

Lab #9 : Running a map-reduce job on a Hadoop cluster

Exercise 1: Running map-reduce jobs on the cloud

Lab report screen-shot #1:

The screenshot shows the AWS Cloud9 IDE interface. On the left, the file tree displays a directory structure for 'Danddank Lab7' containing 'Danddank-MR-WordCount' and 'input.txt'. The code editor shows a Python script named 'databaseconnection.py' with the following content:

```

1 This is a test input.
2 This is mentioned three times.
3 However input is not mentioned three times in this file.
4 input is mentioned four times because
5 here is another input.

```

Below the code editor, there are tabs for 'DanddankLab7', 'bash - *ip-172...', 'danddank-db', and 'databaseconn'. A 'Run' button is visible at the bottom.

Lab report screen-shot #2:

The screenshot shows the AWS Cloud9 IDE interface. On the left, the file tree displays a directory structure for 'Danddank Lab7' containing 'Danddank-MR-WordCount' and 'input.txt'. The code editor shows a Python script named 'mapper.py' with the following content:

```

1 #!/usr/bin/env python
2
3 import sys
4
5 for line in sys.stdin:
6     line = line.strip()
7     words = line.split()
8     for word in words:
9         print("%s\t%s" % (word, 1))

```

Below the code editor, there are tabs for 'Danddank', 'bash - *ip-x', 'danddank', 'database', and 'bash - *ip-x'. A terminal window at the bottom shows the execution of the mapper script on the input file:

```

vocstartsoft:~/environment/danddank-mr-wordcount (master)$ ls
input.txt
vocstartsoft:~/environment/danddank-mr-wordcount (master)$ cat input.txt | python3 mapper.py
This 1
is 1
a 1
test 1
input. 1
This 1
is 1
not 1
mentioned 1
three 1
times 1
However 1
input 1
is 1
is 1
not 1
mentioned 1
three 1
times 1
this 1
file. 1
input 1
is 1
mentioned 1
four 1
times 1
because 1
here 1
is 1
another 1
input. 1

```

Lab report screen-shot #3:

The screenshot shows a browser-based terminal interface with multiple tabs and panes. The tabs include "My Class", "Worker", "Your env", "Danddan", "Browser", and "Run". The main pane displays a Python script named `mapper.py` which reads from standard input and prints words and their counts. The terminal output shows the script running and producing word counts for the input file `input.txt`.

```
#!/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))
```

Another terminal window shows the command `cat input.txt | python3 mapper.py | sort` being run, resulting in the following word count output:

```
a 1
another 1
because 1
file. 1
four 1
here 1
However 1
in 1
input. 1
input. 1
input. 1
input. 1
input. 1
is 1
is 1
is 1
is 1
is 1
mentioned 1
mentioned 1
mentioned 1
not 1
test 1
this 1
This 1
three 1
three 1
times. 1
times 1
times 1
```

Lab report screen-shot #4:

The screenshot shows a browser-based development interface for a Lambda function named "Danddark". The top navigation bar includes tabs for "My Class", "Workben", "Your env", "Danddar", "Browser", and a "+" button. The main area has a "File" menu with "Edit", "Find", "View", "Go", "Run", "Tools", "Window", "Support", and "Preview" options. A "Run" button is located in the top right. On the left, there's a sidebar with a search bar ("Go to Anything (% P)"), file navigation icons, and a tree view of the project structure:

- Danddark Lab7 Er
- danddark-mr-wordcount
 - input.txt
 - mapper.py
 - reducer.py
- DanddarkLab7
- ssl
- danddark-db.py
- danddark-db1.py
- danddark_hello_world.t
- databaseconnection.py
- README.md

The central workspace contains three tabs: "databaselc", "input.txt", and "mapper.py". The "mapper.py" tab displays the following Python code:

```
#!/usr/bin/env python
import sys
current_word = None
current_count = 0
word = None

for line in sys.stdin:
    line = line.strip()
    (word, count) = line.split('\t')
    count = int(count)
    if current_word == word:
        current_count = current_count + 1
    else:
        if current_word:
            print(current_word, current_count)
        current_count = count
        current_word = word
    if current_word == word:
        print(current_word, current_count)
```

The status bar at the bottom indicates "Python Spaces: 4". Below the workspace, a terminal window shows the execution of the application:

```
times 1
vocstartsoft:~/environment/danddark-mr-wordcount (master)$ cat input.txt | python mapper.py | sort | python reducer.py
a 1
another 1
because 1
file. 1
four 1
here 1
However 1
in 1
input. 2
input 2
is 5
mentioned 3
not 1
test 1
this 1
This 2
three 2
times. 1
times 2
vocstartsoft:~/environment/danddark-mr-wordcount (master)$
```

Lab report screen-shot #5:

The screenshot shows the AWS Cloud9 IDE interface. The left sidebar displays a file tree for a project named 'DanddankLab7' containing subfolders 'danddank-mr-wordcount' and 'ssl', and files 'danddank-db.py', 'danddank-db1.py', 'danddank_hello_world.t', 'databaseconnection.py', and 'README.md'. The main workspace has tabs for 'mapper.py' and 'reducer.py'. The code editor shows a Python script 'mapper.py' with the following content:#!/usr/bin/env python
import sys
for line in sys.stdin:
 line = line.strip().lower()
 words = line.split()
 for word in words:
 print("%s\t%s" % (word, 1))The terminal window below shows the output of running the script on an input file 'input.txt' containing Shakespearean text. The output shows word counts like 'here 1', 'However 1', 'in 1', etc.

Lab report screen-shot #6:

This screenshot is similar to the previous one but shows a different input file, 'input.txt', which contains a list of words from a Shakespeare play. The terminal output shows word counts such as 'youth! 7', 'youth? 5', 'youth.' 1', 'youth. 16', 'youth 155', 'you--that 1', 'you'--that 1', 'youthful 1', 'youthful 31', 'youth's 7', 'youths 5', 'you've 5', 'you--well, 1', 'you--well 1', 'you--why 1', 'y-ravished 1', 'y-slaked 1', 'zanies. 1', 'zany, 1', 'zeal, 6', 'zeal: 2', 'zeal! 1', 'zeal. 2', 'zeal 23', 'zealous 6', 'zeals, 1', 'zed! 1', 'zenelophon; 1', 'zenith 1', 'zephyrs 1', 'zir, 1', 'zir: 1', 'zo 1', 'zodiac 1', 'zodiacs 1', 'zone, 1', 'ounds, 15', 'ounds! 3', 'ounds, 1', 'ounds! 1', 'zaggered 1'. The command 'vocstartsoft:~/environment/danddank-mr-wordcount (master) \$' is visible at the bottom of the terminal.

Lab report screen-shot #7:

The screenshot shows the Amazon EMR console interface. On the left, there's a sidebar with navigation links for 'Amazon EMR', 'EMR Studio', 'EMR on EC2' (selected), 'Clusters', 'Notebooks', 'Git repositories', 'Security configurations', 'Block public access', 'VPC subnets', 'Events', and 'EMR on EKS'. The main content area displays the following information for the cluster 'danddankcluster':

- Cluster ID:** j-1WSH3RPZJYSMZ
- Creation date:** 2021-10-30 16:09 (UTC-5)
- Elapsed time:** 2 minutes
- After last step completes:** Cluster waits
- Termination protection:** On Change
- Tags:** -- View All / Edit
- Master public DNS:** ec2-3-86-140-85.compute-1.amazonaws.com
- Connect to the Master Node Using SSH**

Configuration details

- Release label:** emr-5.32.0
- Hadoop distribution:** Amazon 2.10.1
- Applications:** Hive 2.3.7, Pig 0.17.0, Hue 4.8.0, JupyterHub 1.1.0, Oozie 5.2.0, HBase 1.4.13, Spark 2.4.7, Sqoop 1.4.7
- Log URI:** s3://aws-logs-196580907486-us-east-1/elasticmapreduce/
- EMRFS consistent view:** Disabled
- Custom AMI ID:** --

Application user interfaces

- Persistent user interfaces:** --
- On-cluster user interfaces:** Not Enabled | Enable an SSH Connection

At the bottom, there are links for 'Feedback', 'English (US)', 'Privacy Policy', 'Terms of Use', and 'Cookie preferences'.

Lab report screen-shot #8:

The screenshot shows the AWS EC2 Instances page. On the left, there's a sidebar with navigation links for 'New EC2 Experience' (selected), 'EC2 Dashboard', 'EC2 Global View', 'Events', 'Tags', 'Limits', 'Instances' (selected), 'Images', 'AMIs', and 'Elastic Block Store' (Volumes and Snapshots). The main content area displays a table of 21 terminated instances:

Name	Instance ID	Instance state	Instance type	Status check	Alarm
-	i-096a0c06ade8d099b	Terminated	m5.xlarge	-	No alarm
-	i-07287cba18e6a1cf9	Terminated	m5.xlarge	-	No alarm
-	i-0fba79568ec6219ca	Running	m5.xlarge	Initializing	No alarm
-	i-024a1892fff880984	Running	m5.xlarge	Initializing	No alarm
-	i-08515c52f8b2aea5d	Running	m5.xlarge	Initializing	No alarm
aws-cloud9-Danddank-Lab9-En...	i-0317aa4ca415791c3	Stopped	t2.micro	-	No alarm
danddank-ics432website-instance	i-0a4aff20a30b1214a	Terminated	t2.micro	-	No alarm

A message at the bottom says 'Select an instance above'.

At the bottom, there are links for 'Feedback', 'English (US)', 'Privacy Policy', 'Terms of Use', and 'Cookie preferences'.

Lab report screen-shot #9:

```
zeal: 2
zeal! 1
zeal. 2
zeal 23
zealous 6
zeals, 1
zed! 1
zenelophon; 1
zenith 1
zephyrs 1
zir, 1
zir: 1
zo 1
zodiac 1
zodiacs 1
zone, 1
'zounds, 15
'zounds! 3
zounds, 1
zounds! 1
zwaggeder 1
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ ls
input.txt mapper.py reducer.py shakespeare
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ ls shakespeare/
comedies histories poems tragedies
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$
```

Lab report screen-shot #10:

```
ntonhdfs/shakespeare
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -ls dandda
nkwordcountonhdfs/shakespeare
Found 4 items
-rw-r--r-- 1 hadoop hadoop 1784616 2021-10-30 22:00 danddankwordcou
ntonhdfs/shakespeare/comedies
-rw-r--r-- 1 hadoop hadoop 1479035 2021-10-30 22:00 danddankwordcou
ntonhdfs/shakespeare/histories
-rw-r--r-- 1 hadoop hadoop 268140 2021-10-30 22:00 danddankwordcou
ntonhdfs/shakespeare/poems
-rw-r--r-- 1 hadoop hadoop 1752440 2021-10-30 22:00 danddankwordcou
ntonhdfs/shakespeare/tragedies
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -lsr dandd
ankwordcountonhdfs
lsr: DEPRECATED: Please use 'ls -R' instead.
drwxr-xr-x - hadoop hadoop 0 2021-10-30 22:00 danddankwordcou
ntonhdfs/shakespeare
-rw-r--r-- 1 hadoop hadoop 1784616 2021-10-30 22:00 danddankwordcou
ntonhdfs/shakespeare/comedies
-rw-r--r-- 1 hadoop hadoop 1479035 2021-10-30 22:00 danddankwordcou
ntonhdfs/shakespeare/histories
-rw-r--r-- 1 hadoop hadoop 268140 2021-10-30 22:00 danddankwordcou
ntonhdfs/shakespeare/poems
-rw-r--r-- 1 hadoop hadoop 1752440 2021-10-30 22:00 danddankwordcou
ntonhdfs/shakespeare/tragedies
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$
```

Lab report screen-shot #11:

- a. **hadoop fs -mkdir danddankFolder**
 - b. **hadoop fs -put input.txt danddankFolder/input.txt**
 - c. **hadoop fs -ls danddankFolder**
 - d. **hadoop fs -cat danddankFolder/input.txt**
 - e. **hadoop fs -mv danddankFolder/input.txt danddankFolder/newInput.txt**
 - f. **hadoop fs -mkdir nalongsoneFolder**
- Had oop fs -mv danddankFolder/newInput.txt nalongsoneFolder/newInput.txt
- g. **hadoop fs -du nalongsoneFolder/newInput.txt**
 - h. **hadoop fs -get nalongsoneFolder/newInput.txt newInput.txt**
 - i. **hadoop fs -rm nalongsoneFolder/newInput.txt**
 - j. **hadoop fs -rm -r nalongsoneFolder**

```
hadoop@ip-172-31-90-37:~$ ls
danddankwordcountonmaster
[hadoop@ip-172-31-90-37 ~]$ cd danddankwordcountonmaster/
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ ls
input.txt mapper.py reducer.py shakespeare
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -lsr danddankwordcountonhdfs
lsr: DEPRECATED: Please use 'ls -R' instead.
drwxr-xr-x - hadoop hadoop 0 2021-10-30 22:00 danddankwordcountonhdfs/shakespeare
-rw-r--r-- 1 hadoop hadoop 1784616 2021-10-30 22:00 danddankwordcountonhdfs/shakespeare/comedies
-rw-r--r-- 1 hadoop hadoop 1479035 2021-10-30 22:00 danddankwordcountonhdfs/shakespeare/histories
-rw-r--r-- 1 hadoop hadoop 268140 2021-10-30 22:00 danddankwordcountonhdfs/shakespeare/poems
-rw-r--r-- 1 hadoop hadoop 1752440 2021-10-30 22:00 danddankwordcountonhdfs/shakespeare/tragedies
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -mkdir danddankFolder
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -ls
Found 2 items
drwxr-xr-x - hadoop hadoop 0 2021-10-31 02:50 danddankFolder
drwxr-xr-x - hadoop hadoop 0 2021-10-30 22:00 danddankwordcountonhdfs
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ ls
input.txt mapper.py reducer.py shakespeare
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -put input.txt danddankFolder/input.txt
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -ls danddankFolder
Found 1 items
-rw-r--r-- 1 hadoop hadoop 170 2021-10-31 02:52 danddankFolder/input.txt
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -cat danddankFolder/input.txt
This is a test input.
This is mentioned three times.
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -mv danddankFolder/input.txt danddankFolder/newInput.txt
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -mkdir nalongsoneFolder
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -mv danddankFolder/newInput.txt nalongsoneFolder/newInput.txt
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -du -s nalongsoneFolder/newInput.txt
170 nalongsoneFolder/newInput.txt
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -dus nalongsoneFolder/newInput.txt
dus: DEPRECATED: Please use 'du -s' instead.
170 nalongsoneFolder/newInput.txt
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -du -s nalongsoneFolder/newInput.txt
170 nalongsoneFolder/newInput.txt
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ hadoop fs -get nalongsoneFolder/newInput.txt newInput.txt
get: /home/hadoop/danddankwordcountonmaster/newInput.txt._COPYING_ (Permission denied)
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ ls
input.txt mapper.py reducer.py shakespeare
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ cd ..
[hadoop@ip-172-31-90-37 ~]$ ls
danddankwordcountonmaster
[hadoop@ip-172-31-90-37 ~]$ hadoop fs -get nalongsoneFolder/newInput.txt newInput.txt
[hadoop@ip-172-31-90-37 ~]$ ls
danddankwordcountonmaster newInput.txt
[hadoop@ip-172-31-90-37 ~]$ hadoop fs -rm nalongsoneFolder/newInput.txt
Deleted nalongsoneFolder/newInput.txt
[hadoop@ip-172-31-90-37 ~]$ hadoop fs -rm -r nalongsoneFolder
Deleted nalongsoneFolder
[hadoop@ip-172-31-90-37 ~]$
```

Lab report screen-shot #12:

```
hadoop@ip-172-31-90-37:~/danddankwordcountonmaster
input.txt mapper.py reducer.py shakespeare
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ sudo chmod +x mapper.py
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ sudo chmod +x reducer.py
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ cat input.txt | ./mapper.py
this 1
is 1
a 1
test 1
input. 1
this 1
is 1
mentioned 1
three 1
times. 1
however 1
input 1
is 1
not 1
mentioned 1
three 1
times 1
in 1
this 1
file. 1
input 1
is 1
mentioned 1
four 1
times 1
because 1
here 1
is 1
another 1
input. 1
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$
```

Lab report screen-shot #13:

```
hadoop@ip-172-31-90-37:~/danddankwordcountonmaster
this 1
file. 1
input 1
is 1
mentioned 1
four 1
times 1
because 1
here 1
is 1
another 1
input. 1
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ cat input/file1.txt | ./mapper.py | sort | ./reducer.py
cat: input/file1.txt: No such file or directory
(None, 0)
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ cat input.txt | ./mapper.py | sort | ./reducer.py
('a', 1)
('another', 1)
('because', 1)
('file.', 1)
('four', 1)
('here', 1)
('however', 1)
('in', 1)
('input.', 2)
('input', 2)
('is', 5)
('mentioned', 3)
('not', 1)
('test', 1)
('this', 3)
('three', 2)
('times.', 1)
('times', 2)
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$
```

Lab report screen-shot #14:

```

times 1
because 1
here 1
is 1
another 1
input. 1
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ cat input/file1.txt | ./mapper.py | sort | ./reducer.py
cat: input/file1.txt: No such file or directory
(None, 0)
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ cat input.txt | ./mapper.py | sort | ./reducer.py
('a', 1)
('another', 1)
('because', 1)
('file', 1)
('four', 1)
('here', 1)
('however', 1)
('in', 1)
('input.', 2)
('input!', 2)
('is', 5)
('mentioned', 3)
('not', 1)
('test', 1)
('this', 3)
('three', 2)
('times.', 1)
('times', 2)
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ find /usr/lib/ -name *hadoop*streaming*.jar
/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-0.jar
/usr/lib/hadoop/hadoop-streaming.jar
/usr/lib/hadoop-mapreduce/hadoop-streaming.jar
/usr/lib/hadoop-mapreduce/hadoop-streaming-2.10.1-amzn-0.jar
find: '/usr/lib/hadoop-kms/share/hadoop/kms/tomcat/logs': Permission denied
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ 

```

Lab report screen-shot #15:

How many data-local map tasks were executed?	4
How many rack-local map tasks were executed?	6
How many reduce tasks?	3
Total time spent by all maps?	4296096
Total time spent by all reduce tasks?	9677
How many map-input records?	173126
How many map output records?	939236
How many reduce input groups?	62933

```

hadoop@ip-172-31-90-37:~/danddankwordcountonmaster
FILE: Number of bytes written=5354161
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5647788
HDFS: Number of bytes written=1061542
HDFS: Number of read operations=39
HDFS: Number of large read operations=0
HDFS: Number of write operations=6
Job Counters
    Killed reduce tasks=1
    Launched map tasks=10
    Launched reduce tasks=3
    Data-local map tasks=4
    Rack-local map tasks=6
    Total time spent by all maps in occupied slots (ms)=4296096
    Total time spent by all reduces in occupied slots (ms)=1857984
    Total time spent by all map tasks (ms)=44751
    Total time spent by all reduce tasks (ms)=9677
    Total vcore-milliseconds taken by all map tasks=44751
    Total vcore-milliseconds taken by all reduce tasks=9677
    Total megabyte-milliseconds taken by all map tasks=137457072
    Total megabyte-milliseconds taken by all reduce tasks=59455488
Map-Reduce Framework
    Map input records=173126
    Map output records=939236
    Map output bytes=7011642
    Map output materialized bytes=1542380
    Input split bytes=1493
    Combine input records=0
    Combine output records=0
    Reduce input groups=62933
    Reduce shuffle bytes=1542380
    Reduce input records=939236
    Reduce output records=62933
    Spilled Records=1878472
    Shuffled Maps =30
    Failed Shuffles=0
    Merged Map outputs=30
    GC time elapsed (ms)=1409
    CPU time spent (ms)=32370
    Physical memory (bytes) snapshot=6598761088
    Virtual memory (bytes) snapshot=68663730176
    Total committed heap usage (bytes)=5922357248
Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
File Input Format Counters
    Bytes Read=5646295
File Output Format Counters
    Bytes Written=1061542
21/10/31 03:41:56 INFO streaming.StreamJob: Output directory: danddankcountonhdfs/output
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$ 

```

Lab report screen-shot #16:

```
hadoop@ip-172-31-90-37:~/danddankwordcountonmaster
```

```
Reduce input groups=62933
Reduce shuffle bytes=1542380
Reduce input records=939236
Reduce output records=62933
Spilled Records=1878472
Shuffled Maps =30
Failed Shuffles=0
Merged Map outputs=30
GC time elapsed (ms)=1409
CPU time spent (ms)=32370
Physical memory (bytes) snapshot=6508761088
Virtual memory (bytes) snapshot=68663730176
Total committed heap usage (bytes)=5922357248

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
Bytes Read=5646295
File Output Format Counters
Bytes Written=1061542

21/10/31 03:41:56 INFO streaming.StreamJob: Output directory: danddankcountonhdfs/output
[hadoop@ip-172-31-90-37 danddankcountonmaster]$ hadoop fs -ls danddankcountonhdfs/output
Found 4 items
-rw-r--r-- 1 hadoop hadoop 0 2021-10-31 03:41 danddankcountonhdfs/output/_SUCCESS
-rw-r--r-- 1 hadoop hadoop 354355 2021-10-31 03:41 danddankcountonhdfs/output/part-00000
-rw-r--r-- 1 hadoop hadoop 354120 2021-10-31 03:41 danddankcountonhdfs/output/part-00001
-rw-r--r-- 1 hadoop hadoop 353067 2021-10-31 03:41 danddankcountonhdfs/output/part-00002
[hadoop@ip-172-31-90-37 danddankwordcountonmaster]$
```

Lab report screen-shot #17:

```
hadoop@ip-172-31-90-37:~/danddankwordcountonmaster
```

Word	Count
('willow', 1)	1
('willow:', 3)	3
('wills?', 1)	1
('wilt.', 12)	12
('wilt:', 1)	1
('wiltshire', 4)	4
('wiltshire's', 1)	1
('wiltshire?', 1)	1
('win.', 6)	6
('win:', 3)	3
('winchester's', 1)	1
('winchester.', 5)	5
('wincot.', 1)	1
('wind', 116)	116
('wind!', 4)	4
('wind-pipe', 1)	1
('wind-shaked', 1)	1
('wind-shaken.', 1)	1
('wind?', 1)	1
('winding', 5)	5
('window-bars', 1)	1
('window.', 3)	3
('window:', 1)	1
('windows', 14)	14
('windows.', 6)	6
('winds!', 2)	2
('winds?', 1)	1
('windsor.', 6)	6
('wine,', 20)	20
('wine;', 4)	4
('wing', 17)	17
('wing'd', 5)	5
('willow', 6)	6
('willow,', 11)	11
('willow;', 2)	2
('wills', 14)	14
('wills.', 3)	3
('wills:', 2)	2
('wilt', 17)	17
('wilt:', 4)	4
('wilt?!', 1)	1
('wiltshire', 1)	1
('wiltshire.', 1)	1
('win', 108)	108
('win!', 1)	1
('win?', 1)	1
('wince,', 2)	2
("winchester's", 1)	1
('winchester:', 1)	1
('wind', 45)	45
('wind-instruments?', 1)	1
('wind;', 8)	8
('winded', 3)	3
('winding-sheet;', 1)	1
('windlasses', 1)	1
('windmill', 1)	1
('windmill,', 1)	1
('window!', 1)	1
('window--and', 1)	1
('window?', 2)	2
('window', 4)	4
("windpipe's", 1)	1
('winds', 14)	14
('winds,', 1)	1
('winds?', 1)	1
('windsor', 41)	41
('wine.', 8)	8

Lab report screen-shot #18:

mapper2.py — Week10

The screenshot shows a code editor with two tabs: 'mapper2.py' and 'input.txt'. The 'mapper2.py' tab contains the following Python code:

```
#!/usr/bin/env python
import sys
import re
for line in sys.stdin:
    line = line.strip().lower()
    words = line.split()
    for word in words:
        word = re.sub('[^A-Za-z0-9]+', '', word)
        print("%s\t%s" % (word, 1))
```

The 'input.txt' tab contains the following text:

```
This! is a test input.
This is mentioned three times.
However input) is\ not.. mentioned''' three times in this file.
input is) mentioned four times because
here is another input.
```

Below the code editor is a terminal window showing the command and its output:

```
(base) ping58972@Nalongsones-MacBook-Air Week10 % cat input.txt | python3 mapper2.py | sort
a      1
another 1
because 1
file    1
four   1
here   1
however 1
in     1
input   1
input   1
input   1
input   1
is     1
is     1
is     1
is     1
is     1
mentioned 1
mentioned 1
mentioned 1
not    1
test   1
this   1
this   1
three  1
three  1
times  1
times  1
times  1
(base) ping58972@Nalongsones-MacBook-Air Week10 %
```

Lab report screen-shot #19:

The screenshot shows a code editor with two tabs: 'mapper2.py' and 'input.txt'. The 'mapper2.py' tab contains the same Python code as in the previous screenshot:

```
#!/usr/bin/env python
import sys
import re
for line in sys.stdin:
    line = line.strip().lower()
    words = line.split()
    for word in words:
        word = re.sub('[^A-Za-z0-9]+', '', word)
        print("%s\t%s" % (word, 1))
```

The 'input.txt' tab contains the same text as in the previous screenshot:

```
This! is a test input.
This is mentioned three times.
However input) is\ not.. mentioned''' three times in this file.
input is) mentioned four times because
here is another input.
```

Below the code editor is a terminal window showing the command and its output:

```
(base) ping58972@Nalongsones-MacBook-Air Week10 % cat input.txt | python3 mapper2.py | sort
a      1
another 1
because 1
file    1
four   1
here   1
however 1
in     1
input   1
input   1
input   1
input   1
is     1
is     1
is     1
is     1
is     1
mentioned 1
mentioned 1
mentioned 1
not    1
test   1
this   1
this   1
three  1
three  1
times  1
times  1
times  1
(base) ping58972@Nalongsones-MacBook-Air Week10 %
```

Lab report screen-shot #20:

```
input.txt U mapper2.py U ... mapper2.py U input.txt U
❸ mapper2.py > ...
1  #!/usr/bin/env python
2
3  import sys
4  import re
5
6  for line in sys.stdin:
7      line = line.strip().lower()
8      words = line.split()
9      for word in words:
10         word = re.sub('[^A-Za-z0-9]+', '', word)
11         print("%s\t%s" % (word, 1))

PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE
is      1
mentioned 1
mentioned 1
mentioned 1
not      1
test     1
this     1
this     1
this     1
three    1
three    1
times   1
times   1
times   1
(base) ping58972@Nalongsones-MacBook-Air Week10 % cat input.txt | python3 mapper2.py | sort | python3 reducer.py
a 1
another 1
because 1
file 1
four 1
here 1
however 1
in 1
input 4
is 5
mentioned 3
not 1
test 1
this 3
three 2
times 3
(base) ping58972@Nalongsones-MacBook-Air Week10 %
```

Lab report screen-shot #21:

```
input.txt U mapper2.py U mapper.py U ... mapper2.py U input.txt U reducer.py U ...
```

```
mapper2.py > ...
1 #!/usr/bin/env python
2
3 import sys
4 import re
5
6 for line in sys.stdin:
7     line = line.strip().lower()
8     words = line.split()
9     for word in words:
10         word = re.sub('[^A-Za-z0-9]+', '', word)
11         if(len(word) > 0):
12             print("%s\t%s" % (word, 1))
13

reducer.py > ...
7
8 for line in sys.stdin:
9     line = line.strip()
10    (word, count) = line.split('\t')
11    count = int(count)
12    if current_word == word:
13        current_count = current_count + 1
14    else:
15        if current_word:
16            print(current_word, current_count)
17            current_count = count
18            current_word = word
19    if current_word == word:
20        print(current_word, current_count)
21
```

PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE

```
your 6869
yourbut 1
youre 58
yours 266
yoursand 1
yourself 293
yourselfs 2
yourselves 74
yourwife 1
youspare 1
youth 296
youthat 2
youthful 32
youths 12
youve 5
youwell 2
youwhy 1
yravished 1
yslaked 1
zanies 1
zany 1
zeal 34
zealous 6
zeals 1
zed 1
zenelophon 1
zenith 1
zephyrs 1
zir 2
zo 1
zodiac 1
zodiacs 1
zone 1
zounds 20
zwaggered 1
```

```
(base) ping58972@Nalongsones-MacBook-Air Week10 % cat shakespeare/* | python3 mapper2.py | sort | python3 reducer.py
```

Lab report screen-shot 22:

	mapper.py	mapper2.py
How many data-local map tasks were executed?	4	6
How many rack-local map tasks were executed?	6	4
How many reduce tasks?	3	3
Total time spent by all maps?	4296096	4356480
Total time spent by all reduce tasks?	9677	9159
How many map-input records?	173126	173126
How many map output records?	939236	938573
How many reduce input groups?	62933	28941

```
hadoop@ip-172-31-90-37:~/danddankwordcountonmaster
hadoop@ip-172-31-90-37:~/danddankwordcountonmaster (ssh)  #1 | -zsh  #2 | +
FILE: Number of bytes written=4667030
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5647788
HDFS: Number of bytes written=485505
HDFS: Number of read operations=39
HDFS: Number of large read operations=0
HDFS: Number of write operations=6
Job Counters
    Killed reduce tasks=1
    Launched map tasks=10
    Launched reduce tasks=3
    Data-local map tasks=6
    Rack-local map tasks=4
    Total time spent by all maps in occupied slots (ms)=4356480
    Total time spent by all reduces in occupied slots (ms)=1758528
    Total time spent by all map tasks (ms)=45380
    Total time spent by all reduce tasks (ms)=9159
    Total vcore-milliseconds taken by all map tasks=45380
    Total vcore-milliseconds taken by all reduce tasks=9159
    Total megabyte-milliseconds taken by all map tasks=139407360
    Total megabyte-milliseconds taken by all reduce tasks=56272896
Map-Reduce Framework
    Map input records=173126
    Map output records=938573
    Map output bytes=6777584
    Map output materialized bytes=1104519
    Input split bytes=1493
    Combine input records=0
    Combine output records=0
    Reduce input groups=28941
    Reduce shuffle bytes=1104519
    Reduce input records=938573
    Reduce output records=28941
    Spilled Records=1877146
    Shuffled Maps =30
    Failed Shuffles=0
    Merged Map outputs=30
    GC time elapsed (ms)=1442
    CPU time spent (ms)=28690
    Physical memory (bytes) snapshot=6431883264
    Virtual memory (bytes) snapshot=68598767616
    Total committed heap usage (bytes)=5818548224
Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
File Input Format Counters
    Bytes Read=5646295
File Output Format Counters
    Bytes Written=485505
21/10/31 04:47:48 INFO streaming.StreamJob: Output directory: danddankcountonhdfs/output2
```

Lab report screen-shot #23:

The three terminal windows show the following word frequency counts:

- Terminal 1 (hadoop@ip-172-31-90-37):**

```
('wiles', 3)
('will', 5166)
('willeth', 2)
('willful', 1)
('willfully', 1)
('willing', 40)
('willingly', 35)
('willowtree', 1)
('willtis', 1)
('wiltshire', 7)
('winchesters', 2)
('wincot', 1)
('wind', 196)
('winding', 5)
('windingsheet', 2)
('windinstrument', 1)
('windowbars', 1)
('windowd', 2)
('windows', 21)
('windshaked', 1)
('wing', 28)
('winked', 1)
('winkst', 1)
('wins', 9)
('wipe', 26)
('wiping', 1)
```
- Terminal 2 (hadoop@ip-172-31-90-37):**

```
('wills', 32)
('wiltshires', 1)
('wince', 2)
('winchester', 69)
('windchanging', 1)
('window', 31)
('windpipes', 1)
('windring', 1)
('winds', 56)
('windshaken', 1)
('windsor', 56)
('windswift', 1)
('windy', 9)
('wine', 85)
('wingd', 5)
('wingham', 1)
('winghave', 1)
('wings', 62)
('wink', 32)
('winking', 11)
('winners', 2)
('winning', 4)
('winnowd', 3)
('winnowed', 1)
('winnows', 1)
('winters', 46)
```
- Terminal 3 (hadoop@ip-172-31-90-37):**

```
('willgive', 1)
('william', 85)
('willingness', 3)
('willingst', 1)
('willt', 19)
('willthough', 1)
('wilt', 334)
('win', 125)
('winded', 4)
('windinstruments', 1)
('windlasses', 1)
('windmill', 2)
('windobeying', 1)
('windowand', 1)
('windpipe', 1)
('wingdalls', 1)
('winged', 15)
('wingfield', 1)
('winkd', 1)
('winks', 4)
('winner', 5)
('winnow', 1)
('winter', 55)
('wintercricket', 1)
('winterground', 1)
```

The AWS EMR Cluster details page shows the following information for the cluster `dandankcluster`:

- Cluster Status:** Terminating (Terminated by user request)
- Summary:**
 - ID: j-1WSH3RPZJYSMZ
 - Creation date: 2021-10-30 16:09 (UTC-5)
 - Elapsed time: 7 hours, 56 minutes
 - After last step completes: Cluster waits
 - Termination protection: Off
 - Tags: --
 - Master public DNS: ec2-3-86-140-85.compute-1.amazonaws.com
 - Connect to the Master Node Using SSH
- Configuration details:**
 - Release label: emr-5.32.0
 - Hadoop distribution: Amazon 2.10.1
 - Applications: Hive 2.3.7, Pig 0.17.0, Hue 4.8.0, JupyterHub 1.1.0, Oozie 5.2.0, HBase 1.4.13, Spark 2.4.7, Sqoop 1.4.7
 - Log URL: s3://aws-logs-195580907486-us-east-1/elastictmapreduce/
 - EMRFS consistent view: Disabled
 - Custom AMI ID: --
- Network and hardware:**
 - Availability zone: us-east-1d
 - Subnet ID: subnet-86dcf5a7
 - Master: Running 1 m5.xlarge
 - Core: Running 2 m5.xlarge
 - Task: --
 - Cluster scaling: Not enabled
 - Auto-termination: Not enabled
- Security and access:**
 - Key name: awsemrkeypair
 - EC2 instance profile: EMR_EC2_DefaultRole
 - EMR role: EMR_DefaultRole
 - Auto Scaling role: EMR_AutoScaling_DefaultRole
 - Visible to all users: All
 - Security groups for Master: sg-06997c6c77ce1e499 (ElasticMapReduce-master)
 - Security groups for Core & Task: sg-0e54d5335beb16e3f (ElasticMapReduce-Task: slave)