*Project 2 – USA Consumer Forum Data Analysis*

*By: Krushna Chandra Debashram*

*21-Aug-2017*

## Introduction:

The file ‘Consumer\_Complaints.csv’ has in csv format and contains the attributes pertaining to resolution of the consumer complaints. It contains data points as below:

**Dataset Description:**

|  |  |  |  |
| --- | --- | --- | --- |
| Sl.No. | Column Heading | Index | Description |
| 1 | Date received | 0 | date on which consumer filed the complaint |
| 2 | Product | 1 | Type of the product |
| 3 | Sub-product | 2 | Sub product type |
| 4 | Issue | 3 | Issue faced by the consumer |
| 5 | Sub-issue | 4 | Any sub issues if exists |
| 6 | Consumer complaint narrative | 5 | Detailed description of complaint |
| 7 | Company public response | 6 | Company’s public response to the complaint |
| 8 | Company | 7 | Name of the company |
| 9 | State | 8 | State from which consumer filed the complaint |
| 10 | ZIP code | 9 | Zip code |
| 11 | Submitted via | 10 | Channel from which complaint was submitted |
| 12 | Date sent to company | 11 | Date on which consumer forum forwarded the complaint to company |
| 13 | Company response to consumer | 12 | Company’s response to the consumer |
| 14 | Timely response? | 13 |  |
| 15 | Consumer disputed? | 14 |  |
| 16 | Complaint ID | 15 | Unique complaint id |

## Objective:

1. Load the data to HDFS using Flume.
2. Analysis the data using Pig. Store the result in HDFS file.
3. Export the result in HDFS to MySQL RDBMS using Sqoop.

Following to analysis from consumer data using Pig.

1. Find number of complaints which got timely response
2. Find number of complaints where consumer forum forwarded the complaint same day they received to respective company
3. Find list of companies toping in complaint chart (companies with maximum number of complaints)
4. Find no of complaints filed with product type has “ Debt collection” for the year 2015

## Prerequisites:

1. Hadoop cluster installed in the system.
2. Flume
3. Sqoop
4. MySQL

## Program used:

Flume : To load the csv in HDFS

Pig : Analyses the data

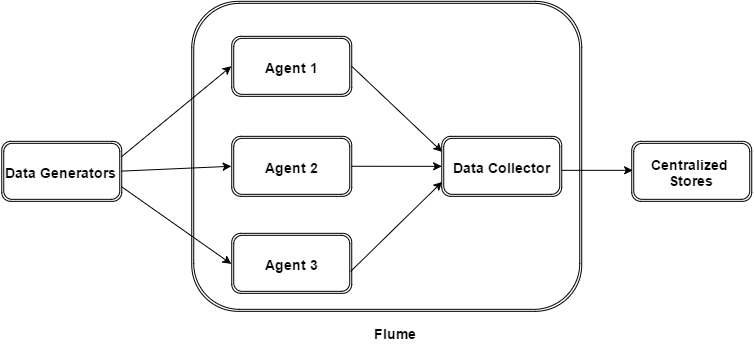
Sqoop : Export the result of Pig analysis store in HDFS file

### Flume

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows.

Architecture of Flume is as below:

**Data generators** (such as Facebook, Twitter) generate data which gets collected by individual Flume **agents** running on them. Thereafter, a **data collector** (which is also an agent) collects the data from the agents which is aggregated and pushed into a centralized store such as HDFS or HBase.



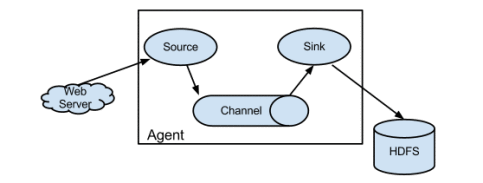
Flume Event

It is the basic unit of the data transported inside **Flume**. It contains a payload of byte array that is to be transported from the source to the destination accompanied by optional headers. A typical Flume event would have the following structure:



Flume Agent

It is an independent daemon process (JVM) in Flume. It receives the data (events) from clients or other agents and forwards it to its next destination (sink or agent).



Agent has three main parts

1. Source
2. Channel
3. Sink

Source

A **source** is the component of an Agent which receives data from the data generators and transfers it to one or more channels in the form of Flume events.

**Example**− Avro source, Thrift source, twitter 1% source etc.

Channel

A **channel** is a transient store which receives the events from the source and buffers them till they are consumed by sinks. It acts as a bridge between the sources and the sinks.

**Example** − JDBC channel, File system channel, Memory channel, etc.

Sink

A **sink** stores the data into centralized stores like HBase and HDFS. It consumes the data (events) from the channels and delivers it to the destination. The destination of the sink might be another agent or the central stores.

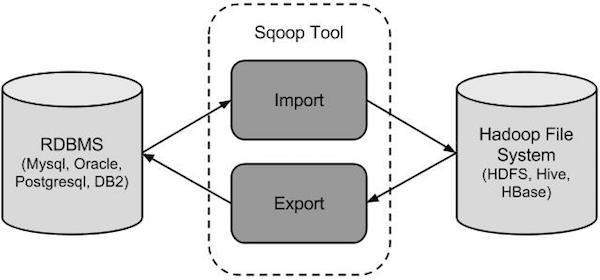
Example − HDFS sink

### Sqoop

Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. This is a brief tutorial that explains how to make use of Sqoop in Hadoop ecosystem.

It allows users to specify the target location inside of Hadoop and instruct Sqoop to move data from Oracle, Teradata or other relational databases to the target.

The following image describes the workflow of Sqoop.



**Sqoop Import**

The import tool imports individual tables from RDBMS to HDFS. Each row in a table is treated as a record in HDFS. All records are stored as text data in text files or as binary data in Avro and Sequence files.

**Sqoop Export**

The export tool exports a set of files from HDFS back to an RDBMS. The files given as input to Sqoop contain records, which are called as rows in table. Those are read and parsed into a set of records and delimited with user-specified delimiter.

### Pig

Apache Pig is an abstraction over MapReduce. It is a tool/platform which is used to analyze larger sets of data representing them as data flows. Pig is generally used with **Hadoop** to perform all the data manipulation operations.

## Start Hadoop daemon:

**[**acadgild@localhost ~]$ start-all.sh

[acadgild@localhost ~]$ mr-jobhistory-daemon.sh start historyserver

[acadgild@localhost ~]$ jps

3122 ResourceManager

3301 Jps

3669 JobHistoryServer

2966 SecondaryNameNode

3256 NodeManager

2733 NameNode

2830 DataNode

## Start MySql :

[acadgild@localhost ~]$ sudo service mysqld start

Starting mysqld: [ OK ]

## Create DB in MySQL

[acadgild@localhost ~]$ mysql -u kduser -p

mysql>Create Database Test;

## Create a directory in Hadoop:

hadoop dfs -mkdir /user/krushnadebashram\_project2\_2\_ConsumerComplaint

hadoop dfs -mkdir /user/krushnadebashram\_project2\_2\_ConsumerComplaint/Source\_ConsumerComplaint

hadoop dfs -mkdir /user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result

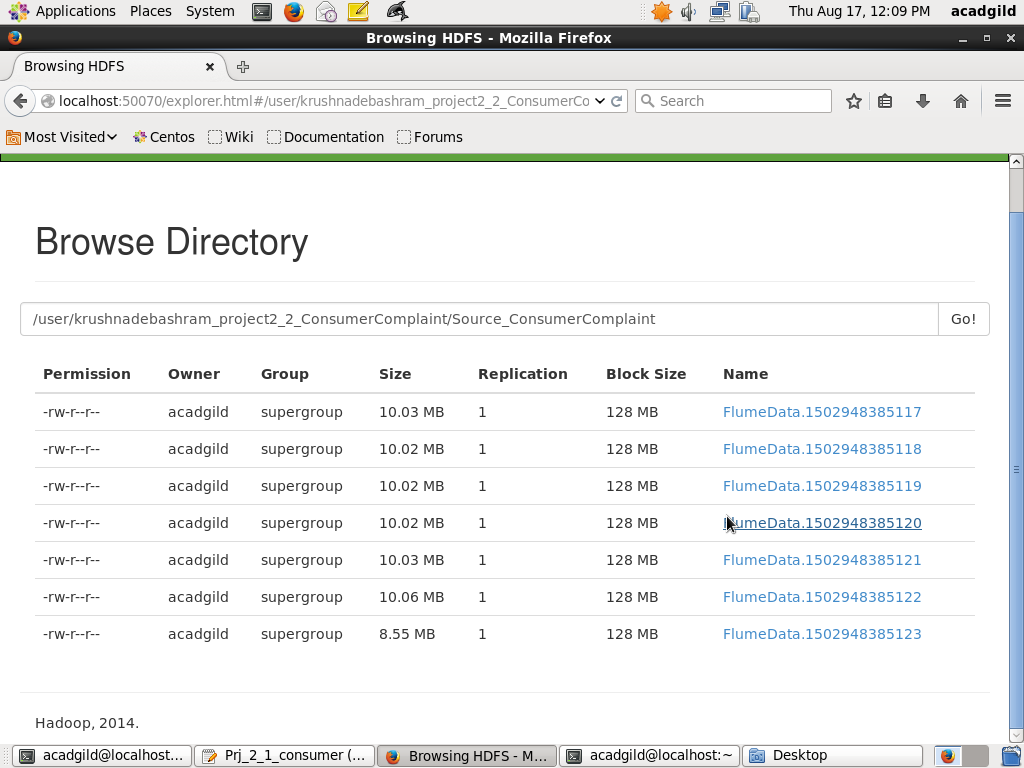
## Create Flume Conf file:

Create a conf file flume\_ConsumerComplaints.CONF

|  |
| --- |
| ConsumerComplaintAgent.channels = channel1  ConsumerComplaintAgent.channels.channel1.type=File  ConsumerComplaintAgent.channels.channel1.capacity = 50000  ConsumerComplaintAgent.channels.channel1.transactionCapacity=10000  ConsumerComplaintAgent.channels.channel1.byteCapacity = 644245  ConsumerComplaintAgent.channels.channel1.byteCapacityBufferPercentage = 10  # Source definition  ConsumerComplaintAgent.sources = source1  ConsumerComplaintAgent.sources.source1.channels = channel1  ConsumerComplaintAgent.sources.source1.type = exec  ConsumerComplaintAgent.sources.source1.command = head -n -1 /home/acadgild/KrushnaDebashram/Consumer\_Complaints.csv  #sink definition  ConsumerComplaintAgent.sinks=sink1  ConsumerComplaintAgent.sinks.sink1.channel=channel1  ConsumerComplaintAgent.sinks.sink1.type=hdfs  ConsumerComplaintAgent.sinks.sink1.hdfs.useLocalTimeStamp = true  ConsumerComplaintAgent.sinks.sink1.hdfs.path = /user/krushnadebashram\_project2\_2\_ConsumerComplaint/Source\_ConsumerComplaint  ConsumerComplaintAgent.sinks.sink1.hdfs.writeFormat=Text  ConsumerComplaintAgent.sinks.sink1.hdfs.fileType = DataStream  ConsumerComplaintAgent.sinks.sink1.hdfs.batchSize=10000  ConsumerComplaintAgent.sinks.sink1.hdfs.rollSize=10485760  ConsumerComplaintAgent.sinks.sink1.hdfs.rollCount=0  ConsumerComplaintAgent.sinks.sink1.hdfs.rollInterval=0 |

## Upload the file in Hadoop using flume:

flume-ng agent -n ConsumerComplaintAgent -c conf -f /usr/lib/apache-flume-1.4.0-bin/conf/flume\_ConsumerComplaints.Conf

****

## Go to Pig shell:

Go to pig shell in mapreduce mode

[acadgild@localhost ~]$ pig -x mapreduce

grunt>

**Register the piggybank jars in pig:**

|  |
| --- |
| grunt>REGISTER /usr/local/pig/lib/piggybank.jar; |

## Problem 1:

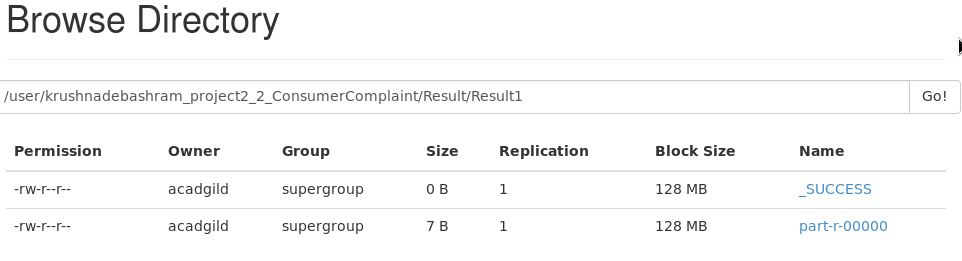
Write a pig script to find no of complaints which got timely response

**Solution 1:**

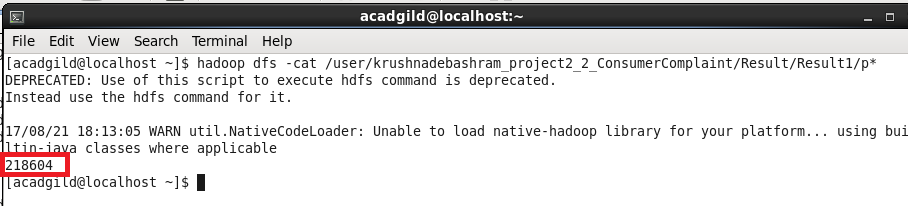
**Pig Statement**

|  |
| --- |
| DEFINE CSVoader org.apache.pig.piggybank.storage.CSVLoader();  ConsumerComplain = LOAD '/user/krushnadebashram\_project2\_2\_ConsumerComplaint/Source\_ConsumerComplaint/F\*.\*'  USING org.apache.pig.piggybank.storage.CSVExcelStorage() as (Datereceived:chararray,Product:chararray,Subproduct:chararray,Issue:chararray,Subissue:chararray  ,Consumercomplaintnarrative:chararray  ,Companypublicresponse:chararray,Company:chararray,State:chararray,ZIPcode:chararray,Submittedvia:chararray  ,Datesenttocompany:chararray  ,Companyresponsetoconsumer:chararray,IsTimelyresponse:chararray,IsConsumerdisputed:chararray,ComplaintID:chararray  );  ConsumerComplain\_excludeHeader = filter ConsumerComplain by (Datereceived!='Date received') ;  filterTimelyResponse = filter ConsumerComplain\_excludeHeader by IsTimelyresponse=='Yes';  ConsumerComplainGenerate = FOREACH filterTimelyResponse GENERATE 1 as Constant, ComplaintID;  groupByConstant = GROUP ConsumerComplainGenerate BY Constant;  countAll = FOREACH groupByConstant GENERATE COUNT(ConsumerComplainGenerate.ComplaintID);  STORE countAll INTO '/user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result1' using PigStorage(','); |

**Result stored in HDFS**

****

[acadgild@localhost ~]$ hadoop dfs -cat /user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result1/p\*

****

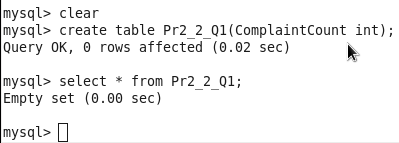
**MySQL create table:**

Go to MySQL shell. Go to the database ‘Test’

[acadgild@localhost ~]$ mysql -u kduser -p

mysql>use Test;

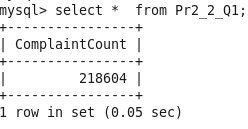
mysql>create table Pr2\_2\_Q1(ComplaintCount int);

****

**Sqoop export to move result to MySQL**

[acadgild@localhost ~]$ sqoop export -connect jdbc:mysql://localhost/test -username kduser -password userkd -table Pr2\_2\_Q1 -export-dir /user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result1/p\*

**View Result in MySQL:**



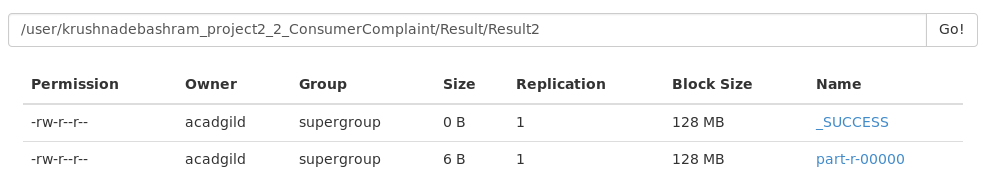
## Problem 2:

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

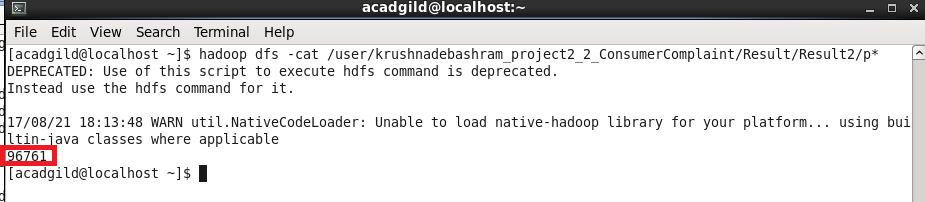
**Solution 2:**

|  |
| --- |
| DEFINE CSVoader org.apache.pig.piggybank.storage.CSVLoader();  ConsumerComplain = LOAD '/user/krushnadebashram\_project2\_2\_ConsumerComplaint/Source\_ConsumerComplaint/F\*.\*'  USING org.apache.pig.piggybank.storage.CSVExcelStorage() as (Datereceived:chararray,Product:chararray,Subproduct:chararray,Issue:chararray,Subissue:chararray  ,Consumercomplaintnarrative:chararray  ,Companypublicresponse:chararray,Company:chararray,State:chararray,ZIPcode:chararray,Submittedvia:chararray  ,Datesenttocompany:chararray  ,Companyresponsetoconsumer:chararray,IsTimelyresponse:chararray,IsConsumerdisputed:chararray,ComplaintID:chararray  );  ConsumerComplain\_excludeHeader = filter ConsumerComplain by (Datereceived!='Date received') ;  filterSentToComapanySameDay= filter ConsumerComplain\_excludeHeader by ToDate(Datereceived,'mm/dd/yyyy')==ToDate(Datesenttocompany,'mm/dd/yyyy');  ConsumerComplainGenerate = FOREACH filterSentToComapanySameDay GENERATE 1 as Constant, ComplaintID;  groupByConstant = GROUP ConsumerComplainGenerate BY Constant;  countAll = FOREACH groupByConstant GENERATE COUNT(ConsumerComplainGenerate.ComplaintID);  STORE countAll INTO '/user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result2' using PigStorage(','); |

**Result stored in HDFS**

****

[acadgild@localhost ~]$ hadoop dfs -cat /user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result2/p\*



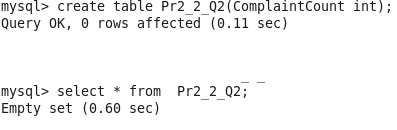
**MySQL create table:**

Go to MySQL shell. Go to the database ‘Test’

[acadgild@localhost ~]$ mysql -u kduser -p

mysql>use Test;

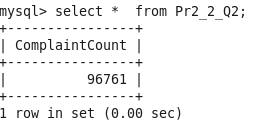
mysql> create table Pr2\_2\_Q2(ComplaintCount int);



**Sqoop export to move result to MySQL**

[acadgild@localhost ~]$ sqoop export -connect jdbc:mysql://localhost/test -username kduser -password userkd -table Pr2\_2\_Q2 -export-dir /user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result2/p\*

**View Result in MySQL:**



## Problem 3:

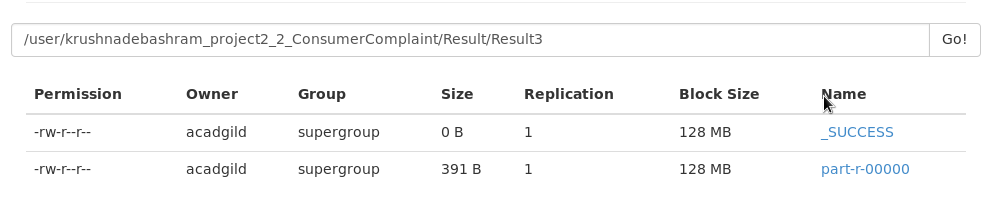
Write a pig script to find list of companies toping in complaint chart (companies with maximum number of complaints)

**Solution 3:**

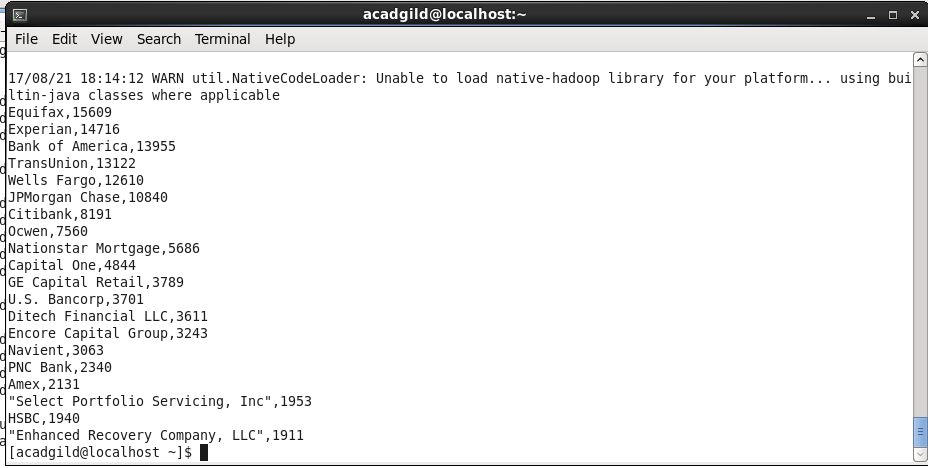
|  |
| --- |
| DEFINE CSVoader org.apache.pig.piggybank.storage.CSVLoader();  ConsumerComplain = LOAD '/user/krushnadebashram\_project2\_2\_ConsumerComplaint/Source\_ConsumerComplaint/F\*.\*'  USING org.apache.pig.piggybank.storage.CSVExcelStorage() as (Datereceived:chararray,Product:chararray,Subproduct:chararray,Issue:chararray,Subissue:chararray  ,Consumercomplaintnarrative:chararray  ,Companypublicresponse:chararray,Company:chararray,State:chararray,ZIPcode:chararray,Submittedvia:chararray  ,Datesenttocompany:chararray  ,Companyresponsetoconsumer:chararray,IsTimelyresponse:chararray,IsConsumerdisputed:chararray,ComplaintID:chararray  );  ConsumerComplain\_excludeHeader = filter ConsumerComplain by (Datereceived!='Date received') ;  ConsumerComplainGenerate = FOREACH ConsumerComplain\_excludeHeader GENERATE Company,ComplaintID;  groupByCompany = GROUP ConsumerComplainGenerate BY Company;  countByCompany = FOREACH groupByCompany GENERATE group as Company,COUNT(ConsumerComplainGenerate.Company) as ComplaintCnt;  orderByDesc = ORDER countByCompany By ComplaintCnt Desc;  limitDesc = LIMIT orderByDesc 20;  STORE limitDesc INTO '/user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result3' using org.apache.pig.piggybank.storage.CSVExcelStorage(); |

**Note**: As the company name has comma as value so instead of using PigStorage(',') applied org.apache.pig.piggybank.storage.CSVExcelStorage() to sore the result in csv format.

**Result stored in HDFS**

****

[acadgild@localhost ~]$ hadoop dfs -cat /user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result3/p\*

****

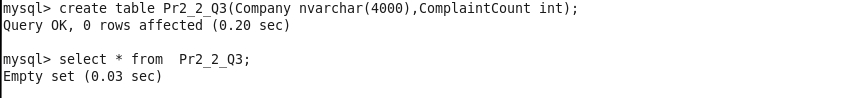
**MySQL create table:**

Go to MySQL shell. Go to the database ‘Test’

[acadgild@localhost ~]$ mysql -u kduser -p

mysql>use Test;

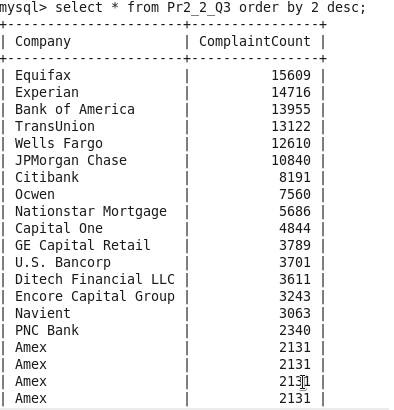
mysql> create table Pr2\_2\_Q3(Company nvarchar(4000),ComplaintCount int);

****

**Sqoop export to move result to MySQL**

[acadgild@localhost ~]$ sqoop export -connect jdbc:mysql://localhost/test -username kduser -password userkd -table Pr2\_2\_Q3 -export-dir /user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result3/p\*

**View Result in MySQL:**



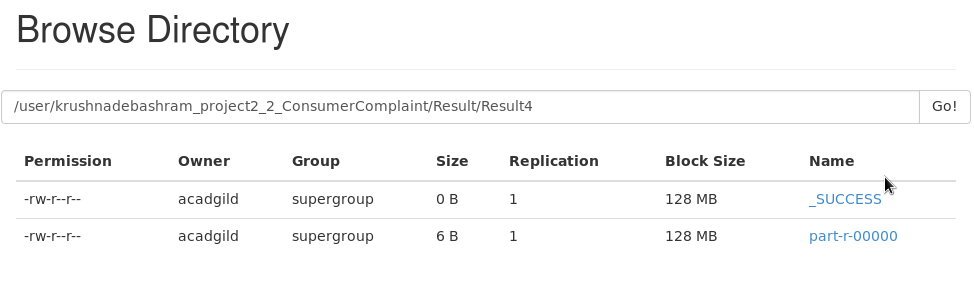
## Problem 4:

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

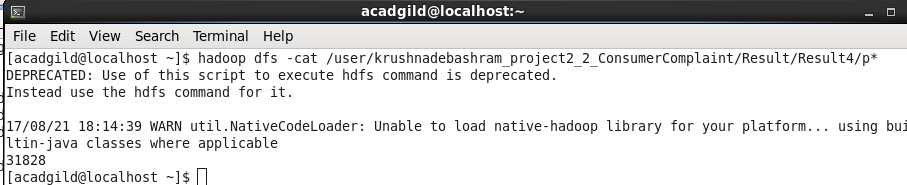
**Solution 4:**

|  |
| --- |
| DEFINE CSVoader org.apache.pig.piggybank.storage.CSVLoader();  ConsumerComplain = LOAD '/user/krushnadebashram\_project2\_2\_ConsumerComplaint/Source\_ConsumerComplaint/F\*.\*'  USING org.apache.pig.piggybank.storage.CSVExcelStorage() as (Datereceived:chararray,Product:chararray,Subproduct:chararray,Issue:chararray,Subissue:chararray  ,Consumercomplaintnarrative:chararray  ,Companypublicresponse:chararray,Company:chararray,State:chararray,ZIPcode:chararray,Submittedvia:chararray  ,Datesenttocompany:chararray  ,Companyresponsetoconsumer:chararray,IsTimelyresponse:chararray,IsConsumerdisputed:chararray,ComplaintID:chararray  );  ConsumerComplain\_excludeHeader = filter ConsumerComplain by (Datereceived!='Date received') ;  filterYear2015withDebtCollection= filter ConsumerComplain\_excludeHeader by GetYear(ToDate(Datereceived,'mm/dd/yyyy'))==2015 and Product=='Debt collection';  Consumer2015withDebtCollectionGenerate = FOREACH filterYear2015withDebtCollection GENERATE 1 as Constant,ComplaintID;  groupBy2015withDebtCollection = GROUP Consumer2015withDebtCollectionGenerate BY Constant;  countBy2015withDebtCollection = FOREACH groupBy2015withDebtCollection GENERATE COUNT(Consumer2015withDebtCollectionGenerate.ComplaintID) as ComplaintCnt;  STORE countBy2015withDebtCollection INTO '/user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result4' using PigStorage(','); |

**Result stored in HDFS**

****

[acadgild@localhost ~]$ hadoop dfs -cat /user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result4/p\*



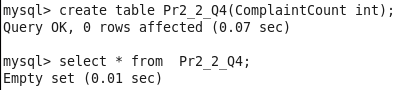
**MySQL create table:**

Go to MySQL shell. Go to the database ‘Test’

[acadgild@localhost ~]$ mysql -u kduser -p

mysql>use Test;

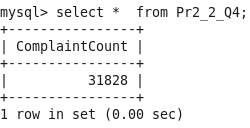
mysql> create table Pr2\_2\_Q4(ComplaintCount int);

****

**Sqoop export to move result to MySQL**

[acadgild@localhost ~]$ sqoop export -connect jdbc:mysql://localhost/test -username kduser -password userkd -table Pr2\_2\_Q4 -export-dir /user/krushnadebashram\_project2\_2\_ConsumerComplaint/Result/Result4/p\*

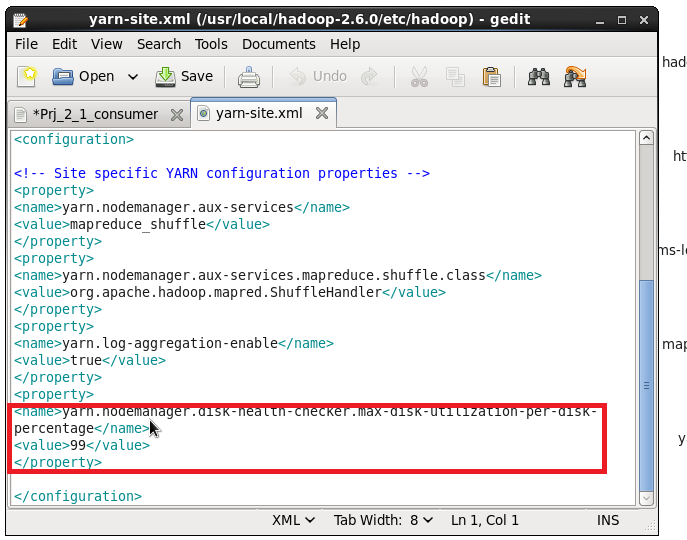
**View Result in MySQL:**



## Troubleshoot

1.In pig dump works fine but when do store to get result store in HDFS then zero percent complete and it waits for infinite.

Solution:This is due to the disk utilization. Increase the disk utilization in the yarn-site.xml.



## Known Issue

1.Flume copies 70mb data to HDFS ,instead of 103mb in the csv file.

2.Sqoop export the pig output store in hdfs file part-r-00000

In the output company name has comma as value

‘Select Portfolio Servicing, Inc’

Exporting this to MySql has issue.