**USA Consumer Forum Data Analysis**

**Problem Statement**

Note: You need to copy the dataset into HDFS using Flume and the results of the problem statements should be exported into RDBMS(Mysql) using sqoop.

The aim of this project is to analyze performance of various companies on aspects like:

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 toping in complaint chart (companies with maximum number of complaints)

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

**Solution**

flume-ng agent -n agent1 -f /usr/local/flume/conf/LocalToHdfs\_Proj2.conf

**Problem Statement 1:**

A = LOAD 'flume\_import\_Proj2/FlumeData.1469729870468' USING PigStorage('\t') AS (f1:chararray, f2:chararray, f3:chararray, f4:chararray, f5:chararray, f6:chararray, f7:chararray, f8:chararray, f9:chararray, f10:chararray, f11:chararray, f12:chararray, f13:chararray, f14:chararray, f15:chararray, f16:chararray);

B = FILTER A BY f14 == 'Yes';

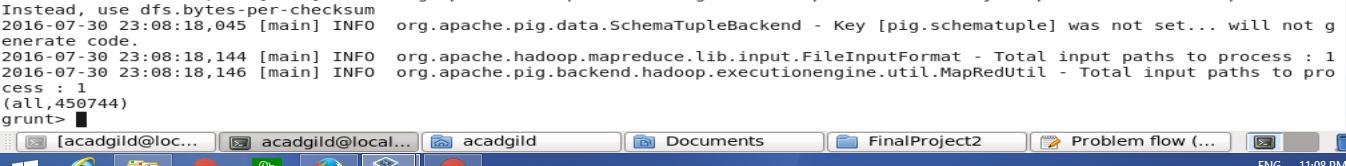
C = GROUP B ALL;

D = FOREACH C GENERATE group, COUNT(B.f14);

DUMP D;

Output:

(all,450744)



**Problem Statement 2:**

A = LOAD 'flume\_import\_Proj2/FlumeData.1469729870468' USING PigStorage('\t') AS (f1:chararray, f2:chararray, f3:chararray, f4:chararray, f5:chararray, f6:chararray, f7:chararray, f8:chararray, f9:chararray, f10:chararray, f11:chararray, f12:chararray, f13:chararray, f14:chararray, f15:chararray, f16:chararray);

F = FILTER A BY f1 == f12;

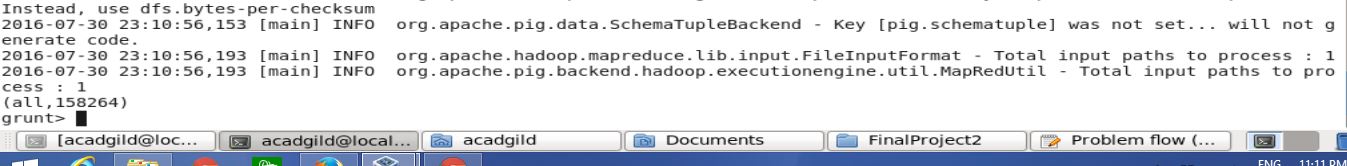
G = GROUP F ALL;

H = FOREACH G GENERATE group, COUNT(F.f1);

DUMP H;

Output:

(all,158264)



**Problem Statement 3:**

A = LOAD 'flume\_import\_Proj2/FlumeData.1469729870468' USING PigStorage('\t') AS (f1:chararray, f2:chararray, f3:chararray, f4:chararray, f5:chararray, f6:chararray, f7:chararray, f8:chararray, f9:chararray, f10:chararray, f11:chararray, f12:chararray, f13:chararray, f14:chararray, f15:chararray, f16:chararray);

I = FOREACH A GENERATE f8, f16;

J = FILTER I BY NOT(f8 is NULL);

K = GROUP J BY f8;

L = FOREACH K GENERATE group, COUNT(J.f16);

M = ORDER L BY $1 DESC;

N = LIMIT M 10;

DUMP N;

Output:

(Bank of America,50722)

(Wells Fargo,36782)

(JPMorgan Chase,29330)

(Equifax,24560)

(Experian,24527)

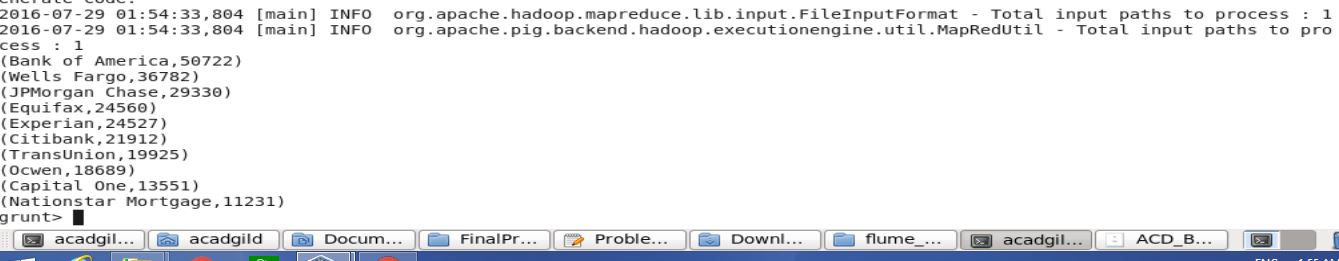
(Citibank,21912)

(TransUnion,19925)

(Ocwen,18689)

(Capital One,13551)

(Nationstar Mortgage,11231)



**Problem Statement 4:**

A = LOAD 'flume\_import\_Proj2/FlumeData.1469729870468' USING PigStorage('\t') AS (f1:chararray, f2:chararray, f3:chararray, f4:chararray, f5:chararray, f6:chararray, f7:chararray, f8:chararray, f9:chararray, f10:chararray, f11:chararray, f12:chararray, f13:chararray, f14:chararray, f15:chararray, f16:chararray);

O = FILTER A BY f2 == 'Debt collection';

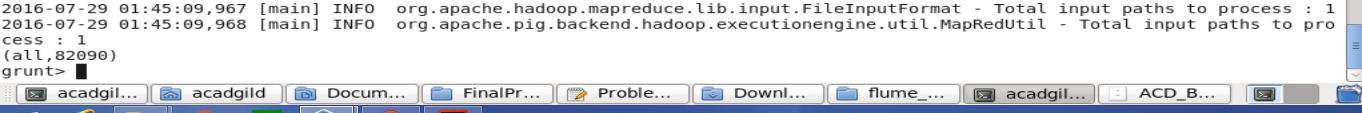
P = GROUP O ALL;

Q = FOREACH P GENERATE group, COUNT(O.f2);

DUMP Q;

Output:

(all,82090)

****