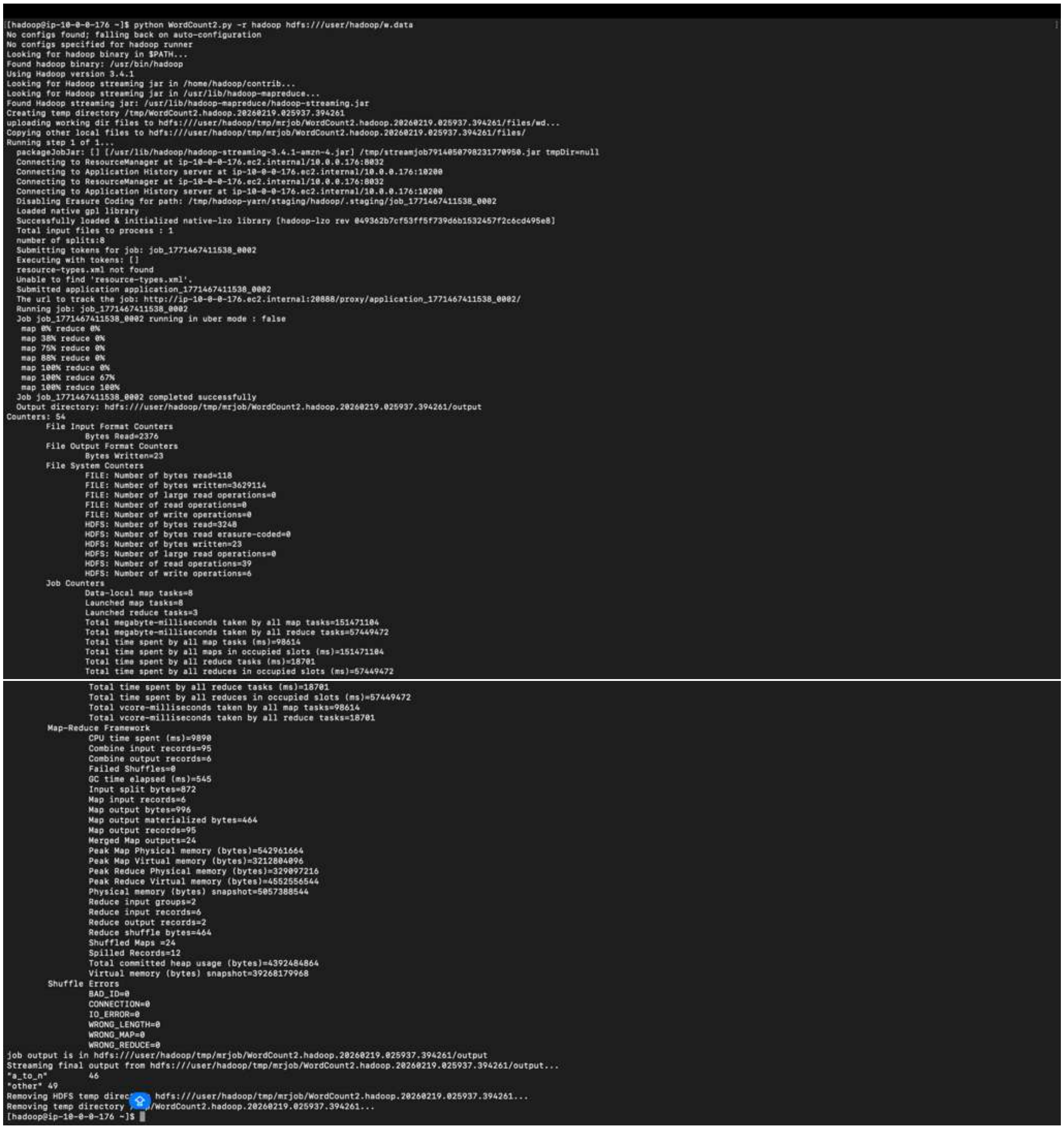
Nishant Khandhar

A20581012 | CSP 554 - Big Data Technologies | Spring 2026 | Assignment 3

Q6. WordCount2.py results:

`a_to_n = 46`

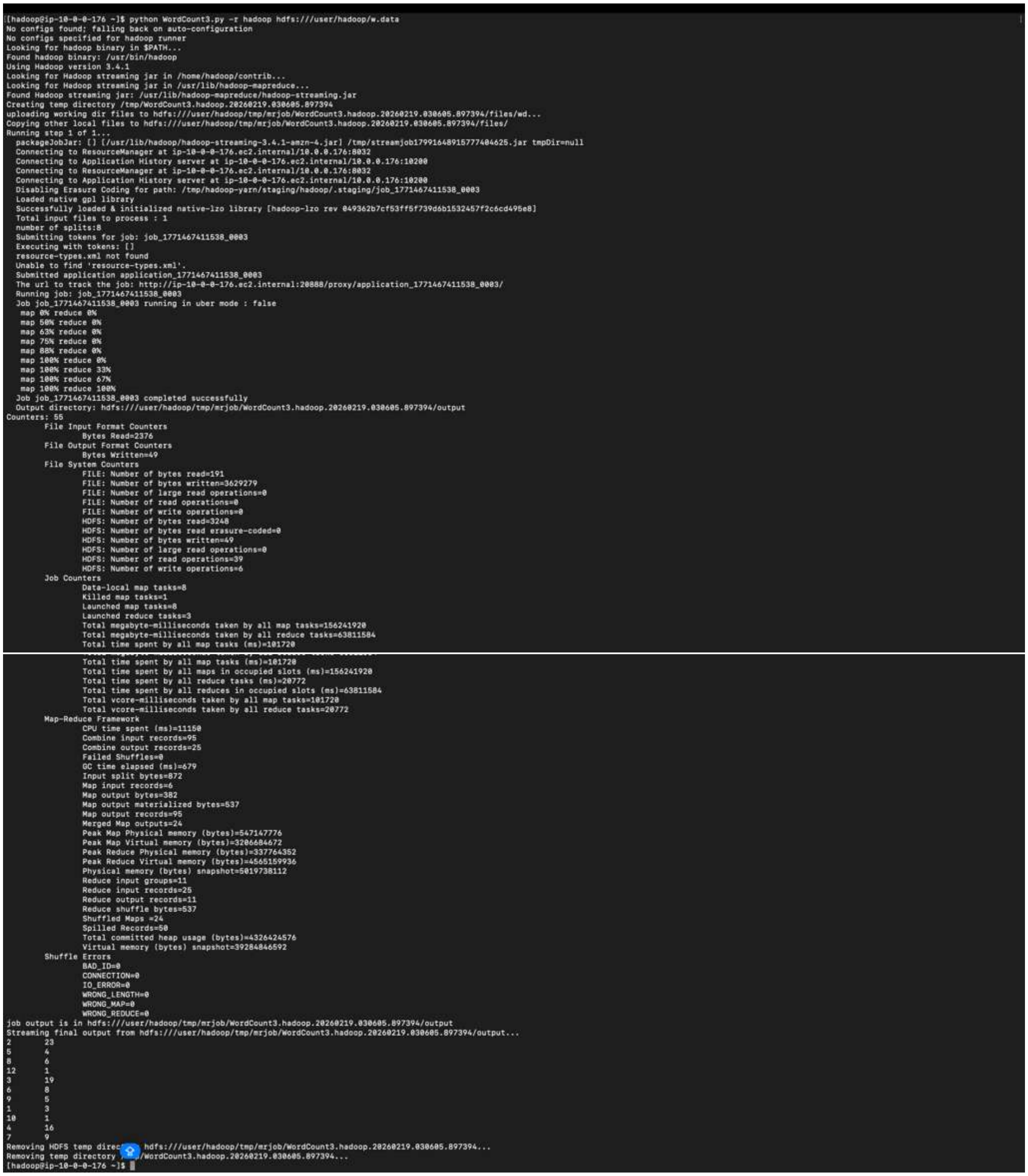
`other = 49`



```
[hadoop@ip-10-0-0-176 ~]$ 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 3.4.1
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.20260219.025937.394261
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20260219.025937.394261/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20260219.025937.394261/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.4.1-amzn-4.jar] /tmp/streamjob7914050798231770950.jar tmpDir=null
  Connecting to ResourceManager at ip-10-0-0-176.ec2.internal/10.0.0.176:8032
  Connecting to Application History server at ip-10-0-0-176.ec2.internal/10.0.0.176:10200
  Connecting to ResourceManager at ip-10-0-0-176.ec2.internal/10.0.0.176:8032
  Connecting to Application History server at ip-10-0-0-176.ec2.internal/10.0.0.176:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1771467411538_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_1771467411538_0002
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1771467411538_0002
  The url to track the job: http://ip-10-0-0-176.ec2.internal:20888/proxy/application_1771467411538_0002/
  Running job: job_1771467411538_0002
  Job job_1771467411538_0002 running in uber mode : false
   map 0% reduce 0%
   map 38% reduce 0%
   map 75% reduce 0%
   map 88% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1771467411538_0002 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20260219.025937.394261/output
Counters: 54
        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=3629114
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3248
                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
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=151471104
                Total megabyte-milliseconds taken by all reduce tasks=57449472
                Total time spent by all map tasks (ms)=98614
                Total time spent by all maps in occupied slots (ms)=151471104
                Total time spent by all reduce tasks (ms)=18701
                Total time spent by all reduces in occupied slots (ms)=57449472
                Total time spent by all reduce tasks (ms)=18701
                Total time spent by all reduces in occupied slots (ms)=57449472
                Total vcore-milliseconds taken by all map tasks=98614
                Total vcore-milliseconds taken by all reduce tasks=18701
        Map-Reduce Framework
                CPU time spent (ms)=9890
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=545
                Input split bytes=872
                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)=542961664
                Peak Map Virtual memory (bytes)=3212804096
                Peak Reduce Physical memory (bytes)=329097216
                Peak Reduce Virtual memory (bytes)=4552556544
                Physical memory (bytes) snapshot=5057388544
                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)=4392484864
                Virtual memory (bytes) snapshot=39268179968
        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.20260219.025937.394261/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20260219.025937.394261/output...
"a_to_n"        46
"other" 49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20260219.025937.394261...
Removing temp directory /tmp/WordCount2.hadoop.20260219.025937.394261...
[hadoop@ip-10-0-0-176 ~]$ 
```

Q8. WordCount3.py results:

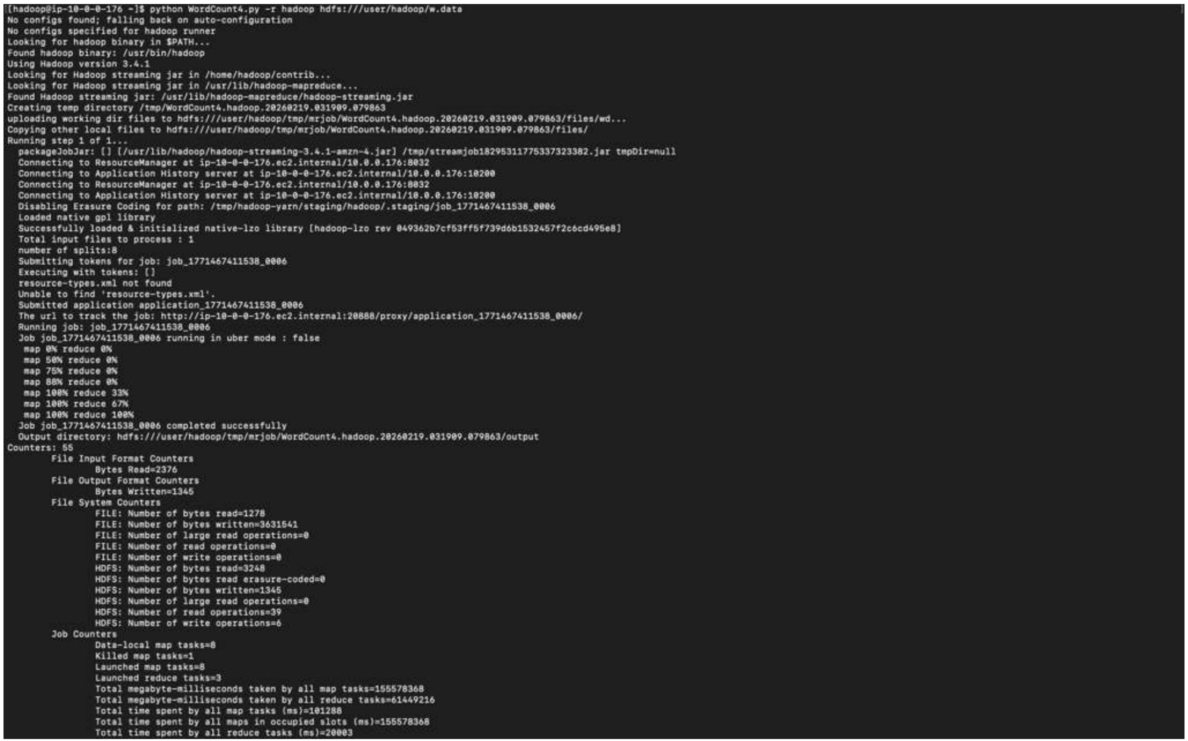
Output



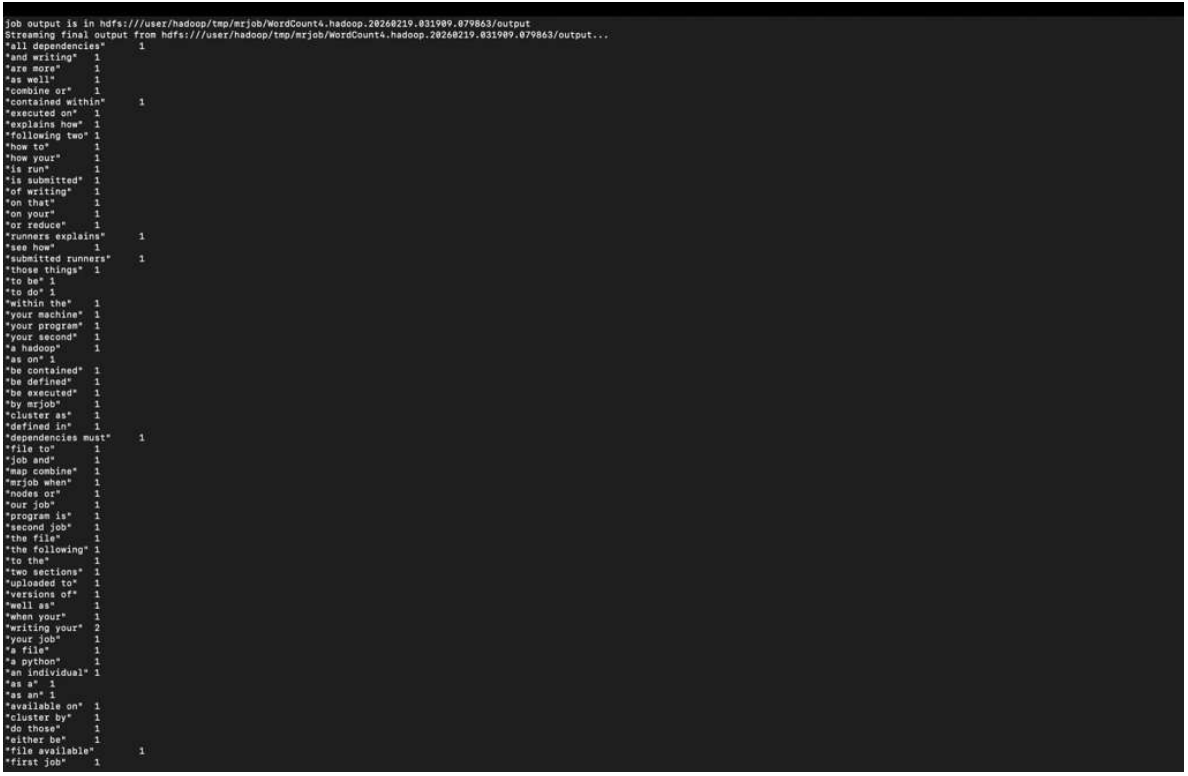
```
[hadoop@ip-10-0-0-176 ~]$ python WordCount3.py -r hadoop hdfs:///user/hadoop/w.data
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.4.1
Looking for Hadoop 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/WordCount3.hadoop.20260219.030605.897394
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20260219.030605.897394/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20260219.030605.897394/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.4.1-amzn-4.jar] /tmp/streamjob17991648915777404625.jar tmpDir=null
  Connecting to ResourceManager at ip-10-0-0-176.ec2.internal/10.0.0.176:8032
  Connecting to Application History server at ip-10-0-0-176.ec2.internal/10.0.0.176:10200
  Connecting to ResourceManager at ip-10-0-0-176.ec2.internal/10.0.0.176:8032
  Connecting to Application History server at ip-10-0-0-176.ec2.internal/10.0.0.176:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1771467411538_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_1771467411538_0003
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1771467411538_0003
  The url to track the job: http://ip-10-0-0-176.ec2.internal:20888/proxy/application_1771467411538_0003/
  Running job: job_1771467411538_0003
  Job job_1771467411538_0003 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 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1771467411538_0003 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20260219.030605.897394/output
Counters: 55
        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=3629279
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3248
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=49
                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=156241920
                Total megabyte-milliseconds taken by all reduce tasks=63811584
                Total time spent by all map tasks (ms)=101720
                Total time spent by all map tasks (ms)=101720
                Total time spent by all maps in occupied slots (ms)=156241920
                Total time spent by all reduce tasks (ms)=20772
                Total time spent by all reduces in occupied slots (ms)=63811584
                Total vcore-milliseconds taken by all map tasks=101720
                Total vcore-milliseconds taken by all reduce tasks=20772
        Map-Reduce Framework
                CPU time spent (ms)=11160
                Combine input records=95
                Combine output records=25
                Failed Shuffles=0
                GC time elapsed (ms)=679
                Input split bytes=872
                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)=547147776
                Peak Map Virtual memory (bytes)=3206684672
                Peak Reduce Physical memory (bytes)=337764352
                Peak Reduce Virtual memory (bytes)=4565159936
                Physical memory (bytes) snapshot=5019738112
                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)=4326424576
                Virtual memory (bytes) snapshot=39284846592
        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.20260219.030605.897394/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20260219.030605.897394/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 direc[illegible] hdfs:///user/hadoop/tmp/mrjob/WordCount3.hadoop.20260219.030605.897394...
Removing temp directory [illegible]/WordCount3.hadoop.20260219.030605.897394...
[hadoop@ip-10-0-0-176 ~]$
```

Q10. WordCount4.py results:

Output



```
[hadoop@ip-10-0-0-176 ~]$ 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 3.4.1
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/WordCount4.hadoop.20260219.031909.079863
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20260219.031909.079863/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20260219.031909.079863/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.4.1-amzn-4.jar] /tmp/streamjob18295311775337323382.jar tmpDir=null
  Connecting to ResourceManager at ip-10-0-0-176.ec2.internal/10.0.0.176:8032
  Connecting to Application History server at ip-10-0-0-176.ec2.internal/10.0.0.176:10200
  Connecting to ResourceManager at ip-10-0-0-176.ec2.internal/10.0.0.176:8032
  Connecting to Application History server at ip-10-0-0-176.ec2.internal/10.0.0.176:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1771467411538_0006
  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_1771467411538_0006
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1771467411538_0006
  The url to track the job: http://ip-10-0-0-176.ec2.internal:20888/proxy/application_1771467411538_0006/
  Running job: job_1771467411538_0006
  Job job_1771467411538_0006 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 75% reduce 0%
   map 88% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1771467411538_0006 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20260219.031909.079863/output
Counters: 55
        File Input Format Counters
                Bytes Read=2376
        File Output Format Counters
                Bytes Written=1345
        File System Counters
                FILE: Number of bytes read=1278
                FILE: Number of bytes written=3631541
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3248
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=1345
                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=155578368
                Total megabyte-milliseconds taken by all reduce tasks=61449216
                Total time spent by all map tasks (ms)=101288
                Total time spent by all maps in occupied slots (ms)=155578368
                Total time spent by all reduce tasks (ms)=20003
```



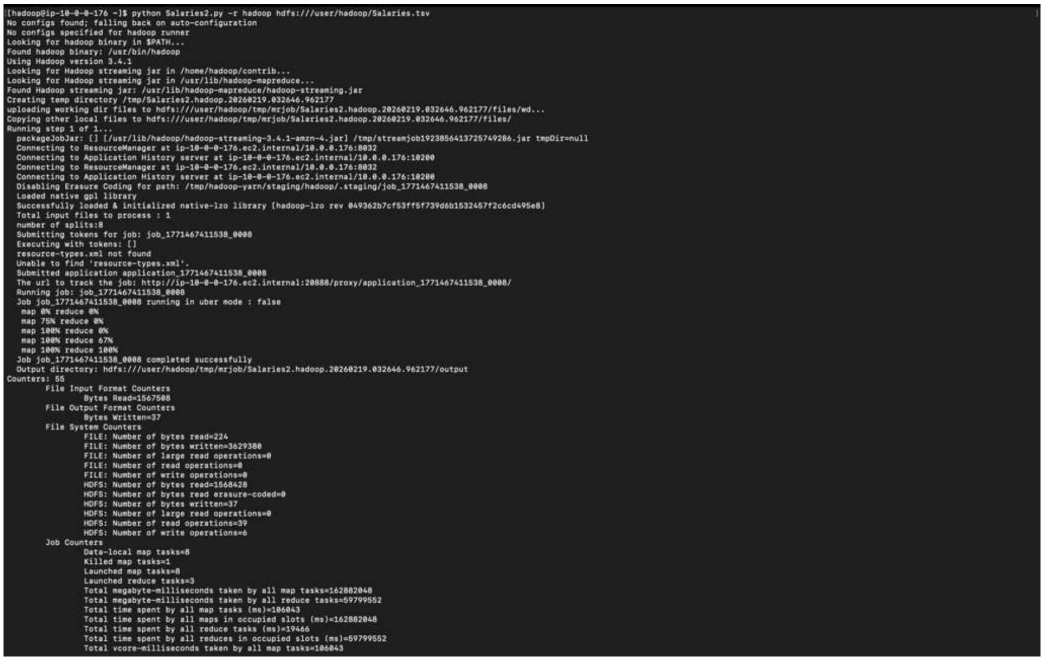
```
job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20260219.031909.079863/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount4.hadoop.20260219.031909.079863/output...
"all dependencies"      1
"and writing"   1
"are more"      1
"as well"       1
"combine or"    1
"contained within"      1
"executed on"   1
"explains how"  1
"following two" 1
"how to"        1
"how your"      1
"is run"        1
"is submitted"  1
"of writing"    1
"on that"       1
"on your"       1
"or reduce"     1
"runners explains"      1
"see how"       1
"submitted runners"     1
"those things"  1
"to be" 1
"to do" 1
"within the"    1
"your machine"  1
"your program"  1
"your second"   1
"a hadoop"      1
"as on" 1
"be contained"  1
"be defined"    1
"be executed"   1
"by mrjob"      1
"cluster as"    1
"defined in"    1
"dependencies must"     1
"file to"       1
"job and"       1
"map combine"   1
"mrjob when"    1
"nodes or"      1
"our job"       1
"program is"    1
"second job"    1
"the file"      1
"the following" 1
"to the"        1
"two sections"  1
"uploaded to"   1
"versions of"   1
"well as"       1
"when your"     1
"writing your"  2
"your job"      1
"a file"        1
"a python"      1
"an individual" 1
"as a"  1
"as an" 1
"available on"  1
"cluster by"    1
"do those"      1
"either be"     1
"file available"        1
"first job"     1
```

Q14. Salaries2.py results:

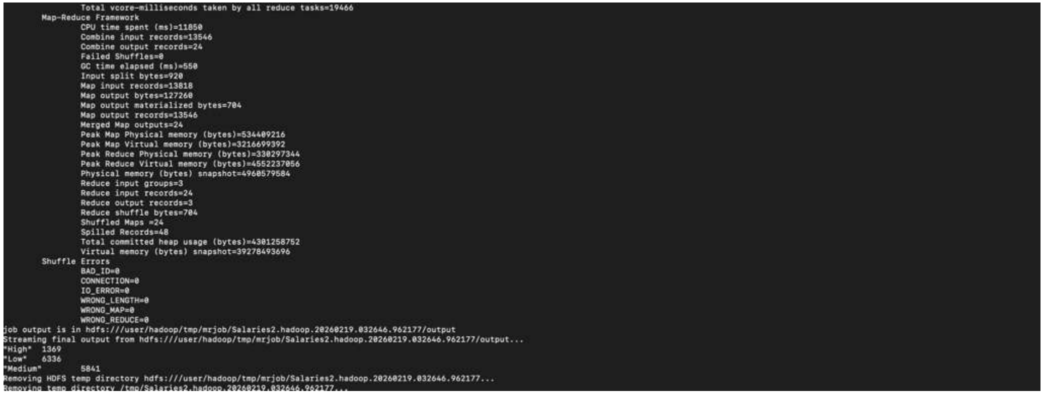
```
high = 1369

low = 6336

medium = 5841
```



```
[hadoop@ip-10-0-0-176 ~]$ 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 3.4.1
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.20260219.032646.962177
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20260219.032646.962177/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20260219.032646.962177/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.4.1-amzn-4.jar] /tmp/streamjob1923856413725749286.jar tmpDir=null
  Connecting to ResourceManager at ip-10-0-0-176.ec2.internal/10.0.0.176:8032
  Connecting to Application History server at ip-10-0-0-176.ec2.internal/10.0.0.176:10200
  Connecting to ResourceManager at ip-10-0-0-176.ec2.internal/10.0.0.176:8032
  Connecting to Application History server at ip-10-0-0-176.ec2.internal/10.0.0.176:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1771467411538_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_1771467411538_0008
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1771467411538_0008
  The url to track the job: http://ip-10-0-0-176.ec2.internal:20888/proxy/application_1771467411538_0008/
  Running job: job_1771467411538_0008
  Job job_1771467411538_0008 running in uber mode : false
   map 0% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1771467411538_0008 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20260219.032646.962177/output
Counters: 55
        File Input Format Counters
                Bytes Read=1567508
        File Output Format Counters
                Bytes Written=37
        File System Counters
                FILE: Number of bytes read=224
                FILE: Number of bytes written=3629380
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1568428
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=37
                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=162882048
                Total megabyte-milliseconds taken by all reduce tasks=59799552
                Total time spent by all map tasks (ms)=106043
                Total time spent by all maps in occupied slots (ms)=162882048
                Total time spent by all reduce tasks (ms)=19466
                Total time spent by all reduces in occupied slots (ms)=59799552
                Total vcore-milliseconds taken by all map tasks=106043
```



```
                Total vcore-milliseconds taken by all reduce tasks=19466
        Map-Reduce Framework
                CPU time spent (ms)=11850
                Combine input records=13546
                Combine output records=24
                Failed Shuffles=0
                GC time elapsed (ms)=550
                Input split bytes=920
                Map input records=13818
                Map output bytes=127260
                Map output materialized bytes=704
                Map output records=13546
                Merged Map outputs=24
                Peak Map Physical memory (bytes)=534409216
                Peak Map Virtual memory (bytes)=3216699392
                Peak Reduce Physical memory (bytes)=330297344
                Peak Reduce Virtual memory (bytes)=4552237056
                Physical memory (bytes) snapshot=4960579584
                Reduce input groups=3
                Reduce input records=24
                Reduce output records=3
                Reduce shuffle bytes=704
                Shuffled Maps =24
                Spilled Records=48
                Total committed heap usage (bytes)=4301258752
                Virtual memory (bytes) snapshot=39278493696
        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.20260219.032646.962177/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20260219.032646.962177/output...
"High"  1369
"Low"   6336
"Medium"        5841
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20260219.032646.962177...
Removing temp directory /tmp/Salaries2.hadoop.20260219.032646.962177...
```