**Assignment 3.2**

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.

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

context.

**INPUT**

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

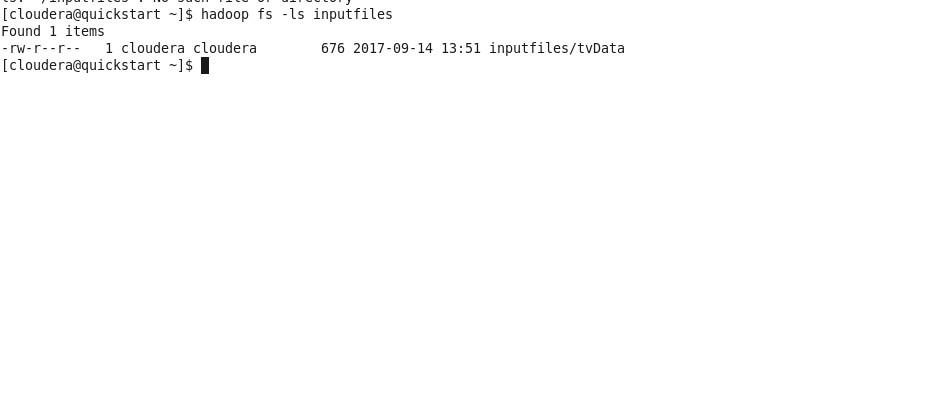
Lava|Attention|20|Assam|454601|24200

Samsung|Super|14|Maharashtra|619082|9200

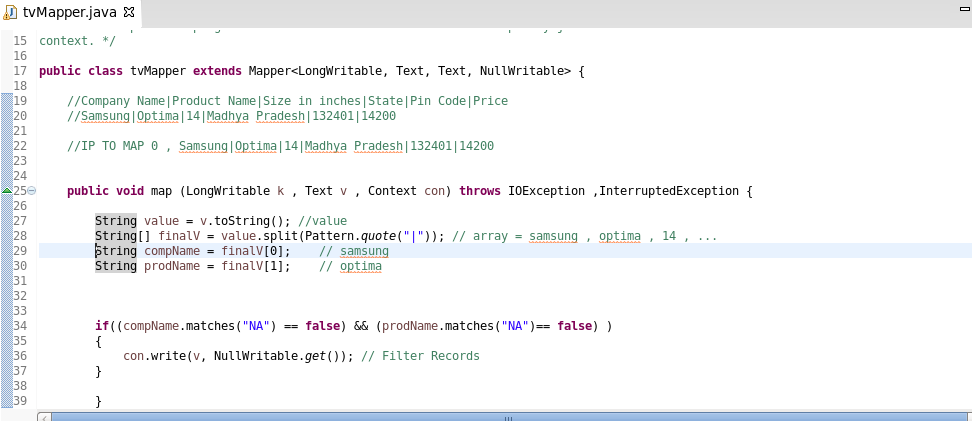
Samsung|Super|14|Maharashtra|619082|9200

Samsung|Super|14|Maharashtra|619082|9200

**We have put the input file in the HDFS folder /inputfiles/tvData**



**Mapper CODE :**

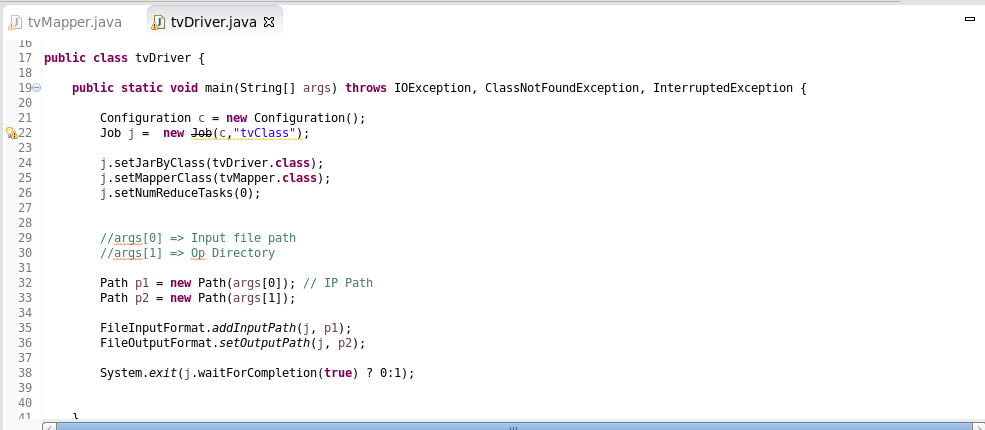


We have use the logic for Filtering out the Rows which have “NA” in Comp Name or in Prodcut Name

Note: We will not be using Reduce Function in this Code because we wont any reducing means like grouping and aggregation.

We are passing only those to Mapper output which doesn’t contain “NA” which are valid records

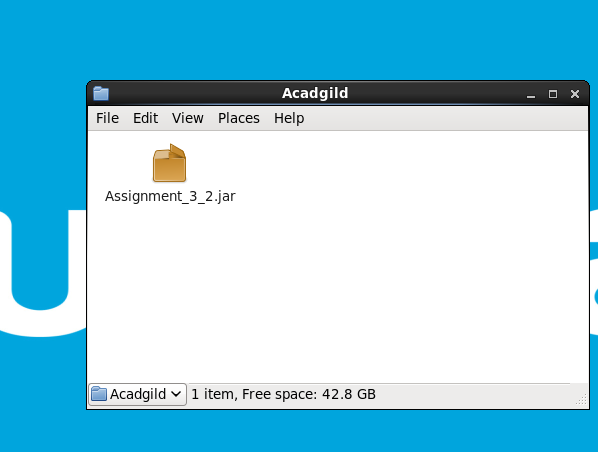
**Driver Code :**



In this Driver Code we have set reduce task to zero because we doesn’t need Reducer.

**Running the Code in the Hadoop Cluster**

We have exported a Hadoop Jar file for this code : Assignment\_3\_2.jar





**hadoop jar Desktop/Acadgild/Assignment\_3\_2.jar acadgild\_3\_2.tvDriver inputfiles/tvData output/Result\_assignment\_3\_2**

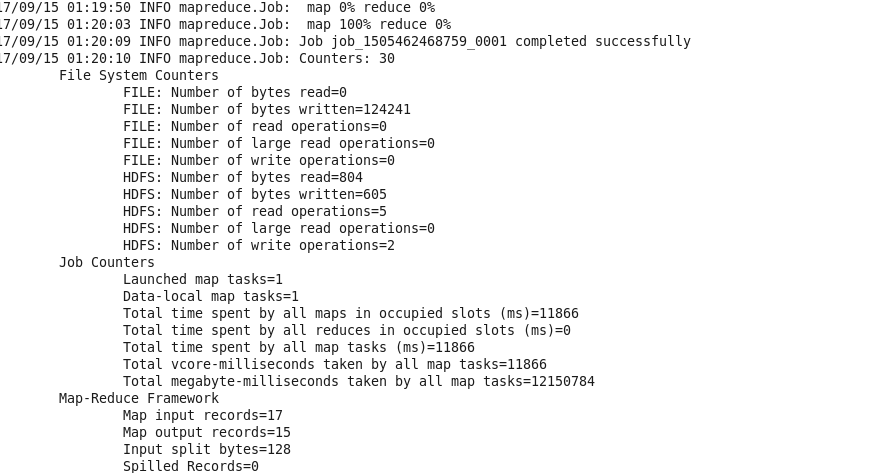
**Desktop/Acadgild/Assignment\_3\_2.jar :**  Jar file location which we want to run

**acadgild\_3\_2.tvDriver:** Driver Class , where the main function resides

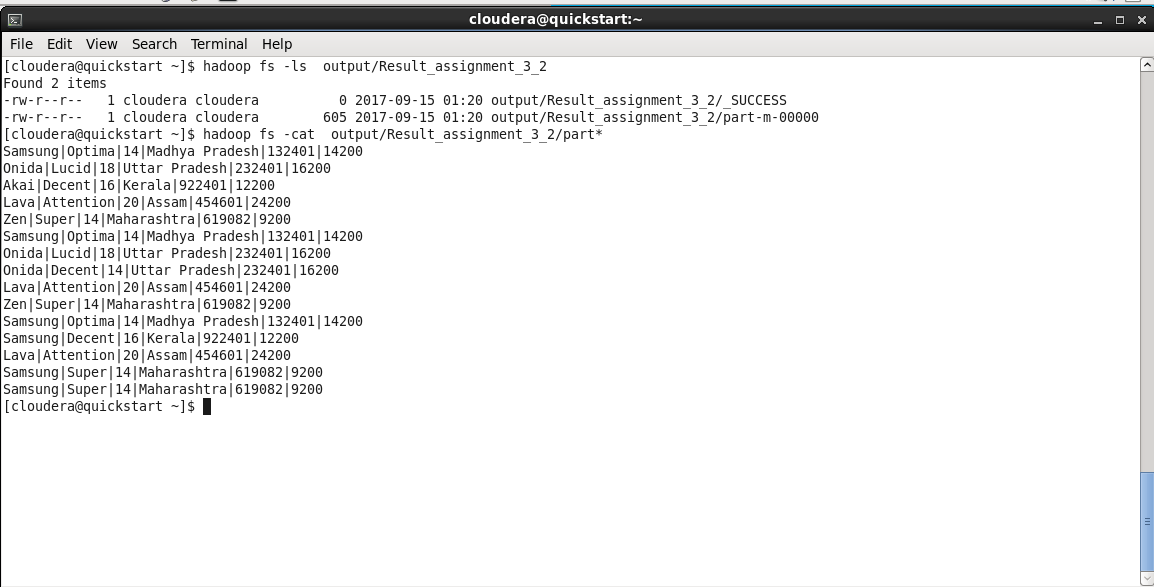
**inputfiles/tvData :** Input File

**output/Result\_assignment\_3\_2:** Output file or result Locatio

**Job Completion :**



**Output File**



File Uploaded in the Github :

1. Jar file
2. Document containing the flow

Successfully Completed the Problem