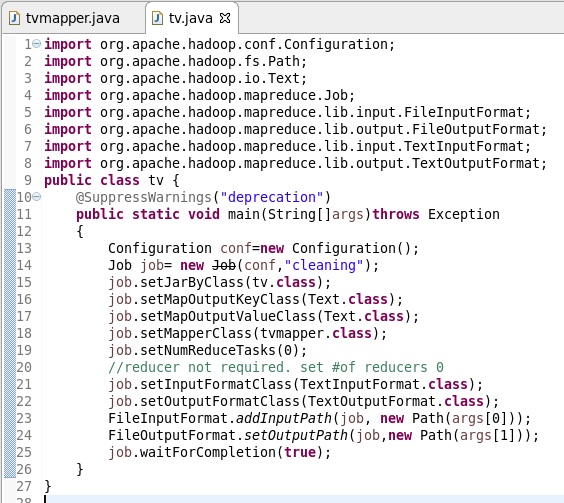
**Project 2:**

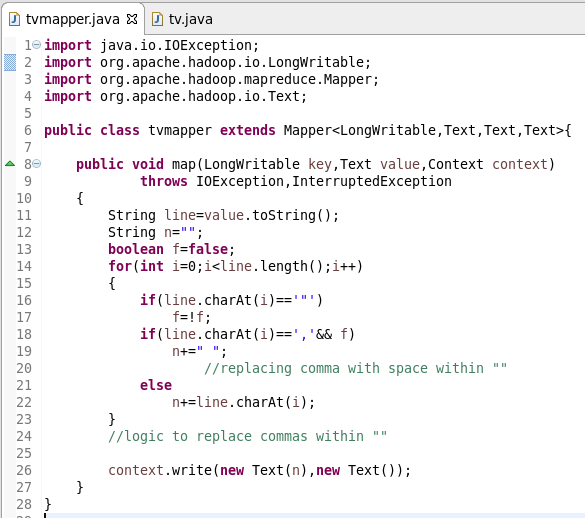
Part I:

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.

Task class:



Map class:



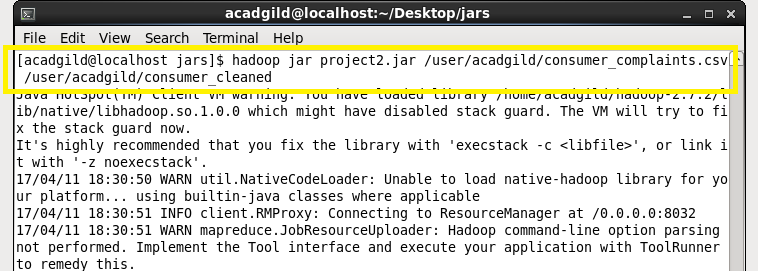
Mapper class is used to generate key and value pair where key is the text i.e. the lines which are read by recordreader and value is null.

Mapper is required to eliminate unnecessary commas which are present within “ “.

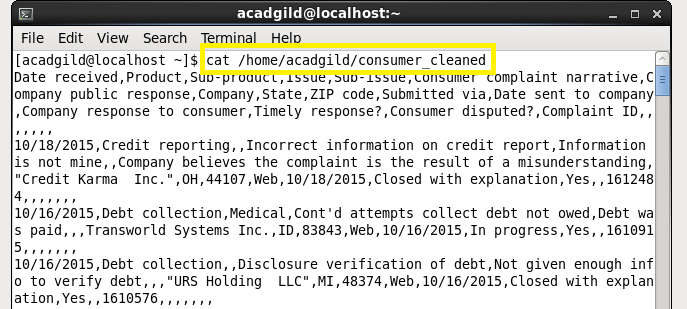
The simple for loop is used to traverse through the line.

Inside for loop the logic is written to replace commas present in “ “ with single space.

Running jar:



Output of jar (cleaned file):



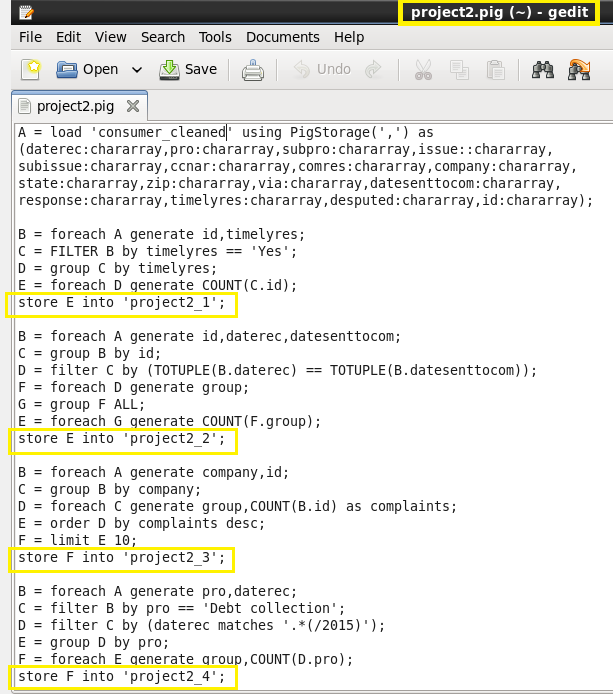
This file ‘consumer\_cleaned’ is the file which contains the data in required format.

It is given as input to the pig script and rest of the data processing is done on this file.

Part 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

Pig script:



Problem 1:

* We generated B such that it contains ID and response details.
* Filter B such that C contains records those have timely response (timelyres==’Yes’).
* We grouped C by timelyres so now all the ids belongs to same group.
* Now take the count of all IDs.

Problem 2:

* Generate B such that it contains id,daterec and datesenttocom.
* Group B by id
* Filter C such that it contains records where daterec and datesenttocom are same.
* Generate F such that it contains ids from D.
* Group F by all and take count of ids

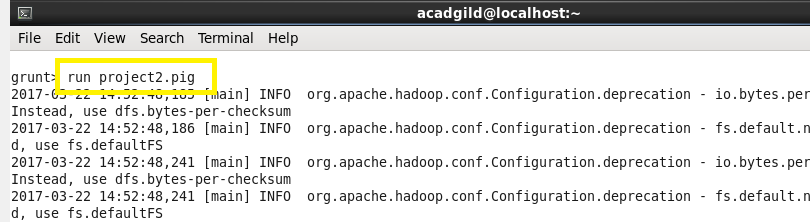
Prolem 3:

* Generate B such that it contains company and id.
* Group B by id so that we will get all case ids for each company in C.
* Now take count of cases and arrange them in descening order.
* Limit E to 10 so that F contains top 10 companies.

Problem 4:

* Create B such that is contains pro and daterec.
* Filter B such that it contains all records where pro is ‘debt collection’ and daterec is in year 2015 (daterec matches ‘.\*(/2015)’)
* Group D by pro so that all cases are in same group
* Take count of records belonging to the group.

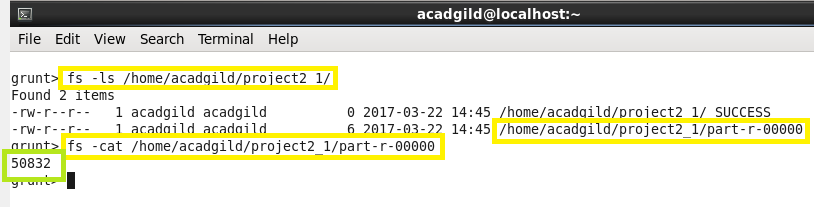
Running pig script:



We are running the .pig extension file on grunt shell in which the script is written.

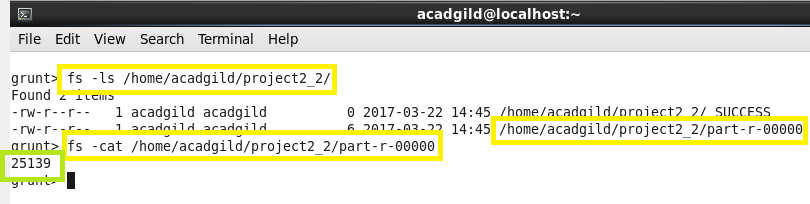
**Outputs:**

**Problem 1:**



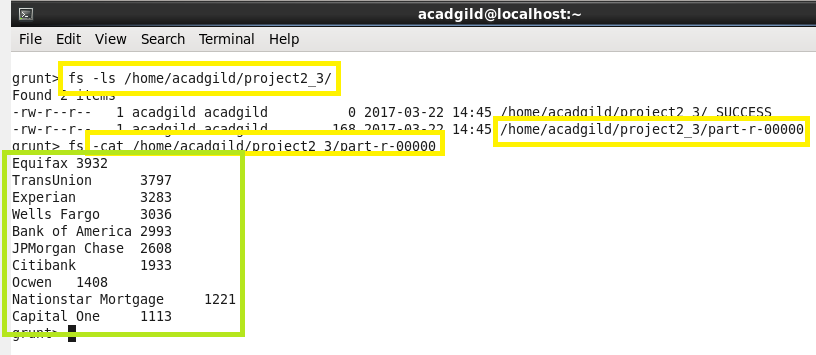
There are total 50832 complaints which got timely response.

**Problem 2:**



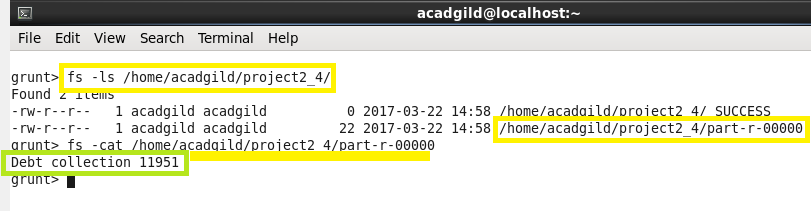
Total 25139 complaints were forwarded on the same day they were registered.

**Problem 3:**



These are the top 10 companies who have maximum number of complaints.

**Problem 4:**



Total 11951 complaints were filed for the product type ‘debt collection’ in year ‘2015’.