Assignment No 4

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.

Solution:

Create Maven project with name TvDatasetExample, add new package `AcadgildAssignment4.TvDataSetExample`

Add new java class named “TvDatasetmapper.java”.

TvDatasetmapper.java:

```
package AcadgildAssignment4.TvDataSetExample;

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class TvSetMapper extends Mapper<LongWritable, Text, LongWritable, Text>
{


    public void map(LongWritable key, Text value, Context context) throws
IOException, InterruptedException{

    if(recordIsInvalid(value)==false){
        Text record = new Text();
      record = value;
        context.write(key,record);
    }
}
    private boolean recordIsInvalid(Text record){
        String[] lineArray = record.toString().split("\\|");
    boolean isInvalid = false;
    for(int i=0;i<lineArray.length;i++){
        if(lineArray[i].equals("NA")){
            isInvalid = true;
        }
    }
    return isInvalid;
 }
public static void main(String[] args) throws Exception{

          Configuration conf = new Configuration();

          Job job = Job.getInstance(conf, "Tv Sales Invalid REcords");
          job.setJarByClass(TvSetMapper.class);

          job.setMapOutputKeyClass(LongWritable.class);
          job.setMapOutputValueClass(Text.class);

          job.setMapperClass(TvSetMapper.class);

          FileInputFormat.addInputPath(job, new Path(args[0]));
          FileOutputFormat.setOutputPath(job, new Path(args[1]));

          System.exit(job.waitForCompletion(true) ? 0 : 1);
      }
}
```

Compile it and run maven install . It create jar file in target folder .

- Put television.txt file in HDFS at /user/root/television.txt

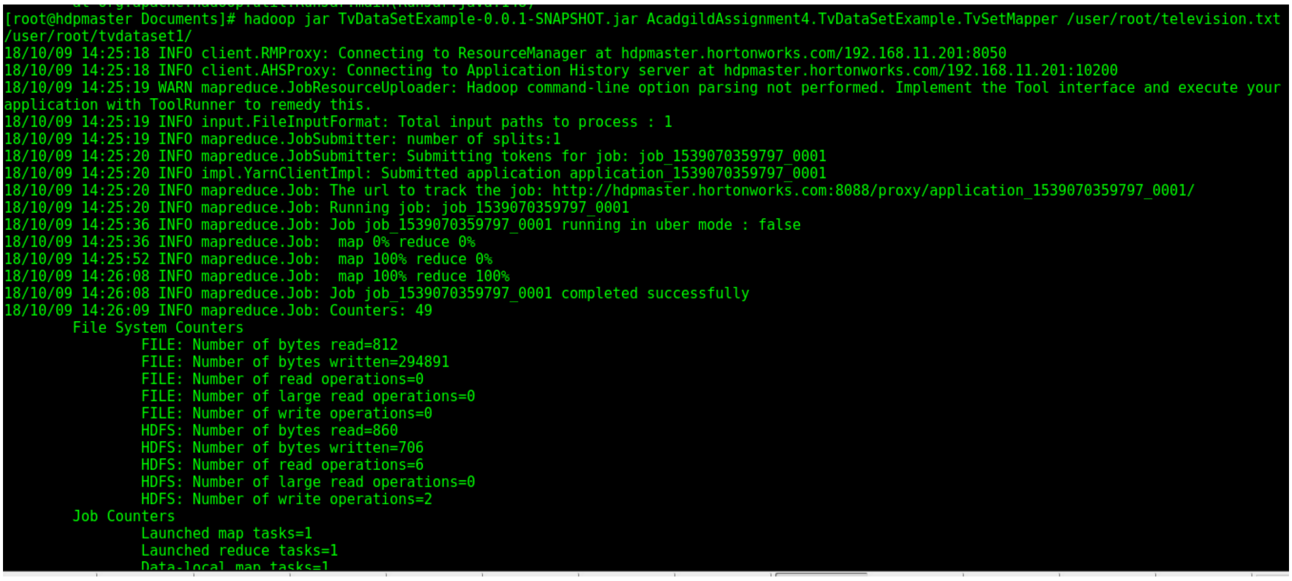
-Copy TvDataSetExample-0.0.1-SNAPSHOT.jar to VM Using File Zilla

Run below command :

hadoop jar TvDataSetExample-0.0.1-SNAPSHOT.jar
AcadgildAssignment4.TvDataSetExample.TvSetMapper /user/root/television.txt
/user/root/tvdataset1/

Commad description:

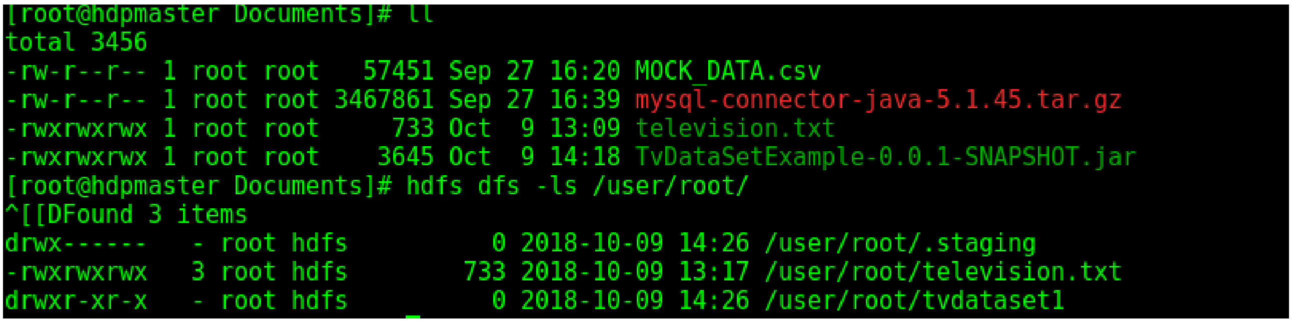
1. Run hadoop jar : hadoop jar hadoop jar TvDataSetExample-0.0.1-SNAPSHOT.jar
2. Classes required to run : AcadgildAssignment4.TvDataSetExample.TvSetMapper
3. Input file hdfs path: /user/root/television.txt
4. output directory hdfs path : /user/root/tvdataset1/



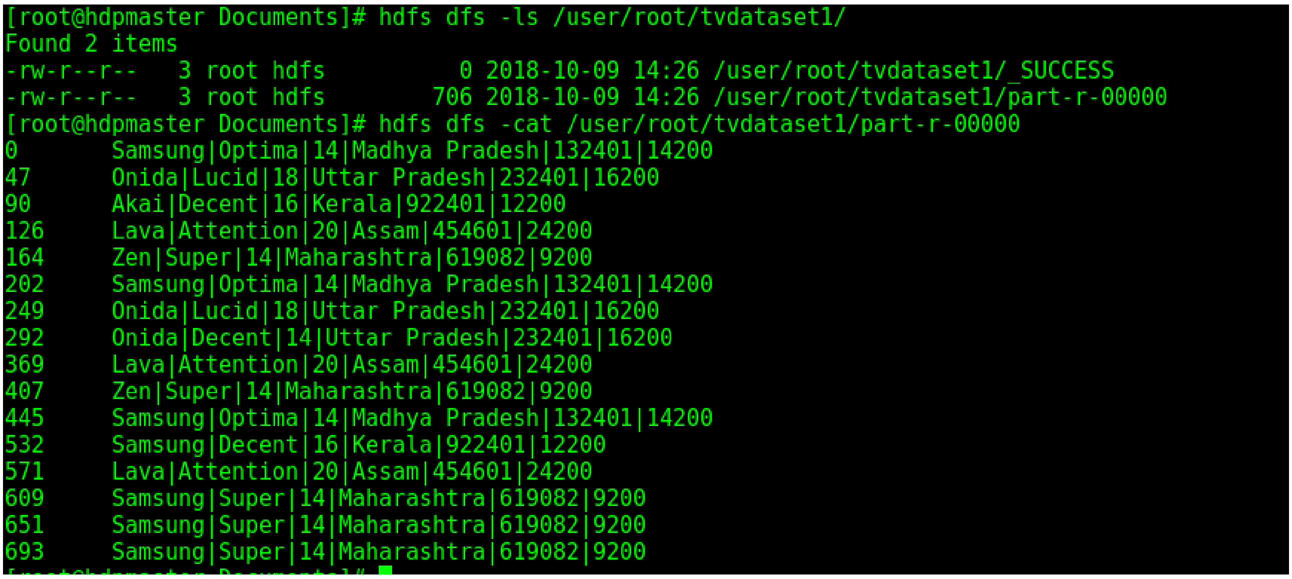
```
[root@hdpmaster Documents]# hadoop jar TvDataSetExample-0.0.1-SNAPSHOT.jar AcadgildAssignment4.TvDataSetExample.TvSetMapper /user/root/television.txt
/user/root/tvdataset1/
18/10/09 14:25:18 INFO client.RMProxy: Connecting to ResourceManager at hdpmaster.hortonworks.com/192.168.11.201:8050
18/10/09 14:25:18 INFO client.AHSProxy: Connecting to Application History server at hdpmaster.hortonworks.com/192.168.11.201:10200
18/10/09 14:25:19 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your
application with ToolRunner to remedy this.
18/10/09 14:25:19 INFO input.FileInputFormat: Total input paths to process : 1
18/10/09 14:25:19 INFO mapreduce.JobSubmitter: number of splits:1
18/10/09 14:25:20 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1539070359797_0001
18/10/09 14:25:20 INFO impl.YarnClientImpl: Submitted application application_1539070359797_0001
18/10/09 14:25:20 INFO mapreduce.Job: The url to track the job: http://hdpmaster.hortonworks.com:8088/proxy/application_1539070359797_0001/
18/10/09 14:25:20 INFO mapreduce.Job: Running job: job_1539070359797_0001
18/10/09 14:25:36 INFO mapreduce.Job: Job job_1539070359797_0001 running in uber mode : false
18/10/09 14:25:36 INFO mapreduce.Job:  map 0% reduce 0%
18/10/09 14:25:52 INFO mapreduce.Job:  map 100% reduce 0%
18/10/09 14:26:08 INFO mapreduce.Job:  map 100% reduce 100%
18/10/09 14:26:08 INFO mapreduce.Job: Job job_1539070359797_0001 completed successfully
18/10/09 14:26:09 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=812
                FILE: Number of bytes written=294891
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=860
                HDFS: Number of bytes written=706
                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
```

It create file **/user/root/tvdataset1/part-r-00000**

To view File :
hdfs dfs -cat /user/root/tvdataset1/part-r-00000



```
[root@hdpmaster Documents]# ll
total 3456
-rw-r--r-- 1 root root   57451 Sep 27 16:20 MOCK_DATA.csv
-rw-r--r-- 1 root root 3467861 Sep 27 16:39 mysql-connector-java-5.1.45.tar.gz
-rwxrwxrwx 1 root root     733 Oct  9 13:09 television.txt
-rwxrwxrwx 1 root root    3645 Oct  9 14:18 TvDataSetExample-0.0.1-SNAPSHOT.jar
[root@hdpmaster Documents]# hdfs dfs -ls /user/root/
^[[DFound 3 items
drwx------   - root hdfs          0 2018-10-09 14:26 /user/root/.staging
-rwxrwxrwx   3 root hdfs        733 2018-10-09 13:17 /user/root/television.txt
drwxr-xr-x   - root hdfs          0 2018-10-09 14:26 /user/root/tvdataset1
```



```
[root@hdpmaster Documents]# hdfs dfs -ls /user/root/tvdataset1/
Found 2 items
-rw-r--r--   3 root hdfs          0 2018-10-09 14:26 /user/root/tvdataset1/_SUCCESS
-rw-r--r--   3 root hdfs        706 2018-10-09 14:26 /user/root/tvdataset1/part-r-00000
[root@hdpmaster Documents]# hdfs dfs -cat /user/root/tvdataset1/part-r-00000
0       Samsung|Optima|14|Madhya Pradesh|132401|14200
47      Onida|Lucid|18|Uttar Pradesh|232401|16200
90      Akai|Decent|16|Kerala|922401|12200
126     Lava|Attention|20|Assam|454601|24200
164     Zen|Super|14|Maharashtra|619082|9200
202     Samsung|Optima|14|Madhya Pradesh|132401|14200
249     Onida|Lucid|18|Uttar Pradesh|232401|16200
292     Onida|Decent|14|Uttar Pradesh|232401|16200
369     Lava|Attention|20|Assam|454601|24200
407     Zen|Super|14|Maharashtra|619082|9200
445     Samsung|Optima|14|Madhya Pradesh|132401|14200
532     Samsung|Decent|16|Kerala|922401|12200
571     Lava|Attention|20|Assam|454601|24200
609     Samsung|Super|14|Maharashtra|619082|9200
651     Samsung|Super|14|Maharashtra|619082|9200
693     Samsung|Super|14|Maharashtra|619082|9200
```

It will sort out records which contains “**NA**”.

Task 2:
Write a Map Reduce program to calculate the total units sold for each Company.

Solution:

For task 2 and 3 . There is separate maven project.

For task 2 , company is key and value is 1. So after split record into array . We get name of compant at index 0.

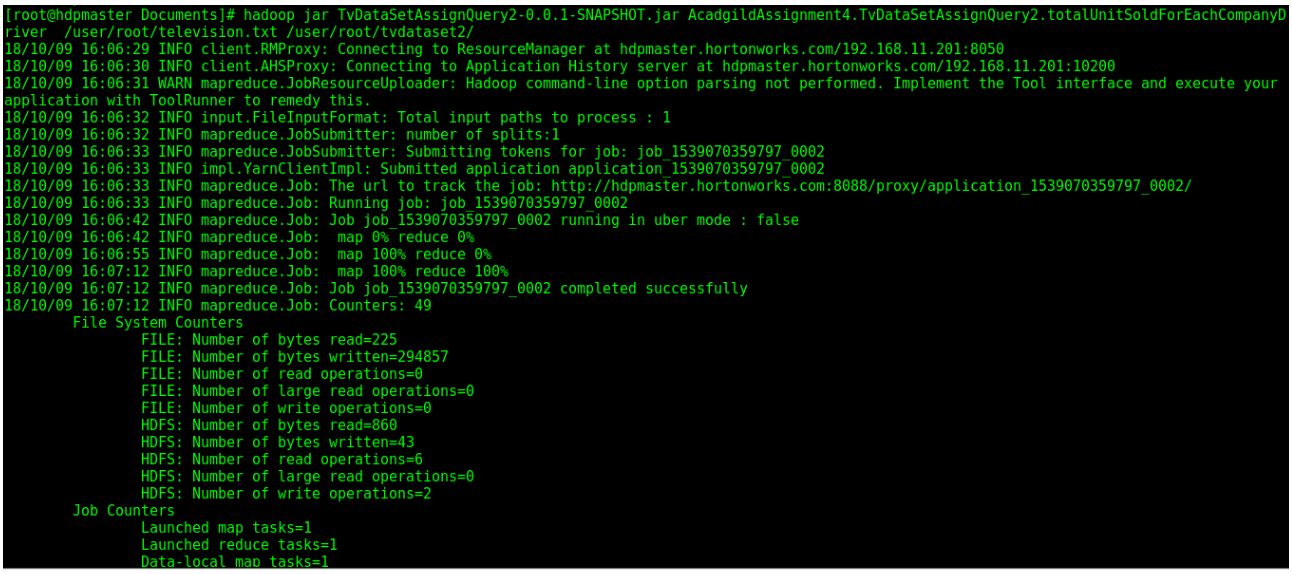
Source of project in Task2and3 folder.

Run command :

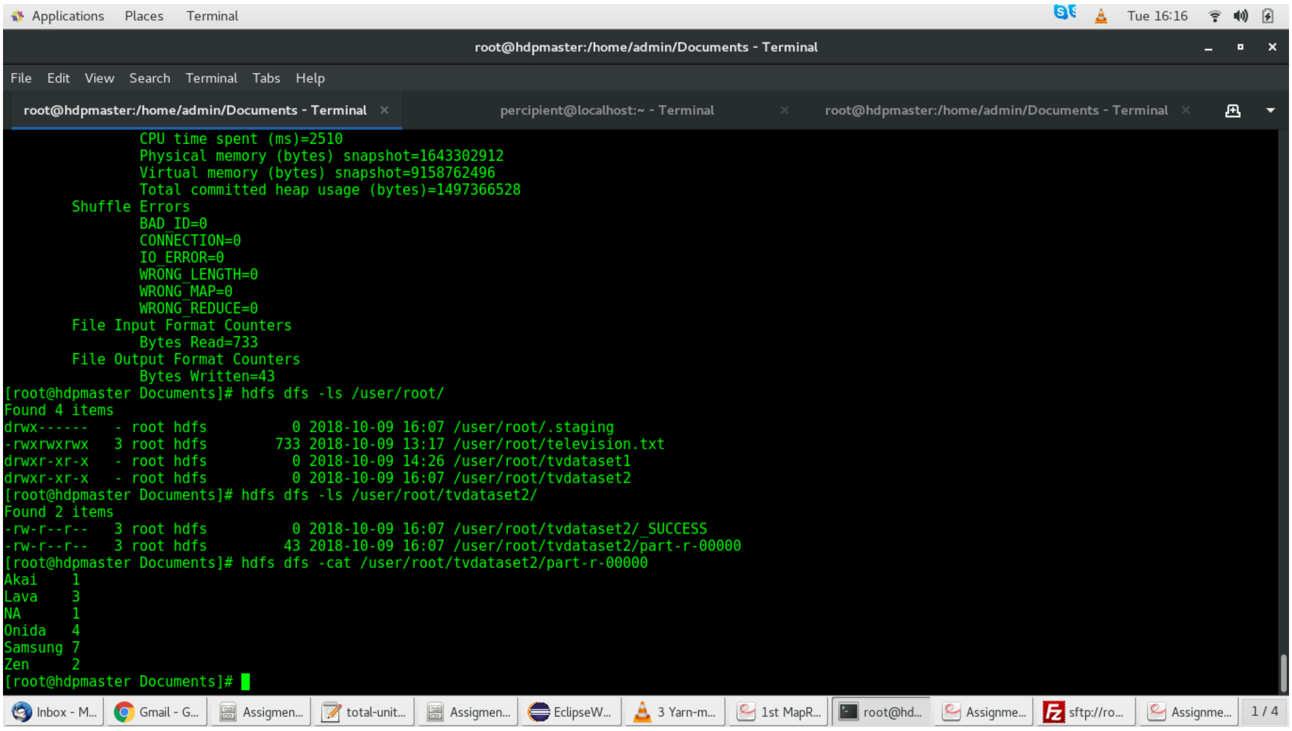
hadoop jar TvDataSetAssignQuery2-0.0.1-SNAPSHOT.jar AcadgildAssignment4.TvDataSetAssignQuery2.totalUnitSoldForEachCompanyDriver /user/root/television.txt /user/root/tvdataset2/

It will create file in /user/root/tvdataset2/

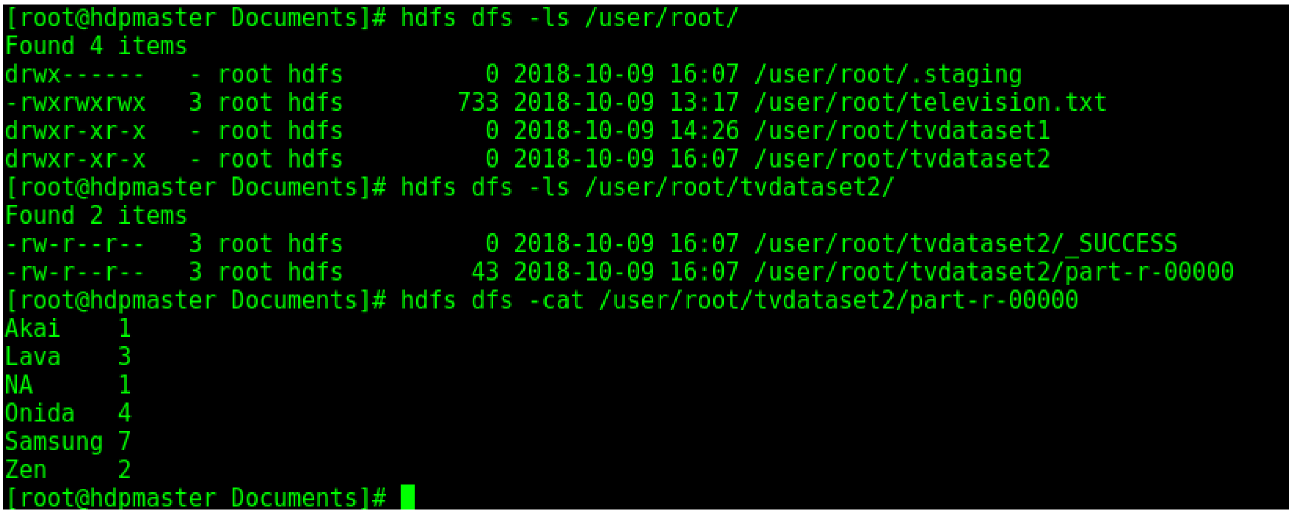
hdfs dfs -cat /user/root/tvdataset2/part-r-00000



```
[root@hdpmaster Documents]# hadoop jar TvDataSetAssignQuery2-0.0.1-SNAPSHOT.jar AcadgildAssignment4.TvDataSetAssignQuery2.totalUnitSoldForEachCompanyD
river  /user/root/television.txt /user/root/tvdataset2/
18/10/09 16:06:29 INFO client.RMProxy: Connecting to ResourceManager at hdpmaster.hortonworks.com/192.168.11.201:8050
18/10/09 16:06:30 INFO client.AHSProxy: Connecting to Application History server at hdpmaster.hortonworks.com/192.168.11.201:10200
18/10/09 16:06:31 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your
application with ToolRunner to remedy this.
18/10/09 16:06:32 INFO input.FileInputFormat: Total input paths to process : 1
18/10/09 16:06:32 INFO mapreduce.JobSubmitter: number of splits:1
18/10/09 16:06:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1539070359797_0002
18/10/09 16:06:33 INFO impl.YarnClientImpl: Submitted application application_1539070359797_0002
18/10/09 16:06:33 INFO mapreduce.Job: The url to track the job: http://hdpmaster.hortonworks.com:8088/proxy/application_1539070359797_0002/
18/10/09 16:06:33 INFO mapreduce.Job: Running job: job_1539070359797_0002
18/10/09 16:06:42 INFO mapreduce.Job: Job job_1539070359797_0002 running in uber mode : false
18/10/09 16:06:42 INFO mapreduce.Job:  map 0% reduce 0%
18/10/09 16:06:55 INFO mapreduce.Job:  map 100% reduce 0%
18/10/09 16:07:12 INFO mapreduce.Job:  map 100% reduce 100%
18/10/09 16:07:12 INFO mapreduce.Job: Job job_1539070359797_0002 completed successfully
18/10/09 16:07:12 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=225
                FILE: Number of bytes written=294857
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=860
                HDFS: Number of bytes written=43
                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
```



```
                CPU time spent (ms)=2510
                Physical memory (bytes) snapshot=1643302912
                Virtual memory (bytes) snapshot=9158762496
                Total committed heap usage (bytes)=1497366528
        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
[root@hdpmaster Documents]# hdfs dfs -ls /user/root/
Found 4 items
drwx------   - root hdfs          0 2018-10-09 16:07 /user/root/.staging
-rwxrwxrwx   3 root hdfs        733 2018-10-09 13:17 /user/root/television.txt
drwxr-xr-x   - root hdfs          0 2018-10-09 14:26 /user/root/tvdataset1
drwxr-xr-x   - root hdfs          0 2018-10-09 16:07 /user/root/tvdataset2
[root@hdpmaster Documents]# hdfs dfs -ls /user/root/tvdataset2/
Found 2 items
-rw-r--r--   3 root hdfs          0 2018-10-09 16:07 /user/root/tvdataset2/_SUCCESS
-rw-r--r--   3 root hdfs         43 2018-10-09 16:07 /user/root/tvdataset2/part-r-00000
[root@hdpmaster Documents]# hdfs dfs -cat /user/root/tvdataset2/part-r-00000
Akai    1
Lava    3
NA      1
Onida   4
Samsung 7
Zen     2
[root@hdpmaster Documents]# 
```



```
[root@hdpmaster Documents]# hdfs dfs -ls /user/root/
Found 4 items
drwx------   - root hdfs          0 2018-10-09 16:07 /user/root/.staging
-rwxrwxrwx   3 root hdfs        733 2018-10-09 13:17 /user/root/television.txt
drwxr-xr-x   - root hdfs          0 2018-10-09 14:26 /user/root/tvdataset1
drwxr-xr-x   - root hdfs          0 2018-10-09 16:07 /user/root/tvdataset2
[root@hdpmaster Documents]# hdfs dfs -ls /user/root/tvdataset2/
Found 2 items
-rw-r--r--   3 root hdfs          0 2018-10-09 16:07 /user/root/tvdataset2/_SUCCESS
-rw-r--r--   3 root hdfs         43 2018-10-09 16:07 /user/root/tvdataset2/part-r-00000
[root@hdpmaster Documents]# hdfs dfs -cat /user/root/tvdataset2/part-r-00000
Akai    1
Lava    3
NA      1
Onida   4
Samsung 7
Zen     2
[root@hdpmaster Documents]# 
```

output of Task 2

Task 3:

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

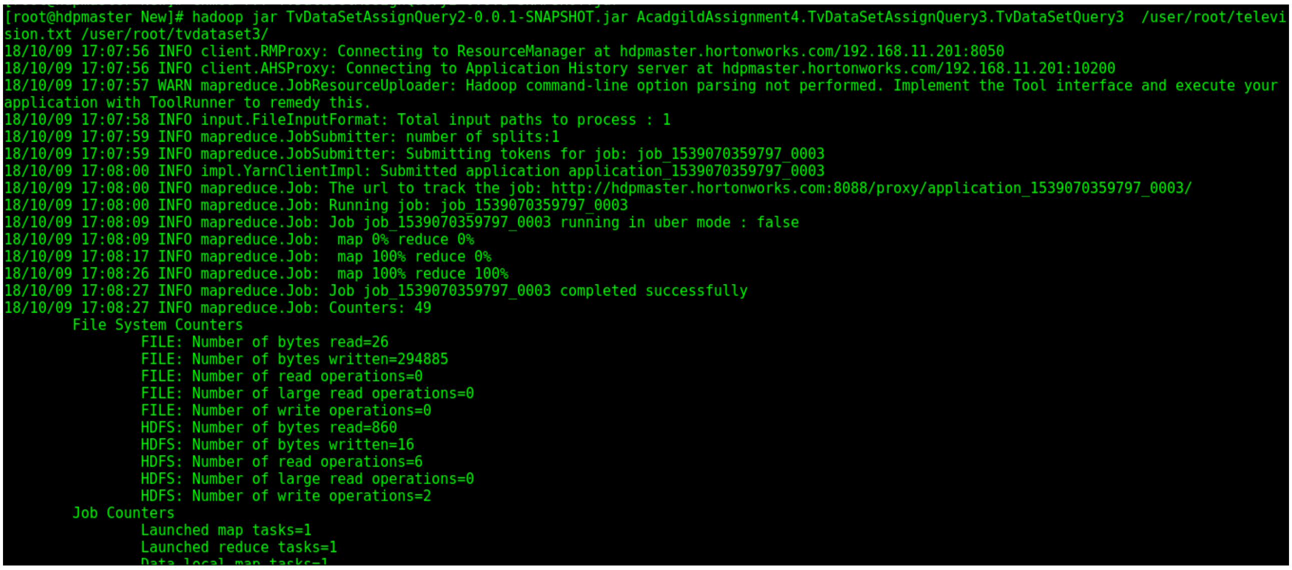
Solution:

As in task 2 , we have to sort record in map phase . After record split we get company name at index 0 and state at index 3. Here state is key and integer 1 is value.

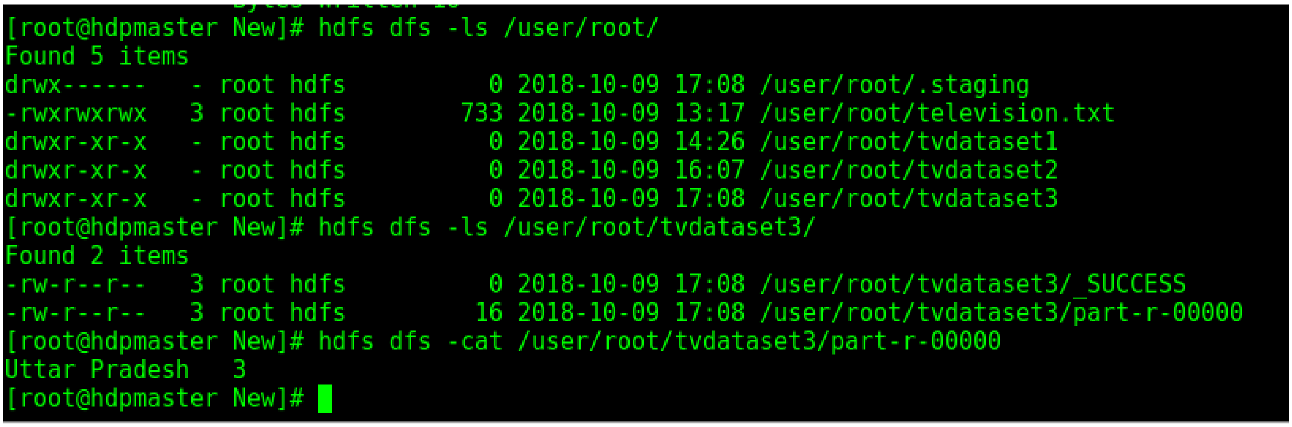
Run below command :

hadoop jar TvDataSetAssignQuery2-0.0.1-SNAPSHOT.jar AcadgildAssignment4.TvDataSetAssignQuery3.TvDataSetQuery3 /user/root/television.txt /user/root/tvdataset3/

It will create file at **/user/root/tvdataset3/** as shown in below screen shot.



```
[root@hdpmaster New]# hadoop jar TvDataSetAssignQuery2-0.0.1-SNAPSHOT.jar AcadgildAssignment4.TvDataSetAssignQuery3.TvDataSetQuery3  /user/root/televi
sion.txt /user/root/tvdataset3/
18/10/09 17:07:56 INFO client.RMProxy: Connecting to ResourceManager at hdpmaster.hortonworks.com/192.168.11.201:8050
18/10/09 17:07:56 INFO client.AHSProxy: Connecting to Application History server at hdpmaster.hortonworks.com/192.168.11.201:10200
18/10/09 17:07:57 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your
application with ToolRunner to remedy this.
18/10/09 17:07:58 INFO input.FileInputFormat: Total input paths to process : 1
18/10/09 17:07:59 INFO mapreduce.JobSubmitter: number of splits:1
18/10/09 17:07:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1539070359797_0003
18/10/09 17:08:00 INFO impl.YarnClientImpl: Submitted application application_1539070359797_0003
18/10/09 17:08:00 INFO mapreduce.Job: The url to track the job: http://hdpmaster.hortonworks.com:8088/proxy/application_1539070359797_0003/
18/10/09 17:08:00 INFO mapreduce.Job: Running job: job_1539070359797_0003
18/10/09 17:08:09 INFO mapreduce.Job: Job job_1539070359797_0003 running in uber mode : false
18/10/09 17:08:09 INFO mapreduce.Job:  map 0% reduce 0%
18/10/09 17:08:17 INFO mapreduce.Job:  map 100% reduce 0%
18/10/09 17:08:26 INFO mapreduce.Job:  map 100% reduce 100%
18/10/09 17:08:27 INFO mapreduce.Job: Job job_1539070359797_0003 completed successfully
18/10/09 17:08:27 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=26
                FILE: Number of bytes written=294885
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=860
                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
                Launched reduce tasks=1
                Data-local map tasks=1
```



```
                Bytes Written=16
[root@hdpmaster New]# hdfs dfs -ls /user/root/
Found 5 items
drwx------   - root hdfs          0 2018-10-09 17:08 /user/root/.staging
-rwxrwxrwx   3 root hdfs        733 2018-10-09 13:17 /user/root/television.txt
drwxr-xr-x   - root hdfs          0 2018-10-09 14:26 /user/root/tvdataset1
drwxr-xr-x   - root hdfs          0 2018-10-09 16:07 /user/root/tvdataset2
drwxr-xr-x   - root hdfs          0 2018-10-09 17:08 /user/root/tvdataset3
[root@hdpmaster New]# hdfs dfs -ls /user/root/tvdataset3/
Found 2 items
-rw-r--r--   3 root hdfs          0 2018-10-09 17:08 /user/root/tvdataset3/_SUCCESS
-rw-r--r--   3 root hdfs         16 2018-10-09 17:08 /user/root/tvdataset3/part-r-00000
[root@hdpmaster New]# hdfs dfs -cat /user/root/tvdataset3/part-r-00000
Uttar Pradesh   3
[root@hdpmaster New]# █
```