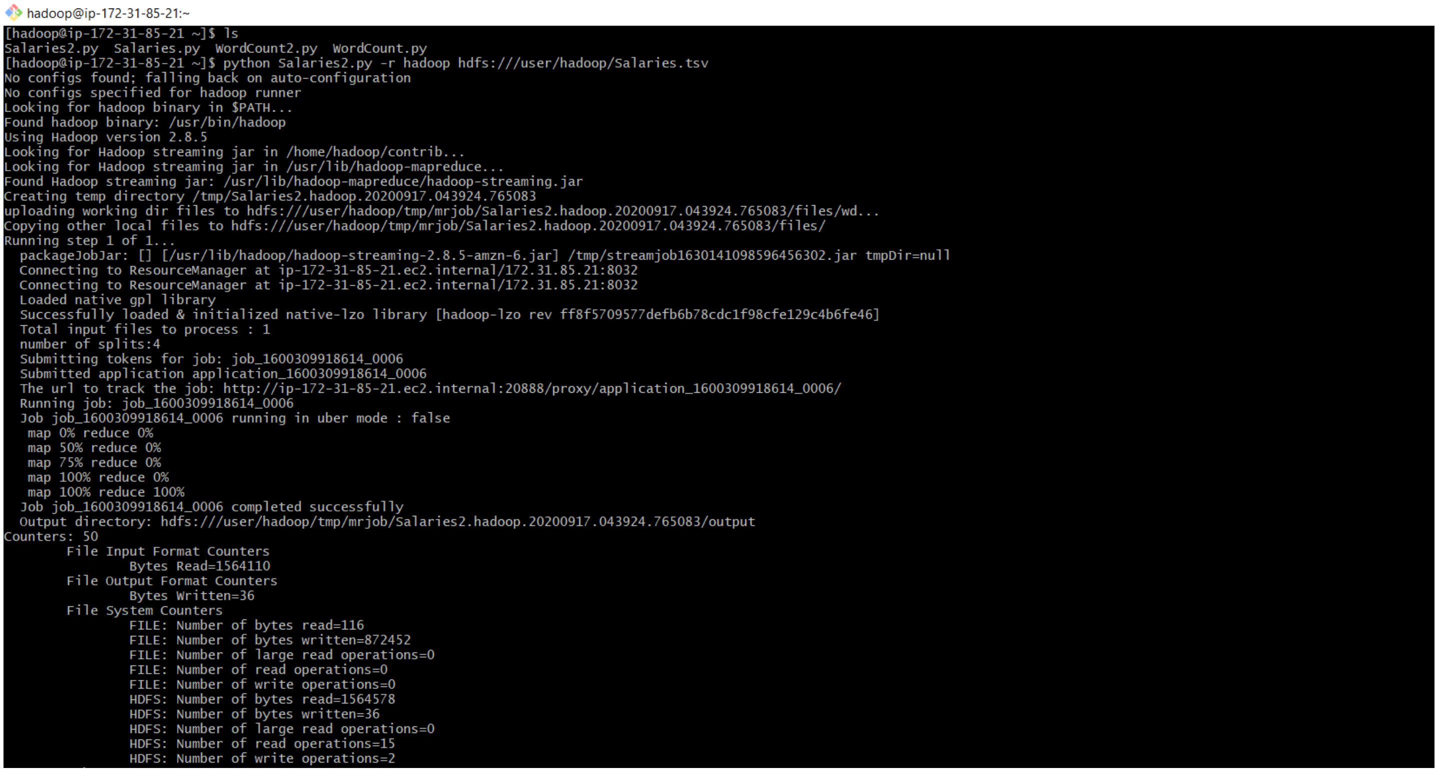
Que 10)

Ans:

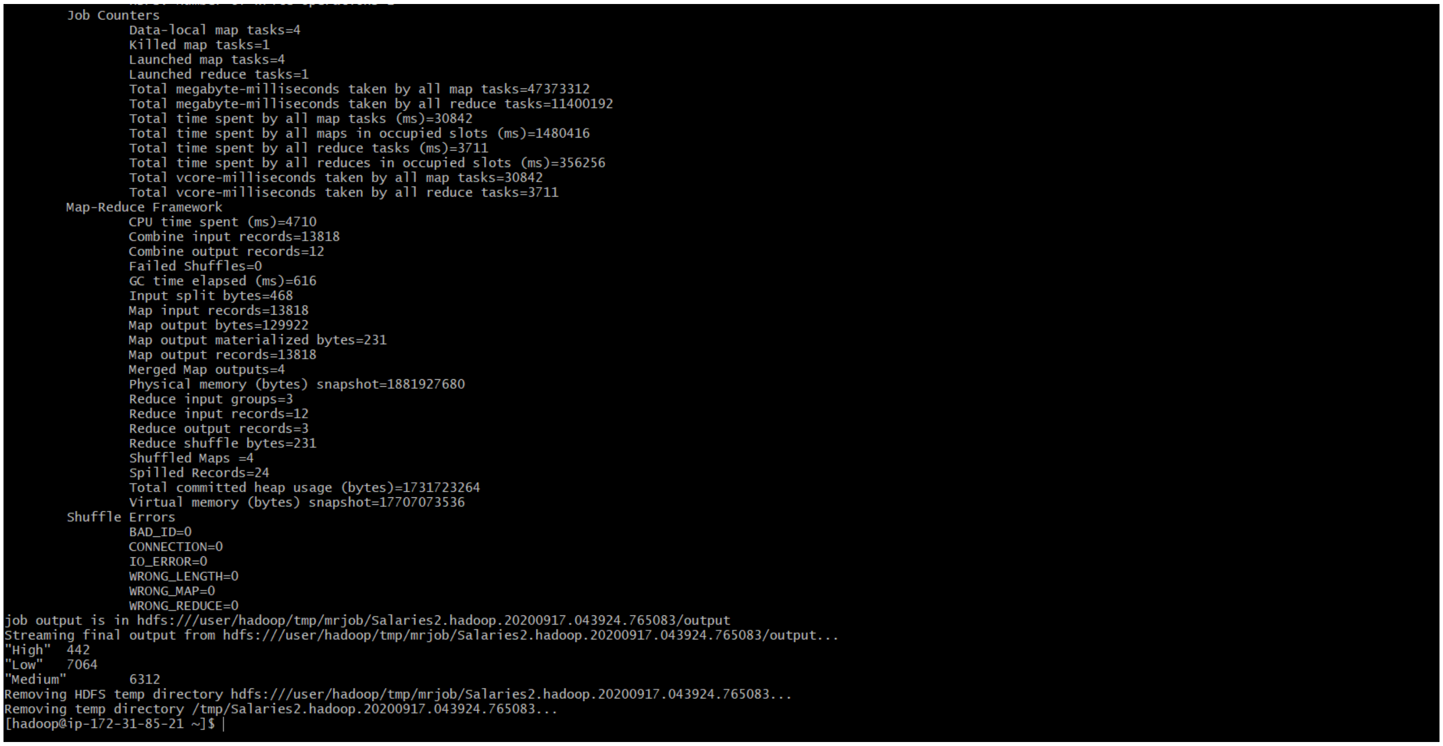
Low=7064

Medium=6312

High=442



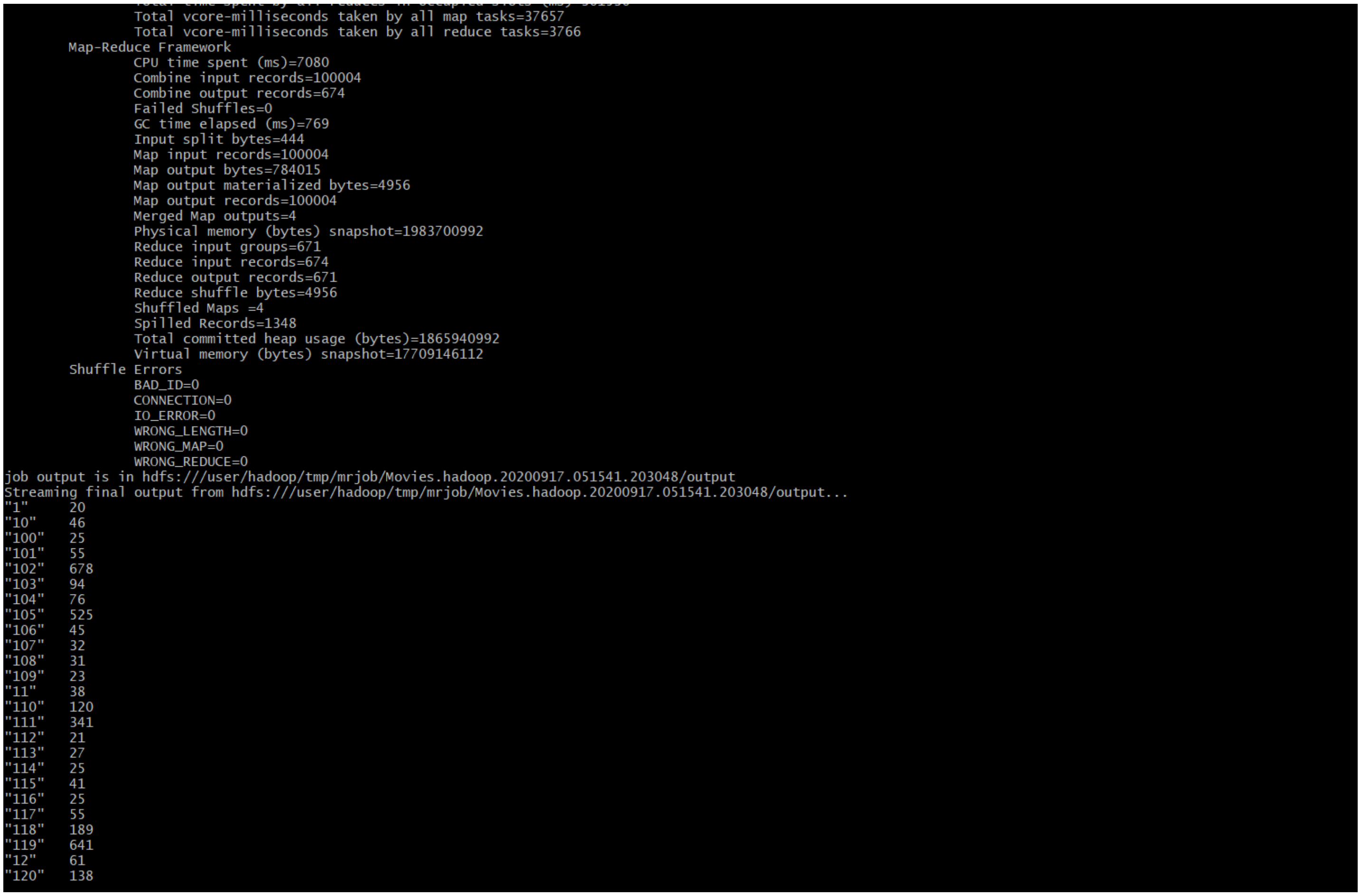
```
[hadoop@ip-172-31-85-21 ~]$ ls
Salaries2.py  Salaries.py  WordCount2.py  WordCount.py
[hadoop@ip-172-31-85-21 ~]$ python Salaries2.py -r hadoop hdfs:///user/hadoop/Salaries.tsv
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.8.5
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.20200917.043924.765083
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20200917.043924.765083/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20200917.043924.765083/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-6.jar] /tmp/streamjob1630141098596456302.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-85-21.ec2.internal/172.31.85.21:8032
  Connecting to ResourceManager at ip-172-31-85-21.ec2.internal/172.31.85.21:8032
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev ff8f5709577defb6b78cdc1f98cfe129c4b6fe46]
  Total input files to process : 1
  number of splits:4
  Submitting tokens for job: job_1600309918614_0006
  Submitted application application_1600309918614_0006
  The url to track the job: http://ip-172-31-85-21.ec2.internal:20888/proxy/application_1600309918614_0006/
  Running job: job_1600309918614_0006
  Job job_1600309918614_0006 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1600309918614_0006 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20200917.043924.765083/output
Counters: 50
        File Input Format Counters
                Bytes Read=1564110
        File Output Format Counters
                Bytes Written=36
        File System Counters
                FILE: Number of bytes read=116
                FILE: Number of bytes written=872452
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1564578
                HDFS: Number of bytes written=36
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=15
                HDFS: Number of write operations=2
```



```
        Job Counters
                Data-local map tasks=4
                Killed map tasks=1
                Launched map tasks=4
                Launched reduce tasks=1
                Total megabyte-milliseconds taken by all map tasks=47373312
                Total megabyte-milliseconds taken by all reduce tasks=11400192
                Total time spent by all map tasks (ms)=30842
                Total time spent by all maps in occupied slots (ms)=1480416
                Total time spent by all reduce tasks (ms)=3711
                Total time spent by all reduces in occupied slots (ms)=356256
                Total vcore-milliseconds taken by all map tasks=30842
                Total vcore-milliseconds taken by all reduce tasks=3711
        Map-Reduce Framework
                CPU time spent (ms)=4710
                Combine input records=13818
                Combine output records=12
                Failed Shuffles=0
                GC time elapsed (ms)=616
                Input split bytes=468
                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=1881927680
                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)=1731723264
                Virtual memory (bytes) snapshot=17707073536
        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.20200917.043924.765083/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20200917.043924.765083/output...
"High"  442
"Low"   7064
"Medium"        6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20200917.043924.765083...
Removing temp directory /tmp/Salaries2.hadoop.20200917.043924.765083...
[hadoop@ip-172-31-85-21 ~]$ |
```

Que 12)

Ans:



```
                Total time spent by all reduces in occupied slots (ms)=[illegible]
                Total vcore-milliseconds taken by all map tasks=37657
                Total vcore-milliseconds taken by all reduce tasks=3766
        Map-Reduce Framework
                CPU time spent (ms)=7080
                Combine input records=100004
                Combine output records=674
                Failed Shuffles=0
                GC time elapsed (ms)=769
                Input split bytes=444
                Map input records=100004
                Map output bytes=784015
                Map output materialized bytes=4956
                Map output records=100004
                Merged Map outputs=4
                Physical memory (bytes) snapshot=1983700992
                Reduce input groups=671
                Reduce input records=674
                Reduce output records=671
                Reduce shuffle bytes=4956
                Shuffled Maps =4
                Spilled Records=1348
                Total committed heap usage (bytes)=1865940992
                Virtual memory (bytes) snapshot=17709146112
        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/Movies.hadoop.20200917.051541.203048/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20200917.051541.203048/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
```