**PROJECT 2**

**1. Introduction:**

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.

**2. Prerequisites:**

You should have Hadoop cluster installed in your system.

**3. Associated Data Files:**

[**https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/consumer\_complai nts.csv**](https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/consumer_complai%20nts.csv)

**3.1. Dataset Description:**

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 split into many columns.

For example: 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 split 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.

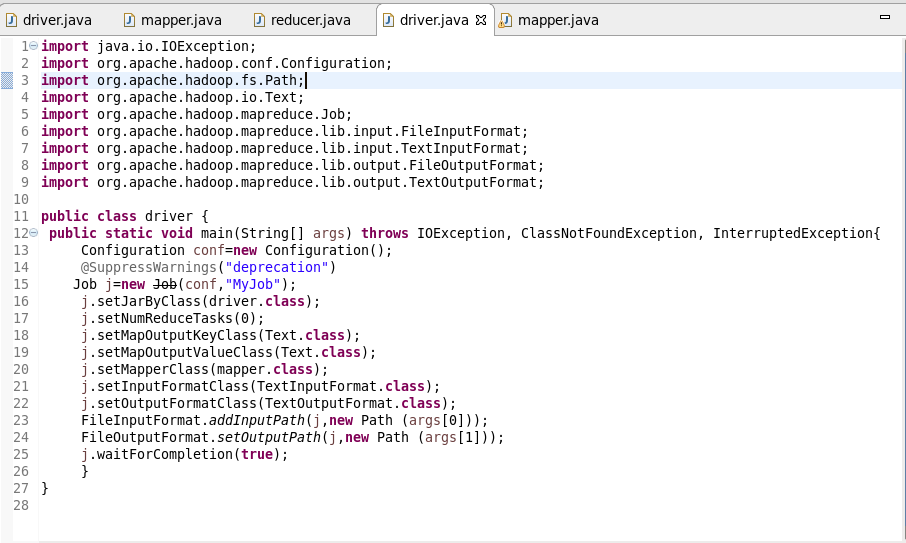
**4. Problem Statement 1:**

* 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.

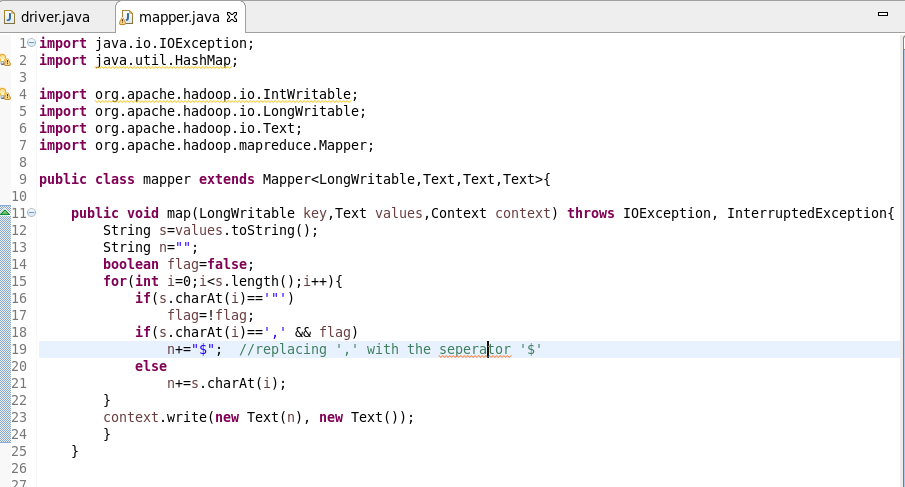
**5. Program:**

**5.1. MAPREDUCE PROGRAM:**

* **Driver class:**



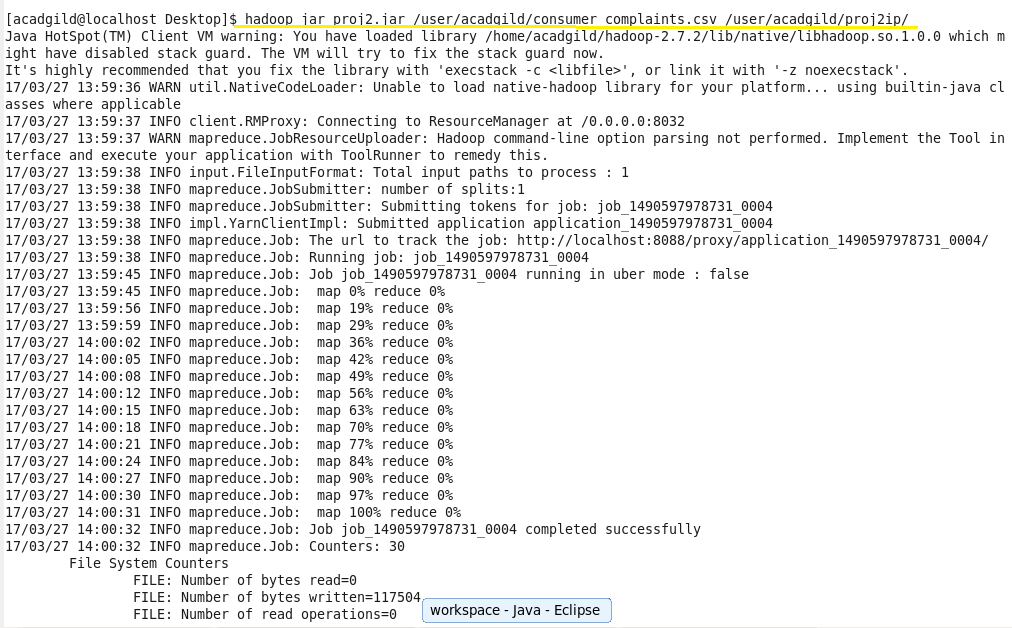
* In driver class job j is initiated to perform the task in order to solve the problem.
* The path arguments are set for input file and output files.
* In this class mapper class is set. The mapper class holds the logic to replace remove the commas present only inside double quotes.
* **Mapper class:**

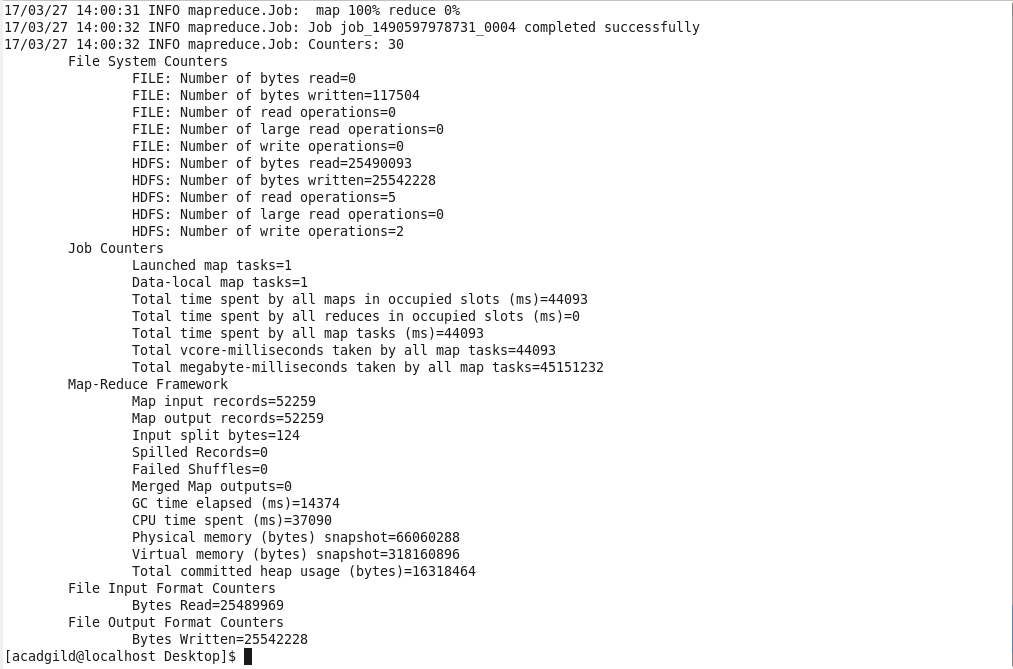
****

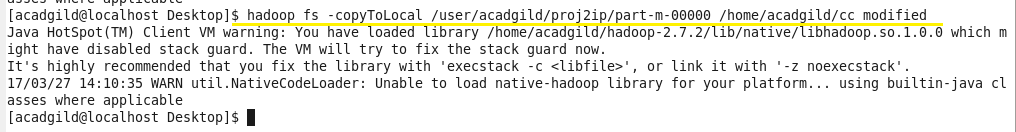
The mapper class is executed for each line in the input database.

* The input data given 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 split into many columns.
* To solve this problem we need to replace the commas between double quotes with some alternative symbol.
* In above code the commas which are enclosed in double quotes are replaced by ‘$’ while rest of the commas are not replaced.
* In this we check if the commas are present within the data enclosed by double quotes.
* If comma is present within double quotes then it is removed and replaced with ‘$’ and the modified data is stored in string ‘n’.

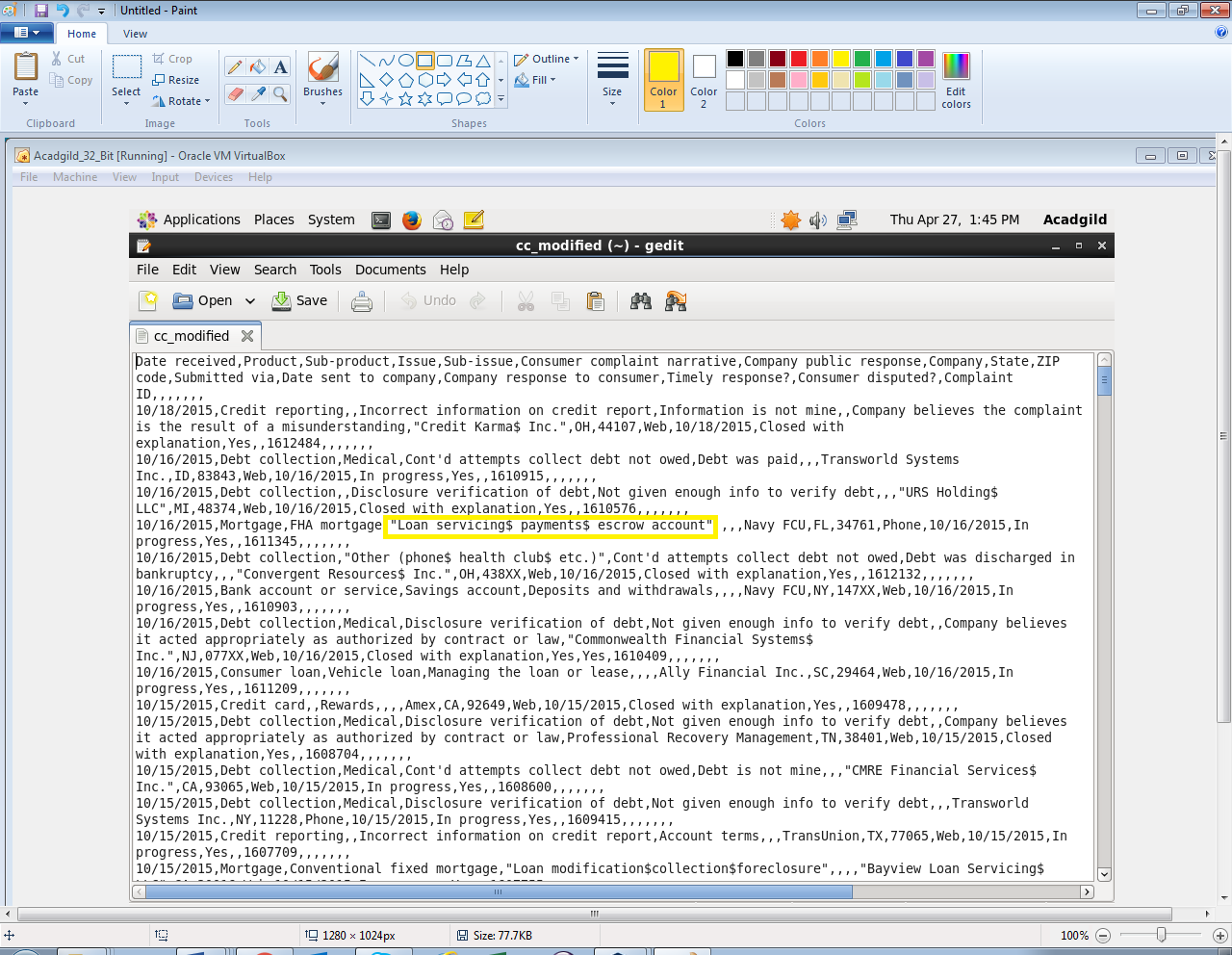
**5.2. OUTPUT:**

****

****

****

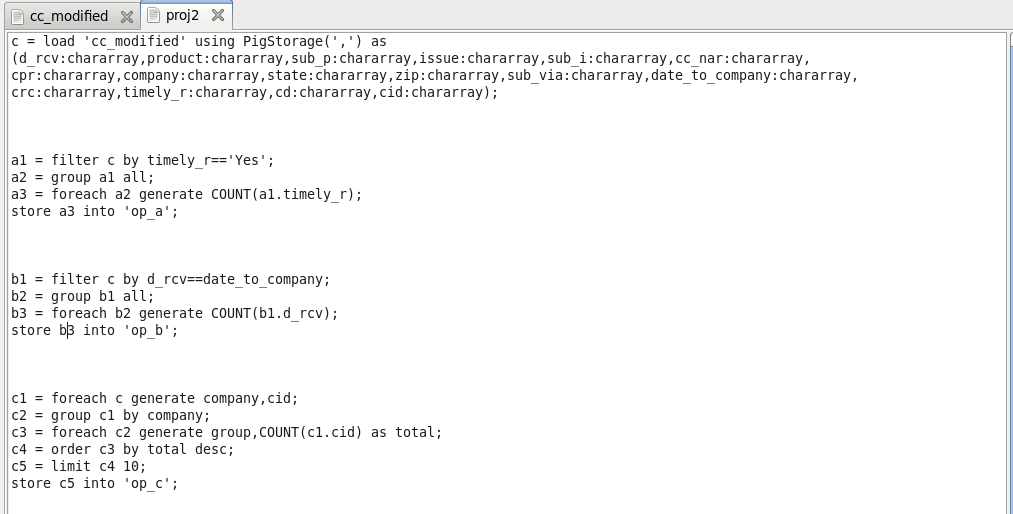
* The output is stored in a part file inside the folder named as “proj2\_ip”.
* The part file is then copied to the local system using “copyToLocal”.
* The file cc\_modified is the modified input file in which data enclosed by double quotes is separated by ‘$’.

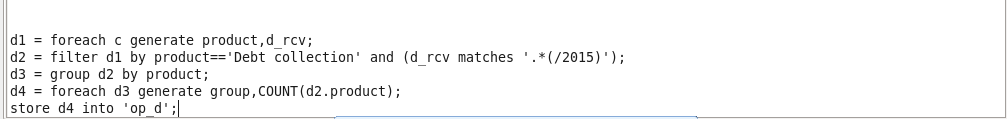


**6. Problem Statement 2:**

* **Write a pig script to find no of complaints which got timely response.**
* **Write a pig script to find no of complaints where consumer forum forwarded the complaint same day they received to respective company.**
* **Write a pig script to find list of companies topping in complaint chart (companies with maximum number of complaints).**
* **Write a pig script to find no of complaints filed with product type has "Debt collection" for the year 2015.**

**7. PIG SCRIPT:**

****

****

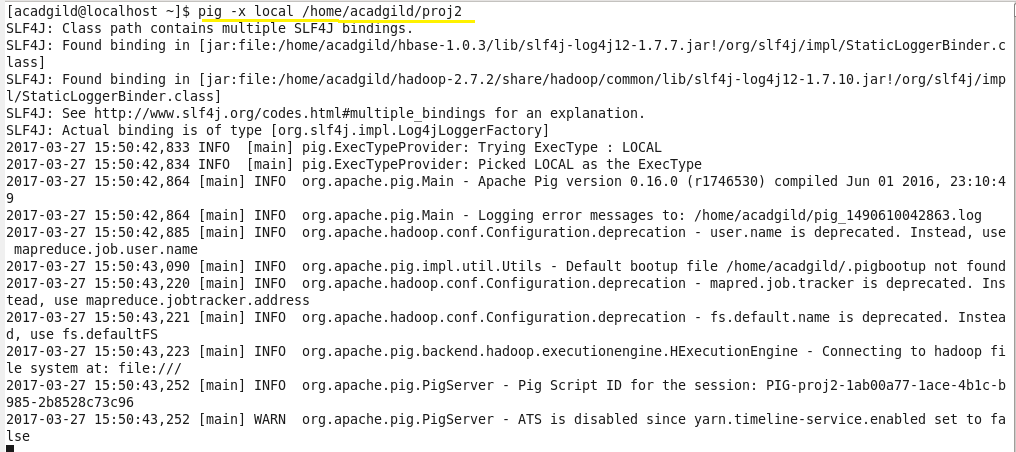
**7.1. DESCRIPTION FOR PIG SCRIPT:**

This script has solution for all the four problems in a sequence of problem statements.

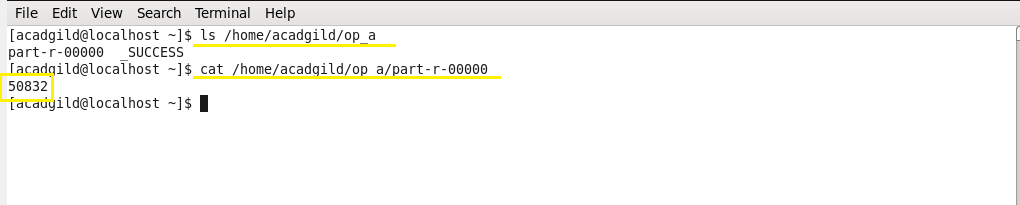
* c: All the contents of the input dataset are loaded in ‘c’. The structure of the input data is defined in here. As the data is separated using ‘,’ it is provided in PigStorage().
* **Solution for problem 1:**
* a1: The records are filtered by timely\_r where timely\_r is “yes” which indicates that the complaint got timely response.
* a2: The all records are then grouped.
* a3: The total count of records is calculated.
* op\_a: The output is stored in part file inside this folder.
* **Solution for problem 2:**
* b1: The records are filtered where consumer forum forwarded the complaint same day they received to respective company by comparing the d\_date and date\_to\_company fields.
* b2: The records are then grouped with respect to company.
* b3: The total count of records is calculated.
* op\_b: The output is stored in part file inside this folder.
* **Solution for problem 3:**
* c1: The records of required fields (company and cid) are loaded.
* c2: The records are grouped company wise.
* c3: The total count of records is calculated for each company.
* c4: The companies are arranged in descending order of the count of complaints.
* c5: The number of records to be stored or displayed is limited to 10.
* op\_c: The output is stored in part file inside this folder.
* **Solution for problem 4:**
* d1: The records of required fields (product and d\_rcv) are loaded.
* d2: The records are filtered where product field contains value as “Debt collection” and d\_rcv is of year 2015.
* d3: The records then are grouped by product.
* d4: The count of arrests is generated in this step.
* op\_d: The output is stored in part file inside this folder.

**7.2. OUTPUT:**

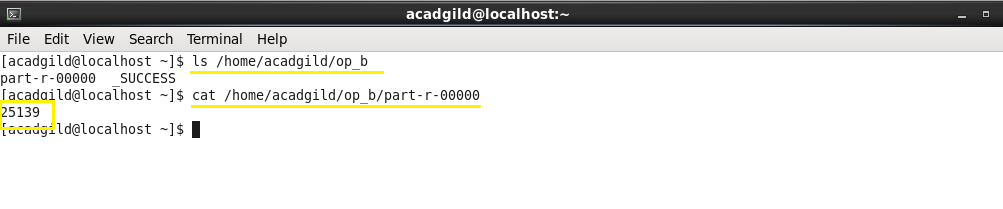
The pig script is executed using following command. The contents of the output files are displayed using cat command:

****

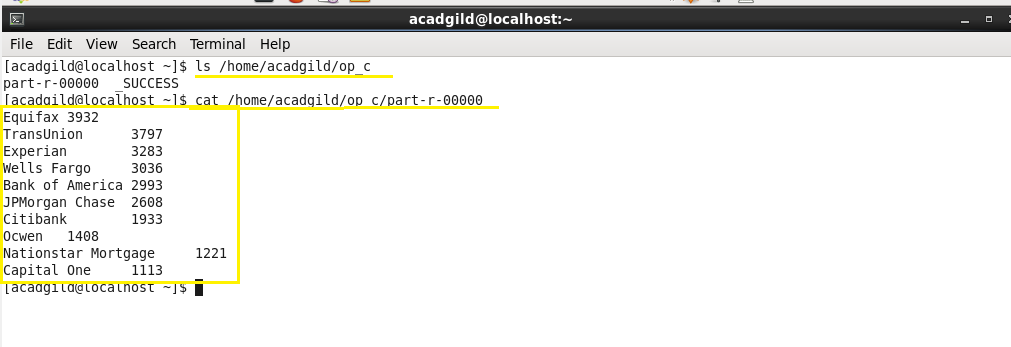
* **Output 1:**
* op\_1 contains part file which holds the number of complaints which got timely response.



* **Output 2:**
* op\_b has part file which contains the number of complaints where consumer forum forwarded the complaint same day they received to respective company.



* **Output 3:**
* op\_c contains the part file which has list of companies topping in complaint chart (companies with maximum number of complaints).



* **Output 4:**
* op\_d contains the part file which has the number of complaints filed with product type has "Debt collection" for the year 2015.

