

1. Problem Statement

We have a dataset of sales of different TV sets across different locations.

Records look like:

Samsung|Optima|14|Madhya Pradesh|132401|14200

The fields are arranged like:

Company Name|Product Name|Size in inches|State|Pin Code|Price

There are some invalid records which contain 'NA' in either Company Name or Product Name.

Task 1:

Write a Map Reduce program to filter out the invalid records. Map only job will fit for this context.

Task 2:

Write a Map Reduce program to calculate the total units sold for each Company.

Task 3:

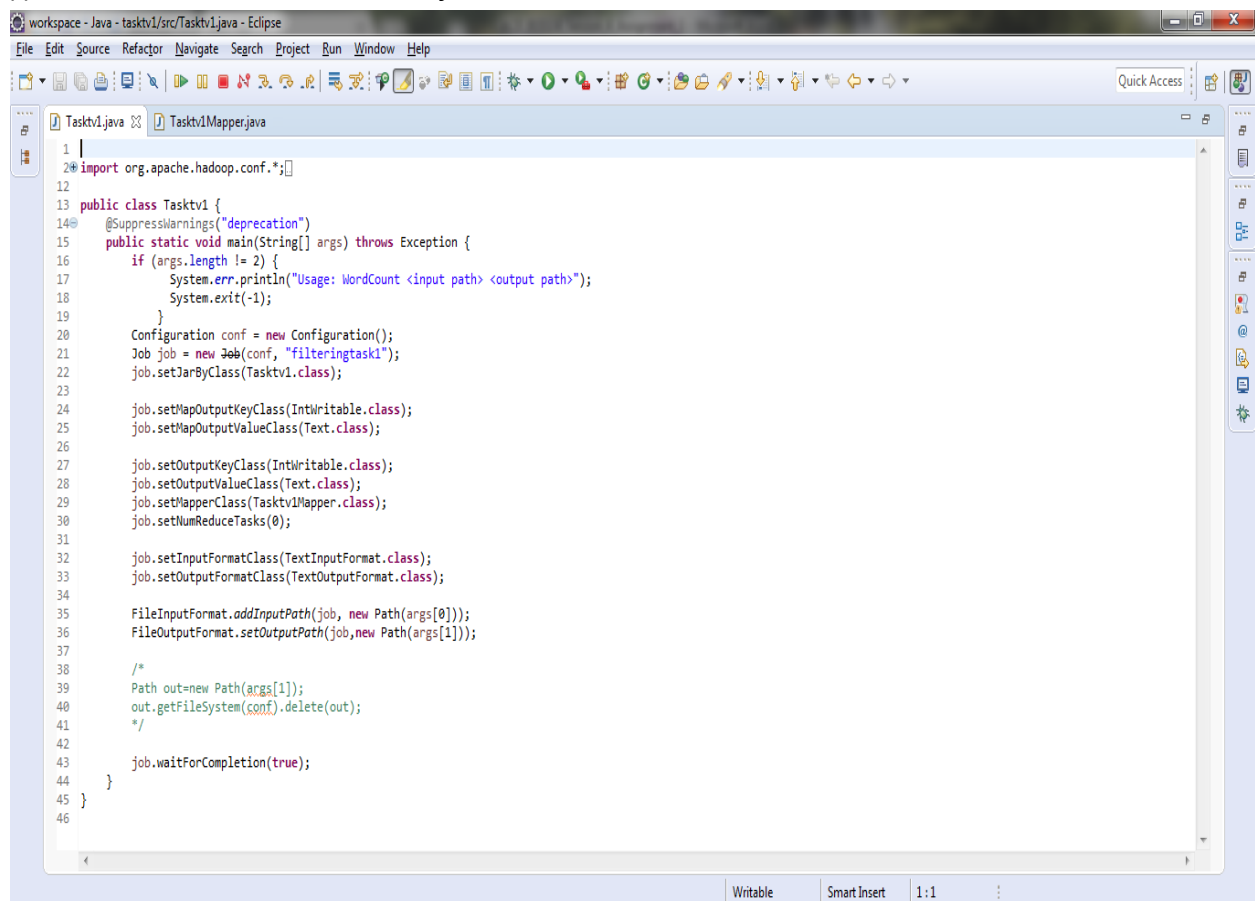
Write a Map Reduce program to calculate the total units sold in each state for Onida company.

Task1:-

COMMAND:-

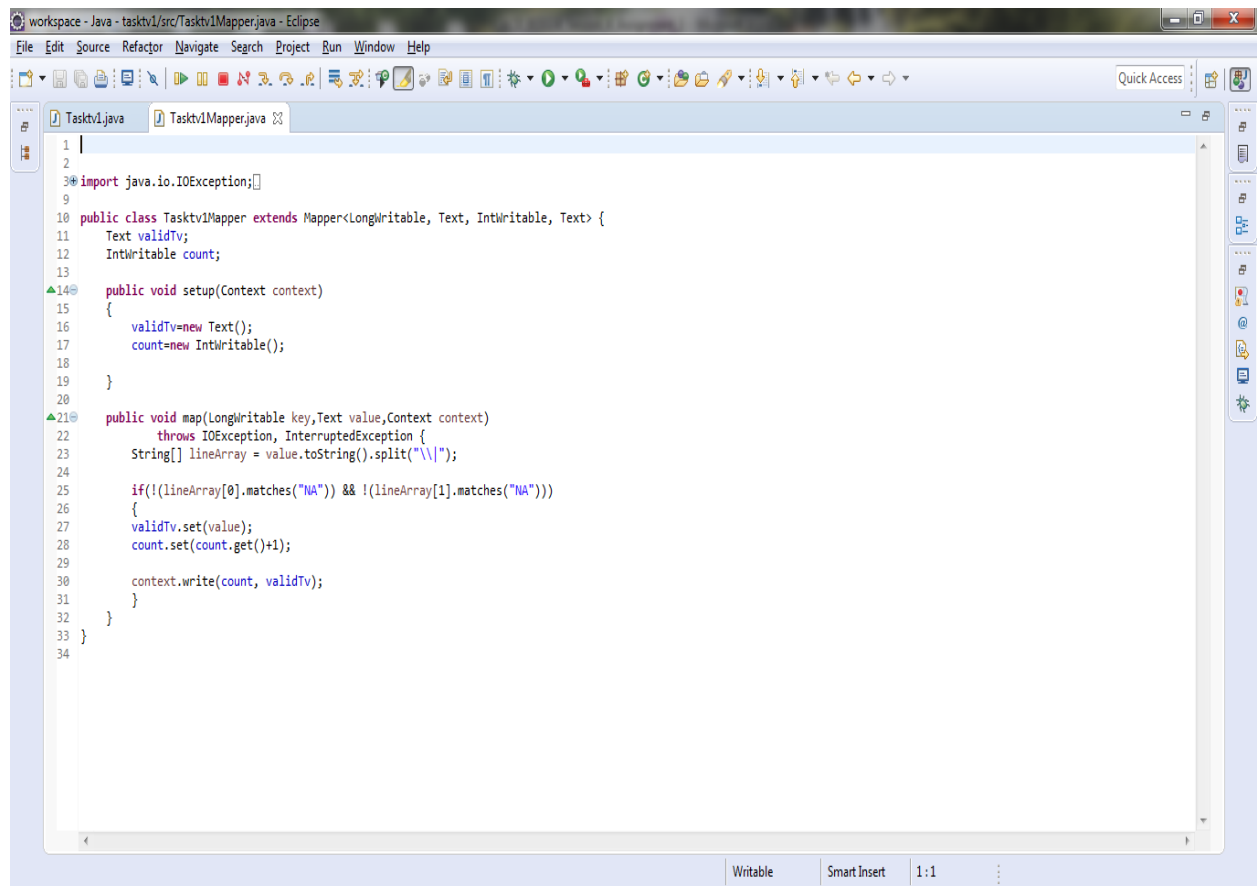
create two java files,

(i)driver code file named Tasktv1.java.



```
1 |
2 | import org.apache.hadoop.conf.*;
3 |
4 |
5 | public class Tasktv1 {
6 |     @SuppressWarnings("deprecation")
7 |     public static void main(String[] args) throws Exception {
8 |         if (args.length != 2) {
9 |             System.err.println("Usage: WordCount <input path> <output path>");
10 |            System.exit(-1);
11 |        }
12 |        Configuration conf = new Configuration();
13 |        Job job = new Job(conf, "filteringtask1");
14 |        job.setJarByClass(Tasktv1.class);
15 |
16 |        job.setMapOutputKeyClass(IntWritable.class);
17 |        job.setMapOutputValueClass(Text.class);
18 |
19 |        job.setOutputKeyClass(IntWritable.class);
20 |        job.setOutputValueClass(Text.class);
21 |        job.setMapperClass(Tasktv1Mapper.class);
22 |        job.setNumReduceTasks(0);
23 |
24 |        job.setInputFormatClass(TextInputFormat.class);
25 |        job.setOutputFormatClass(TextOutputFormat.class);
26 |
27 |        FileInputFormat.addInputPath(job, new Path(args[0]));
28 |        FileOutputFormat.setOutputPath(job, new Path(args[1]));
29 |
30 |        /*
31 |         Path out=new Path(args[1]);
32 |         out.getFileSystem(conf).delete(out);
33 |         */
34 |
35 |        job.waitForCompletion(true);
36 |    }
37 | }
```

(ii)Mapper Class file named Taskv1Mapper.java



```
1 |
2 |
3 | import java.io.IOException;
4 |
5 |
6 |
7 |
8 |
9 |
10 | public class Taskv1Mapper extends Mapper<LongWritable, Text, IntWritable, Text> {
11 |     Text validTv;
12 |     IntWritable count;
13 |
14 |     public void setup(Context context)
15 |     {
16 |         validTv=new Text();
17 |         count=new IntWritable();
18 |     }
19 |
20 |
21 |     public void map(LongWritable key,Text value,Context context)
22 |         throws IOException, InterruptedException {
23 |         String[] lineArray = value.toString().split("\\|");
24 |
25 |         if(!(lineArray[0].matches("NA")) && !(lineArray[1].matches("NA")))
26 |         {
27 |             validTv.set(value);
28 |             count.set(count.get()+1);
29 |
30 |             context.write(count, validTv);
31 |         }
32 |     }
33 | }
34 |
```

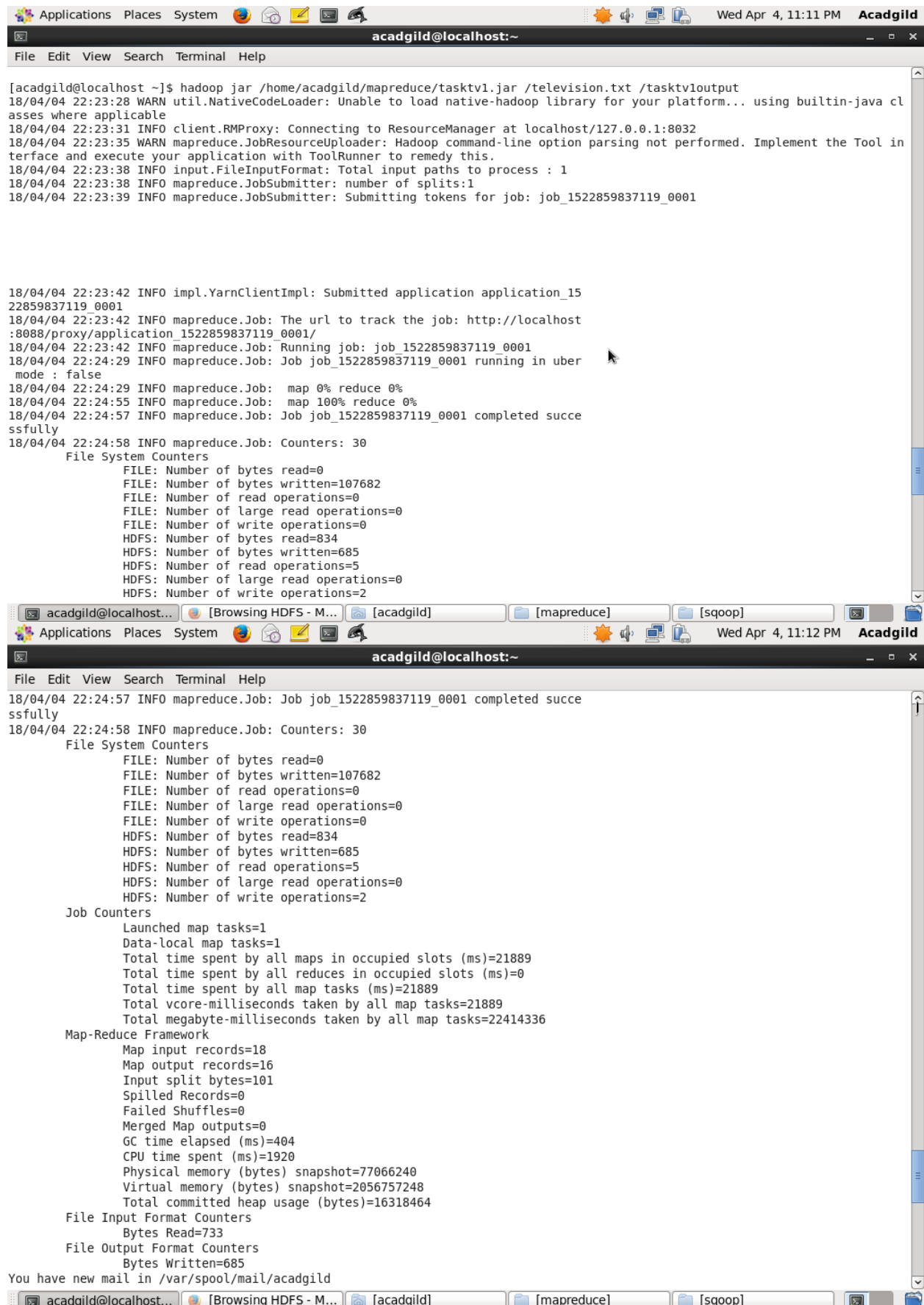
EXPLANATION:-

Map Reduce program to filter out the invalid records. Map only job will fit for this context.

Here I Use the television.txt file which is attached to the git .
the tasktv1.jar file also created for run at VM,this jar file also attached to git.

Here all "NA" records rows are deleted.

OUTPUT:-

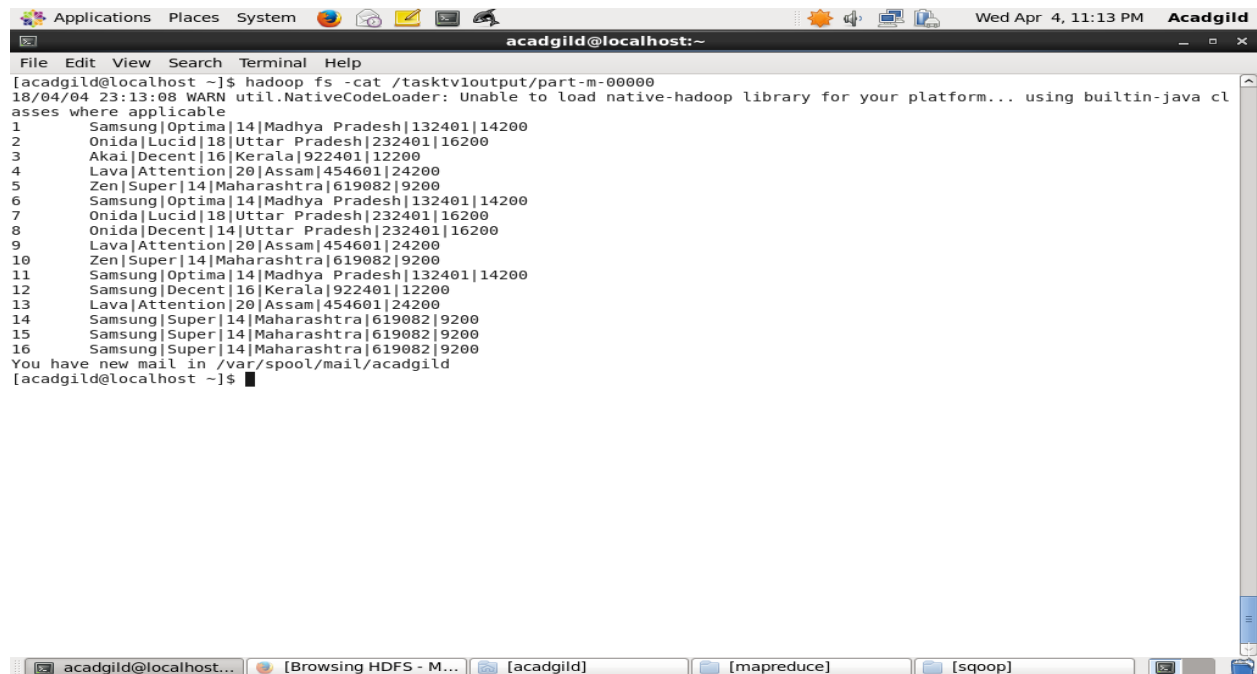


The screenshot shows a terminal window titled 'acadgild@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal displays the output of a Hadoop MapReduce job. The job is identified by ID 1522859837119_0001. The output includes various status messages, progress updates, and a detailed summary of counters at the end.

```
[acadgild@localhost ~]$ hadoop jar /home/acadgild/mapreduce/tasktv1.jar /television.txt /tasktv1output
18/04/04 22:23:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/04/04 22:23:31 INFO client.RMPProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/04/04 22:23:35 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in interface and execute your application with ToolRunner to remedy this.
18/04/04 22:23:38 INFO input.FileInputFormat: Total input paths to process : 1
18/04/04 22:23:38 INFO mapreduce.JobSubmitter: number of splits:1
18/04/04 22:23:39 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1522859837119_0001

18/04/04 22:23:42 INFO impl.YarnClientImpl: Submitted application application_1522859837119_0001
18/04/04 22:23:42 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1522859837119_0001/
18/04/04 22:23:42 INFO mapreduce.Job: Running job: job_1522859837119_0001
18/04/04 22:24:29 INFO mapreduce.Job: Job job_1522859837119_0001 running in uber mode : false
18/04/04 22:24:29 INFO mapreduce.Job: map 0% reduce 0%
18/04/04 22:24:55 INFO mapreduce.Job: map 100% reduce 0%
18/04/04 22:24:57 INFO mapreduce.Job: Job job_1522859837119_0001 completed successfully
18/04/04 22:24:58 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=107682
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=834
    HDFS: Number of bytes written=685
    HDFS: Number of read operations=5
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2

acadgild@localhost... [Browsing HDFS - M... [acadgild] [mapreduce] [sqoop]
18/04/04 22:24:57 INFO mapreduce.Job: Job job_1522859837119_0001 completed successfully
18/04/04 22:24:58 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=107682
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=834
    HDFS: Number of bytes written=685
    HDFS: Number of read operations=5
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=21889
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=21889
    Total vcore-milliseconds taken by all map tasks=21889
    Total megabyte-milliseconds taken by all map tasks=22414336
  Map-Reduce Framework
    Map input records=18
    Map output records=16
    Input split bytes=101
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=404
    CPU time spent (ms)=1920
    Physical memory (bytes) snapshot=77066240
    Virtual memory (bytes) snapshot=2056757248
    Total committed heap usage (bytes)=16318464
  File Input Format Counters
    Bytes Read=733
  File Output Format Counters
    Bytes Written=685
You have new mail in /var/spool/mail/acadgild
acadgild@localhost... [Browsing HDFS - M... [acadgild] [mapreduce] [sqoop]
```



The terminal window shows a user running the command `hadoop fs -cat /tasktv1output/part-m-00000`. The output lists 16 lines of data, each containing a device name, a location, and two numerical values. A warning message from `util.NativeCodeLoader` is also visible. The terminal window is titled `acadmild@localhost:~` and has a menu bar with `File Edit View Search Terminal Help`. The bottom of the terminal window shows a taskbar with several open windows: `acadmild@localhost...`, `[Browsing HDFS - M...`, `[acadmild]`, `[mapreduce]`, and `[sqoop]`.

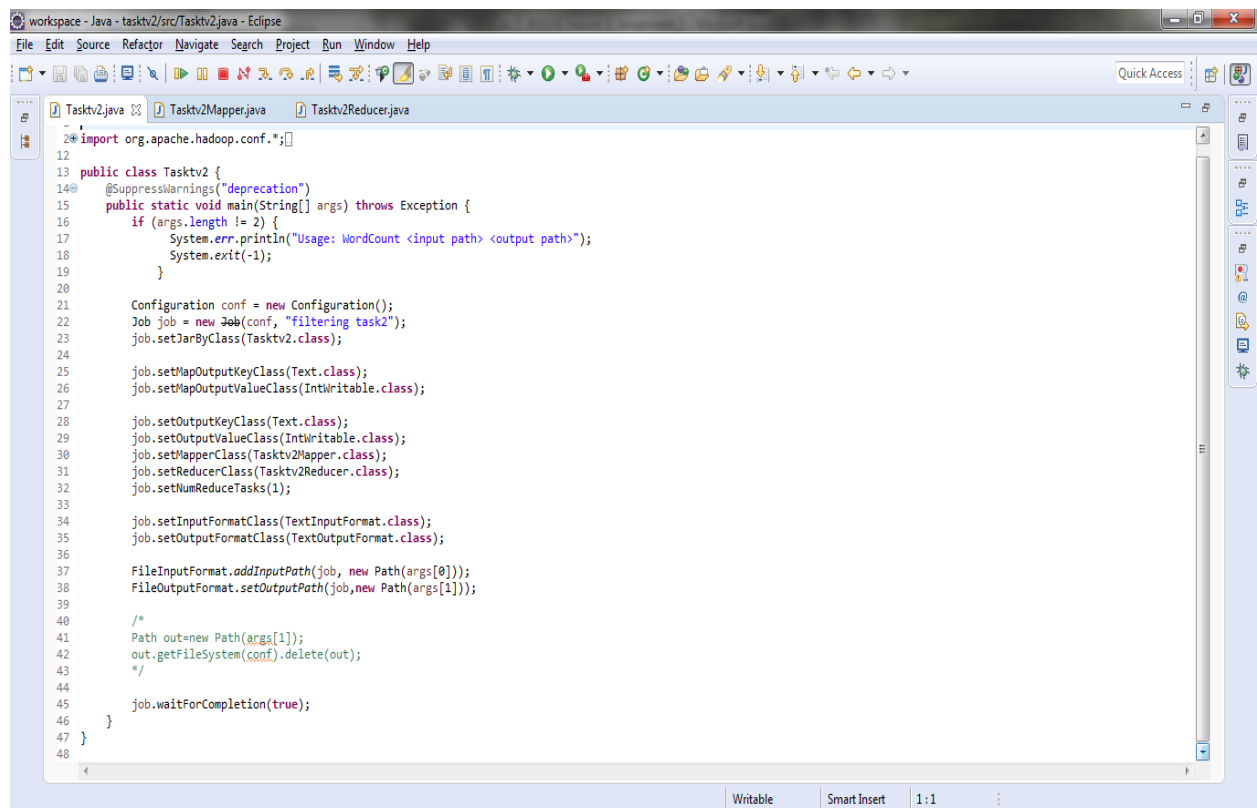
```
[acadmild@localhost ~]$ hadoop fs -cat /tasktv1output/part-m-00000
18/04/04 23:13:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
1 Samsung|Optima|14|Madhya Pradesh|132401|14200
2 Onida|Lucid|18|Uttar Pradesh|232401|16200
3 Akai|Decent|16|Kerala|922401|12200
4 Lava|Attention|20|Assam|454601|24200
5 Zen|Super|14|Maharashtra|619082|9200
6 Samsung|Optima|14|Madhya Pradesh|132401|14200
7 Onida|Lucid|18|Uttar Pradesh|232401|16200
8 Onida|Decent|14|Uttar Pradesh|232401|16200
9 Lava|Attention|20|Assam|454601|24200
10 Zen|Super|14|Maharashtra|619082|9200
11 Samsung|Optima|14|Madhya Pradesh|132401|14200
12 Samsung|Decent|16|Kerala|922401|12200
13 Lava|Attention|20|Assam|454601|24200
14 Samsung|Super|14|Maharashtra|619082|9200
15 Samsung|Super|14|Maharashtra|619082|9200
16 Samsung|Super|14|Maharashtra|619082|9200
You have new mail in /var/spool/mail/acadmild
[acadmild@localhost ~]$
```

Task2:-

COMMAND:-

Three files are created,

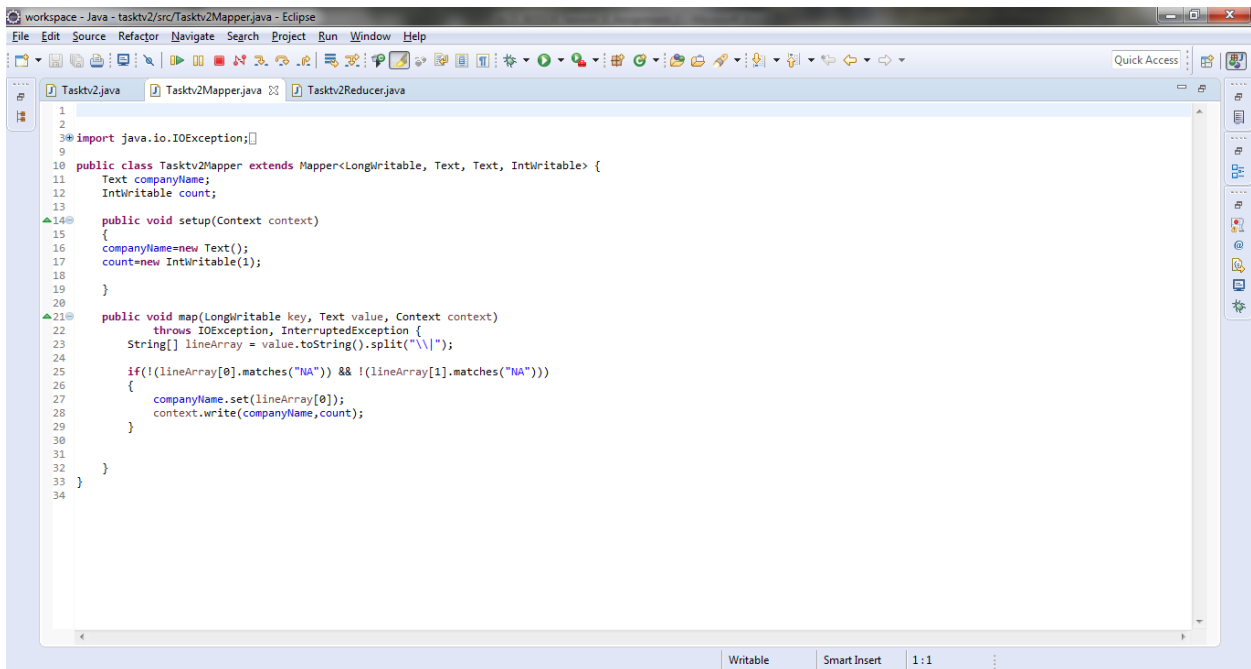
(i) Driver code file named Taskv2.java.



The Eclipse IDE shows the code for `Taskv2.java`. The code is a Java class that implements a Hadoop MapReduce job. It includes imports for `org.apache.hadoop.conf.*` and `org.apache.hadoop.mapreduce.*`. The `main` method sets up the job configuration, including the input and output paths, and runs the job. The code is as follows:

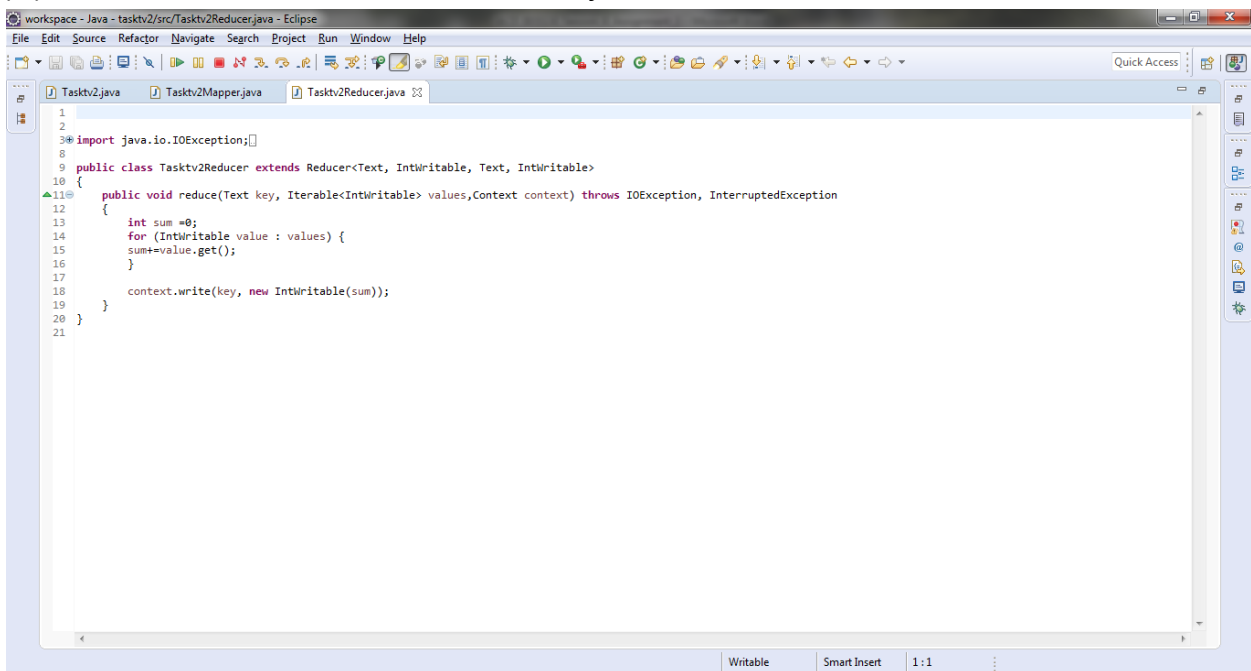
```
1 import org.apache.hadoop.conf.*;
2
3 public class Taskv2 {
4     @SuppressWarnings("deprecation")
5     public static void main(String[] args) throws Exception {
6         if (args.length != 2) {
7             System.err.println("Usage: WordCount <input path> <output path>");
8             System.exit(-1);
9         }
10
11         Configuration conf = new Configuration();
12         Job job = new Job(conf, "filtering task2");
13         job.setJarByClass(Taskv2.class);
14
15         job.setMapOutputKeyClass(Text.class);
16         job.setMapOutputValueClass(IntWritable.class);
17
18         job.setOutputKeyClass(Text.class);
19         job.setOutputValueClass(IntWritable.class);
20         job.setMapperClass(Taskv2Mapper.class);
21         job.setReducerClass(Taskv2Reducer.class);
22         job.setNumReduceTasks(1);
23
24         job.setInputFormatClass(TextInputFormat.class);
25         job.setOutputFormatClass(TextOutputFormat.class);
26
27         FileInputFormat.addInputPath(job, new Path(args[0]));
28         FileOutputFormat.setOutputPath(job, new Path(args[1]));
29
30         /*
31          * Path out=new Path(args[1]);
32          * out.getFileSystem(conf).delete(out);
33          */
34
35         job.waitForCompletion(true);
36     }
37 }
```

(ii) Mapper Class file named Tasktv2Mapper.java.



```
1
2
3 import java.io.IOException;
4
5
6
7
8
9 public class Tasktv2Mapper extends Mapper<LongWritable, Text, Text, IntWritable> {
10     Text companyName;
11     IntWritable count;
12
13
14     public void setup(Context context)
15     {
16         companyName=new Text();
17         count=new IntWritable(1);
18     }
19
20
21     public void map(LongWritable key, Text value, Context context)
22     throws IOException, InterruptedException {
23         String[] lineArray = value.toString().split("\\\\");
24
25         if(!lineArray[0].matches("NA")) && !(lineArray[1].matches("NA"))
26         {
27             companyName.set(lineArray[0]);
28             context.write(companyName,count);
29         }
30
31     }
32 }
33
34
```

(iii)Reducer Class file named Tasktv2Reducer.java.



```
1
2
3 import java.io.IOException;
4
5
6
7
8 public class Tasktv2Reducer extends Reducer<Text, IntWritable, Text, IntWritable>
9 {
10     public void reduce(Text key, Iterable<IntWritable> values,Context context) throws IOException, InterruptedException
11     {
12         int sum =0;
13         for (IntWritable value : values) {
14             sum+=value.get();
15         }
16
17         context.write(key, new IntWritable(sum));
18     }
19 }
20
21
```

EXPLANATION:-

television.txt file is used here .It is attached with git.

tasktv2.jar file is created.It is also attached with git.

Total unit sold for each Company is found here.

OUTPUT:-

```
Applications Places System Thu Apr 5, 9:47 AM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ hadoop jar /home/acadgild/mapreduce/tasktv2.jar /television.txt /tasktv2output
18/04/05 09:43:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/04/05 09:43:57 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/04/05 09:44:05 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
18/04/05 09:44:10 INFO input.FileInputFormat: Total input paths to process : 1
18/04/05 09:44:11 INFO mapreduce.JobSubmitter: number of splits:1
18/04/05 09:44:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1522901587164_0001
18/04/05 09:44:17 INFO impl.YarnClientImpl: Submitted application application_1522901587164_0001
18/04/05 09:44:17 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1522901587164_0001/
18/04/05 09:44:17 INFO mapreduce.Job: Running job: job_1522901587164_0001
18/04/05 09:44:59 INFO mapreduce.Job: Job job_1522901587164_0001 running in uber mode : false
18/04/05 09:45:00 INFO mapreduce.Job: map 0% reduce 0%
18/04/05 09:45:22 INFO mapreduce.Job: map 100% reduce 0%
18/04/05 09:45:46 INFO mapreduce.Job: map 100% reduce 100%
18/04/05 09:45:48 INFO mapreduce.Job: Job job_1522901587164_0001 completed successfully
18/04/05 09:45:49 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=204
    FILE: Number of bytes written=216347
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=834
    HDFS: Number of bytes written=38
    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

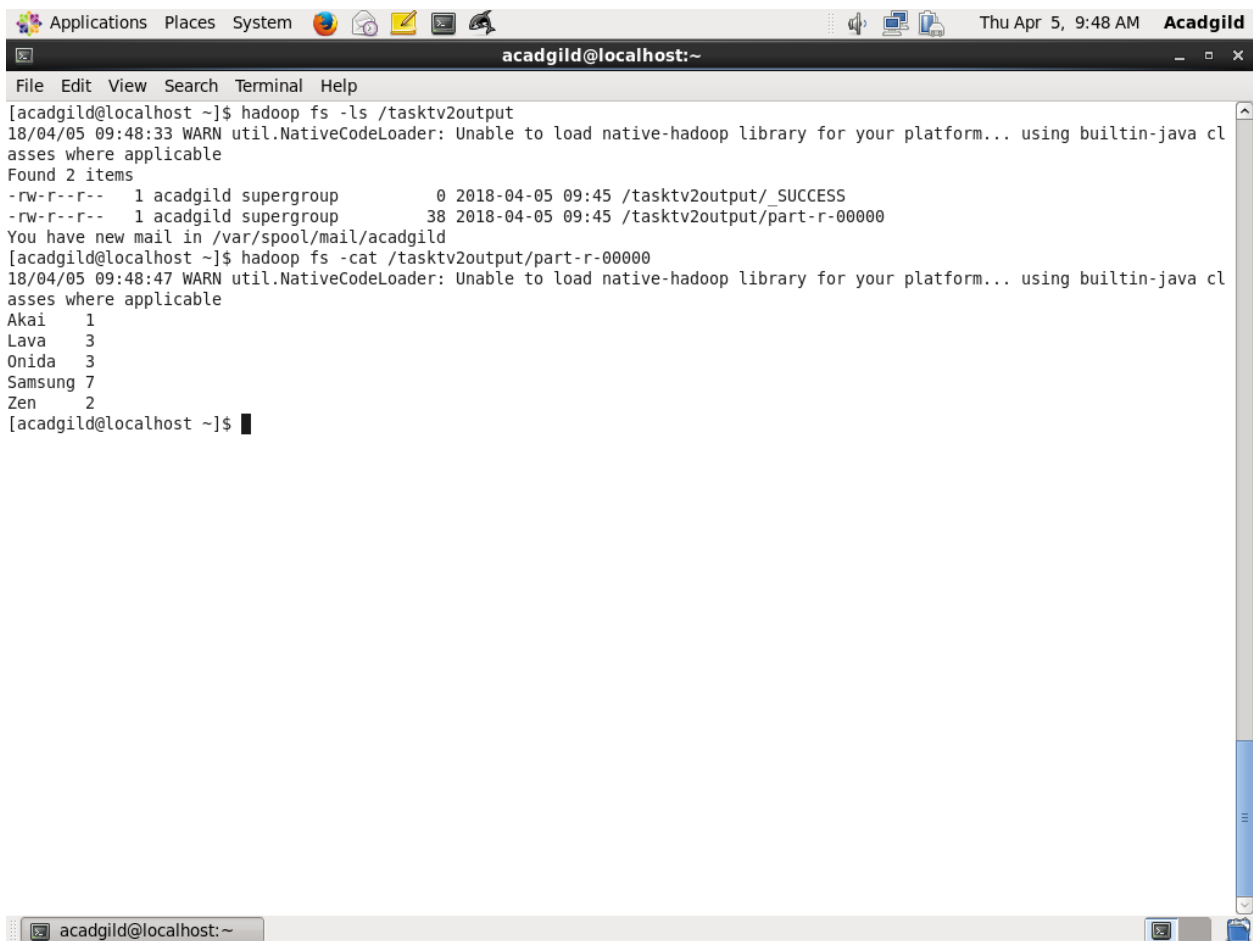
  Total time spent by all map tasks (ms)=19817
  Total time spent by all reduce tasks (ms)=20363
  Total vcore-milliseconds taken by all map tasks=19817
  Total vcore-milliseconds taken by all reduce tasks=20363
  Total megabyte-milliseconds taken by all map tasks=20292608
  Total megabyte-milliseconds taken by all reduce tasks=20851712

  Map-Reduce Framework
    Map input records=18
    Map output records=16
    Map output bytes=166
    Map output materialized bytes=204
    Input split bytes=101
    Combine input records=0
    Combine output records=0
    Reduce input groups=5
    Reduce shuffle bytes=204
    Reduce input records=16
    Reduce output records=5
    Spilled Records=32
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=505
    CPU time spent (ms)=4530
    Physical memory (bytes) snapshot=272134144
    Virtual memory (bytes) snapshot=4118179840
    Total committed heap usage (bytes)=138432512

  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=733
  File Output Format Counters
    Bytes Written=38

You have new mail in /var/spool/mail/acadgild
acadgild@localhost:~
```



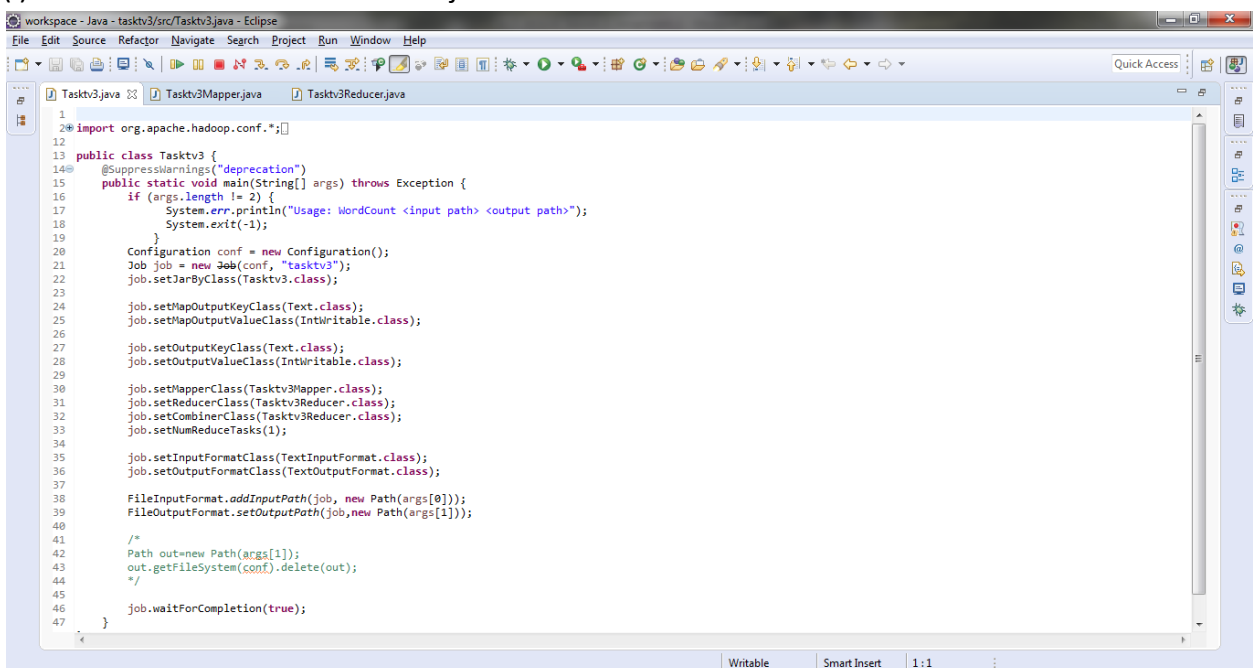
```
acadmild@localhost:~  
File Edit View Search Terminal Help  
[acadmild@localhost ~]$ hadoop fs -ls /tasktv2output  
18/04/05 09:48:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl  
asses where applicable  
Found 2 items  
-rw-r--r-- 1 acadmild supergroup 0 2018-04-05 09:45 /tasktv2output/_SUCCESS  
-rw-r--r-- 1 acadmild supergroup 38 2018-04-05 09:45 /tasktv2output/part-r-00000  
You have new mail in /var/spool/mail/acadmild  
[acadmild@localhost ~]$ hadoop fs -cat /tasktv2output/part-r-00000  
18/04/05 09:48:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl  
asses where applicable  
Akai 1  
Lava 3  
Onida 3  
Samsung 7  
Zen 2  
[acadmild@localhost ~]$
```

Task3:-

COMMAND:-

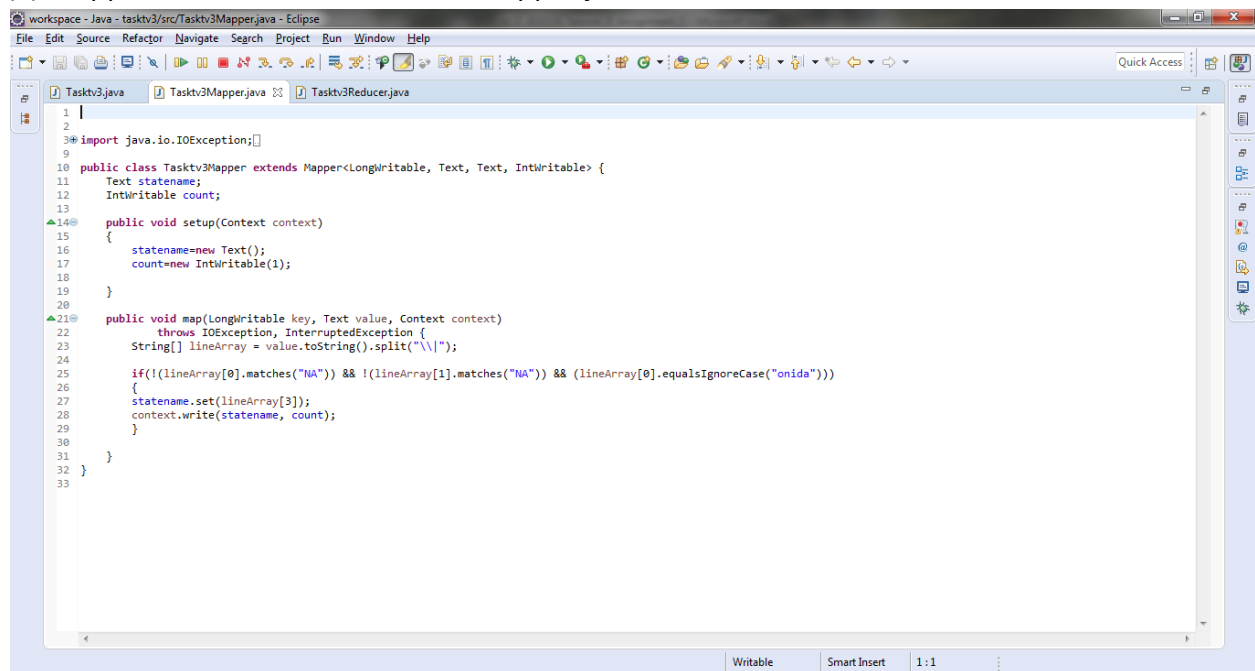
Three files are created,

(i) Driver Code file named Tasktv3.java.



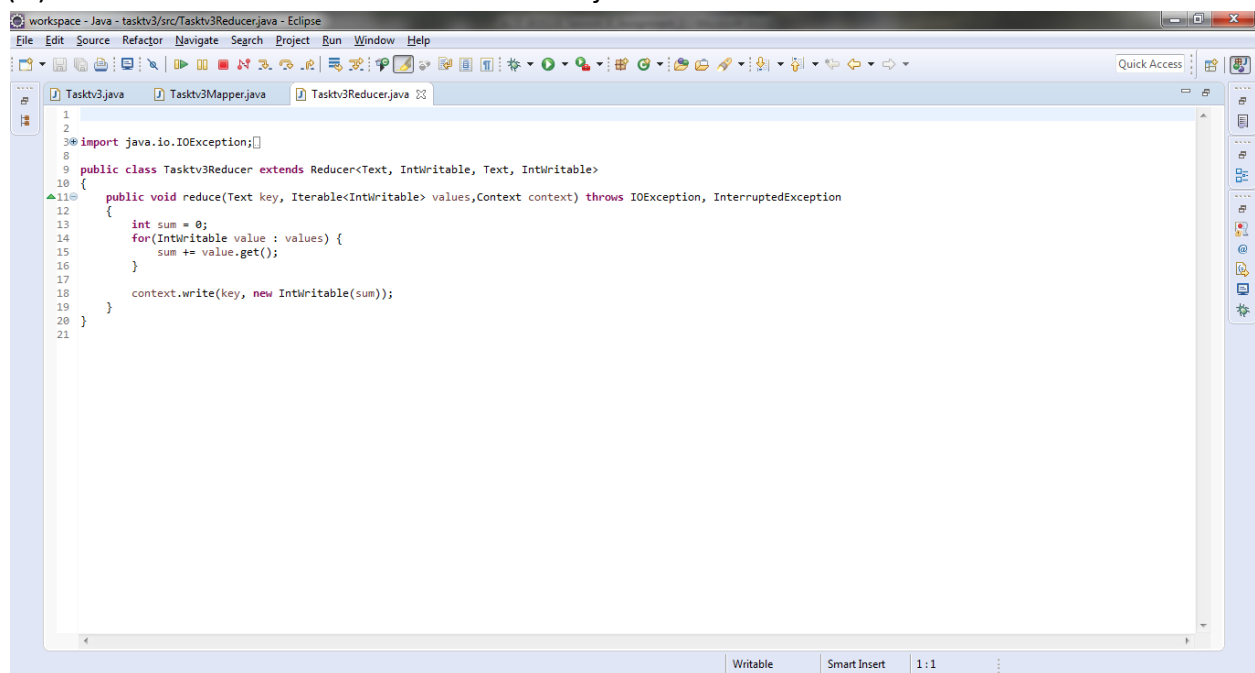
```
workspace - Java - tasktv3/src/Tasktv3.java - Eclipse  
File Edit Source Refactor Navigate Search Project Run Window Help  
Tasktv3.java Tasktv3Mapper.java Tasktv3Reducer.java  
1  
2* import org.apache.hadoop.conf.*;  
3  
4 public class Tasktv3 {  
5     @SuppressWarnings("deprecation")  
6     public static void main(String[] args) throws Exception {  
7         if (args.length != 2) {  
8             System.err.println("Usage: WordCount <input path> <output path>");  
9             System.exit(-1);  
10        }  
11        Configuration conf = new Configuration();  
12        Job job = new Job(conf, "tasktv3");  
13        job.setJarByClass(Tasktv3.class);  
14  
15        job.setMapOutputKeyClass(Text.class);  
16        job.setMapOutputValueClass(IntWritable.class);  
17  
18        job.setOutputKeyClass(Text.class);  
19        job.setOutputValueClass(IntWritable.class);  
20  
21        job.setMapperClass(Tasktv3Mapper.class);  
22        job.setReducerClass(Tasktv3Reducer.class);  
23        job.setCombinerClass(Tasktv3Reducer.class);  
24        job.setNumReduceTasks(1);  
25  
26        job.setInputFormatClass(TextInputFormat.class);  
27        job.setOutputFormatClass(TextOutputFormat.class);  
28  
29        FileInputFormat.addInputPath(job, new Path(args[0]));  
30        FileOutputFormat.setOutputPath(job, new Path(args[1]));  
31  
32        /*  
33        Path out=new Path(args[1]);  
34        out.getFileSystem(conf).delete(out);  
35        */  
36  
37        job.waitForCompletion(true);  
38    }  
39 }
```

(ii) Mapper Class file named Tasktv3Mapper.java.



```
1 |
2 |
3 | import java.io.IOException;
4 |
5 |
6 |
7 |
8 |
9 | public class Tasktv3Mapper extends Mapper<LongWritable, Text, Text, IntWritable> {
10 |     Text statename;
11 |     IntWritable count;
12 |
13 |
14 |     public void setup(Context context)
15 |     {
16 |         statename=new Text();
17 |         count=new IntWritable(1);
18 |
19 |     }
20 |
21 |     public void map(LongWritable key, Text value, Context context)
22 |     throws IOException, InterruptedException {
23 |         String[] lineArray = value.toString().split("\\\\");
24 |
25 |         if(!(lineArray[0].matches("NA")) && !(lineArray[1].matches("NA")) && (lineArray[0].equalsIgnoreCase("onida")))
26 |         {
27 |             statename.set(lineArray[3]);
28 |             context.write(statename, count);
29 |         }
30 |     }
31 | }
32 |
33 |
```

(iii)Reducer Class file named Tasktv3Reducer.java.



```
1 |
2 |
3 | import java.io.IOException;
4 |
5 |
6 |
7 |
8 | public class Tasktv3Reducer extends Reducer<Text, IntWritable, Text, IntWritable>
9 | {
10 |     public void reduce(Text key, Iterable<IntWritable> values,Context context) throws IOException, InterruptedException
11 |     {
12 |         int sum = 0;
13 |         for(IntWritable value : values) {
14 |             sum += value.get();
15 |         }
16 |
17 |         context.write(key, new IntWritable(sum));
18 |     }
19 | }
20 |
21 |
```

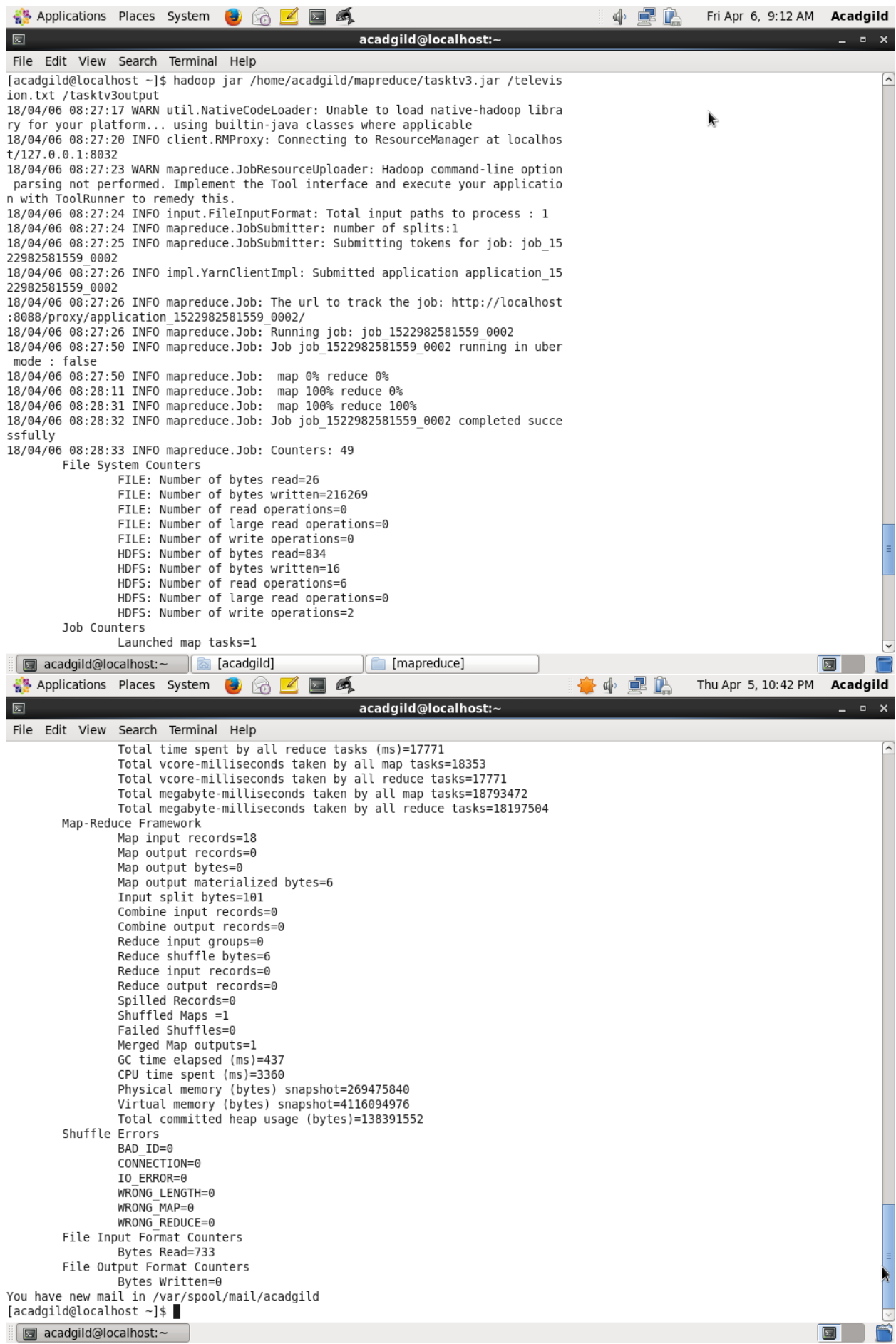
EXPLANATION:-

television.txt file is used here.It is attached to the git.

tasktv3.jar file is used here.It is attached to the git.

Total unit sold for each state for Onida Comapny, and "NA" records will be deleted here.

OUTPUT:-



```
[acadgild@localhost ~]$ hadoop jar /home/acadgild/mapreduce/tasktv3.jar /televis
ion.txt /tasktv3output
18/04/06 08:27:17 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
18/04/06 08:27:20 INFO client.RMPProxy: Connecting to ResourceManager at localhos
t/127.0.0.1:8032
18/04/06 08:27:23 WARN mapreduce.JobResourceUploader: Hadoop command-line option
 parsing not performed. Implement the Tool interface and execute your applicatio
n with ToolRunner to remedy this.
18/04/06 08:27:24 INFO input.FileInputFormat: Total input paths to process : 1
18/04/06 08:27:24 INFO mapreduce.JobSubmitter: number of splits:1
18/04/06 08:27:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
22982581559_0002
18/04/06 08:27:26 INFO impl.YarnClientImpl: Submitted application application_15
22982581559_0002
18/04/06 08:27:26 INFO mapreduce.Job: The url to track the job: http://localhost
:8088/proxy/application_1522982581559_0002/
18/04/06 08:27:26 INFO mapreduce.Job: Running job: job_1522982581559_0002
18/04/06 08:27:50 INFO mapreduce.Job: Job job_1522982581559_0002 running in uber
mode : false
18/04/06 08:27:50 INFO mapreduce.Job: map 0% reduce 0%
18/04/06 08:28:11 INFO mapreduce.Job: map 100% reduce 0%
18/04/06 08:28:31 INFO mapreduce.Job: map 100% reduce 100%
18/04/06 08:28:32 INFO mapreduce.Job: Job job_1522982581559_0002 completed succe
ssfully
18/04/06 08:28:33 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=26
    FILE: Number of bytes written=216269
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=834
    HDFS: Number of bytes written=16
    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
    Total time spent by all reduce tasks (ms)=17771
    Total vcore-milliseconds taken by all map tasks=18353
    Total vcore-milliseconds taken by all reduce tasks=17771
    Total megabyte-milliseconds taken by all map tasks=18793472
    Total megabyte-milliseconds taken by all reduce tasks=18197504
  Map-Reduce Framework
    Map input records=18
    Map output records=0
    Map output bytes=0
    Map output materialized bytes=6
    Input split bytes=101
    Combine input records=0
    Combine output records=0
    Reduce input groups=0
    Reduce shuffle bytes=6
    Reduce input records=0
    Reduce output records=0
    Spilled Records=0
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=437
    CPU time spent (ms)=3360
    Physical memory (bytes) snapshot=269475840
    Virtual memory (bytes) snapshot=4116094976
    Total committed heap usage (bytes)=138391552
  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=733
  File Output Format Counters
    Bytes Written=0
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

