

This assignment is made by Abhishek Parekh UFID 6333-3803 (abhishekiparekh@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://letsdobydata.wordpress.com/2014/01/13/setting-up-hadoop-multi-node-cluster-on-amazon-ec2-part-1/>

Part 2: <https://letsdobydata.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.

The screenshot displays the AWS Management Console interface for the EC2 service. The left-hand navigation pane shows the 'INSTANCES' section selected. The main content area displays a table of EC2 instances, with the following data:

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm State	Public DNS (IPv4)	IPv4 Public IP
HadoopNameNode	i-020bf625160fcc329	t2.micro	us-east-2c	running	2/2 checks ...	None	ec2-13-59-238-71.us-east-2.compute.amazonaws.com	13.59.238.71
HadoopSecondaryNode	i-031258311930dd2d8	t2.micro	us-east-2c	running	2/2 checks ...	None	ec2-18-188-184-237.us-east-2.compute.amazonaws.com	18.188.184.237
HadoopSlave1	i-07439ab5bb2e997...	t2.micro	us-east-2c	running	2/2 checks ...	None	ec2-18-216-43-196.us-east-2.compute.amazonaws.com	18.216.43.196
HadoopSlave2	i-09b58ad707fea0aa1	t2.micro	us-east-2c	running	2/2 checks ...	None	ec2-18-224-122-102.us-east-2.compute.amazonaws.com	18.224.122.102
HadoopSlave3	i-0bb39570809c37374	t2.micro	us-east-2c	running	2/2 checks ...	None	ec2-18-224-155-38.us-east-2.compute.amazonaws.com	18.224.155.38
Hadoop-namenode	i-01c62cf7b1da1c1f	t2.micro	us-east-2c	running	2/2 checks ...	None	ec2-18-222-238-14.us-east-2.compute.amazonaws.com	18.222.238.14

Below the table, the details for the selected instance 'HadoopNameNode' (Instance ID: i-020bf625160fcc329) are shown. The instance is in a 'running' state, and its Elastic IP is 13.59.238.71. The 'Description' tab is active, showing the instance's configuration details.

Setting up Hadoop multi-node cl... Hadoop NameNode ec2-13-59-238-71.us-east-2.compute.amazonaws.com:8020/dfshealth.jsp

## NameNode 'ec2-13-59-238-71.us-east-2.compute.amazonaws.com:8020'

Started: Sun Sep 30 18:33:56 UTC 2018  
Version: 1.2.1, r1503152  
Compiled: Mon Jul 22 15:23:09 PDT 2013 by mattf  
Upgrades: There are no upgrades in progress.

[Browse the filesystem](#)  
[NameNode Logs](#)

### Cluster Summary

56 files and directories, 25 blocks = 81 total. Heap Size is 33.86 MB / 966.69 MB (3%)

Configured Capacity	: 23.08 GB
DFS Used	: 189.5 MB
Non DFS Used	: 5.18 GB
DFS Remaining	: 17.71 GB
DFS Used%	: 0.8 %
DFS Remaining%	: 76.75 %
<a href="#">Live Nodes</a>	: 3
<a href="#">Dead Nodes</a>	: 0
<a href="#">Decommissioning Nodes</a>	: 0
Number of Under-Replicated Blocks	: 2

### NameNode Storage:

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

```
ubuntu@ip-172-31-41-196:~$ hadoop dfs -ls
Found 1 items
drwxr-xr-x - ubuntu supergroup          0 2018-09-30 19:02 /user/ubuntu/wordcount
ubuntu@ip-172-31-41-196:~$ hadoop dfs -ls /user/ubuntu/wordcount
Found 2 items
drwxr-xr-x - ubuntu supergroup          0 2018-09-30 19:11 /user/ubuntu/wordcount/input
drwxr-xr-x - ubuntu supergroup          0 2018-09-30 19:02 /user/ubuntu/wordcount/output
ubuntu@ip-172-31-41-196:~$ hadoop dfs -ls /user/ubuntu/wordcount/input
Found 2 items
-rw-r--r-- 2 ubuntu supergroup    9068074 2018-09-30 19:11 /user/ubuntu/wordcount/input/bible
-rw-r--r-- 2 ubuntu supergroup     23 2018-09-30 18:59 /user/ubuntu/wordcount/input/file01.txt
ubuntu@ip-172-31-41-196:~$ hadoop dfs -rm /user/ubuntu/wordcount/input/file01.txt
Deleted hdfs://ec2-13-59-238-71.us-east-2.compute.amazonaws.com:8020/user/ubuntu/wordcount/input/file01.txt
ubuntu@ip-172-31-41-196:~$ hadoop jar /usr/joe/wordcount.jar org.myorg.WordCount /usr/joe/wordcount/input /usr/joe/wordcount/output
```

```
MTPuTTY (Multi-Tabbed PuTTY)
ubuntu@ip-172-31-41-196: ~ X  ubuntu@ip-172-31-47-20: ~/hadoop/conf X  ubuntu@ip-172-31-38-21: ~/hadoop/conf X  ubuntu@ip-172-31-34-223: ~/hadoop/conf X  ubuntu@ip-172-31-36-95: ~/hadoop/conf X
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:~$
```

```
MTPuTTY (Multi-Tabbed PuTTY)
ubuntu@ip-172-31-41-196: ~ X  ubuntu@ip-172-31-47-20: ~/hadoop/conf X  ubuntu@ip-172-31-38-21: ~/hadoop/conf X  ubuntu@ip-172-31-34-223: ~/hadoop/conf X  ubuntu@ip-172-31-36-95: ~/hadoop/conf X
ziz 1
ziza 2
zizah 1
zo 1
zoan 7
zoar 10
zoba 2
zobah 11
zobebah 1
zodiac 1
zodiacs 1
zohar 4
zohelath 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
zswaggered 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:~$
```

then 5  
'therefore 1  
'they 1  
'think 1  
'this 4  
'thou 7  
'thricefairer 1  
'thus 4  
'thy 1  
'time's 1  
'tis 18  
'to 5  
'torches 1  
'touch 1  
'truth 1  
'twas 1  
'tween 1  
'twixt 1  
'unless 1  
'unruly 1

[Download this file](#)  
[Tail this file](#)  
Chunk size to view (in bytes, up to file's DFS block size):

Total number of blocks: 1  
-5064470552668424351: [172.31.34.223:50010](#) [172.31.36.95:50010](#)

[Go back to DFS home](#)

### Local logs

[Log directory](#)

[https://drive.google.com/open?id=1\\_3h1qgfGa0ilMxcqNKbcTY-IS4nZswAS](https://drive.google.com/open?id=1_3h1qgfGa0ilMxcqNKbcTY-IS4nZswAS)

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](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

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

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

```
ubuntu@ec2-13-59-238-71: ~/WordCount/Double Word Count
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$ ~/hadoop/bin/hadoop dfs -ls /user/ubuntu/wordcount/input
Found 10 items
-rw-r--r-- 2 ubuntu supergroup 9068074 2018-09-30 19:11 /user/ubuntu/wordcount/input/bible
-rw-r--r-- 2 ubuntu supergroup 9068074 2018-09-30 19:19 /user/ubuntu/wordcount/input/bible01
-rw-r--r-- 2 ubuntu supergroup 9068074 2018-09-30 19:19 /user/ubuntu/wordcount/input/bible02
-rw-r--r-- 2 ubuntu supergroup 9068074 2018-09-30 19:19 /user/ubuntu/wordcount/input/bible03
-rw-r--r-- 2 ubuntu supergroup 9068074 2018-09-30 19:19 /user/ubuntu/wordcount/input/bible04
-rw-r--r-- 2 ubuntu supergroup 9068074 2018-09-30 19:19 /user/ubuntu/wordcount/input/bible05
-rw-r--r-- 2 ubuntu supergroup 9068074 2018-09-30 19:19 /user/ubuntu/wordcount/input/bible06
-rw-r--r-- 2 ubuntu supergroup 9068074 2018-09-30 19:19 /user/ubuntu/wordcount/input/bible07
-rw-r--r-- 2 ubuntu supergroup 9068074 2018-09-30 19:19 /user/ubuntu/wordcount/input/bible08
-rw-r--r-- 2 ubuntu supergroup 9068074 2018-09-30 19:19 /user/ubuntu/wordcount/input/bible09
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$
```

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

```
ubuntu@ec2-13-59-238-71: ~/WordCount/Double Word Count
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$ ls
bible double output2.txt double wc classes DoubleWordCount.java SumReducer.java
double bible output1 doublewordcount.jar DoubleWordMapper.java
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$ bin/hadoop jar doublewordcount.jar com.ufl.DoubleWordCount /user/ubuntu/wordcount
t/input /user/ubuntu/wordcount/output-for-double
-bash: bin/hadoop: No such file or directory
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$ ~/bin/hadoop jar doublewordcount.jar com.ufl.DoubleWordCount /user/ubuntu/wordco
unt/input /user/ubuntu/wordcount/output-for-double
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$ ~/
.cache/      hadoop/      .nano/      WordCount/
.git/        hdfsmp/      .ssh/       wordcount_classes/
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$ ~/bin/hadoop jar doublewordcount.jar com.ufl.DoubleWordCount /user/ubuntu/wordco
unt/input /user/ubuntu/wordcount/output-for-double
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$ ~/
.cache/      hadoop/      .nano/      WordCount/
.git/        hdfsmp/      .ssh/       wordcount_classes/
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$ ~/hadoop/bin/hadoop jar doublewordcount.jar com.ufl.DoubleWordCount /user/ubuntu
/wordcount/input /user/ubuntu/wordcount/output-for-double
18/09/30 22:43:54 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the
same.
18/09/30 22:43:55 INFO input.FileInputFormat: Total input paths to process : 10
18/09/30 22:43:55 INFO util.NativeCodeLoader: Loaded the native-hadoop library
18/09/30 22:43:55 WARN snappy.LoadSnappy: Snappy native library not loaded
18/09/30 22:43:55 INFO mapred.JobClient: Running job: job_201809301834_0008
18/09/30 22:43:56 INFO mapred.JobClient: map 0% reduce 0%
18/09/30 22:44:08 INFO mapred.JobClient: map 31% reduce 0%
18/09/30 22:44:11 INFO mapred.JobClient: map 60% reduce 0%
18/09/30 22:44:19 INFO mapred.JobClient: map 80% reduce 0%
18/09/30 22:44:23 INFO mapred.JobClient: map 90% reduce 0%
18/09/30 22:44:26 INFO mapred.JobClient: map 97% reduce 0%
18/09/30 22:44:27 INFO mapred.JobClient: map 97% reduce 16%
18/09/30 22:44:29 INFO mapred.JobClient: map 100% reduce 16%
18/09/30 22:44:33 INFO mapred.JobClient: map 100% reduce 26%
18/09/30 22:44:39 INFO mapred.JobClient: map 100% reduce 68%
18/09/30 22:44:42 INFO mapred.JobClient: map 100% reduce 90%
18/09/30 22:44:44 INFO mapred.JobClient: map 100% reduce 100%
```

```
ubuntu@ec2-13-59-238-71: ~/WordCount/Double Word Count
'show me 10
'since thou 10
'small show 10
'so in 10
'so let 10
'so many 10
'so on 10
'so shall 10
'so so 10
'so then 10
'so thy 10
'sometime he 10
'such devils 10
'sweet boy 10
't may 20
'tereu tereu 10
'that he 10
'that horse 10
'that life 10
'that not 10
'the aged 10
'the baser 10
'the boar 10
'the crow 10
'the field's 10
'the more 10
'the nurse 10
'the patient 10
'the sun 10
'the tender 10
'then be 10
'then childish 10
'then for 10
'then love 10
'then shalt 10

234,1 0%

ubuntu@ec2-13-59-238-71: ~/WordCount/Double Word Count
18/09/30 22:44:45 INFO mapred.JobClient: Launched map tasks=10
18/09/30 22:44:45 INFO mapred.JobClient: Launched reduce tasks=1
18/09/30 22:44:45 INFO mapred.JobClient: SLOTS MILLIS REDUCES=31875
18/09/30 22:44:45 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0
18/09/30 22:44:45 INFO mapred.JobClient: SLOTS MILLIS MAPS=135738
18/09/30 22:44:45 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
18/09/30 22:44:45 INFO mapred.JobClient: Rack-local map tasks=1
18/09/30 22:44:45 INFO mapred.JobClient: Data-local map tasks=9
18/09/30 22:44:45 INFO mapred.JobClient: File Output Format Counters
18/09/30 22:44:45 INFO mapred.JobClient: Bytes Written=6474891
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$ ~/hadoop/bin/hadoop dfs -ls /user/ubuntu/wordcount/output-for-double
Found 3 items
-rw-r--r-- 2 ubuntu supergroup 0 2018-09-30 22:44 /user/ubuntu/wordcount/output-for-double/ SUCCESS
drwxr-xr-x - ubuntu supergroup 0 2018-09-30 22:43 /user/ubuntu/wordcount/output-for-double/ logs
-rw-r--r-- 2 ubuntu supergroup 6474891 2018-09-30 22:44 /user/ubuntu/wordcount/output-for-double/part-r-000000
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$ ~/hadoop/bin/hadoop dfs -cat /user/ubuntu/wordcount/output-for-double/part-r-000000 > output-for-double.txt
ubuntu@ec2-13-59-238-71:~/WordCount/Double Word Count$ vi output-for-double.txt
#c about 10
#c draw 10
#c french 10
#c in 10
#c it 10
#c lay 10
#c lord 10
#c marcus 10
#c pass 10
#c passing 10
#c speak 10
#c this 10
#c thus 10
#c to 10
#c train 10
#c we'll 10
#c weapons 10
'all my 10
```

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



```
ubuntu@ec2-13-59-238-71: ~/WordCount/Double Word Count
1 package com.uf1;
2
3 import java.io.IOException;
4
5 import org.apache.hadoop.conf.Configuration;
6 import org.apache.hadoop.fs.Path;
7 import org.apache.hadoop.io.IntWritable;
8 import org.apache.hadoop.io.Text;
9 import org.apache.hadoop.mapreduce.Job;
10 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
11 import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
12 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
13 import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
14
15 public class DoubleWordCount {
16
17     public static void main(String[] args) throws IOException, ClassNotFoundException, InterruptedException {
18         Job job = Job.getInstance(new Configuration()); //initialize the MapReduce job with new configuration
19         job.setJobName("Double Word Count");
20         job.setOutputKeyClass(Text.class); //set the type of final output key from reducer
21         job.setOutputValueClass(IntWritable.class); //set the type of final output value from reducer
22
23         job.setMapperClass(DoubleWordMapper.class); //set the Mapper
24         job.setReducerClass(SumReducer.class); //set the Reducer
25
26         job.setInputFormatClass(TextInputFormat.class); // set the type of input to the Job
27         job.setOutputFormatClass(TextOutputFormat.class); //set the type of Output expected from the job
28
29         FileInputFormat.setInputPaths(job, new Path(args[0])); // set the input file path from command-line
30         FileOutputFormat.setOutputPath(job, new Path(args[1])); //set the output file path from command-line
31
32         job.setJarByClass(DoubleWordCount.class); //set the main driver class
33
34         System.exit(job.waitForCompletion(true) ? 0 : 1); //run the job and wait for its completion
35
36     }
37 }
```

```
ubuntu@ec2-13-59-238-71: ~/WordCount/Double Word Count
package com.uf1;

import java.io.IOException;
import java.util.Iterator;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class SumReducer extends Reducer<Text, IntWritable, Text, IntWritable> {

    private IntWritable totalWordCount = new IntWritable();

    @Override
    public void reduce(Text key, Iterable<IntWritable> values, Context context)
        throws IOException, InterruptedException {
        int wordCount = 0;
        Iterator<IntWritable> it = values.iterator();
        while (it.hasNext()) {
            wordCount += it.next().get();
        }
        totalWordCount.set(wordCount);
        context.write(key, totalWordCount);
    }
}
```

"SumReducer.java" 25L, 674C

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](https://github.com/abhishekparekh1/hadoop_cloud)



```
ubuntu@ec2-13-59-238-71: ~/WordCount/Double Word Count
package com.ufl;

import java.io.IOException;
import java.util.StringTokenizer;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class DoubleWordMapper extends Mapper<Object, Text, Text, IntWritable> {

    private Text word = new Text();
    private final static IntWritable one = new IntWritable(1);

    @Override
    public void map(Object key, Text value, Context context) throws IOException,
        InterruptedException {
        // Break line into words for processing
        StringTokenizer wordList = new StringTokenizer(value.toString());
        String prevToken = null;
        if (wordList.hasMoreTokens()) {
            prevToken = wordList.nextToken(); // save the previous word in the line
        }
        String currentToken = null;
        while (wordList.hasMoreTokens()) {
            currentToken = wordList.nextToken();
            word.set(prevToken + " " + currentToken); // key would be 'previousWord currentWord'
            context.write(word, one);
            prevToken = currentToken; // assign current word as previous word for next word
        }
    }
}

"DoubleWordMapper.java" 32L, 1054C
1, 1 All
```

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

```
ubuntu@ec2-13-59-238-71: ~/WordCount
cache /user/ubuntu/wordcount/wordpatterns/word-patterns.txt
18/10/01 00:07:44 INFO util.NativeCodeLoader: Loaded the native-hadoop library
18/10/01 00:07:44 WARN snappy.LoadSnappy: Snappy native library not loaded
18/10/01 00:07:44 INFO mapred.FileInputFormat: Total input paths to process : 10
18/10/01 00:07:45 INFO mapred.JobClient: Running job: job_201809301834_0010
18/10/01 00:07:46 INFO mapred.JobClient: map 0% reduce 0%
18/10/01 00:07:57 INFO mapred.JobClient: map 50% reduce 0%
18/10/01 00:07:58 INFO mapred.JobClient: map 60% reduce 0%
18/10/01 00:08:02 INFO mapred.JobClient: map 80% reduce 0%
18/10/01 00:08:08 INFO mapred.JobClient: map 90% reduce 10%
18/10/01 00:08:09 INFO mapred.JobClient: map 100% reduce 10%
18/10/01 00:08:11 INFO mapred.JobClient: map 100% reduce 30%
18/10/01 00:08:14 INFO mapred.JobClient: map 100% reduce 33%
18/10/01 00:08:17 INFO mapred.JobClient: map 100% reduce 68%
18/10/01 00:08:19 INFO mapred.JobClient: map 100% reduce 100%
18/10/01 00:08:21 INFO mapred.JobClient: Job complete: job_201809301834_0010
18/10/01 00:08:21 INFO mapred.JobClient: Counters: 31
18/10/01 00:08:21 INFO mapred.JobClient: Map-Reduce Framework
18/10/01 00:08:21 INFO mapred.JobClient: Spilled Records=20391540
18/10/01 00:08:21 INFO mapred.JobClient: Map output materialized bytes=66497080
18/10/01 00:08:21 INFO mapred.JobClient: Reduce input records=6797180
18/10/01 00:08:21 INFO mapred.JobClient: Virtual memory (bytes) snapshot=20948062208
18/10/01 00:08:21 INFO mapred.JobClient: Map input records=1562150
18/10/01 00:08:21 INFO mapred.JobClient: SPLIT_RAW_BYTES=1488
18/10/01 00:08:21 INFO mapred.JobClient: Map output bytes=52902660
18/10/01 00:08:21 INFO mapred.JobClient: Reduce shuffle bytes=66497080
18/10/01 00:08:21 INFO mapred.JobClient: Physical memory (bytes) snapshot=2415546368
18/10/01 00:08:21 INFO mapred.JobClient: Map input bytes=90680740
18/10/01 00:08:21 INFO mapred.JobClient: Reduce input groups=88
18/10/01 00:08:21 INFO mapred.JobClient: Combine output records=0
18/10/01 00:08:21 INFO mapred.JobClient: Reduce output records=88
18/10/01 00:08:21 INFO mapred.JobClient: Map output records=6797180
18/10/01 00:08:21 INFO mapred.JobClient: Combine input records=0
18/10/01 00:08:21 INFO mapred.JobClient: CPU time spent (ms)=28900
18/10/01 00:08:21 INFO mapred.JobClient: Total committed heap usage (bytes)=1729757184
18/10/01 00:08:21 INFO mapred.JobClient: File Input Format Counters
```

```
ubuntu@ec2-13-59-238-71: ~/WordCount
18/10/01 00:08:21 INFO mapred.JobClient: Reduce shuffle bytes=66497080
18/10/01 00:08:21 INFO mapred.JobClient: Physical memory (bytes) snapshot=2415546368
18/10/01 00:08:21 INFO mapred.JobClient: Map input bytes=90680740
18/10/01 00:08:21 INFO mapred.JobClient: Reduce input groups=88
18/10/01 00:08:21 INFO mapred.JobClient: Combine output records=0
18/10/01 00:08:21 INFO mapred.JobClient: Reduce output records=88
18/10/01 00:08:21 INFO mapred.JobClient: Map output records=6797180
18/10/01 00:08:21 INFO mapred.JobClient: Combine input records=0
18/10/01 00:08:21 INFO mapred.JobClient: CPU time spent (ms)=28900
18/10/01 00:08:21 INFO mapred.JobClient: Total committed heap usage (bytes)=1729757184
18/10/01 00:08:21 INFO mapred.JobClient: File Input Format Counters
18/10/01 00:08:21 INFO mapred.JobClient:   Bytes Read=90680740
18/10/01 00:08:21 INFO mapred.JobClient: FileSystemCounters
18/10/01 00:08:21 INFO mapred.JobClient:   HDFS_BYTES_READ=90682228
18/10/01 00:08:21 INFO mapred.JobClient:   FILE_BYTES_WRITTEN=200126004
18/10/01 00:08:21 INFO mapred.JobClient:   FILE_BYTES_READ=132994226
18/10/01 00:08:21 INFO mapred.JobClient:   HDFS_BYTES_WRITTEN=964
18/10/01 00:08:21 INFO mapred.JobClient: File Output Format Counters
18/10/01 00:08:21 INFO mapred.JobClient:   Bytes Written=964
18/10/01 00:08:21 INFO mapred.JobClient: Job Counters
18/10/01 00:08:21 INFO mapred.JobClient:   Launched map tasks=10
18/10/01 00:08:21 INFO mapred.JobClient:   Launched reduce tasks=1
18/10/01 00:08:21 INFO mapred.JobClient:   SLOTS_MILLIS_REDUCES=22190
18/10/01 00:08:21 INFO mapred.JobClient:   Total time spent by all reduces waiting after reserving slots (ms)=0
18/10/01 00:08:21 INFO mapred.JobClient:   SLOTS_MILLIS_MAPS=83504
18/10/01 00:08:21 INFO mapred.JobClient:   Total time spent by all maps waiting after reserving slots (ms)=0
18/10/01 00:08:21 INFO mapred.JobClient:   Rack-local map tasks=1
18/10/01 00:08:21 INFO mapred.JobClient:   Data-local map tasks=9
ubuntu@ec2-13-59-238-71:~/WordCount$ ~/hadoop/bin/hadoop dfs -cat /user/ubuntu/wordcount/output-cache/part-r-00000 > output-cache.txt
cat: File does not exist: /user/ubuntu/wordcount/output-cache/part-r-00000
ubuntu@ec2-13-59-238-71:~/WordCount$ ~/hadoop/bin/hadoop dfs -ls /user/ubuntu/wordcount/output-cache/
Found 3 items
-rw-r--r--  2 ubuntu supergroup          0 2018-10-01 00:08 /user/ubuntu/wordcount/output-cache/_SUCCESS
drwxr-xr-x  - ubuntu supergroup          0 2018-10-01 00:07 /user/ubuntu/wordcount/output-cache/_logs
-rw-r--r--  2 ubuntu supergroup       964 2018-10-01 00:08 /user/ubuntu/wordcount/output-cache/part-00000
ubuntu@ec2-13-59-238-71:~/WordCount$
```

```
ubuntu@ec2-13-59-238-71: ~/WordCount
ago 235040
also 430
am 18060
an 30890
and 35800
another 791820
any 9010
as 17500
ask 95320
at 2780
before 42070
behold 26550
bright 14990
but 1110
call 105710
came 7820
clothing 24520
come 200
common 45620
company 1770
cornelius 2930
days 330
fasting 10690
for 300
found 169410
four 6360
gainsaying 4780
god 40
hath 52290
have 42910
he 99900
him 170870
hour 120690
house 4020
"output-cache.txt" 88L, 964C
```

Link to the output

<https://drive.google.com/open?id=1VZPuDHvMulejdONnrbyYG7JRnEK9kj6cc>

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.

The screenshot shows the Hadoop NameNode web interface for the instance `ec2-13-59-238-71.us-east-2.compute.amazonaws.com:8020`. The interface includes a header with the instance name and a navigation bar with links to the filesystem and logs. The main content area displays the **Cluster Summary** and **NameNode Storage** sections. The cluster summary shows 74 files and directories, 39 blocks, and a total heap size of 966.69 MB (3% used). The storage section shows the directory `/logs/` with a list of log files and their sizes.

**NameNode 'ec2-13-59-238-71.us-east-2.compute.amazonaws.com:8020'**

Started: Sun Sep 30 18:33:56 UTC 2018  
Version: 1.2.1, r1503152  
Compiled: Mon Jul 22 15:23:09 PDT 2013 by mattf  
Upgrades: There are no upgrades in progress.

[Browse the filesystem](#)  
[Namenode Logs](#)

---

**Cluster Summary**

74 files and directories, 39 blocks = 113 total. Heap Size is 33.86 MB / 966.69 MB (3%)

Configured Capacity	: 23.08 GB
DFS Used	: 202.41 MB
Non DFS Used	: 5.19 GB
DFS Remaining	: 17.69 GB
DFS Used%	: 0.86 %
DFS Remaining%	: 76.67 %
<a href="#">Live Nodes</a>	: 3
<a href="#">Dead Nodes</a>	: 0
<a href="#">Decommissioning Nodes</a>	: 0
Number of Under-Replicated Blocks :	3

---

**NameNode Storage:**

Directory: /logs/

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





## Directory: /logs/

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

