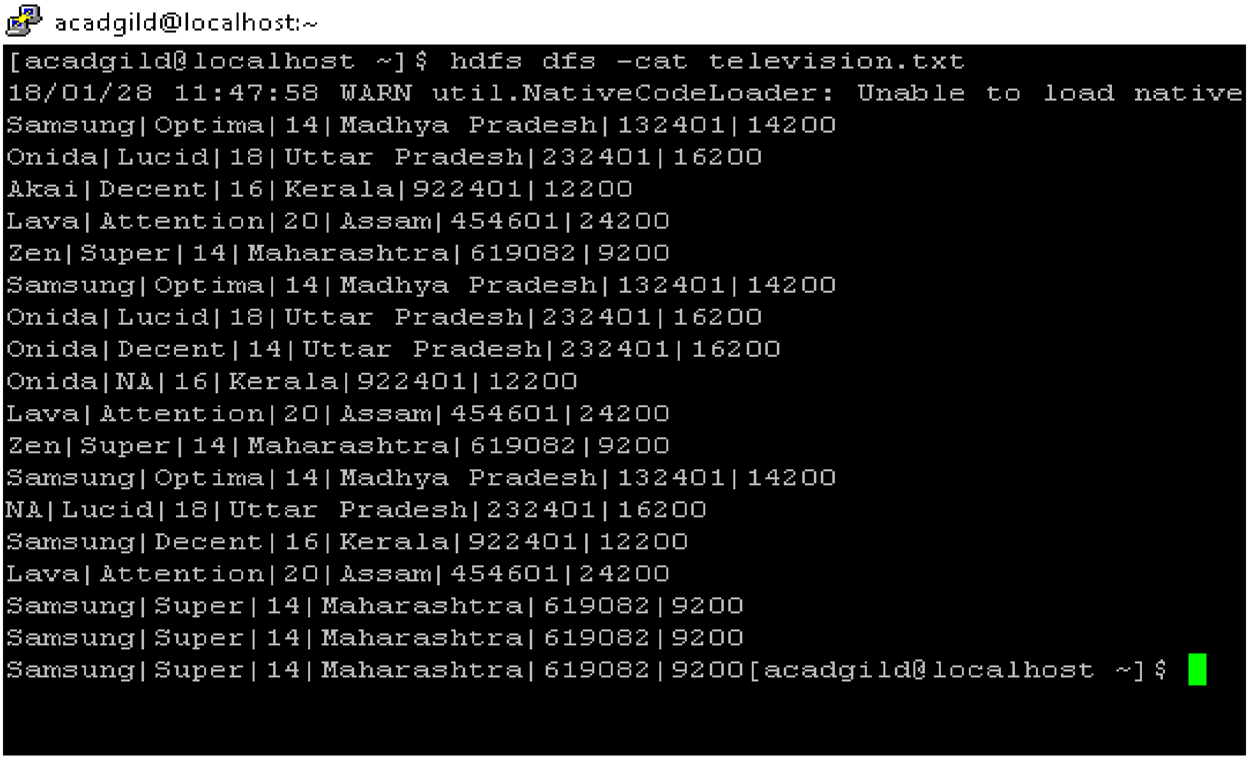
ASSIGNMENT 4.1

Task 1: Write a Map-Reduce Program to filter invalid record.

Input command: hdfs dfs –cat television.txt

Above command shows the input file television.txt

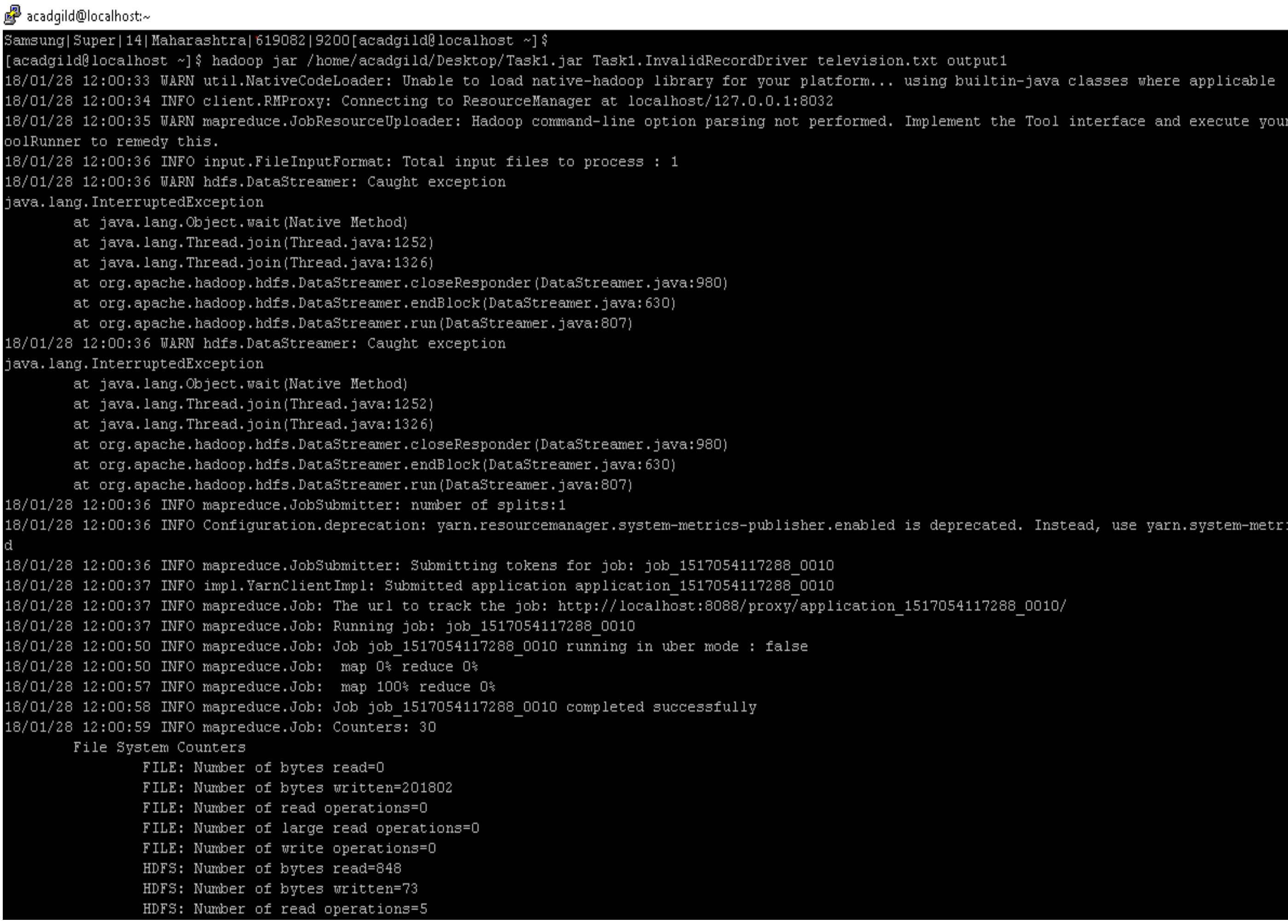


```
[acadgild@localhost ~]$ hdfs dfs -cat television.txt
18/01/28 11:47:58 WARN util.NativeCodeLoader: Unable to load native
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Akai|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Onida|Decent|14|Uttar Pradesh|232401|16200
Onida|NA|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
NA|Lucid|18|Uttar Pradesh|232401|16200
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 ~]$
```

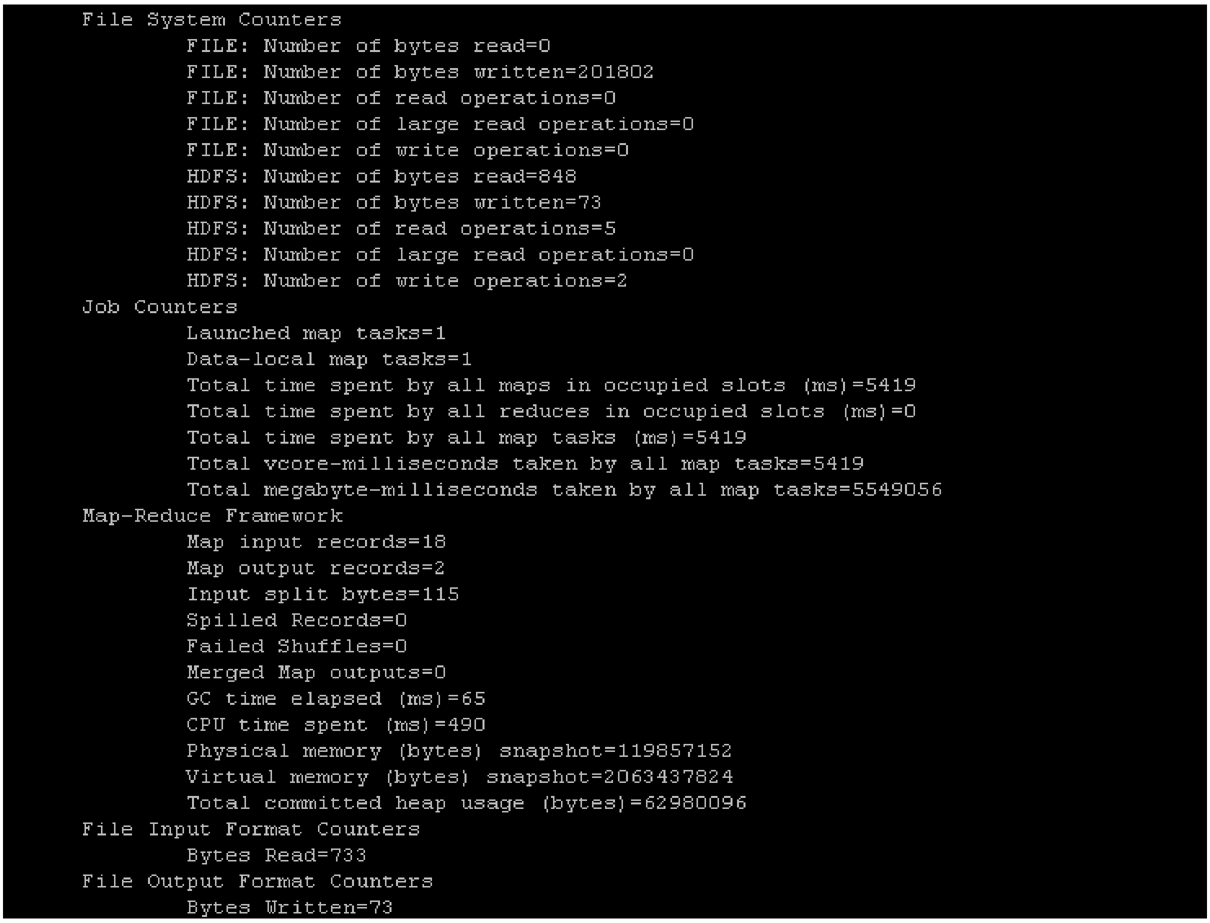
<u>Output Command</u>: hadoop jar /home/acadgild/Desktop/Task1.jar Task1.

InvalidRecordDriver television.txt output1

Above command is used to run the map reduce program to find invalid records. Task1.jar contains the required map-reduce program in which driver class is InvalidRecordDriver.



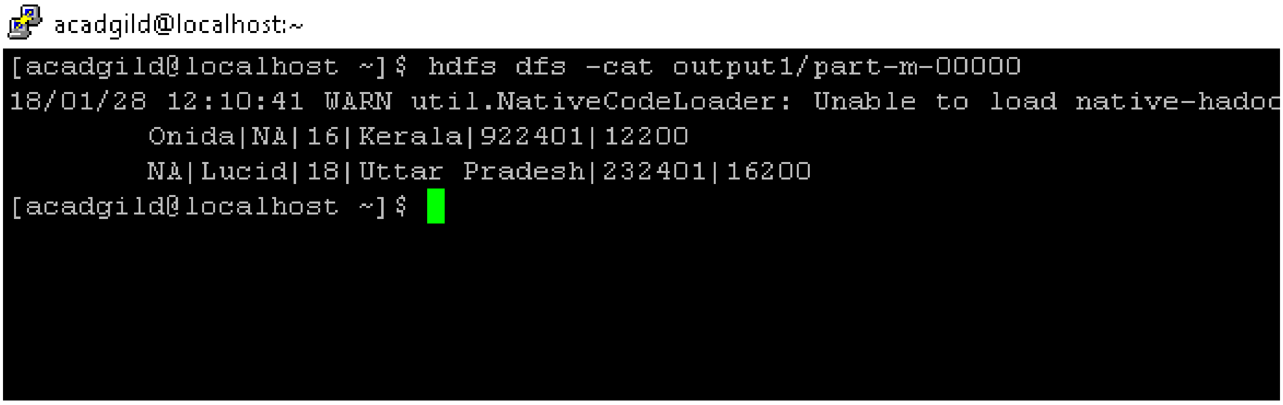
```
Samsung|Super|14|Maharashtra|619082|9200[acadgild@localhost ~]$
[acadgild@localhost ~]$ hadoop jar /home/acadgild/Desktop/Task1.jar Task1.InvalidRecordDriver television.txt output1
18/01/28 12:00:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/01/28 12:00:34 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/01/28 12:00:35 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your
oolRunner to remedy this.
18/01/28 12:00:36 INFO input.FileInputFormat: Total input files to process : 1
18/01/28 12:00:36 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:980)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:630)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:807)
18/01/28 12:00:36 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:980)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:630)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:807)
18/01/28 12:00:36 INFO mapreduce.JobSubmitter: number of splits:1
18/01/28 12:00:36 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metri
d
18/01/28 12:00:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1517054117288_0010
18/01/28 12:00:37 INFO impl.YarnClientImpl: Submitted application application_1517054117288_0010
18/01/28 12:00:37 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1517054117288_0010/
18/01/28 12:00:37 INFO mapreduce.Job: Running job: job_1517054117288_0010
18/01/28 12:00:50 INFO mapreduce.Job: Job job_1517054117288_0010 running in uber mode : false
18/01/28 12:00:50 INFO mapreduce.Job:  map 0% reduce 0%
18/01/28 12:00:57 INFO mapreduce.Job:  map 100% reduce 0%
18/01/28 12:00:58 INFO mapreduce.Job: Job job_1517054117288_0010 completed successfully
18/01/28 12:00:59 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=201802
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=848
                HDFS: Number of bytes written=73
                HDFS: Number of read operations=5
```



```
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=201802
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=848
                HDFS: Number of bytes written=73
                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)=5419
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=5419
                Total vcore-milliseconds taken by all map tasks=5419
                Total megabyte-milliseconds taken by all map tasks=5549056
        Map-Reduce Framework
                Map input records=18
                Map output records=2
                Input split bytes=115
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=65
                CPU time spent (ms)=490
                Physical memory (bytes) snapshot=119857152
                Virtual memory (bytes) snapshot=2063437824
                Total committed heap usage (bytes)=62980096
        File Input Format Counters
                Bytes Read=733
        File Output Format Counters
                Bytes Written=73
```

Output: hdfs dfs –cat output1/part-m-00000

Above command will shows the output data.

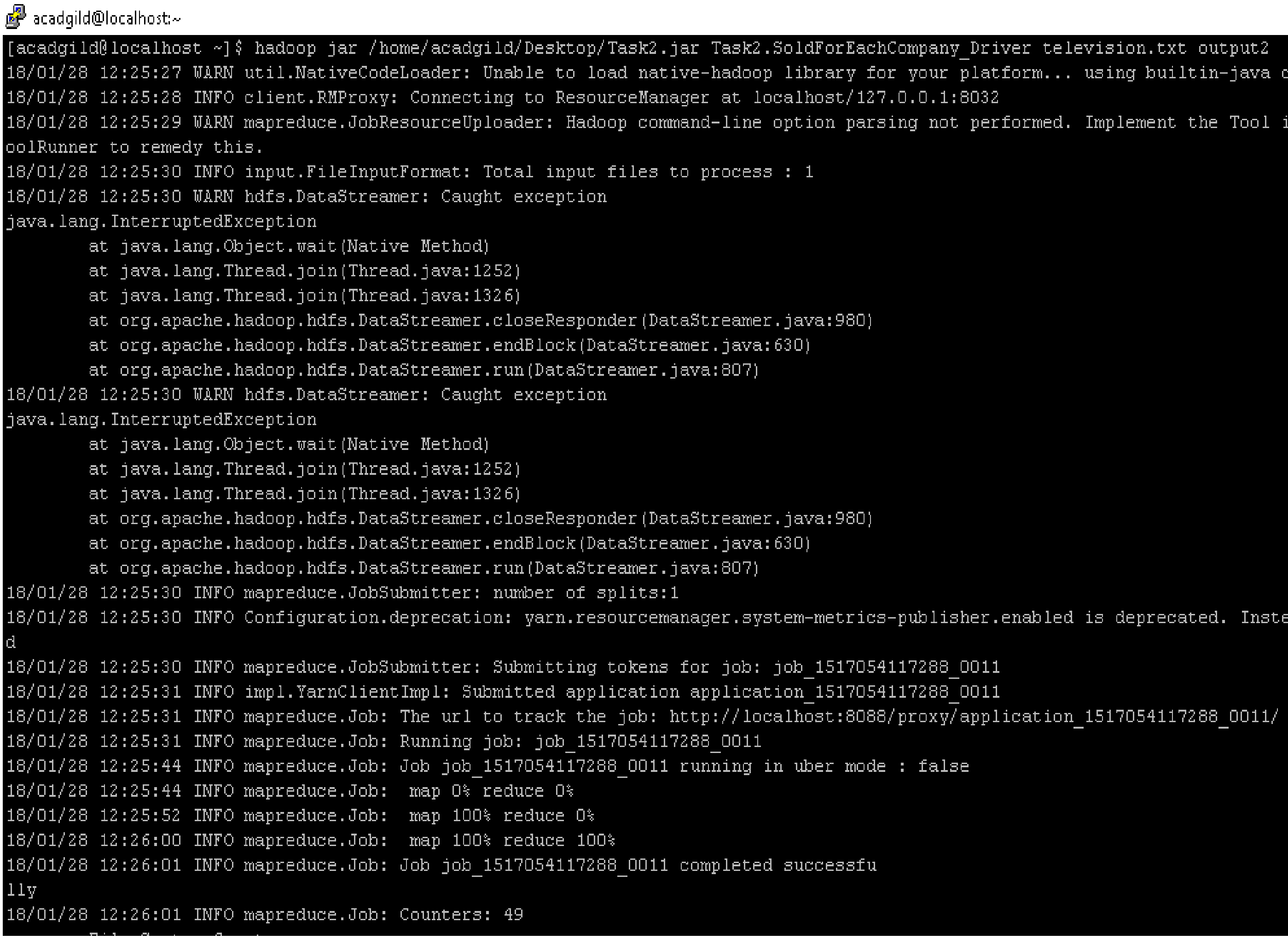


```
[acadgild@localhost ~]$ hdfs dfs -cat output1/part-m-00000
18/01/28 12:10:41 WARN util.NativeCodeLoader: Unable to load native-hadoc
        Onida|NA|16|Kerala|922401|12200
        NA|Lucid|18|Uttar Pradesh|232401|16200
[acadgild@localhost ~]$
```

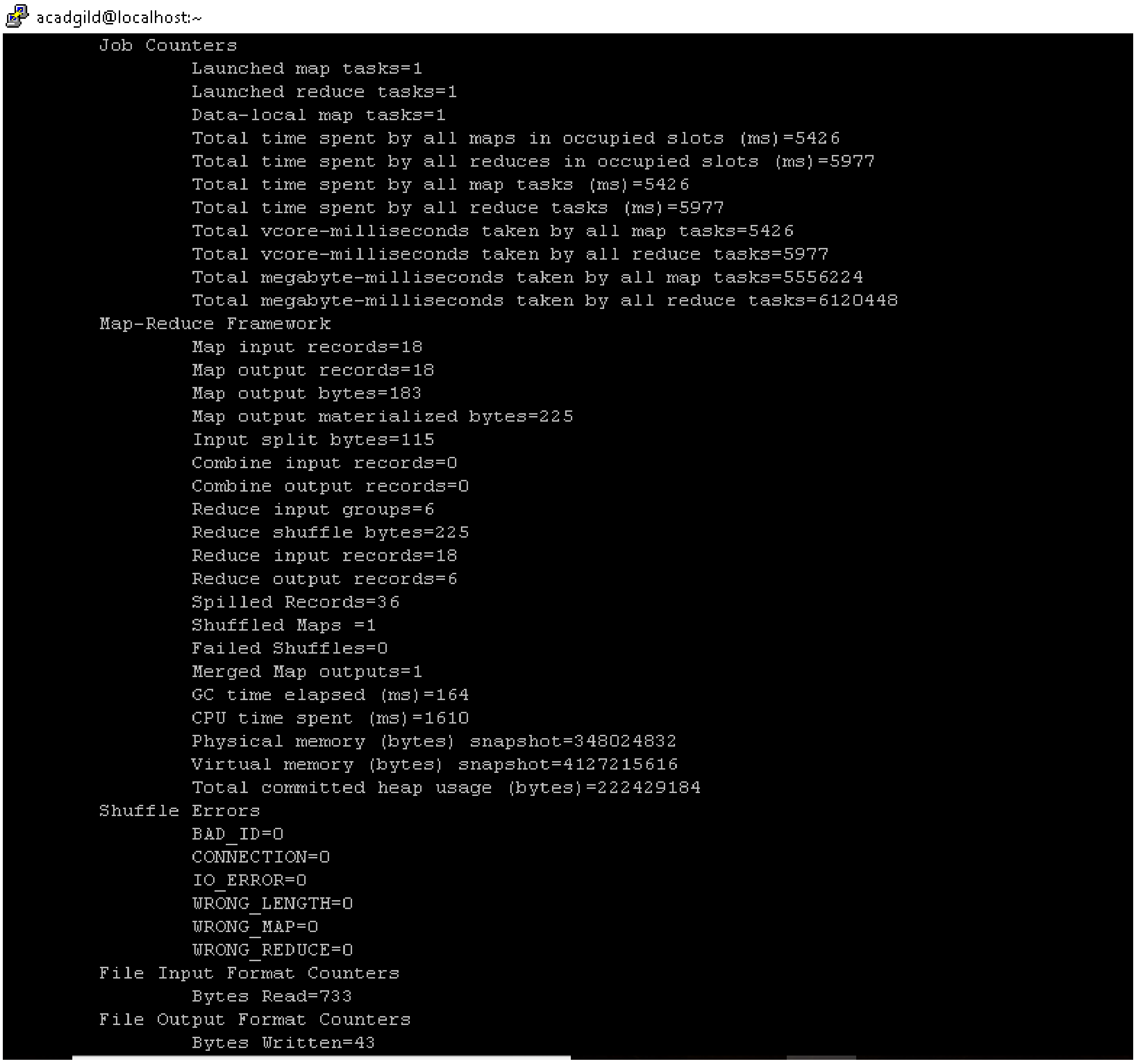
Task 2: Write a Map-Reduce Program to calculate the total unit sold for each company.

Output Command: hadoop jar /home/acadgild/Desktop/Task2.jar Task2.SoldForEachCompany_Driver television.txt output2

Above command is used to run the map reduce program to calculate total units sold for the company. Task1.jar contains the required map-reduce program in which driver class is SoldForEachCompany_Driver.

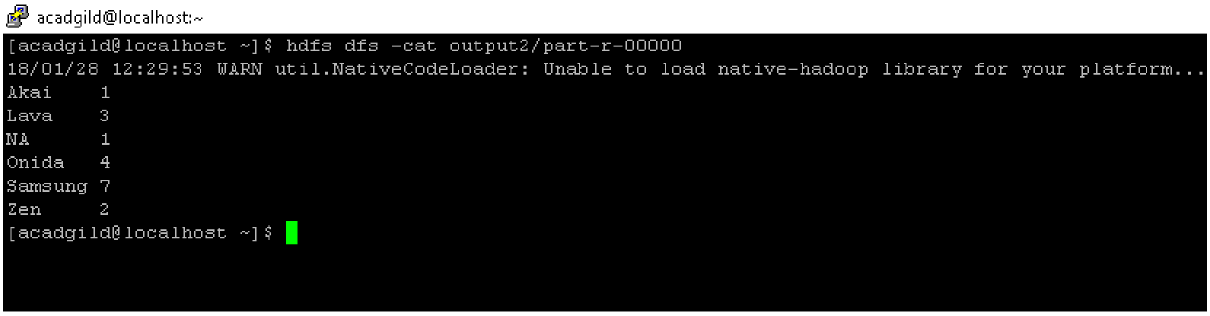


```
[acadgild@localhost ~]$ hadoop jar /home/acadgild/Desktop/Task2.jar Task2.SoldForEachCompany_Driver television.txt output2
18/01/28 12:25:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
18/01/28 12:25:28 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/01/28 12:25:29 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool i
oolRunner to remedy this.
18/01/28 12:25:30 INFO input.FileInputFormat: Total input files to process : 1
18/01/28 12:25:30 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:980)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:630)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:807)
18/01/28 12:25:30 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:980)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:630)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:807)
18/01/28 12:25:30 INFO mapreduce.JobSubmitter: number of splits:1
18/01/28 12:25:30 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Inste
d
18/01/28 12:25:30 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1517054117288_0011
18/01/28 12:25:31 INFO impl.YarnClientImpl: Submitted application application_1517054117288_0011
18/01/28 12:25:31 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1517054117288_0011/
18/01/28 12:25:31 INFO mapreduce.Job: Running job: job_1517054117288_0011
18/01/28 12:25:44 INFO mapreduce.Job: Job job_1517054117288_0011 running in uber mode : false
18/01/28 12:25:44 INFO mapreduce.Job:  map 0% reduce 0%
18/01/28 12:25:52 INFO mapreduce.Job:  map 100% reduce 0%
18/01/28 12:26:00 INFO mapreduce.Job:  map 100% reduce 100%
18/01/28 12:26:01 INFO mapreduce.Job: Job job_1517054117288_0011 completed successfu
lly
18/01/28 12:26:01 INFO mapreduce.Job: Counters: 49
```



```
        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)=5426
                Total time spent by all reduces in occupied slots (ms)=5977
                Total time spent by all map tasks (ms)=5426
                Total time spent by all reduce tasks (ms)=5977
                Total vcore-milliseconds taken by all map tasks=5426
                Total vcore-milliseconds taken by all reduce tasks=5977
                Total megabyte-milliseconds taken by all map tasks=5556224
                Total megabyte-milliseconds taken by all reduce tasks=6120448
        Map-Reduce Framework
                Map input records=18
                Map output records=18
                Map output bytes=183
                Map output materialized bytes=225
                Input split bytes=115
                Combine input records=0
                Combine output records=0
                Reduce input groups=6
                Reduce shuffle bytes=225
                Reduce input records=18
                Reduce output records=6
                Spilled Records=36
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=164
                CPU time spent (ms)=1610
                Physical memory (bytes) snapshot=348024832
                Virtual memory (bytes) snapshot=4127215616
                Total committed heap usage (bytes)=222429184
        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=43
```

Output: hdfs dfs –cat output2/part-r-00000

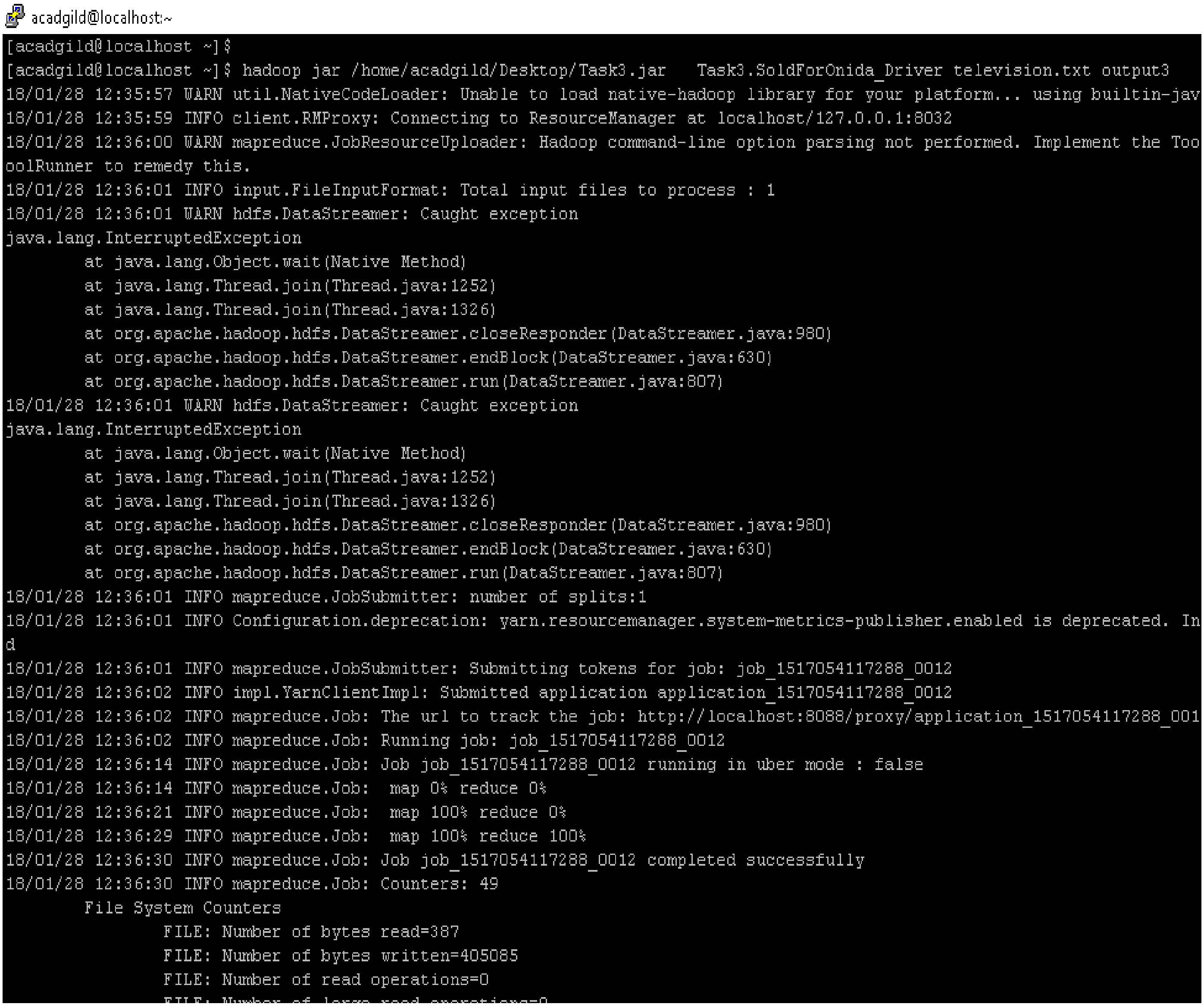


```
[acadgild@localhost ~]$ hdfs dfs -cat output2/part-r-00000
18/01/28 12:29:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Akai    1
Lava    3
NA      1
Onida   4
Samsung 7
Zen     2
[acadgild@localhost ~]$
```

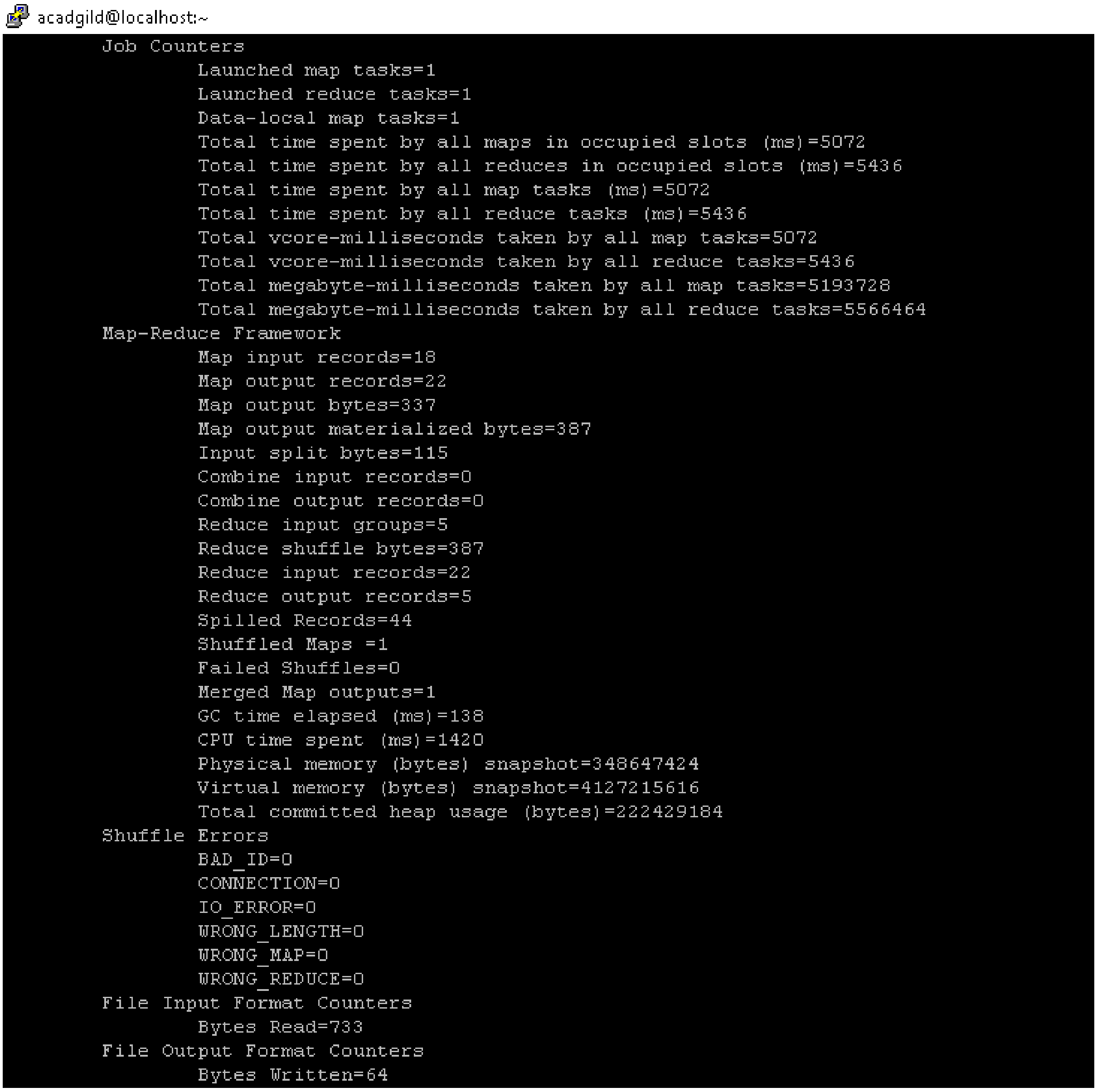
Task 3: Write a Map-Reduce Program to calculate the total unit sold in each state for Onida company.

Output Command: hadoop jar /home/acadgild/Desktop/Task3.jar Task3.SoldForOnida_Driver television.txt output3

Above command is used to run the map reduce program to calculate total units sold for onida company. Task1.jar contains the required map-reduce program in which driver class is SoldForOnida_Driver.

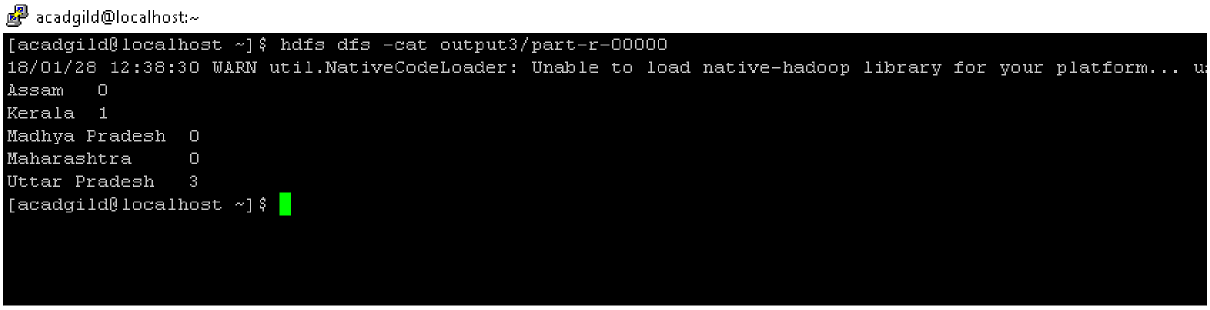


```
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hadoop jar /home/acadgild/Desktop/Task3.jar   Task3.SoldForOnida_Driver television.txt output3
18/01/28 12:35:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
18/01/28 12:35:59 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/01/28 12:36:00 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Too
oolRunner to remedy this.
18/01/28 12:36:01 INFO input.FileInputFormat: Total input files to process : 1
18/01/28 12:36:01 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:980)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:630)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:807)
18/01/28 12:36:01 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:980)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:630)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:807)
18/01/28 12:36:01 INFO mapreduce.JobSubmitter: number of splits:1
18/01/28 12:36:01 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. In
d
18/01/28 12:36:01 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1517054117288_0012
18/01/28 12:36:02 INFO impl.YarnClientImpl: Submitted application application_1517054117288_0012
18/01/28 12:36:02 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1517054117288_001
18/01/28 12:36:02 INFO mapreduce.Job: Running job: job_1517054117288_0012
18/01/28 12:36:14 INFO mapreduce.Job: Job job_1517054117288_0012 running in uber mode : false
18/01/28 12:36:14 INFO mapreduce.Job:  map 0% reduce 0%
18/01/28 12:36:21 INFO mapreduce.Job:  map 100% reduce 0%
18/01/28 12:36:29 INFO mapreduce.Job:  map 100% reduce 100%
18/01/28 12:36:30 INFO mapreduce.Job: Job job_1517054117288_0012 completed successfully
18/01/28 12:36:30 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=387
                FILE: Number of bytes written=405085
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
```



```
        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)=5072
                Total time spent by all reduces in occupied slots (ms)=5436
                Total time spent by all map tasks (ms)=5072
                Total time spent by all reduce tasks (ms)=5436
                Total vcore-milliseconds taken by all map tasks=5072
                Total vcore-milliseconds taken by all reduce tasks=5436
                Total megabyte-milliseconds taken by all map tasks=5193728
                Total megabyte-milliseconds taken by all reduce tasks=5566464
        Map-Reduce Framework
                Map input records=18
                Map output records=22
                Map output bytes=337
                Map output materialized bytes=387
                Input split bytes=115
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=387
                Reduce input records=22
                Reduce output records=5
                Spilled Records=44
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=138
                CPU time spent (ms)=1420
                Physical memory (bytes) snapshot=348647424
                Virtual memory (bytes) snapshot=4127215616
                Total committed heap usage (bytes)=222429184
        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=64
```

Output: hdfs dfs –cat output3/part-r-00000



```
[acadgild@localhost ~]$ hdfs dfs -cat output3/part-r-00000
18/01/28 12:38:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... u
Assam   0
Kerala  1
Madhya Pradesh  0
Maharashtra     0
Uttar Pradesh   3
[acadgild@localhost ~]$
```