Name – Rishabh Jain
Mail – rjain35@hawk.iit.edu
Course – CSP 554

Assignment – 3
CSP – 554

1. ssh to the master node.

```
➜  ~ ssh -i ~/Desktop/new-key-pair-emr.pem hadoop@ec2-44-197-114-95.compute-1.amazonaws.com
The authenticity of host 'ec2-44-197-114-95.compute-1.amazonaws.com (44.197.114.95)' can't be established.
ED25519 key fingerprint is SHA256:bec0Z1x//bI88ATzFox1164Blb83lFeoEPvM5hHF3nY.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-44-197-114-95.compute-1.amazonaws.com' (ED25519) to the list of known hosts.

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
21 package(s) needed for security, out of 26 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR:::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-65-168 ~]$ ls
```

2. Moving WordCount.py and w.data to /home/hadoop

```
➜  ~ scp -i /Users/rishabhjain/Desktop/new-key-pair-emr.pem /Users/rishabhjain/Desktop/W.data hadoop@ec2-44-197-114-95.compute-1.amazonaws.com:/home/hadoop
W.data                                        100%  528    12.3KB/s   00:00
```

```
➜  ~ scp -i /Users/rishabhjain/Desktop/new-key-pair-emr.pem /Users/rishabhjain/Desktop/WordCount.py hadoop@ec2-44-197-114-95.compute-1.amazonaws.com:/home/hadoop
WordCount.py                                  100%  402    11.4KB/s   00:00
```

3. Moving W.data to /user/hadoop

```
[hadoop@ip-172-31-65-168 ~]$ hadoop fs -copyFromLocal /home/hadoop/W.data /user/hadoop/W.data
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]
[hadoop@ip-172-31-65-168 ~]$ hadoop fs -ls /user/hadoop/
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]
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup        528 2022-01-30 15:42 /user/hadoop/W.data
```

Name – Rishabh Jain
Mail – rjain35@hawk.iit.edu
Course – CSP 554

4. Executing the first Job from unmodified WordCount.py code

```
[hadoop@ip-172-31-65-168 ~]$ python WordCount.py -r hadoop hdfs:///user/hadoop/W.data --output-dir /user/hadoop/dout
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/WordCount.hadoop.20220130.154528.747500
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20220130.154528.747500/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20220130.154528.747500/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/streamjob7548467840542247310.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-65-168.ec2.internal/172.31.65.168:8032
  Connecting to Application History server at ip-172-31-65-168.ec2.internal/172.31.65.168:10200
  Connecting to ResourceManager at ip-172-31-65-168.ec2.internal/172.31.65.168:8032
  Connecting to Application History server at ip-172-31-65-168.ec2.internal/172.31.65.168:10200
  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_1643557320946_0001
  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_1643557320946_0001
  The url to track the job: http://ip-172-31-65-168.ec2.internal:20888/proxy/application_1643557320946_0001/
  Running job: job_1643557320946_0001
  Job job_1643557320946_0001 running in uber mode : false
```

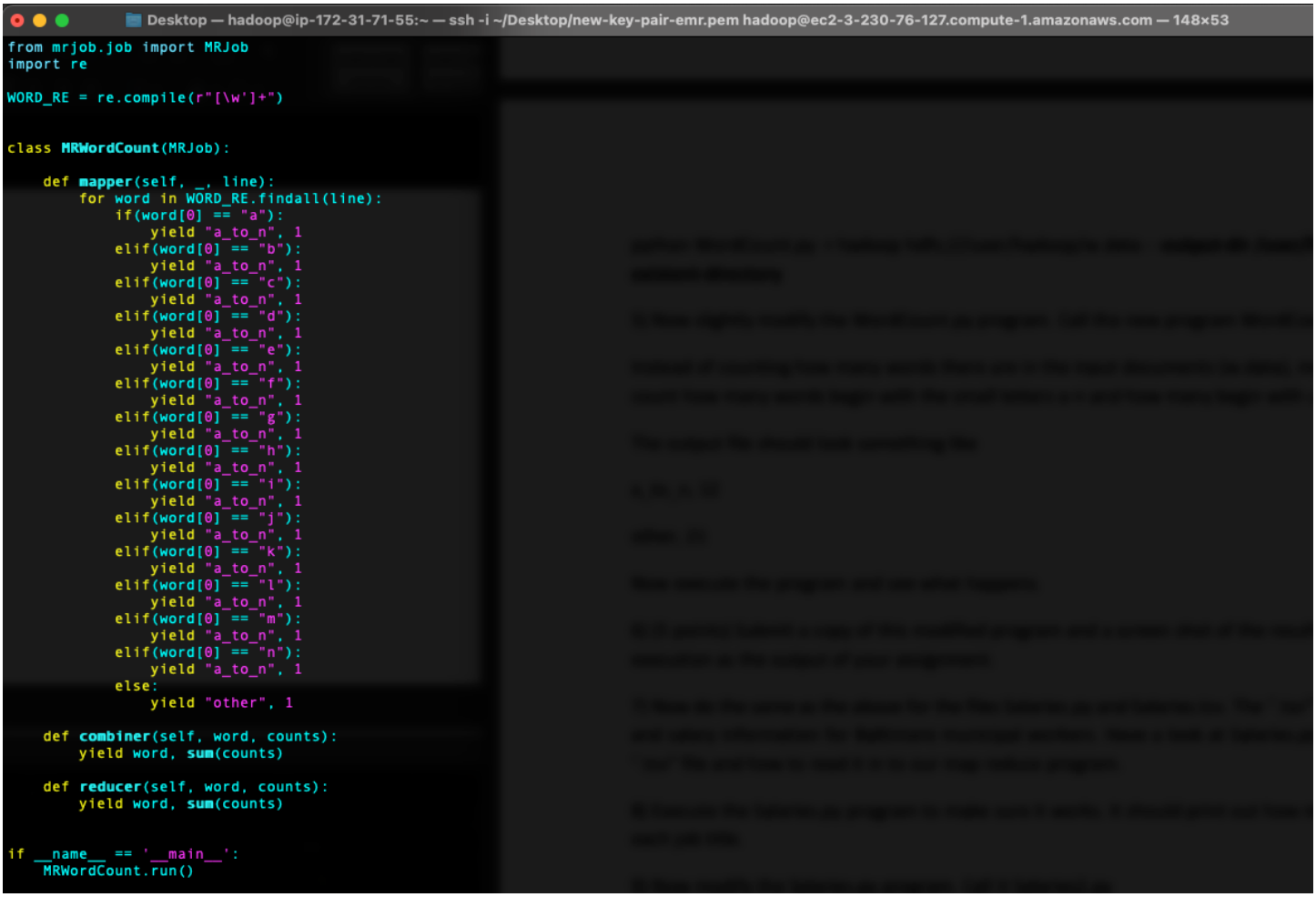
5. Output file in /user/hadoop/dout from unmodified WordCount.py code

```
[hadoop@ip-172-31-65-168 ~]$ hadoop fs -cat /user/hadoop/dout/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"     3
"all"   1
"an"    1
"and"   1
"are"   1
"as"    4
"available"     1
"be"    3
"by"    1
"cluster"       2
"combine"       1
"contained"     1
"defined"       1
"dependencies"  1
"do"    1
"either"        1
"executed"      1
"explains"      1
"file"  2
"first" 1
"following"     1
"for"   1
"hadoop"        1
"how"   2
"in"    1
"individual"    1
"is"    2
"job"   4
"machine"       1
"map"   1
"more"  2
"mrjob" 1
"must"  1
"nodes" 1
"of"    1
"on"    4
"or"    2
"oriented"      1
"our"   1
"program"       1
"python"        1
"reduce"        1
```

Name – Rishabh Jain
Mail – rjain35@hawk.iit.edu
Course – CSP 554

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.

➔ Screen shot of modified WordCount.py code further saved to new file named "**WordCount2.py**" for finding **words staring from a to n and other words.**




```
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] == "a"):
                yield "a_to_n", 1
            elif(word[0] == "b"):
                yield "a_to_n", 1
            elif(word[0] == "c"):
                yield "a_to_n", 1
            elif(word[0] == "d"):
                yield "a_to_n", 1
            elif(word[0] == "e"):
                yield "a_to_n", 1
            elif(word[0] == "f"):
                yield "a_to_n", 1
            elif(word[0] == "g"):
                yield "a_to_n", 1
            elif(word[0] == "h"):
                yield "a_to_n", 1
            elif(word[0] == "i"):
                yield "a_to_n", 1
            elif(word[0] == "j"):
                yield "a_to_n", 1
            elif(word[0] == "k"):
                yield "a_to_n", 1
            elif(word[0] == "l"):
                yield "a_to_n", 1
            elif(word[0] == "m"):
                yield "a_to_n", 1
            elif(word[0] == "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()
```

➔ Executing the new modified code "**WordCount2.py**"

```
[hadoop@ip-172-31-71-55 ~]$ python WordCount2.py -r hadoop hdfs:///user/hadoop/W.data --output-dir /user/hadoop/WordCount2
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.20220130.221820.340299
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220130.221820.340299/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220130.221820.340299/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/streamjob2481673610621627866.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-71-55.ec2.internal/172.31.71.55:8032
  Connecting to Application History server at ip-172-31-71-55.ec2.internal/172.31.71.55:10200
  Connecting to ResourceManager at ip-172-31-71-55.ec2.internal/172.31.71.55:8032
```

Name – Rishabh Jain
Mail – rjain35@hawk.iit.edu
Course – CSP 554

➔ Output [a_to_n = 46 and other = 49]

```
[hadoop@ip-172-31-71-55 ~]$ hadoop fs -cat /user/hadoop/WordCount2/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"        46
"other" 49
[hadoop@ip-172-31-71-55 ~]$
```

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.

➔ Screen shot of modified Salaries.py code further saved to new file named "**Salaries2.py**" for finding **the number of workers having High, Medium or Low annual salaries.**

Desktop — hadoop@ip-172-31-71-55:~ — ssh -i ~/Desktop/new-key-pair-emr.pem hadoop@ec2-3-230-76-127.compute-1.amazonaws.com — 132×47

```
from mrjob.job import MRJob

class MRSalaries(MRJob):

    def mapper(self, _, line):
        (name,jobTitle,agencyID,agency,hireDate,annualSalary,grossPay) = line.split('\t')
        salary = float(annualSalary)
        if(salary >= 100000.00):
                yield 'High', 1
        elif(salary >= 50000.00):
                yield 'Medium', 1
        elif(salary >= 0):
                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()
```

➔ Executing the new modified code "**Salaries2.py**"

```
[hadoop@ip-172-31-71-55 ~]$ python Salaries2.py -r hadoop hdfs:///user/hadoop/Salaries.tsv --output-dir /user/hadoop/Salaries2
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/Salaries2.hadoop.20220130.225210.335620
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220130.225210.335620/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220130.225210.335620/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/streamjob1591891980741433167.jar tmpDir=null
```

➔ Output

```
[hadoop@ip-172-31-71-55 ~]$ hadoop fs -cat /user/hadoop/Salaries2/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
```

Name – Rishabh Jain
Mail – rjain35@hawk.iit.edu
Course – CSP 554

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.

➔ Screen shot of code named as **"ratemovie.py"**

Desktop — hadoop@ip-172-31-71-55:~ — ssh -i ~/Desktop/new-key-pair-emr.pem hadoop@ec2-3-230-76-127.compute-1.amazonaws.com — 132×47

```
from mrjob.job import MRJob

class MRRatings(MRJob):

    def mapper(self, _, line):
        (userid, movieid, rating, timestamp) = line.split(',')
        yield userid, 1

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

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


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

➔ Executing the code "**ratemovie.py**"

```
[hadoop@ip-172-31-71-55 ~]$ python ratemovie.py -r hadoop hdfs:///user/hadoop/u.data --output-dir /user/hadoop/ratemovie
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/ratemovie.hadoop.20220130.231709.427464
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/ratemovie.hadoop.20220130.231709.427464/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/ratemovie.hadoop.20220130.231709.427464/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/streamjob9324347129704657l3.jar tmpDir=null
```

➔ Output

```
[hadoop@ip-172-31-71-55 ~]$ hadoop fs -cat /user/hadoop/ratemovie/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]
"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
```