Assignment 4:

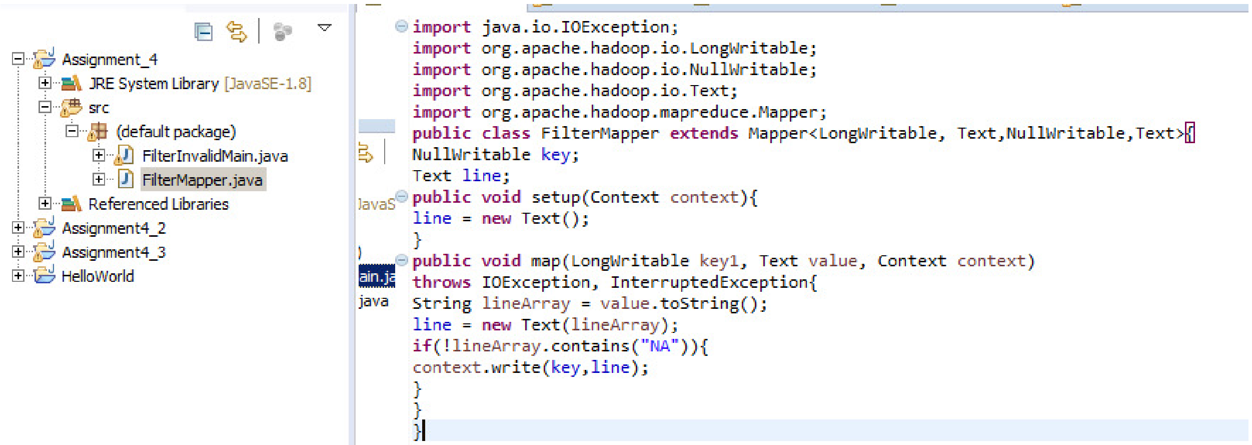
Submitted by Jyoti Jha
DOS: 25th July'18

Task 1:

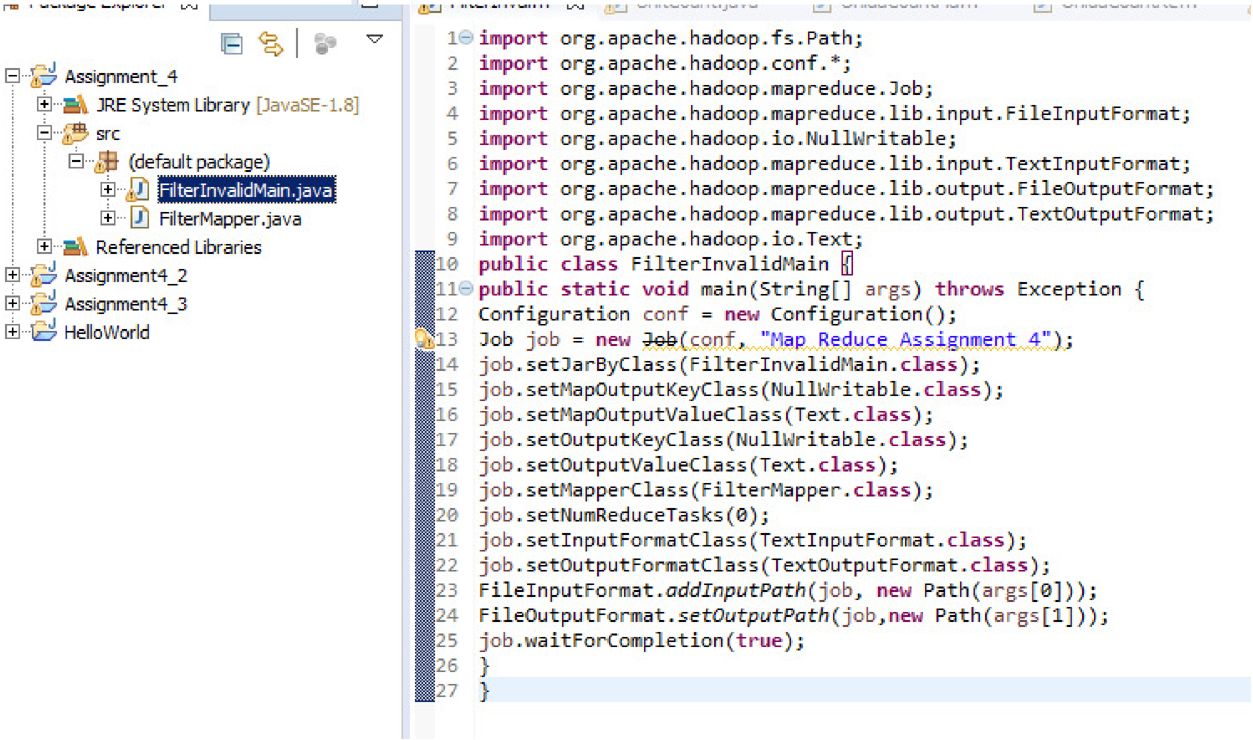
Map Reduce program to filter out Invalid Records; i.e.; records which contain 'NA' in Company or Product name.

Solution:

Code:

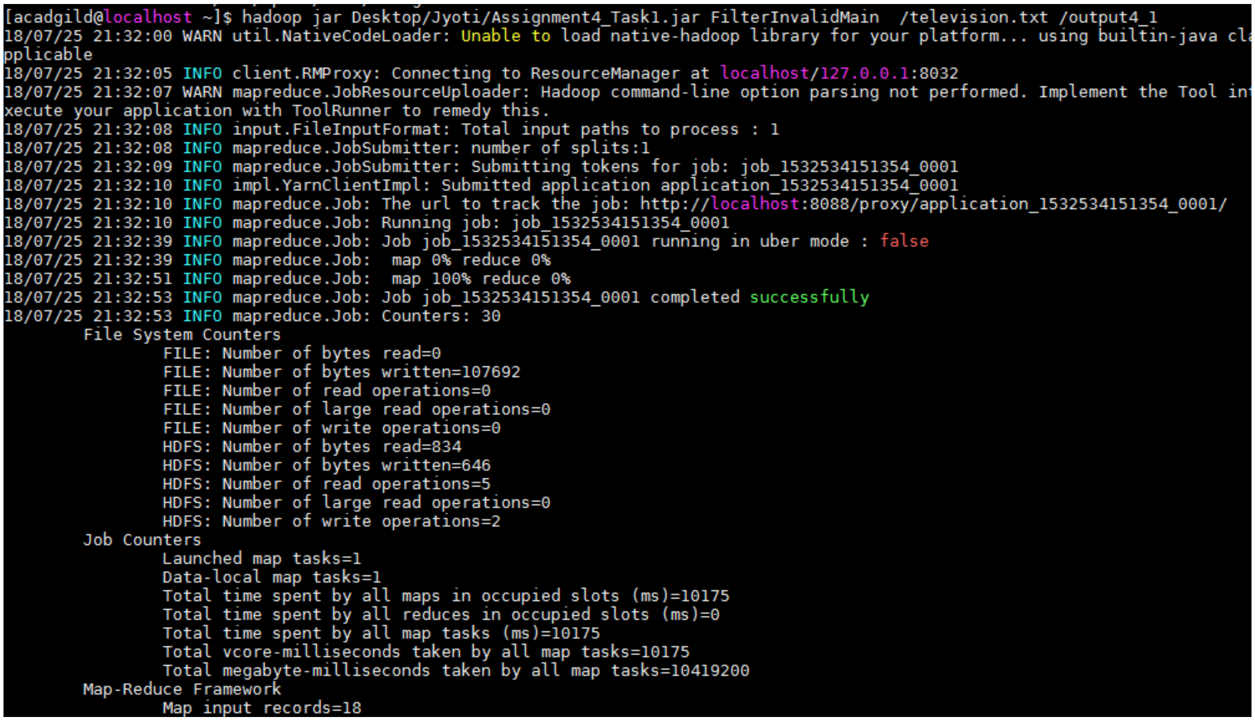


```java
import java.io.IOException;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
public class FilterMapper extends Mapper<LongWritable, Text,NullWritable,Text>{
NullWritable key;
Text line;
public void setup(Context context){
line = new Text();
}
public void map(LongWritable key1, Text value, Context context)
throws IOException, InterruptedException{
String lineArray = value.toString();
line = new Text(lineArray);
if(!lineArray.contains("NA")){
context.write(key,line);
}
}
}
```

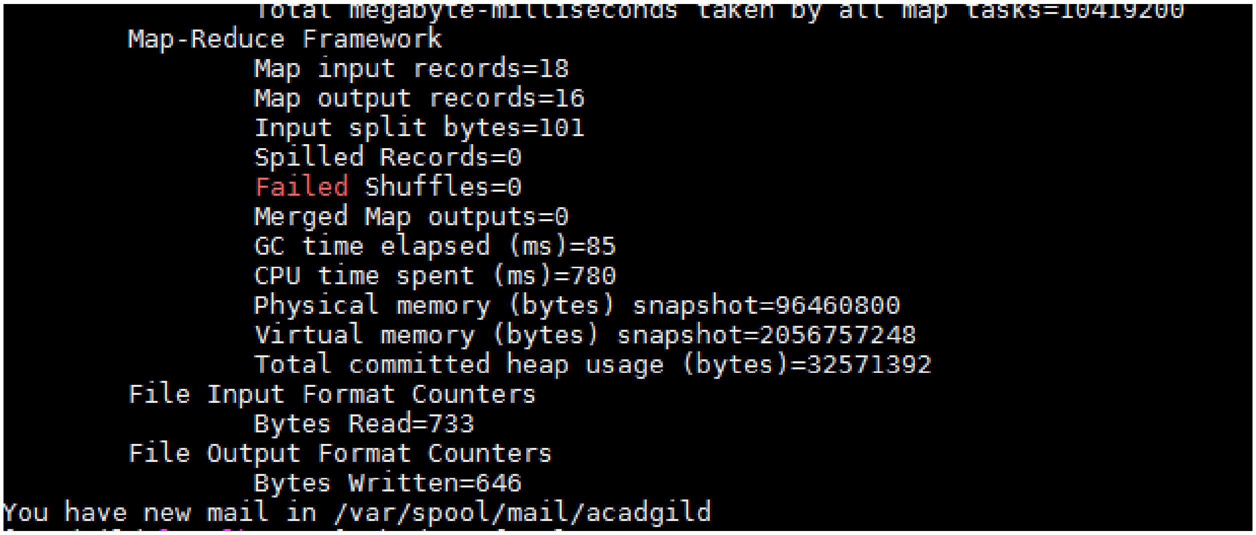


```java
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.io.Text;
public class FilterInvalidMain {
public static void main(String[] args) throws Exception {
Configuration conf = new Configuration();
Job job = new Job(conf, "Map Reduce Assignment 4");
job.setJarByClass(FilterInvalidMain.class);
job.setMapOutputKeyClass(NullWritable.class);
job.setMapOutputValueClass(Text.class);
job.setOutputKeyClass(NullWritable.class);
job.setOutputValueClass(Text.class);
job.setMapperClass(FilterMapper.class);
job.setNumReduceTasks(0);
job.setInputFormatClass(TextInputFormat.class);
job.setOutputFormatClass(TextOutputFormat.class);
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job,new Path(args[1]));
job.waitForCompletion(true);
}
}
```

Execution of Jar File:



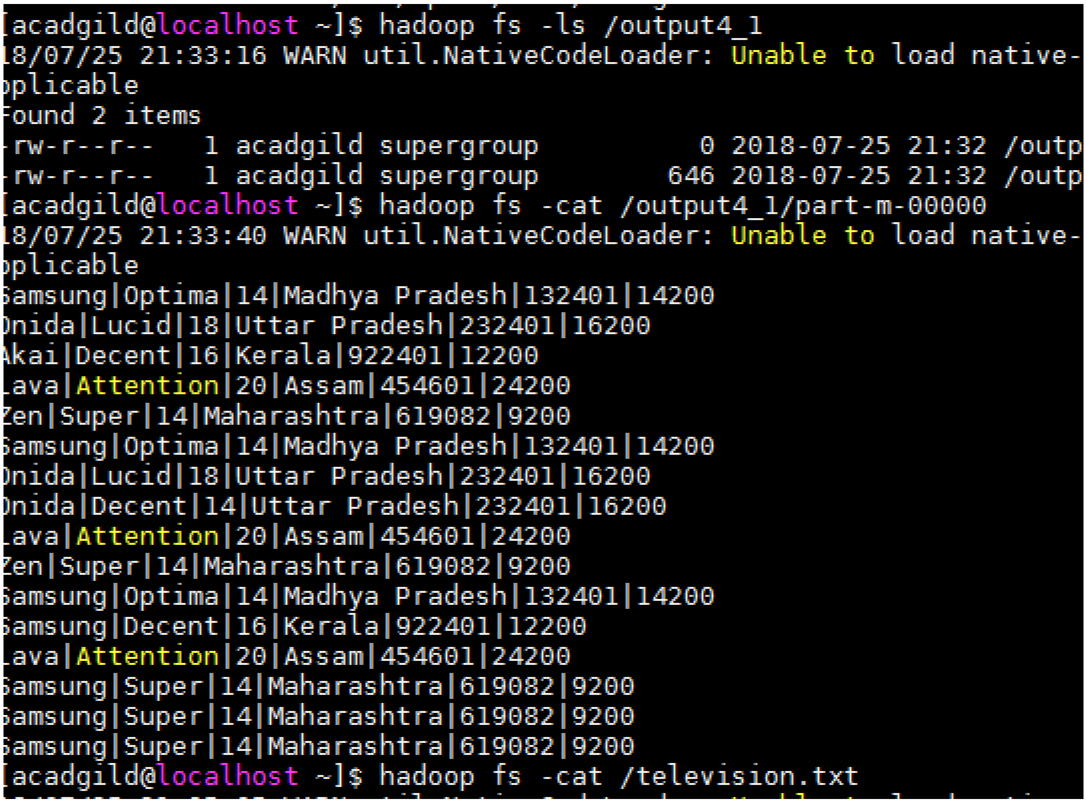
```
[acadgild@localhost ~]$ hadoop jar Desktop/Jyoti/Assignment4_Task1.jar FilterInvalidMain  /television.txt /output4_1
18/07/25 21:32:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cla
pplicable
18/07/25 21:32:05 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/07/25 21:32:07 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in
xecute your application with ToolRunner to remedy this.
18/07/25 21:32:08 INFO input.FileInputFormat: Total input paths to process : 1
18/07/25 21:32:08 INFO mapreduce.JobSubmitter: number of splits:1
18/07/25 21:32:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1532534151354_0001
18/07/25 21:32:10 INFO impl.YarnClientImpl: Submitted application application_1532534151354_0001
18/07/25 21:32:10 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1532534151354_0001/
18/07/25 21:32:10 INFO mapreduce.Job: Running job: job_1532534151354_0001
18/07/25 21:32:39 INFO mapreduce.Job: Job job_1532534151354_0001 running in uber mode : false
18/07/25 21:32:39 INFO mapreduce.Job:  map 0% reduce 0%
18/07/25 21:32:51 INFO mapreduce.Job:  map 100% reduce 0%
18/07/25 21:32:53 INFO mapreduce.Job: Job job_1532534151354_0001 completed successfully
18/07/25 21:32:53 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=107692
                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=646
                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)=10175
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=10175
                Total vcore-milliseconds taken by all map tasks=10175
                Total megabyte-milliseconds taken by all map tasks=10419200
        Map-Reduce Framework
                Map input records=18
```



```
                Total megabyte-milliseconds taken by all map tasks=10419200
        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)=85
                CPU time spent (ms)=780
                Physical memory (bytes) snapshot=96460800
                Virtual memory (bytes) snapshot=2056757248
                Total committed heap usage (bytes)=32571392
        File Input Format Counters
                Bytes Read=733
        File Output Format Counters
                Bytes Written=646
You have new mail in /var/spool/mail/acadgild
```

Output1:

Output doesn't have records with Product/Company name as 'NA".



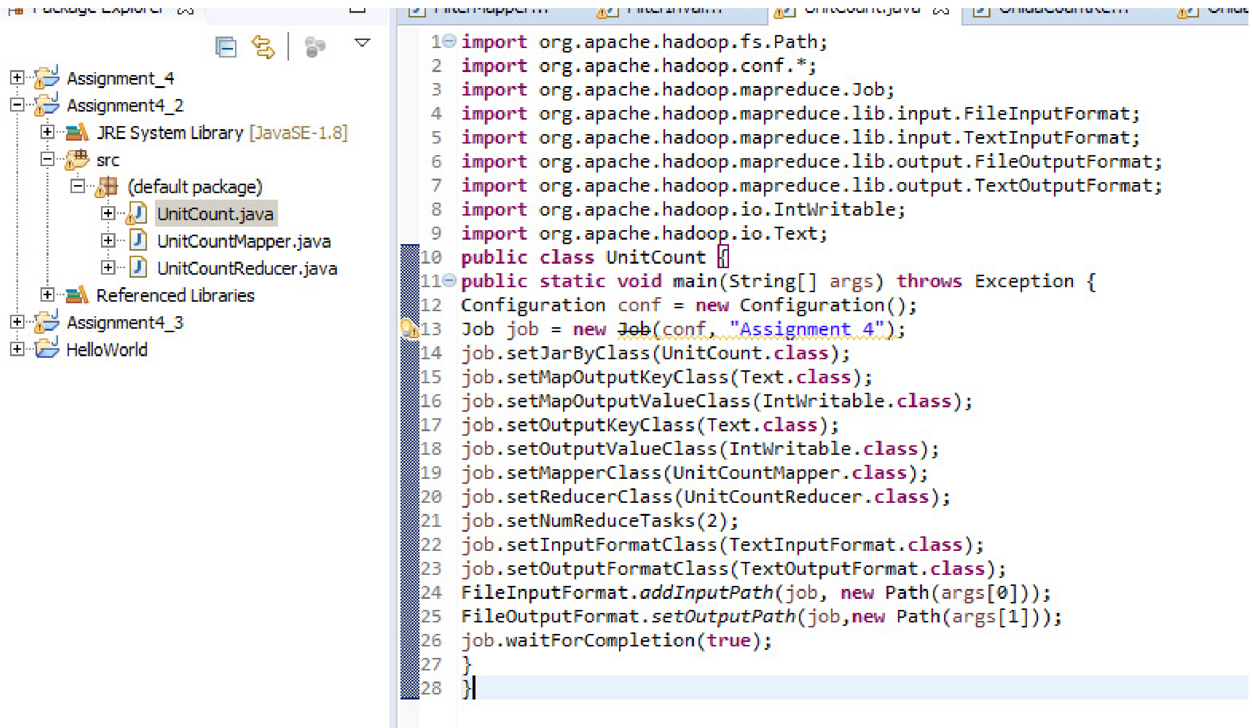
```
[acadgild@localhost ~]$ hadoop fs -ls /output4_1
18/07/25 21:33:16 WARN util.NativeCodeLoader: Unable to load native-
pplicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-07-25 21:32 /outp
-rw-r--r--   1 acadgild supergroup        646 2018-07-25 21:32 /outp
[acadgild@localhost ~]$ hadoop fs -cat /output4_1/part-m-00000
18/07/25 21:33:40 WARN util.NativeCodeLoader: Unable to load native-
pplicable
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Akai|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Onida|Decent|14|Uttar Pradesh|232401|16200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Samsung|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
[acadgild@localhost ~]$ hadoop fs -cat /television.txt
```

Task 2:

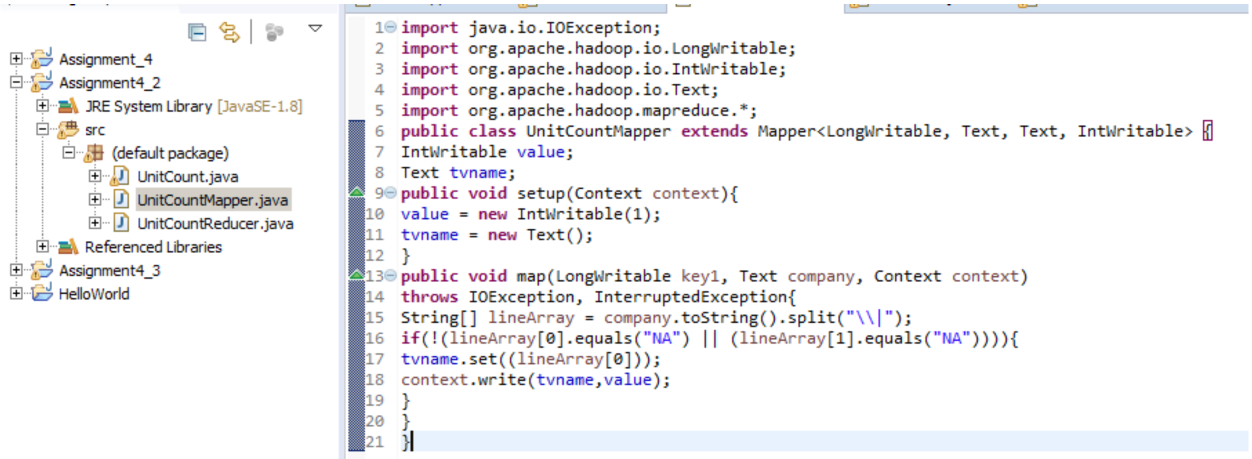
Map Reduce program to calculate total units sold for each Company.

Solution:

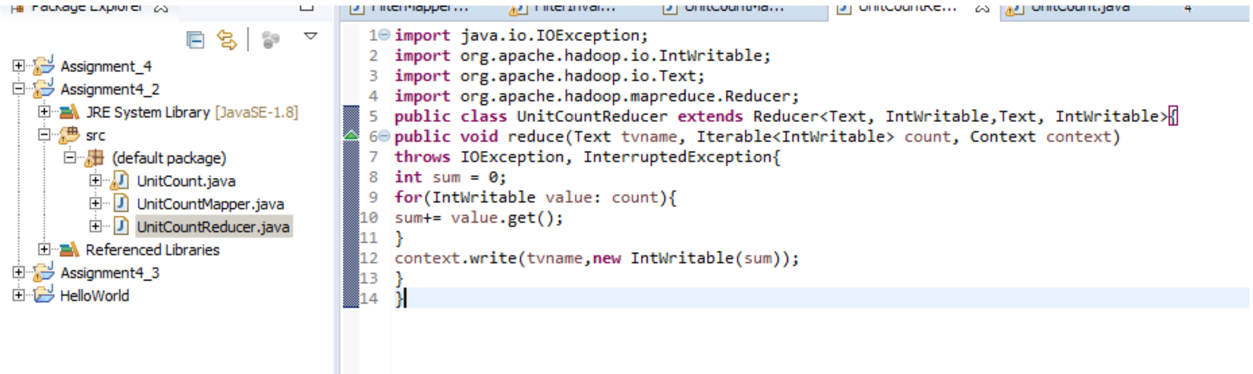
Code:



```
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
public class UnitCount {
public static void main(String[] args) throws Exception {
Configuration conf = new Configuration();
Job job = new Job(conf, "Assignment 4");
job.setJarByClass(UnitCount.class);
job.setMapOutputKeyClass(Text.class);
job.setMapOutputValueClass(IntWritable.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
job.setMapperClass(UnitCountMapper.class);
job.setReducerClass(UnitCountReducer.class);
job.setNumReduceTasks(2);
job.setInputFormatClass(TextInputFormat.class);
job.setOutputFormatClass(TextOutputFormat.class);
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job,new Path(args[1]));
job.waitForCompletion(true);
}
}
```

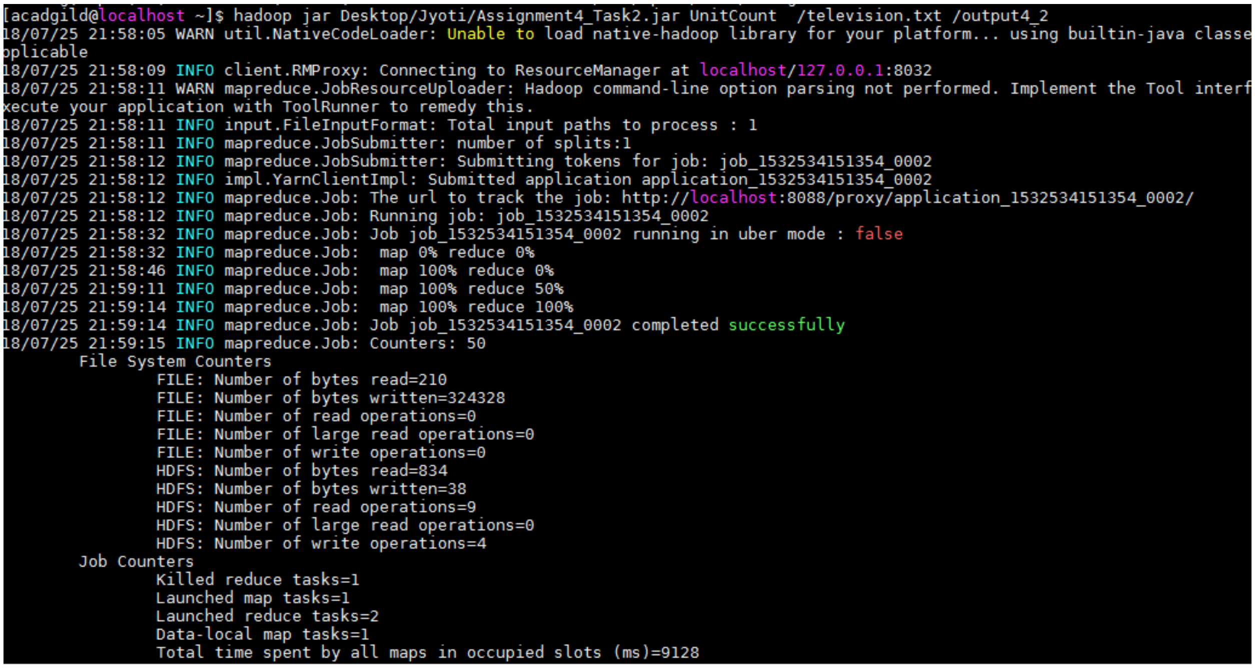


```
import java.io.IOException;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.*;
public class UnitCountMapper extends Mapper<LongWritable, Text, Text, IntWritable> {
IntWritable value;
Text tvname;
public void setup(Context context){
value = new IntWritable(1);
tvname = new Text();
}
public void map(LongWritable key1, Text company, Context context)
throws IOException, InterruptedException{
String[] lineArray = company.toString().split("\\|");
if(!(lineArray[0].equals("NA") || (lineArray[1].equals("NA")))){
tvname.set((lineArray[0]));
context.write(tvname,value);
}
}
}
```

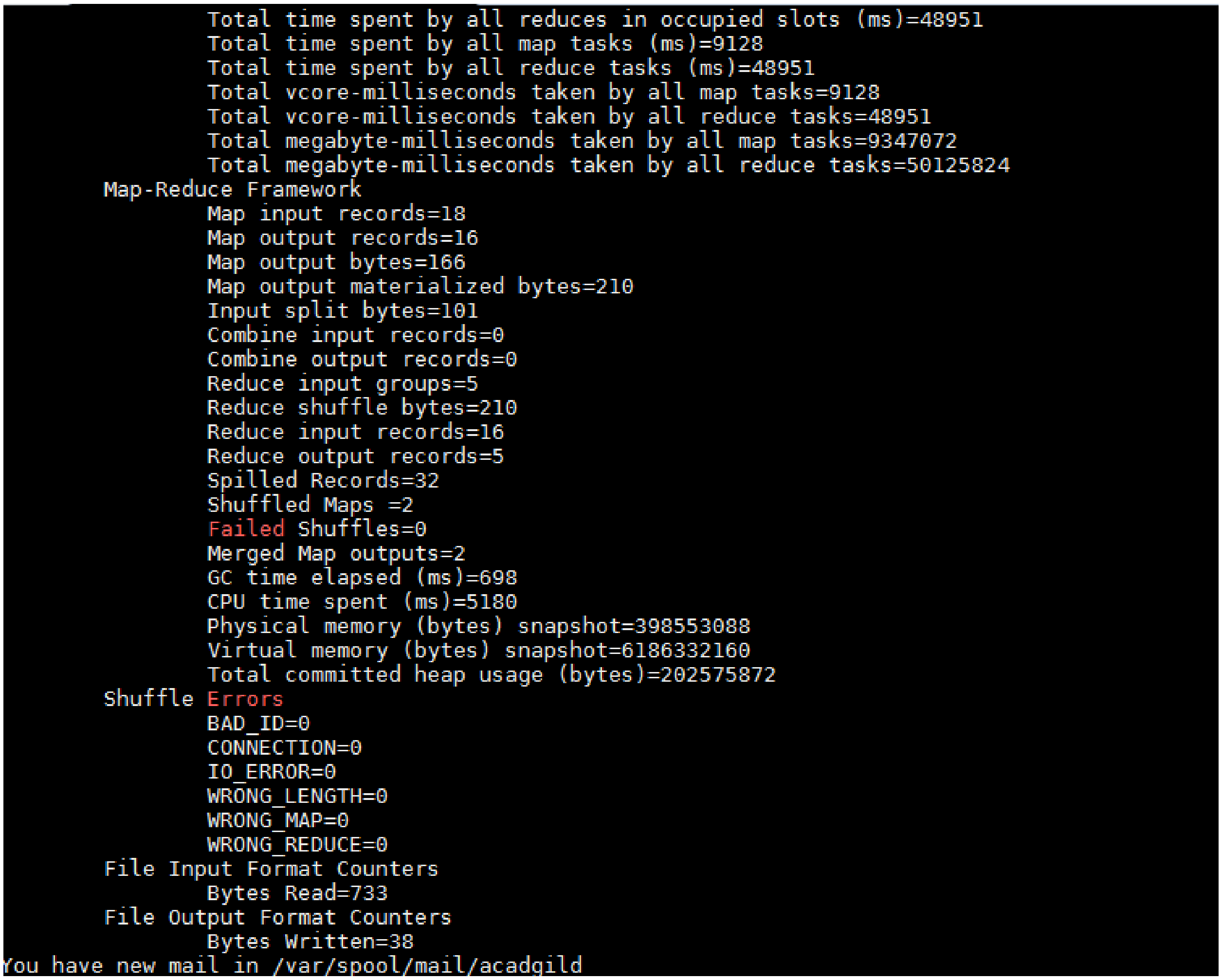


```
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
public class UnitCountReducer extends Reducer<Text, IntWritable,Text, IntWritable>{
public void reduce(Text tvname, Iterable<IntWritable> count, Context context)
throws IOException, InterruptedException{
int sum = 0;
for(IntWritable value: count){
sum+= value.get();
}
context.write(tvname,new IntWritable(sum));
}
}
```

Execution of Jar File:

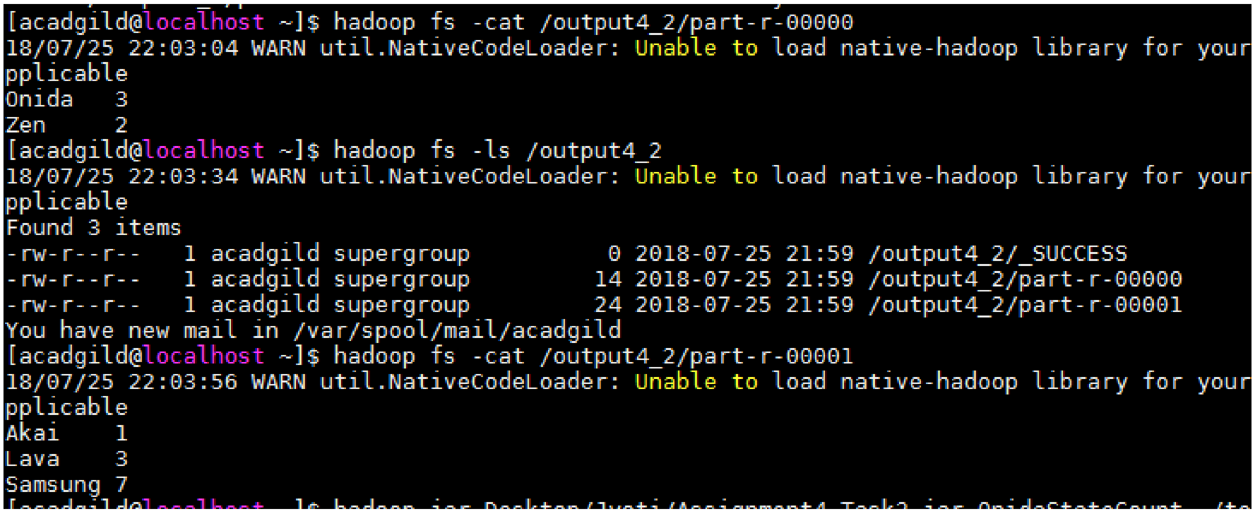


```
[acadgild@localhost ~]$ hadoop jar Desktop/Jyoti/Assignment4_Task2.jar UnitCount  /television.txt /output4_2
18/07/25 21:58:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
oplicable
18/07/25 21:58:09 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/07/25 21:58:11 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interf
xecute your application with ToolRunner to remedy this.
18/07/25 21:58:11 INFO input.FileInputFormat: Total input paths to process : 1
18/07/25 21:58:11 INFO mapreduce.JobSubmitter: number of splits:1
18/07/25 21:58:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1532534151354_0002
18/07/25 21:58:12 INFO impl.YarnClientImpl: Submitted application application_1532534151354_0002
18/07/25 21:58:12 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1532534151354_0002/
18/07/25 21:58:12 INFO mapreduce.Job: Running job: job_1532534151354_0002
18/07/25 21:58:32 INFO mapreduce.Job: Job job_1532534151354_0002 running in uber mode : false
18/07/25 21:58:32 INFO mapreduce.Job:  map 0% reduce 0%
18/07/25 21:58:46 INFO mapreduce.Job:  map 100% reduce 0%
18/07/25 21:59:11 INFO mapreduce.Job:  map 100% reduce 50%
18/07/25 21:59:14 INFO mapreduce.Job:  map 100% reduce 100%
18/07/25 21:59:14 INFO mapreduce.Job: Job job_1532534151354_0002 completed successfully
18/07/25 21:59:15 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=210
                FILE: Number of bytes written=324328
                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=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Killed reduce tasks=1
                Launched map tasks=1
                Launched reduce tasks=2
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=9128
```



```
                Total time spent by all reduces in occupied slots (ms)=48951
                Total time spent by all map tasks (ms)=9128
                Total time spent by all reduce tasks (ms)=48951
                Total vcore-milliseconds taken by all map tasks=9128
                Total vcore-milliseconds taken by all reduce tasks=48951
                Total megabyte-milliseconds taken by all map tasks=9347072
                Total megabyte-milliseconds taken by all reduce tasks=50125824
        Map-Reduce Framework
                Map input records=18
                Map output records=16
                Map output bytes=166
                Map output materialized bytes=210
                Input split bytes=101
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=210
                Reduce input records=16
                Reduce output records=5
                Spilled Records=32
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=698
                CPU time spent (ms)=5180
                Physical memory (bytes) snapshot=398553088
                Virtual memory (bytes) snapshot=6186332160
                Total committed heap usage (bytes)=202575872
        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
```

Output2:



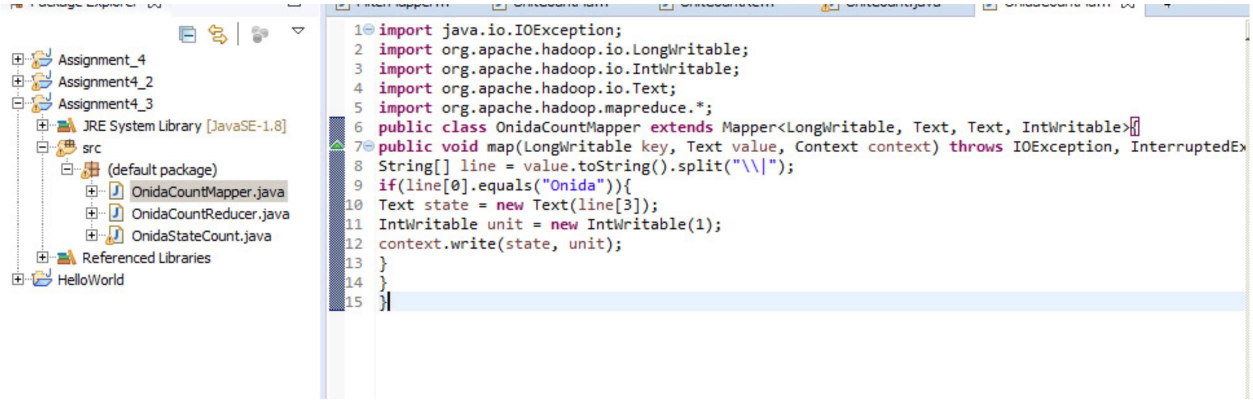
```
[acadgild@localhost ~]$ hadoop fs -cat /output4_2/part-r-00000
18/07/25 22:03:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
pplicable
Onida   3
Zen     2
[acadgild@localhost ~]$ hadoop fs -ls /output4_2
18/07/25 22:03:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
pplicable
Found 3 items
-rw-r--r--   1 acadgild supergroup          0 2018-07-25 21:59 /output4_2/_SUCCESS
-rw-r--r--   1 acadgild supergroup         14 2018-07-25 21:59 /output4_2/part-r-00000
-rw-r--r--   1 acadgild supergroup         24 2018-07-25 21:59 /output4_2/part-r-00001
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -cat /output4_2/part-r-00001
18/07/25 22:03:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
pplicable
Akai    1
Lava    3
Samsung 7
```

Task 3:

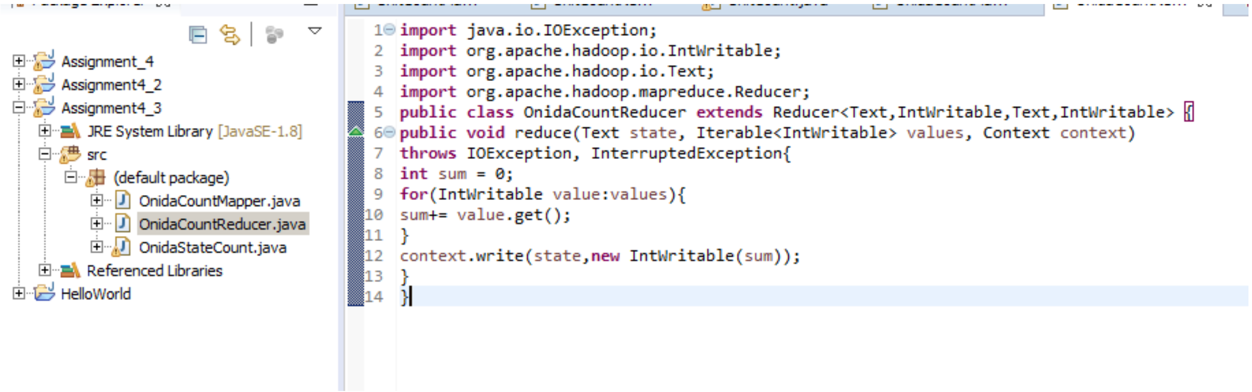
Map Reduce program to calculate total units sold in each state for Onida Company.

Solution:

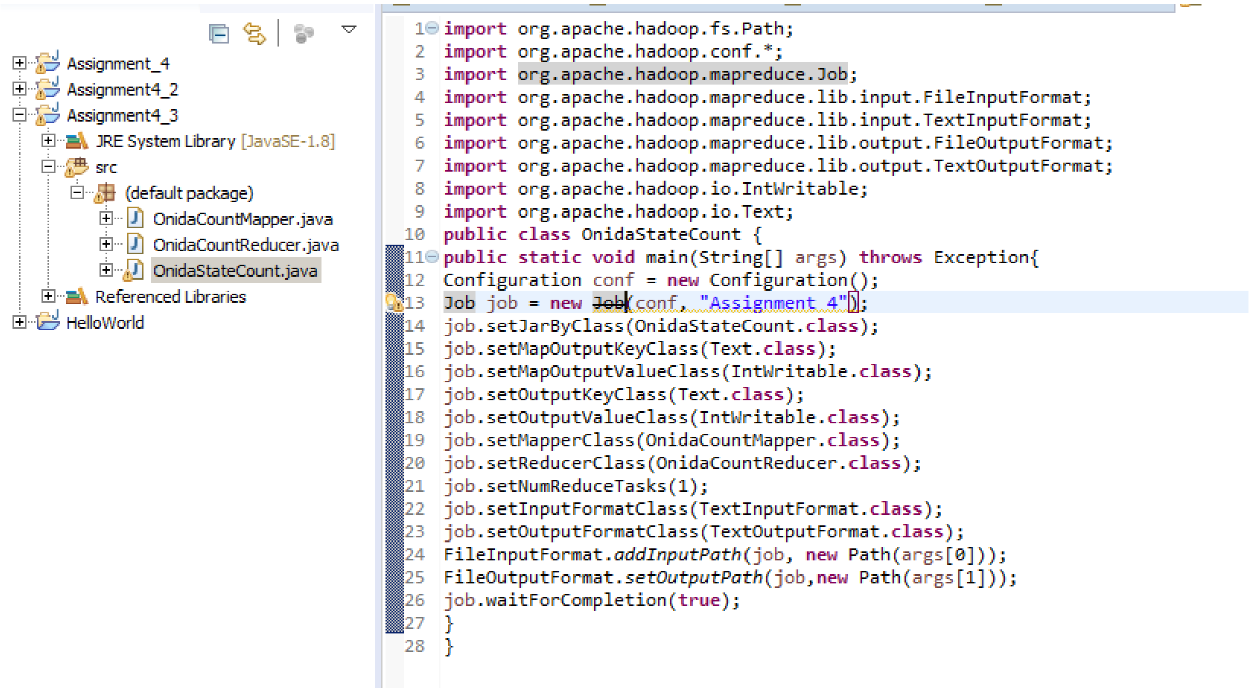
Code:



```java
import java.io.IOException;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.*;
public class OnidaCountMapper extends Mapper<LongWritable, Text, Text, IntWritable>{
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedEx
String[] line = value.toString().split("\\|");
if(line[0].equals("Onida")){
Text state = new Text(line[3]);
IntWritable unit = new IntWritable(1);
context.write(state, unit);
}
}
}
```

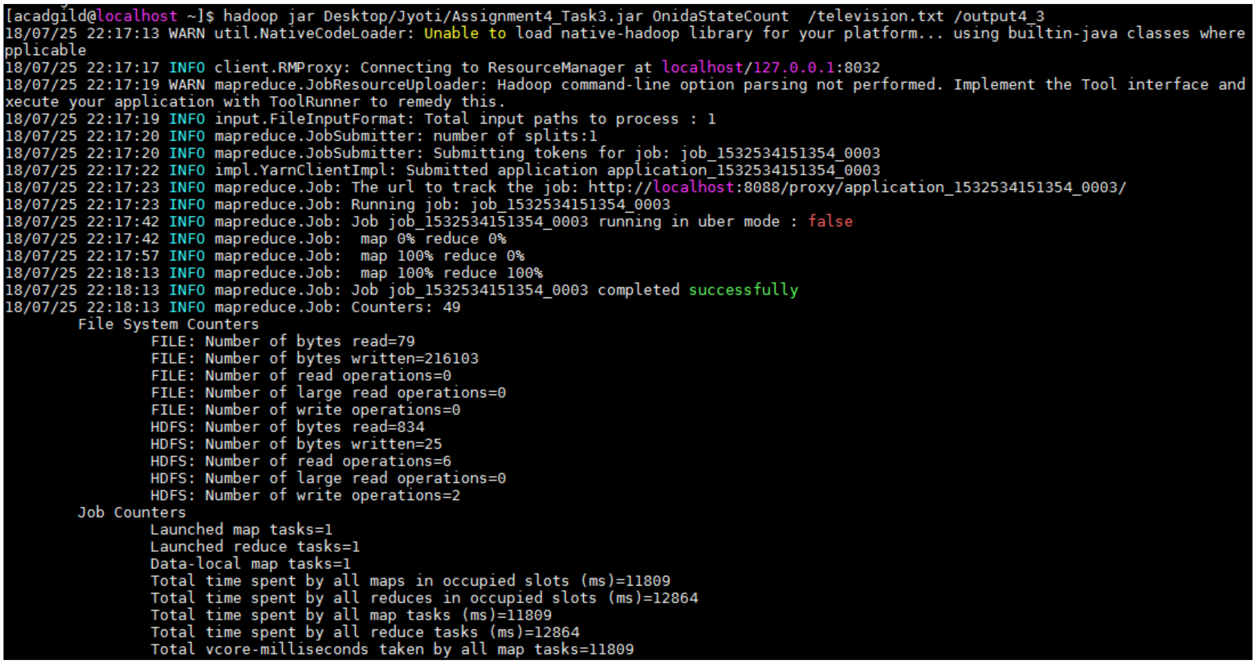


```java
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
public class OnidaCountReducer extends Reducer<Text,IntWritable,Text,IntWritable> {
public void reduce(Text state, Iterable<IntWritable> values, Context context)
throws IOException, InterruptedException{
int sum = 0;
for(IntWritable value:values){
sum+= value.get();
}
context.write(state,new IntWritable(sum));
}
}
```

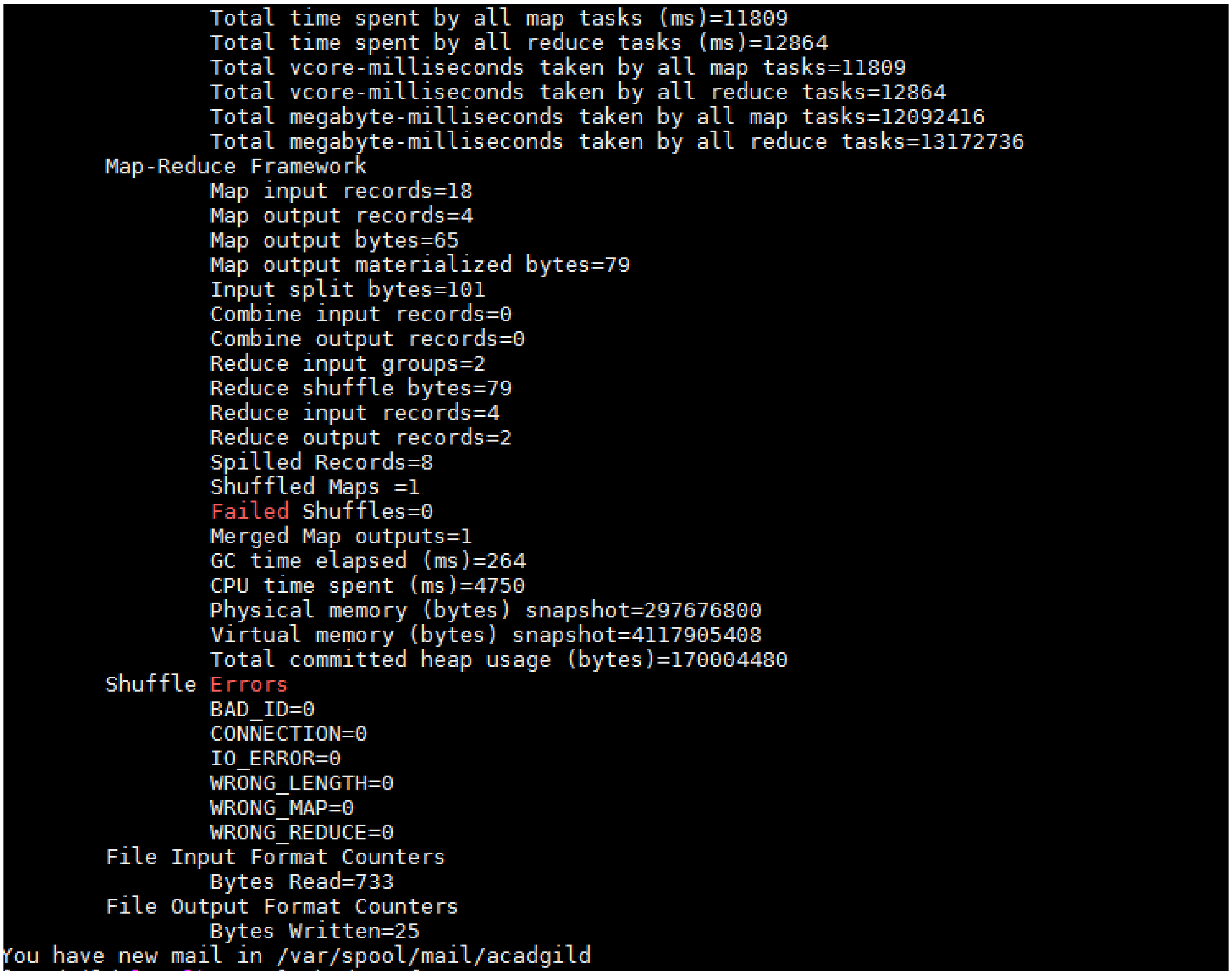


```java
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
public class OnidaStateCount {
public static void main(String[] args) throws Exception{
Configuration conf = new Configuration();
Job job = new Job(conf, "Assignment 4");
job.setJarByClass(OnidaStateCount.class);
job.setMapOutputKeyClass(Text.class);
job.setMapOutputValueClass(IntWritable.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
job.setMapperClass(OnidaCountMapper.class);
job.setReducerClass(OnidaCountReducer.class);
job.setNumReduceTasks(1);
job.setInputFormatClass(TextInputFormat.class);
job.setOutputFormatClass(TextOutputFormat.class);
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job,new Path(args[1]));
job.waitForCompletion(true);
}
}
```

Execution of Jar File:

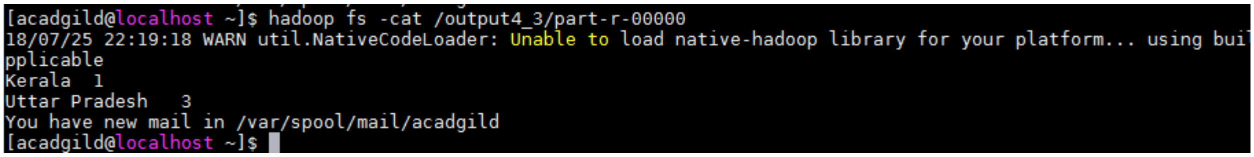


```
[acadgild@localhost ~]$ hadoop jar Desktop/Jyoti/Assignment4_Task3.jar OnidaStateCount  /television.txt /output4_3
18/07/25 22:17:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
pplicable
18/07/25 22:17:17 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/07/25 22:17:19 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and
xecute your application with ToolRunner to remedy this.
18/07/25 22:17:19 INFO input.FileInputFormat: Total input paths to process : 1
18/07/25 22:17:20 INFO mapreduce.JobSubmitter: number of splits:1
18/07/25 22:17:20 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1532534151354_0003
18/07/25 22:17:22 INFO impl.YarnClientImpl: Submitted application application_1532534151354_0003
18/07/25 22:17:23 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1532534151354_0003/
18/07/25 22:17:23 INFO mapreduce.Job: Running job: job_1532534151354_0003
18/07/25 22:17:42 INFO mapreduce.Job: Job job_1532534151354_0003 running in uber mode : false
18/07/25 22:17:42 INFO mapreduce.Job:  map 0% reduce 0%
18/07/25 22:17:57 INFO mapreduce.Job:  map 100% reduce 0%
18/07/25 22:18:13 INFO mapreduce.Job:  map 100% reduce 100%
18/07/25 22:18:13 INFO mapreduce.Job: Job job_1532534151354_0003 completed successfully
18/07/25 22:18:13 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=79
                FILE: Number of bytes written=216103
                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=25
                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)=11809
                Total time spent by all reduces in occupied slots (ms)=12864
                Total time spent by all map tasks (ms)=11809
                Total time spent by all reduce tasks (ms)=12864
                Total vcore-milliseconds taken by all map tasks=11809
```



```
                Total time spent by all map tasks (ms)=11809
                Total time spent by all reduce tasks (ms)=12864
                Total vcore-milliseconds taken by all map tasks=11809
                Total vcore-milliseconds taken by all reduce tasks=12864
                Total megabyte-milliseconds taken by all map tasks=12092416
                Total megabyte-milliseconds taken by all reduce tasks=13172736
        Map-Reduce Framework
                Map input records=18
                Map output records=4
                Map output bytes=65
                Map output materialized bytes=79
                Input split bytes=101
                Combine input records=0
                Combine output records=0
                Reduce input groups=2
                Reduce shuffle bytes=79
                Reduce input records=4
                Reduce output records=2
                Spilled Records=8
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=264
                CPU time spent (ms)=4750
                Physical memory (bytes) snapshot=297676800
                Virtual memory (bytes) snapshot=4117905408
                Total committed heap usage (bytes)=170004480
        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=25
You have new mail in /var/spool/mail/acadgild
```

Output3:



```
[acadgild@localhost ~]$ hadoop fs -cat /output4_3/part-r-00000
18/07/25 22:19:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using bui
pplicable
Kerala  1
Uttar Pradesh   3
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```