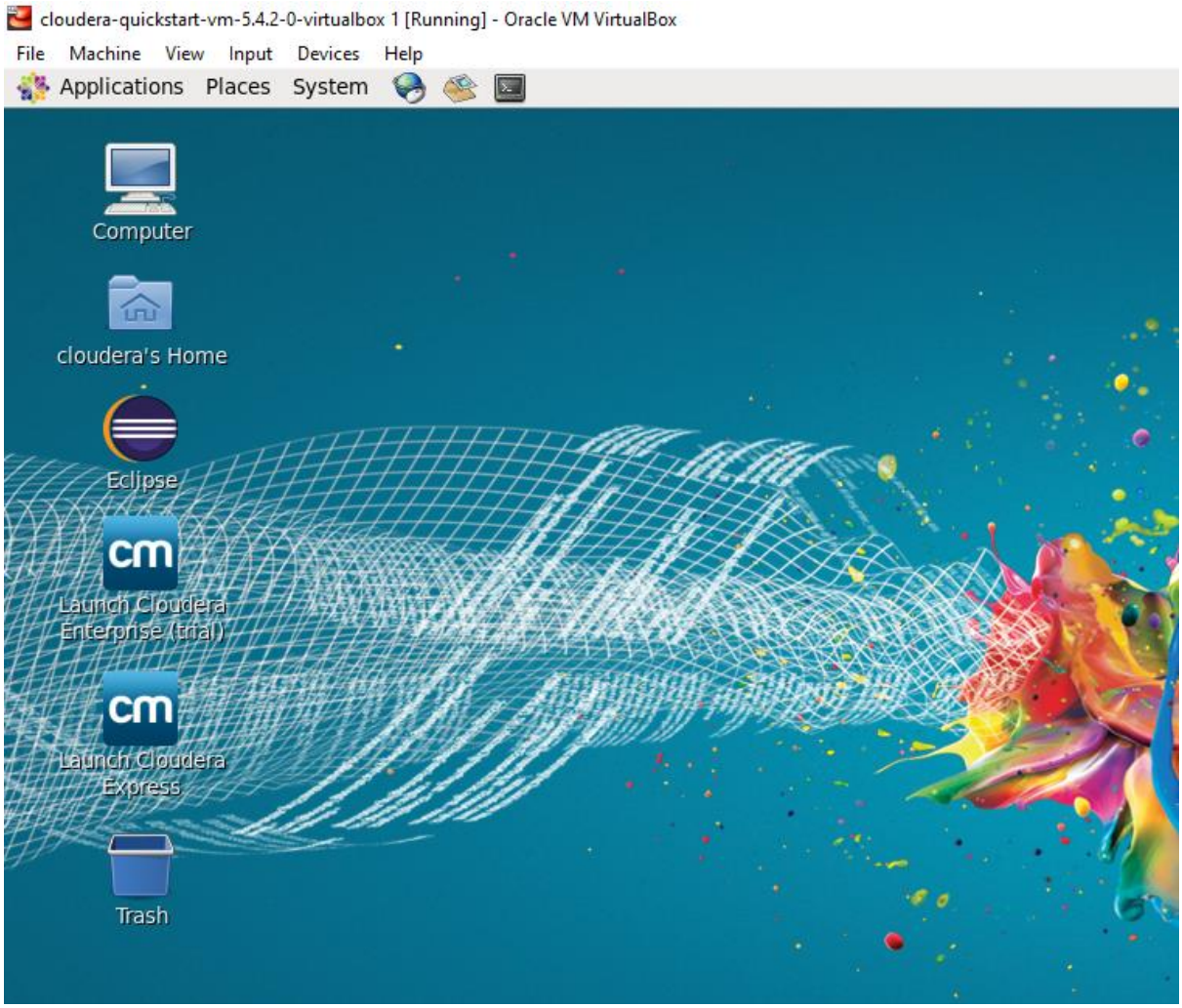
Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

Subject: Big Data Technology
Practical_3: Map Reduce

Steps for Word Count in Cloudera: (Without Combiner)

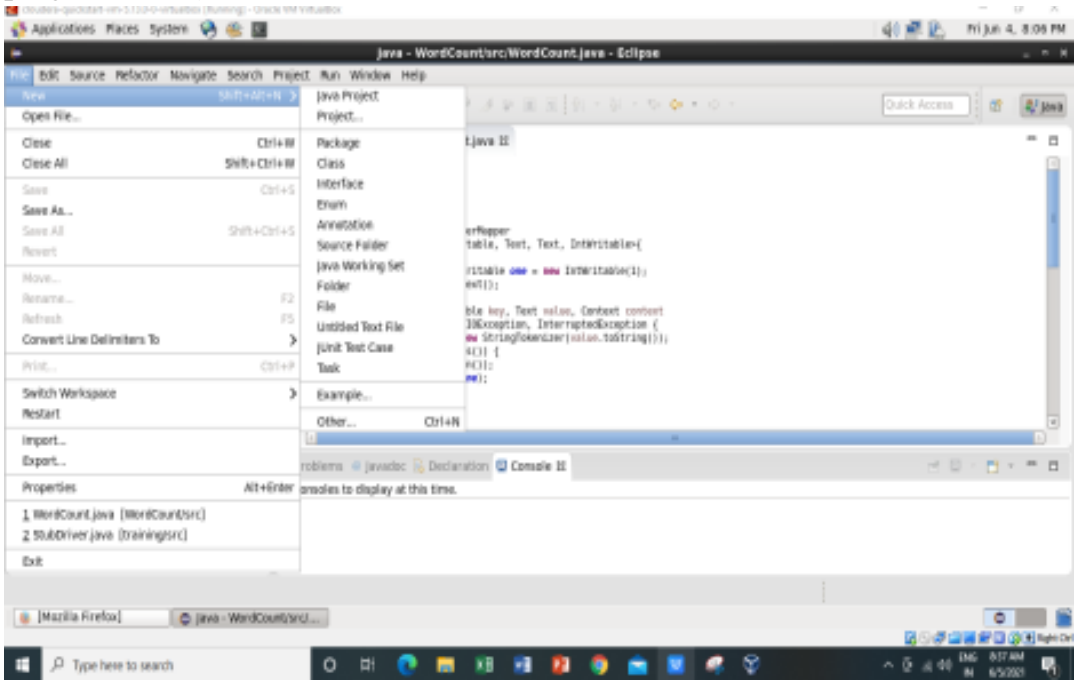
1) Open virtual box and then start cloudera quickstart



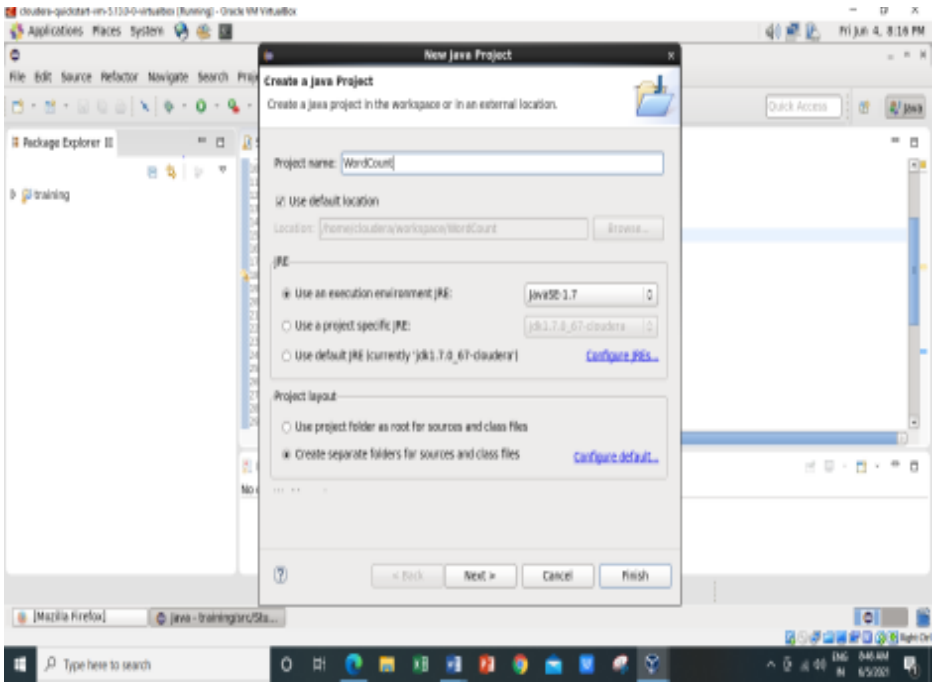
Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

2)Open Eclipse present on the cloudera desktop

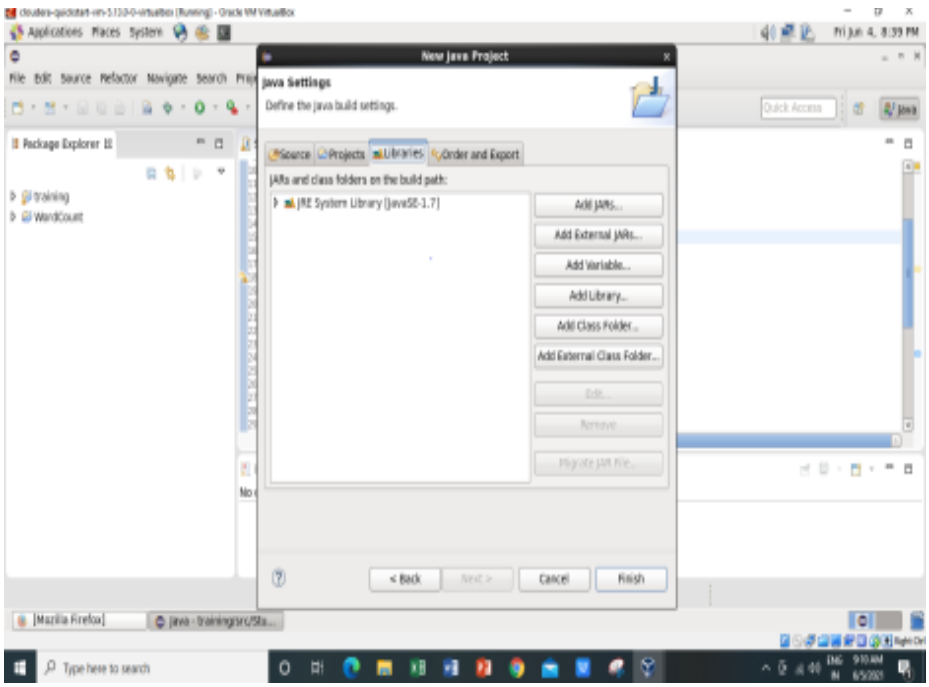
3) Create a new Java project clicking: File -> New -> Project -> Java Project -> Next ("WordCount" is the project name).



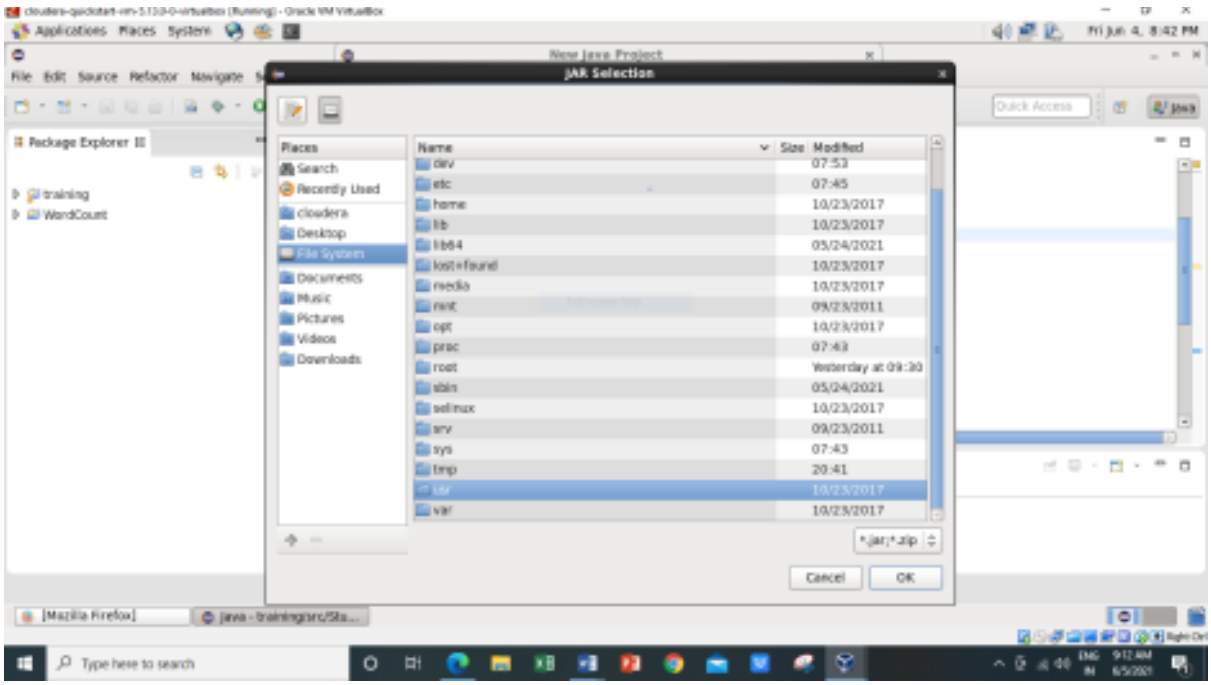
Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

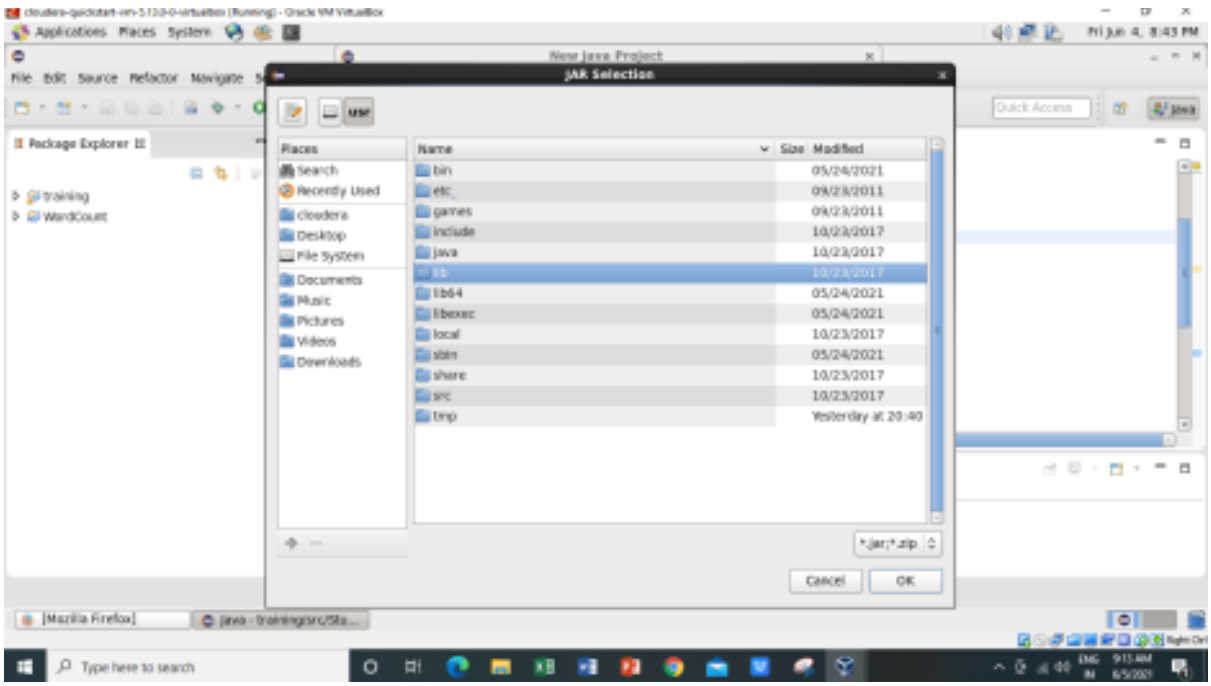


4) Adding the Hadoop libraries to the project Click on Libraries -> Add External JARs Click on File System -> usr -> lib -> hadoop Select all the libraries (JAR Files) -> click OK Click on Add External jars, -> client -> select all jar files -> ok -> Finish

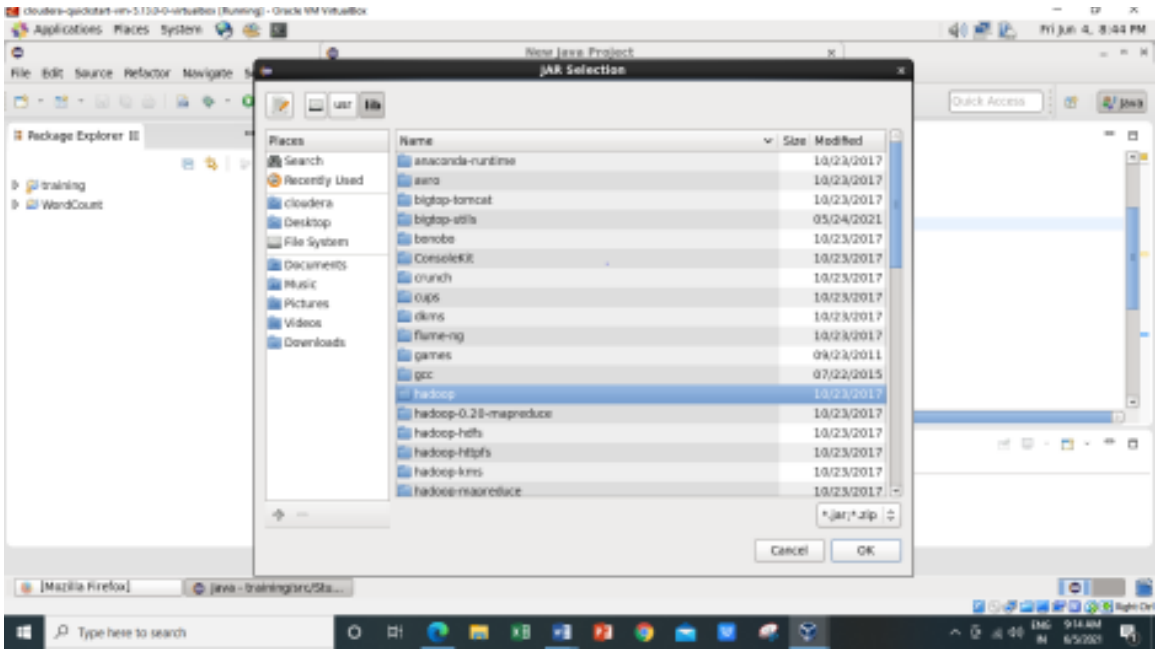


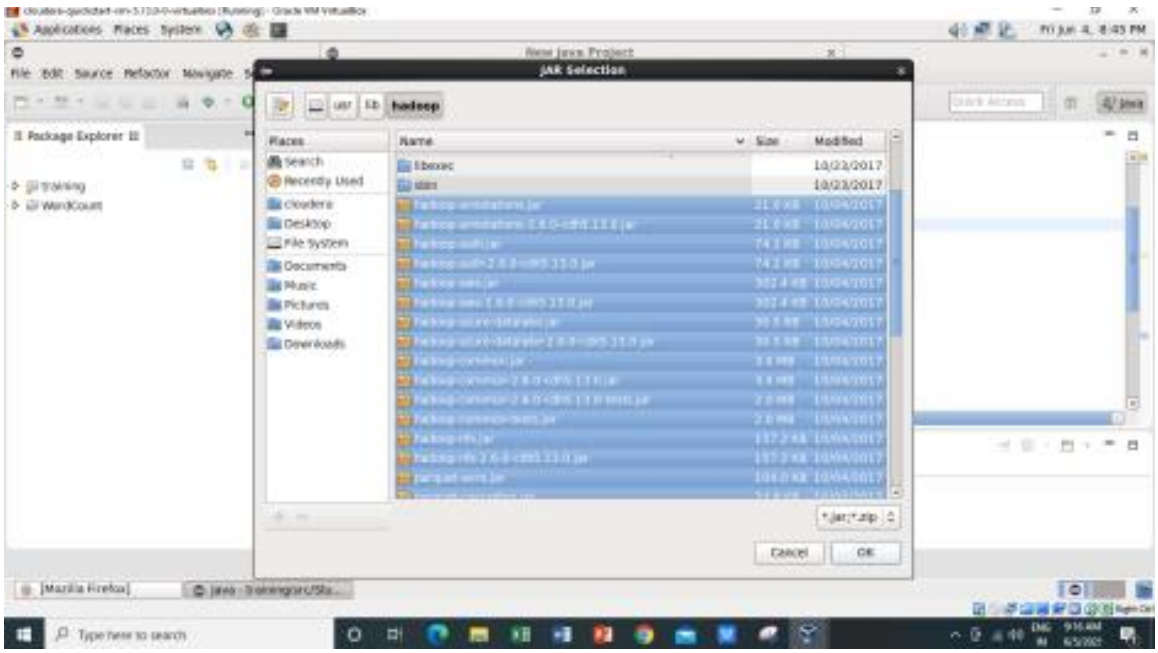
Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2



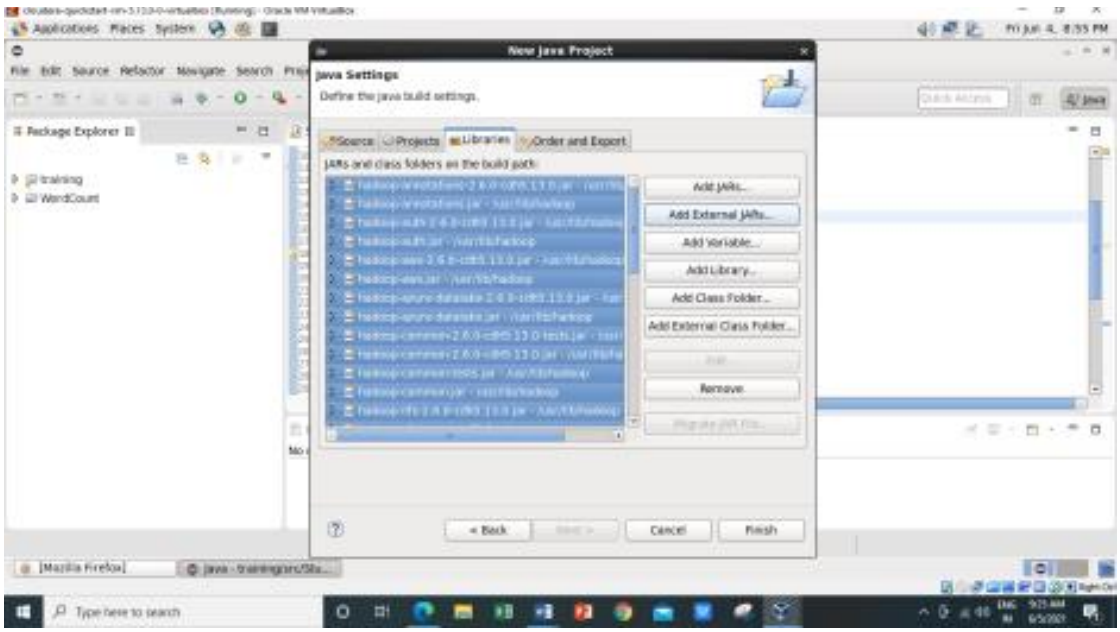


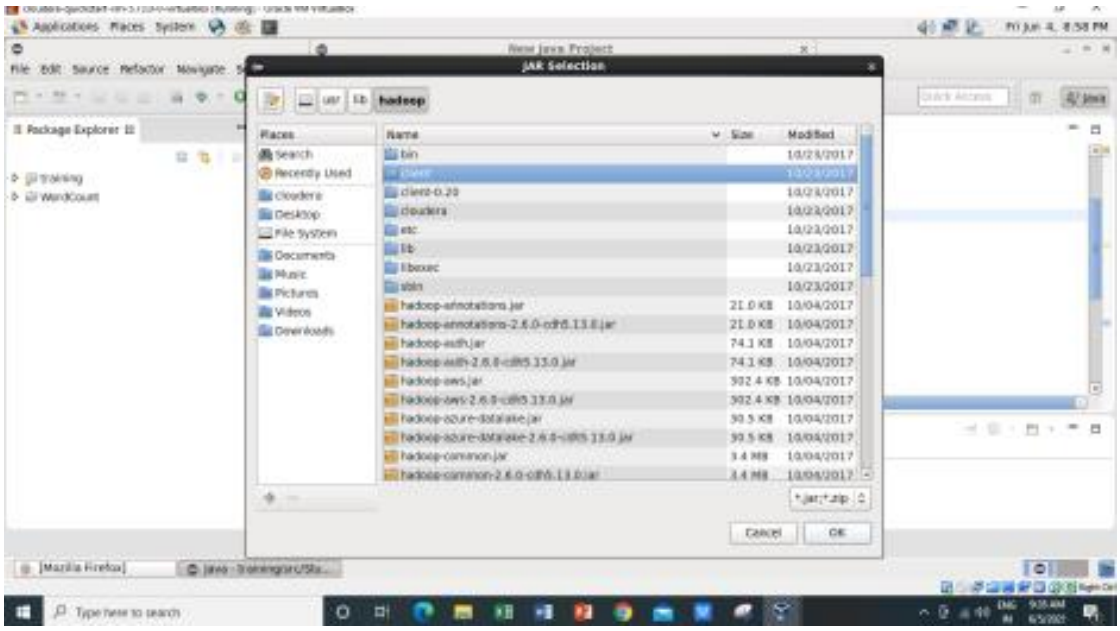
Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2



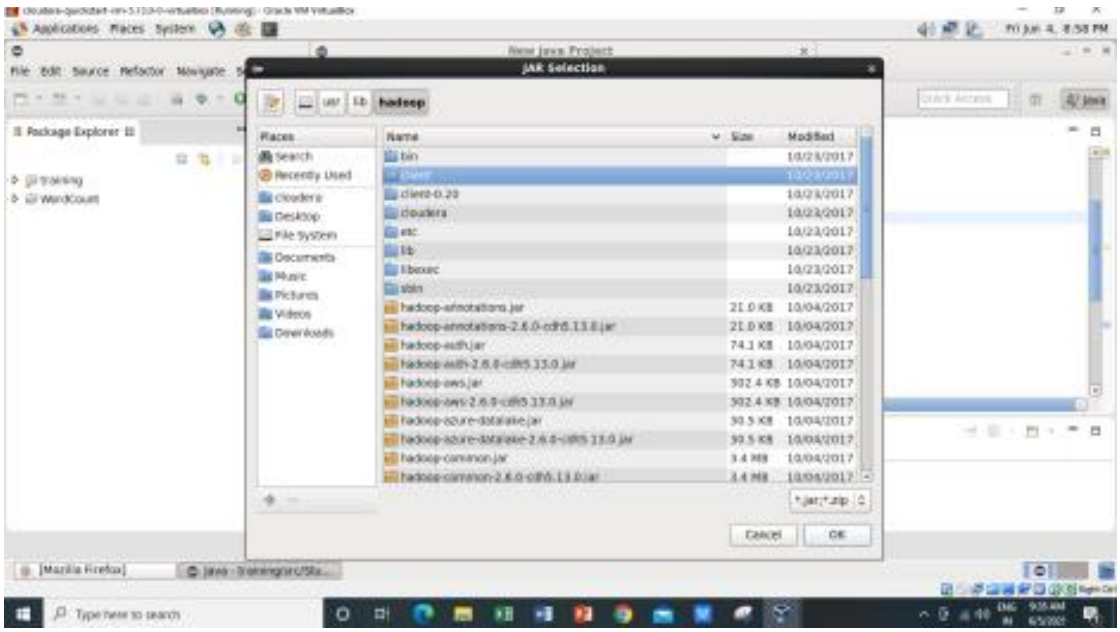


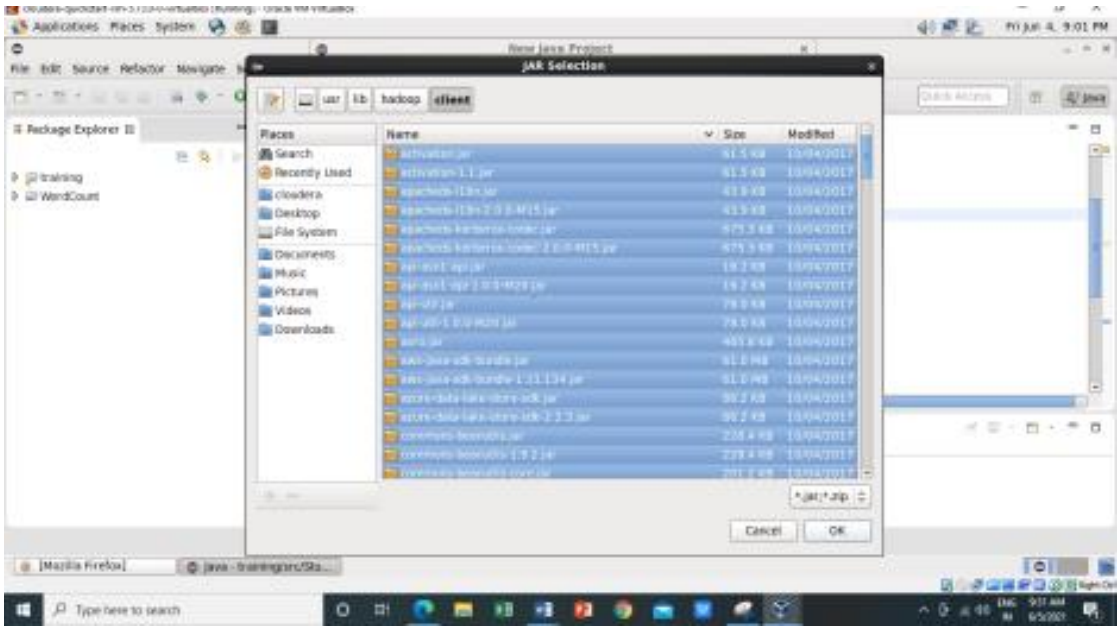
Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2



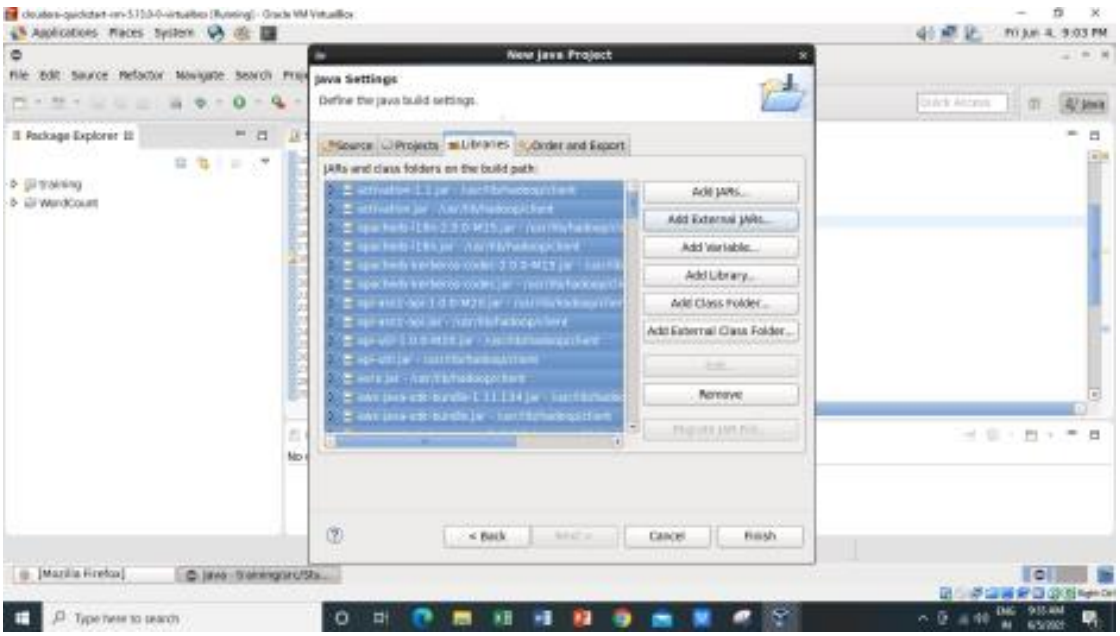


Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

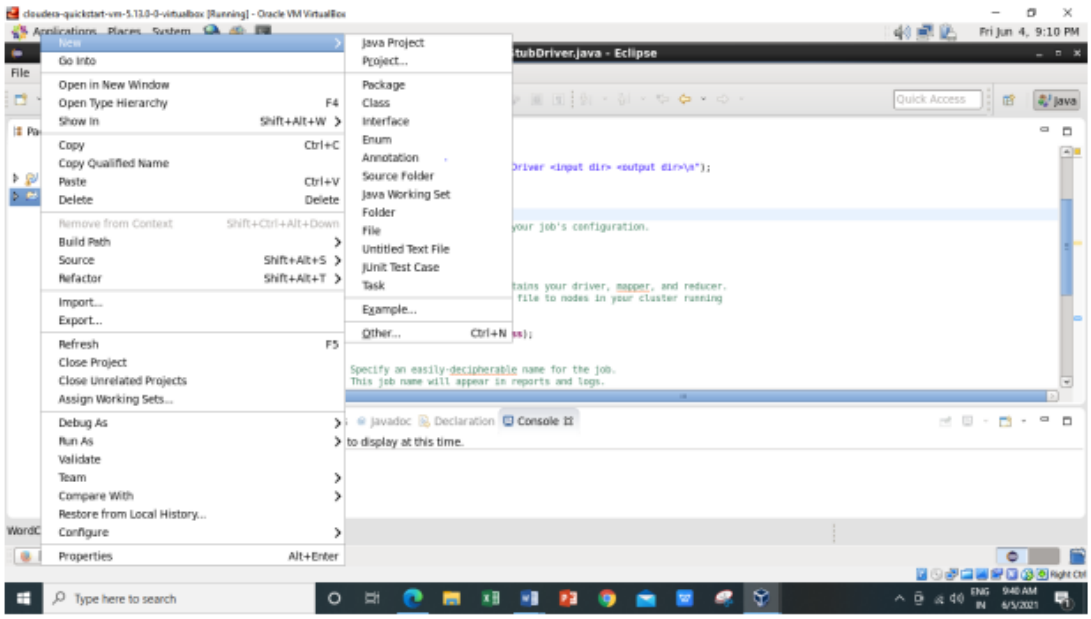




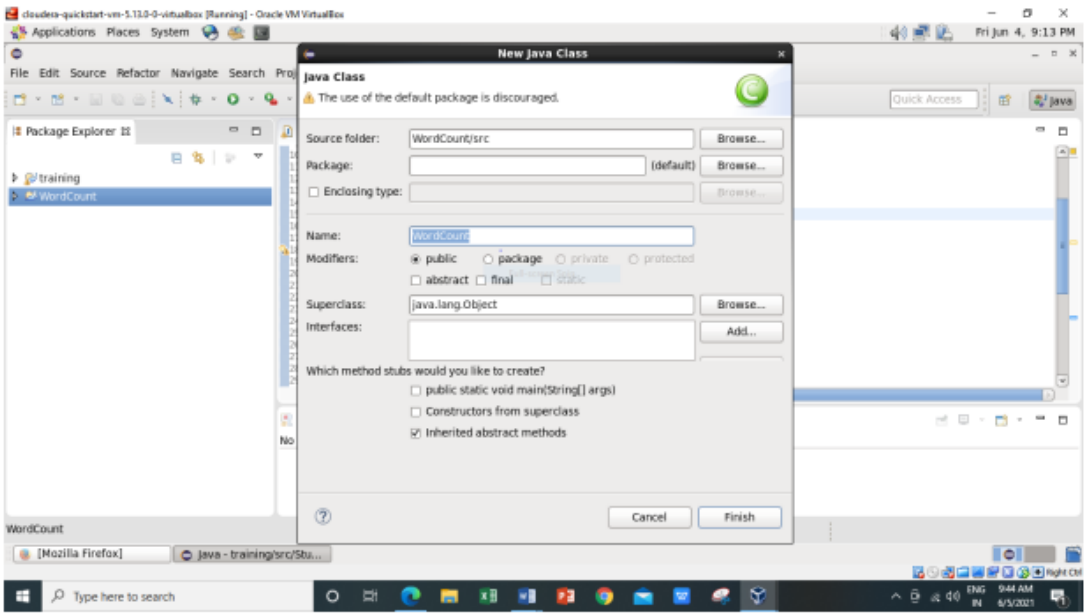
Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

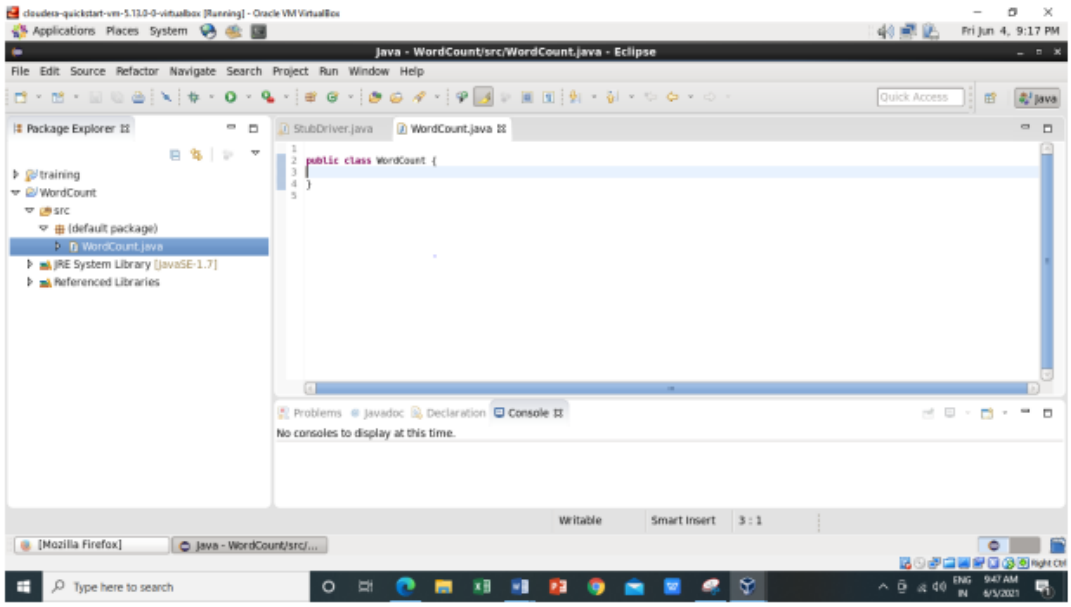


5) Right Click on the name of Project “WordCount” -> New -> class Don’t write anything for package Write Name Textbox write “WordCount” -> Finish Then WordCount.java window will pop up

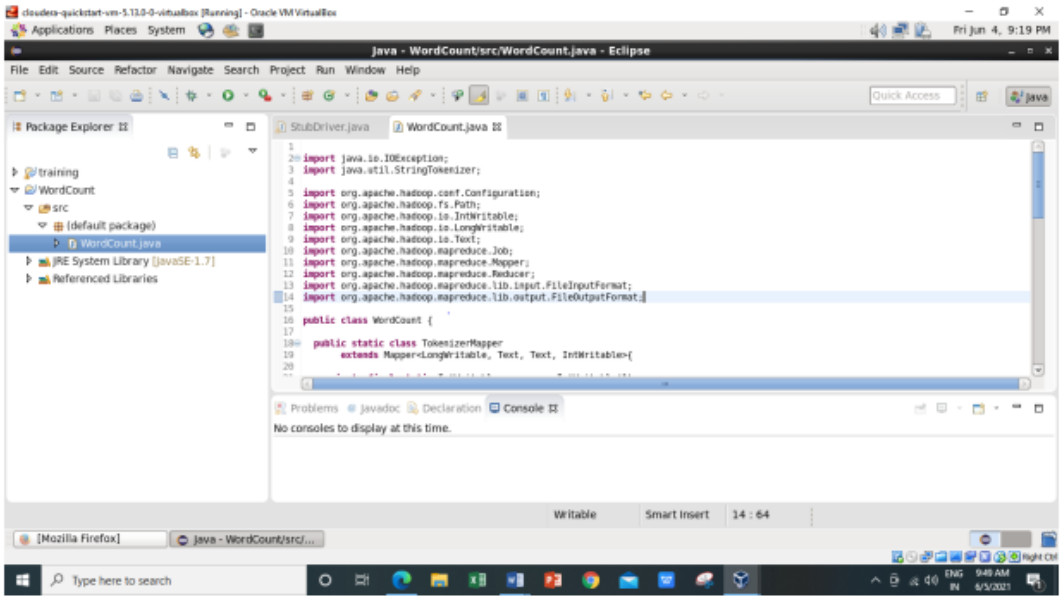


Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2





Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2



Source code:

Packages

```
1  import java.io.IOException;
2  import java.util.StringTokenizer;
3
4  import org.apache.hadoop.conf.Configuration;
5  import org.apache.hadoop.fs.Path;
6  import org.apache.hadoop.io.IntWritable;
7  import org.apache.hadoop.io.LongWritable;
8  import org.apache.hadoop.io.Text;
9  import org.apache.hadoop.mapreduce.Job;
0  import org.apache.hadoop.mapreduce.Mapper;
1  import org.apache.hadoop.mapreduce.Reducer;
2  import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
3  import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
```

Mapper Logic

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

```
15
16  public class WordCount {
17
18    public static class TokenizerMapper
19         extends Mapper<LongWritable, Text, Text, IntWritable>{
20
21      private final static IntWritable one = new IntWritable(1);
22      private Text word = new Text();
23
24      public void map(LongWritable key, Text value, Context context
25                      ) throws IOException, InterruptedException {
26        StringTokenizer itr = new StringTokenizer(value.toString());
27        while (itr.hasMoreTokens()) {
28          word.set(itr.nextToken());
29          context.write(word, one);
30        }
31      }
32    }
33
```

Reducer logic

```
34    public static class IntSumReducer
35         extends Reducer<Text,IntWritable,Text,IntWritable> {
36      private IntWritable result = new IntWritable();
37
38      public void reduce(Text key, Iterable<IntWritable> values,
39                         Context context
40                         ) throws IOException, InterruptedException {
41        int sum = 0;
42        for (IntWritable val : values) {
43          sum += val.get();
44        }
45        result.set(sum);
46        context.write(key, result);
47      }
48    }
```

Main function

We are running the code without combiner. That is why we commented the combiner line in main function.

```
50    public static void main(String[] args) throws Exception {
51      Configuration conf = new Configuration();
52      Job job = Job.getInstance(conf, "word count");
53      job.setJarByClass(WordCount.class);
54      job.setMapperClass(TokenizerMapper.class);
55      //job.setCombinerClass(IntSumReducer.class);
56      job.setReducerClass(IntSumReducer.class);
57      job.setOutputKeyClass(Text.class);
58      job.setOutputValueClass(IntWritable.class);
59      FileInputFormat.addInputPath(job, new Path(args[0]));
60      FileOutputFormat.setOutputPath(job, new Path(args[1]));
61      System.exit(job.waitForCompletion(true) ? 0 : 1);
62    }
63  }
```

6) Right Click on the project name WordCount -> Export -> Java ->

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

JAR File -> Next -> for select the export destination for JAR file: browse -> Name : WordCount.jar -> save in folder -> cloudera -> Finish -> OK

Here listing all the directory present in hdfs using **hdfs dfs -ls /** command.

```
[cloudera@quickstart ~]$ ls
aaa               Documents                   lib              Videos
aakansha          Downloads                   Music            WordCount.jar
ABC               eclipse                     Pictures         workspace
bbb               enterprise-deployment.json  Public           x
cloudera-manager  express-deployment.json     Templates
cm_api.py         kerberos                    untitled folder
[cloudera@quickstart ~]$ hdfs dfs - ls /
ls: Unknown command
Did you mean -ls?  This command begins with a dash.
[cloudera@quickstart ~]$ hdfs dfs-ls/
Error: Could not find or load main class dfs-ls.
[cloudera@quickstart ~]$ hdfs dfs -ls/
-ls/: Unknown command
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 8 items
drwxr-xr-x   - hbase    supergroup          0 2022-02-17 19:19 /hbase
drwxr-xr-x   - cloudera supergroup          0 2022-02-14 20:56 /rjc2122
drwxr-xr-x   - cloudera supergroup          0 2022-02-14 20:10 /rjclocal
drwxr-xr-x   - cloudera supergroup          0 2022-02-14 19:58 /rjcnew
drwxr-xr-x   - solr     solr                0 2015-06-09 03:38 /solr
drwxrwxrwx   - hdfs     supergroup          0 2022-02-07 21:21 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:38 /user
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:36 /var
[cloudera@quickstart ~]$ hdfs dfs -mkdir /inputdir
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 8 items
drwxr-xr-x   - hbase    supergroup          0 2022-02-17 19:19 /hbase
drwxr-xr-x   - cloudera supergroup          0 2022-02-14 20:56 /rjc2122
drwxr-xr-x   - cloudera supergroup          0 2022-02-14 20:10 /rjclocal
drwxr-xr-x   - cloudera supergroup          0 2022-02-14 19:58 /rjcnew
drwxr-xr-x   - solr     solr                0 2015-06-09 03:38 /solr
drwxrwxrwx   - hdfs     supergroup          0 2022-02-07 21:21 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:38 /user
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:36 /var
[cloudera@quickstart ~]$ hdfs dfs -mkdir /inputdir
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 9 items
drwxr-xr-x   - hbase    supergroup          0 2022-02-17 19:19 /hbase
drwxr-xr-x   - cloudera supergroup          0 2022-02-17 20:05 /inputdir
drwxr-xr-x   - cloudera supergroup          0 2022-02-14 20:56 /rjc2122
drwxr-xr-x   - cloudera supergroup          0 2022-02-14 20:10 /rjclocal
drwxr-xr-x   - cloudera supergroup          0 2022-02-14 19:58 /rjcnew
drwxr-xr-x   - solr     solr                0 2015-06-09 03:38 /solr
drwxrwxrwx   - hdfs     supergroup          0 2022-02-07 21:21 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:38 /user
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:36 /var
```

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

```
[cloudera@quickstart ~]$ hdfs dfs -ls /inputdir
Found 1 items
-rw-r--r--   1 cloudera supergroup        864 2022-02-17 20:09 /inputdir/ABC
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hdfs dfs-cat/inputdir/ABC command
Error: Could not find or load main class dfs-cat.inputdir.ABC
[cloudera@quickstart ~]$ hdfs dfs -cat /inputdir/ABC
Greed will always lead to downfall.
There once was a king named Midas who did a good deed for a Satyr. And he was th
en granted a wish by Dionysus, the god of wine.
For his wish, Midas asked that whatever he touched would turn to gold. Despite D
ionysus' efforts to prevent it, Midas pleaded that this was a fantastic wish, an
d so, it was bestowed.
Excited about his newly-earned powers, Midas started touching all kinds of thing
s, turning each item into pure gold.
But soon, Midas became hungry. As he picked up a piece of food, he found he coul
dn't eat it. It had turned to gold in his hand.
Hungry, Midas groaned, "I'll starve! Perhaps this was not such an excellent wish
 after all!"
Seeing his dismay, Midas' beloved daughter threw her arms around him to comfort
him, and she, too, turned to gold. "The golden touch is no blessing," Midas crie
d.
```

Running Mapreduce Program on Hadoop, syntax is hadoop jar jarFileName.jar ClassName /InputFileAddress /outputdir

i.e. **hadoop jar /home/cloudera/WordCount.jar WordCount /inputdir/ABC /outputdir**

```
[cloudera@quickstart ~]$ hadoop jar /home/cloudera/WordCount.jar WordCount /inpu
tdir/ABC /outputdir
22/02/17 20:19:34 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
22/02/17 20:19:35 WARN mapreduce.JobSubmitter: Hadoop command-line option parsin
g not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
22/02/17 20:19:35 INFO input.FileInputFormat: Total input paths to process : 1
22/02/17 20:19:36 INFO mapreduce.JobSubmitter: number of splits:1
22/02/17 20:19:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
45154347193_0001
22/02/17 20:19:36 INFO impl.YarnClientImpl: Submitted application application_16
45154347193_0001
22/02/17 20:19:37 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1645154347193_0001/
22/02/17 20:19:37 INFO mapreduce.Job: Running job: job_1645154347193_0001
22/02/17 20:19:49 INFO mapreduce.Job: Job job_1645154347193_0001 running in uber
 mode : false
22/02/17 20:19:49 INFO mapreduce.Job:  map 0% reduce 0%
22/02/17 20:19:57 INFO mapreduce.Job:  map 100% reduce 0%
22/02/17 20:20:04 INFO mapreduce.Job:  map 100% reduce 100%
22/02/17 20:20:04 INFO mapreduce.Job: Job job_1645154347193_0001 completed succe
ssfully
22/02/17 20:20:04 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=1421
                FILE: Number of bytes written=223429
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=973
                HDFS: Number of bytes written=947
                HDFS: Number of read operations=6
```

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

```
        HDFS: Number of write operations=2
Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=5718
        Total time spent by all reduces in occupied slots (ms)=4872
        Total time spent by all map tasks (ms)=5718
        Total time spent by all reduce tasks (ms)=4872
        Total vcore-seconds taken by all map tasks=5718
        Total vcore-seconds taken by all reduce tasks=4872
        Total megabyte-seconds taken by all map tasks=5855232
        Total megabyte-seconds taken by all reduce tasks=4988928
Map-Reduce Framework
        Map input records=7
        Map output records=154
        Map output bytes=1480
        Map output materialized bytes=1421
        Input split bytes=109
        Combine input records=154
        Combine output records=117
        Reduce input groups=117
        Reduce shuffle bytes=1421
        Reduce input records=117
        Reduce output records=117
        Spilled Records=234
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=107
        CPU time spent (ms)=1050
        Physical memory (bytes) snapshot=334733312
        Virtual memory (bytes) snapshot=3007062016
        Total committed heap usage (bytes)=226365440
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=864
File Output Format Counters
        Bytes Written=947
```

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

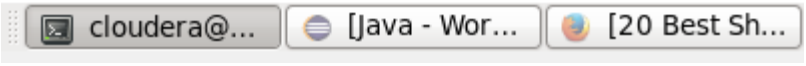
```
[cloudera@quickstart ~]$ hdfs dfs -cat /outputdir/part-r-00000
And     1
As      1
But     1
Despite 1
Dionysus,       1
Dionysus’       1
Excited 1
For     1
Greed   1
Hungry, 1
It      1
Midas   7
Midas’  1
Perhaps 1
Satyr.  1
Seeing  1
There   1
a       6
about   1
after   1
all     1
all!”   1
always  1
an      1
and     2
arms    1
around  1
asked   1
became  1
beloved 1
bestowed.       1
blessing,”      1
by      1
comfort 1
couldn’t        1
cried.  1
daughter        1
deed    1
did     1
dismay, 1
downfall.       1
each    1
eat     1
efforts 1
excellent       1
fantastic       1
food,   1
for     1
found   1
```

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

```
golden  1
good    1
granted 1
groaned,        1
had     1
hand.   1
he      5
her     1
him     1
him,    1
his     4
hungry. 1
in      1
into    1
is      1
it      1
it,     1
it.     1
item    1
kinds   1
king    1
lead    1
named   1
newly-earned    1
no      1
not     1
of      3
once    1
picked  1
piece   1
pleaded 1
powers, 1
prevent 1
pure    1
she,    1
so,     1
soon,   1
started 1
starve! 1
such    1
that    2
the     1
then    1
things, 1
this    2
threw   1
to      6
too,    1
touch   1
touched 1
touching        1
```

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

```
touching        1
turn    1
turned  2
turning 1
up      1
was     5
whatever        1
who     1
will    1
wine.   1
wish    2
wish,   2
would   1
“I’ll   1
“The    1
[cloudera@quickstart ~]$
```



As we can see in the above output,
Combine input records=0
Combine output records=0
We are getting this because we have commented the Combiner line in main function.
And Reduce shuffle bytes coming as,
Reduce shuffle bytes=1876
So when we are not using combiner 1876 bytes acting as an input for the reducer.

Implementation of WordCount problem using Hadoop MapReduce (With Combiner) in Eclipse:

We will perform the same steps as we have done above for WordCount (without using combiner) in that we just uncommenting the combiner line in main function.
And will delete the WordCount.jar file in which all jar files are present from **/home/cloudera.**
We have successfully deleted the WordCount.jar file

Running Mapreduce Program on Hadoop, syntax is
hadoop jar jarFileName.jar ClassName /InputFileAddress /outputdir
i.e. hadoop jar /home/cloudera/WordCount.jar WordCount /inputdir/abc /XYZ

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

here I am using the same input file 'abc' which I have created earlier for WordCount example (Without Combiner). For every execution of this program we need to delete the output directory or give a new name to the output directory every time.
So here I am giving the new name to the output directory as 'XYZ'.

```
[cloudera@quickstart ~]$ pwd
/home/cloudera
[cloudera@quickstart ~]$ hadoop jar /home/cloudera/WordCount.jar WordCount /inpu
tdir/ABC /XYZ
22/02/17 20:41:44 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
22/02/17 20:41:45 WARN mapreduce.JobSubmitter: Hadoop command-line option parsin
g not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
22/02/17 20:41:45 INFO input.FileInputFormat: Total input paths to process : 1
22/02/17 20:41:45 INFO mapreduce.JobSubmitter: number of splits:1
22/02/17 20:41:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
45154347193_0002
22/02/17 20:41:45 INFO impl.YarnClientImpl: Submitted application application_16
45154347193_0002
22/02/17 20:41:45 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1645154347193_0002/
22/02/17 20:41:45 INFO mapreduce.Job: Running job: job_1645154347193_0002
22/02/17 20:41:53 INFO mapreduce.Job: Job job_1645154347193_0002 running in uber
 mode : false
22/02/17 20:41:53 INFO mapreduce.Job:  map 0% reduce 0%
22/02/17 20:42:01 INFO mapreduce.Job:  map 100% reduce 0%
22/02/17 20:42:09 INFO mapreduce.Job:  map 100% reduce 100%
22/02/17 20:42:09 INFO mapreduce.Job: Job job_1645154347193_0002 completed successfully
22/02/17 20:42:09 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=1421
                FILE: Number of bytes written=223417
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=973
                HDFS: Number of bytes written=947
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=5978
                Total time spent by all reduces in occupied slots (ms)=4482
                Total time spent by all map tasks (ms)=5978
                Total time spent by all reduce tasks (ms)=4482
                Total vcore-seconds taken by all map tasks=5978
                Total vcore-seconds taken by all reduce tasks=4482
                Total megabyte-seconds taken by all map tasks=6121472
                Total megabyte-seconds taken by all reduce tasks=4589568
        Map-Reduce Framework
                Map input records=7
                Map output records=154
```

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

```
        Map-Reduce Framework
                Map input records=7
                Map output records=154
                Map output bytes=1480
                Map output materialized bytes=1421
                Input split bytes=109
                Combine input records=154
                Combine output records=117
                Reduce input groups=117
                Reduce shuffle bytes=1421
                Reduce input records=117
                Reduce output records=117
                Spilled Records=234
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=132
                CPU time spent (ms)=1040
                Physical memory (bytes) snapshot=329506816
                Virtual memory (bytes) snapshot=3007062016
                Total committed heap usage (bytes)=226365440
        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=864
        File Output Format Counters
                Bytes Written=947
```

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

```
fantastic       1
food,   1
for     1
found   1
god     1
gold    1
gold.   3
golden  1
good    1
granted 1
groaned,        1
had     1
hand.   1
he      5
her     1
him     1
him,    1
his     4
hungry. 1
in      1
into    1
is      1
it      1
it,     1
it.     1
item    1
kinds   1
king    1
lead    1
named   1
newly-earned    1
no      1
not     1
of      3
once    1
picked  1
piece   1
pleaded 1
powers, 1
prevent 1
pure    1
she,    1
so,     1
soon,   1
started 1
starve! 1
such    1
that    2
the     1
then    1
things, 1
```

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

```
                Bytes Written=947
[cloudera@quickstart ~]$ hdfs dfs -cat /XYZ/part-r-00000
And     1
As      1
But     1
Despite 1
Dionysus,       1
Dionysus’       1
Excited 1
For     1
Greed   1
Hungry, 1
It      1
Midas   7
Midas’  1
Perhaps 1
Satyr.  1
Seeing  1
There   1
a       6
about   1
after   1
all     1
all!”   1
always  1
an      1
and     2
arms    1
around  1
asked   1
became  1
beloved 1
bestowed.       1
blessing,”      1
by      1
comfort 1
couldn’t        1
cried.  1
daughter        1
deed    1
did     1
dismay, 1
downfall.       1
each    1
eat     1
efforts 1
excellent       1
fantastic       1
food,   1
for     1
```

Name: Pooja Pathak
Roll No:14
MSc Part -I Sem - 2

```
such    1
that    2
the     1
then    1
things, 1
this    2
threw   1
to      6
too,    1
touch   1
touched 1
touching        1
turn    1
turned  2
turning 1
up      1
was     5
whatever        1
who     1
will    1
wine.   1
wish    2
wish,   2
would   1
“I’ll   1
“The    1
[cloudera@quickstart ~]$
```