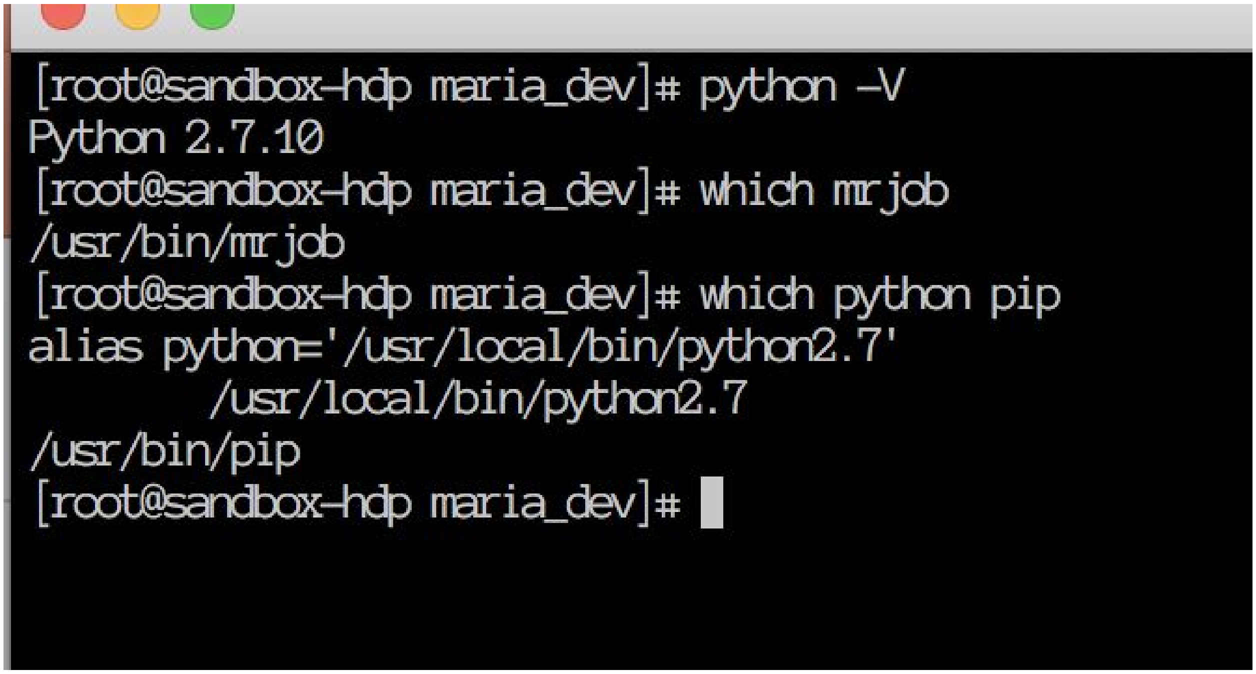
CSP554—Big Data Technologies

Saptarshi Chatterjee (A20413922)

Assignment #3

4) Install the python mrjob library on your Hadoop sandbox.

- Log on to the maria_dev account
- Enter “su root”
    - You will be asked for the root password, enter the word: hadoop
    - You will then be asked again for this password, and finally asked to supply a new root password, which you should remember. You will need to have this password for future assignments.
    - Now you need to make a small change to the following file /etc/yum.repos.d/sandbox.repo
    - Use the vi editor to open this file by entering ‘vi /etc/yum.repos.d/sandbox.repo’
    - If you don’t know how to use the vi editor, use google to find a tutorial
    - Then change the line ‘enabled=1’ to ‘enabled=0’
    - Save the file using the ‘wq’ command
- Enter “yum install python-pip”
- Enter “yum install nano”
- Enter “pip install mrjob==0.5.11”
- Enter “exit”



```
[root@sandbox-hdp maria_dev]# python -V
Python 2.7.10
[root@sandbox-hdp maria_dev]# which mrjob
/usr/bin/mrjob
[root@sandbox-hdp maria_dev]# which python pip
alias python='/usr/local/bin/python2.7'
        /usr/local/bin/python2.7
/usr/bin/pip
[root@sandbox-hdp maria_dev]# 
```

4) Next you will set up to execute the provided WordCount mapreduce program found in the “Assignments” section of the Blackboard. This is the exact same program we saw in class.

Step 1:

Copy the two files “cs595words.txt” and “WordCount.py” to your PC or Mac. They are part of the documents included with the assignment.

Step 2:

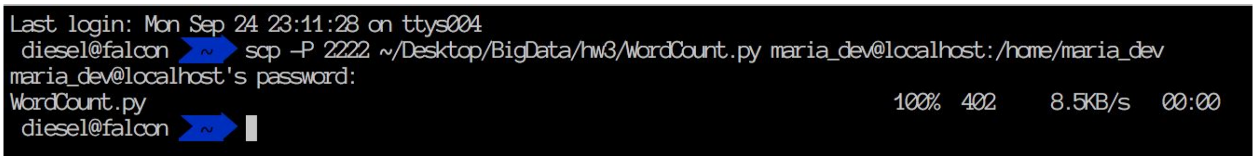
Log on to your Hadoop environment and use the secure copy (scp) program to move the WordCount.py and cs595words.txt files to the home directory of the maria_dev account which should be “/home/maria_dev”

If we assume that you have downloaded the WordCount.py file to /my/dir/WordCount.py on your mac or pc the command would be something like

scp –P 2222 /my/mydir/WordCount.py maria_dev@localhost:/home/maria_dev

Note that you need to use a capital “-P”.:

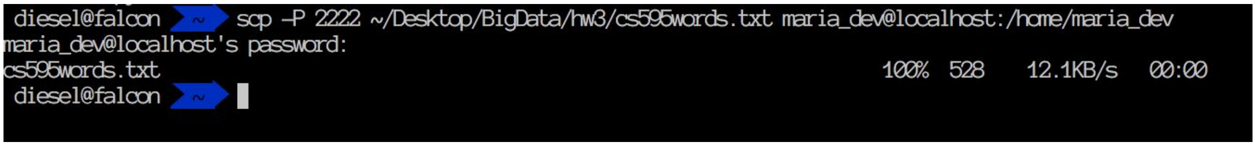
Step 4



```
Last login: Mon Sep 24 23:11:28 on ttys004
 diesel@falcon  ~  scp -P 2222 ~/Desktop/BigData/hw3/WordCount.py maria_dev@localhost:/home/maria_dev
maria_dev@localhost's password:
WordCount.py                                                    100%  402     8.5KB/s   00:00
 diesel@falcon  ~  
```

:

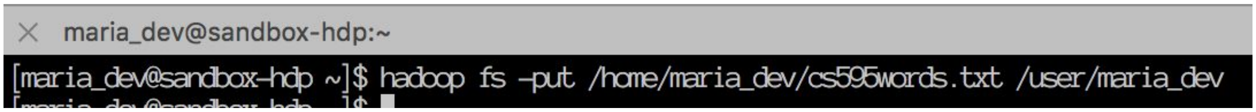
Do the same for the assignment file cs595words.txt



```
 diesel@falcon  ~  scp -P 2222 ~/Desktop/BigData/hw3/cs595words.txt maria_dev@localhost:/home/maria_dev
maria_dev@localhost's password:
cs595words.txt                                                  100%  528    12.1KB/s   00:00
 diesel@falcon  ~  
```

In this case move the file from “/home/maria_dev” to the Hadoop file system, say to the directory “/user/maria_dev”

Step

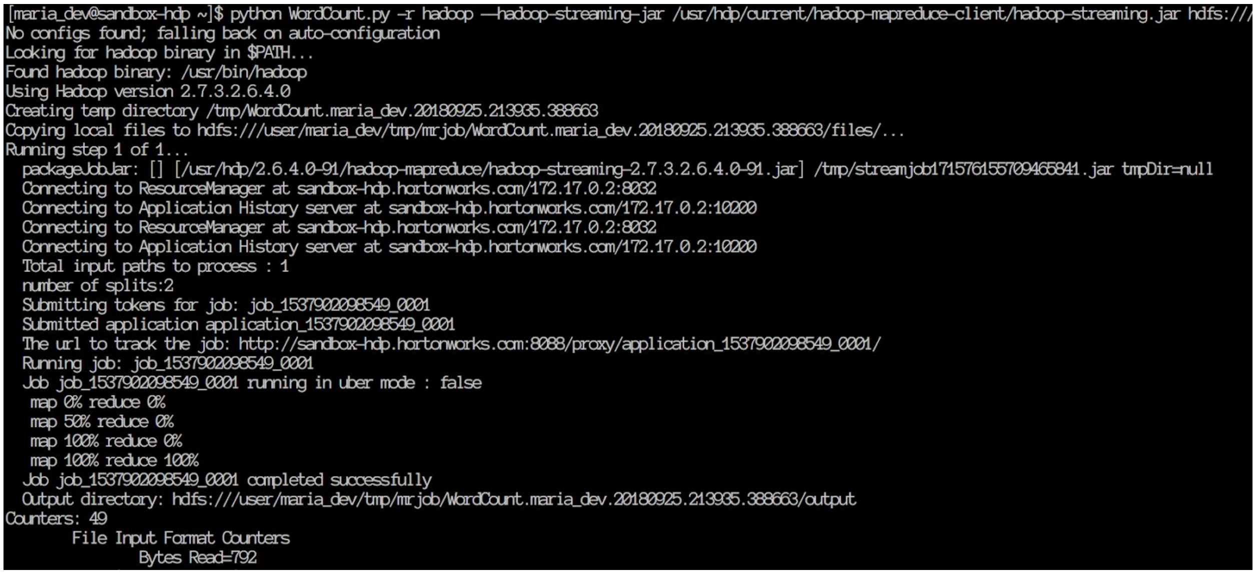


```
[maria_dev@sandbox-hdp ~]$ hadoop fs -put /home/maria_dev/cs595words.txt /user/maria_dev
[maria_dev@sandbox-hdp ~]$ 
```

5:

Now execute the following

python WordCount.py -r hadoop --hadoop-streaming-jar
/usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar
hdfs:///user/maria_dev/cs595words.txt



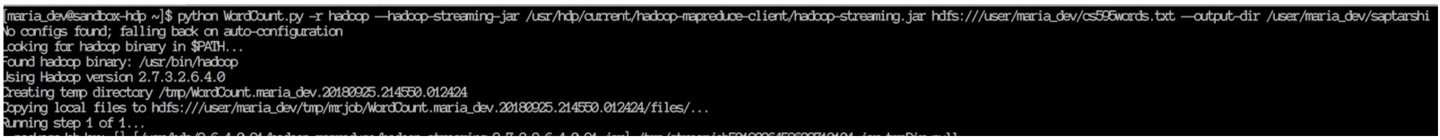
```
[maria_dev@sandbox-hdp ~]$ python WordCount.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar hdfs:///
No configs found; falling back on auto-configuration
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.7.3.2.6.4.0
Creating temp directory /tmp/WordCount.maria_dev.20180925.213935.388663
Copying local files to hdfs:///user/maria_dev/tmp/mrjob/WordCount.maria_dev.20180925.213935.388663/files/...
Running step 1 of 1...
  packageJobJar: [] [/usr/hdp/2.6.4.0-91/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.4.0-91.jar] /tmp/streamjob171576155709465841.jar tmpDir=null
  Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.17.0.2:8032
  Connecting to Application History server at sandbox-hdp.hortonworks.com/172.17.0.2:10200
  Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.17.0.2:8032
  Connecting to Application History server at sandbox-hdp.hortonworks.com/172.17.0.2:10200
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1537902098549_0001
  Submitted application application_1537902098549_0001
  The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1537902098549_0001/
  Running job: job_1537902098549_0001
  Job job_1537902098549_0001 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1537902098549_0001 completed successfully
  Output directory: hdfs:///user/maria_dev/tmp/mrjob/WordCount.maria_dev.20180925.213935.388663/output
Counters: 49
        File Input Format Counters
                Bytes Read=792
```

Note there must be three slashes in “hdfs:///” as “hdfs://” indicates that the file you are reading from is in the hadoop file system and the “/user” is the first part of the path to that file. Also note that sometimes copying and pasting this command from the assignment document does not work and it needs to be entered manually.

Check that it produces some reasonable output.

Note, the above command will erase all output files in hdfs. If you want to keep the output use the following command instead:

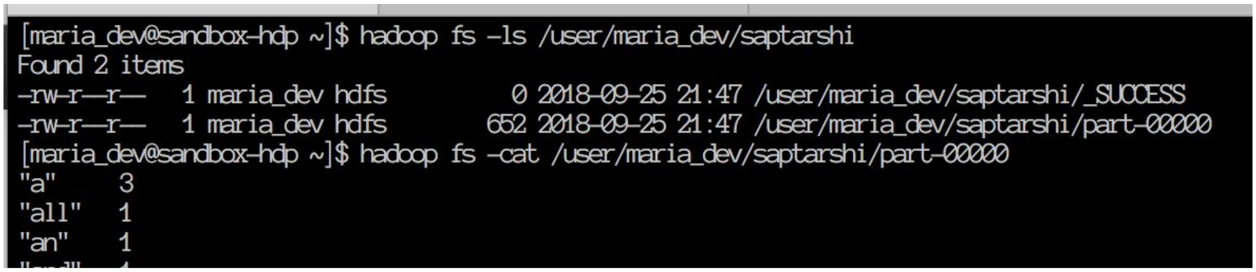
python WordCount.py -r hadoop --hadoop-streaming-jar
/usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar
hdfs:///user/maria_dev/cs595words.txt **- -output-dir /user/maria_dev/some-non-existent-directory**



```
[maria_dev@sandbox-hdp ~]$ python WordCount.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar hdfs:///user/maria_dev/cs595words.txt --output-dir /user/maria_dev/saptarshi
o configs found; falling back on auto-configuration
ooking for hadoop binary in $PATH...
ound hadoop binary: /usr/bin/hadoop
sing Hadoop version 2.7.3.2.6.4.0
reating temp directory /tmp/WordCount.maria_dev.20180925.214550.012424
opying local files to hdfs:///user/maria_dev/tmp/mrjob/WordCount.maria_dev.20180925.214550.012424/files/...
unning step 1 of 1...
```

To see which all file got created - hadoop fs -ls /user/maria_dev/saptarshi

To see content of the file - hadoop fs -cat /user/maria_dev/saptarshi/part-00000



```
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls /user/maria_dev/saptarshi
Found 2 items
-rw-r--r--   1 maria_dev hdfs          0 2018-09-25 21:47 /user/maria_dev/saptarshi/_SUCCESS
-rw-r--r--   1 maria_dev hdfs        652 2018-09-25 21:47 /user/maria_dev/saptarshi/part-00000
[maria_dev@sandbox-hdp ~]$ hadoop fs -cat /user/maria_dev/saptarshi/part-00000
"a"     3
"all"   1
"an"    1
"and"   1
```

5) Now slightly modify the WordCount program. Call the new program WordCount2.py.

Instead of counting how many words there are in the input documents, modify the program to count how many words begin with the small letters a-n and how many begin with anything else.

The output file should look something like

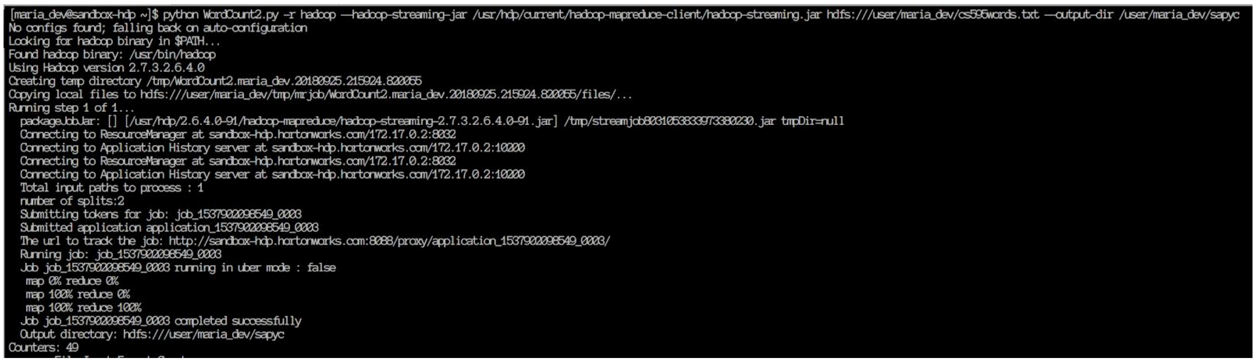
a_to_n, 12

other, 21

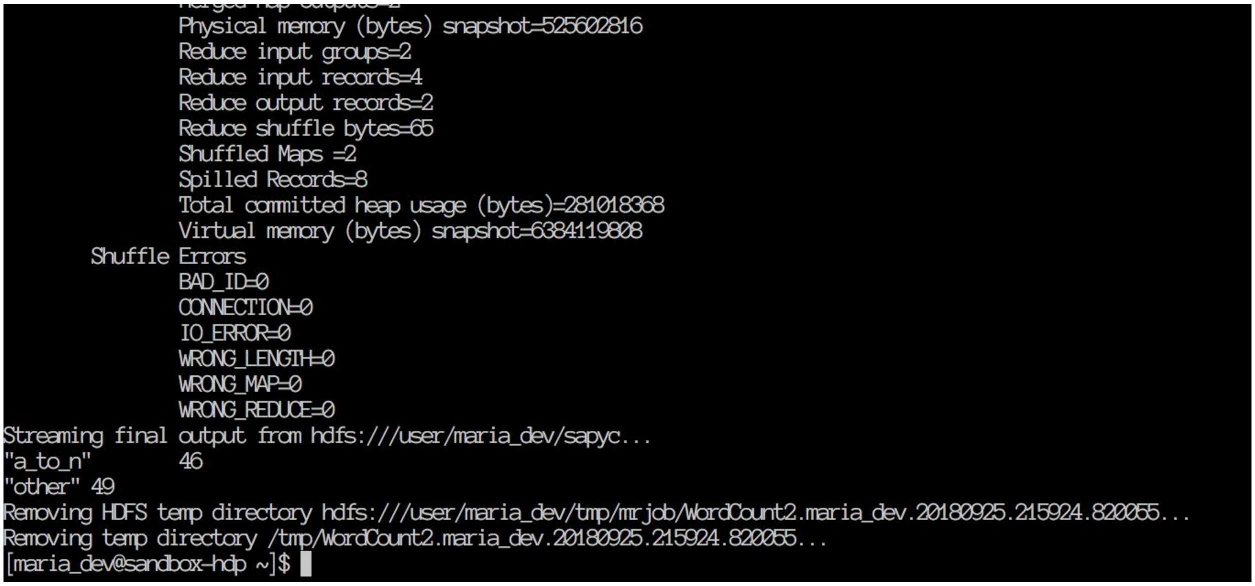
Now execute the program and see what happens.

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.

Command - *python WordCount2.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar hdfs:///user/maria_dev/cs595words.txt --output-dir /user/maria_dev/sapyc*

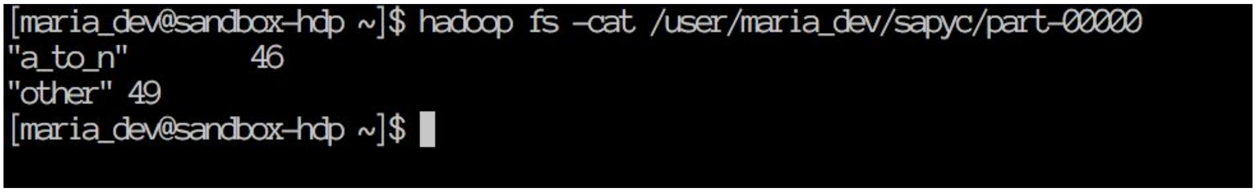


```
[maria_dev@sandbox-hdp ~]$ python WordCount2.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar hdfs:///user/maria_dev/cs595words.txt --output-dir /user/maria_dev/sapyc
No configs found; falling back on auto-configuration
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.7.3.2.6.4.0
Creating temp directory /tmp/WordCount2.maria_dev.20180925.215924.820055
Copying local files to hdfs:///user/maria_dev/tmp/mrjob/WordCount2.maria_dev.20180925.215924.820055/files/...
Running step 1 of 1...
  packageJobJar: [] [/usr/hdp/2.6.4.0-91/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.4.0-91.jar] /tmp/streamjob8031053833973380230.jar tmpDir=null
  Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.17.0.2:8032
  Connecting to Application History server at sandbox-hdp.hortonworks.com/172.17.0.2:10200
  Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.17.0.2:8032
  Connecting to Application History server at sandbox-hdp.hortonworks.com/172.17.0.2:10200
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1537902098549_0003
  Submitted application application_1537902098549_0003
  The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1537902098549_0003/
  Running job: job_1537902098549_0003
  Job job_1537902098549_0003 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1537902098549_0003 completed successfully
  Output directory: hdfs:///user/maria_dev/sapyc
Counters: 49
```



```
                Physical memory (bytes) snapshot=525602816
                Reduce input groups=2
                Reduce input records=4
                Reduce output records=2
                Reduce shuffle bytes=65
                Shuffled Maps =2
                Spilled Records=8
                Total committed heap usage (bytes)=281018368
                Virtual memory (bytes) snapshot=6384119808
        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/sapyc...
"a_to_n"        46
"other" 49
Removing HDFS temp directory hdfs:///user/maria_dev/tmp/mrjob/WordCount2.maria_dev.20180925.215924.820055...
Removing temp directory /tmp/WordCount2.maria_dev.20180925.215924.820055...
[maria_dev@sandbox-hdp ~]$
```

To see file content - hadoop fs -cat /user/maria_dev/sapyc/part-00000

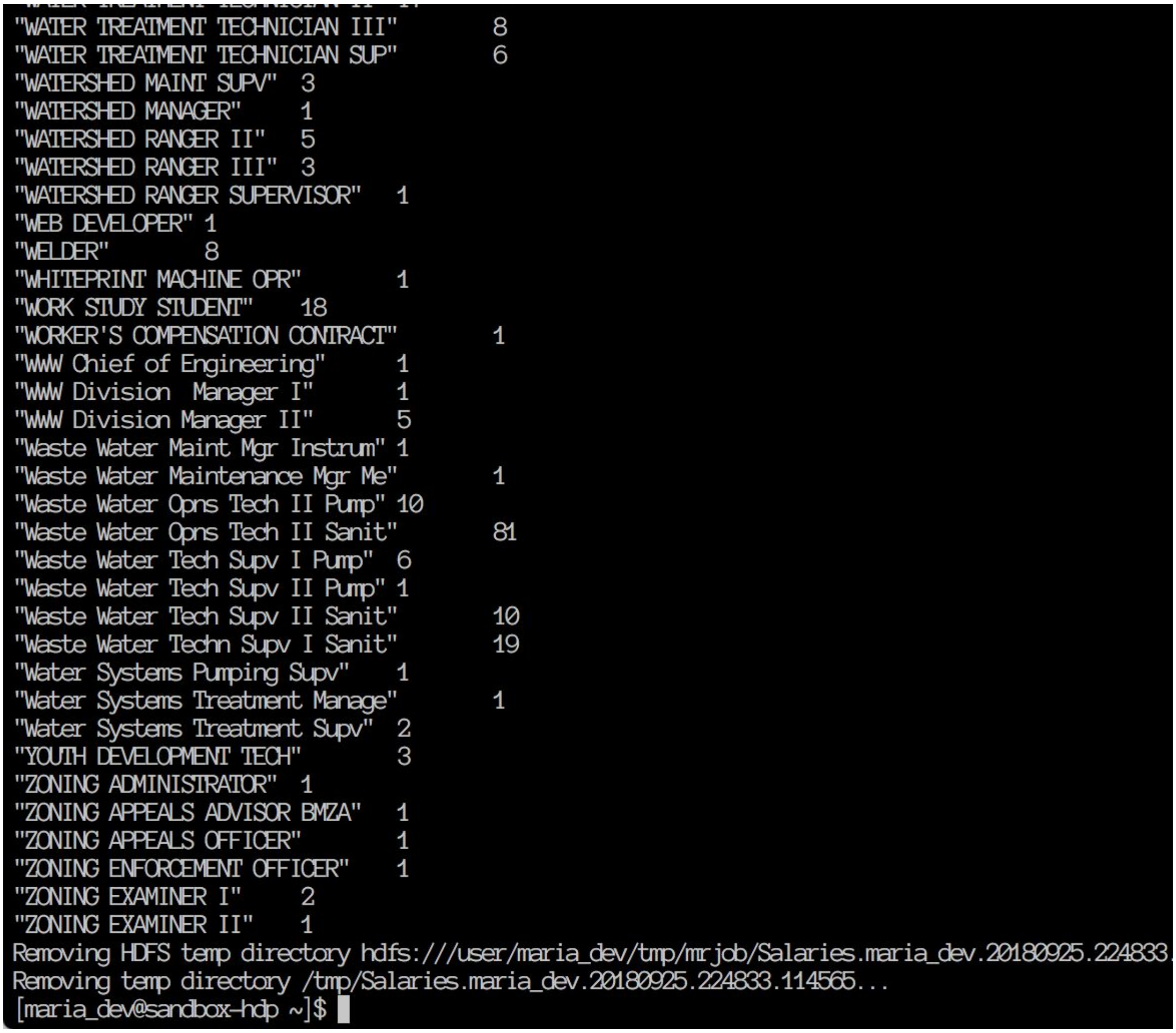


```
[maria_dev@sandbox-hdp ~]$ hadoop fs -cat /user/maria_dev/sapyc/part-00000
"a_to_n"        46
"other" 49
[maria_dev@sandbox-hdp ~]$
```

7) Now do the same as the above for the files Salaries.py and Salaries.tsv. The ".tsv" file holds department and salary information for Baltimore municipal workers. Have a look at Salaries.py for the layout of the ".tsv" file and how to read it in to our map reduce program.

8) Execute the Salaries.py program to make sure it works. It should print out how many workers share each job title.

python Salaries.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar hdfs:///user/maria_dev/Salaries.tsv --output-dir /user/maria_dev/sap_salaries



```
"WATER TREATMENT TECHNICIAN III"        8
"WATER TREATMENT TECHNICIAN SUP"        6
"WATERSHED MAINT SUPV"  3
"WATERSHED MANAGER"     1
"WATERSHED RANGER II"   5
"WATERSHED RANGER III"  3
"WATERSHED RANGER SUPERVISOR"   1
"WEB DEVELOPER" 1
"WELDER"        8
"WHITEPRINT MACHINE OPR"        1
"WORK STUDY STUDENT"    18
"WORKER'S COMPENSATION CONTRACT"        1
"WWW Chief of Engineering"      1
"WWW Division  Manager I"       1
"WWW Division Manager II"       5
"Waste Water Maint Mgr Instrum" 1
"Waste Water Maintenance Mgr Me"        1
"Waste Water Opns Tech II Pump" 10
"Waste Water Opns Tech II Sanit"        81
"Waste Water Tech Supv I Pump"  6
"Waste Water Tech Supv II Pump" 1
"Waste Water Tech Supv II Sanit"        10
"Waste Water Techn Supv I Sanit"        19
"Water Systems Pumping Supv"    1
"Water Systems Treatment Manage"        1
"Water Systems Treatment Supv"  2
"YOUTH DEVELOPMENT TECH"        3
"ZONING ADMINISTRATOR"  1
"ZONING APPEALS ADVISOR BMZA"   1
"ZONING APPEALS OFFICER"        1
"ZONING ENFORCEMENT OFFICER"    1
"ZONING EXAMINER I"     2
"ZONING EXAMINER II"    1
Removing HDFS temp directory hdfs:///user/maria_dev/tmp/mrjob/Salaries.maria_dev.20180925.224833.
Removing temp directory /tmp/Salaries.maria_dev.20180925.224833.114565...
[maria_dev@sandbox-hdp ~]$ █
```

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

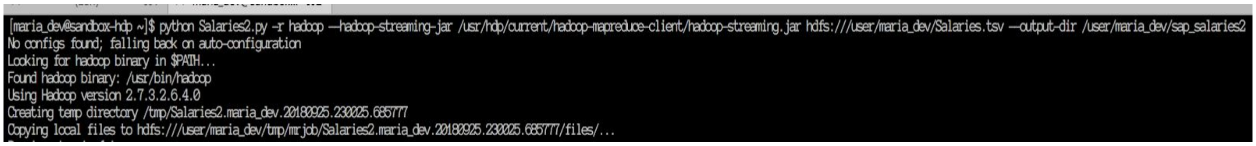
Low 10

Some important hints:

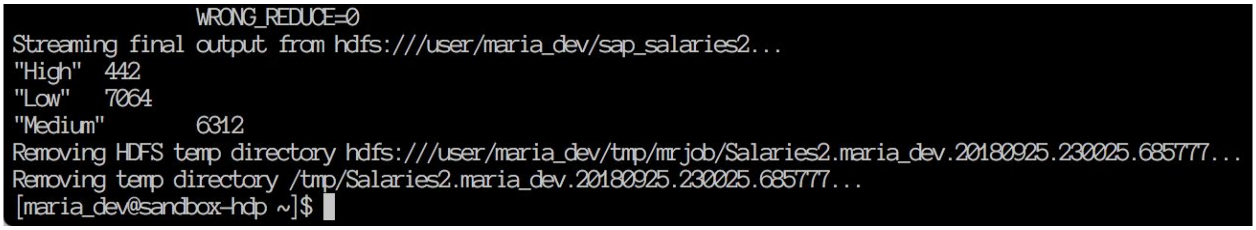
- The annual salary is a string that will need to be converted to a float.
- The mapper should output tuples with one of three keys depending on the annual salary: High, Medium and Low
- The value part of the tuple is not a salary. (What should it be?)

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.



```
[maria_dev@sandbox-hdp ~]$ python Salaries2.py -r hadoop —hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar hdfs:///user/maria_dev/Salaries.tsv —output-dir /user/maria_dev/sap_salaries2
No configs found; falling back on auto-configuration
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.7.3.2.6.4.0
Creating temp directory /tmp/Salaries2.maria_dev.20180925.230025.685777
Copying local files to hdfs:///user/maria_dev/tmp/mrjob/Salaries2.maria_dev.20180925.230025.685777/files/...
```



```
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/maria_dev/sap_salaries2...
"High"  442
"Low"   7064
"Medium"        6312
Removing HDFS temp directory hdfs:///user/maria_dev/tmp/mrjob/Salaries2.maria_dev.20180925.230025.685777...
Removing temp directory /tmp/Salaries2.maria_dev.20180925.230025.685777...
[maria_dev@sandbox-hdp ~]$
```

python Salaries2.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar hdfs:///user/maria_dev/Salaries.tsv --output-dir /user/maria_dev/sap_salaries2

11) Now copy the file u.data to /user/maria_dev. 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 17-22 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.

Output might look something like the following:

186: 2

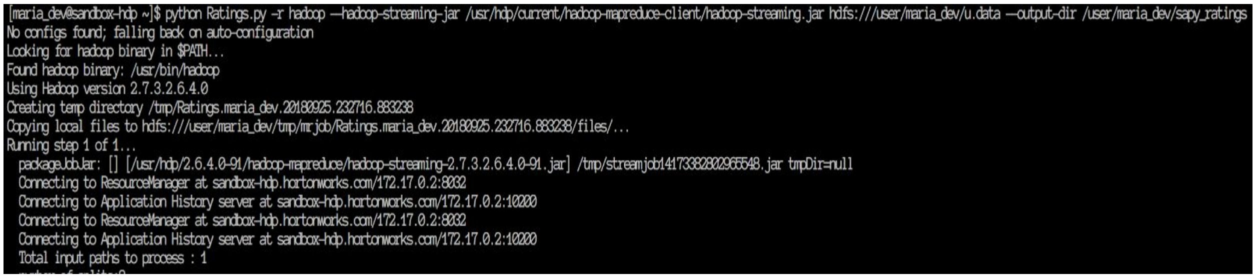
192: 2

112: 1

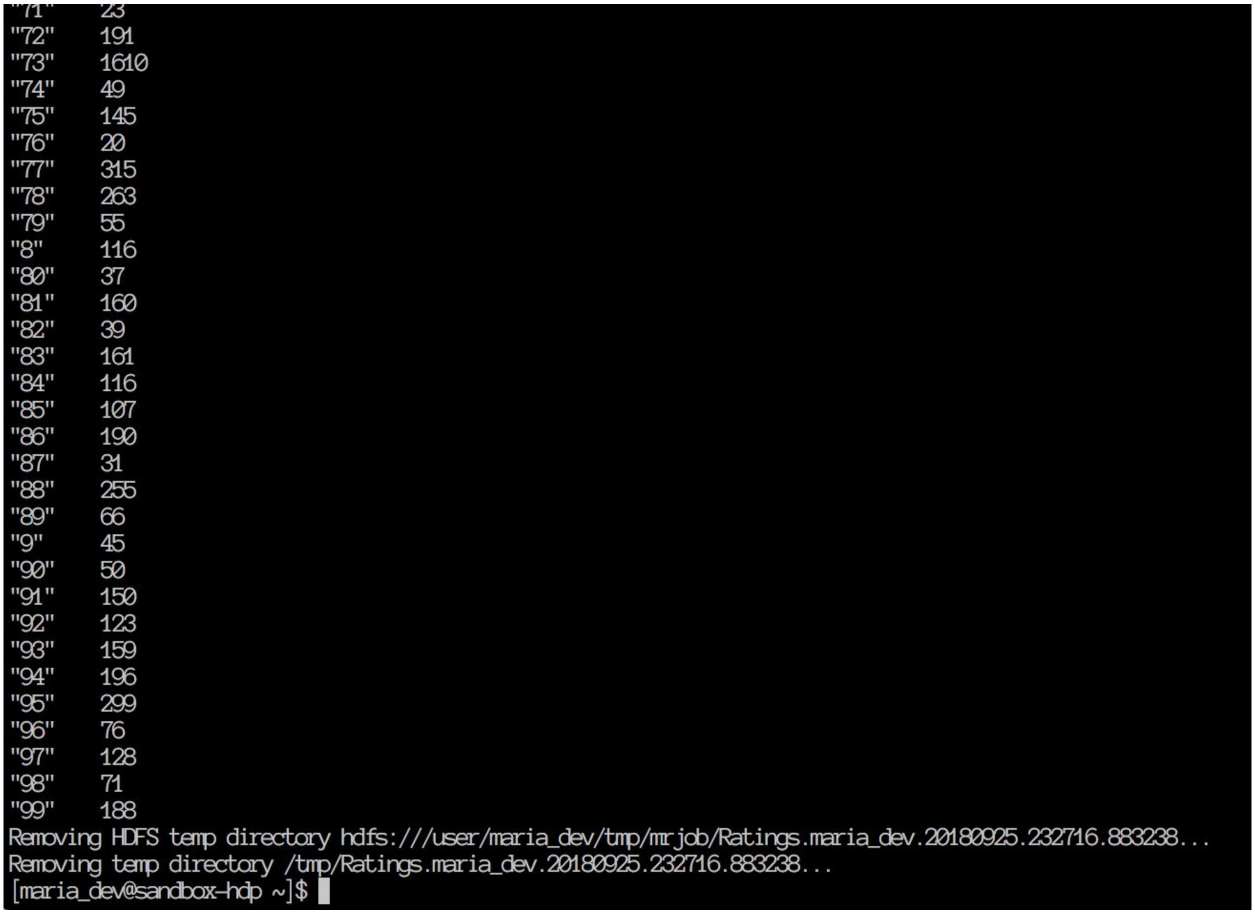
etc.

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.

Command - python Ratings.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar hdfs:///user/maria_dev/u.data --output-dir /user/maria_dev/ratings

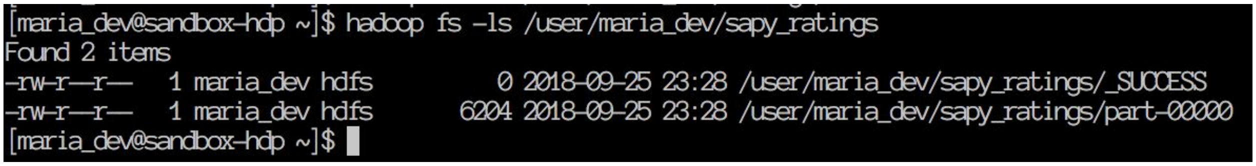


```
[maria_dev@sandbox-hdp ~]$ python Ratings.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar hdfs:///user/maria_dev/u.data --output-dir /user/maria_dev/sapy_ratings
No configs found; falling back on auto-configuration
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.7.3.2.6.4.0
Creating temp directory /tmp/Ratings.maria_dev.20180925.232716.883238
Copying local files to hdfs:///user/maria_dev/tmp/mrjob/Ratings.maria_dev.20180925.232716.883238/files/...
Running step 1 of 1...
  packageJobJar: [] [/usr/hdp/2.6.4.0-91/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.4.0-91.jar] /tmp/streamjob1417338280296548.jar tmpDir=null
  Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.17.0.2:8032
  Connecting to Application History server at sandbox-hdp.hortonworks.com/172.17.0.2:10200
  Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.17.0.2:8032
  Connecting to Application History server at sandbox-hdp.hortonworks.com/172.17.0.2:10200
  Total input paths to process : 1
```



```
"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/Ratings.maria_dev.20180925.232716.883238...
Removing temp directory /tmp/Ratings.maria_dev.20180925.232716.883238...
[maria_dev@sandbox-hdp ~]$
```

hadoop fs -ls /user/maria_dev/sapy_ratings



```
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls /user/maria_dev/sapy_ratings
Found 2 items
-rw-r--r--   1 maria_dev hdfs          0 2018-09-25 23:28 /user/maria_dev/sapy_ratings/_SUCCESS
-rw-r--r--   1 maria_dev hdfs       6204 2018-09-25 23:28 /user/maria_dev/sapy_ratings/part-00000
[maria_dev@sandbox-hdp ~]$ 
```