**Project 2**

**USA Consumer Forum Data Analysis**

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.

Below is the description of the data set

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

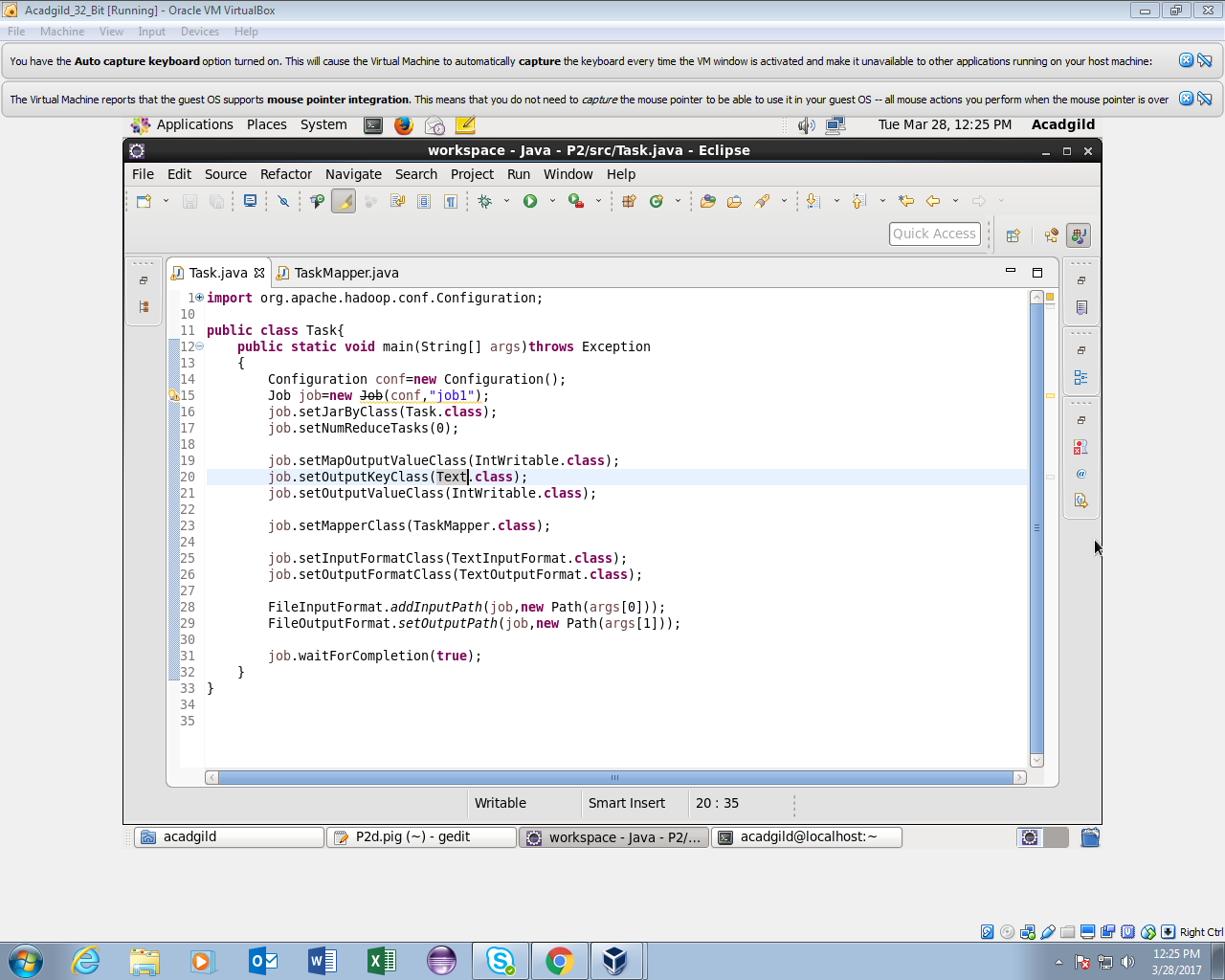
This data is comma delimited. In some rows there are few columns which are enclosed in double quotes and have many commas and due to this the same column gets spitted into many columns for ex: Sample record: 10/16/2015,Debt collection,"Other (phone, health club, etc.)",Cont'd attempts collect debt not 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.

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.

**Driver class**

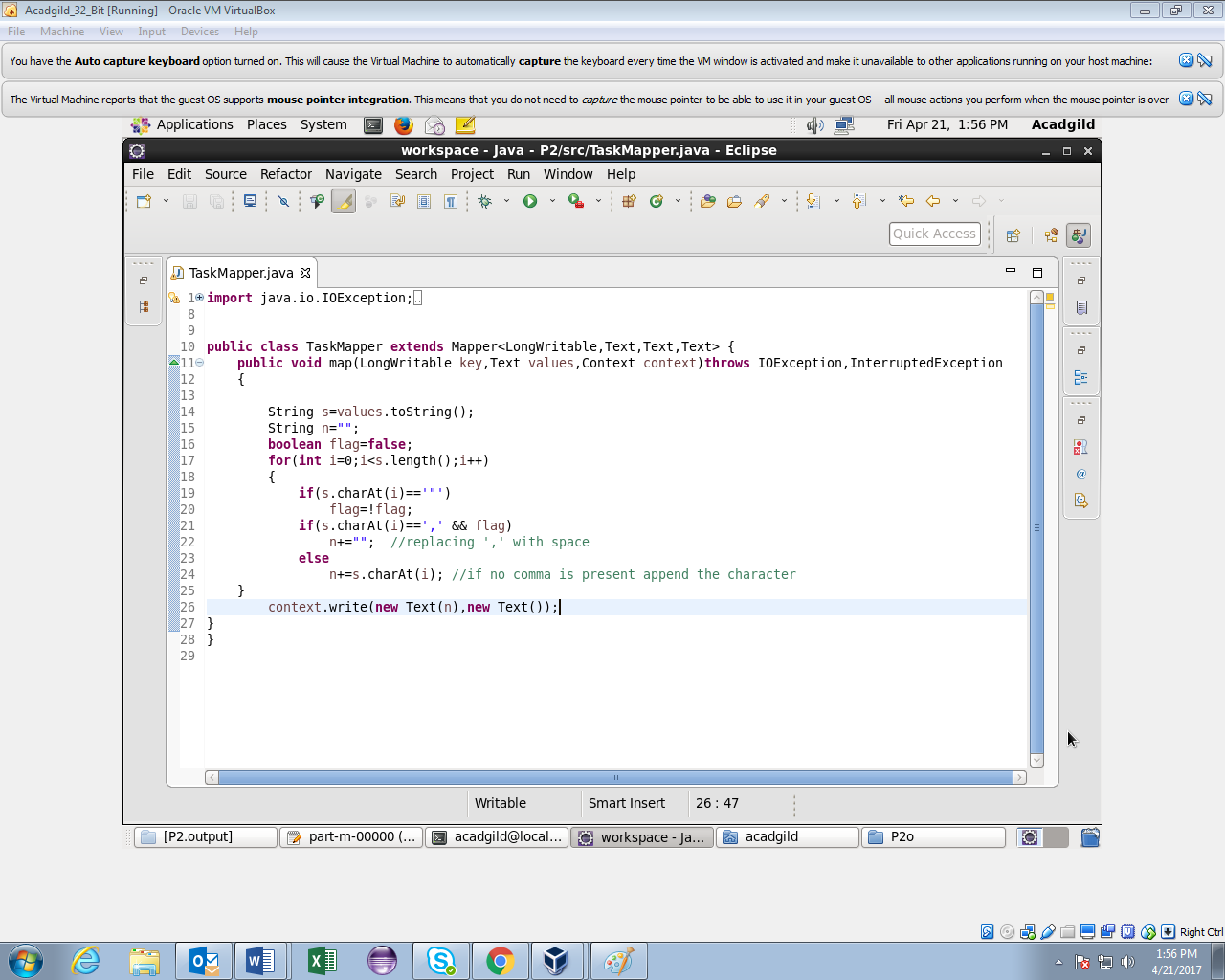


**Mapper class**

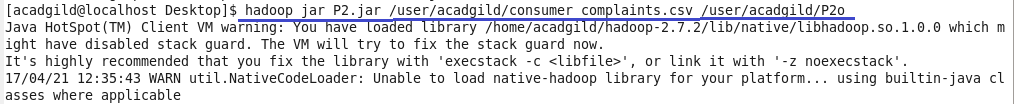
**Logic**

A boolean variable flag is set as false ansd it reverses its state wherever there is double quote.

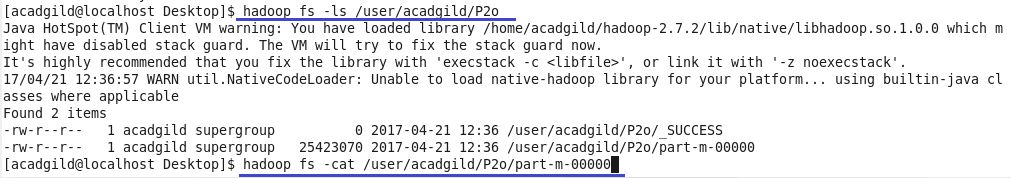
If there is a comma enclosed in double quotes and flag is set to true,comma is replaced by space.Else if there is no comma,character will be appended as it is.

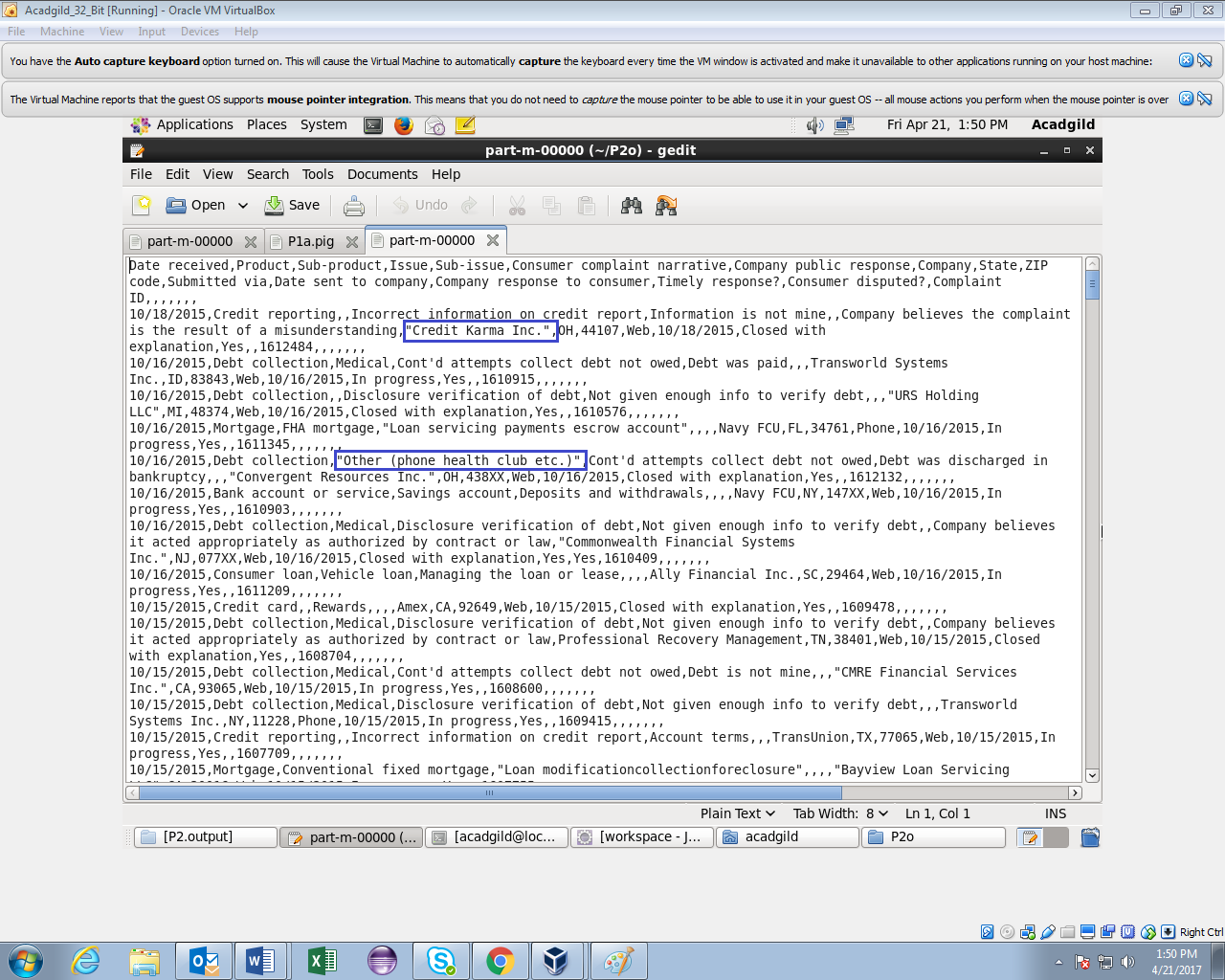


**Running jar**



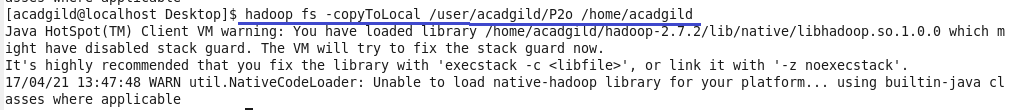
**Output**





The comma within double quotes is replaced by space in the highlighted parts of the output.

The copyToLocal command is used to copy the output file to the local file system after removing the commas between double quotes to perform pig script operations.



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

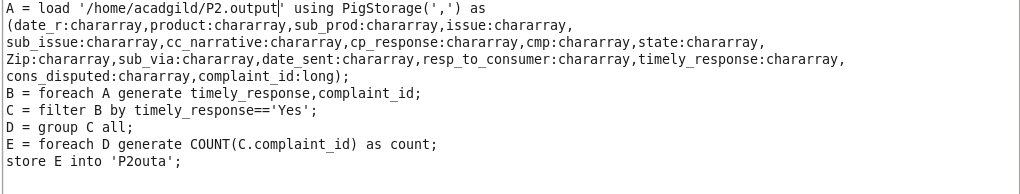
**Logic:**

1.Loaded the output file after removing commas.

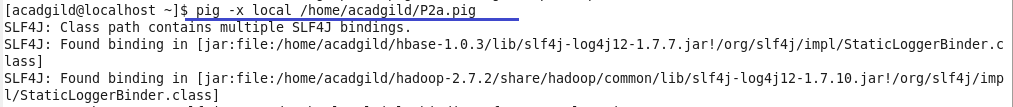
2.filtered complaints for which timely response is yes.

3.grouped by all and made a count of complaint id.

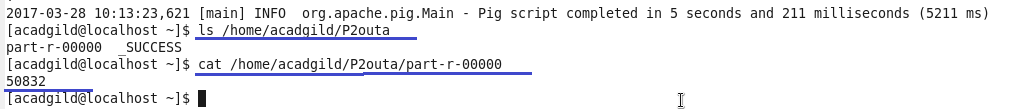
**Pig Script**



**Running Pig Script**



**Output**



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

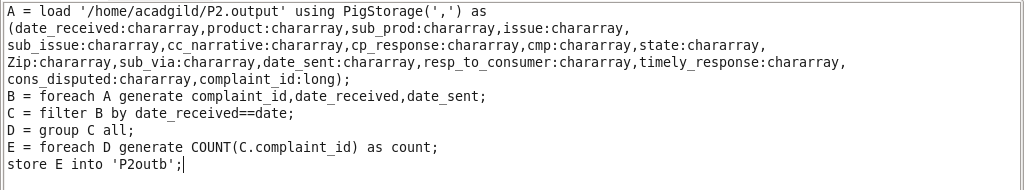
**Logic**

1.Loaded the output file.

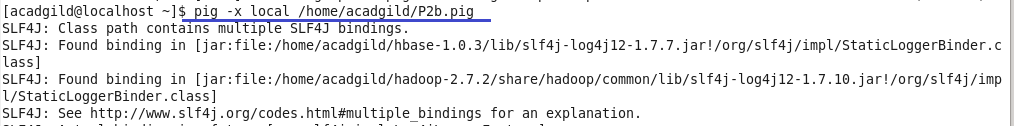
2.filtered the complaints for which date received=date sent to company.

3.grouped by all and made count by complaint id.

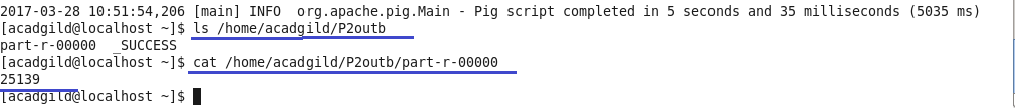
**Pig Script**



**Running Pig Script**



**Output**



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

**Logic:**

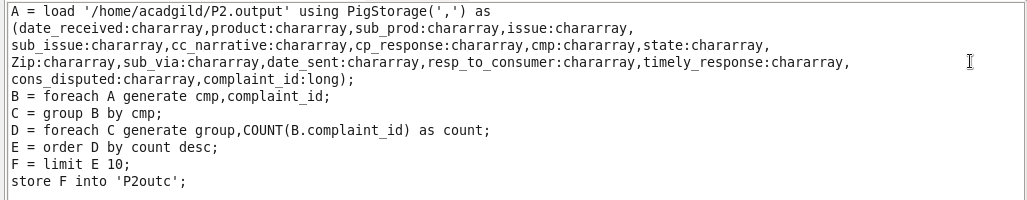
1.Loaded the output file.

2.grouped by company name.

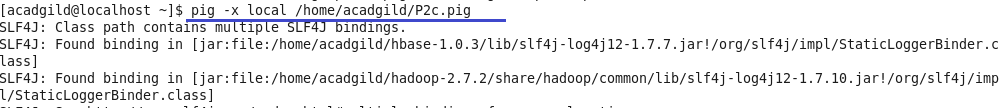
3.made count of complaint id for each company

4.ordered by descending order of count and limited them by 10 to generate list of top 10 companies.

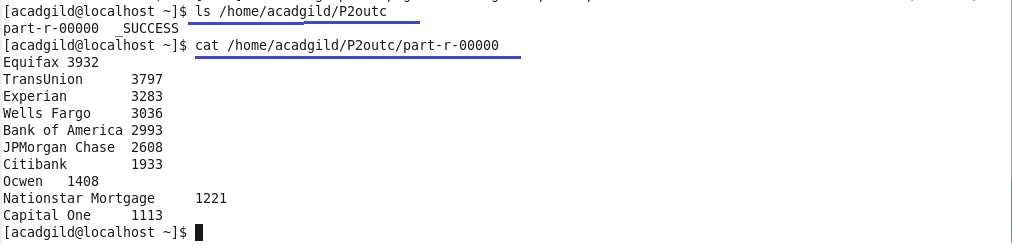
**Pig Script**



**Running Pig Script**



**Output**



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

**Logic:**

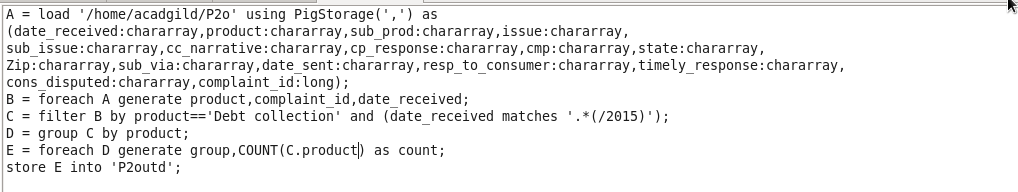
1.loaded the output file.

2.filtered the complaints for which product type =Debt collection and used matches function to filter date with year 2015.

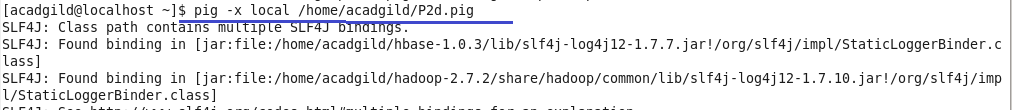
3.grouped by product type.

4.made count of product type.

**Pig Script**



**Running Pig Script**



**Output**

