

Assignment 3.2

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.

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

For this task I created a method that checks if a record is invalid. To achieve this, we split every record using “\” character as rule, into an array. Then we check every part of the array and we validate if that part contains “NA”, then we return a Boolean variable:

```
private boolean recordIsValid(Text record){
    String[] lineArray = record.toString().split("\\|");
    boolean isValid = false;
    for(int i=0;i<lineArray.length;i++){
        if(lineArray[i].equals("NA")){
            isValid = true;
        }
    }
    return isValid;
}
```

The mapper uses this method to filter out invalid records:

```
public void map(LongWritable key, Text value, Context context)
throws IOException, InterruptedException{
    if(recordIsValid(value)==false){
        Text record = new Text();
        record = value;
        context.write(key,record);
    }
}
```

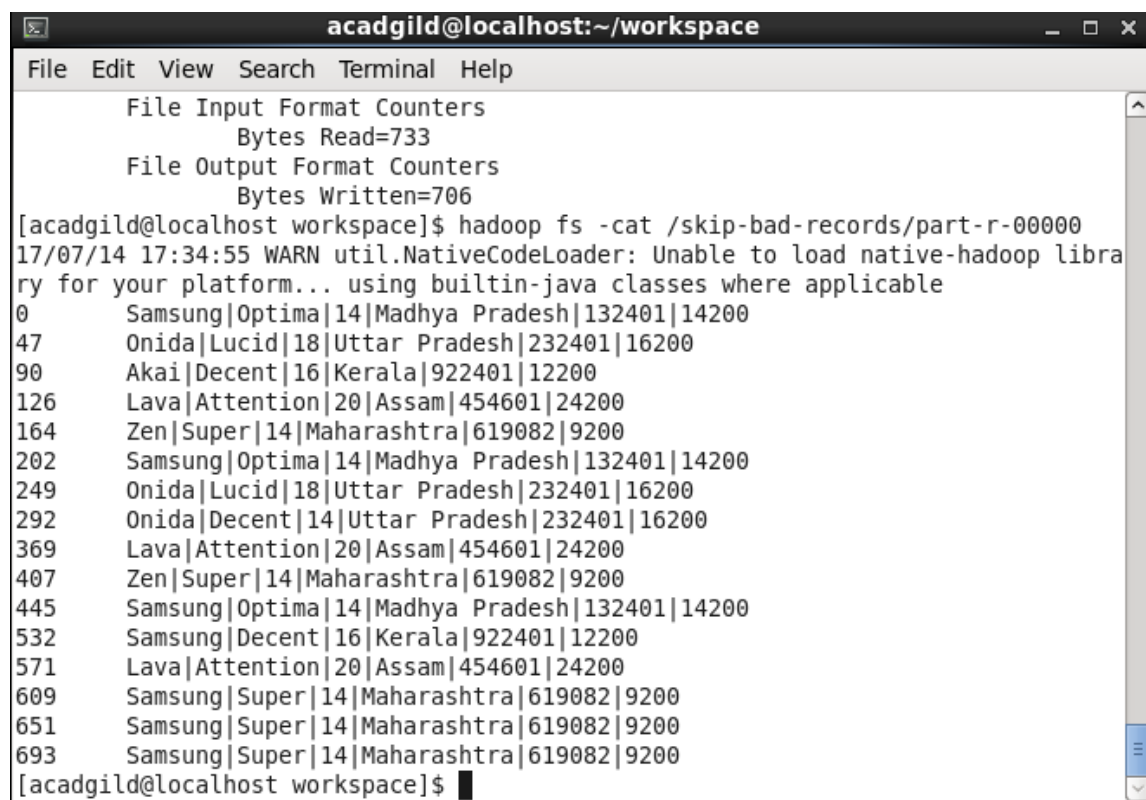
Then we need to export the .jar file from Eclipse and in the command shell we type:

```
hadoop jar /home/acadgild/workspace/TvSales.jar television.txt /skip-bad-records
```

- **jar** is the command to execute the mapping and reducing
- **/home/acadgild/workspace/TvSales.jar** location on file system of the jar exported before
- **television.txt** is the file location on HDFS containing the records
- **/skip-bad-records** is the destination directory on HDFS after reducing the records. Since there is no reducer for this task, Hadoop uses identity Reducer

To check the results obtained on HDFS we can type

```
hadoop fs -cat /filter-bad-records/part-r-00000
```



```
acadgild@localhost:~/workspace
File Edit View Search Terminal Help
File Input Format Counters
  Bytes Read=733
File Output Format Counters
  Bytes Written=706
[acadgild@localhost workspace]$ hadoop fs -cat /skip-bad-records/part-r-00000
17/07/14 17:34:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
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
[acadgild@localhost workspace]$
```

The complete code is in **filter-invalid-records-program.txt** and the filtered records in **filtered-records.txt** obtained from HDFS using **hadoop fs -copyToLocal** command