

SMIT R PATEL

19162121031

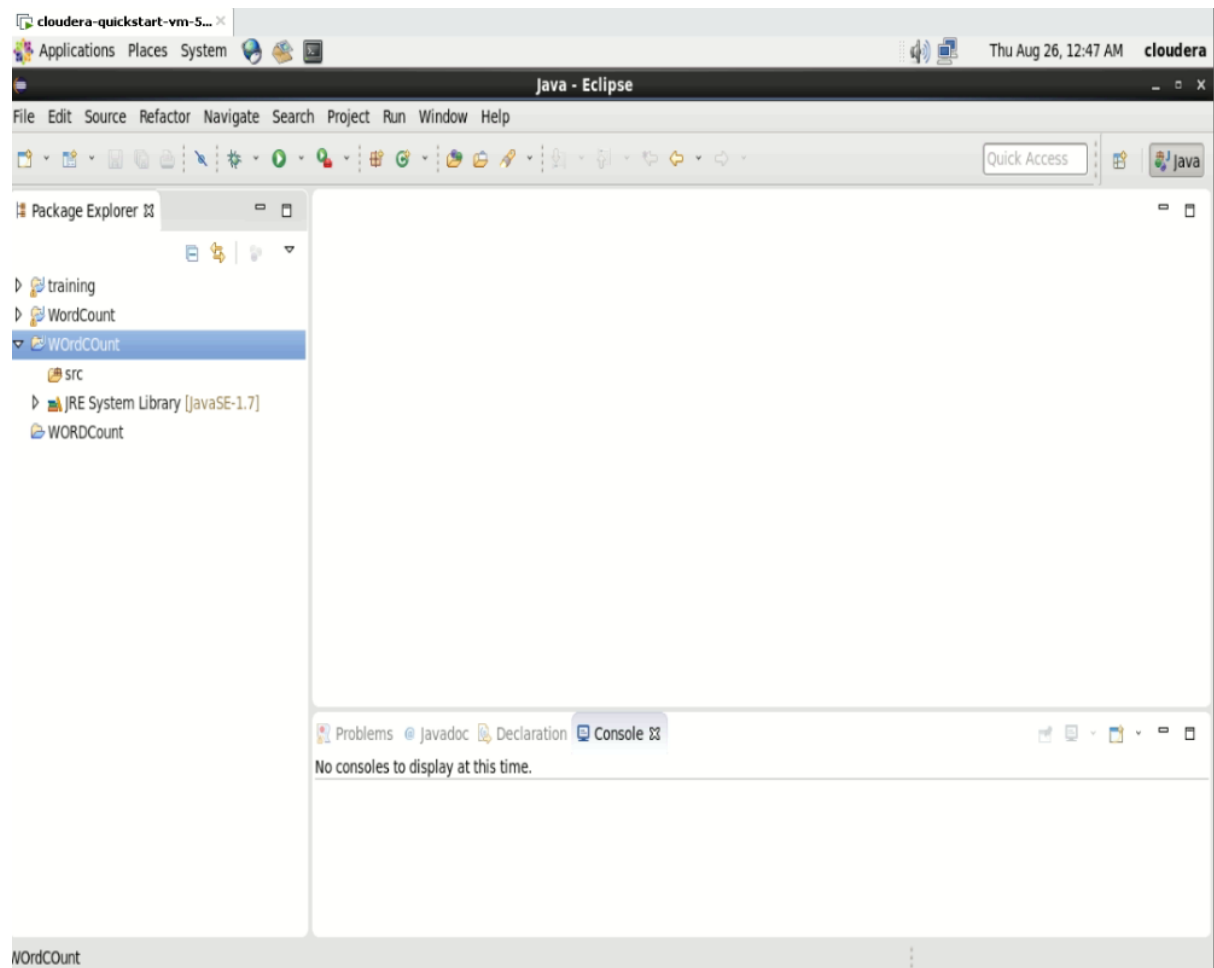
SEM 5

PRACTICAL 4

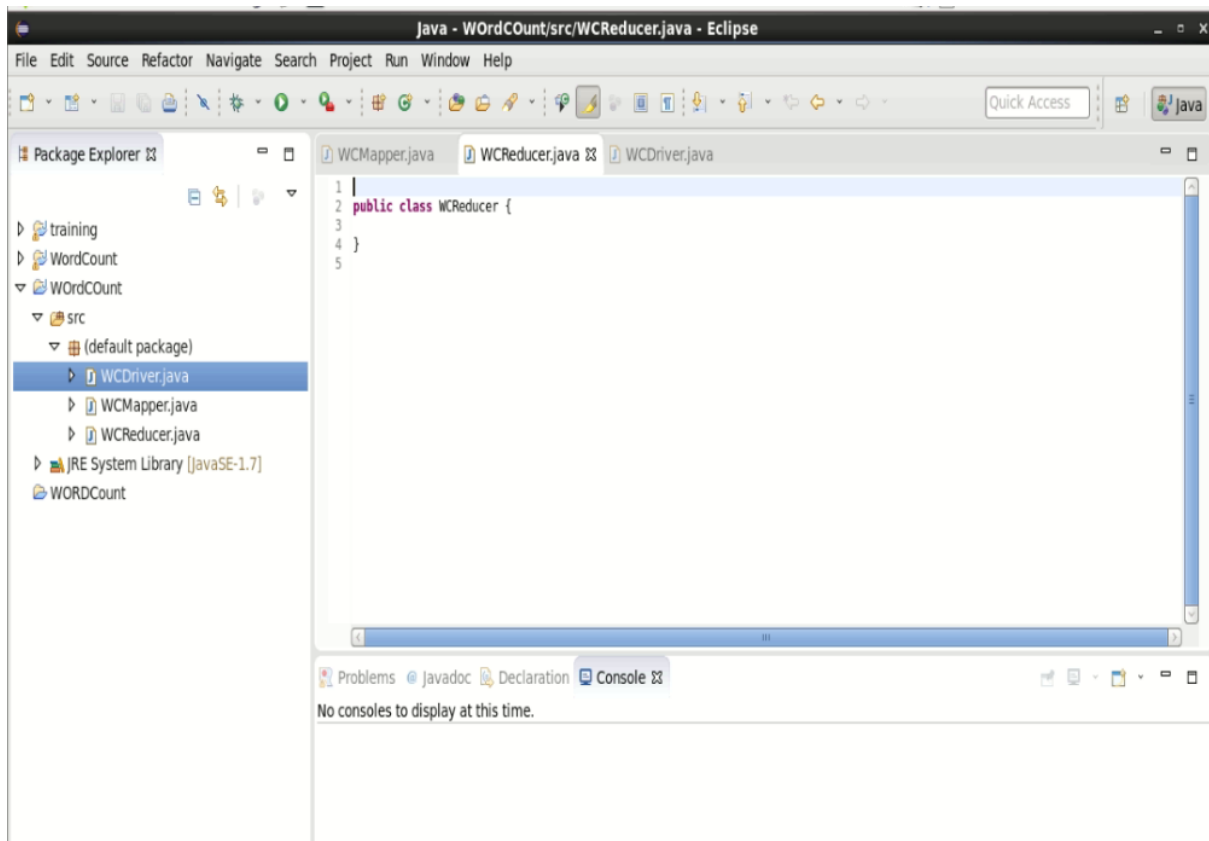
BIG DATA AND ANALYTICS

Aim: Understanding how to Execute WordCount Program in MapReduce using Cloudera Distribution Hadoop (CDH). You are deployed as a trainer in a multi-national company that is planning to adapt to big data practices. The batch of employees you are training- are freshers to Hadoop. You need to teach them to execute a basic Word Count MapReduce program. Counting the number of words in any language is a piece of cake like in C, C++, Python, Java, etc. MapReduce also uses Java but it is very easy if you know the syntax on how to write it. It is the basic of MapReduce. You will first learn how to execute this code similar to "Hello World" program in other languages. So here are the steps which show how to write a MapReduce code for Word Count.

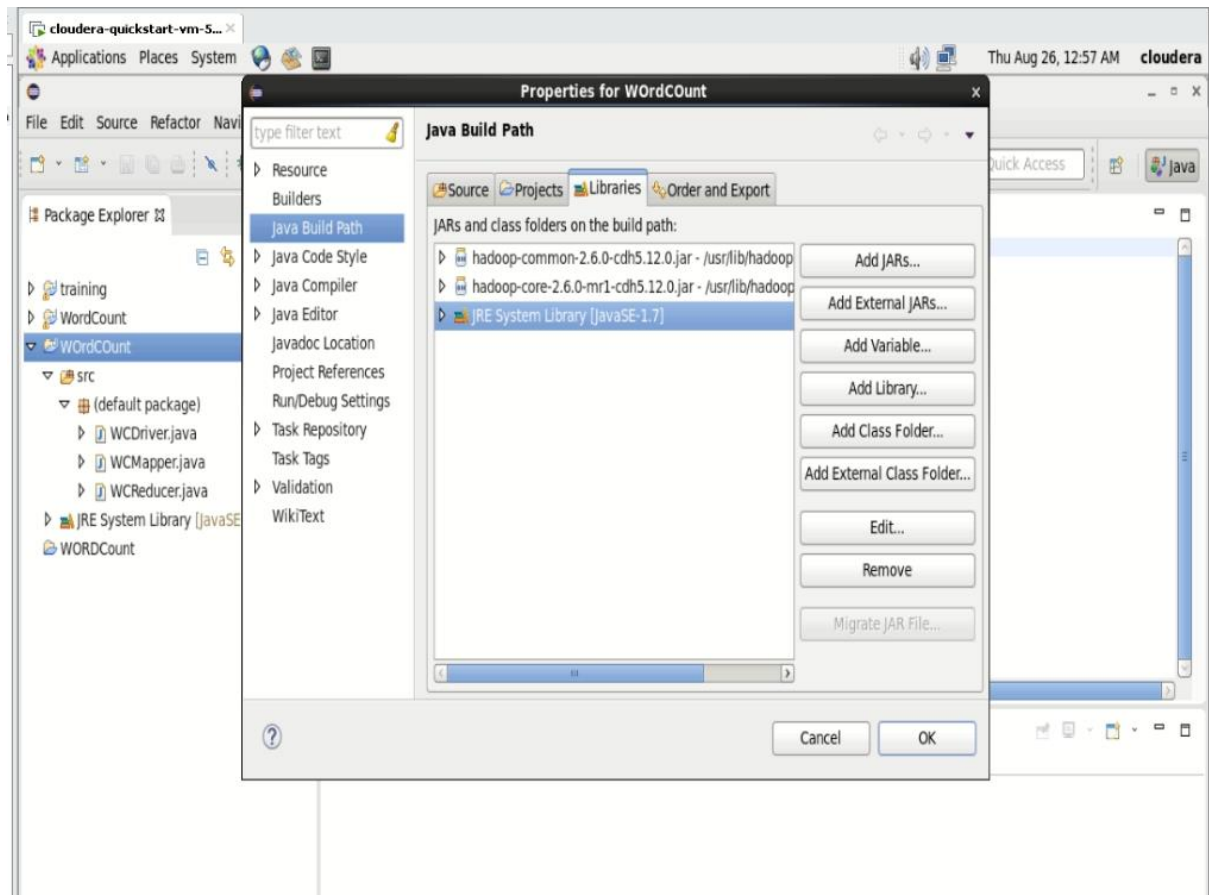
Tasks: 1. First Open Eclipse -> then select File -> New -> Java Project -> Name it WordCount -> then Finish.



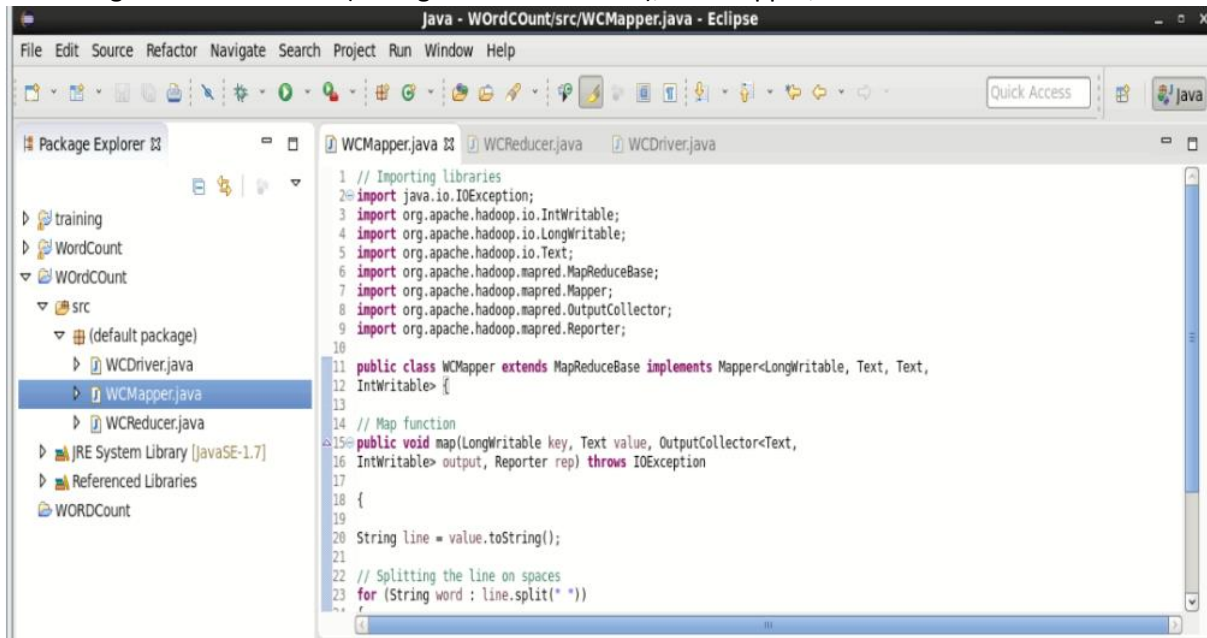
2. Create Three Java Classes into the project. Name them WCDriver(having the main function), WCMapper, WCReducer.



3. You have to include two Reference Libraries for that: Right Click on Project -> then select Build Path-> Click on Configure Build Path. You will see an “Add External JARs” option on the Right-Hand Side. Click on it and add the below mention files. You can find these files in /usr/lib/ 1. /usr/lib/hadoop-0.20-mapreduce/hadoop-core-2.6.0-mr1-cdh5.13.0.jar 2. /usr/lib/hadoop/hadoop-common-2.6.0-cdh5.13.0.jar

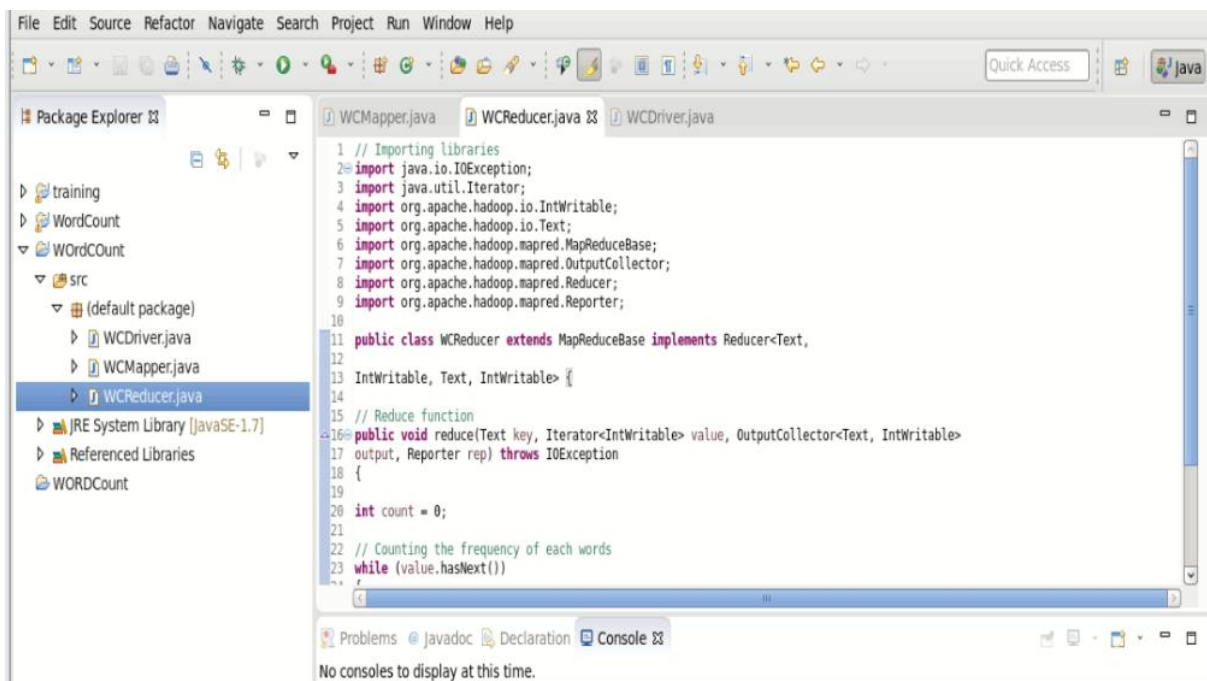


4. Writing Code in WCDriver(having the main function), WCMapper, WCReducer.



The screenshot shows the Eclipse IDE with the 'WCMapper.java' file open. The Package Explorer on the left shows the project structure: 'training' > 'WordCount' > 'src' > 'WCMapper.java'. The code in the editor is as follows:

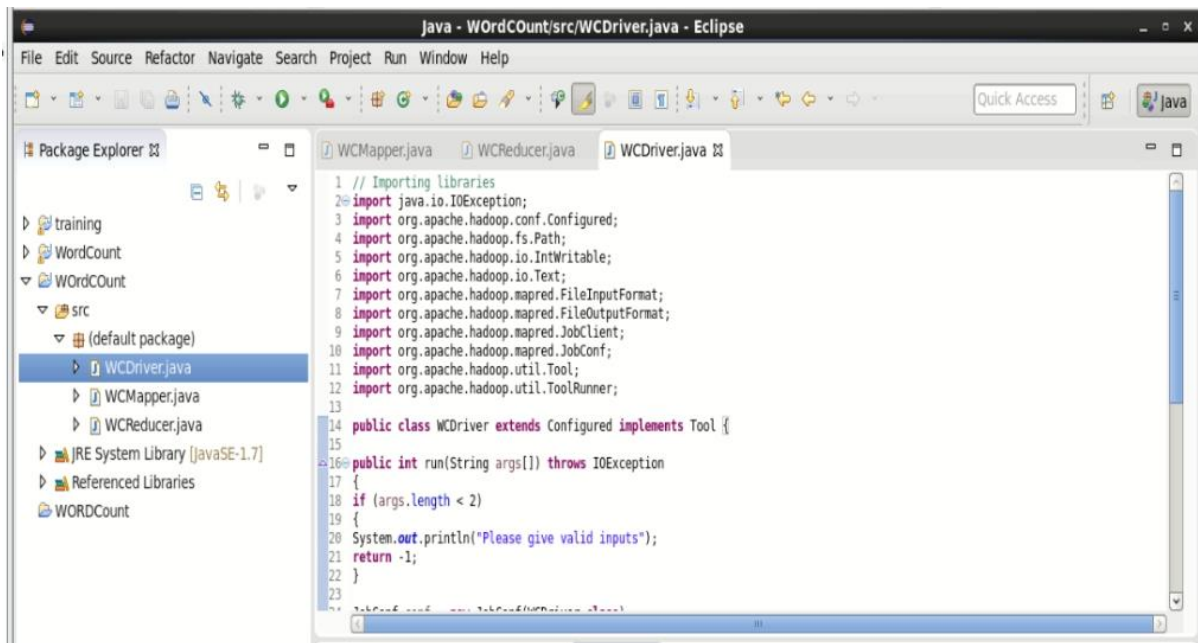
```
1 // Importing libraries
2 import java.io.IOException;
3 import org.apache.hadoop.io.IntWritable;
4 import org.apache.hadoop.io.LongWritable;
5 import org.apache.hadoop.io.Text;
6 import org.apache.hadoop.mapred.MapReduceBase;
7 import org.apache.hadoop.mapred.Mapper;
8 import org.apache.hadoop.mapred.OutputCollector;
9 import org.apache.hadoop.mapred.Reporter;
10
11 public class WCMapper extends MapReduceBase implements Mapper<LongWritable, Text, Text,
12 IntWritable> {
13
14 // Map function
15 public void map(LongWritable key, Text value, OutputCollector<Text,
16 IntWritable> output, Reporter rep) throws IOException
17 {
18 {
19 String line = value.toString();
20
21 // Splitting the line on spaces
22 for (String word : line.split(" "))
```



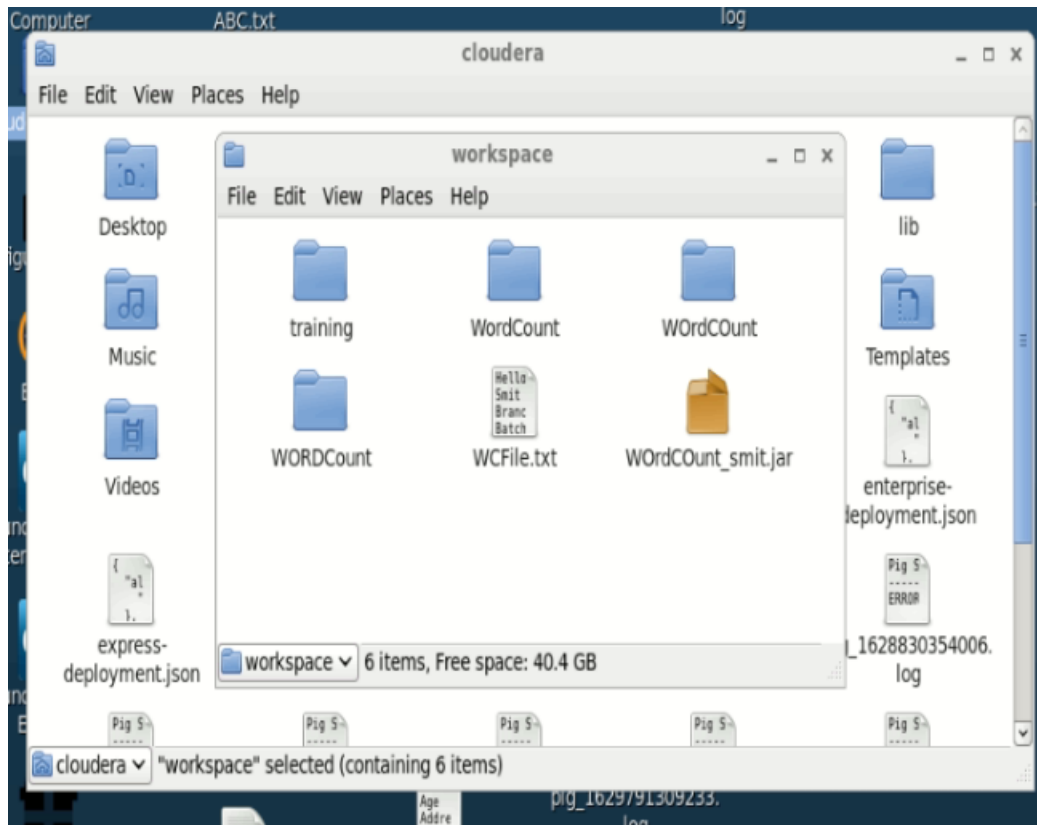
The screenshot shows the Eclipse IDE with the 'WCReducer.java' file open. The Package Explorer on the left shows the project structure: 'training' > 'WordCount' > 'src' > 'WCReducer.java'. The code in the editor is as follows:

```
1 // Importing libraries
2 import java.io.IOException;
3 import java.util.Iterator;
4 import org.apache.hadoop.io.IntWritable;
5 import org.apache.hadoop.io.Text;
6 import org.apache.hadoop.mapred.MapReduceBase;
7 import org.apache.hadoop.mapred.OutputCollector;
8 import org.apache.hadoop.mapred.Reducer;
9 import org.apache.hadoop.mapred.Reporter;
10
11 public class WCReducer extends MapReduceBase implements Reducer<Text,
12 IntWritable, Text, IntWritable> {
13
14 // Reduce function
15 public void reduce(Text key, Iterator<IntWritable> value, OutputCollector<Text, IntWritable>
16 output, Reporter rep) throws IOException
17 {
18 {
19 int count = 0;
20
21 // Counting the frequency of each words
22 while (value.hasNext())
```

At the bottom of the IDE, the 'Console' tab is selected, showing the message: 'No consoles to display at this time.'



5. After writing the MapReduce code into the classes, make a jar file. Right Click on Project -> Click on Export -> Select export destination as Jar File -> Name the jar File (WordCount.jar) -> Click on next -> at last Click on Finish. Now copy this file into the Workspace directory of Cloudera.



6. Open the terminal on CDH and change the directory to the workspace. You can do this by using "cd workspace/" command. Now, create a text file (WCFile.txt) and move it to HDFS. For that open terminal and write this code (remember you should be in the same directory as jar file you have created just now). Then, run this command to copy the file input file into the HDFS. "hadoop fs -put WCFile.txt WCFile.txt"

```
cloudera@quickstart:~/workspace
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ cd workspace/
bash: cd: workspace/: No such file or directory
[cloudera@quickstart Desktop]$ cd ..
[cloudera@quickstart ~]$ cd workspace/
[cloudera@quickstart workspace]$ touch WCFile.txt
[cloudera@quickstart workspace]$ gedit WCFile.txt
[cloudera@quickstart workspace]$ hadoop fs -put WCFile.txt WCFile.txt
[cloudera@quickstart workspace]$ hadoop fs -ls
Found 20 items
drwxr-xr-x - cloudera cloudera      0 2021-07-26 02:32 Dhrumil
-rw-r--r-- 1 cloudera cloudera      0 2021-07-26 02:05 DhrumilXYZ.txt
-rw-r--r-- 1 cloudera cloudera      0 2021-07-19 03:28 Documentsmy_details
.txt
drwxrwxrwx - cloudera cloudera      0 2021-08-19 02:34 ICT
-rw-r--r-- 1 cloudera cloudera      0 2021-08-17 02:20 Just_Empty_File.txt
-rw-r--r-- 1 cloudera cloudera      0 2021-08-19 03:16 Prac3
drwxr-xr-x - cloudera cloudera      0 2021-08-09 21:17 Prac6
drwxr-xr-x - cloudera cloudera      0 2021-08-10 00:26 Practical6
-rw-r--r-- 1 cloudera cloudera 53024484 2021-08-09 20:12 SalesJan2009.csv.zip
-rw-r--r-- 1 cloudera cloudera    103 2021-08-26 00:34 WCfile.txt
-rw-r--r-- 1 cloudera cloudera      0 2021-08-12 01:50 XYZ.txt
-rw-r--r-- 1 cloudera cloudera      0 2021-08-12 01:50 aaa.txt
-rw-r--r-- 1 cloudera cloudera 91162 2021-03-11 22:25 ad_analysis.csv
```



```
cloudera@quickstart:~/workspace
File Edit View Search Terminal Help
[cloudera@quickstart workspace]$ hadoop fs -cat WCFFile.txt
cat: 'WCFFile.txt': No such file or directory
[cloudera@quickstart workspace]$ hadoop fs -put WCFFile.txt WCFFile.txt
[cloudera@quickstart workspace]$ hadoop fs -ls
Found 21 items
drwxr-xr-x - cloudera cloudera      0 2021-07-26 02:32 Dhrumil
-rw-r--r-- 1 cloudera cloudera      0 2021-07-26 02:05 DhrumilXYZ.txt
-rw-r--r-- 1 cloudera cloudera      0 2021-07-19 03:28 Documentsmy_details
.txt
drwxrwxrwx - cloudera cloudera      0 2021-08-19 02:34 ICT
-rw-r--r-- 1 cloudera cloudera      0 2021-08-17 02:20 Just_Empty_File.txt
-rw-r--r-- 1 cloudera cloudera      0 2021-08-19 03:16 Prac3
drwxr-xr-x - cloudera cloudera      0 2021-08-09 21:17 Prac6
drwxr-xr-x - cloudera cloudera      0 2021-08-10 00:26 Practical6
-rw-r--r-- 1 cloudera cloudera 53024484 2021-08-09 20:12 SalesJan2009.csv.zip
-rw-r--r-- 1 cloudera cloudera    103 2021-08-26 00:35 WCFFile.txt
-rw-r--r-- 1 cloudera cloudera    103 2021-08-26 00:34 WCFfile.txt
-rw-r--r-- 1 cloudera cloudera      0 2021-08-12 01:50 XYZ.txt
-rw-r--r-- 1 cloudera cloudera      0 2021-08-12 01:50 aaa.txt
-rw-r--r-- 1 cloudera cloudera 91162 2021-03-11 22:25 ad_analysis.csv
drwxr-xr-x - cloudera cloudera      0 2021-07-19 03:19 dhrumil
-rw-r--r-- 1 cloudera cloudera      0 2021-07-19 03:04 my_details
drwxr-xr-x - cloudera cloudera      0 2021-08-09 21:21 pig_output_sales
```

```
cloudera@quickstart:~/workspace
File Edit View Search Terminal Help
drwxr-xr-x - cloudera cloudera      0 2021-08-09 21:17 Prac6
drwxr-xr-x - cloudera cloudera      0 2021-08-10 00:26 Practical6
-rw-r--r-- 1 cloudera cloudera 53024484 2021-08-09 20:12 SalesJan2009.csv.zip
-rw-r--r-- 1 cloudera cloudera    103 2021-08-26 00:35 WCFFile.txt
-rw-r--r-- 1 cloudera cloudera    103 2021-08-26 00:34 WCFfile.txt
-rw-r--r-- 1 cloudera cloudera      0 2021-08-12 01:50 XYZ.txt
-rw-r--r-- 1 cloudera cloudera      0 2021-08-12 01:50 aaa.txt
-rw-r--r-- 1 cloudera cloudera 91162 2021-03-11 22:25 ad_analysis.csv
drwxr-xr-x - cloudera cloudera      0 2021-07-19 03:19 dhrumil
-rw-r--r-- 1 cloudera cloudera      0 2021-07-19 03:04 my_details
drwxr-xr-x - cloudera cloudera      0 2021-08-09 21:21 pig_output_sales
drwxr-xr-x - cloudera cloudera      0 2021-08-10 01:15 pig_output_sales1
drwxr-xr-x - cloudera cloudera      0 2021-08-12 22:56 prac_7
drwxr-xr-x - cloudera cloudera      0 2021-08-23 20:31 prac_8
drwxr-xr-x - cloudera cloudera      0 2021-08-24 02:15 smitpatel
[cloudera@quickstart workspace]$ hadoop fs -cat WCFFile.txt
Hello Brother,
Smit here,
Branch : Data Science
Batch : 54
Subject : Big data and analytics
Goal : IAS
[cloudera@quickstart workspace]$
```

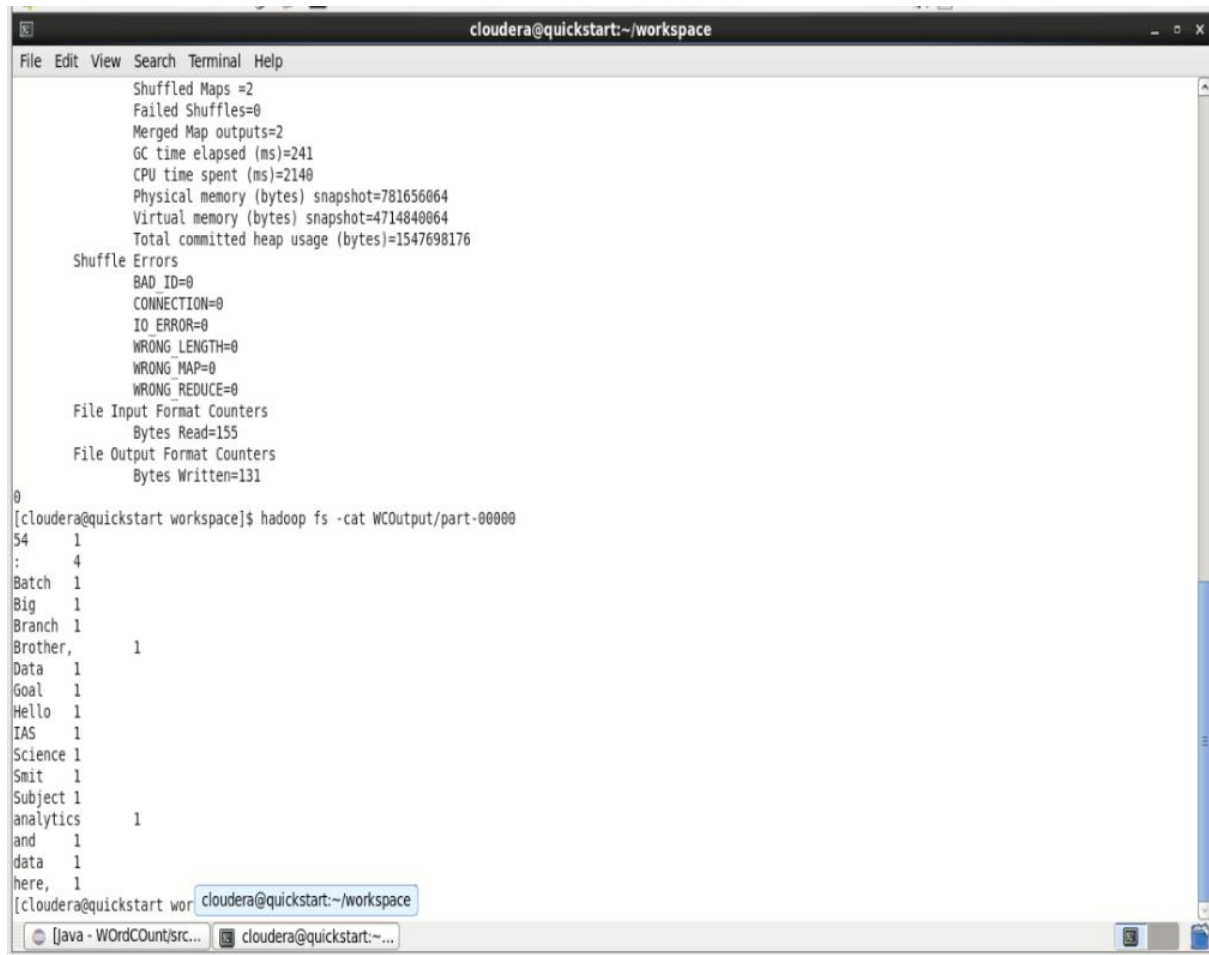
7. Run this command to store the output of map reduce in hadoop's directory. Command: `hadoop jar WordCount.jar WCDriver WCFile.txt WCOOutput`



```
cloudera-quickstart-vm-5...
Applications Places System Thu Aug 26, 1:18 AM cloudera
cloudera@quickstart:~/workspace
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ cd ..
[cloudera@quickstart ~]$ cd workspace
[cloudera@quickstart workspace]$ hadoop jar WordCount_smit.jar WCDriver
Please give valid inputs
-1
[cloudera@quickstart workspace]$ hadoop jar WordCount_smit.jar WCDriver WCFile.txt WCOOutput
21/08/26 01:13:40 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
21/08/26 01:13:41 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
21/08/26 01:13:41 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
21/08/26 01:13:41 INFO mapred.FileInputFormat: Total input paths to process : 1
21/08/26 01:13:41 INFO mapreduce.JobSubmitter: number of splits:2
21/08/26 01:13:41 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1629960043242_0001
21/08/26 01:13:42 INFO impl.YarnClientImpl: Submitted application application_1629960043242_0001
21/08/26 01:13:42 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1629960043242_0001/
21/08/26 01:13:42 INFO mapreduce.Job: Running job: job_1629960043242_0001
21/08/26 01:13:50 INFO mapreduce.Job: Job job_1629960043242_0001 running in uber mode : false
21/08/26 01:13:50 INFO mapreduce.Job: map 0% reduce 0%
21/08/26 01:13:56 INFO mapreduce.Job: map 50% reduce 0%
21/08/26 01:13:57 INFO mapreduce.Job: map 100% reduce 0%
21/08/26 01:14:02 INFO mapreduce.Job: map 100% reduce 100%
21/08/26 01:14:03 INFO mapreduce.Job: Job job_1629960043242_0001 completed successfully
21/08/26 01:14:03 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=229
    FILE: Number of bytes written=375973
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=371
    HDFS: Number of bytes written=131
    HDFS: Number of read operations=9
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=2
    Launched reduce tasks=1
    Data-local map tasks=2
    Total time spent by all maps in occupied slots (ms)=8167
```

```
cloudera@quickstart:~/workspace
File Edit View Search Terminal Help
Total time spent by all map tasks (ms)=8167
Total time spent by all reduce tasks (ms)=2788
Total vcore-milliseconds taken by all map tasks=8167
Total vcore-milliseconds taken by all reduce tasks=2788
Total megabyte-milliseconds taken by all map tasks=8363008
Total megabyte-milliseconds taken by all reduce tasks=2854912
Map-Reduce Framework
  Map input records=6
  Map output records=20
  Map output bytes=183
  Map output materialized bytes=235
  Input split bytes=216
  Combine input records=0
  Combine output records=0
  Reduce input groups=17
  Reduce shuffle bytes=235
  Reduce input records=20
  Reduce output records=17
  Spilled Records=40
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=241
  CPU time spent (ms)=2140
  Physical memory (bytes) snapshot=781656064
  Virtual memory (bytes) snapshot=4714840064
  Total committed heap usage (bytes)=1547698176
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=155
File Output Format Counters
  Bytes Written=131
0
[java - WOrdCount/src... cloudera@quickstart:~/...
```

8. After Executing the code, you can see the result in WCOOutput file or by writing following command on terminal. Command: `hadoop fs -cat WCOOutput/part-00000`



```
cloudera@quickstart:~/workspace
File Edit View Search Terminal Help
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=241
CPU time spent (ms)=2140
Physical memory (bytes) snapshot=781656064
Virtual memory (bytes) snapshot=4714840064
Total committed heap usage (bytes)=1547698176
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=155
File Output Format Counters
Bytes Written=131
0
[cloudera@quickstart workspace]$ hadoop fs -cat WCOOutput/part-00000
54      1
:        4
Batch   1
Big     1
Branch  1
Brother,      1
Data      1
Goal      1
Hello     1
IAS       1
Science   1
Smit      1
Subject   1
analytics      1
and        1
data       1
here,      1
[cloudera@quickstart wor cloudera@quickstart:~/workspace
[java - WordCount/src... cloudera@quickstart:~/...
```