This assignment is made by Abhishek Parekh UFID 6333-3803 (abhishekparekh@ufl.edu) in regards to submission for the Programming Assignment 1 for EEL 6761 Cloud Computing and Storage.

Task 1 (10 points). Count one-word frequency as in the wordcount example. (This is required. Copy/paste source code is allowed.)

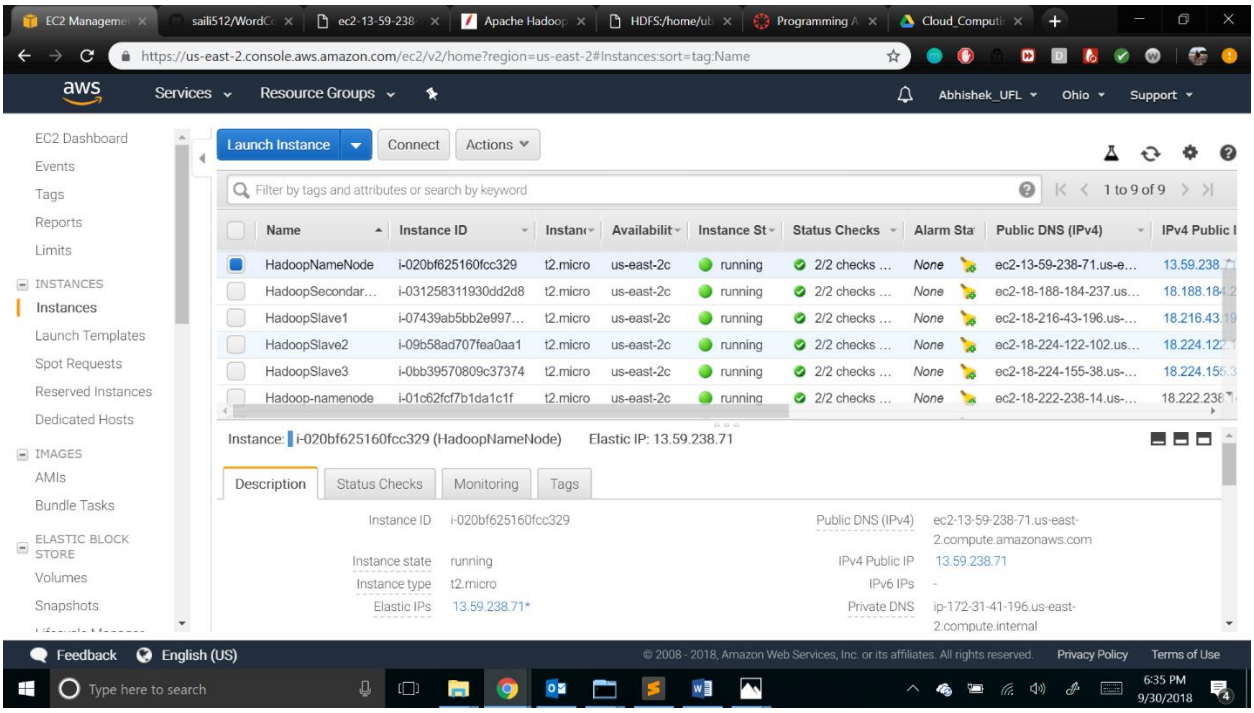
This task has been completed by setting up the Hadoop cluster from the link:

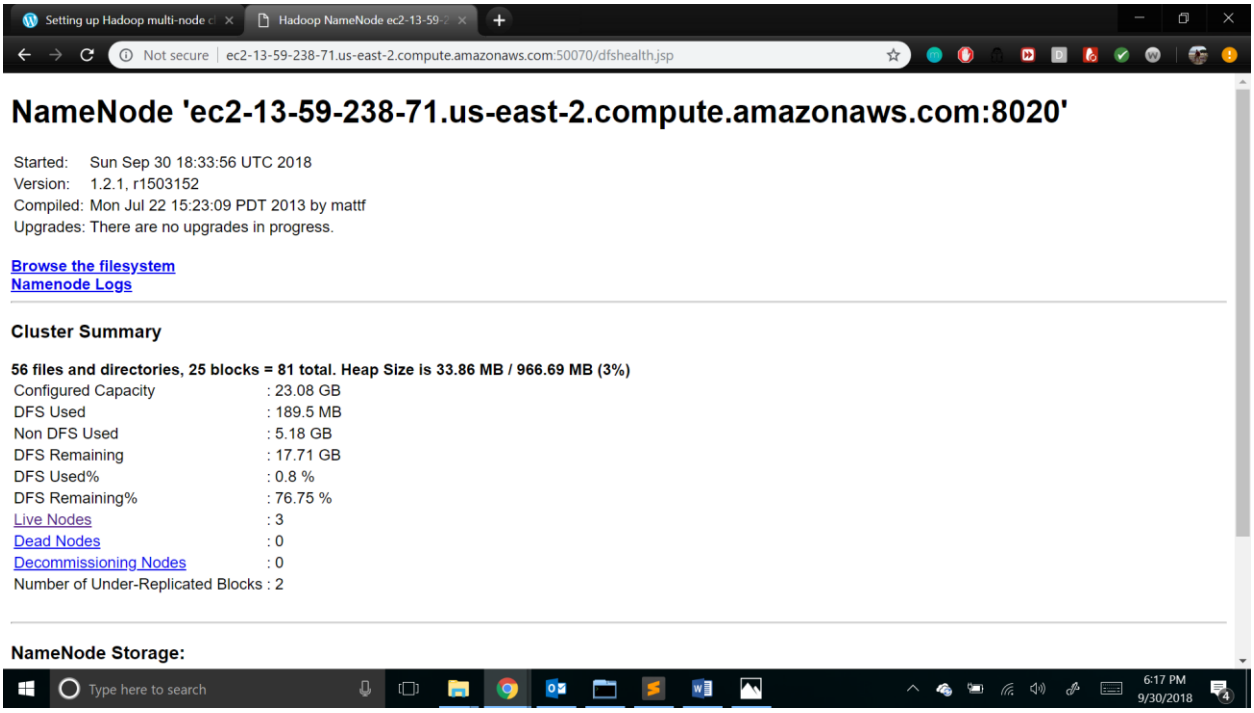
Part 1: https://letsdobigdata.wordpress.com/2014/01/13/setting-up-hadoop-multi-node-cluster-on-amazon-ec2-part-1/

Part 2: https://letsdobigdata.wordpress.com/2014/01/13/setting-up-hadoop-1-2-1-multi-node-cluster-on-amazon-ec2-part-2/

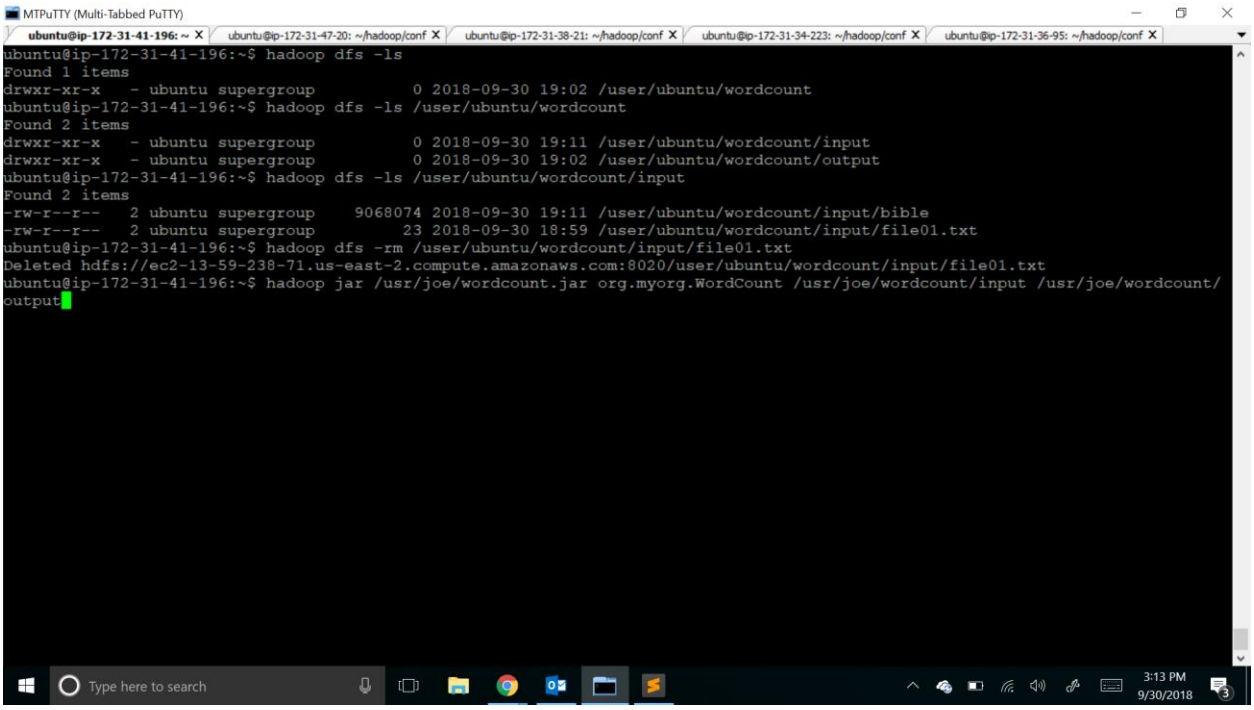
Part 1 is the basic launching up of EC2 instance and setting up key pairs and accessing with putty on windows.

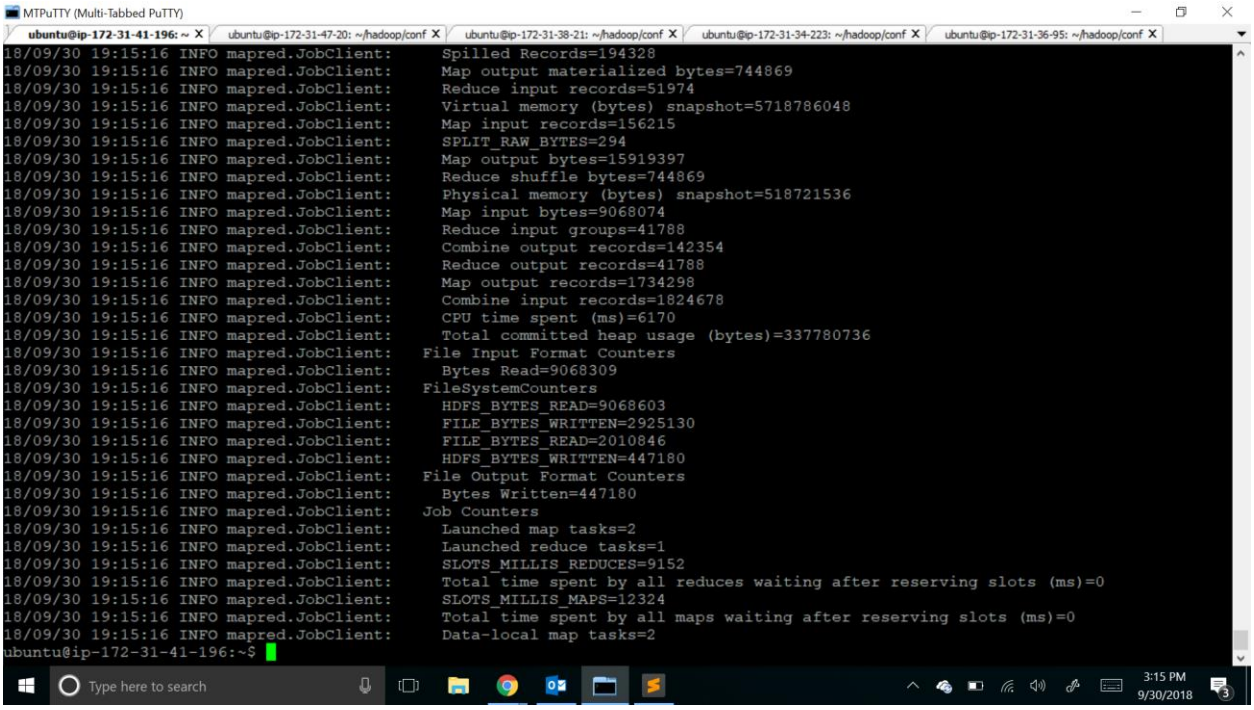
Part 2 involves a more sophisticated approach for setting up a multi-node cluster of Hadoop on EC2.

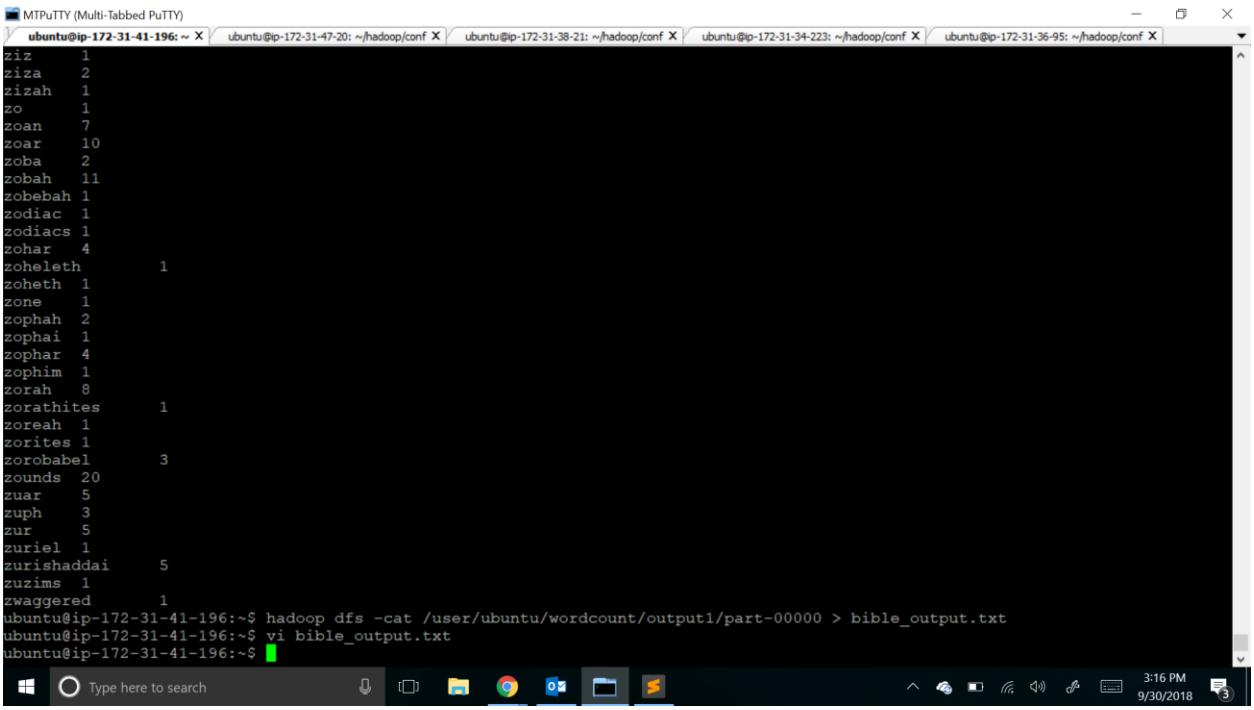


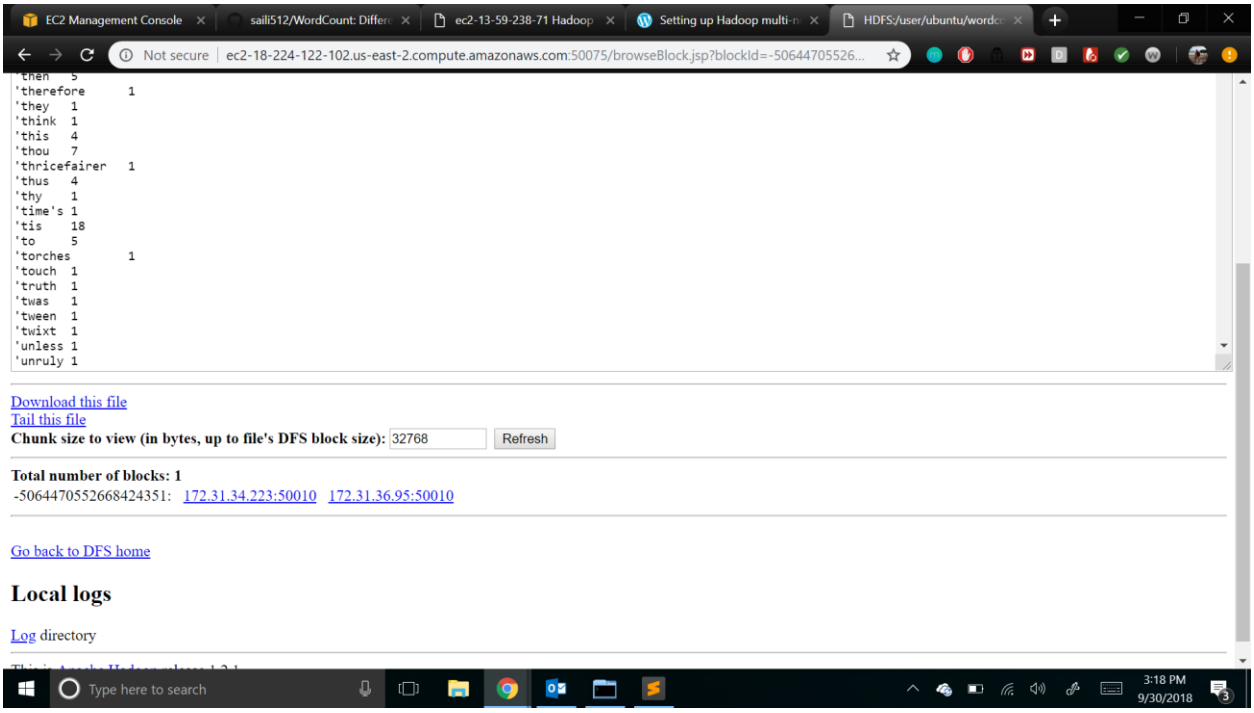


As you can see in the above screenshot the cluster is running version 1.2.1 and there are 3 live nodes.




MTPuTTY (Multi-Tabbed PuTTY)
ubuntu@ip-172-31-41-196: ~
ubuntu@ip-172-31-47-20: ~/hadoop/conf
ubuntu@ip-172-31-38-21: ~/hadoop/conf
ubuntu@ip-172-31-34-223: ~/hadoop/conf
ubuntu@ip-172-31-36-95: ~/hadoop/conf
18/09/30 19:15:16 INFO mapred.JobClient: Spilled Records=194328
18/09/30 19:15:16 INFO mapred.JobClient: Map output materialized bytes=744869
18/09/30 19:15:16 INFO mapred.JobClient: Reduce input records=51974
18/09/30 19:15:16 INFO mapred.JobClient: Virtual memory (bytes) snapshot=5718786048
18/09/30 19:15:16 INFO mapred.JobClient: Map input records=156215
18/09/30 19:15:16 INFO mapred.JobClient: SPLIT_RAW_BYTES=294
18/09/30 19:15:16 INFO mapred.JobClient: Map output bytes=15919397
18/09/30 19:15:16 INFO mapred.JobClient: Reduce shuffle bytes=744869
18/09/30 19:15:16 INFO mapred.JobClient: Physical memory (bytes) snapshot=518721536
18/09/30 19:15:16 INFO mapred.JobClient: Map input bytes=9068074
18/09/30 19:15:16 INFO mapred.JobClient: Reduce input groups=41788
18/09/30 19:15:16 INFO mapred.JobClient: Combine output records=142354
18/09/30 19:15:16 INFO mapred.JobClient: Reduce output records=41788
18/09/30 19:15:16 INFO mapred.JobClient: Map output records=1734298
18/09/30 19:15:16 INFO mapred.JobClient: Combine input records=1824678
18/09/30 19:15:16 INFO mapred.JobClient: CPU time spent (ms)=6170
18/09/30 19:15:16 INFO mapred.JobClient: Total committed heap usage (bytes)=337780736
18/09/30 19:15:16 INFO mapred.JobClient: File Input Format Counters
18/09/30 19:15:16 INFO mapred.JobClient: Bytes Read=9068309
18/09/30 19:15:16 INFO mapred.JobClient: FileSystemCounters
18/09/30 19:15:16 INFO mapred.JobClient: HDFS_BYTES_READ=9068603
18/09/30 19:15:16 INFO mapred.JobClient: FILE_BYTES_WRITTEN=2925130
18/09/30 19:15:16 INFO mapred.JobClient: FILE_BYTES_READ=2010846
18/09/30 19:15:16 INFO mapred.JobClient: HDFS_BYTES_WRITTEN=447180
18/09/30 19:15:16 INFO mapred.JobClient: File Output Format Counters
18/09/30 19:15:16 INFO mapred.JobClient: Bytes Written=447180
18/09/30 19:15:16 INFO mapred.JobClient: Job Counters
18/09/30 19:15:16 INFO mapred.JobClient: Launched map tasks=2
18/09/30 19:15:16 INFO mapred.JobClient: Launched reduce tasks=1
18/09/30 19:15:16 INFO mapred.JobClient: SLOTS_MILLIS_REDUCES=9152
18/09/30 19:15:16 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0
18/09/30 19:15:16 INFO mapred.JobClient: SLOTS_MILLIS_MAPS=12324
18/09/30 19:15:16 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
18/09/30 19:15:16 INFO mapred.JobClient: Data-local map tasks=2
ubuntu@ip-172-31-41-196:~$
Type here to search
3:15 PM
9/30/2018

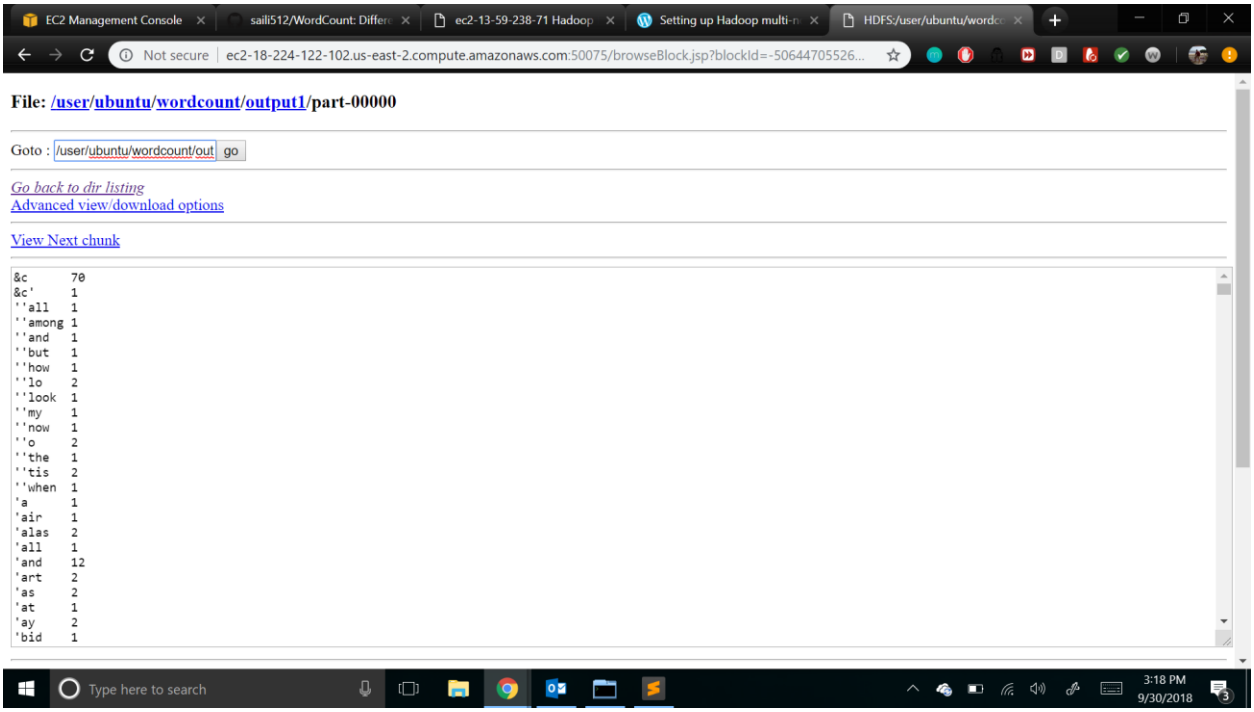

MTPuTTY (Multi-Tabbed PuTTY)
ubuntu@ip-172-31-41-196: ~
ubuntu@ip-172-31-47-20: ~/hadoop/conf
ubuntu@ip-172-31-38-21: ~/hadoop/conf
ubuntu@ip-172-31-34-223: ~/hadoop/conf
ubuntu@ip-172-31-36-95: ~/hadoop/conf
ziz 1
ziza 2
zizah 1
zo 1
zoan 7
zoar 10
zoba 2
zobah 11
zobebah 1
zodiac 1
zodiacs 1
zohar 4
zoheleth 1
zoheth 1
zone 1
zophah 2
zophai 1
zophar 4
zophim 1
zorah 8
zorathites 1
zoreah 1
zorites 1
zorobabel 3
zounds 20
zuar 5
zuph 3
zur 5
zuriel 1
zurishaddai 5
zuzims 1
zwaggered 1
ubuntu@ip-172-31-41-196:~$ hadoop dfs -cat /user/ubuntu/wordcount/output1/part-00000 > bible_output.txt
ubuntu@ip-172-31-41-196:~$ vi bible_output.txt
ubuntu@ip-172-31-41-196:~$
Type here to search
3:16 PM
9/30/2018



https://drive.google.com/open?id=1_3h1qgfGa0ilMxcqNKbcTY-lS4nZswAS

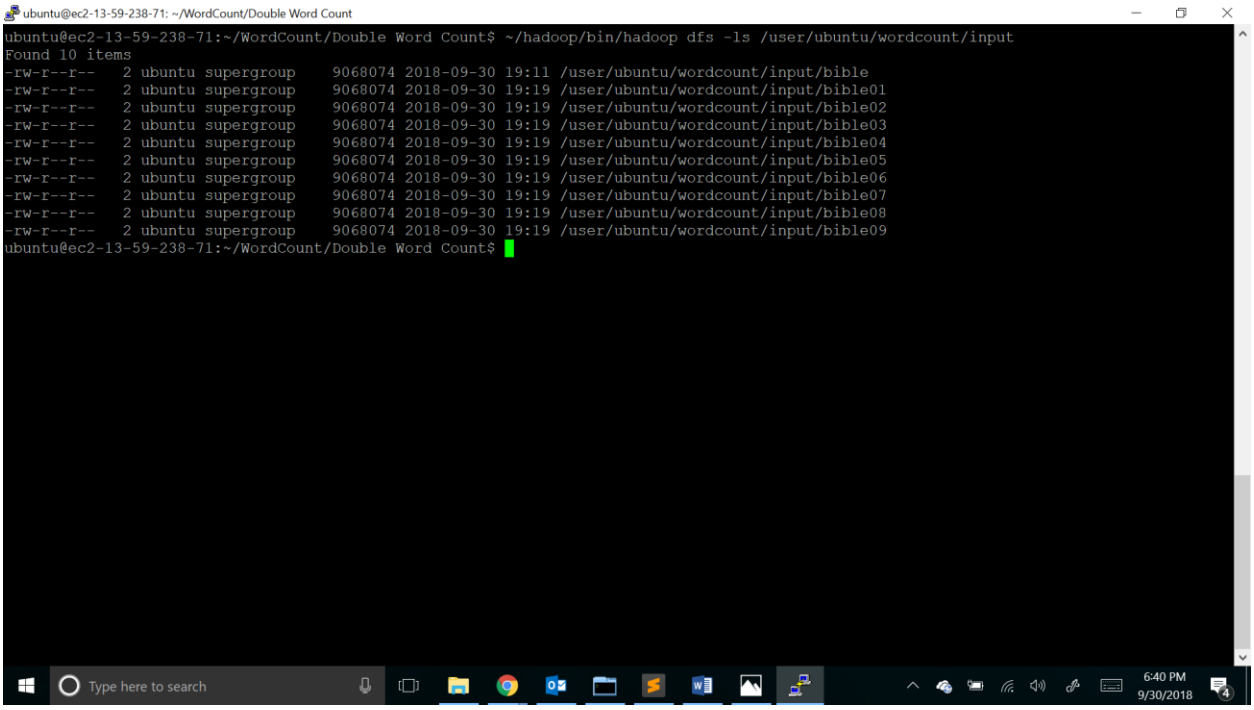
The Screenshot has been shown for wordcount on 1 copy of bible.

The above is the link to a text file where the bible 10 copies are present in the input file and the output is the word count. The code has been copied from word count example in this link https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

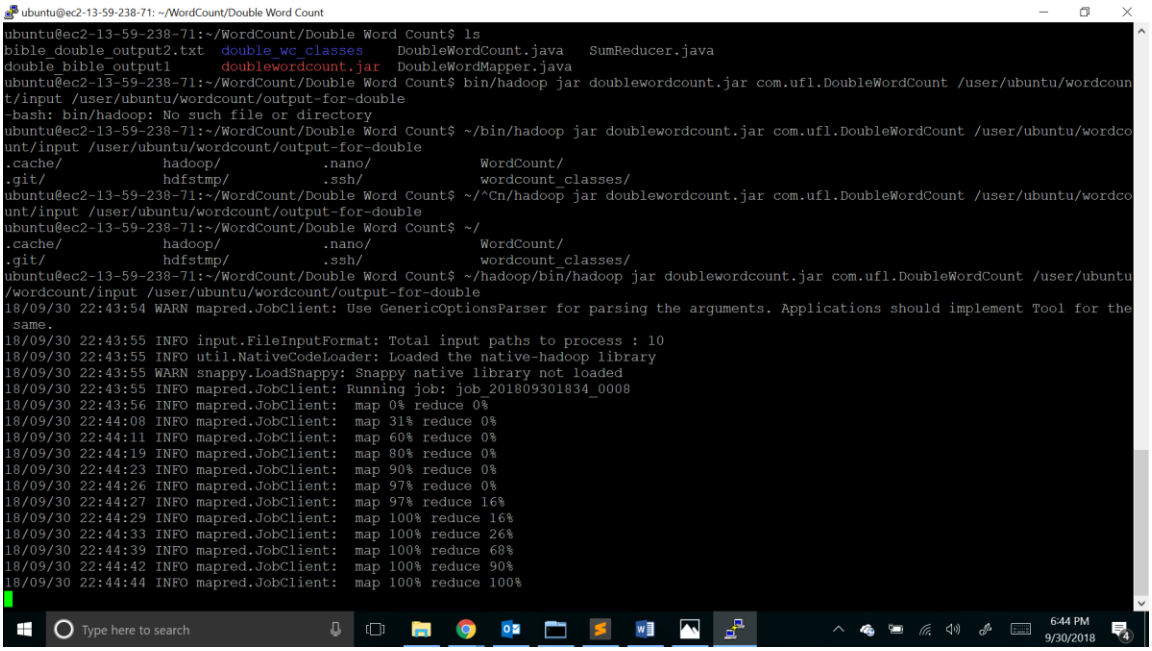

EC2 Management Console
saili512/WordCount: Differ
ec2-13-59-238-71 Hadoop
Setting up Hadoop multi-n
HDFS:/user/ubuntu/wordc
Not secure
ec2-18-224-122-102.us-east-2.compute.amazonaws.com:50075/browseBlock.jsp?blockId=-50644705526...
File: /user/ubuntu/wordcount/output1/part-00000
Goto : /user/ubuntu/wordcount/out go
Go back to dir listing
Advanced view/download options
View Next chunk
&c 70
&c' 1
''all 1
''among 1
''and 1
''but 1
''how 1
''lo 2
''look 1
''my 1
''now 1
''o 2
''the 1
''tis 2
''when 1
'a 1
'air 1
'alas 2
'all 1
'and 12
'art 2
'as 2
'at 1
'ay 2
'bid 1
Type here to search
3:18 PM
9/30/2018

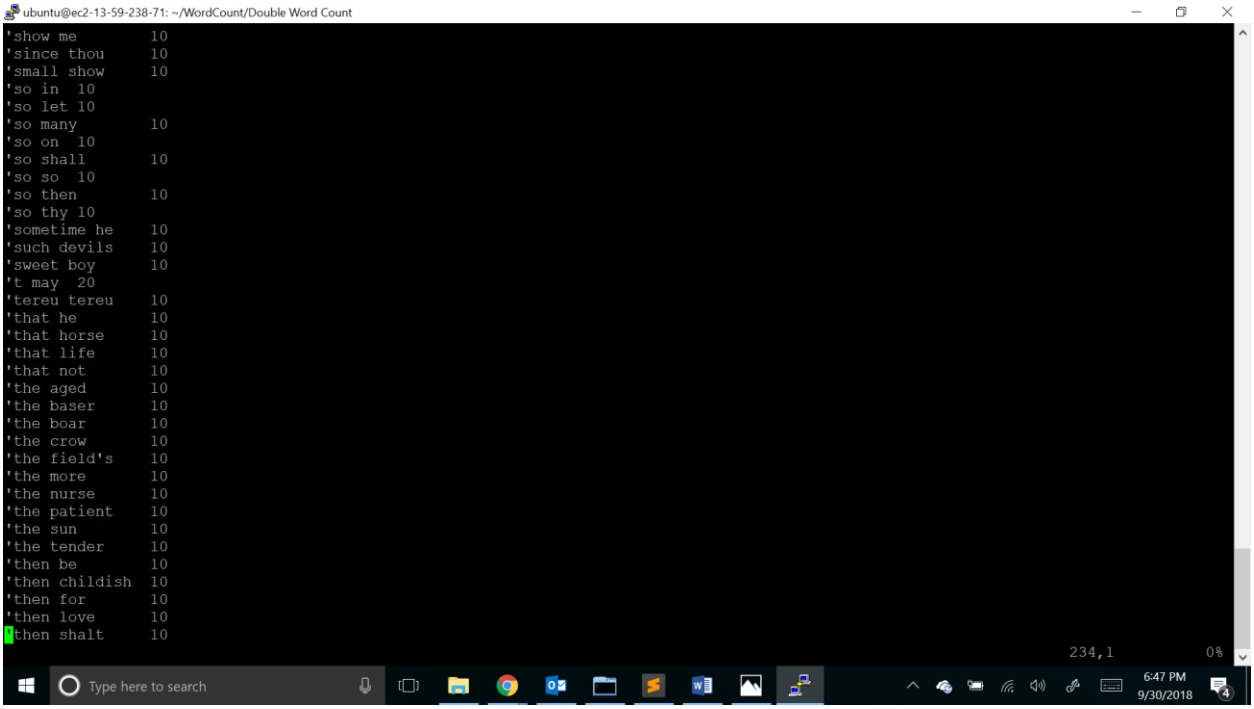
Task 2 (50 points). Count double-word frequency.

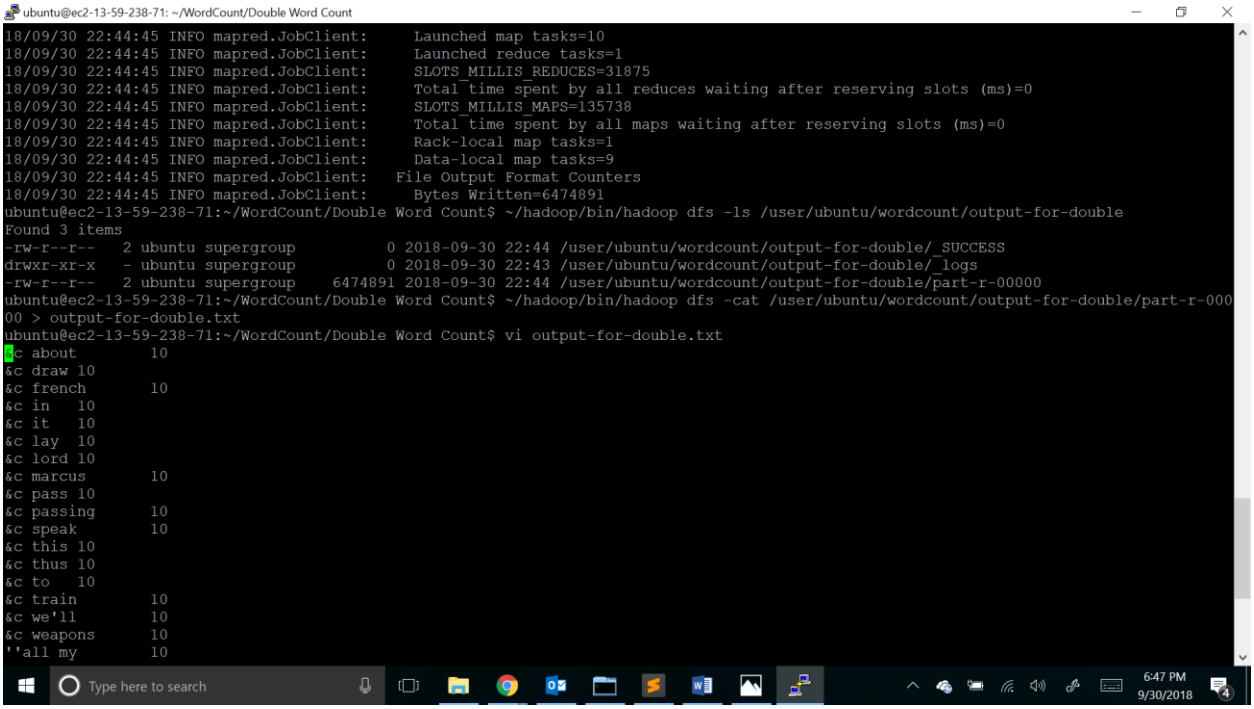
The following screenshot shows the 10 bible files in hdfs input.



The next screenshot shows the map reduce task for the Double Word Count Program for 10 bible input files shown above



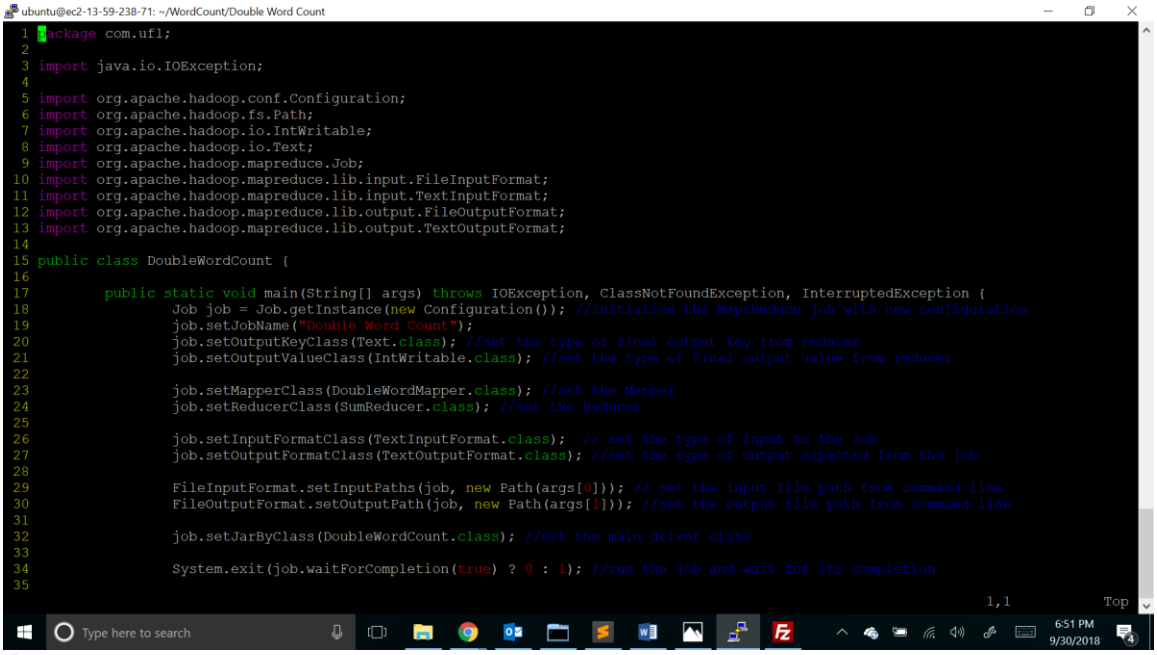


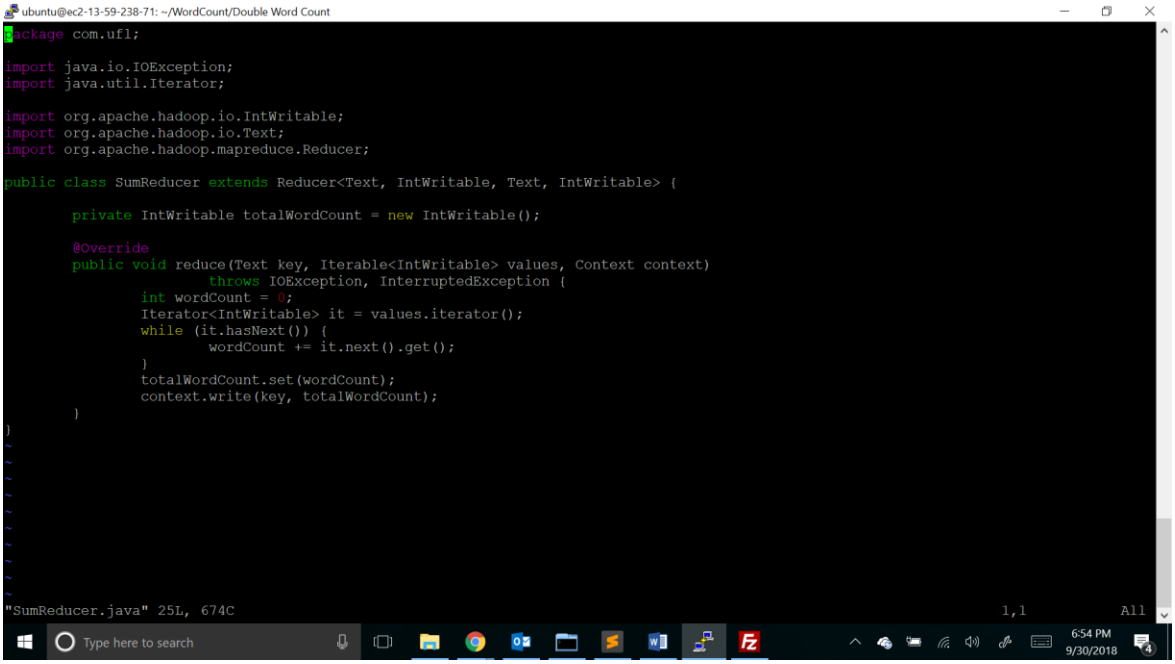


These above 2 give a brief output.

The full output for double count program on 10 bible files is in this link

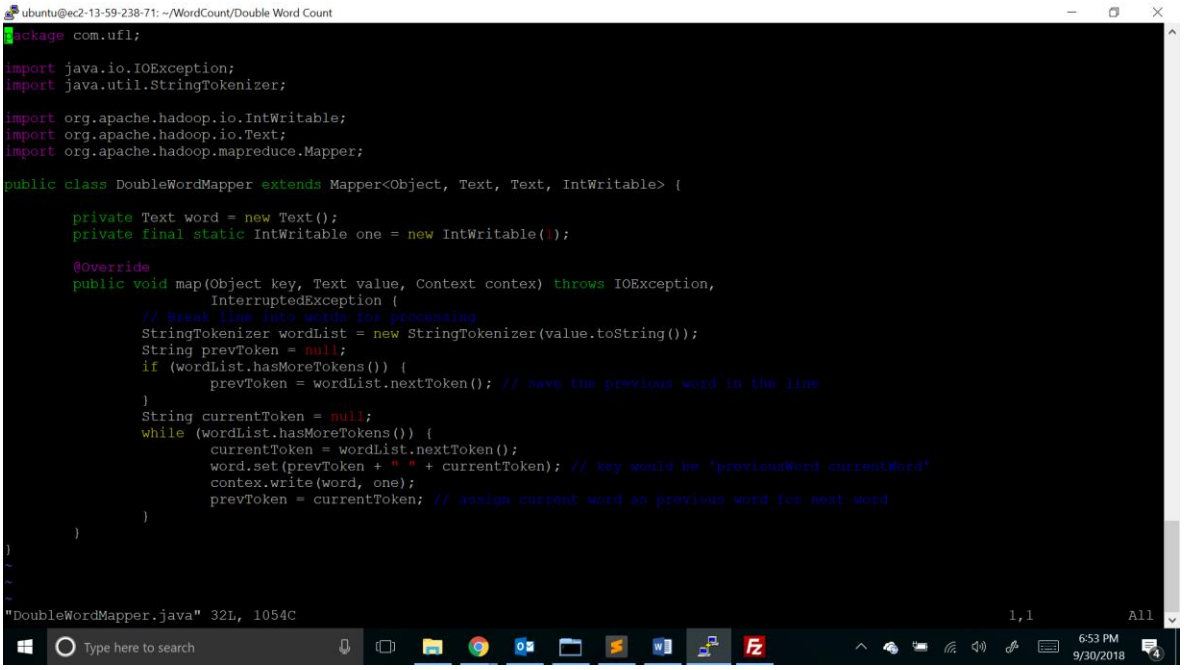
https://drive.google.com/open?id=1_RpSHAkS2gDKK1DY9MGNRpkAvrYUOY35



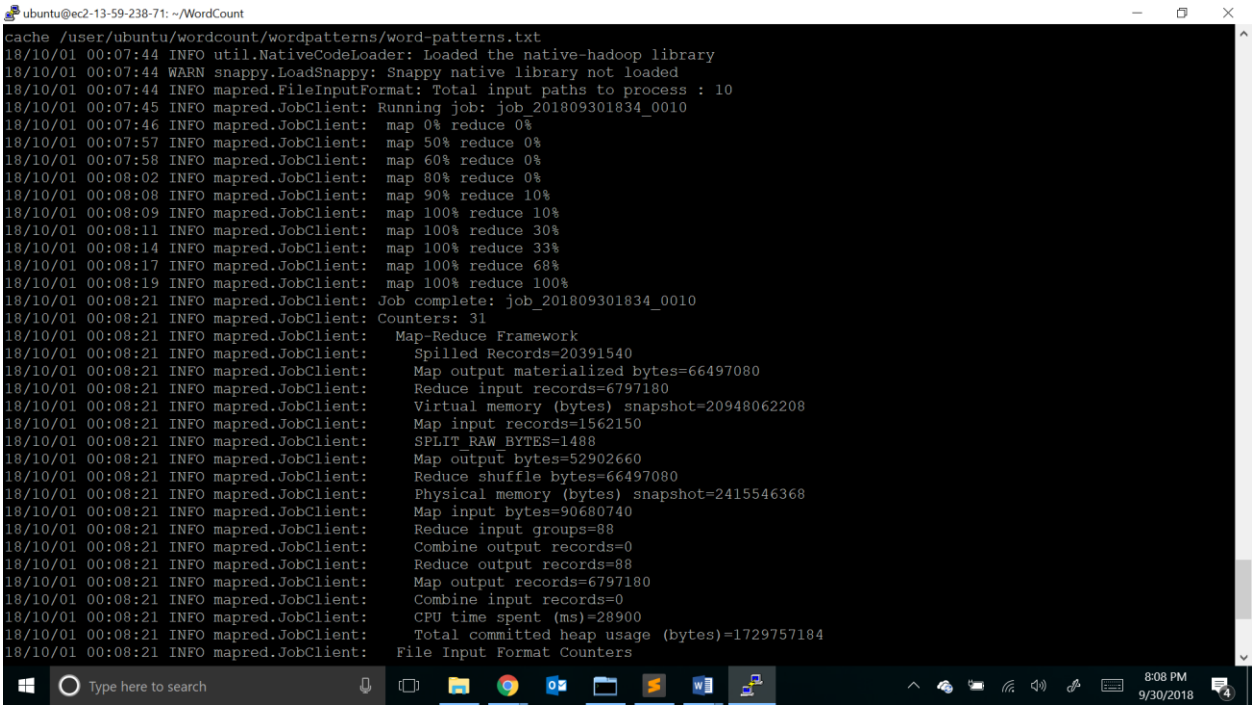


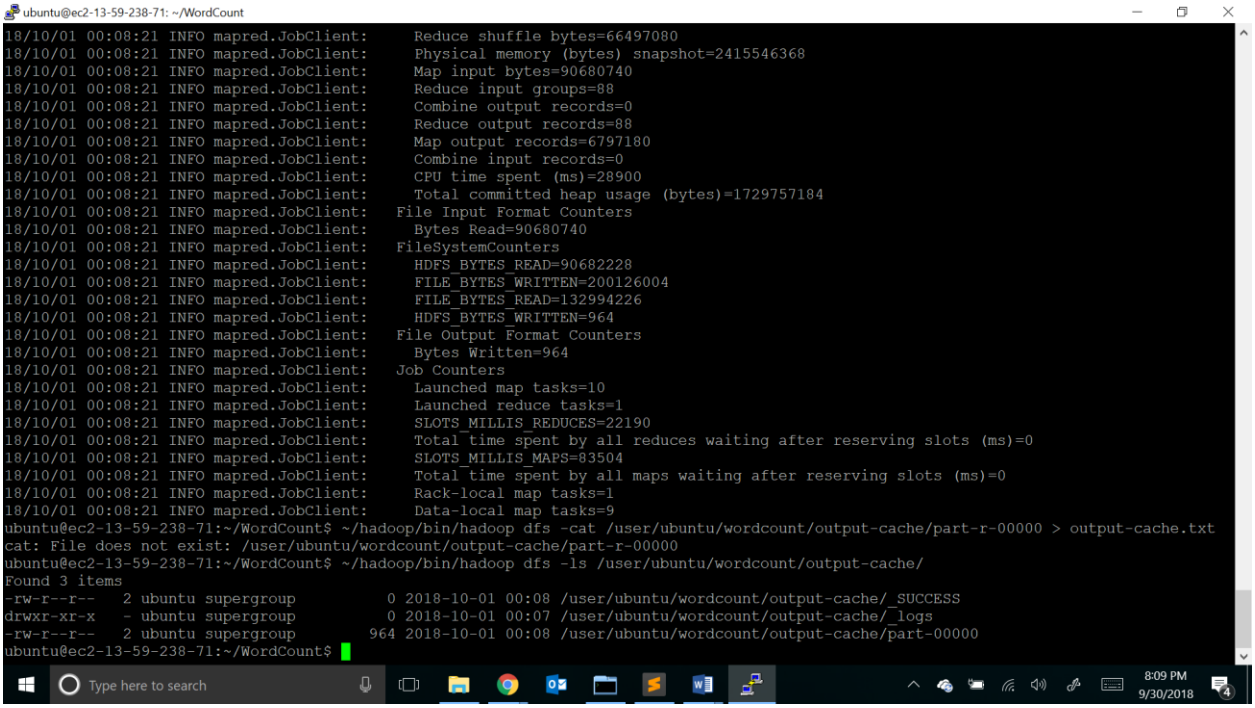
The difference in the mapper is that we include 2 tokens at the same time to count the frequency of the double word occurrences in the given file. This helps to check if the 2 words simultaneously occur in the same file. The code has been uploaded to github.

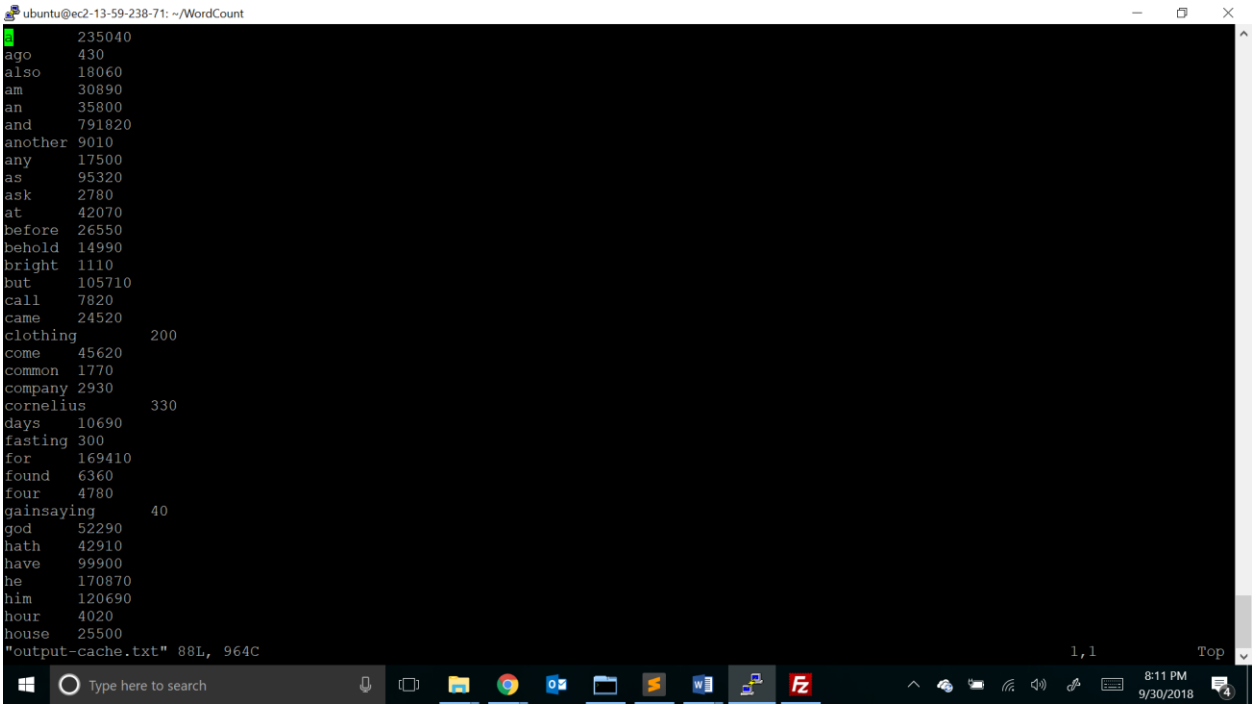
https://github.com/abhishekparekh1/hadoop_cloud



Task 3 (40 points). Using Distributed Cache (cache file).







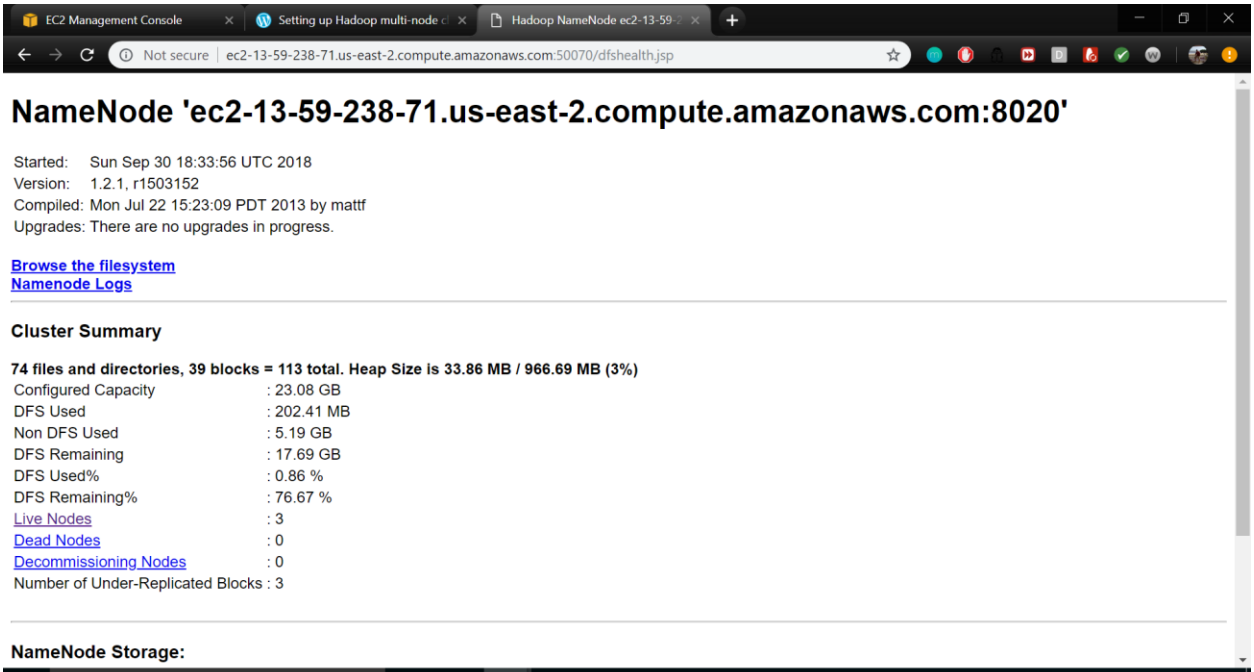
Link to the output

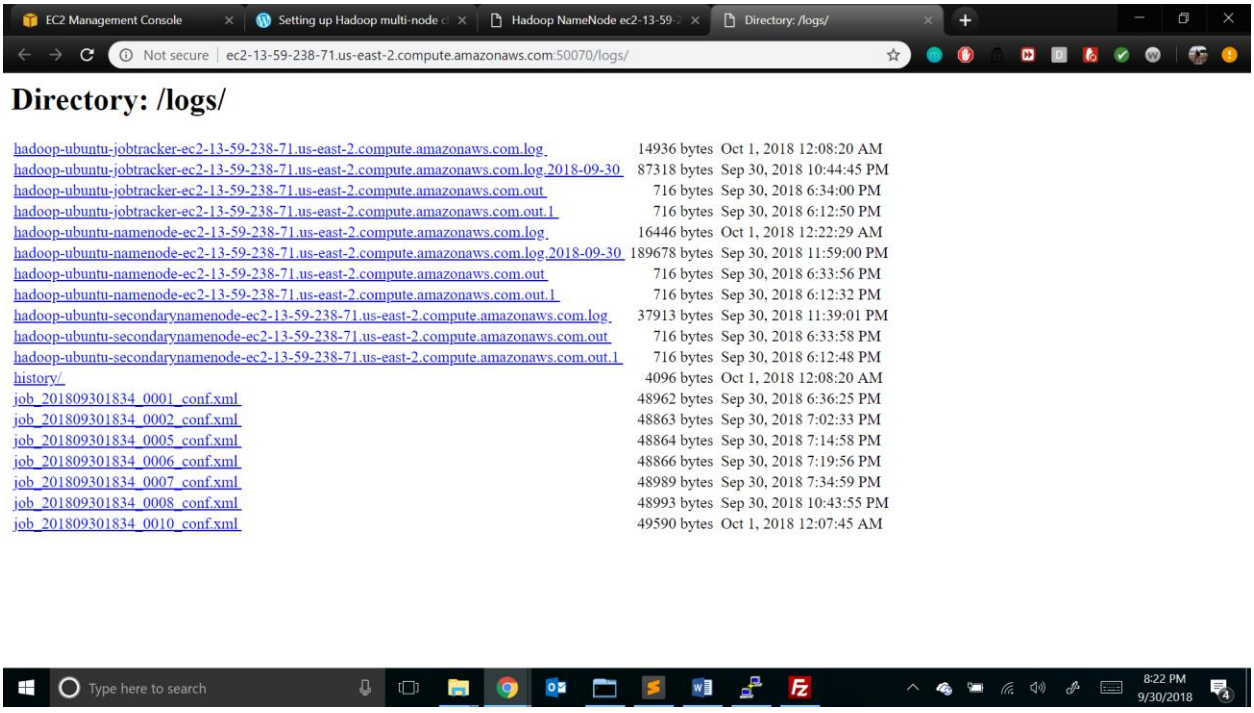
https://drive.google.com/open?id=1VZPuDHvMulejdONnrbYG7JRnEK9kj6cc

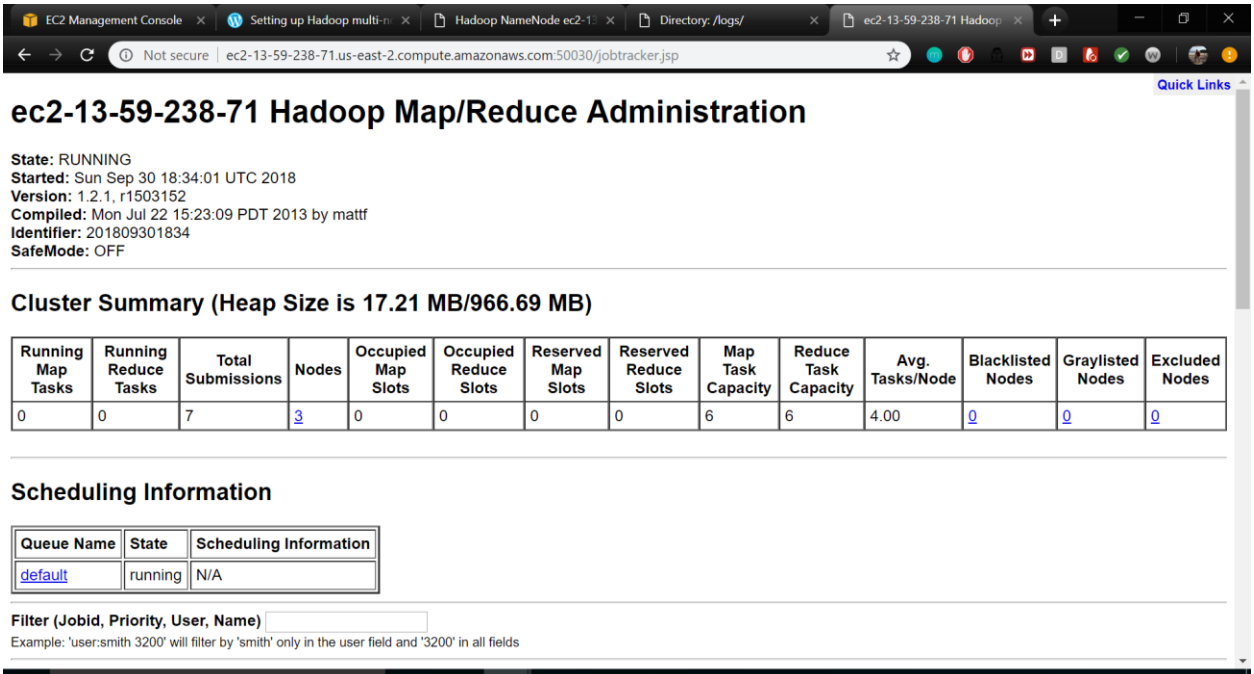
The above code shows the execution of the bible 10 files present in the input directory to have the whole word-patterns.txt described above. The distributed cache as in this example is used.

http://stackoverflow.com/questions/21239722/hadoop-distributedcache-is-deprecated-what-is-the-preferred-api#21240883

Final Status of all the nodes after all 3 tasks are completed.







ec2-13-59-238-71 Hadoop Map/Reduce Administration

State: RUNNING
Started: Sun Sep 30 18:34:01 UTC 2018
Version: 1.2.1, r1503152
Compiled: Mon Jul 22 15:23:09 PDT 2013 by mattf
Identifier: 201809301834
SafeMode: OFF

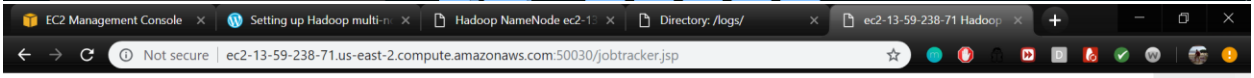
Cluster Summary (Heap Size is 17.21 MB/966.69 MB)

Running Map Tasks	Running Reduce Tasks	Total Submissions	Nodes	Occupied Map Slots	Occupied Reduce Slots	Reserved Map Slots	Reserved Reduce Slots	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node	Blacklisted Nodes	Graylisted Nodes	Excluded Nodes
0	0	7	3	0	0	0	0	6	6	4.00	0	0	0

Scheduling Information

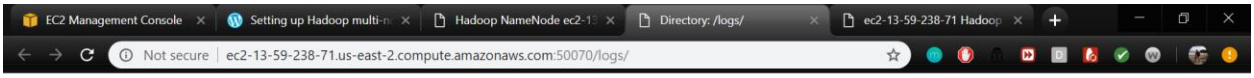
Queue Name	State	Scheduling Information
default	running	N/A

Filter (Jobid, Priority, User, Name)
Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields



Completed Jobs

Jobid	Started	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201809301834_0001	Sun Sep 30 18:36:25 UTC 2018	NORMAL	ubuntu	PiEstimator	100.00%	10	10	100.00%	1	1	NA	NA
job_201809301834_0002	Sun Sep 30 19:02:33 UTC 2018	NORMAL	ubuntu	wordcount	100.00%	2	2	100.00%	1	1	NA	NA
job_201809301834_0005	Sun Sep 30 19:14:58 UTC 2018	NORMAL	ubuntu	wordcount	100.00%	2	2	100.00%	1	1	NA	NA
job_201809301834_0006	Sun Sep 30 19:19:56 UTC 2018	NORMAL	ubuntu	wordcount	100.00%	10	10	100.00%	1	1	NA	NA
job_201809301834_0007	Sun Sep 30 19:34:59 UTC 2018	NORMAL	ubuntu	Double Word Count	100.00%	10	10	100.00%	1	1	NA	NA
job_201809301834_0008	Sun Sep 30 22:43:55 UTC 2018	NORMAL	ubuntu	Double Word Count	100.00%	10	10	100.00%	1	1	NA	NA
job_201809301834_0010	Mon Oct 01 00:07:45 UTC 2018	NORMAL	ubuntu	wordcount	100.00%	10	10	100.00%	1	1	NA	NA



Directory: /logs/

hadoop-ubuntu-jobtracker-ec2-13-59-238-71.us-east-2.compute.amazonaws.com.log	14936 bytes	Oct 1, 2018 12:08:20 AM
hadoop-ubuntu-jobtracker-ec2-13-59-238-71.us-east-2.compute.amazonaws.com.log.2018-09-30	87318 bytes	Sep 30, 2018 10:44:45 PM
hadoop-ubuntu-jobtracker-ec2-13-59-238-71.us-east-2.compute.amazonaws.com.out	716 bytes	Sep 30, 2018 6:34:00 PM
hadoop-ubuntu-jobtracker-ec2-13-59-238-71.us-east-2.compute.amazonaws.com.out.1	716 bytes	Sep 30, 2018 6:12:50 PM
hadoop-ubuntu-namenode-ec2-13-59-238-71.us-east-2.compute.amazonaws.com.log	16446 bytes	Oct 1, 2018 12:22:29 AM
hadoop-ubuntu-namenode-ec2-13-59-238-71.us-east-2.compute.amazonaws.com.log.2018-09-30	189678 bytes	Sep 30, 2018 11:59:00 PM
hadoop-ubuntu-namenode-ec2-13-59-238-71.us-east-2.compute.amazonaws.com.out	716 bytes	Sep 30, 2018 6:33:56 PM
hadoop-ubuntu-namenode-ec2-13-59-238-71.us-east-2.compute.amazonaws.com.out.1	716 bytes	Sep 30, 2018 6:12:32 PM
hadoop-ubuntu-secondarynamenode-ec2-13-59-238-71.us-east-2.compute.amazonaws.com.log	37913 bytes	Sep 30, 2018 11:39:01 PM
hadoop-ubuntu-secondarynamenode-ec2-13-59-238-71.us-east-2.compute.amazonaws.com.out	716 bytes	Sep 30, 2018 6:33:58 PM
hadoop-ubuntu-secondarynamenode-ec2-13-59-238-71.us-east-2.compute.amazonaws.com.out.1	716 bytes	Sep 30, 2018 6:12:48 PM
history/	4096 bytes	Oct 1, 2018 12:08:20 AM
job_201809301834_0001_conf.xml	48962 bytes	Sep 30, 2018 6:36:25 PM
job_201809301834_0002_conf.xml	48863 bytes	Sep 30, 2018 7:02:33 PM
job_201809301834_0005_conf.xml	48864 bytes	Sep 30, 2018 7:14:58 PM
job_201809301834_0006_conf.xml	48866 bytes	Sep 30, 2018 7:19:56 PM
job_201809301834_0007_conf.xml	48989 bytes	Sep 30, 2018 7:34:59 PM
job_201809301834_0008_conf.xml	48993 bytes	Sep 30, 2018 10:43:55 PM
job_201809301834_0010_conf.xml	49590 bytes	Oct 1, 2018 12:07:45 AM