**Problem Statement:**

Analyzing Consumer Complaints.

Hadoop can be applied with many projects of various domains:

Data consolidation

Machine learning

Specialized analysis (domain specific)

**This project comes under Data consolidation**

**Data consolidation**

This is like "enterprise data hub" or "data lake." The idea is you have disparate data sources, and you want to perform analysis across them. This type of project consists of getting feeds from all the sources (either real time or as a batch) and transferring them into Hadoop. Sometimes this is the first step for becoming a “data-driven company”. Data lakes usually materialize as files on HDFS and tables in pig/Hive or Impala and then we can create simply pretty reports on the data contained in hdfs.

**About data:**

The dataset consists of consumer complaints details. In the process consumer files the complaints to consumer forum and then consumer forum forwards the complaint to respective company.

**DATA SET DESCRIPTION:**

|  |  |  |
| --- | --- | --- |
| **Column heading** | **index** | **Description** |
| Date received | 0 | date on which consumer filed the complaint |
| Product | 1 | Type of the product |
| Sub-product | 2 | Sub product type |
| Issue | 3 | Issue faced by the consumer |
| Sub-issue | 4 | Any sub issues if exists |
| Consumer complaint narrative | 5 | Detailed description of complaint |
| Company public response | 6 | Company’s public response to the complaint |
| Company | 7 | Name of the company |
| State | 8 | State from which consumer filed the complaint |
| ZIP code | 9 | Zip code |
| Submitted via | 10 | Channel from which complaint was submitted |
| Date sent to company | 11 | Date on which consumer forum forwarded the complaint to company |
| Company response to consumer | 12 | Company’s response to the consumer |
| Timely response? | 13 |  |
| Consumer disputed? | 14 |  |
| Complaint ID | 15 | Unique complaint id |

This data is comma delimited.

In some rows there are few columns which are enclosed in double quotes and have many commas and due to this the same column gets spitted into many columns for ex:

**Sample record:**

10/16/2015,Debt collection,"Other (phone, health club, etc.)",Cont'd attempts collect debt not owed,Debt was discharged in bankruptcy,,,"Convergent Resources, Inc.",OH,438XX,Web,10/16/2015,Closed with explanation,Yes,,1612132

This entire column "Other (phone, health club, etc.)" Should be product but, if we split this file based on comma then this column will be splitted into 3 columns which will result in wrong outputs.

In order to tackle this we should remove the commas present only inside double quotes.

Since Hadoop is used to handle big data it’s recommended to use java map reduce to remove unnecessary commas.

**I**

**1.** Write a mapreduce program to remove commas present inside the double quotes.

Note: Work on this problem statements after doing the data cleaning as mentioned above.

**II**

**1.** Write a pig script to find no of complaints which got timely response

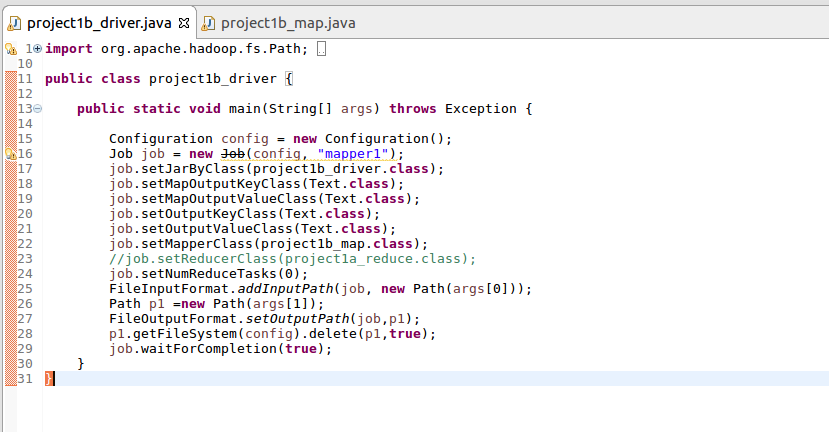
**2.** Write a pig script to find no of complaints where consumer forum forwarded the complaint same day they received to respective company

**3.** Write a pig script to find list of companies topping in complaint chart (companies with maximum number of complaints)

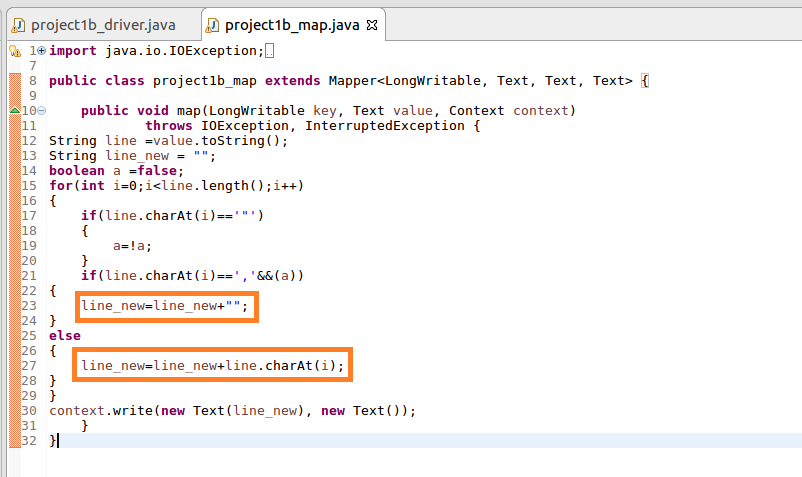
**4.** Write a pig script to find no of complaints filed with product type has "Debt collection" for the year 2015

**MAP REDUCE TO REMOVE COMMA INBETWEEN THE DOUBLE QUOTES:**

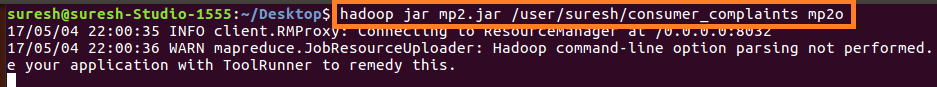
**DRIVER CLASS:**

****

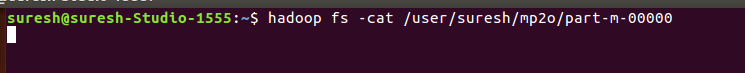
**MAPPER CLASS:**

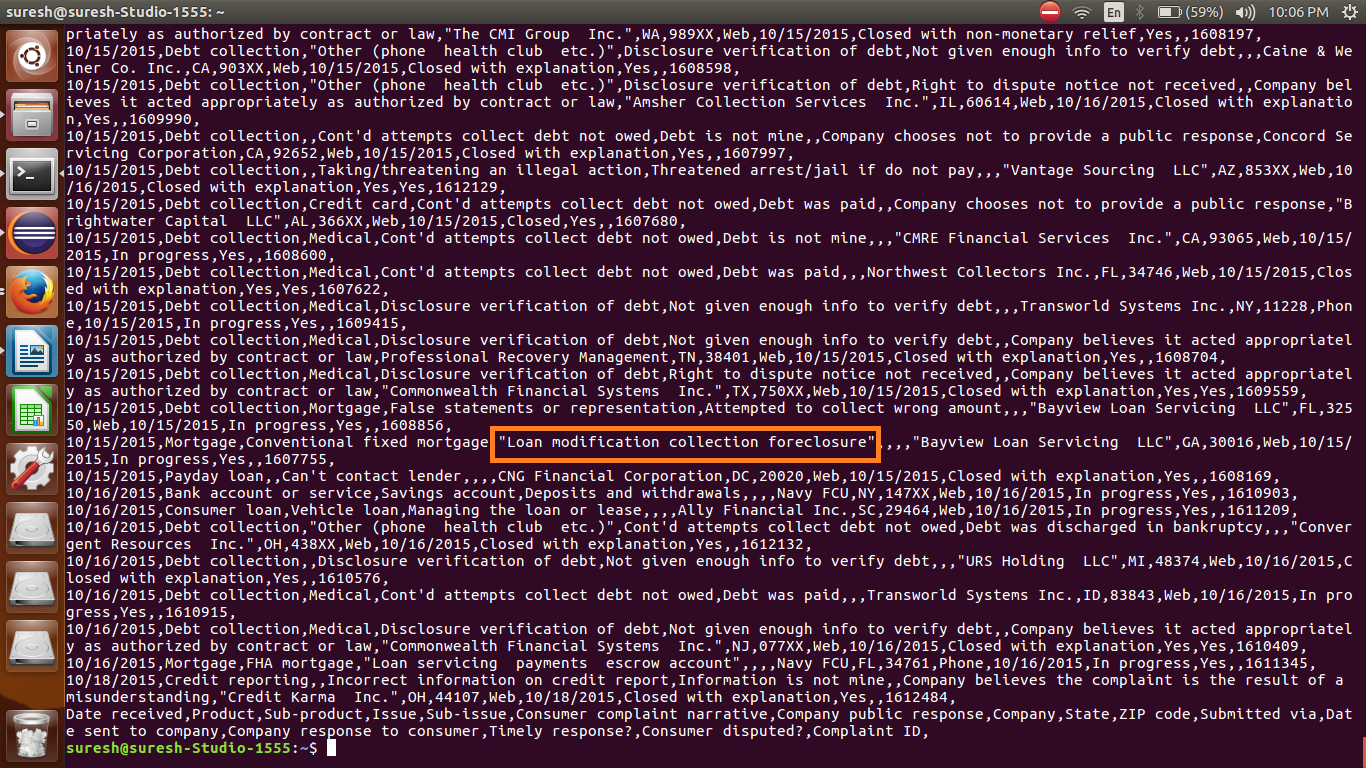


**RUNNING MAPREDUCE JAR:**



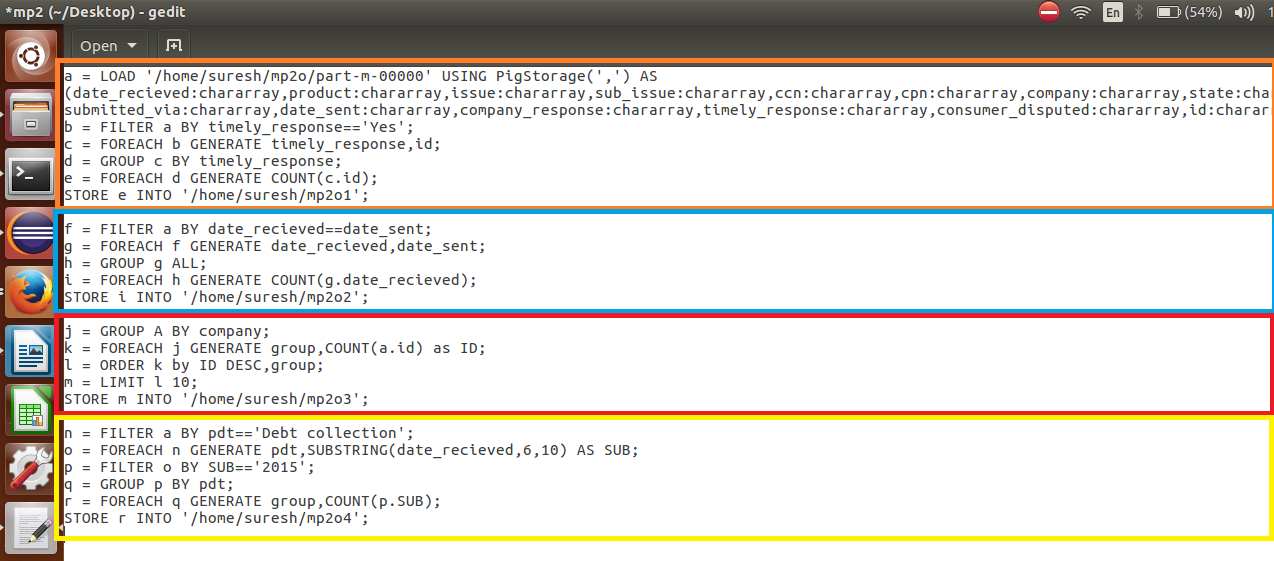
**CAT COMMANDING THE FILE AFTER RUNNING WITH MAPREDUCE:**



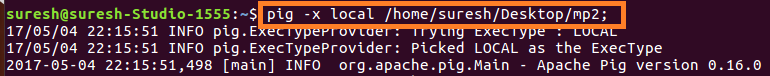
****

**GETTING THE COMMA REMOVED FILE TO THE LOCAL TO PERFORM PIG OPERATION:**

**PIG SCRIPT:**



**RUNNING PIG SCRIPT:**



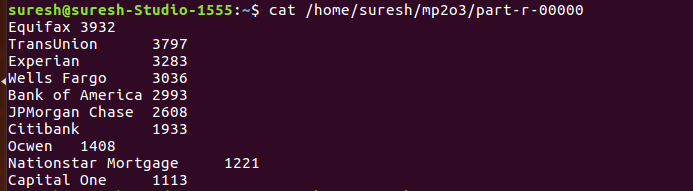
**OUTPUT 1:**

D:\assignment\mp2o1.png

**OUTPUT 2:**

D:\assignment\mp2o2.png

**OUTPUT 3:**



**OUTPUT 4:**

D:\assignment\mp2o4.png