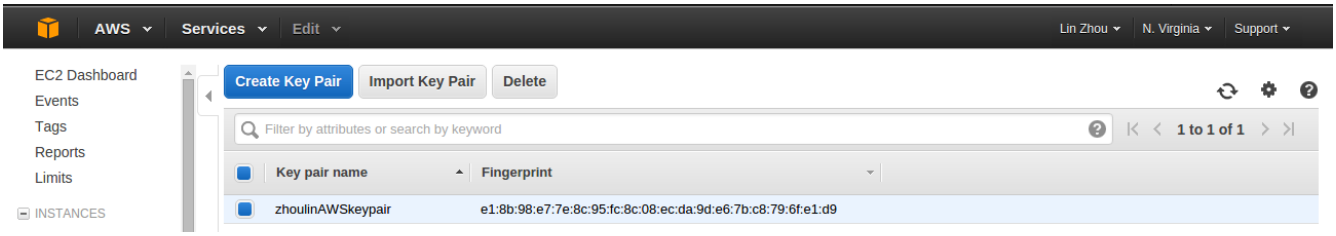
Launch a cluster on EC2:

1. EC2 keypair: create EC2 keypair in order to SSH into a master or slave instances in a Spark cluster. Download private key to local machine and set its permissions to 400.



2. Set environment variables <ACCESS_KEY_ID> <SECRET_ACCESS_KEY>.

export AWS_ACCESS_KEY_ID=...
export AWS_SECRET_ACCESS_KEY=...

3. Launch a Spark cluster with 4 slave nodes and 1 master node.

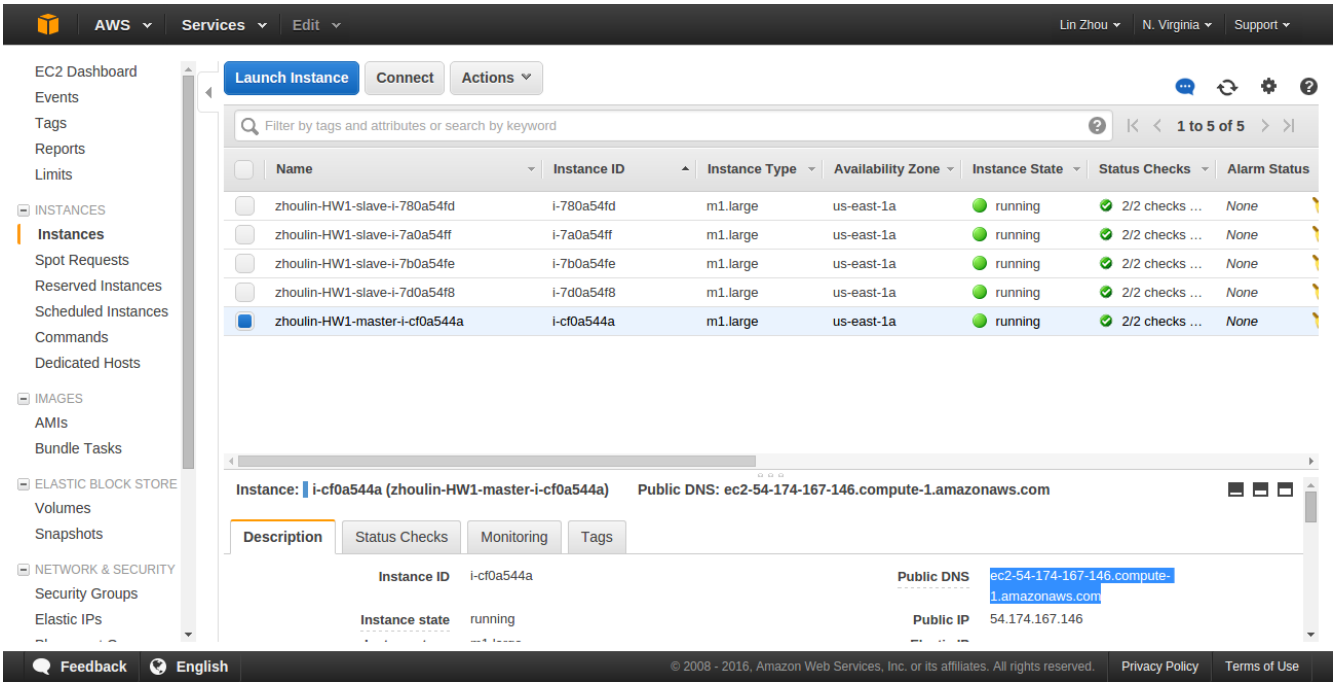
cd .../spark/ec2/spark-ec2 -k <name of EC2 keypair> -i <path of .pem> -s <num-slaves> -r us-east-1 -z us-east-1a launch <cluster-name>
<name of EC2 keypair>: zhoulinAWSkeypair
<path of .pem>: $DOCUMENTS/AWS/zhoulinAWSkeypair.pem
<num-slaves>: 4
<cluster-name>: zhoulin-HW1

```
linux@ZhouLin:/usr/local/spark/ec2$ ./spark-ec2 -k zhoulinAWSkeypair -i $DOCUMENTS/AWS/zhoulinAWSkeypair.pem -s 4 -r us-east-1 -z us-east-1a la
unch zhoulin-HW1
Setting up security groups...
Searching for existing cluster zhoulin-HW1 in region us-east-1...
Spark AMI: ami-5bb18832
Launching instances...
Launched 4 slaves in us-east-1a, regid = r-4dfa1e9f
Launched master in us-east-1a, regid = r-cbfb1f19
Waiting for AWS to propagate instance metadata...
Waiting for cluster to enter 'ssh-ready' state.........
```

master node public DNS: ec2-54-174-167-146.compute-1.amazonaws.com

```
Connection to ec2-54-174-167-146.compute-1.amazonaws.com closed.
Spark standalone cluster started at http://ec2-54-174-167-146.compute-1.amazonaws.com:8080
Ganglia started at http://ec2-54-174-167-146.compute-1.amazonaws.com:5080/ganglia
Done!
```

AWS EC2 console:



4. Copy input (10 bibles) file to the master instance.

scp -i <path to .pem file> <path to trasformed file> ec2-user@<Public DNS>:/home/ec2-user/

```
linux@ZhouLin:~$ scp -i $DOCUMENTS/AWS/zhoulinAWSkeypair.pem $DOCUMENTS/HW1/input/bible+shakes+*.nopunc ec2-user@ec2-54-174-167-146.compute-1.a
mazonaws.com:/home/ec2-user/
The authenticity of host 'ec2-54-174-167-146.compute-1.amazonaws.com (54.174.167.146)' can't be established.
ECDSA key fingerprint is ad:ec:bd:64:e0:d6:25:e5:22:b5:0a:ac:df:e4:9d:d5.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ec2-54-174-167-146.compute-1.amazonaws.com,54.174.167.146' (ECDSA) to the list of known hosts.
bible+shakes+10.nopunc                                                                                100% 8856KB   8.7MB/s   00:01
bible+shakes+1.nopunc                                                                                 100% 8856KB   8.7MB/s   00:01
bible+shakes+2.nopunc                                                                                 100% 8856KB   4.3MB/s   00:02
bible+shakes+3.nopunc                                                                                 100% 8856KB   8.7MB/s   00:01
bible+shakes+4.nopunc                                                                                 100% 8856KB   8.7MB/s   00:01
bible+shakes+5.nopunc                                                                                 100% 8856KB   8.7MB/s   00:01
bible+shakes+6.nopunc                                                                                 100% 8856KB   8.7MB/s   00:01
bible+shakes+7.nopunc                                                                                 100% 8856KB   8.7MB/s   00:01
bible+shakes+8.nopunc                                                                                 100% 8856KB   8.7MB/s   00:01
bible+shakes+9.nopunc                                                                                 100% 8856KB   8.7MB/s   00:01
```

5. SSH into the master instances.

ssh -i <path to .pem file> ec2-user@<Public DNS>

```
linux@ZhouLin:~$ ssh -i $DOCUMENTS/AWS/zhoulinAWSkeypair.pem ec2-user@ec2-54-174-167-146.compute-1.amazonaws.com
Last login: Wed Apr 17 21:59:34 2013 from c-76-21-41-48.hsd1.ca.comcast.net

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

https://aws.amazon.com/amazon-linux-ami/2013.03-release-notes/
Amazon Linux version 2015.09 is available.
```

6. Upload input file to master node HDFS.

sudo /root/ephemeral-hdfs/bin/hadoop fs -mkdir /user/zhoulin/input
sudo /root/ephemeral-hdfs/bin/hadoop fs -put /home/ec2-user/input/* /user/zhoulin/input/

```
[ec2-user@ip-172-31-4-253 ~]$ sudo /root/ephemeral-hdfs/bin/hadoop fs -put /home/ec2-user/bible+shakes+*.nopunc /user/zhoulin/input/
Warning: $HADOOP_HOME is deprecated.

[ec2-user@ip-172-31-4-253 ~]$ sudo /root/ephemeral-hdfs/bin/hadoop fs -ls /user/zhoulin/input
Warning: $HADOOP_HOME is deprecated.

Found 10 items
-rw-r--r--   3 root supergroup    9068074 2016-02-07 00:41 /user/zhoulin/input/bible+shakes+1.nopunc
-rw-r--r--   3 root supergroup    9068074 2016-02-07 00:41 /user/zhoulin/input/bible+shakes+10.nopunc
-rw-r--r--   3 root supergroup    9068074 2016-02-07 00:41 /user/zhoulin/input/bible+shakes+2.nopunc
-rw-r--r--   3 root supergroup    9068074 2016-02-07 00:41 /user/zhoulin/input/bible+shakes+3.nopunc
-rw-r--r--   3 root supergroup    9068074 2016-02-07 00:41 /user/zhoulin/input/bible+shakes+4.nopunc
-rw-r--r--   3 root supergroup    9068074 2016-02-07 00:41 /user/zhoulin/input/bible+shakes+5.nopunc
-rw-r--r--   3 root supergroup    9068074 2016-02-07 00:41 /user/zhoulin/input/bible+shakes+6.nopunc
-rw-r--r--   3 root supergroup    9068074 2016-02-07 00:41 /user/zhoulin/input/bible+shakes+7.nopunc
-rw-r--r--   3 root supergroup    9068074 2016-02-07 00:41 /user/zhoulin/input/bible+shakes+8.nopunc
-rw-r--r--   3 root supergroup    9068074 2016-02-07 00:41 /user/zhoulin/input/bible+shakes+9.nopunc
```

7. Run applications.

/root/spark/bin/spark-submit /home/ec2-user/wordCount.py

8. Check output.

sudo /root/ephemeral-hdfs/bin/hadoop fs -ls /user/zhoulin/output

```
[ec2-user@ip-172-31-4-253 ~]$ sudo /root/ephemeral-hdfs/bin/hadoop fs -ls /user/zhoulin
Warning: $HADOOP_HOME is deprecated.

Found 5 items
drwxr-xr-x   - root     supergroup          0 2016-02-07 01:00 /user/zhoulin/cache
drwxr-xr-x   - root     supergroup          0 2016-02-07 00:41 /user/zhoulin/input
drwxr-xr-x   - ec2-user supergroup          0 2016-02-07 00:52 /user/zhoulin/output-EX1
drwxr-xr-x   - ec2-user supergroup          0 2016-02-07 00:44 /user/zhoulin/output-EX2
drwxr-xr-x   - ec2-user supergroup          0 2016-02-07 01:23 /user/zhoulin/output-EX3
```

9. Copy files from HDFS to master node.

sudo /root/ephemeral-hdfs/bin/hadoop fs -get /user/zhoulin/output/* /home/ec2-user/output

10. Copy output files from remote machine to local machine.

scp -i /path/my-key-pair.pem ec2-user@<public DNS>:home/ec2-user/File.txt $HOME/dir

```
linux@ZhouLin:/usr/local/spark/ec2$ scp -i $DOCUMENTS/AWS/zhoulinAWSkeypair.pem ec2-user@ec2-52-90-150-15.compute-1.amazonaws.com:/home/ec2-use
r/output-EX1/* $DESKTOP/HW1-spark
part-00000                                                                                     100%   72KB  72.3KB/s   00:00
part-00001                                                                                     100%   73KB  72.5KB/s   00:01
part-00002                                                                                     100%   71KB  70.9KB/s   00:00
part-00003                                                                                     100%   74KB  73.6KB/s   00:00
part-00004                                                                                     100%   72KB  72.2KB/s   00:00
part-00005                                                                                     100%   72KB  71.5KB/s   00:01
part-00006                                                                                     100%   71KB  71.1KB/s   00:00
part-00007                                                                                     100%   72KB  72.3KB/s   00:00
part-00008                                                                                     100%   73KB  73.3KB/s   00:00
part-00009                                                                                     100%   73KB  72.7KB/s   00:00
_SUCCESS                                                                                       100%    0     0.0KB/s   00:00
```

EX1: one-word frequency

Copy wordCount1.py from local machine to master instance

```
linux@ZhouLin:~$ scp -i $DOCUMENTS/AWS/zhoulinAWSkeypair.pem $DESKTOP/HW1/wordCount1/wordCount1.py ec2-user@ec2-54-174-167-146.compute-1.amazon
aws.com:/home/ec2-user/
wordCount1.py                                                                                 100% 1048     1.0KB/s   00:00
```

HDFS result file

```
[ec2-user@ip-172-31-4-253 ~]$ sudo /root/ephemeral-hdfs/bin/hadoop fs -ls /user/zhoulin/output-EX1
Warning: $HADOOP_HOME is deprecated.

Found 11 items
-rw-r--r--   3 ec2-user supergroup          0 2016-02-07 00:52 /user/zhoulin/output-EX1/_SUCCESS
-rw-r--r--   3 ec2-user supergroup      74002 2016-02-07 00:52 /user/zhoulin/output-EX1/part-00000
-rw-r--r--   3 ec2-user supergroup      74253 2016-02-07 00:52 /user/zhoulin/output-EX1/part-00001
-rw-r--r--   3 ec2-user supergroup      72628 2016-02-07 00:52 /user/zhoulin/output-EX1/part-00002
-rw-r--r--   3 ec2-user supergroup      75392 2016-02-07 00:52 /user/zhoulin/output-EX1/part-00003
-rw-r--r--   3 ec2-user supergroup      73922 2016-02-07 00:52 /user/zhoulin/output-EX1/part-00004
-rw-r--r--   3 ec2-user supergroup      73233 2016-02-07 00:52 /user/zhoulin/output-EX1/part-00005
-rw-r--r--   3 ec2-user supergroup      72773 2016-02-07 00:52 /user/zhoulin/output-EX1/part-00006
-rw-r--r--   3 ec2-user supergroup      74011 2016-02-07 00:52 /user/zhoulin/output-EX1/part-00007
-rw-r--r--   3 ec2-user supergroup      75052 2016-02-07 00:52 /user/zhoulin/output-EX1/part-00008
-rw-r--r--   3 ec2-user supergroup      74444 2016-02-07 00:52 /user/zhoulin/output-EX1/part-00009
```

Elapse time: approximately 2s

```
16/02/07 00:52:44 INFO scheduler.TaskSetManager: Starting task 8.0 in stage 1.0 (TID 18, ip-172-31-15-3.ec2.internal, partition 8,NODE_LOCAL, 1
949 bytes)
16/02/07 00:52:44 INFO scheduler.TaskSetManager: Finished task 2.0 in stage 1.0 (TID 12) in 1032 ms on ip-172-31-15-3.ec2.internal (1/10)
16/02/07 00:52:44 INFO scheduler.TaskSetManager: Starting task 9.0 in stage 1.0 (TID 19, ip-172-31-15-3.ec2.internal, partition 9,NODE_LOCAL, 1
949 bytes)
16/02/07 00:52:44 INFO scheduler.TaskSetManager: Finished task 6.0 in stage 1.0 (TID 16) in 1036 ms on ip-172-31-15-3.ec2.internal (2/10)
16/02/07 00:52:44 INFO scheduler.TaskSetManager: Finished task 3.0 in stage 1.0 (TID 13) in 1047 ms on ip-172-31-15-6.ec2.internal (3/10)
16/02/07 00:52:44 INFO scheduler.TaskSetManager: Finished task 7.0 in stage 1.0 (TID 17) in 1044 ms on ip-172-31-15-6.ec2.internal (4/10)
16/02/07 00:52:44 INFO scheduler.TaskSetManager: Finished task 5.0 in stage 1.0 (TID 15) in 1062 ms on ip-172-31-15-4.ec2.internal (5/10)
16/02/07 00:52:44 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 1.0 (TID 11) in 1076 ms on ip-172-31-15-4.ec2.internal (6/10)
16/02/07 00:52:44 INFO scheduler.TaskSetManager: Finished task 4.0 in stage 1.0 (TID 14) in 1138 ms on ip-172-31-15-5.ec2.internal (7/10)
16/02/07 00:52:44 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 10) in 1152 ms on ip-172-31-15-5.ec2.internal (8/10)
16/02/07 00:52:45 INFO scheduler.TaskSetManager: Finished task 9.0 in stage 1.0 (TID 19) in 452 ms on ip-172-31-15-3.ec2.internal (9/10)
16/02/07 00:52:45 INFO scheduler.TaskSetManager: Finished task 8.0 in stage 1.0 (TID 18) in 472 ms on ip-172-31-15-3.ec2.internal (10/10)
16/02/07 00:52:45 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
16/02/07 00:52:45 INFO scheduler.DAGScheduler: ResultStage 1 (saveAsTextFile at NativeMethodAccessorImpl.java:-2) finished in 1.506 s
16/02/07 00:52:45 INFO scheduler.DAGScheduler: Job 0 finished: saveAsTextFile at NativeMethodAccessorImpl.java:-2, took 15.340848 s
EX1 elapsed_time is:0:00:01.664955s
```

EX2: double-word frequency

Upload file from local machine to master node:

```
linux@ZhouLin:~$ scp -i $DOCUMENTS/AWS/zhoulinAWSkeypair.pem $DESKTOP/HW1/wordCount2/wordCount2.py ec2-user@ec2-54-174-167-146.compute-1.amazon
aws.com:/home/ec2-user/
wordCount2.py                                                                                 100% 2038     2.0KB/s   00:00
```

HDFS result:

```
[ec2-user@ip-172-31-4-253 bin]$ sudo /root/ephemeral-hdfs/bin/hadoop fs -ls /user/zhoulin/output-EX2
Warning: $HADOOP_HOME is deprecated.

Found 9 items
-rw-r--r--   3 ec2-user supergroup          0 2016-02-07 00:44 /user/zhoulin/output-EX2/_SUCCESS
-rw-r--r--   3 ec2-user supergroup    1335511 2016-02-07 00:44 /user/zhoulin/output-EX2/part-00000
-rw-r--r--   3 ec2-user supergroup    1346694 2016-02-07 00:44 /user/zhoulin/output-EX2/part-00001
-rw-r--r--   3 ec2-user supergroup    1337774 2016-02-07 00:44 /user/zhoulin/output-EX2/part-00002
-rw-r--r--   3 ec2-user supergroup    1343236 2016-02-07 00:44 /user/zhoulin/output-EX2/part-00003
-rw-r--r--   3 ec2-user supergroup    1324474 2016-02-07 00:44 /user/zhoulin/output-EX2/part-00004
-rw-r--r--   3 ec2-user supergroup    1345121 2016-02-07 00:44 /user/zhoulin/output-EX2/part-00005
-rw-r--r--   3 ec2-user supergroup    1333937 2016-02-07 00:44 /user/zhoulin/output-EX2/part-00006
-rw-r--r--   3 ec2-user supergroup    1351395 2016-02-07 00:44 /user/zhoulin/output-EX2/part-00007
```

Elapse time: 58s

```
16/02/07 00:44:31 INFO spark.MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to ip-172-31-15-6.ec2.internal:48
379
16/02/07 00:44:35 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 2.0 (TID 19) in 4846 ms on ip-172-31-15-3.ec2.internal (1/8)
16/02/07 00:44:36 INFO scheduler.TaskSetManager: Finished task 7.0 in stage 2.0 (TID 25) in 5098 ms on ip-172-31-15-5.ec2.internal (2/8)
16/02/07 00:44:36 INFO scheduler.TaskSetManager: Finished task 4.0 in stage 2.0 (TID 22) in 5108 ms on ip-172-31-15-4.ec2.internal (3/8)
16/02/07 00:44:36 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 18) in 5155 ms on ip-172-31-15-4.ec2.internal (4/8)
16/02/07 00:44:36 INFO scheduler.TaskSetManager: Finished task 2.0 in stage 2.0 (TID 20) in 5187 ms on ip-172-31-15-6.ec2.internal (5/8)
16/02/07 00:44:36 INFO scheduler.TaskSetManager: Finished task 6.0 in stage 2.0 (TID 24) in 5188 ms on ip-172-31-15-6.ec2.internal (6/8)
16/02/07 00:44:36 INFO scheduler.TaskSetManager: Finished task 5.0 in stage 2.0 (TID 23) in 5194 ms on ip-172-31-15-3.ec2.internal (7/8)
16/02/07 00:44:36 INFO scheduler.TaskSetManager: Finished task 3.0 in stage 2.0 (TID 21) in 5221 ms on ip-172-31-15-5.ec2.internal (8/8)
16/02/07 00:44:36 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
16/02/07 00:44:36 INFO scheduler.DAGScheduler: ResultStage 2 (saveAsTextFile at NativeMethodAccessorImpl.java:-2) finished in 5.227 s
16/02/07 00:44:36 INFO scheduler.DAGScheduler: Job 1 finished: saveAsTextFile at NativeMethodAccessorImpl.java:-2, took 40.036064 s
EX2 elapsed_time is:0:00:58.006611s
```

EX3: find an exact word in a file

Upload file from local machine to master node:

```
linux@ZhouLin:~$ scp -i $DOCUMENTS/AWS/zhoulinAWSkeypair.pem $DESKTOP/HW1/wordCount3/wordCount3.py ec2-user@ec2-54-174-167-146.compute-1.amazon
aws.com:/home/ec2-user/
wordCount3.py                                                                                 100% 1410     1.4KB/s   00:00
```

Upload cache file from local machine to master node:

```
linux@ZhouLin:~$ scp -i $DOCUMENTS/AWS/zhoulinAWSkeypair.pem $DESKTOP/HW1_handin/wordCount_3/cache_input/* ec2-user@ec2-54-174-167-146.compute-
1.amazonaws.com:/home/ec2-user/
cache_input                                                                                   100%  628     0.6KB/s   00:00
```

HDFS result:

```
[ec2-user@ip-172-31-4-253 ~]$ sudo /root/ephemeral-hdfs/bin/hadoop fs -ls /user/zhoulin/output-EX3
Warning: $HADOOP_HOME is deprecated.

Found 9 items
-rw-r--r--   3 ec2-user supergroup          0 2016-02-07 01:23 /user/zhoulin/output-EX3/_SUCCESS
-rw-r--r--   3 ec2-user supergroup        144 2016-02-07 01:23 /user/zhoulin/output-EX3/part-00000
-rw-r--r--   3 ec2-user supergroup        222 2016-02-07 01:23 /user/zhoulin/output-EX3/part-00001
-rw-r--r--   3 ec2-user supergroup        191 2016-02-07 01:23 /user/zhoulin/output-EX3/part-00002
-rw-r--r--   3 ec2-user supergroup         90 2016-02-07 01:23 /user/zhoulin/output-EX3/part-00003
-rw-r--r--   3 ec2-user supergroup        224 2016-02-07 01:23 /user/zhoulin/output-EX3/part-00004
-rw-r--r--   3 ec2-user supergroup        248 2016-02-07 01:23 /user/zhoulin/output-EX3/part-00005
-rw-r--r--   3 ec2-user supergroup        174 2016-02-07 01:23 /user/zhoulin/output-EX3/part-00006
-rw-r--r--   3 ec2-user supergroup        206 2016-02-07 01:23 /user/zhoulin/output-EX3/part-00007
```

Elapse time: 22 min

```
16/02/07 01:23:30 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 3.0 (TID 20) in 520 ms on ip-172-31-15-3.ec2.internal (1/8)
16/02/07 01:23:30 INFO scheduler.TaskSetManager: Finished task 4.0 in stage 3.0 (TID 24) in 537 ms on ip-172-31-15-3.ec2.internal (2/8)
16/02/07 01:23:30 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 3.0 (TID 23) in 544 ms on ip-172-31-15-4.ec2.internal (3/8)
16/02/07 01:23:30 INFO scheduler.TaskSetManager: Finished task 6.0 in stage 3.0 (TID 26) in 547 ms on ip-172-31-15-6.ec2.internal (4/8)
16/02/07 01:23:30 INFO scheduler.TaskSetManager: Finished task 7.0 in stage 3.0 (TID 27) in 549 ms on ip-172-31-15-4.ec2.internal (5/8)
16/02/07 01:23:30 INFO scheduler.TaskSetManager: Finished task 3.0 in stage 3.0 (TID 22) in 564 ms on ip-172-31-15-6.ec2.internal (6/8)
16/02/07 01:23:30 INFO scheduler.TaskSetManager: Finished task 2.0 in stage 3.0 (TID 21) in 601 ms on ip-172-31-15-5.ec2.internal (7/8)
16/02/07 01:23:30 INFO scheduler.TaskSetManager: Finished task 5.0 in stage 3.0 (TID 25) in 602 ms on ip-172-31-15-5.ec2.internal (8/8)
16/02/07 01:23:30 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
16/02/07 01:23:30 INFO scheduler.DAGScheduler: ResultStage 3 (saveAsTextFile at NativeMethodAccessorImpl.java:-2) finished in 0.603 s
16/02/07 01:23:30 INFO scheduler.DAGScheduler: Job 2 finished: saveAsTextFile at NativeMethodAccessorImpl.java:-2, took 5.496226 s
EX1 elapsed_time is:0:21:51.280308s
```

There are 88 words in bibles that are the same as the given text:

```
linux@ZhouLin:~/Desktop/HW1_handin/wordCount_3/EX3-EC2-output$ find . -name 'part-0000*' | xargs wc -l
  15 ./part-00005
  10 ./part-00006
  11 ./part-00002
   5 ./part-00003
  13 ./part-00001
  13 ./part-00004
  12 ./part-00007
   9 ./part-00000
  88 total
```