

ICE-2

Krishna Reddy Chaduvu

UNT id:11551974

1.By following the step-by-step instructions given,implemented and successfully executed the wordcount program and explained in detail below.

Step:1

Follwed by the instructions given,created a new file called wordp.txt using cat command and given the input data.

File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ cat > wordp.txt
```

The preacher was in fact a thief who had stolen the original manuscript of Hamlet from an exhibit on the Riviera, where he also acquired his tan. The preacher disguised the manuscript as a Bible and had Longworth authenticate it for the Duke, whom the preacher hoped would buy it. Longworth, however, in desperate need of money, killed the preacher with Hamlet's sword and stole the manuscript. Longworth, who does not smoke, planted the German-made cigarette near the victim's body to throw suspicion from himself; but in the process, he accidentally dropped his packet of aspirin.

KILLER: EARL LONGWORTH

WEAPON: SWORD

MOTIVE: MANUSCRIPT

Sir Reginald Cosgrove, in dire need of money, concocted a plot to collect money from his large insurance policy on Silver Patch. Sir Reginald painted over the silver patch on the mane of his prize horse and painted a silver patch on the mane of Night Dancer, another of his horses. He then switched stalls and poisoned Night Dancer, who now looked like Silver Patch. Realizing that he would have to take Oscar Switt into his confidence, Sir Reginald arranged for his trainer to meet him at the stables. When Switt refused to go along with Sir Reginald's plan to kill Night Dancer, Sir Reginald became enraged and killed the trainer, smashing him over the head with an ale bottle, then stabbing him repeatedly with the broken bottle.

KILLER: SIR REGINALD COSGROVE

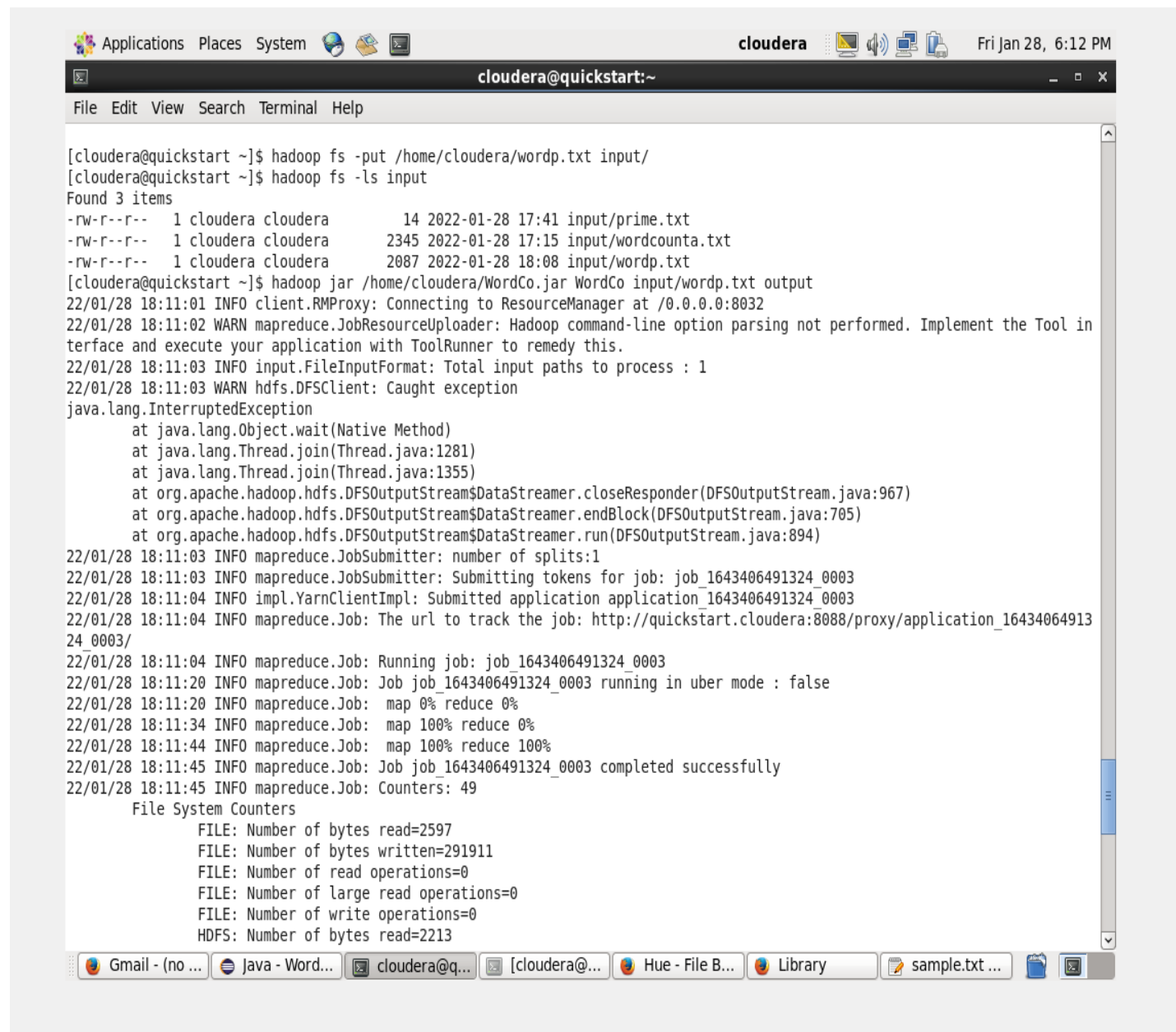
WEAPON: BROKEN BOTTLE

MOTIVE: INSURANCE

Two years after being dishonorably discharged from the British military, Donald Hobson changed his name to Alfred Cooke and tried to start a new life. Neither his eventual employers at Scotland Yard nor his wife and children knew the truth of his past. Manfred Maloney stumbled onto Cooke and attempted to blackmail his old army buddy. To protect his secret, Cooke

Step:2

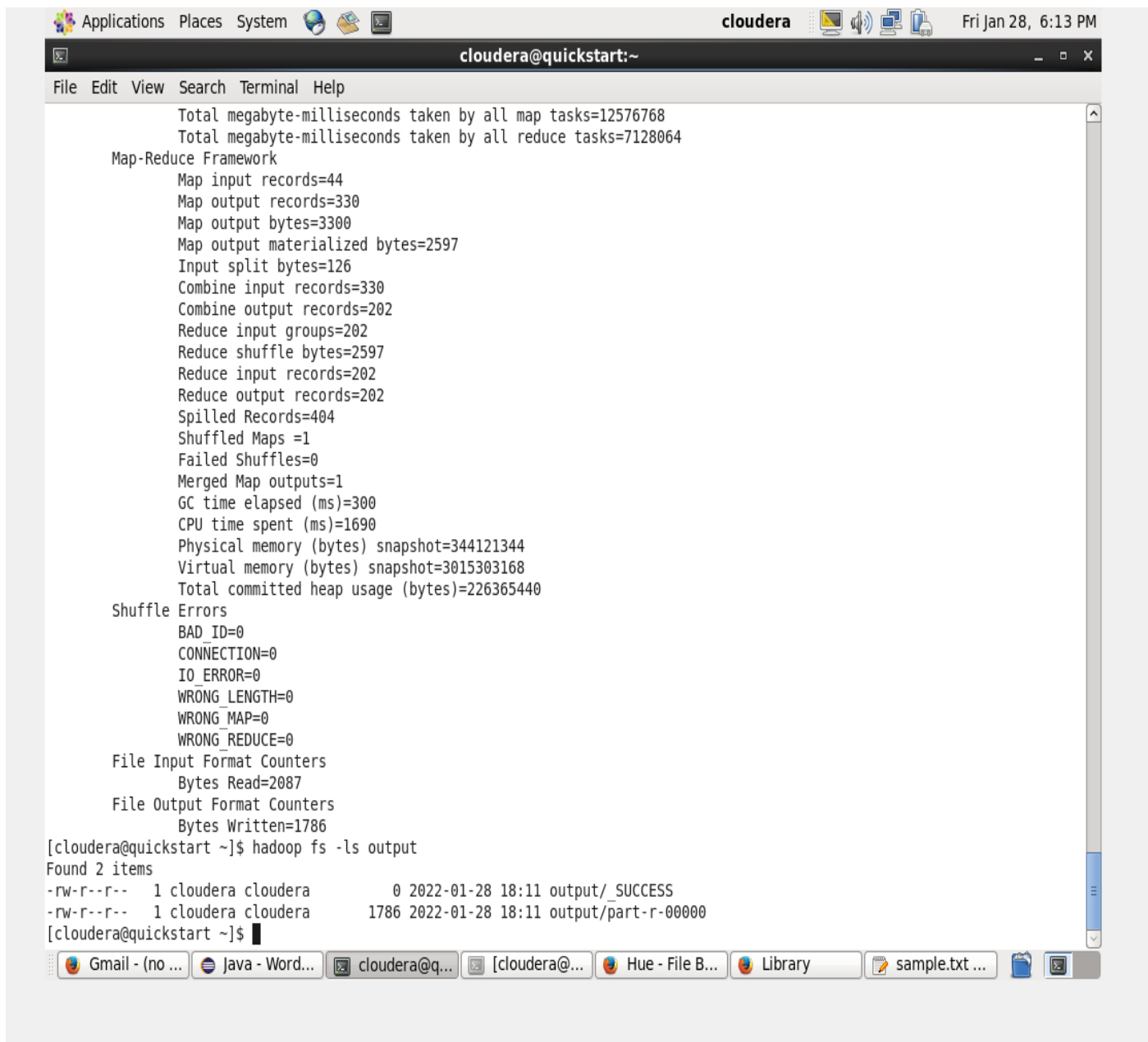
Loading the wordp.txt file into hadoop hdfs and implementing the mapreducing alogorithm.



```
Applications Places System cloudera Fri Jan 28, 6:12 PM
cloudera@quickstart:~
File Edit View Search Terminal Help

[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/wordp.txt input/
[cloudera@quickstart ~]$ hadoop fs -ls input
Found 3 items
-rw-r--r-- 1 cloudera cloudera      14 2022-01-28 17:41 input/prime.txt
-rw-r--r-- 1 cloudera cloudera    2345 2022-01-28 17:15 input/wordcounta.txt
-rw-r--r-- 1 cloudera cloudera    2087 2022-01-28 18:08 input/wordp.txt
[cloudera@quickstart ~]$ hadoop jar /home/cloudera/WordCo.jar WordCo input/wordp.txt output
22/01/28 18:11:01 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/01/28 18:11:02 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in
terface and execute your application with ToolRunner to remedy this.
22/01/28 18:11:03 INFO input.FileInputFormat: Total input paths to process : 1
22/01/28 18:11:03 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/01/28 18:11:03 INFO mapreduce.JobSubmitter: number of splits:1
22/01/28 18:11:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1643406491324_0003
22/01/28 18:11:04 INFO impl.YarnClientImpl: Submitted application application_1643406491324_0003
22/01/28 18:11:04 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_16434064913
24_0003/
22/01/28 18:11:04 INFO mapreduce.Job: Running job: job_1643406491324_0003
22/01/28 18:11:20 INFO mapreduce.Job: Job job_1643406491324_0003 running in uber mode : false
22/01/28 18:11:20 INFO mapreduce.Job:  map 0% reduce 0%
22/01/28 18:11:34 INFO mapreduce.Job:  map 100% reduce 0%
22/01/28 18:11:44 INFO mapreduce.Job:  map 100% reduce 100%
22/01/28 18:11:45 INFO mapreduce.Job: Job job_1643406491324_0003 completed successfully
22/01/28 18:11:45 INFO mapreduce.Job: Counters: 49
    File System Counters
      FILE: Number of bytes read=2597
      FILE: Number of bytes written=291911
      FILE: Number of read operations=0
      FILE: Number of large read operations=0
      FILE: Number of write operations=0
      HDFS: Number of bytes read=2213
```

Map Reducing is done and the given command is success and two items are found



The screenshot shows a terminal window titled "cloudera@quickstart:~" with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal output displays the following statistics:

```
Total megabyte-milliseconds taken by all map tasks=12576768
Total megabyte-milliseconds taken by all reduce tasks=7128064
Map-Reduce Framework
  Map input records=44
  Map output records=330
  Map output bytes=3300
  Map output materialized bytes=2597
  Input split bytes=126
  Combine input records=330
  Combine output records=202
  Reduce input groups=202
  Reduce shuffle bytes=2597
  Reduce input records=202
  Reduce output records=202
  Spilled Records=404
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=300
  CPU time spent (ms)=1690
  Physical memory (bytes) snapshot=344121344
  Virtual memory (bytes) snapshot=3015303168
  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=2087
File Output Format Counters
  Bytes Written=1786
```

Below the statistics, the command `[cloudera@quickstart ~]$ hadoop fs -ls output` is executed, resulting in the following output:

```
Found 2 items
-rw-r--r--  1 cloudera cloudera      0 2022-01-28 18:11 output/ SUCCESS
-rw-r--r--  1 cloudera cloudera    1786 2022-01-28 18:11 output/part-r-00000
```

The terminal prompt is `[cloudera@quickstart ~]$`. The window's taskbar at the bottom shows several open applications: Gmail - (no ...), Java - Word..., cloudera@q..., [cloudera@..., Hue - File B..., Library, and sample.txt

Visualization of the items found.

The screenshot shows the Hue File Browser interface in a Mozilla Firefox browser window. The browser's address bar displays the URL: `quickstart.cloudera:8888/hue/filebrowser/view%3D/user/cloudera/ou`. The Hue interface includes a top navigation bar with links to Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. Below this is a search bar and a 'Query' button. The main content area is titled 'File Browser' and shows a directory listing for the path `Home / user / cloudera / output`. The listing includes a table with columns for Name, Size, User, Group, Permissions, and Date. The table contains four entries: a directory named `.`, a directory named `_SUCCESS`, a file named `part-r-00000`, and a file named `part-r-00000`. The bottom of the interface shows a taskbar with several open applications: Gmail, Java - Word..., cloudera@q..., [cloudera@...], Hue - File B..., Library, and sample.txt ...

Applications Places System cloudera Fri Jan 28, 6:14 PM

Hue - File Browser - Mozilla Firefox

Hue - File Browser x +

quickstart.cloudera:8888/hue/filebrowser/view%3D/user/cloudera/ou Search

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Hue Query Search data and saved documents... Jobs cloudera

File Browser

Search for file name Actions Move to trash Upload New

Home / user / cloudera / output

	Name	Size	User	Group	Permissions	Date
	.		cloudera	cloudera	drwxr-xr-x	January 28, 2022 06:11 PM
	._SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	January 28, 2022 06:11 PM
	part-r-00000	1.7 KB	cloudera	cloudera	-rw-r--r--	January 28, 2022 06:11 PM

Show 45 of 2 items Page 1 of 1

Gmail - (no ... Java - Word... cloudera@q... [cloudera@... Hue - File B... Library sample.txt ...

Step:3

Frequency count of words, of the given input is done and the output is below.

The screenshot shows the Hue File Browser interface in a Mozilla Firefox browser window. The address bar shows the URL: `quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera/output`. The interface includes a top navigation bar with links to Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. The main content area displays the file `part-r-00000` with a word frequency count.

File details:

- View as binary
- Edit file
- Download
- View file location
- Refresh
- Last modified: 01/29/2022 2:11 AM
- User: cloudera
- Group: cloudera
- Size: 1.74 KB
- Mode: 100644

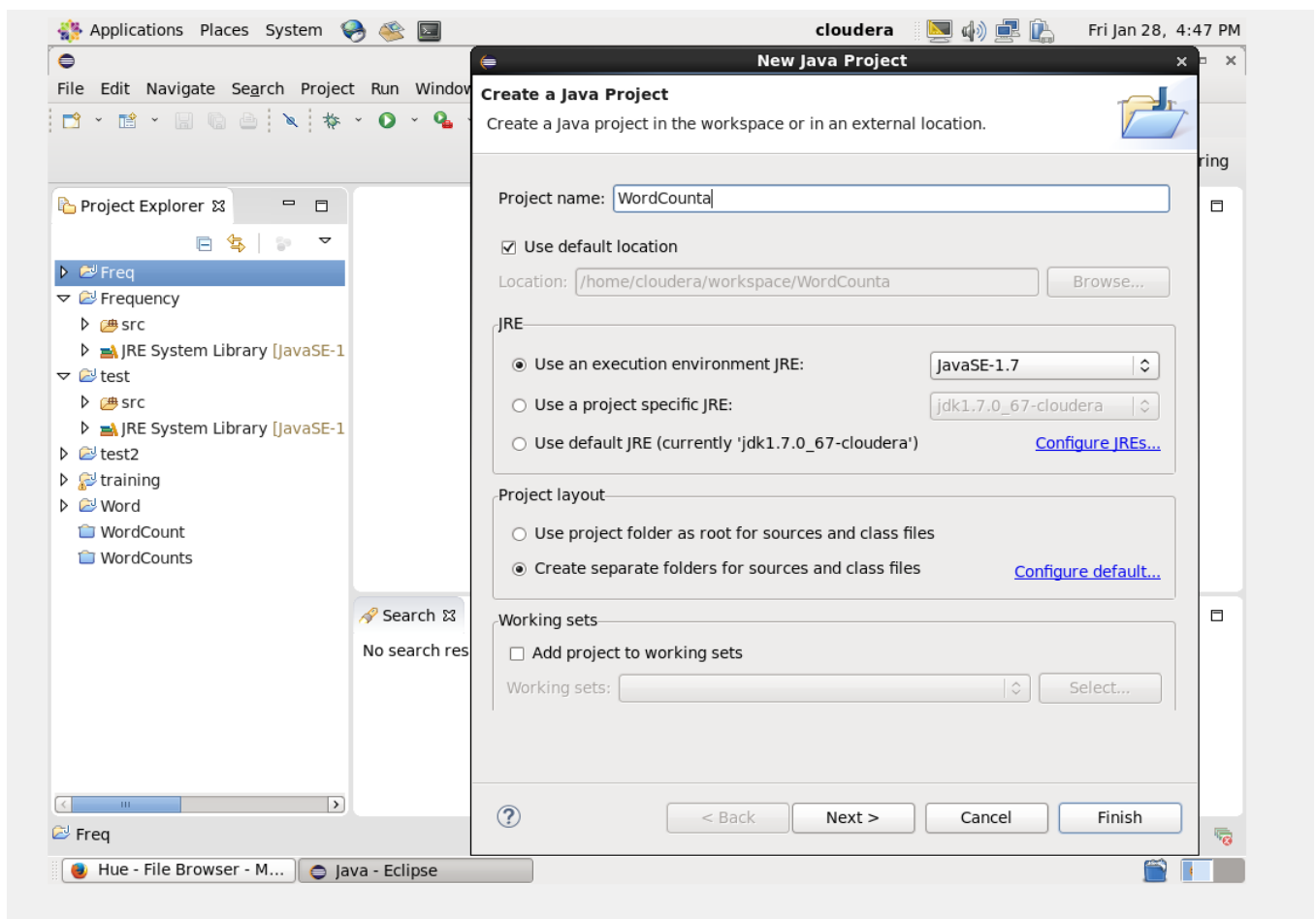
Word frequency count:

(DONALD	1
ALFRED	1
Alfred	1
BLACKMAIL	1
BOTTLE	1
BROKEN	1
Bible	1
British	1
COOKE	1
COSGROVE	1
Cooke	3
Cosgrove,	1
Dancer,	3
Donald	1
Duke,	1
EARL	1
German-made	1
HOBSON)	1
Hamlet	1
Hamlet's	1
He	1
Hobson	1

2.Counting the frequency of words in given text file that starts with 'a'.

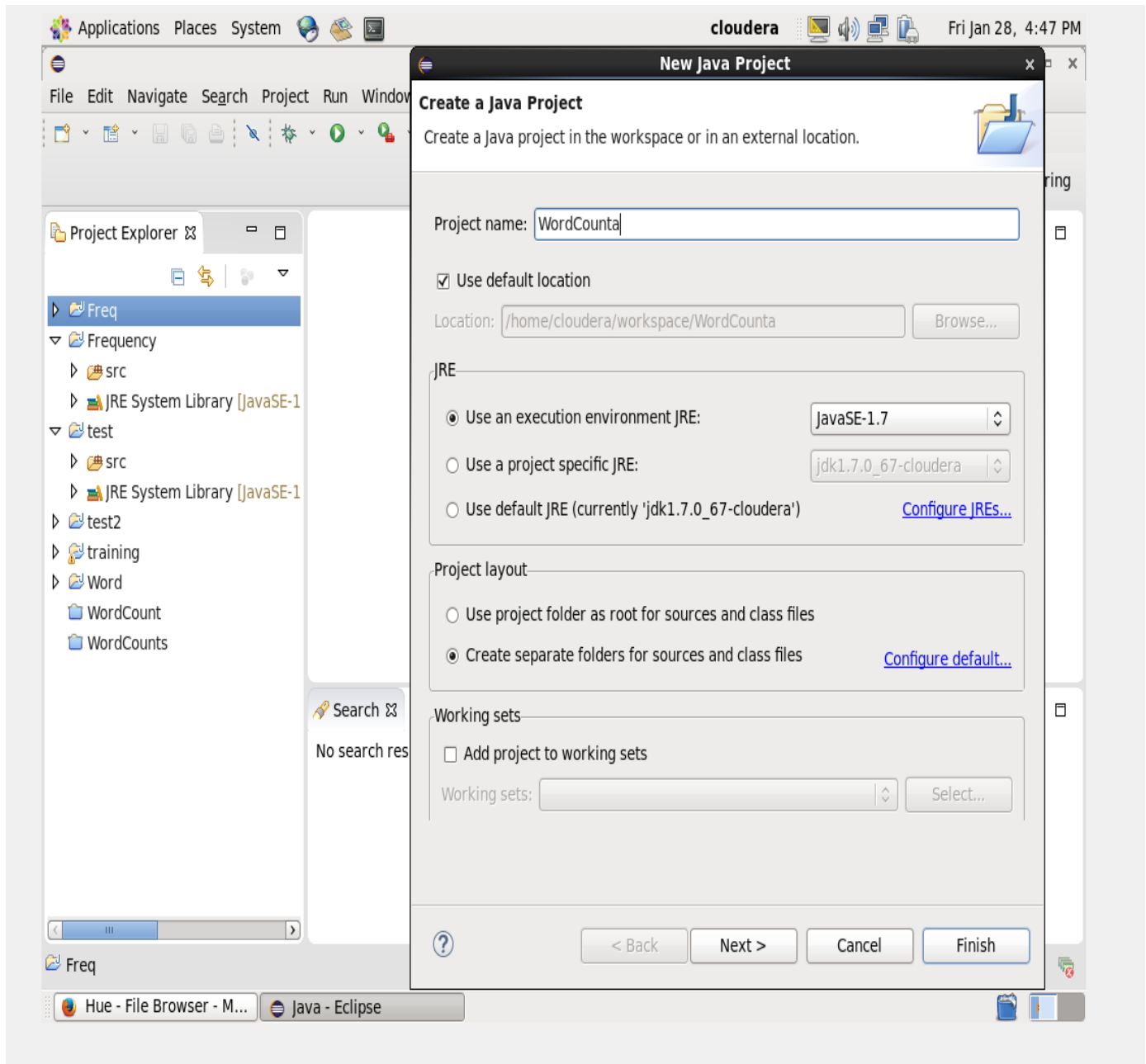
Step:1

Creating a new java project and naming it as WordCounta

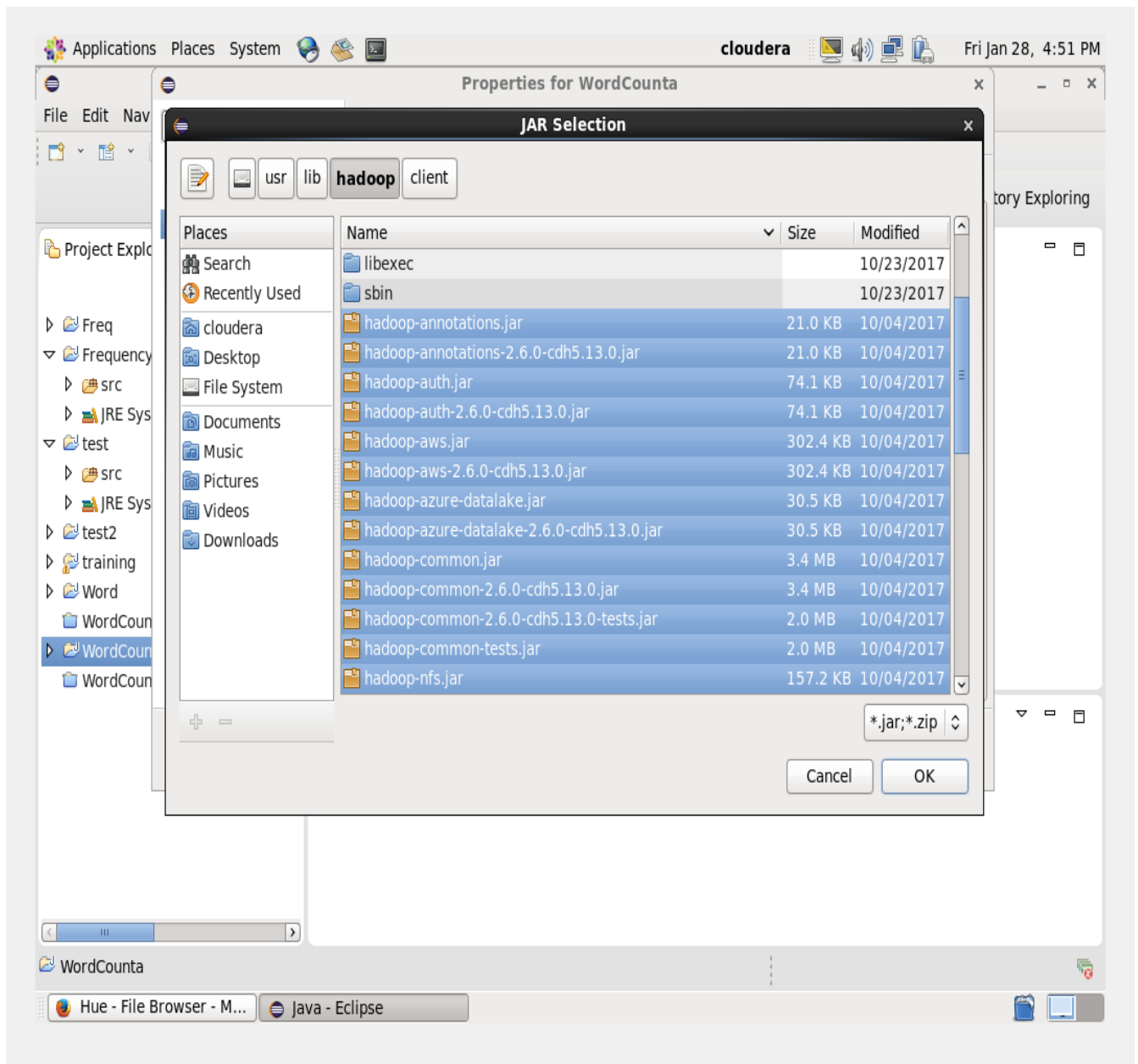


Step:2

Adding the external JARs files

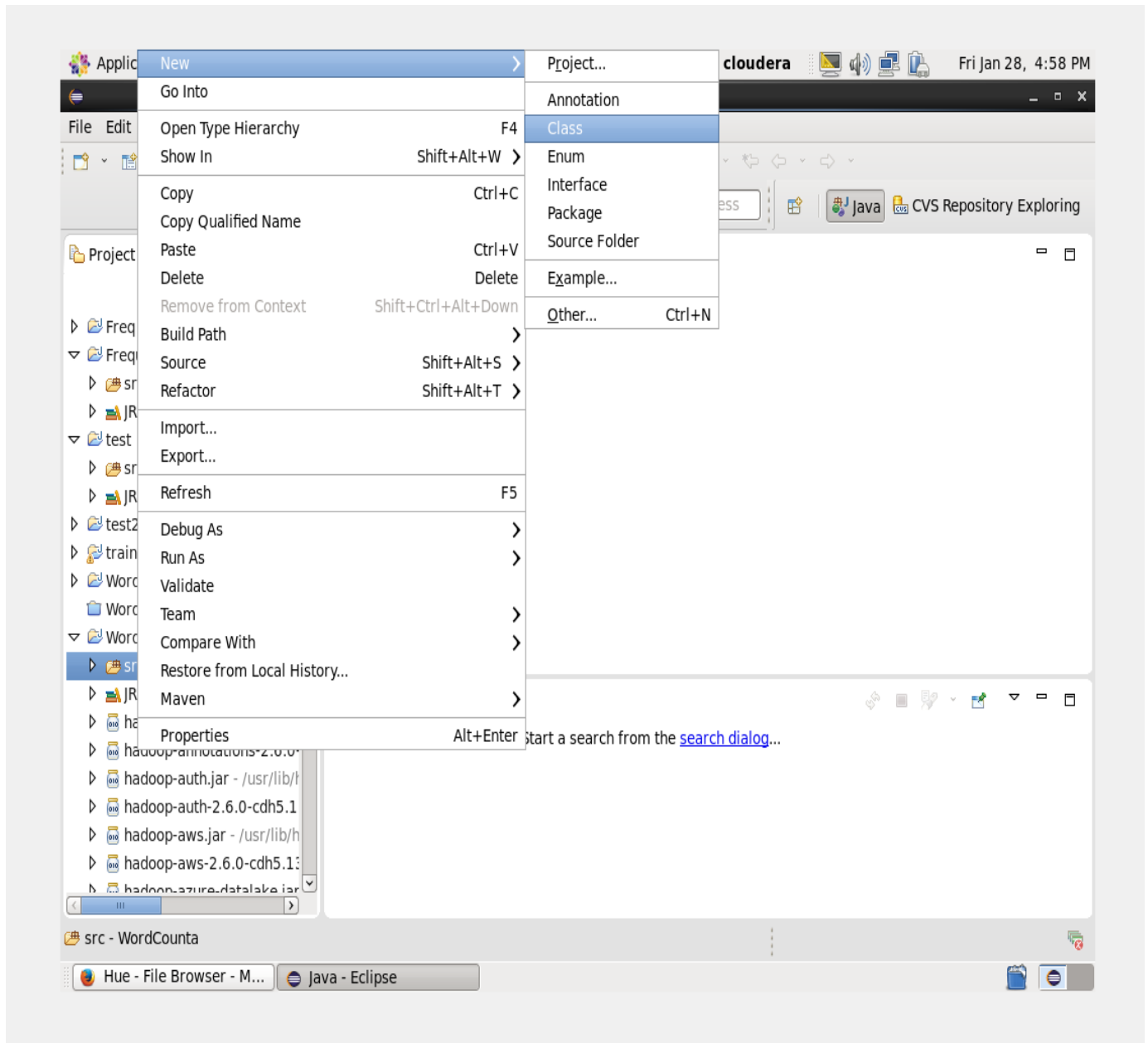


Adding the JARs files of hadoop and client

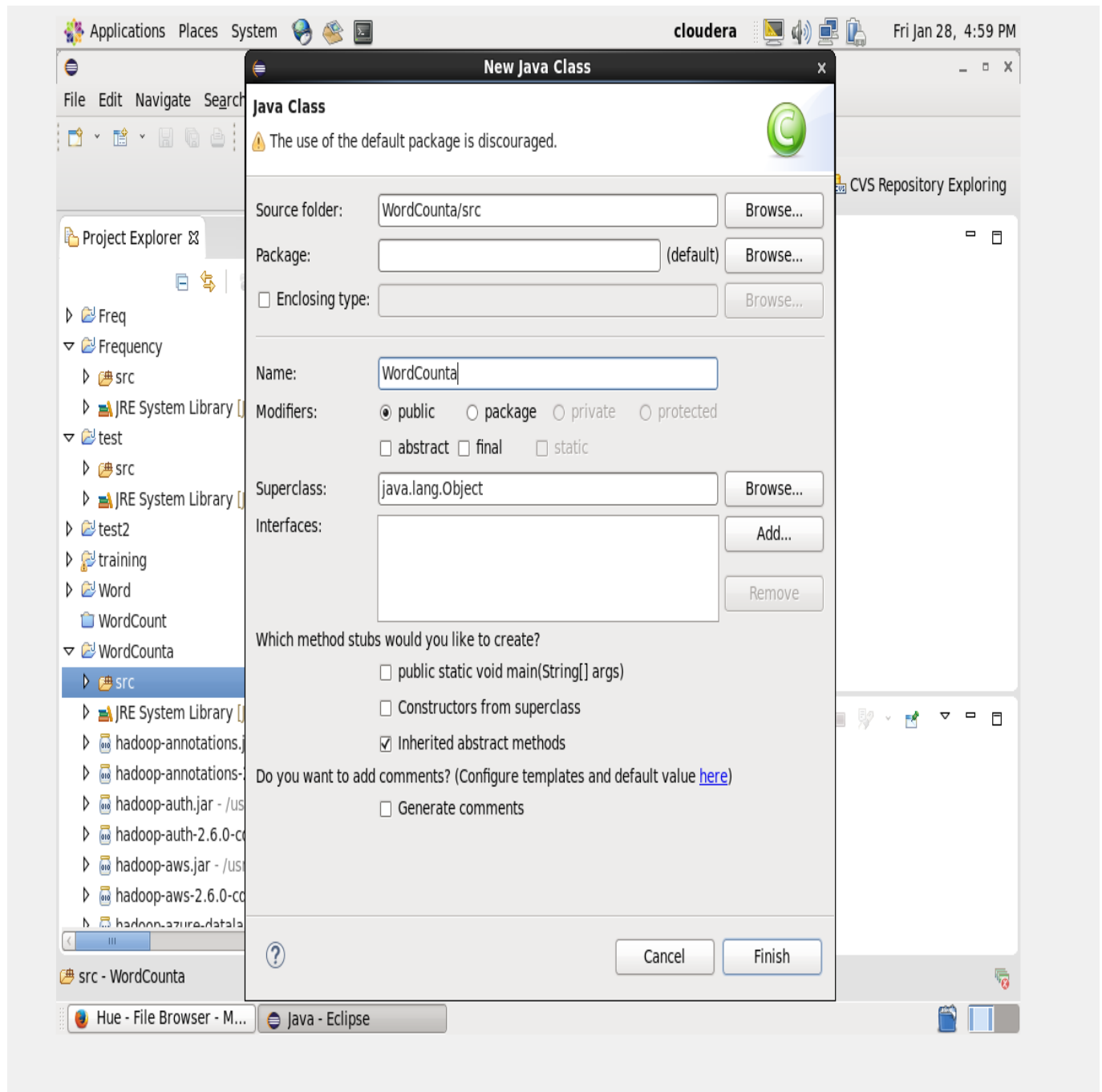


Step:3

Creating the class files. Right click on source, New>Class



Naming the file as WordCounta

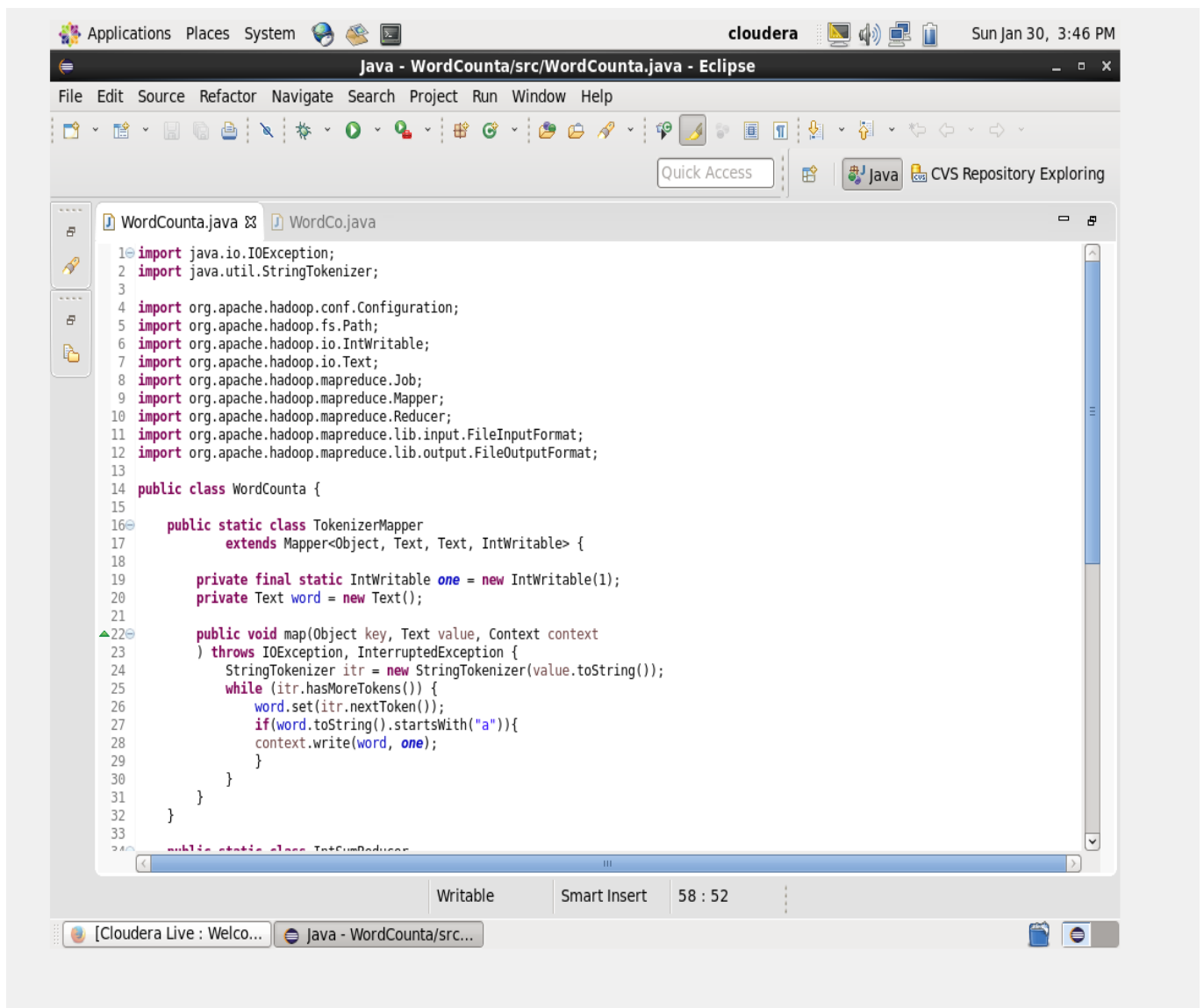


Step:4

Logic: `if(word.toString().startsWith("a"))`

Logic explanation: The code is almost same as the wordcount program, but here only the frequency of words that starts with "a" are counted.

Program code:



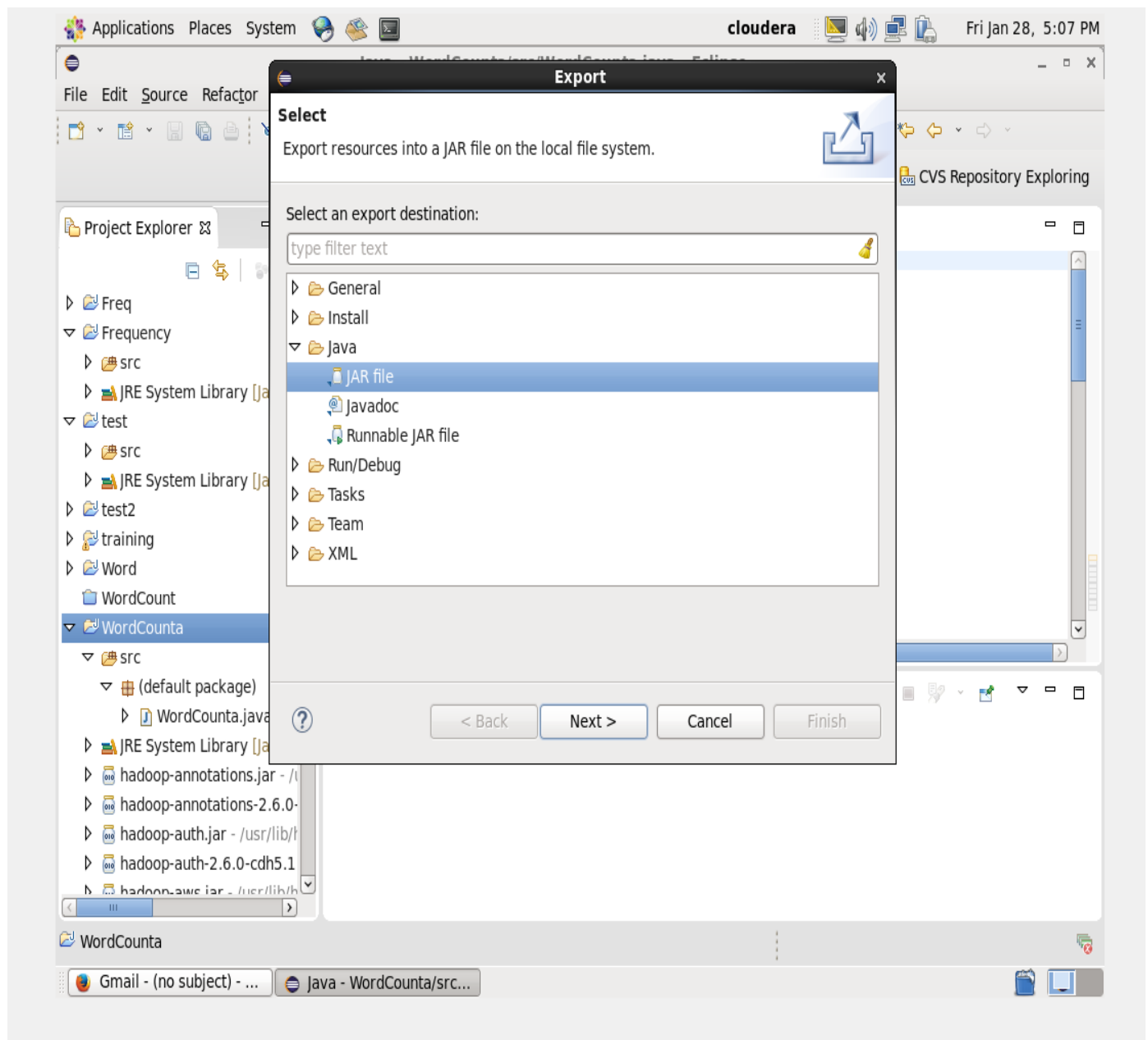
```
1 import java.io.IOException;
2 import java.util.StringTokenizer;
3
4 import org.apache.hadoop.conf.Configuration;
5 import org.apache.hadoop.fs.Path;
6 import org.apache.hadoop.io.IntWritable;
7 import org.apache.hadoop.io.Text;
8 import org.apache.hadoop.mapreduce.Job;
9 import org.apache.hadoop.mapreduce.Mapper;
10 import org.apache.hadoop.mapreduce.Reducer;
11 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
12 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
13
14 public class WordCounta {
15
16     public static class TokenizerMapper
17         extends Mapper<Object, Text, Text, IntWritable> {
18
19         private final static IntWritable one = new IntWritable(1);
20         private Text word = new Text();
21
22     public void map(Object key, Text value, Context context
23         ) throws IOException, InterruptedException {
24         StringTokenizer itr = new StringTokenizer(value.toString());
25         while (itr.hasMoreTokens()) {
26             word.set(itr.nextToken());
27             if(word.toString().startsWith("a")){
28                 context.write(word, one);
29             }
30         }
31     }
32 }
33
34 public static class TestSumReducer
```

WordCounta.java WordCo.java

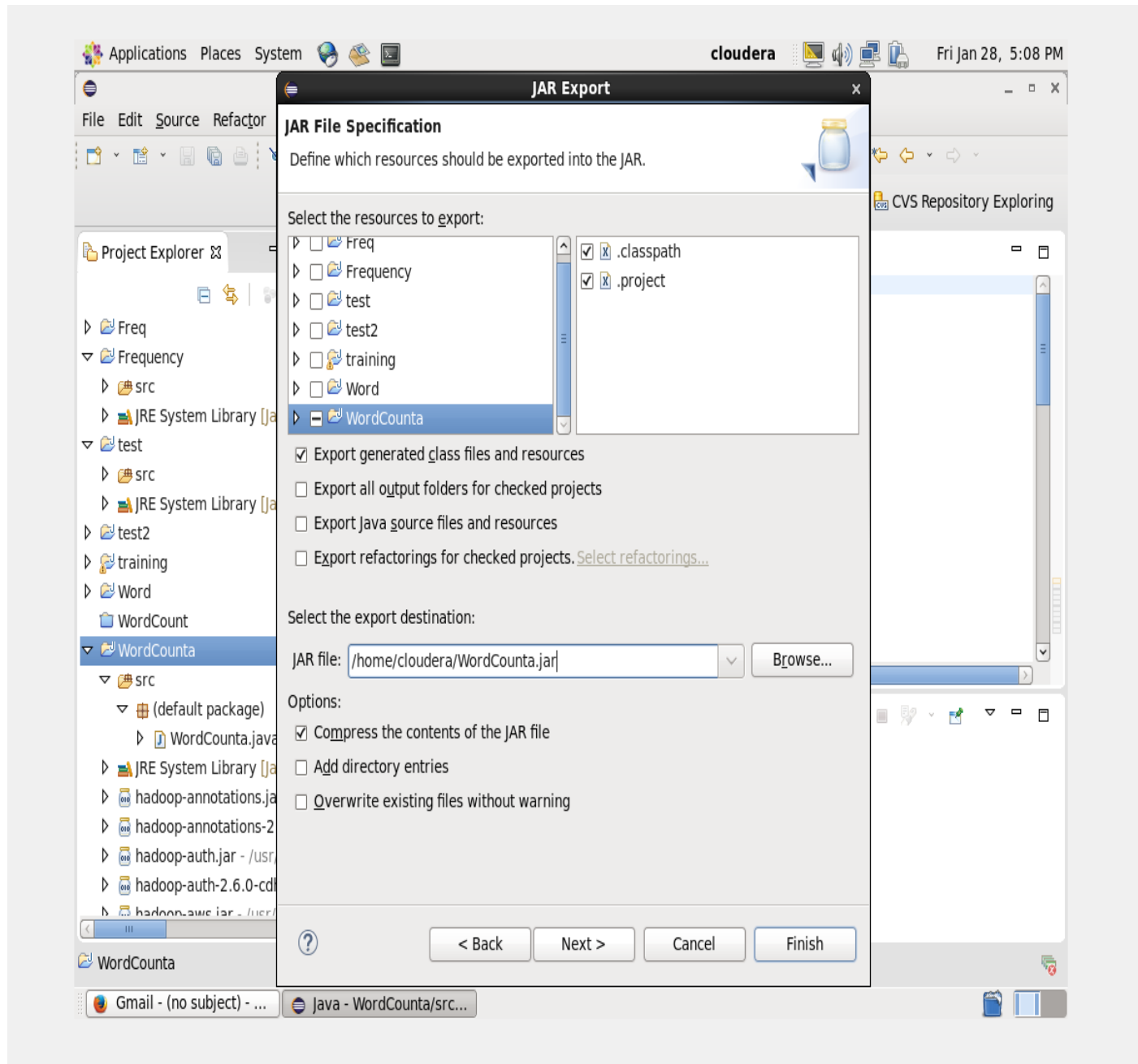
```
30    }
31  }
32 }
33
34 public static class IntSumReducer
35     extends Reducer<Text, IntWritable, Text, IntWritable> {
36     private IntWritable result = new IntWritable();
37
38     public void reduce(Text key, Iterable<IntWritable> values,
39         Context context
40     ) throws IOException, 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 }
49
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(WordCounta.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 }
```

Step5:

Exporting the JAR, right click on WordCounta and select export

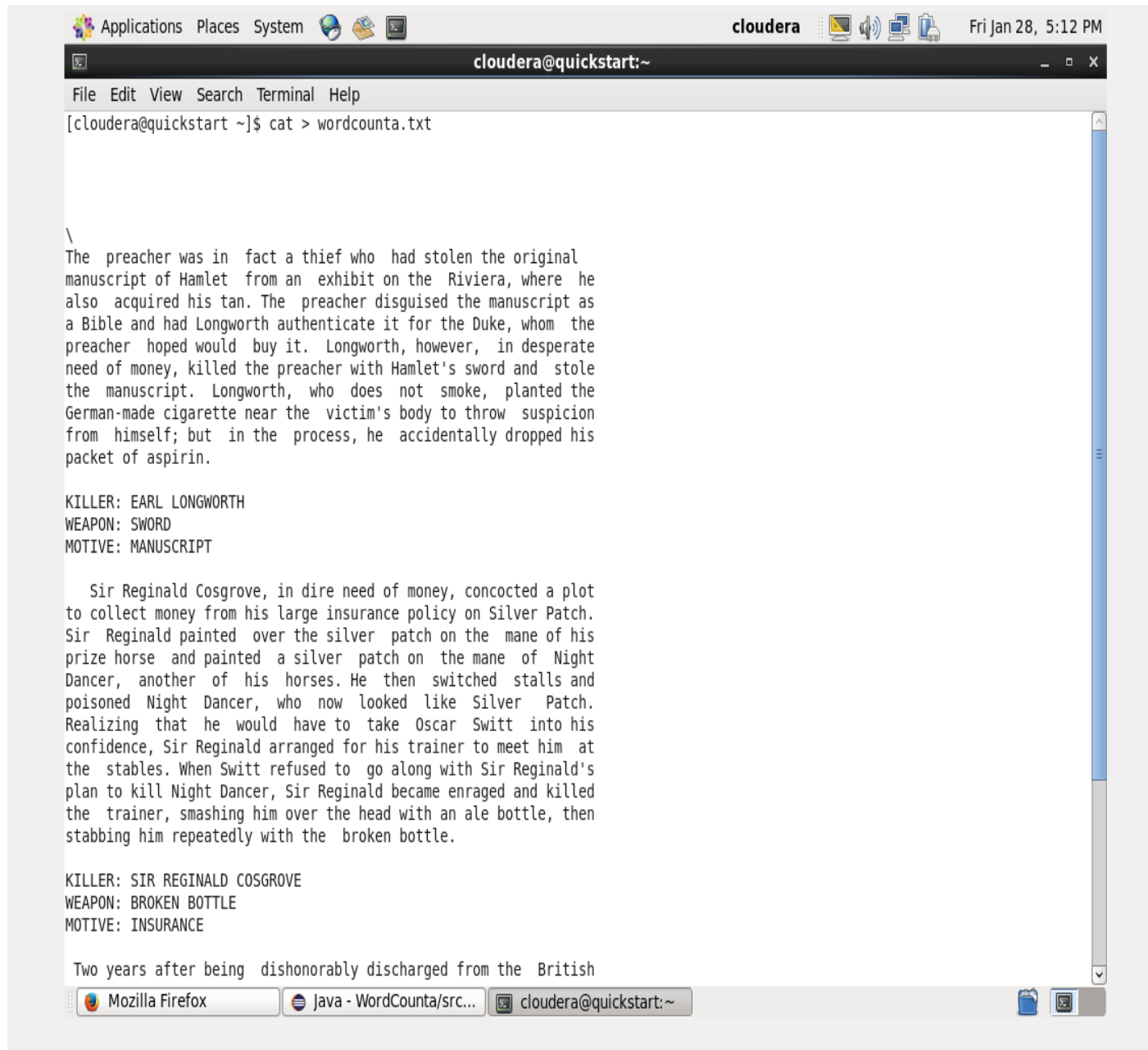


Select the export destination



Step:6

Creating a new file called wordcounta.txt and giving the input



The screenshot shows a Linux desktop environment with a terminal window titled "cloudera@quickstart:~". The terminal displays the command `cat > wordcounta.txt` and the following text content:

```
\
The preacher was in fact a thief who had stolen the original
manuscript of Hamlet from an exhibit on the Riviera, where he
also acquired his tan. The preacher disguised the manuscript as
a Bible and had Longworth authenticate it for the Duke, whom the
preacher hoped would buy it. Longworth, however, in desperate
need of money, killed the preacher with Hamlet's sword and stole
the manuscript. Longworth, who does not smoke, planted the
German-made cigarette near the victim's body to throw suspicion
from himself; but in the process, he accidentally dropped his
packet of aspirin.

KILLER: EARL LONGWORTH
WEAPON: SWORD
MOTIVE: MANUSCRIPT

Sir Reginald Cosgrove, in dire need of money, concocted a plot
to collect money from his large insurance policy on Silver Patch.
Sir Reginald painted over the silver patch on the mane of his
prize horse and painted a silver patch on the mane of Night
Dancer, another of his horses. He then switched stalls and
poisoned Night Dancer, who now looked like Silver Patch.
Realizing that he would have to take Oscar Switt into his
confidence, Sir Reginald arranged for his trainer to meet him at
the stables. When Switt refused to go along with Sir Reginald's
plan to kill Night Dancer, Sir Reginald became enraged and killed
the trainer, smashing him over the head with an ale bottle, then
stabbing him repeatedly with the broken bottle.

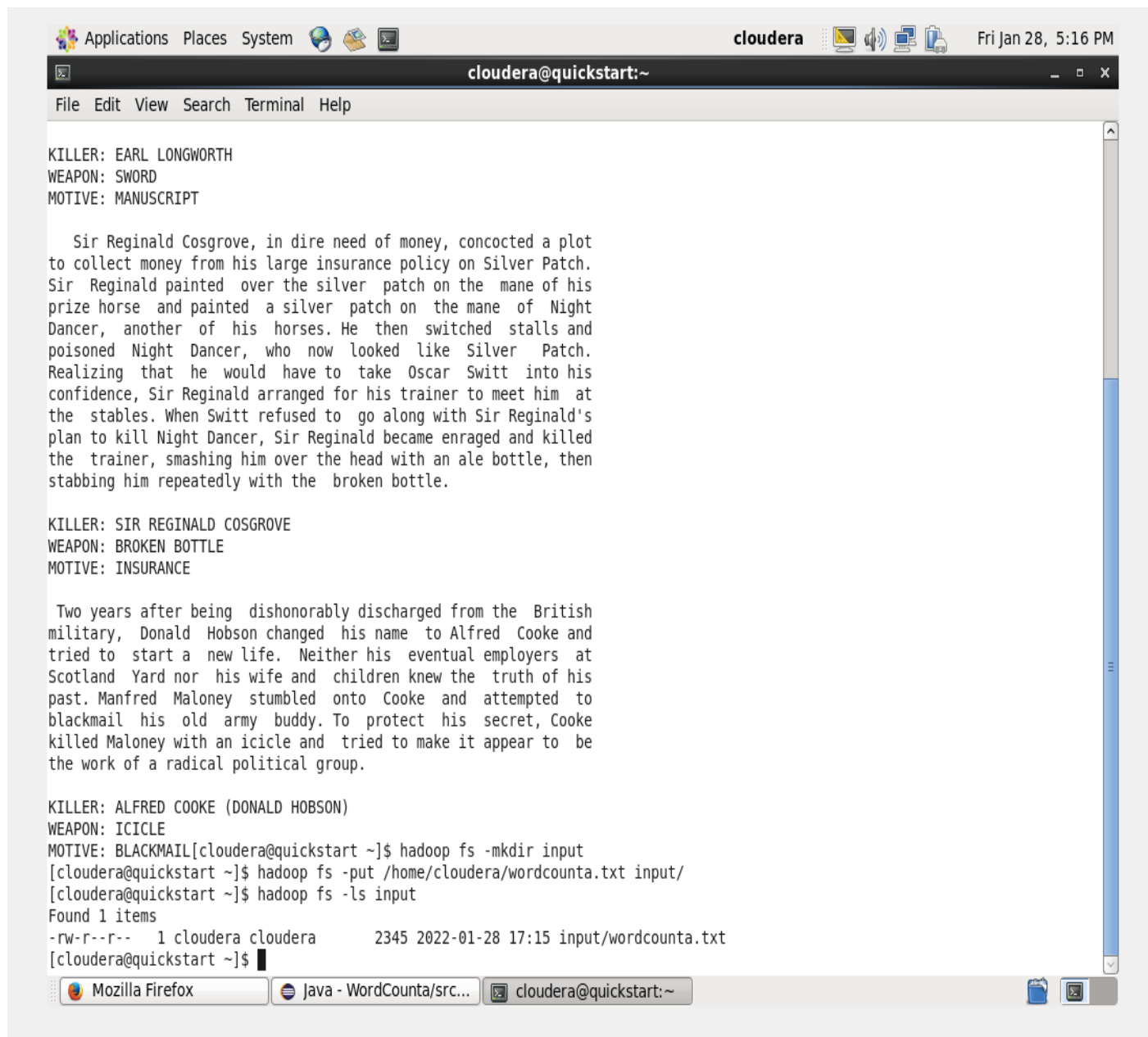
KILLER: SIR REGINALD COSGROVE
WEAPON: BROKEN BOTTLE
MOTIVE: INSURANCE

Two years after being dishonorably discharged from the British
```

The terminal window has a menu bar with "File", "Edit", "View", "Search", "Terminal", and "Help". The desktop background is light gray, and the taskbar at the bottom shows icons for "Mozilla Firefox", "Java - WordCounta/src...", and "cloudera@quickstart:~". The system clock in the top right corner indicates "Fri Jan 28, 5:12 PM".

Step:7

Creating a directory called input and loading the wordcounta.txt file into hadoop hdfs



```
Applications Places System cloudera Fri Jan 28, 5:16 PM
cloudera@quickstart:~
File Edit View Search Terminal Help

KILLER: EARL LONGWORTH
WEAPON: SWORD
MOTIVE: MANUSCRIPT

Sir Reginald Cosgrove, in dire need of money, concocted a plot
to collect money from his large insurance policy on Silver Patch.
Sir Reginald painted over the silver patch on the mane of his
prize horse and painted a silver patch on the mane of Night
Dancer, another of his horses. He then switched stalls and
poisoned Night Dancer, who now looked like Silver Patch.
Realizing that he would have to take Oscar Switt into his
confidence, Sir Reginald arranged for his trainer to meet him at
the stables. When Switt refused to go along with Sir Reginald's
plan to kill Night Dancer, Sir Reginald became enraged and killed
the trainer, smashing him over the head with an ale bottle, then
stabbing him repeatedly with the broken bottle.

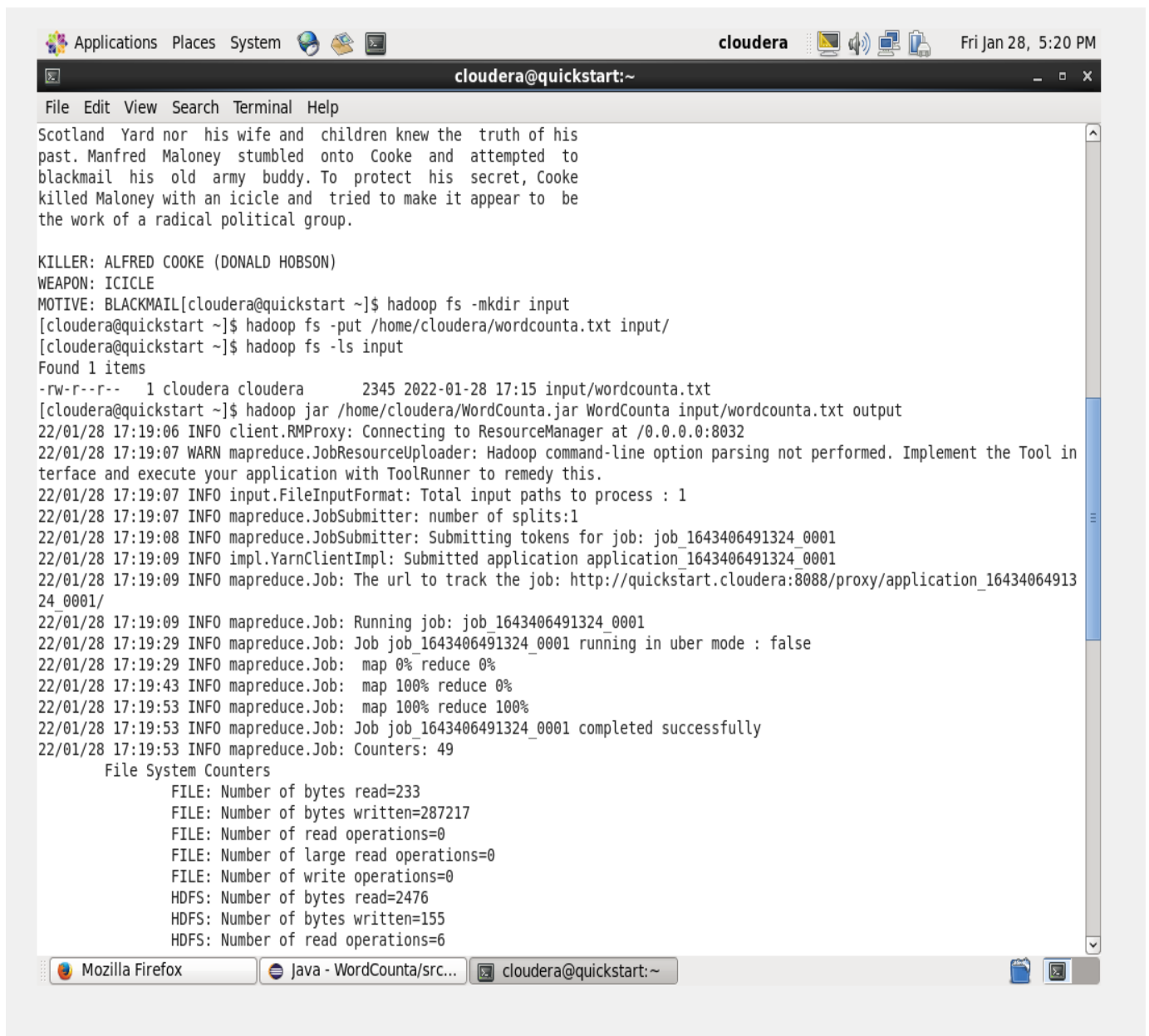
KILLER: SIR REGINALD COSGROVE
WEAPON: BROKEN BOTTLE
MOTIVE: INSURANCE

Two years after being dishonorably discharged from the British
military, Donald Hobson changed his name to Alfred Cooke and
tried to start a new life. Neither his eventual employers at
Scotland Yard nor his wife and children knew the truth of his
past. Manfred Maloney stumbled onto Cooke and attempted to
blackmail his old army buddy. To protect his secret, Cooke
killed Maloney with an icicle and tried to make it appear to be
the work of a radical political group.

KILLER: ALFRED COOKE (DONALD HOBSON)
WEAPON: ICICLE
MOTIVE: BLACKMAIL
[cloudera@quickstart ~]$ hadoop fs -mkdir input
[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/wordcounta.txt input/
[cloudera@quickstart ~]$ hadoop fs -ls input
Found 1 items
-rw-r--r-- 1 cloudera cloudera 2345 2022-01-28 17:15 input/wordcounta.txt
[cloudera@quickstart ~]$
```

Step:8

Performing the mapreducing

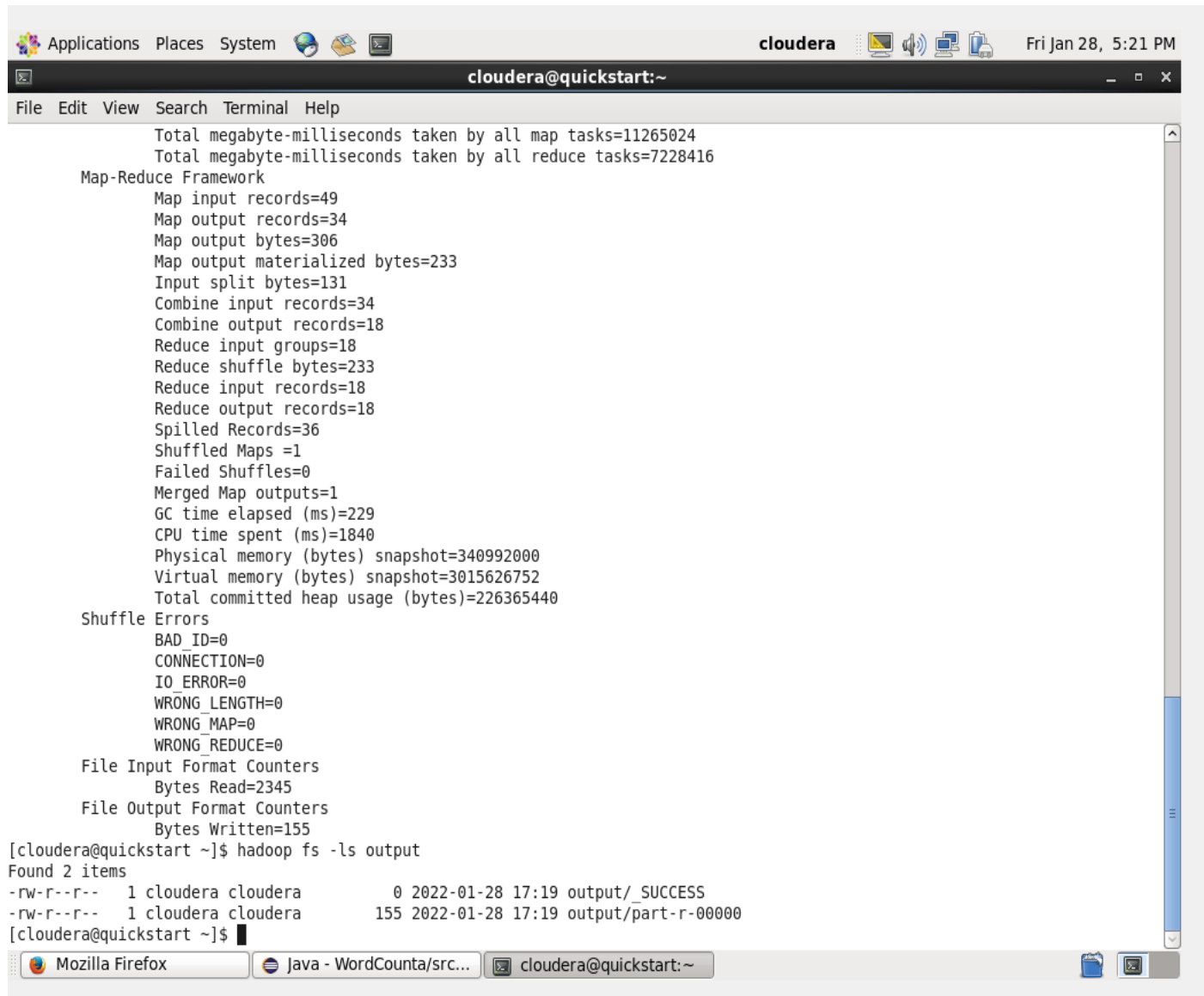


```
Applications Places System cloudera Fri Jan 28, 5:20 PM
cloudera@quickstart:~
File Edit View Search Terminal Help
Scotland Yard nor his wife and children knew the truth of his
past. Manfred Maloney stumbled onto Cooke and attempted to
blackmail his old army buddy. To protect his secret, Cooke
killed Maloney with an icicle and tried to make it appear to be
the work of a radical political group.

KILLER: ALFRED COOKE (DONALD HOBSON)
WEAPON: ICICLE
MOTIVE: BLACKMAIL[cloudera@quickstart ~]$ hadoop fs -mkdir input
[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/wordcounta.txt input/
[cloudera@quickstart ~]$ hadoop fs -ls input
Found 1 items
-rw-r--r-- 1 cloudera cloudera      2345 2022-01-28 17:15 input/wordcounta.txt
[cloudera@quickstart ~]$ hadoop jar /home/cloudera/WordCounta.jar WordCounta input/wordcounta.txt output
22/01/28 17:19:06 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/01/28 17:19:07 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in
terface and execute your application with ToolRunner to remedy this.
22/01/28 17:19:07 INFO input.FileInputFormat: Total input paths to process : 1
22/01/28 17:19:07 INFO mapreduce.JobSubmitter: number of splits:1
22/01/28 17:19:08 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1643406491324_0001
22/01/28 17:19:09 INFO impl.YarnClientImpl: Submitted application application_1643406491324_0001
22/01/28 17:19:09 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_16434064913
24_0001/
22/01/28 17:19:09 INFO mapreduce.Job: Running job: job_1643406491324_0001
22/01/28 17:19:29 INFO mapreduce.Job: Job job_1643406491324_0001 running in uber mode : false
22/01/28 17:19:29 INFO mapreduce.Job: map 0% reduce 0%
22/01/28 17:19:43 INFO mapreduce.Job: map 100% reduce 0%
22/01/28 17:19:53 INFO mapreduce.Job: map 100% reduce 100%
22/01/28 17:19:53 INFO mapreduce.Job: Job job_1643406491324_0001 completed successfully
22/01/28 17:19:53 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=233
    FILE: Number of bytes written=287217
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=2476
    HDFS: Number of bytes written=155
    HDFS: Number of read operations=6

Mozilla Firefox Java - WordCounta/src... cloudera@quickstart:~
```

Map reducing is success and two items are found



The screenshot shows a terminal window titled "cloudera@quickstart:~" with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal displays the output of a Hadoop MapReduce job. The output includes statistics for map and reduce tasks, Map-Reduce Framework details (input/output records, bytes, materialized bytes, splits, combine, shuffle, and reduce counts), Shuffle Errors (all zero), File Input/Output Format Counters, and the results of a "hadoop fs -ls output" command. The command found two items: "output/_SUCCESS" (0 bytes) and "output/part-r-00000" (155 bytes). The terminal window is part of a desktop environment with a taskbar at the bottom showing "Mozilla Firefox", "Java - WordCounta/src...", and "cloudera@quickstart:~". The system tray on the right shows icons for network, volume, and power. The top status bar indicates the user is "cloudera" and the date/time is "Fri Jan 28, 5:21 PM".

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Total megabyte-milliseconds taken by all map tasks=11265024  
Total megabyte-milliseconds taken by all reduce tasks=7228416  
Map-Reduce Framework  
  Map input records=49  
  Map output records=34  
  Map output bytes=306  
  Map output materialized bytes=233  
  Input split bytes=131  
  Combine input records=34  
  Combine output records=18  
  Reduce input groups=18  
  Reduce shuffle bytes=233  
  Reduce input records=18  
  Reduce output records=18  
  Spilled Records=36  
  Shuffled Maps =1  
  Failed Shuffles=0  
  Merged Map outputs=1  
  GC time elapsed (ms)=229  
  CPU time spent (ms)=1840  
  Physical memory (bytes) snapshot=340992000  
  Virtual memory (bytes) snapshot=3015626752  
  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=2345  
File Output Format Counters  
  Bytes Written=155  
[cloudera@quickstart ~]$ hadoop fs -ls output  
Found 2 items  
-rw-r--r--  1 cloudera cloudera      0 2022-01-28 17:19 output/_SUCCESS  
-rw-r--r--  1 cloudera cloudera    155 2022-01-28 17:19 output/part-r-00000  
[cloudera@quickstart ~]$
```

Visualization of the found items

The screenshot shows the Hue File Browser interface in a Mozilla Firefox browser window. The browser's address bar displays the URL `quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera#/use`. The interface includes a top navigation bar with the Hue logo, a search bar, and a sidebar on the left with a 'Tables' section that states 'The database has no tables'. The main content area is titled 'File Browser' and shows the current path as `Home / user / cloudera / output`. A search bar and action buttons ('Upload', 'New') are at the top of the main area. Below, a table lists the contents of the directory:

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	.		cloudera	cloudera	drwxr-xr-x	January 28, 2022 05:19 PM
<input type="checkbox"/>	..		cloudera	cloudera	drwxr-xr-x	January 28, 2022 05:19 PM
<input type="checkbox"/>	_SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	January 28, 2022 05:19 PM
<input type="checkbox"/>	part-r-00000	155 bytes	cloudera	cloudera	-rw-r--r--	January 28, 2022 05:19 PM

At the bottom of the table, it indicates 'Show 45 of 2 items' and 'Page 1 of 1'. The taskbar at the bottom shows the active window 'Hue - File Browser - M...' and other open applications like 'Java - WordCounta/src...' and a terminal window with the prompt `[cloudera@quickstart:~]`.

Frequency count of the words that starts with 'a' in the given dataset is found and output is below

The screenshot shows the Hue File Browser interface in a Mozilla Firefox browser. The browser's address bar displays the URL `quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera/output`. The interface includes a top navigation bar with links to Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. A search bar is present with the text "Search data and saved documents...".

The main content area is titled "File Browser" and shows the file path `/ user / cloudera / output / part-r-00000`. On the left, there are options to "View as binary", "Edit file", "Download", "View file location", and "Refresh". Below these, file metadata is displayed: Last modified 01/29/2022 1:19 AM, User cloudera, Group cloudera, Size 155 B, and Mode 100644.

The file content is displayed in a table format, showing word frequency counts for words starting with 'a':

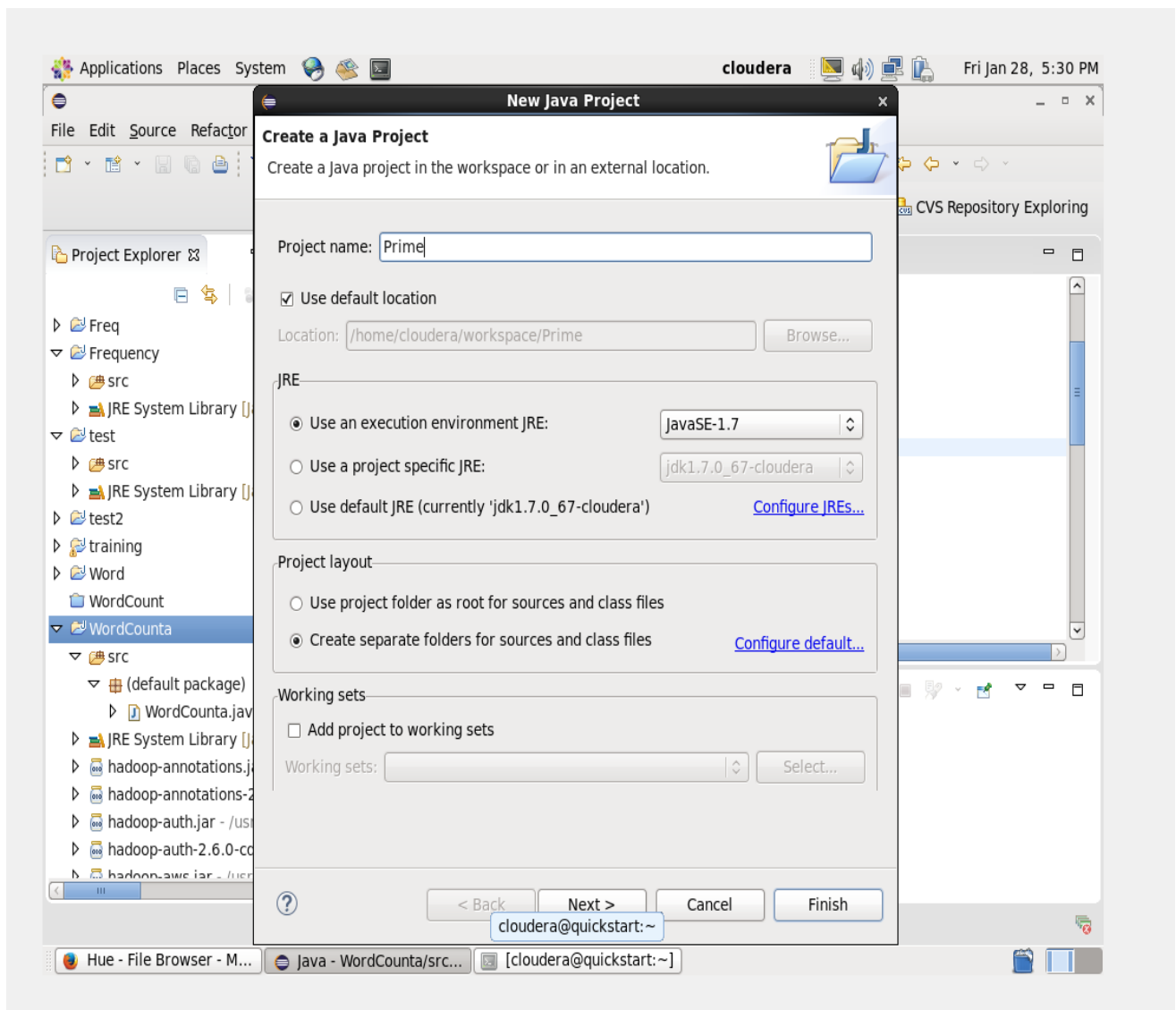
Word	Count
a	6
accidentally	1
acquired	1
after	1
ale	1
along	1
also	1
an	3
and	9
another	1
appear	1
army	1
arranged	1
as	1
aspirin.	1
at	2
attempted	1
authenticate	1

The bottom of the screenshot shows the operating system's taskbar with open applications: Hue - File Browser - M..., Java - WordCounta/src..., and a terminal window [cloudera@quickstart:~].

3. Determining the prime number in input and print number only once

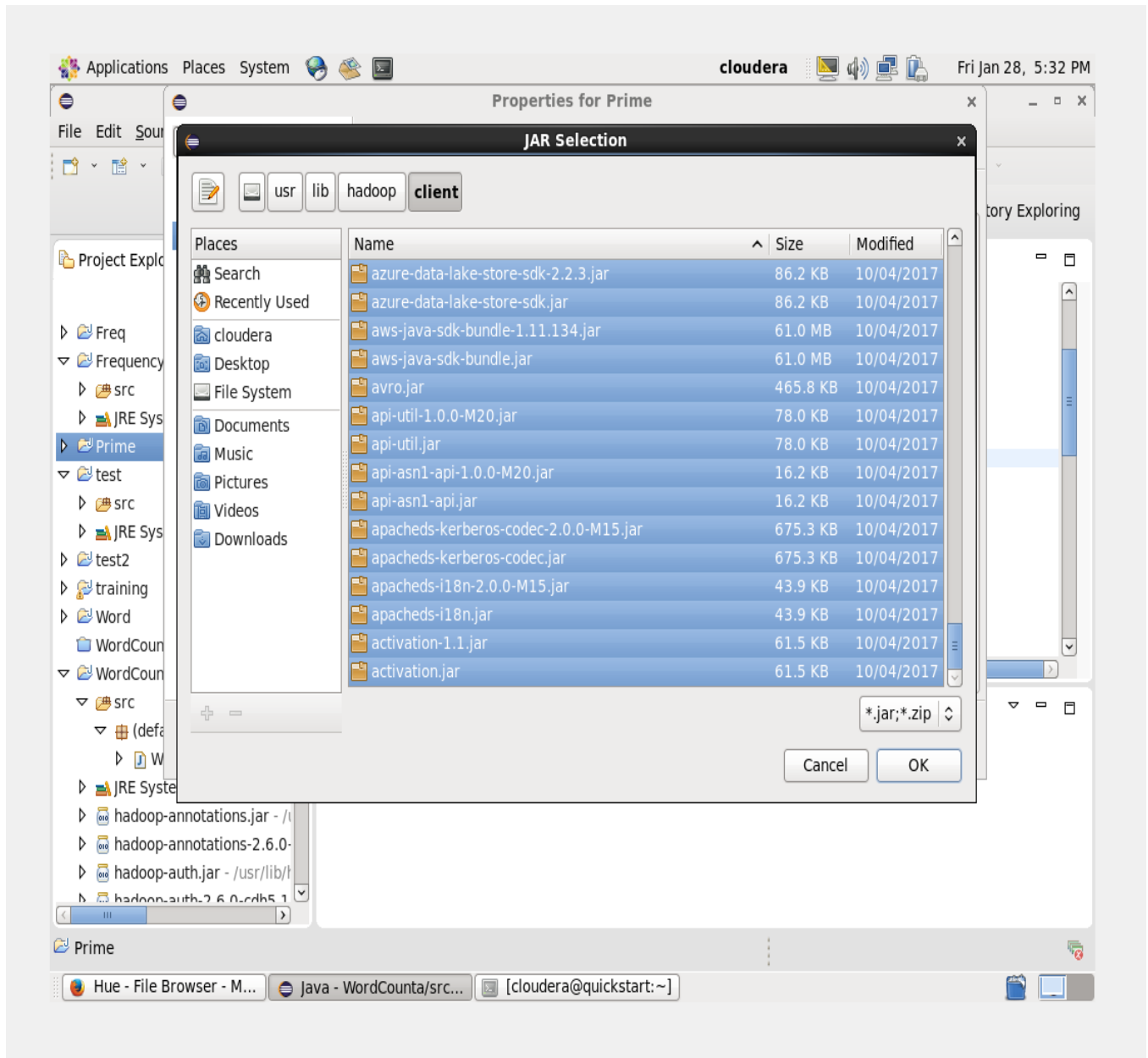
Step:1

Creating a new java project and naming it as Prime



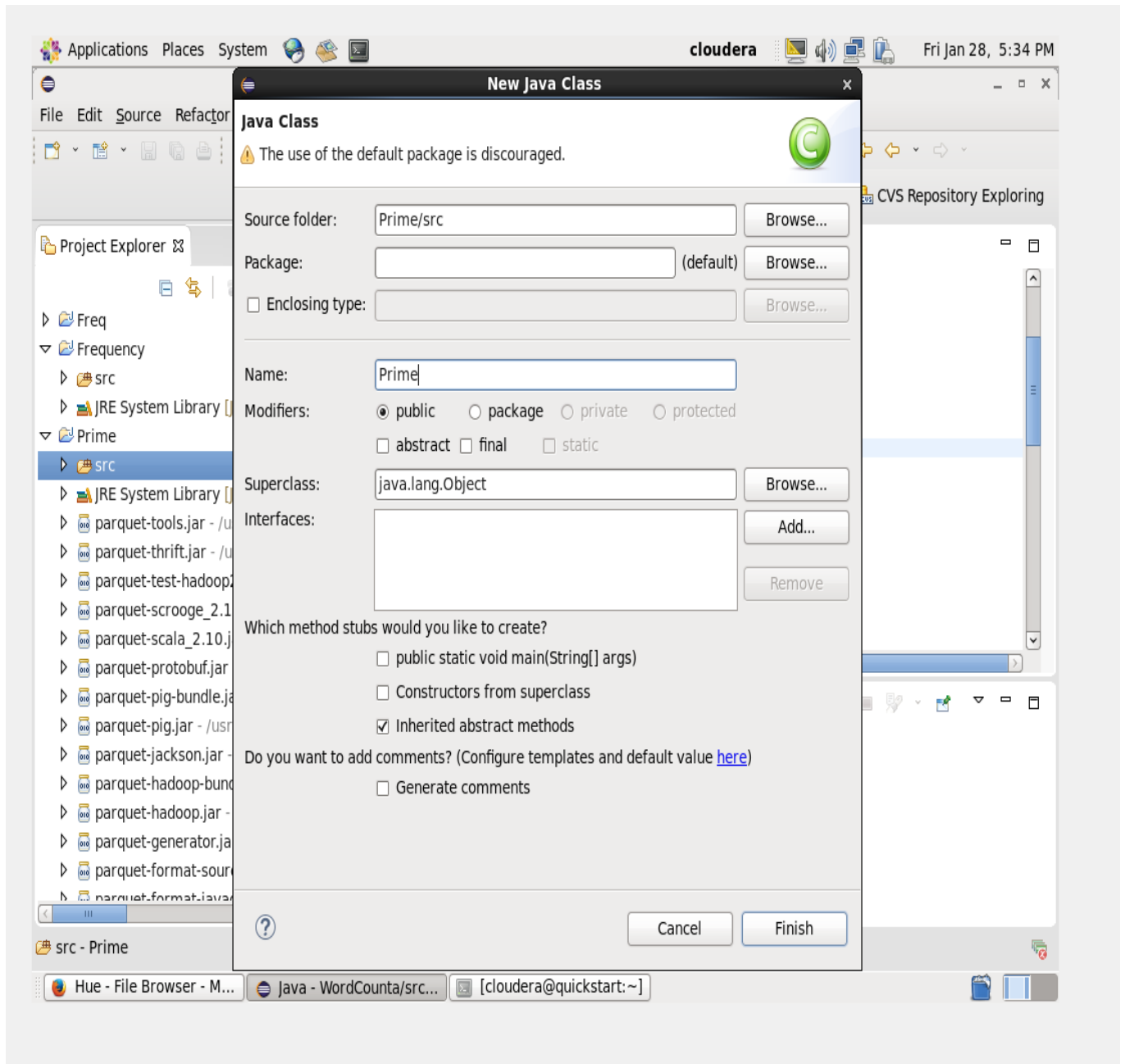
Step:2

Adding the JARs files of hadoop and client



Step:3

Creating the class file



Step:4

Logic:

```
int output =0;
```

```
int num = Integer.parseInt(Key.toString());
```

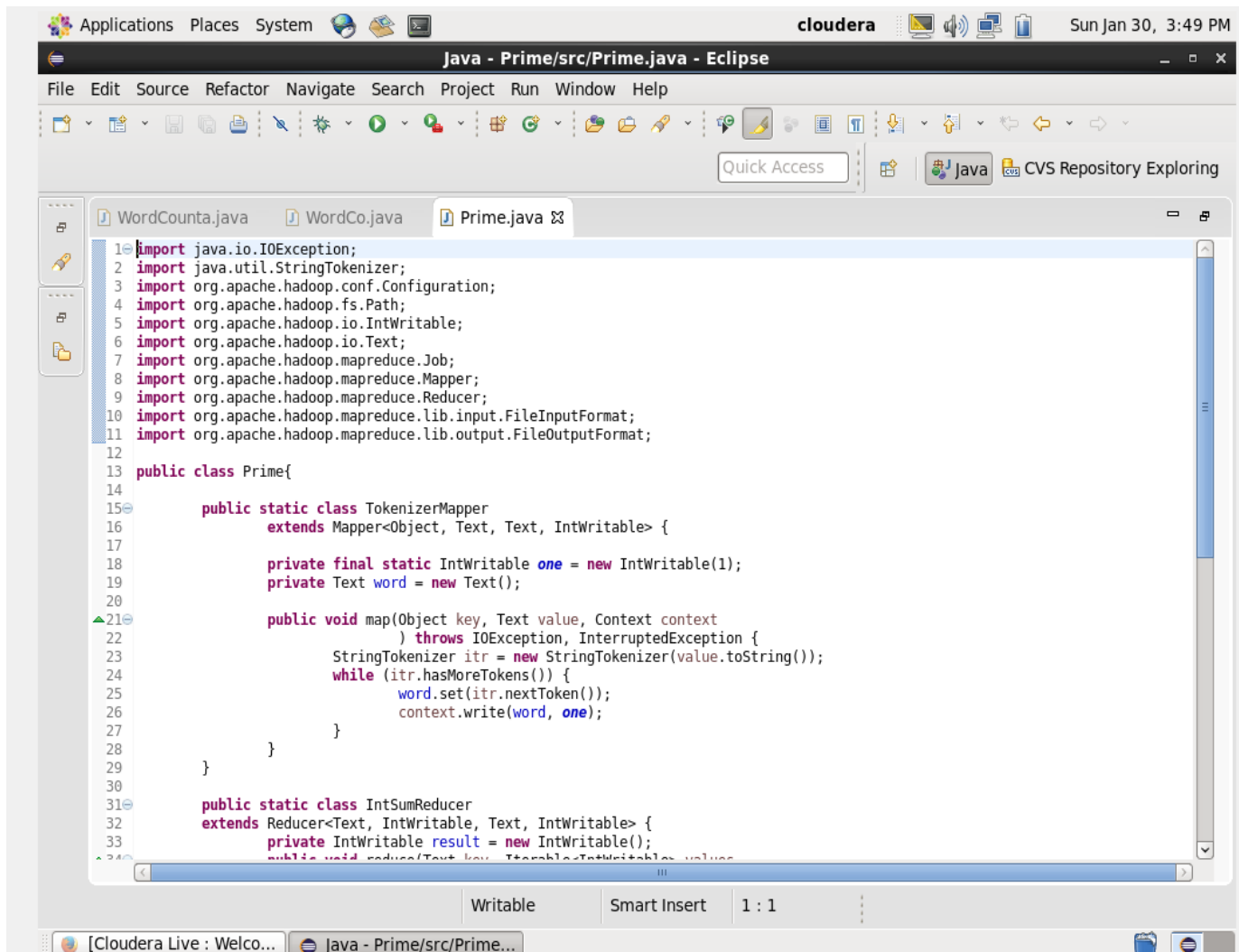
```
for(int i=2;i<num;i++) {
```

```
    if(num%i==0) {
```

```
        Output=1; }
```

```
}
```

Program code:

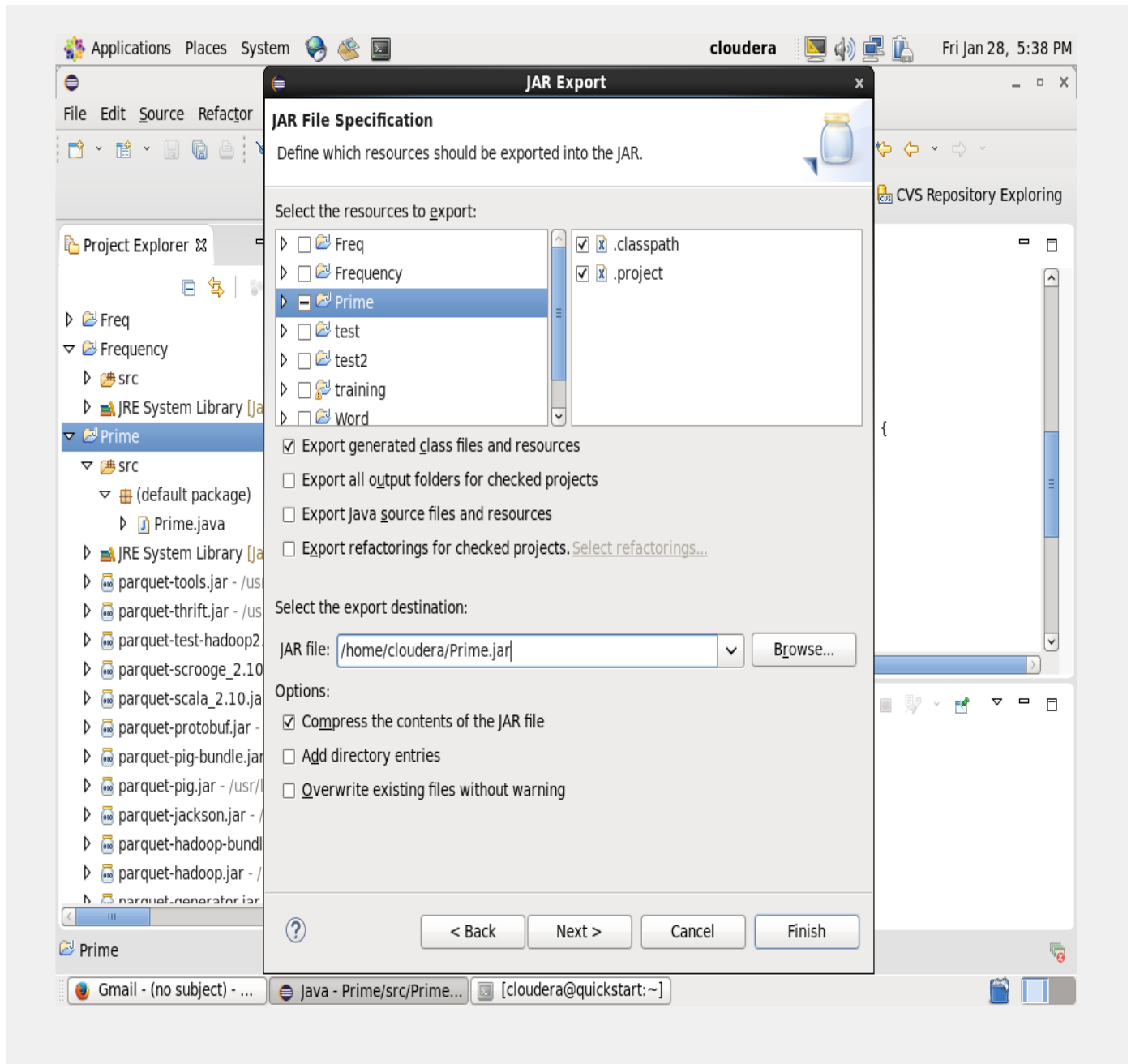


```
1 import java.io.IOException;
2 import java.util.StringTokenizer;
3 import org.apache.hadoop.conf.Configuration;
4 import org.apache.hadoop.fs.Path;
5 import org.apache.hadoop.io.IntWritable;
6 import org.apache.hadoop.io.Text;
7 import org.apache.hadoop.mapreduce.Job;
8 import org.apache.hadoop.mapreduce.Mapper;
9 import org.apache.hadoop.mapreduce.Reducer;
10 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
11 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
12
13 public class Prime{
14
15     public static class TokenizerMapper
16         extends Mapper<Object, Text, Text, IntWritable> {
17
18         private final static IntWritable one = new IntWritable(1);
19         private Text word = new Text();
20
21         public void map(Object key, Text value, Context context
22             ) throws IOException, InterruptedException {
23             StringTokenizer itr = new StringTokenizer(value.toString());
24             while (itr.hasMoreTokens()) {
25                 word.set(itr.nextToken());
26                 context.write(word, one);
27             }
28         }
29     }
30
31     public static class IntSumReducer
32         extends Reducer<Text, IntWritable, Text, IntWritable> {
33         private IntWritable result = new IntWritable();
34         public void reduce(Text key, Iterable<IntWritable> values,
```

```
WordCounta.java WordCo.java Prime.java
32 extends Reducer<Text, IntWritable, Text, IntWritable> {
33     private IntWritable result = new IntWritable();
34     public void reduce(Text key, Iterable<IntWritable> values,
35         Context context
36     ) throws IOException, InterruptedException {
37         int output = 0;
38         int num = Integer.parseInt(key.toString());
39
40         for (int i = 2; i < num; i++) {
41             if(num % i == 0) {
42                 output = 1;
43             }
44         }
45         result.set(output);
46         context.write(key, result);
47     }
48 }
49
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(Prime.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
64
65 }
```

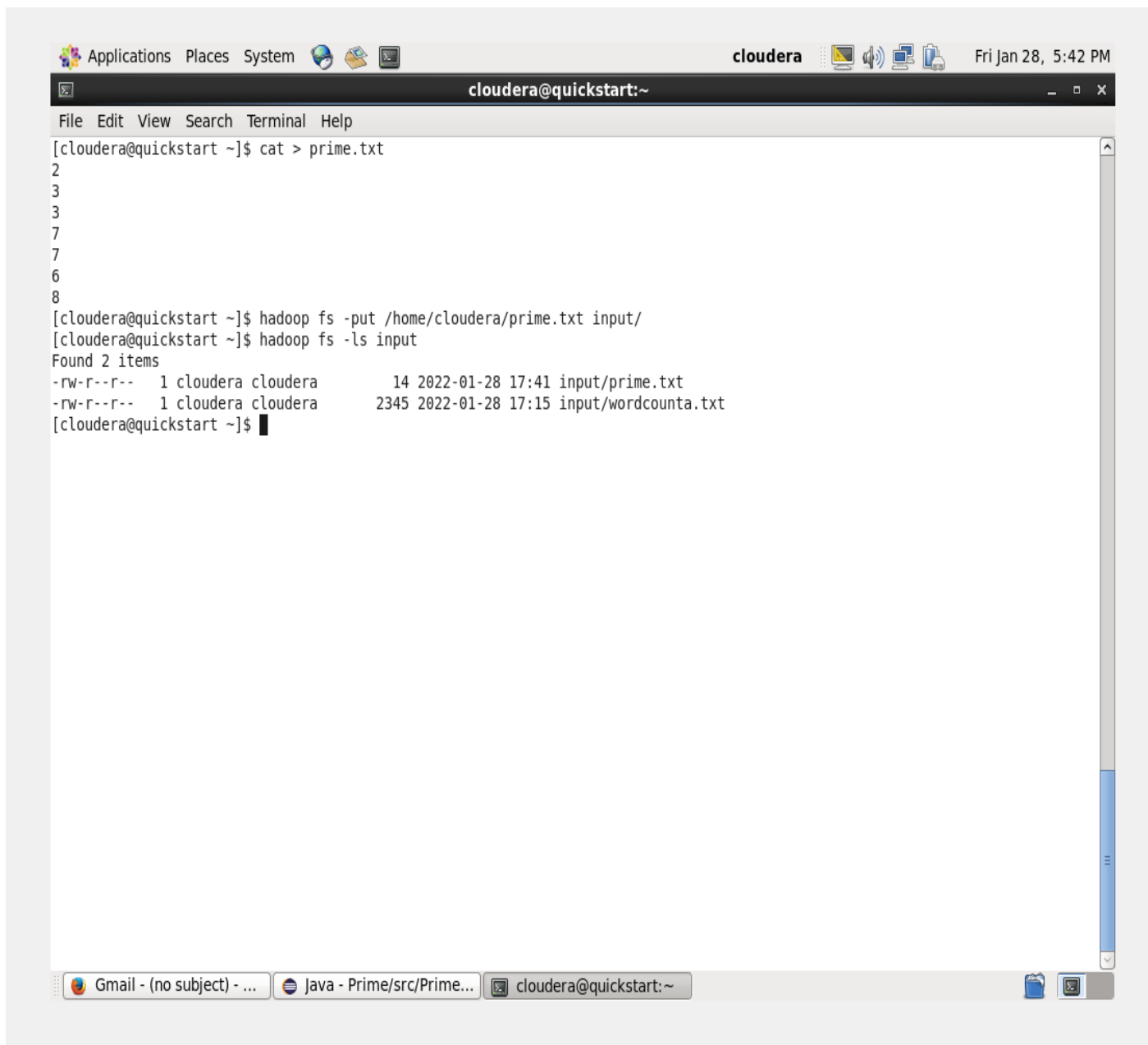
Step:5

Exporting the JARs file



Step:6

Creating a newfile called prime.txt and giving it the input data,then loading it into the hadoop hdfs



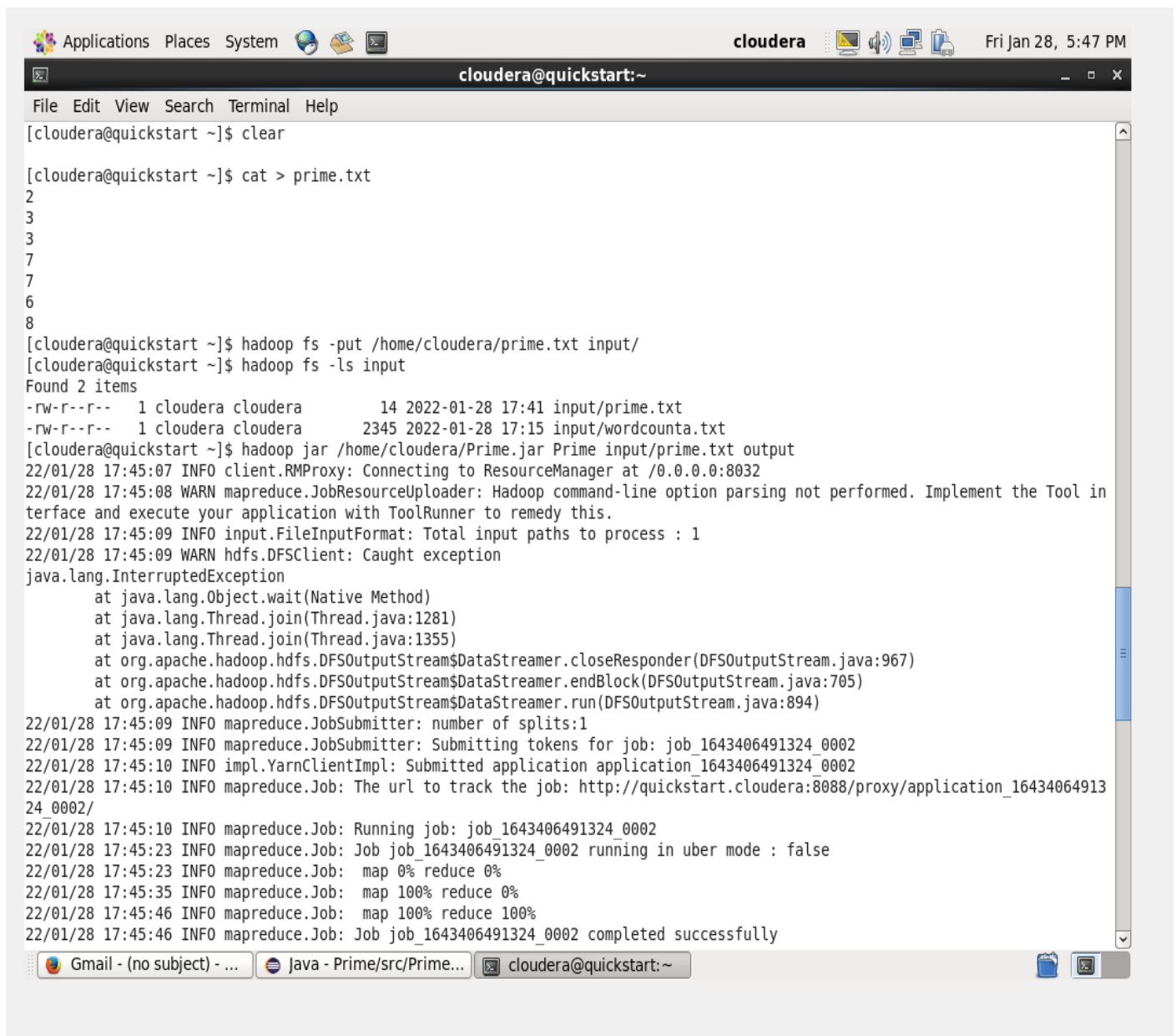
The screenshot shows a terminal window titled "cloudera@quickstart:~" with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal output is as follows:

```
[cloudera@quickstart ~]$ cat > prime.txt
2
3
3
7
7
6
8
[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/prime.txt input/
[cloudera@quickstart ~]$ hadoop fs -ls input
Found 2 items
-rw-r--r--  1 cloudera cloudera      14 2022-01-28 17:41 input/prime.txt
-rw-r--r--  1 cloudera cloudera   2345 2022-01-28 17:15 input/wordcounta.txt
[cloudera@quickstart ~]$
```

The terminal window is part of a desktop environment with a top bar showing "Applications", "Places", "System", and system status "cloudera" with icons for network, volume, and power. The date and time "Fri Jan 28, 5:42 PM" are also displayed. The bottom taskbar shows three open windows: "Gmail - (no subject) - ...", "Java - Prime/src/Prime...", and the active terminal window "cloudera@quickstart:~".

Step:7

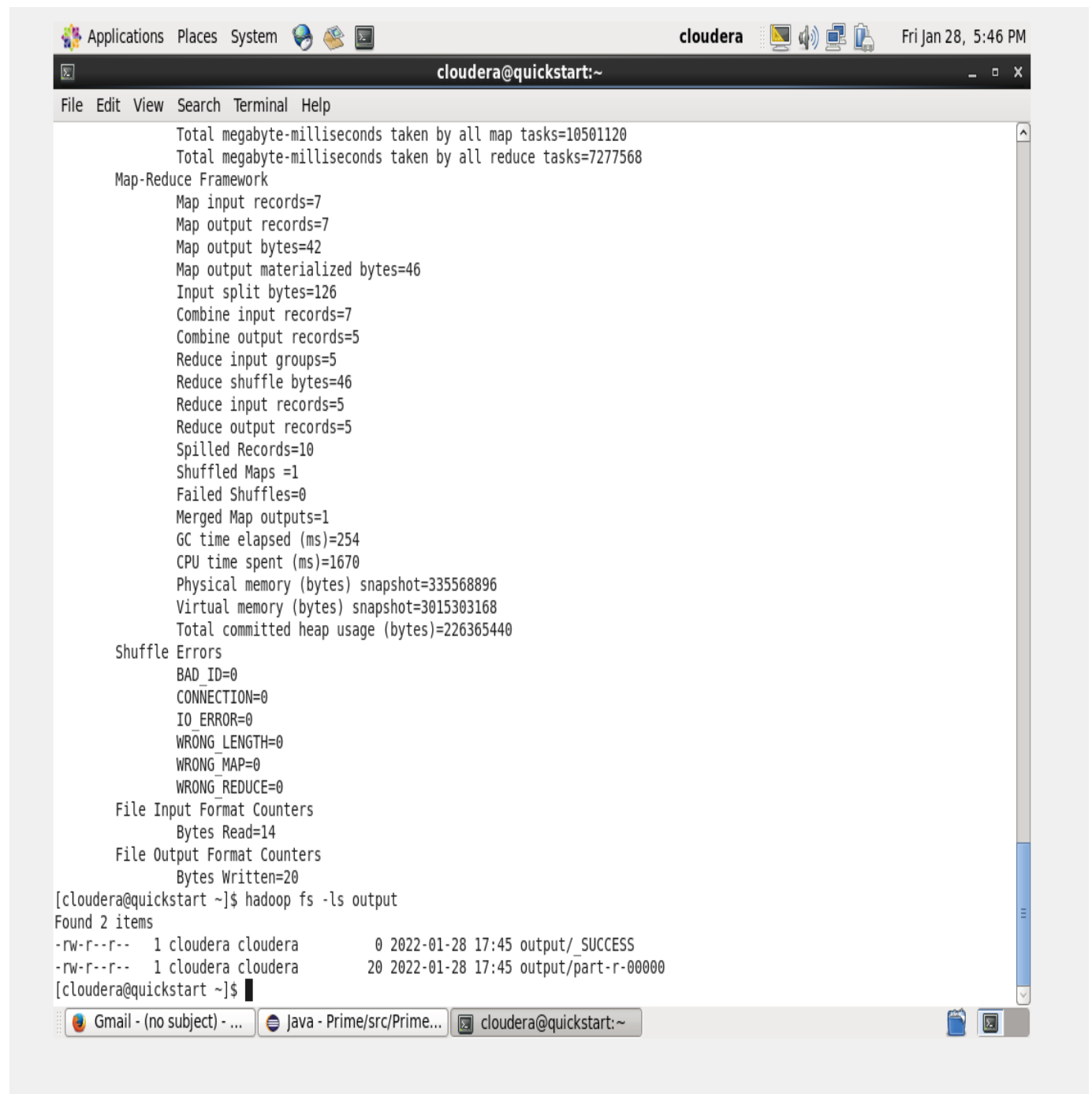
Map reducing



```
Applications Places System cloudera Fri Jan 28, 5:47 PM
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ clear

[cloudera@quickstart ~]$ cat > prime.txt
2
3
3
7
7
6
8
[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/prime.txt input/
[cloudera@quickstart ~]$ hadoop fs -ls input
Found 2 items
-rw-r--r-- 1 cloudera cloudera      14 2022-01-28 17:41 input/prime.txt
-rw-r--r-- 1 cloudera cloudera    2345 2022-01-28 17:15 input/wordcounta.txt
[cloudera@quickstart ~]$ hadoop jar /home/cloudera/Prime.jar Prime input/prime.txt output
22/01/28 17:45:07 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/01/28 17:45:08 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in
terface and execute your application with ToolRunner to remedy this.
22/01/28 17:45:09 INFO input.FileInputFormat: Total input paths to process : 1
22/01/28 17:45:09 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/01/28 17:45:09 INFO mapreduce.JobSubmitter: number of splits:1
22/01/28 17:45:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1643406491324_0002
22/01/28 17:45:10 INFO impl.YarnClientImpl: Submitted application application_1643406491324_0002
22/01/28 17:45:10 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_16434064913
24_0002/
22/01/28 17:45:10 INFO mapreduce.Job: Running job: job_1643406491324_0002
22/01/28 17:45:23 INFO mapreduce.Job: Job job_1643406491324_0002 running in uber mode : false
22/01/28 17:45:23 INFO mapreduce.Job:  map 0% reduce 0%
22/01/28 17:45:35 INFO mapreduce.Job:  map 100% reduce 0%
22/01/28 17:45:46 INFO mapreduce.Job:  map 100% reduce 100%
22/01/28 17:45:46 INFO mapreduce.Job: Job job_1643406491324_0002 completed successfully
```

Map reducing is done and two items are found



The screenshot shows a terminal window titled "cloudera@quickstart:~" with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal output displays the completion of a Hadoop MapReduce job, followed by a file listing command and its results.

```
Total megabyte-milliseconds taken by all map tasks=10501120
Total megabyte-milliseconds taken by all reduce tasks=7277568
Map-Reduce Framework
  Map input records=7
  Map output records=7
  Map output bytes=42
  Map output materialized bytes=46
  Input split bytes=126
  Combine input records=7
  Combine output records=5
  Reduce input groups=5
  Reduce shuffle bytes=46
  Reduce input records=5
  Reduce output records=5
  Spilled Records=10
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=254
  CPU time spent (ms)=1670
  Physical memory (bytes) snapshot=335568896
  Virtual memory (bytes) snapshot=3015303168
  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=14
File Output Format Counters
  Bytes Written=20
[cloudera@quickstart ~]$ hadoop fs -ls output
Found 2 items
-rw-r--r--  1 cloudera cloudera      0 2022-01-28 17:45 output/_SUCCESS
-rw-r--r--  1 cloudera cloudera    20 2022-01-28 17:45 output/part-r-00000
[cloudera@quickstart ~]$
```

The terminal window has a taskbar at the bottom with three open applications: "Gmail - (no subject) - ...", "Java - Prime/src/Prime...", and "cloudera@quickstart:~". The system clock in the top right corner indicates "Fri Jan 28, 5:46 PM".

Visualizing the items found

The screenshot shows the Hue File Browser interface in a Mozilla Firefox browser window. The address bar displays the URL: `quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera#/use`. The interface includes a top navigation bar with links to Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. Below this is a search bar and a 'Query' button. The main content area is titled 'File Browser' and shows a directory listing for the path `Home / user / cloudera / output`. The listing includes a table with columns for Name, Size, User, Group, Permissions, and Date. The table contains four entries: a folder named `.`, a folder named `..`, a file named `._SUCCESS` (0 bytes), and a file named `part-r-00000` (20 bytes). The bottom of the interface shows a pagination bar indicating 'Page 1 of 1' and a total of 2 items.

Applications Places System cloudera Fri Jan 28, 5:49 PM

Hue - File Browser - Mozilla Firefox

Hue - File Browser x

quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera#/use Search

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Hue Query Search data and saved documents... Jobs cloudera

File Browser

Search for file name Actions Move to trash Upload New

Home / user / cloudera / output

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	.		cloudera	cloudera	drwxr-xr-x	January 28, 2022 05:45 PM
<input type="checkbox"/>	..		cloudera	cloudera	drwxr-xr-x	January 28, 2022 05:45 PM
<input type="checkbox"/>	._SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	January 28, 2022 05:45 PM
<input type="checkbox"/>	part-r-00000	20 bytes	cloudera	cloudera	-rw-r--r--	January 28, 2022 05:45 PM

Show 45 of 2 items Page 1 of 1

[Gmail - (no subject...)] [Java - Prime/src/Pr...] [cloudera@quickst...] [cloudera@quicksta...] Hue - File Browser - ...

Prime numbers are determined in the given input and the output is shown in the below figure

The screenshot displays the Hue File Browser interface within a Mozilla Firefox browser window. The address bar shows the URL `quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera/output`. The interface includes a top navigation bar with links to Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. The main content area shows the file browser view for the path `/ user / cloudera / output /`, displaying a file named `part-r-00000`. The file content is shown as a table with two columns, representing prime numbers and their corresponding values.

2	0
3	0
6	1
7	0
8	1

The interface also includes a sidebar with navigation options, a top navigation bar, and a bottom taskbar with open applications.

Overall Explanation of the ICE-2

Using the Hadoop MapReducing technique we found the required output. MapReducing technique basically has two functions 1. Mapper 2. Reducer.

Mapper takes the input data given and forms the key/value pairs for that data.

Reducer in turn, receives the key/value pairs as input from the mapper and performs the functions of shuffling, sorting and reducing.

Thus large datasets are reduced and required data/output is obtained by using hadoop mapreducing algorithm.

