Management of Big Data and Tools – DS8003 – Fall 2016

Assignment 1

NAJLIS, BERNARDO - Student Number #500744793

Submit the Python codes, screenshots that show the execution, and output files generated (use getmerge to convert it into a single file before submission).

1. File: dept_salary.txt. The first column has the id of a dept and the second column a salary amount. Write a map-reduce program to get the minimum salary per department. (5)

2. File: shakespeare_100.txt. In the Word count example we output all words with their count. What if I only wanted the words with top 10 counts (This is very common problem in the industry). Think about using multiple map-reduce steps (5)

Optional:

File: dept_course.txt. The first column is studentID, the second column is Dept, and the third column is Course. Write a map-reduce job to identify students who belong to "Math" course or "QA" department.

Resolution

1. For this item, I used the code base of the Python average calculation, with slight modifications to the reducer to keep the minimum salary per department.

   Mapper Python Code:

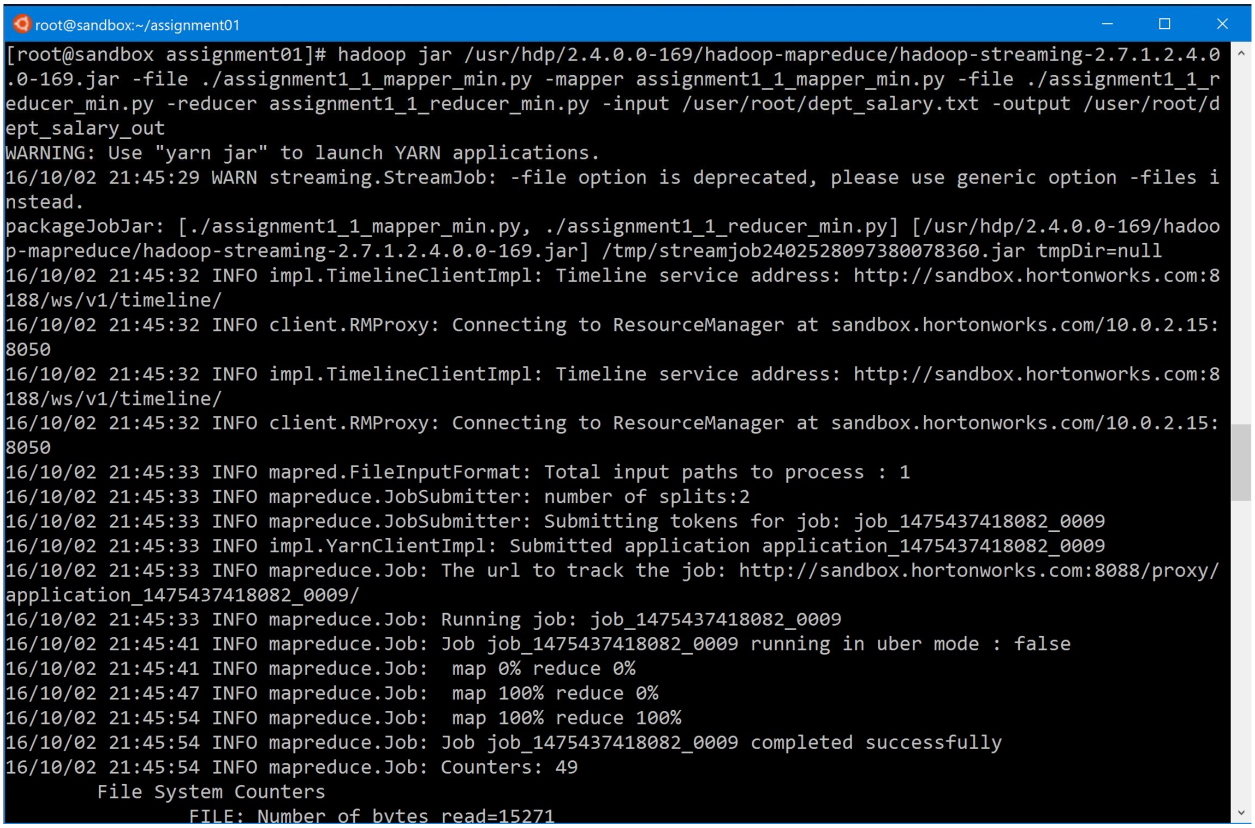
   ```
   #!/usr/bin/env python

   import sys

   for line in sys.stdin:
       line = line.strip()
       salary = line.split()
       if len(salary) == 2:
           try:
               print '%s\t%s' % (salary[0], salary[1])
           except:
               continue
   ```

   Reducer Python Code:

   ```
   #!/usr/bin/env python

   from operator import itemgetter
   import sys

   current_dept = None
   current_min_salary = 0
   dept = None

   # input comes from STDIN
   for line in sys.stdin:
       line = line.strip()

       dept, salary = line.split('\t', 1)

       try:
           salary = int(salary)
       except ValueError:
           # count was not a number, so silently
           # ignore/discard this line
           continue


       if current_dept == dept:
   ```

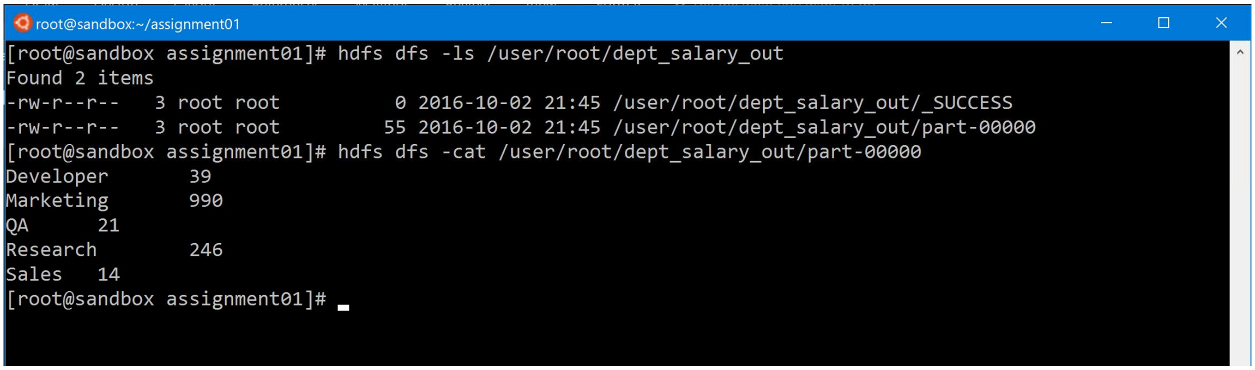
```
        if salary < current_min_salary:
            current_min_salary = salary
    else:
        if current_dept:
            print '%s\t%s' % (current_dept, current_min_salary)
        current_dept = dept
        current_min_salary = salary

if current_dept == dept:
    print '%s\t%s' % (current_dept, current_min_salary)
```

Execution in Hadoop:



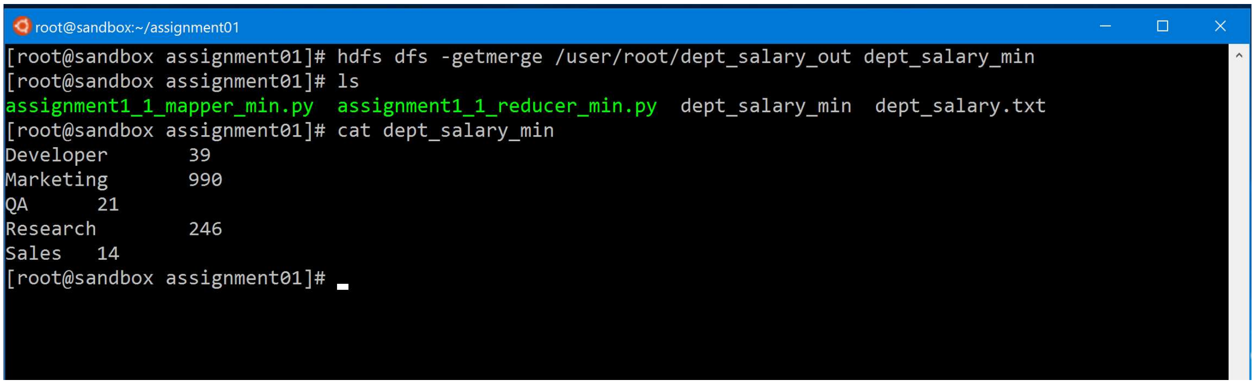
```
[root@sandbox assignment01]# hadoop jar /usr/hdp/2.4.0.0-169/hadoop-mapreduce/hadoop-streaming-2.7.1.2.4.0
.0-169.jar -file ./assignment1_1_mapper_min.py -mapper assignment1_1_mapper_min.py -file ./assignment1_1_r
educer_min.py -reducer assignment1_1_reducer_min.py -input /user/root/dept_salary.txt -output /user/root/d
ept_salary_out
WARNING: Use "yarn jar" to launch YARN applications.
16/10/02 21:45:29 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files i
nstead.
packageJobJar: [./assignment1_1_mapper_min.py, ./assignment1_1_reducer_min.py] [/usr/hdp/2.4.0.0-169/hadoo
p-mapreduce/hadoop-streaming-2.7.1.2.4.0.0-169.jar] /tmp/streamjob2402528097380078360.jar tmpDir=null
16/10/02 21:45:32 INFO impl.TimelineClientImpl: Timeline service address: http://sandbox.hortonworks.com:8
188/ws/v1/timeline/
16/10/02 21:45:32 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:
8050
16/10/02 21:45:32 INFO impl.TimelineClientImpl: Timeline service address: http://sandbox.hortonworks.com:8
188/ws/v1/timeline/
16/10/02 21:45:32 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:
8050
16/10/02 21:45:33 INFO mapred.FileInputFormat: Total input paths to process : 1
16/10/02 21:45:33 INFO mapreduce.JobSubmitter: number of splits:2
16/10/02 21:45:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1475437418082_0009
16/10/02 21:45:33 INFO impl.YarnClientImpl: Submitted application application_1475437418082_0009
16/10/02 21:45:33 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/
application_1475437418082_0009/
16/10/02 21:45:33 INFO mapreduce.Job: Running job: job_1475437418082_0009
16/10/02 21:45:41 INFO mapreduce.Job: Job job_1475437418082_0009 running in uber mode : false
16/10/02 21:45:41 INFO mapreduce.Job:  map 0% reduce 0%
16/10/02 21:45:47 INFO mapreduce.Job:  map 100% reduce 0%
16/10/02 21:45:54 INFO mapreduce.Job:  map 100% reduce 100%
16/10/02 21:45:54 INFO mapreduce.Job: Job job_1475437418082_0009 completed successfully
16/10/02 21:45:54 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=15271
```

Output from bash command line:



```
[root@sandbox assignment01]# hdfs dfs -ls /user/root/dept_salary_out
Found 2 items
-rw-r--r--   3 root root          0 2016-10-02 21:45 /user/root/dept_salary_out/_SUCCESS
-rw-r--r--   3 root root         55 2016-10-02 21:45 /user/root/dept_salary_out/part-00000
[root@sandbox assignment01]# hdfs dfs -cat /user/root/dept_salary_out/part-00000
Developer       39
Marketing       990
QA      21
Research        246
Sales   14
[root@sandbox assignment01]#
```

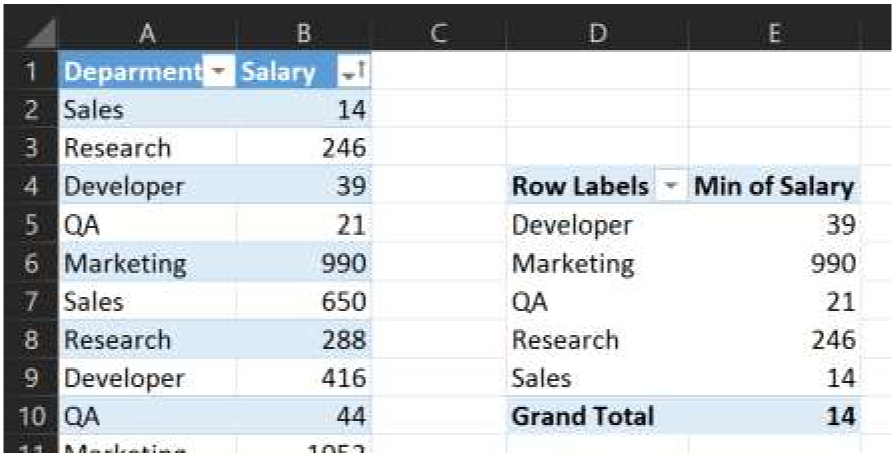
Output retrieved using getmerge:



```
[root@sandbox assignment01]# hdfs dfs -getmerge /user/root/dept_salary_out dept_salary_min
[root@sandbox assignment01]# ls
assignment1_1_mapper_min.py  assignment1_1_reducer_min.py  dept_salary_min  dept_salary.txt
[root@sandbox assignment01]# cat dept_salary_min
Developer       39
Marketing       990
QA      21
Research        246
Sales   14
[root@sandbox assignment01]#
```

```
Developer   39
Marketing   990
QA    21
Research    246
Sales 14
```

Results validation using Excel:



	A	B	C	D	E
1	Deparment	Salary			
2	Sales	14			
3	Research	246			
4	Developer	39		Row Labels	Min of Salary
5	QA	21		Developer	39
6	Marketing	990		Marketing	990
7	Sales	650		QA	21
8	Research	288		Research	246
9	Developer	416		Sales	14
10	QA	44		**Grand Total**	**14**
11	Marketing	1052			

2. The idea is to have two Map Reduce steps; the first step does the count over all words (exactly as in the example provided for the lab during class) and the second step sorts the results obtained by the first map reduce step by word count and emits only the top 10 words.

First Mapper Python Code (Word Count):

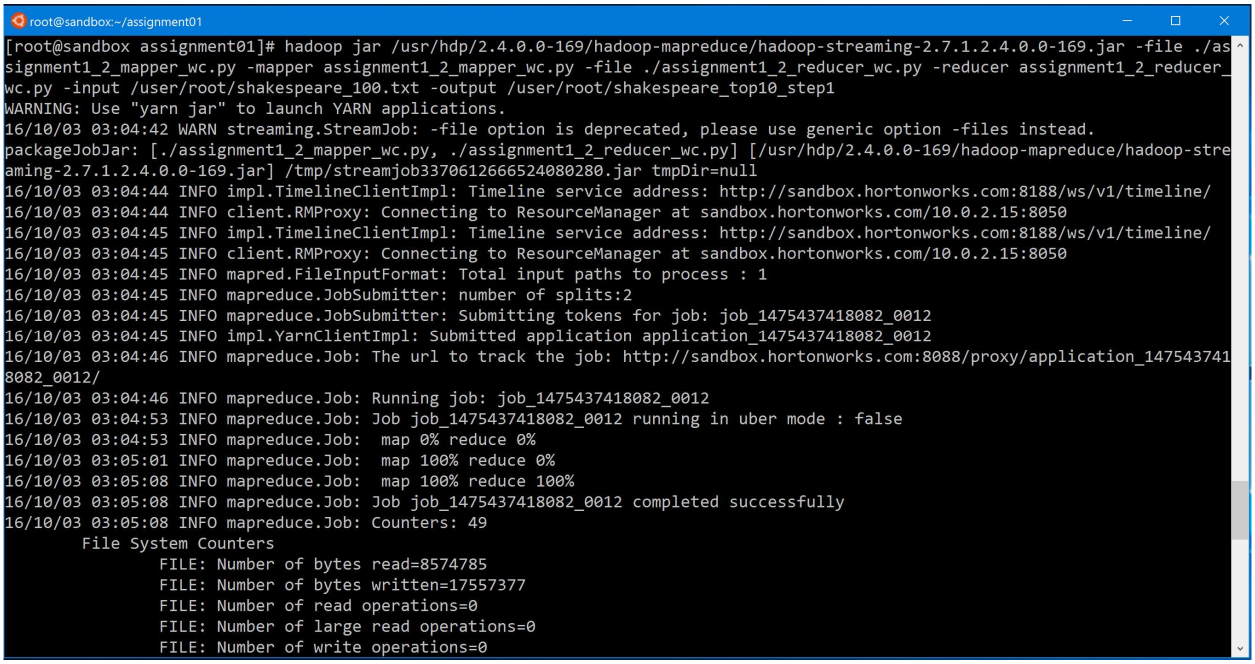
```
#!/usr/bin/env python

import sys

for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        print '%s\t%s' % (word, 1)
```

First Reducer Python Code (Word Count):

```
#!/usr/bin/env python

from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None
topn = 0

# input comes from STDIN
for line in sys.stdin:
    line = line.strip()
    word, count = line.split('\t', 1)

    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue


    if current_word == word:
        current_count += count
    else:
        if current_word:
            print '%s\t%s' % (current_word, current_count)
        current_count = count
        current_word = word
```

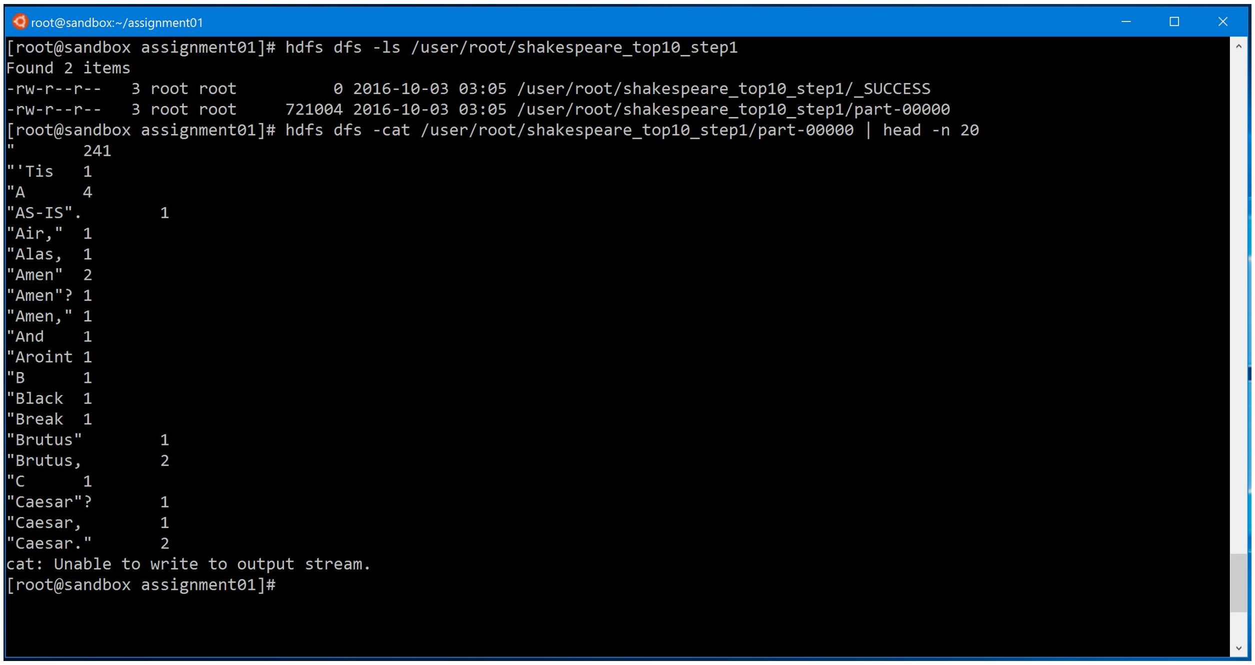
```
if current_word == word:
    print '%s\t%s' % (current_word, current_count)
```

First Map Reduce Execution in Hadoop:



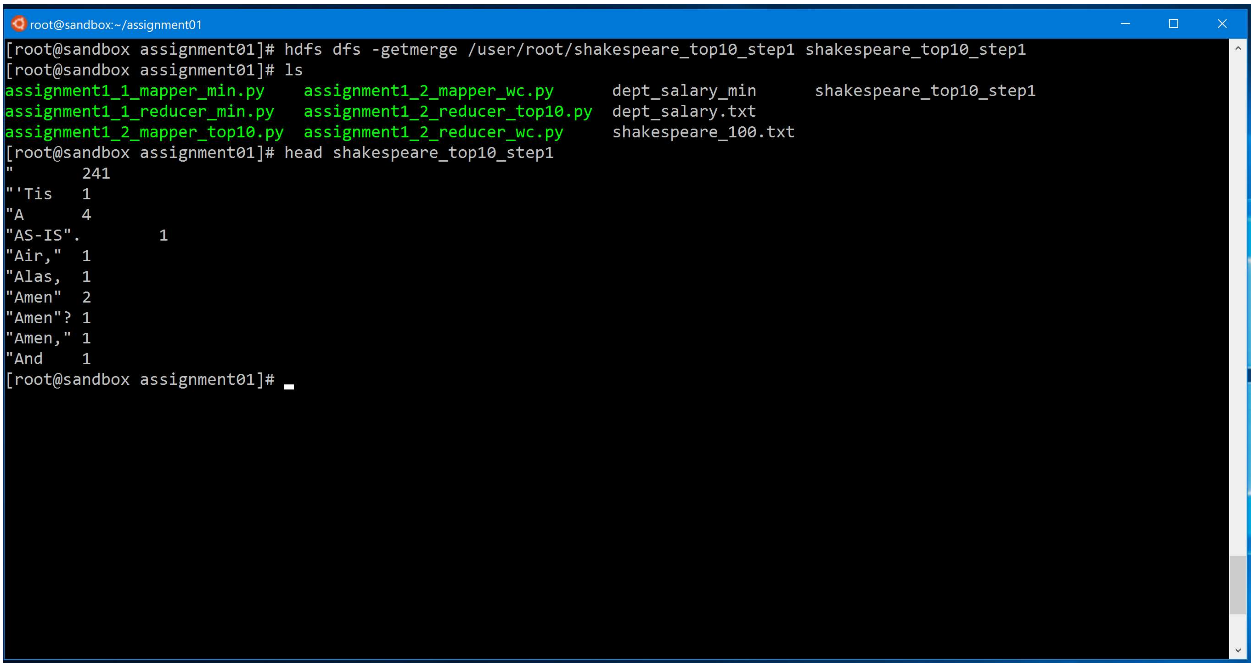
```
[root@sandbox assignment01]# hadoop jar /usr/hdp/2.4.0.0-169/hadoop-mapreduce/hadoop-streaming-2.7.1.2.4.0.0-169.jar -file ./as
signment1_2_mapper_wc.py -mapper assignment1_2_mapper_wc.py -file ./assignment1_2_reducer_wc.py -reducer assignment1_2_reducer_
wc.py -input /user/root/shakespeare_100.txt -output /user/root/shakespeare_top10_step1
WARNING: Use "yarn jar" to launch YARN applications.
16/10/03 03:04:42 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [./assignment1_2_mapper_wc.py, ./assignment1_2_reducer_wc.py] [/usr/hdp/2.4.0.0-169/hadoop-mapreduce/hadoop-stre
aming-2.7.1.2.4.0.0-169.jar] /tmp/streamjob3370612666524080280.jar tmpDir=null
16/10/03 03:04:44 INFO impl.TimelineClientImpl: Timeline service address: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
16/10/03 03:04:44 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
16/10/03 03:04:45 INFO impl.TimelineClientImpl: Timeline service address: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
16/10/03 03:04:45 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
16/10/03 03:04:45 INFO mapred.FileInputFormat: Total input paths to process : 1
16/10/03 03:04:45 INFO mapreduce.JobSubmitter: number of splits:2
16/10/03 03:04:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1475437418082_0012
16/10/03 03:04:45 INFO impl.YarnClientImpl: Submitted application application_1475437418082_0012
16/10/03 03:04:46 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_147543741
8082_0012/
16/10/03 03:04:46 INFO mapreduce.Job: Running job: job_1475437418082_0012
16/10/03 03:04:53 INFO mapreduce.Job: Job job_1475437418082_0012 running in uber mode : false
16/10/03 03:04:53 INFO mapreduce.Job:  map 0% reduce 0%
16/10/03 03:05:01 INFO mapreduce.Job:  map 100% reduce 0%
16/10/03 03:05:08 INFO mapreduce.Job:  map 100% reduce 100%
16/10/03 03:05:08 INFO mapreduce.Job: Job job_1475437418082_0012 completed successfully
16/10/03 03:05:08 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=8574785
                FILE: Number of bytes written=17557377
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

First Map Reduce output in bash command line:



```
[root@sandbox assignment01]# hdfs dfs -ls /user/root/shakespeare_top10_step1
Found 2 items
-rw-r--r--   3 root root          0 2016-10-03 03:05 /user/root/shakespeare_top10_step1/_SUCCESS
-rw-r--r--   3 root root     721004 2016-10-03 03:05 /user/root/shakespeare_top10_step1/part-00000
[root@sandbox assignment01]# hdfs dfs -cat /user/root/shakespeare_top10_step1/part-00000 | head -n 20
"       241
"'Tis   1
"A      4
"AS-IS".        1
"Air,"  1
"Alas,  1
"Amen"  2
"Amen"? 1
"Amen," 1
"And    1
"Aroint 1
"B      1
"Black  1
"Break  1
"Brutus"        1
"Brutus,        2
"C      1
"Caesar"?       1
"Caesar,        1
"Caesar."       2
cat: Unable to write to output stream.
[root@sandbox assignment01]#
```

First Map Reduce output with getmerge:



```
[root@sandbox assignment01]# hdfs dfs -getmerge /user/root/shakespeare_top10_step1 shakespeare_top10_step1
[root@sandbox assignment01]# ls
assignment1_1_mapper_min.py    assignment1_2_mapper_wc.py      dept_salary_min      shakespeare_top10_step1
assignment1_1_reducer_min.py   assignment1_2_reducer_top10.py  dept_salary.txt
assignment1_2_mapper_top10.py  assignment1_2_reducer_wc.py     shakespeare_100.txt
[root@sandbox assignment01]# head shakespeare_top10_step1
"       241
"'Tis   1
"A      4
"AS-IS".        1
"Air,"  1
"Alas,  1
"Amen"  2
"Amen"? 1
"Amen," 1
"And    1
[root@sandbox assignment01]#
```

For the second map reducer job, the code for the mapper and reducer are the same, just return a list of words sorted by count number in descendent order.

Second Mapper Python Code (Top 10 Words by Count):

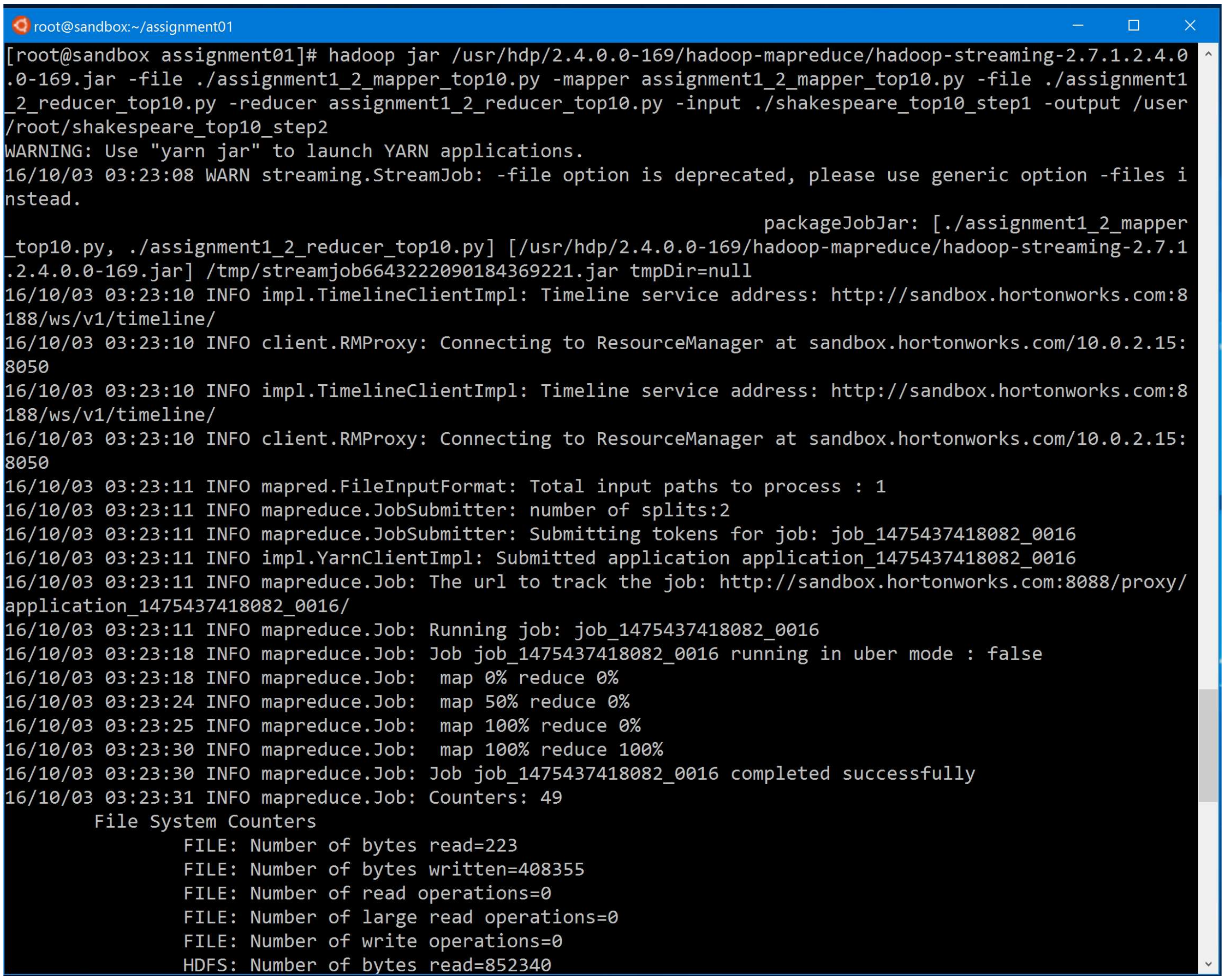
```
#!/usr/bin/env python

import sys
import operator

topnwords = {}  #dictionary to sort the words

for line in sys.stdin:
    line = line.strip()
    words = line.split()
    if len(words) == 2:
        try:
            count = int(words[1])    # word count
            topnwords[words[0]] = count # add word and count to
dictonary
        except:
            continue

# list of words sorted by count
sorted = sorted(topnwords, key=topnwords.__getitem__, reverse=True)
```

```
n = 0   #counter to limit print of top 10 words
for w in sorted:    #iterate through all words sorted by count
    if n < 10:      # print only top 10 words
        print '%s\t%s' % (w, topnwords[w])
        n = n + 1
    else:
        continue
```

Second Reducer Python Code (Top 10 Words by Count):

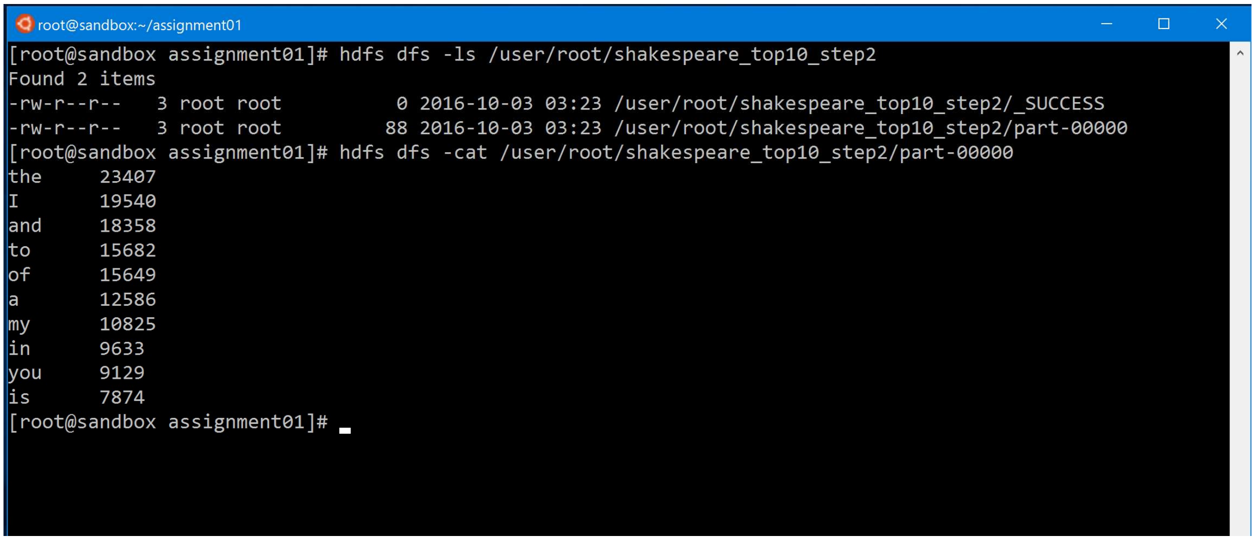
```
#!/usr/bin/env python


import sys
import operator

topnwords = {}  #dictionary to sort the words

for line in sys.stdin:
    line = line.strip()
    words = line.split()
    if len(words) == 2:
        try:
            count = int(words[1])   # word count
            topnwords[words[0]] = count # add word and count to
dictonary
        except:
            continue

# list of words sorted by count
sorted = sorted(topnwords, key=topnwords.__getitem__, reverse=True)

n = 0   #counter to limit print of top 10 words
for w in sorted:    #iterate through all words sorted by count
    if n < 10:      # print only top 10 words
        print '%s\t%s' % (w, topnwords[w])
        n = n + 1
    else:
        continue
```

Second Map Reduce Execution in Hadoop:



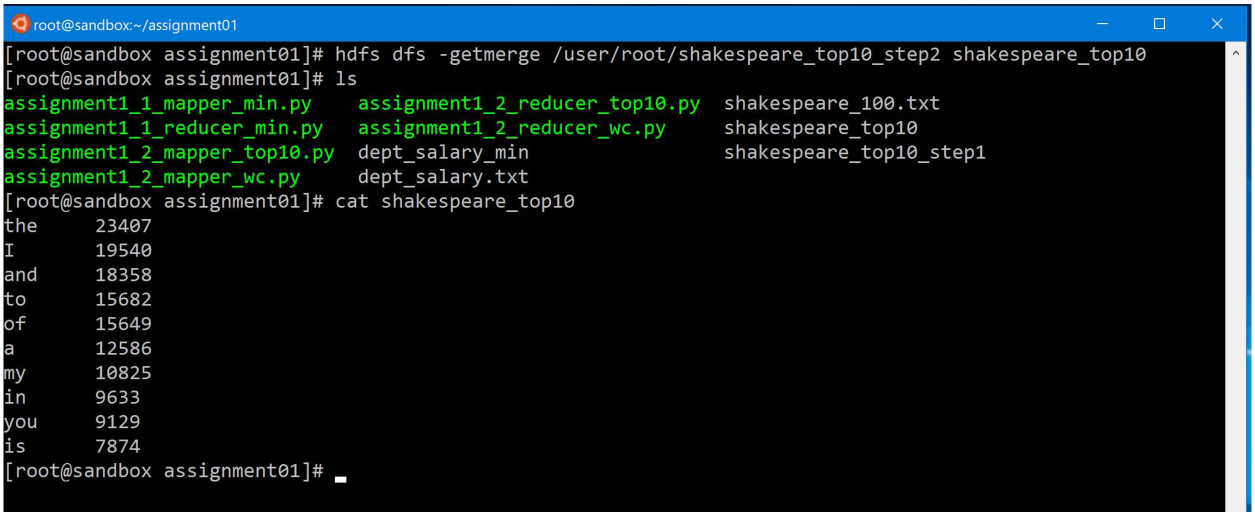
```
[root@sandbox assignment01]# hadoop jar /usr/hdp/2.4.0.0-169/hadoop-mapreduce/hadoop-streaming-2.7.1.2.4.0
.0-169.jar -file ./assignment1_2_mapper_top10.py -mapper assignment1_2_mapper_top10.py -file ./assignment1
_2_reducer_top10.py -reducer assignment1_2_reducer_top10.py -input ./shakespeare_top10_step1 -output /user
/root/shakespeare_top10_step2
WARNING: Use "yarn jar" to launch YARN applications.
16/10/03 03:23:08 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files i
nstead.
                                                                     packageJobJar: [./assignment1_2_mapper
_top10.py, ./assignment1_2_reducer_top10.py] [/usr/hdp/2.4.0.0-169/hadoop-mapreduce/hadoop-streaming-2.7.1
.2.4.0.0-169.jar] /tmp/streamjob6643222090184369221.jar tmpDir=null
16/10/03 03:23:10 INFO impl.TimelineClientImpl: Timeline service address: http://sandbox.hortonworks.com:8
188/ws/v1/timeline/
16/10/03 03:23:10 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:
8050
16/10/03 03:23:10 INFO impl.TimelineClientImpl: Timeline service address: http://sandbox.hortonworks.com:8
188/ws/v1/timeline/
16/10/03 03:23:10 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:
8050
16/10/03 03:23:11 INFO mapred.FileInputFormat: Total input paths to process : 1
16/10/03 03:23:11 INFO mapreduce.JobSubmitter: number of splits:2
16/10/03 03:23:11 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1475437418082_0016
16/10/03 03:23:11 INFO impl.YarnClientImpl: Submitted application application_1475437418082_0016
16/10/03 03:23:11 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/
application_1475437418082_0016/
16/10/03 03:23:11 INFO mapreduce.Job: Running job: job_1475437418082_0016
16/10/03 03:23:18 INFO mapreduce.Job: Job job_1475437418082_0016 running in uber mode : false
16/10/03 03:23:18 INFO mapreduce.Job:  map 0% reduce 0%
16/10/03 03:23:24 INFO mapreduce.Job:  map 50% reduce 0%
16/10/03 03:23:25 INFO mapreduce.Job:  map 100% reduce 0%
16/10/03 03:23:30 INFO mapreduce.Job:  map 100% reduce 100%
16/10/03 03:23:30 INFO mapreduce.Job: Job job_1475437418082_0016 completed successfully
16/10/03 03:23:31 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=223
                FILE: Number of bytes written=408355
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=852340
```

Second Map Reduce output in bash command line:



```
[root@sandbox assignment01]# hdfs dfs -ls /user/root/shakespeare_top10_step2
Found 2 items
-rw-r--r--   3 root root          0 2016-10-03 03:23 /user/root/shakespeare_top10_step2/_SUCCESS
-rw-r--r--   3 root root         88 2016-10-03 03:23 /user/root/shakespeare_top10_step2/part-00000
[root@sandbox assignment01]# hdfs dfs -cat /user/root/shakespeare_top10_step2/part-00000
the     23407
I       19540
and     18358
to      15682
of      15649
a       12586
my      10825
in      9633
you     9129
is      7874
[root@sandbox assignment01]#
```

Second Map Reduce output with getmerge:



```
[root@sandbox assignment01]# hdfs dfs -getmerge /user/root/shakespeare_top10_step2 shakespeare_top10
[root@sandbox assignment01]# ls
assignment1_1_mapper_min.py    assignment1_2_reducer_top10.py  shakespeare_100.txt
assignment1_1_reducer_min.py   assignment1_2_reducer_wc.py     shakespeare_top10
assignment1_2_mapper_top10.py  dept_salary_min                 shakespeare_top10_step1
assignment1_2_mapper_wc.py     dept_salary.txt
[root@sandbox assignment01]# cat shakespeare_top10
the     23407
I       19540
and     18358
to      15682
of      15649
a       12586
my      10825
in      9633
you     9129
is      7874
[root@sandbox assignment01]# 
```

```
the	23407
I	19540
and	18358
to	15682
of	15649
a	12586
my	10825
in	9633
you	9129
is	7874
```