Shraddha Patel A20499171

CSP554—Big Data Technologies

Assignment-3

6) (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.

WordCount2.py

```
from mrjob.job import MRJob
import re

WORD_RE = re.compile(r"[\w']+")

class MRWordCount(MRJob):

def mapper(self, _, line):
for word in WORD_RE.findall(line):
if word[0].lower() >= 'a' and word[0].lower() <= 'n':
yield "a_to_n", 1
else:
yield "other", 1

def combiner(self, word, counts):
yield word, sum(counts)

def reducer(self, word, counts):
yield word, sum(counts)

if __name__ == '__main__':
MRWordCount.run()
```

Execution Snapshot:

```
[hadoop@ip-172-31-25-241 ~]$ python WordCount2.py -r hadoop hdfs:///user/hadoop/w.data --output-dir /user/hadoop/Q5
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.20220131.201154.913713
Uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220131.201154.913713/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220131.201154.913713/files/
Running step 1 of 1...
  SLF4J: Class path contains multiple SLF4J bindings.
  SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
  SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
  SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
  SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-2.jar] /tmp/streamjob697715484561796325.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-25-241.ec2.internal/172.31.25.241:8032
  Connecting to Application History server at ip-172-31-25-241.ec2.internal/172.31.25.241:10200
  Connecting to ResourceManager at ip-172-31-25-241.ec2.internal/172.31.25.241:8032
  Connecting to Application History server at ip-172-31-25-241.ec2.internal/172.31.25.241:10200
  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_1643653747885_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_1643653747885_0004
  The url to track the job: http://ip-172-31-25-241.ec2.internal:20888/proxy/application_1643653747885_0004/
  Running job: job_1643653747885_0004
  Job job_1643653747885_0004 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_1643653747885_0004 completed successfully
  Output directory: hdfs:///user/hadoop/Q5
Counters: 50
        File Input Format Counters
                Bytes Read=1320
        File Output Format Counters
```

```
 Output directory: hdfs:///user/hadoop/Q5
Counters: 50
        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=1126492
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1768
                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
                Killed map tasks=1
                Launched map tasks=4
                Launched reduce tasks=1
                Total megabyte-milliseconds taken by all map tasks=59771904
                Total megabyte-milliseconds taken by all reduce tasks=12847104
                Total time spent by all map tasks (ms)=38914
                Total time spent by all maps in occupied slots (ms)=1867872
                Total time spent by all reduce tasks (ms)=4182
                Total time spent by all reduces in occupied slots (ms)=401472
                Total vcore-milliseconds taken by all map tasks=38914
                Total vcore-milliseconds taken by all reduce tasks=4182
        Map-Reduce Framework
                CPU time spent (ms)=5250
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=1007
                Input split bytes=448
                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=2141683712
                Reduce input groups=2
                Reduce input records=6
                Reduce output records=2
                Reduce shuffle bytes=144
```

```
                Total time spent by all map tasks (ms)=38914
                Total time spent by all maps in occupied slots (ms)=1867872
                Total time spent by all reduce tasks (ms)=4182
                Total time spent by all reduces in occupied slots (ms)=401472
                Total vcore-milliseconds taken by all map tasks=38914
                Total vcore-milliseconds taken by all reduce tasks=4182
        Map-Reduce Framework
                CPU time spent (ms)=5250
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=1007
                Input split bytes=448
                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=2141683712
                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)=1760559104
                Virtual memory (bytes) snapshot=17868836864
        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/Q5
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220131.201154.913713...
Removing temp directory /tmp/WordCount2.hadoop.20220131.201154.913713...
[hadoop@ip-172-31-25-241 ~]$ hdfs dfs -cat Q5/part-00000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
"a_to_n"        49
"other" 46
```

Output:

```
[hadoop@ip-172-31-25-241 ~]$ hdfs dfs -cat Q5/part-00000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
"a_to_n"        49
"other" 46
```

9) Now modify the Salaries.py program. Call it Salaries2.py

Instead of counting the number of workers per department, change the program to provide the number of workers having High, Medium or Low annual salaries. This is defined as follows:

High	100,000.00 and above
Medium	50,000.00 to 99,999.99
Low	0.00 to 49,999.99

The output of the program should be something like the following (in any order):

High 20

Medium 30

Low 10

Some important hints:

- The annual salary is a string that will need to be converted to a float.
- The mapper should output tuples with one of three keys depending on the annual salary: High, Medium and Low
- The value part of the tuple is not a salary. (What should it be?)

Now execute the program and see what happens.

10) (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.

Salaries2.py

```
from mrjob.job import MRJob

class MRSalaries(MRJob):

    def mapper(self, _, line):
        (name,jobTitle,agencyID,agency,hireDate,annualSalary,grossPay) = line.split('\t')
        if float(annualSalary) >= float(100000.00):
```

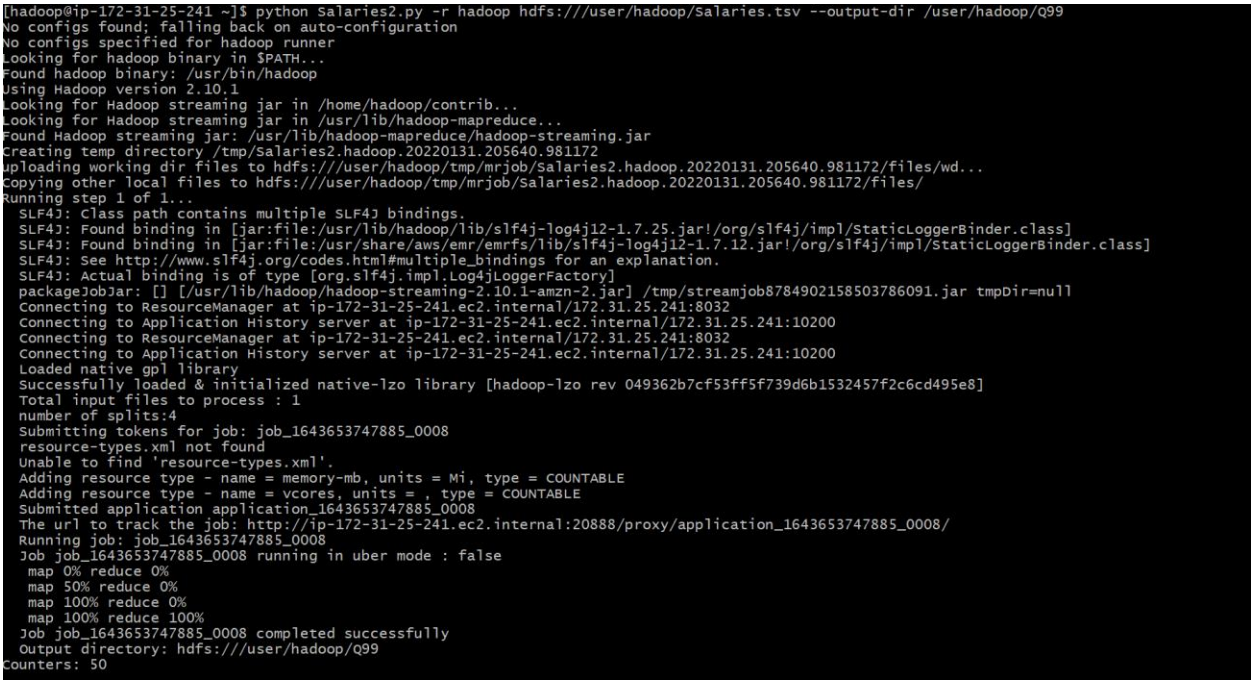
```
            yield "High", 1
        elif float(annualSalary) >= float(50000.00) and float(annualSalary) < float(100000.00):
            yield "Medium", 1
        else:
            yield "Low", 1

    def combiner(self, jobTitle, counts):
        yield jobTitle, sum(counts)

    def reducer(self, jobTitle, counts):
        yield jobTitle, sum(counts)


if __name__ == '__main__':
    MRSalaries.run()
```

Execution Snapshot:



```
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=1126587
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1564582
                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=61019136
                Total megabyte-milliseconds taken by all reduce tasks=13031424
                Total time spent by all map tasks (ms)=39726
                Total time spent by all maps in occupied slots (ms)=1906848
                Total time spent by all reduce tasks (ms)=4242
                Total time spent by all reduces in occupied slots (ms)=407232
                Total vcore-milliseconds taken by all map tasks=39726
                Total vcore-milliseconds taken by all reduce tasks=4242
        Map-Reduce Framework
                CPU time spent (ms)=5420
                Combine input records=13818
                Combine output records=12
                Failed Shuffles=0
                GC time elapsed (ms)=994
                Input split bytes=472
                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=2156347392
```

```
                Total vcore-milliseconds taken by all map tasks=39726
                Total vcore-milliseconds taken by all reduce tasks=4242
        Map-Reduce Framework
                CPU time spent (ms)=5420
                Combine input records=13818
                Combine output records=12
                Failed Shuffles=0
                GC time elapsed (ms)=994
                Input split bytes=472
                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=2156347392
                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)=1765801984
                Virtual memory (bytes) snapshot=17868263424
        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/Q99
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220131.205640.981172...
Removing temp directory /tmp/Salaries2.hadoop.20220131.205640.981172...
[hadoop@ip-172-31-25-241 ~]$ hdfs dfs -cat Q99/part-00000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
"High"  442
"Low"   7064
"Medium"        6312
```

Output:

```
[hadoop@ip-172-31-25-241 ~]$ hdfs dfs -cat Q99/part-00000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
"High"  442
"Low"   7064
"Medium"        6312
```

11) Now copy the file u.data from the assignment to /user/hadoop. This is similar to the file used for some examples in Module 03b. **NOTE: unlike the slide deck examples, this version of u.data has fields separated by commas and not tabs.**

12) (5 points) Review the slides 22-29 in lecture notes Module 3b. Now write a program to perform the task of outputting a count of the number of movies each user (identified via their user id) reviewed.

Movies.py

```
from mrjob.job import MRJob

class MRMoviesReviewed(MRJob):

    def mapper(self, _, line):
        (userID,movieID,rating,timestamp) = line.split(',')
        yield userID, movieID

    def combiner(self, userID, movies):
        yield userID, len(list(movies))

    def reducer(self, userID, count):
        yield userID, sum(count)


if __name__ == '__main__':
    MRMoviesReviewed.run()
```

Execution Snapshot:

```
[hadoop@ip-172-31-25-241 ~]$ python Movies.py -r hadoop hdfs:///user/hadoop/u.data --output-dir /user/hadoop/Q12
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/Movies.hadoop.20220131.211214.229917
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20220131.211214.229917/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20220131.211214.229917/files/
Running step 1 of 1...
  SLF4J: Class path contains multiple SLF4J bindings.
  SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
  SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
  SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
  SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-2.jar] /tmp/streamjob3099403433591341912.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-25-241.ec2.internal/172.31.25.241:8032
  Connecting to Application History server at ip-172-31-25-241.ec2.internal/172.31.25.241:10200
  Connecting to ResourceManager at ip-172-31-25-241.ec2.internal/172.31.25.241:8032
  Connecting to Application History server at ip-172-31-25-241.ec2.internal/172.31.25.241:10200
  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_1643653747885_0009
  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_1643653747885_0009
  The url to track the job: http://ip-172-31-25-241.ec2.internal:20888/proxy/application_1643653747885_0009/
  Running job: job_1643653747885_0009
  Job job_1643653747885_0009 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1643653747885_0009 completed successfully
  Output directory: hdfs:///user/hadoop/Q12
Counters: 50
        File Input Format Counters
                Bytes Read=2575317
        File Output Format Counters
                Bytes Written=6204
```

```
Counters: 50
        File Input Format Counters
                Bytes Read=2575317
        File Output Format Counters
                Bytes Written=6204
        File System Counters
                FILE: Number of bytes read=4636
                FILE: Number of bytes written=1135592
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2575765
                HDFS: Number of bytes written=6204
                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=72399360
                Total megabyte-milliseconds taken by all reduce tasks=15925248
                Total time spent by all map tasks (ms)=47135
                Total time spent by all maps in occupied slots (ms)=2262480
                Total time spent by all reduce tasks (ms)=5184
                Total time spent by all reduces in occupied slots (ms)=497664
                Total vcore-milliseconds taken by all map tasks=47135
                Total vcore-milliseconds taken by all reduce tasks=5184
        Map-Reduce Framework
                CPU time spent (ms)=8890
                Combine input records=100004
                Combine output records=674
                Failed Shuffles=0
                GC time elapsed (ms)=996
                Input split bytes=448
                Map input records=100004
                Map output bytes=1273035
                Map output materialized bytes=4956
                Map output records=100004
                Merged Map outputs=4
                Physical memory (bytes) snapshot=2079760384
                Reduce input groups=671
                Reduce input records=674
                Reduce output records=671
                Reduce shuffle bytes=4956
                Shuffled Maps =4
```

```
        Map-Reduce Framework
                CPU time spent (ms)=8890
                Combine input records=100004
                Combine output records=674
                Failed Shuffles=0
                GC time elapsed (ms)=996
                Input split bytes=448
                Map input records=100004
                Map output bytes=1273035
                Map output materialized bytes=4956
                Map output records=100004
                Merged Map outputs=4
                Physical memory (bytes) snapshot=2079760384
                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)=1736441856
                Virtual memory (bytes) snapshot=17838104576
        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/Q12
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20220131.211214.229917...
Removing temp directory /tmp/Movies.hadoop.20220131.211214.229917...
```

Output:

```
[hadoop@ip-172-31-25-241 ~]$ hdfs dfs -cat Q12/part-00000 | head
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
"1"     20
"10"    46
"100"   25
"101"   55
"102"   678
"103"   94
"104"   76
"105"   525
"106"   45
"107"   32
[hadoop@ip-172-31-25-241 ~]$
```