# 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 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;
}
```

The mapper uses this method to filter out invalid records:

```
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);
        }
}
```

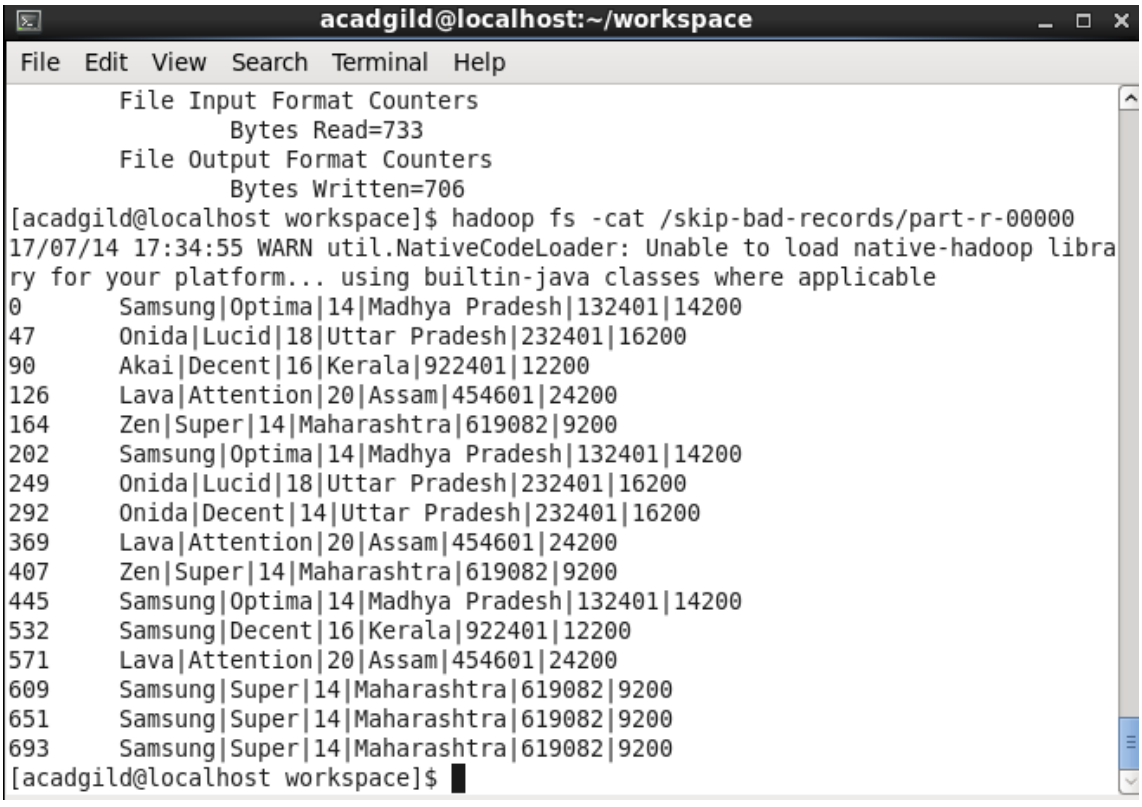
Then we need to export the .jar file from Eclipse and in the command shell we type:

**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 befofe
- **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 the 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**




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