**Mini Project 2 - Analyzing Consumer Complaints**

**Dataset Description:**

Date received, Product, Sub-product, Issue, Sub-issue, Consumer Complaint Narrative, Company Public Response, Company, State, ZIP code, Submitted via, Date sent to company, Company response to consumer, Timely response, Consumer disputed, Complaint ID.

**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.

**Task To Be Done:**

The entire column "Other (phone, health club, etc.)" should be product but is spitted by comma. If the file is spitted based on comma then this column will be spitted into 3 columns which will result in wrong outputs.

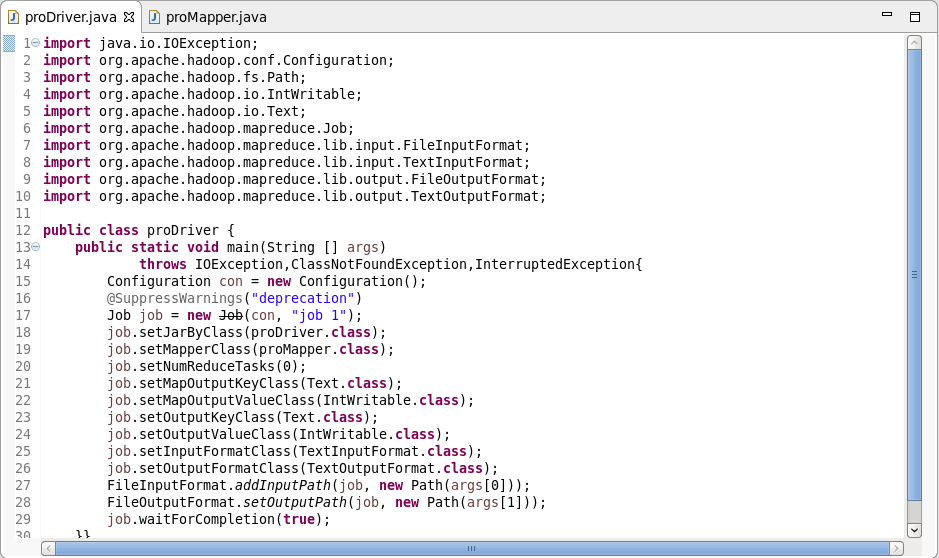
So in order to tackle this, the commas should be removed present only inside double quotes. Since Hadoop is used to handle big data it is used in java map reduce to remove unnecessary commas.

**Part 1:**

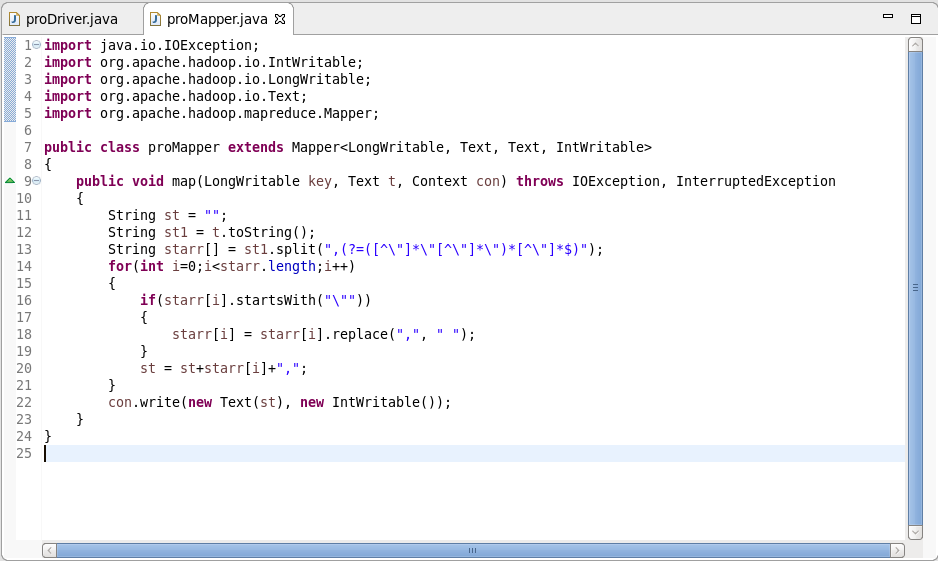
A MapReduce program is written to remove commas present inside the double quotes and replace them with space.

**MapReduce Program:**

**Driver Class:**

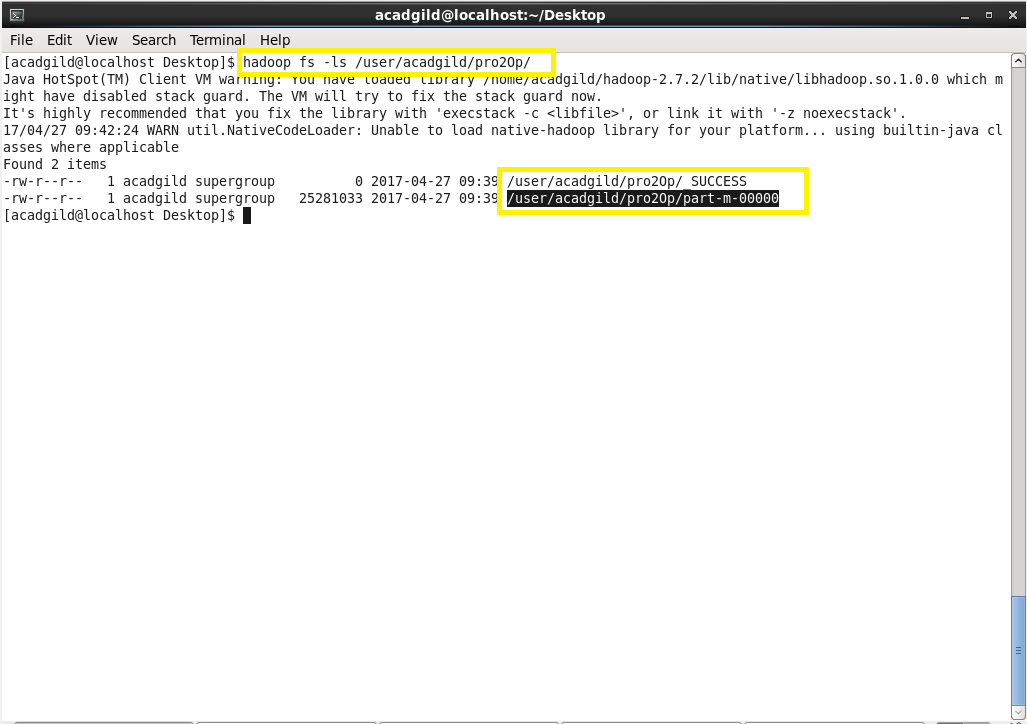


**Mapper Class:**



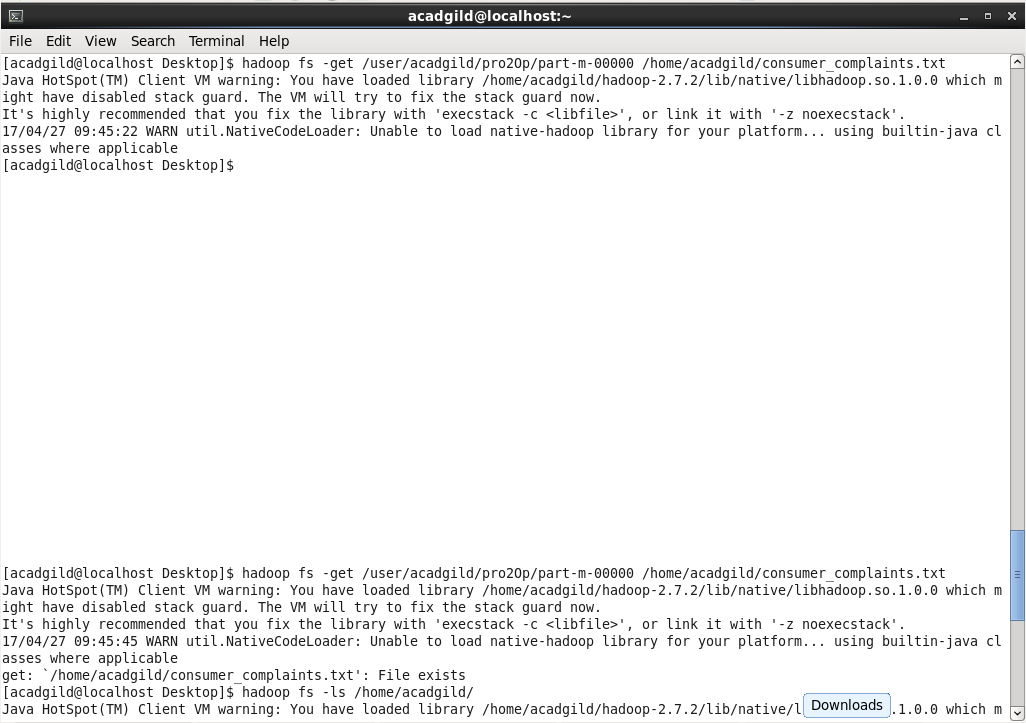
**Output showing success of the MapReduce program:**

1. By using hadoop fs –ls /user/acadgild/pro2Op/ the files inside the pro2Op can be listed.



**Saving output of MapReduce program into a Text File:**

1. By using -get command, we can copy the file from hdfs to local.
2. Part-m-00000 is copied to the local in the name of consumer\_complaints.txt with all the commas inside the double quotes replaced with space.



**Part 2: Pig Program:**

Using the output from MapReduce program, different Pig Command Sequences are executed to obtain different outputs.

**Pig programs to be executed to obtain different results:**

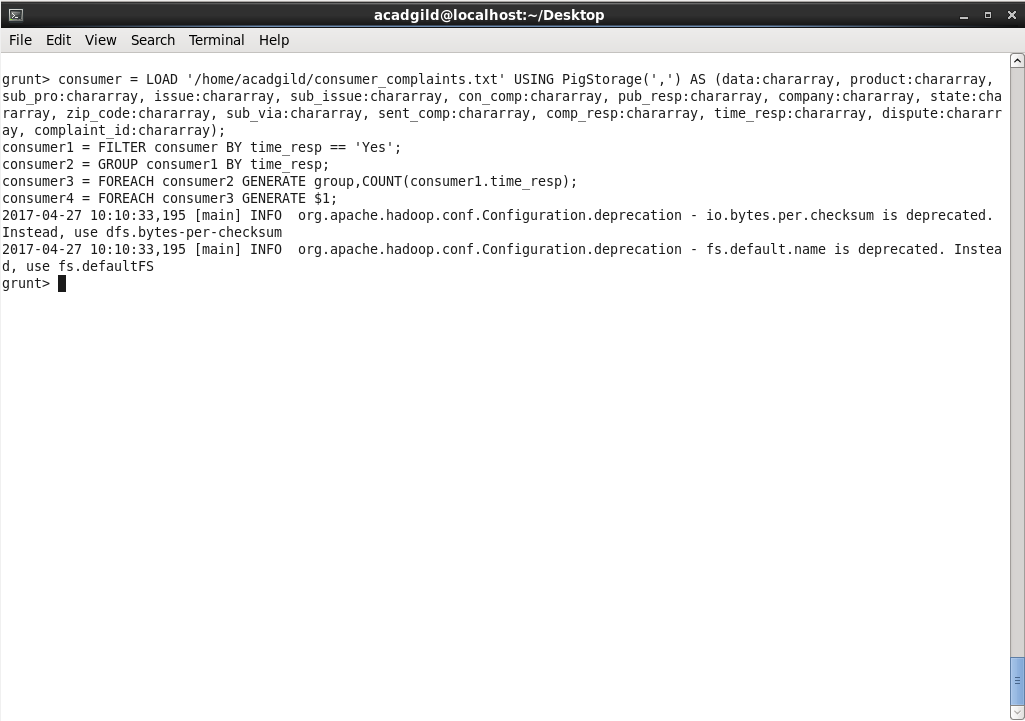
1. Write a pig script to find no of complaints which got timely response.
2. Write a pig script to find no of complaints where consumer forum forwarded the complaint same day they received to respective company
3. Write a pig script to find list of companies topping in complaint chart (companies with 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.

**NOTE:** The dataset used in the Pig Script is the output obtained by executing the MapReduce program.

**Program 1:**

Write a pig script to find number of complaints which got timely response.

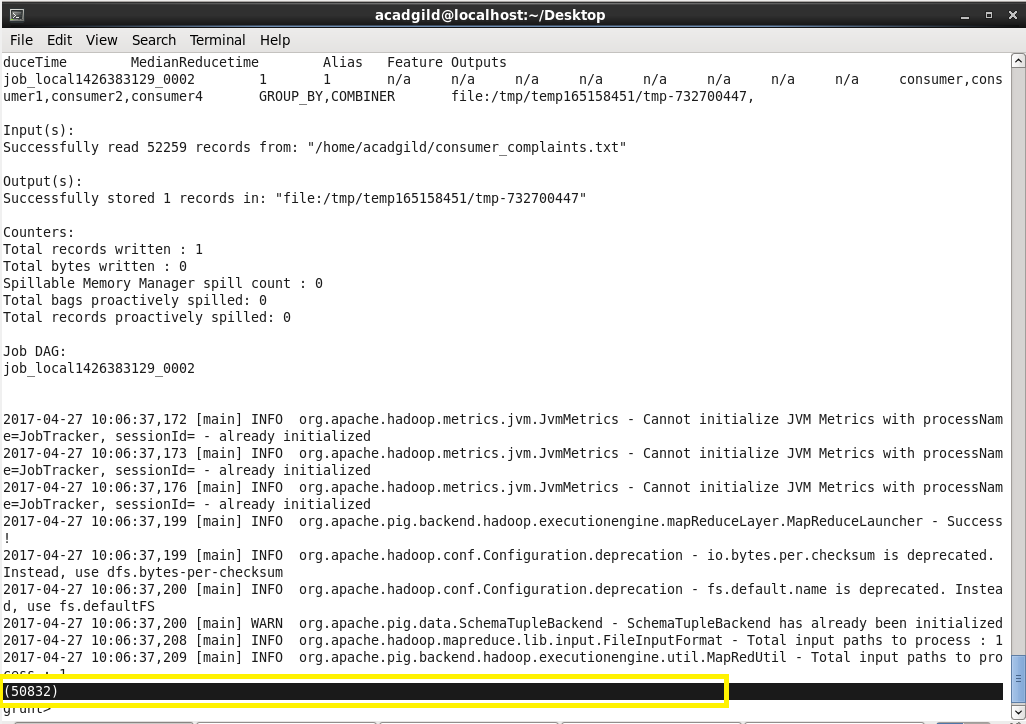
**Pig Command lines:**



**Logic:**

* Loading the dataset into the grunt shell using LOAD command.
* Filtering the dataset by checking the column 14(Timely Response) is ‘Yes’.
* Grouping the resultant dataset based on column 14(Timely Response) using Group By command.
* Counting the number of complaints after grouping using COUNT command.
  + At this stage o/p will be like (Yes, 50832).
* Generating only the count using GENERATE command by neglecting the ‘Yes’.

**Output:**



**Program 2:**

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

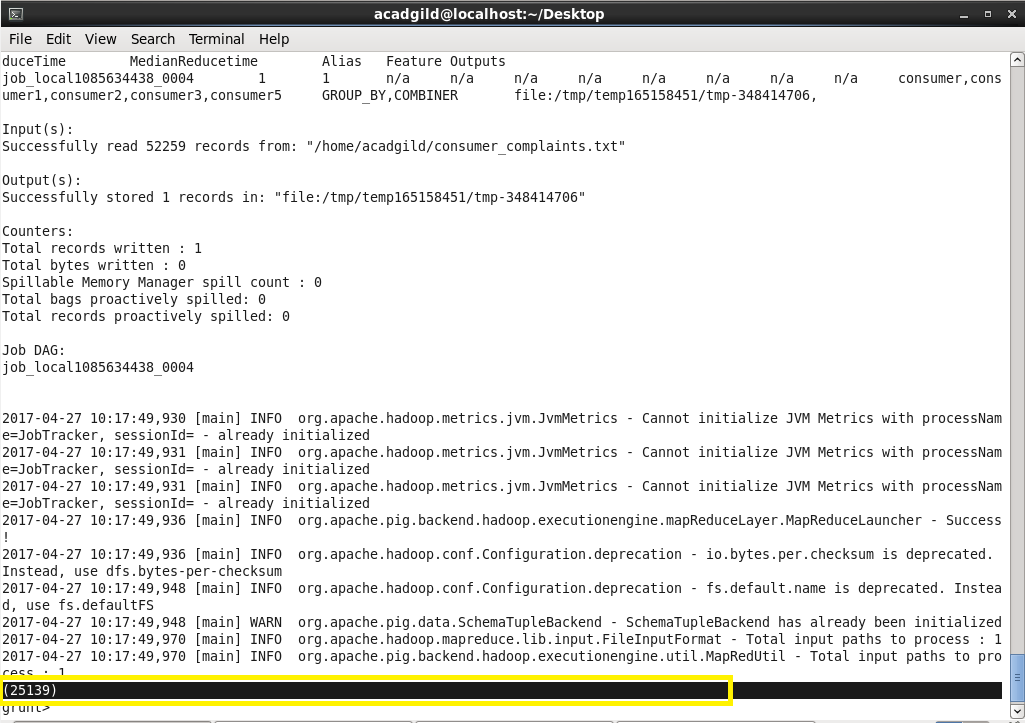
**Pig Command lines:**



**Logic:**

* Loading the dataset into the grunt shell using LOAD command.
* Filtering the dataset by checking whether column 1(Date received) and column 12(Date sent to company) are same.
* Generating only column 1(Date received) and column 12(Date sent to company) from the filtered dataset using GENERATE command.
* Grouping the generated dataset by all columns using GROUP ALL command.
* Generating the count using COUNT command and displaying it.

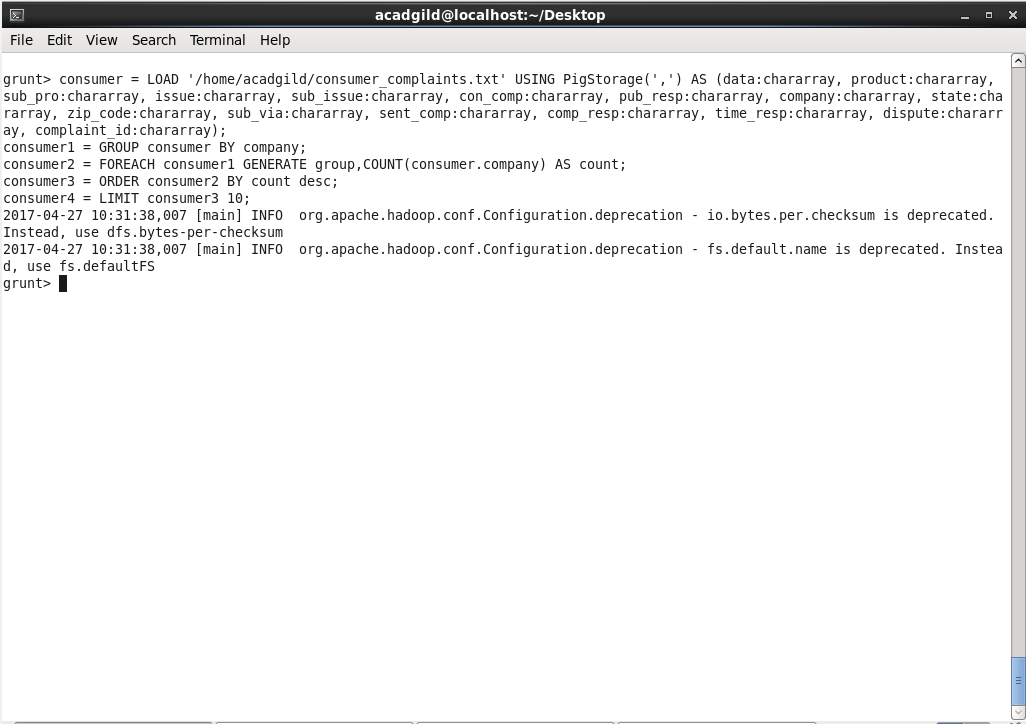
**Output:**



**Program 3:**

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

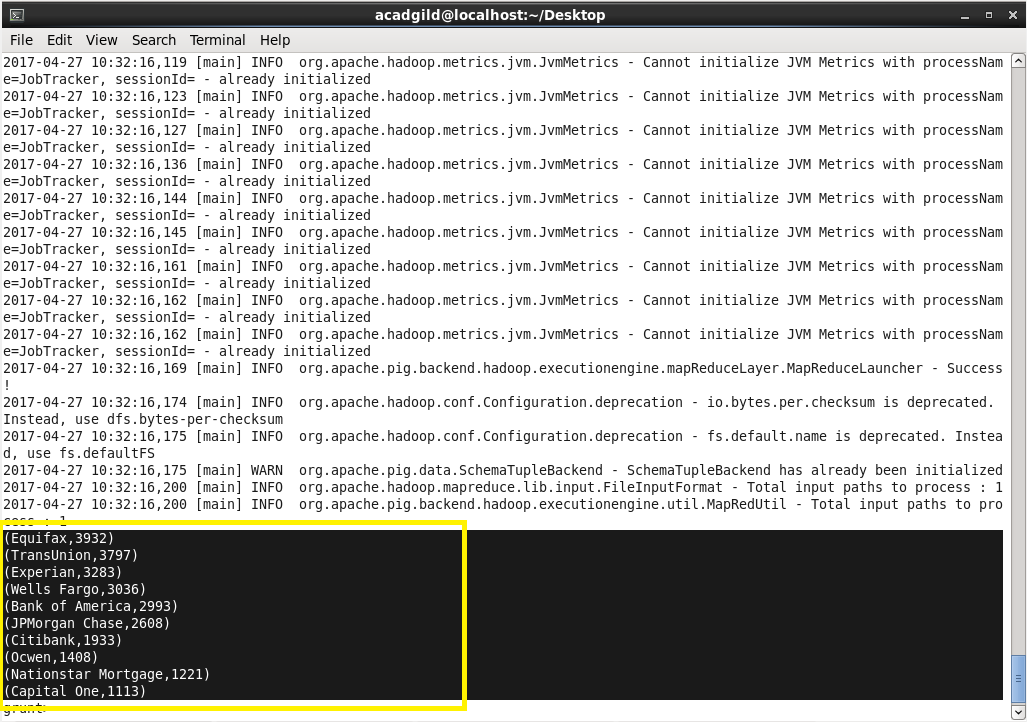
**Pig Command lines:**



**Logic:**

* Loading the dataset into the grunt shell using LOAD command.
* Grouping the dataset by the column Company name.
* Generating the count from the grouped dataset using COUNT command.
* Ordering the content based on the count in Descending order using ORDER BY command.
* Displaying only Top 10 companies using LIMIT command that is the companies that have maximum number of complaints from the consumers.

Output:



**Program 4:**

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

**Pig Command lines:**



**Logic:**

* Filtering the dataset by checking whether column 2(Product) contains ‘Debt collection’ using FILTER command.
* Generating only the year (yyyy) from the column ‘Date received’ using GENERATE and SUBSTRING command.
* Filtering the processed dataset for year =’2015’.
* Grouping the dataset by the column Product for ‘Debt collection’ using GROUP BY command.
* Generating the count using COUNT command.
* Displaying only the count as output.

**Output:**

