Project 2: USA Consumer Forum Data Analysis

Copying data into HDFS using FLUME:

**Consumer.conf:**

agent1.sources = source1

agent1.sinks = sink1

agent1.channels = channel1

agent1.sources.source1.type = exec

agent1.sources.source1.command = hadoop dfs -put /home/acadgild/flume/Consumer\_Complaints.csv /user/acadgild/hadoop/

agent1.sinks.sink1.type = hdfs

agent1.sinks.sink1.hdfs.path = hdfs://localhost:9000/user/acadgild/hadoop/

agent1.channels.channel1.type = memory

agent1.sources.source1.channels = channel1

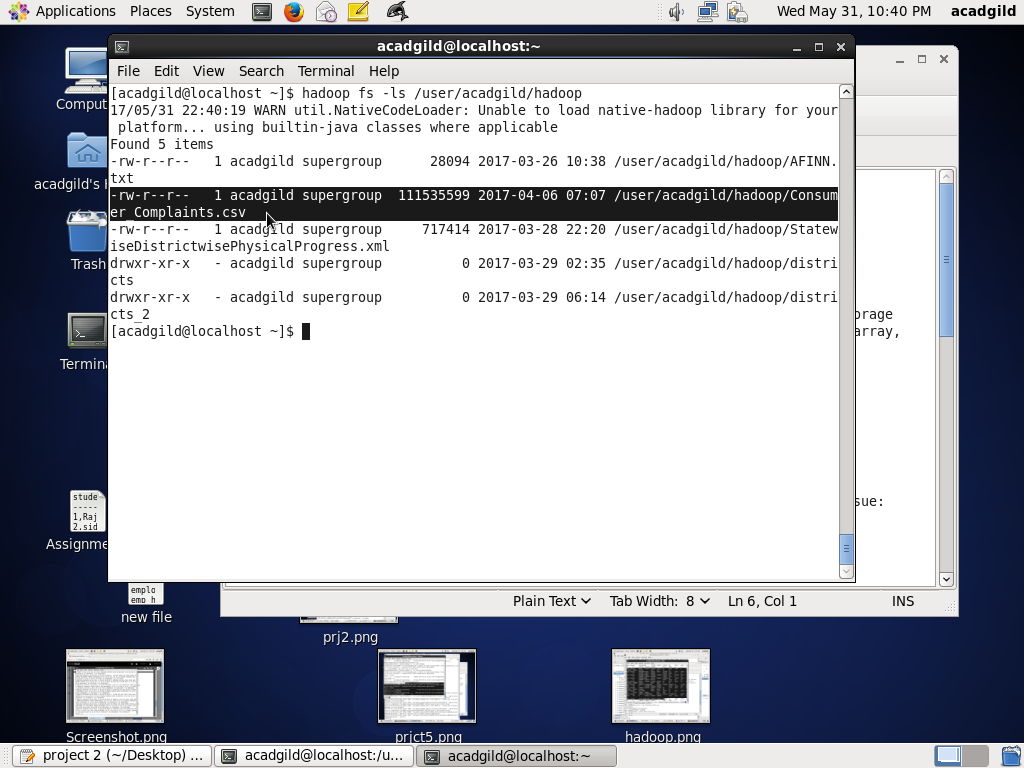
agent1.sinks.sink1.channel = channel1

**Run consumer.conf:**

flume-ng agent --name agent1 --conf-file /home/acadgild/flume/consumer.conf

**Checking HDFS location:**

hadoop fs -ls /user/acadgild/hadoop/



**Loading dataset into pig:**

As the file is in csv format, we use CSVExcelStorage to load it.

**REGISTER** '/usr/local/pig/lib/piggybank.jar';

**consumerData**= LOAD '/user/acadgild/hadoop/Consumer\_Complaints.csv' using org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER') as (date\_rec:chararray,product:chararray,subproduct:chararray,issue:chararray,

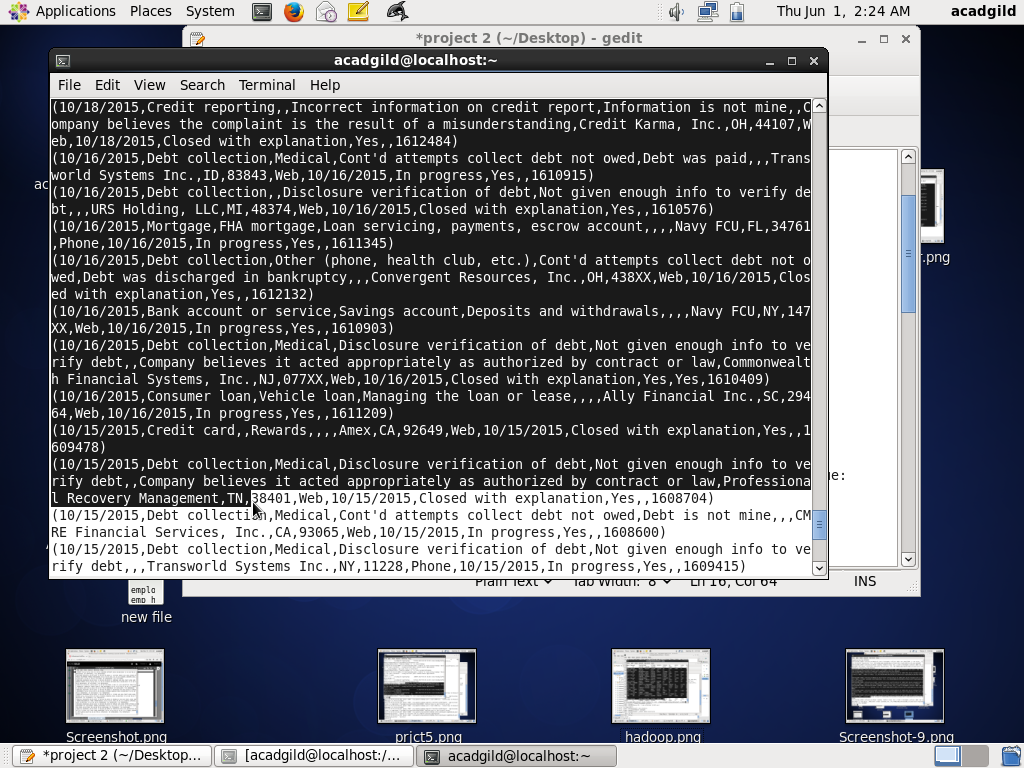
subissue:chararray,consumer\_complaint:chararray,company\_response:chararray,

company:chararray,state:chararray,zip:chararray,

submit\_via:chararray,date\_sent:chararray,company\_res:chararray,

timely:chararray,consumer\_dispute:chararray,complaint\_id:int);

**DUMP consumerData;**



**DESCRIBE consumerData;**

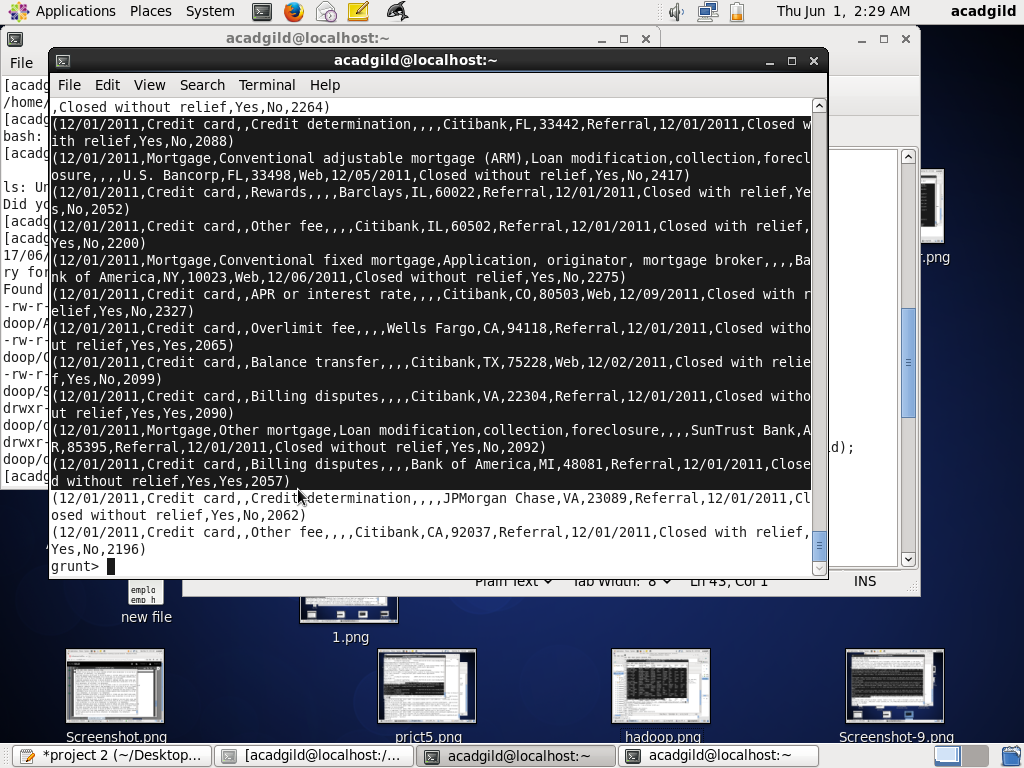
consumerData: {date\_rec: chararray,product: chararray,subproduct: chararray,issue: chararray,subissue: chararray,consumer\_complaint: chararray,company\_response: chararray,company: chararray,state: chararray,zip: chararray,submit\_via: chararray,date\_sent: chararray,company\_res: chararray,timely: chararray,consumer\_dispute: chararray,complaint\_id: int}

**Problem statement 1: Write a pig script to find no of complaints which got timely response**

**Filtering based on timely response ‘Yes’:**

filterResponse= FILTER consumerData BY timely=='Yes';

**DUMP filterResponse;**



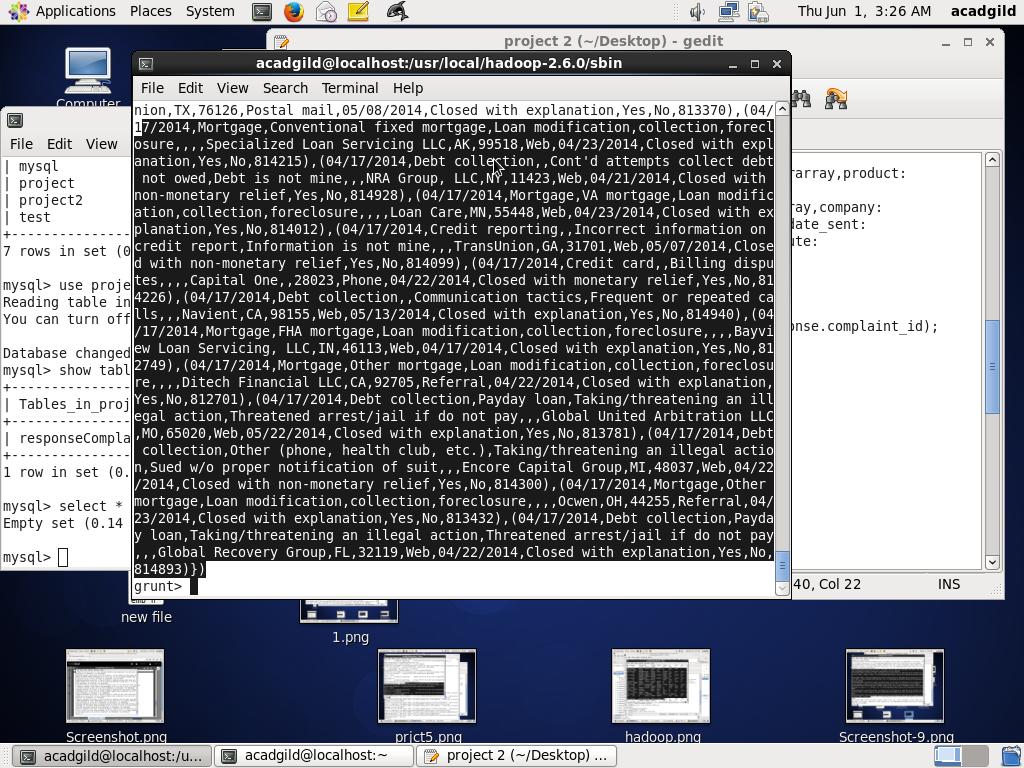
**Grouping:**

groupResponse= GROUP filterResponse all;

**DESCRIBE groupResponse;**

groupResponse: {group: chararray,filterResponse: {(date\_rec: chararray,product: chararray,subproduct: chararray,issue: chararray,subissue: chararray,consumer\_complaint: chararray,company\_response: chararray,company: chararray,state: chararray,zip: chararray,submit\_via: chararray,date\_sent: chararray,company\_res: chararray,timely: chararray,consumer\_dispute: chararray,complaint\_id: int)}}

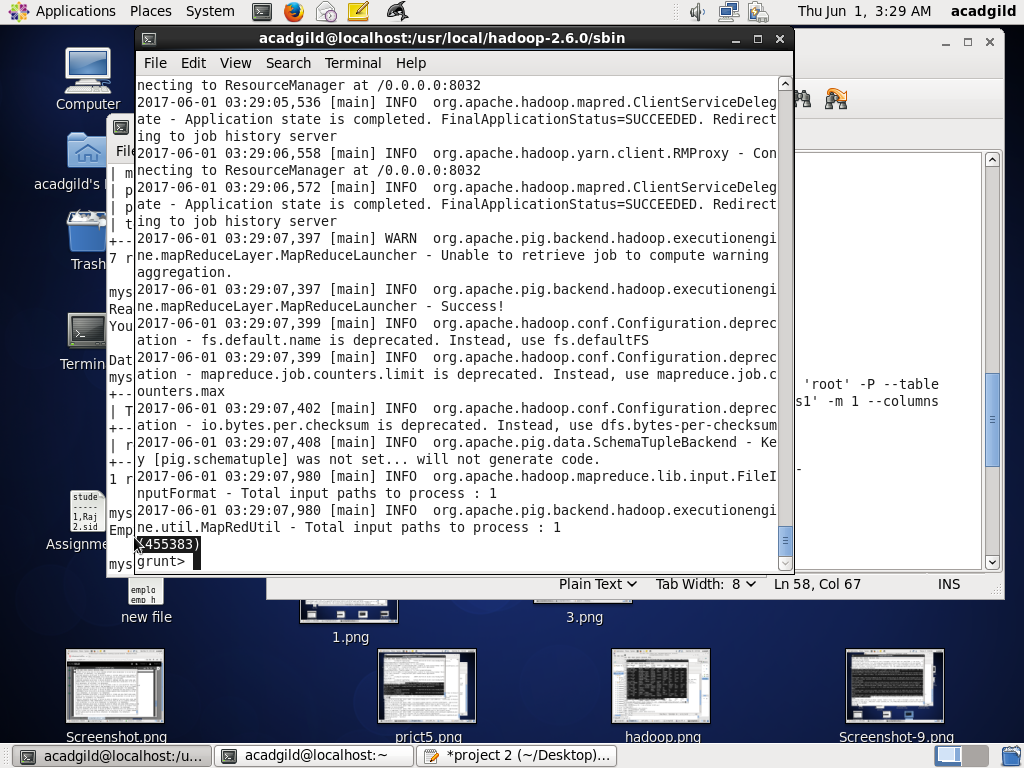
**DUMP groupResponse;**



**Getting count of complaints:**

**totalComplaints**= FOREACH groupResponse GENERATE COUNT(filterResponse.complaint\_id);

**DUMP totalComplaints;**

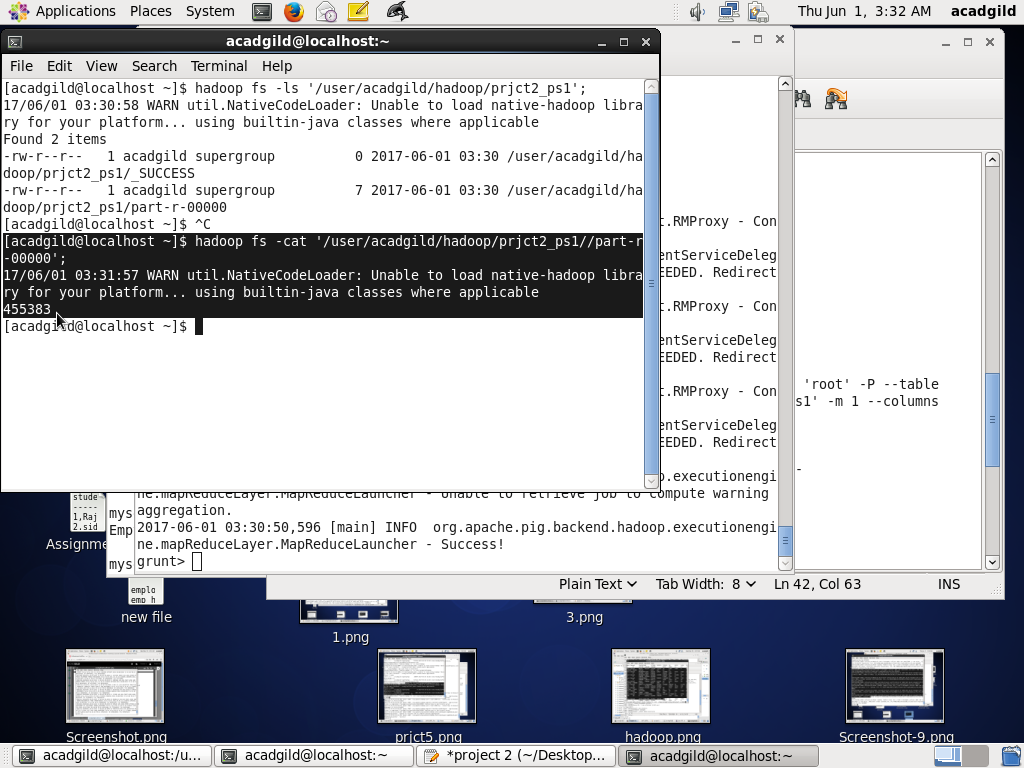


**Storing output into HDFS:**

STORE totalComplaints INTO '/user/acadgild/hadoop/ps1';

**Checking in HDFS location:**

hadoop fs -cat '/user/acadgild/hadoop/prjct2\_ps1/part-r-00000';



**Creating database in mysql:**

mysql> **create database project2;**

Query OK, 1 row affected (0.11 sec)

mysql> **use project2;**

Database changed

**Creating table:**

create table responseComplaints(

complaints int

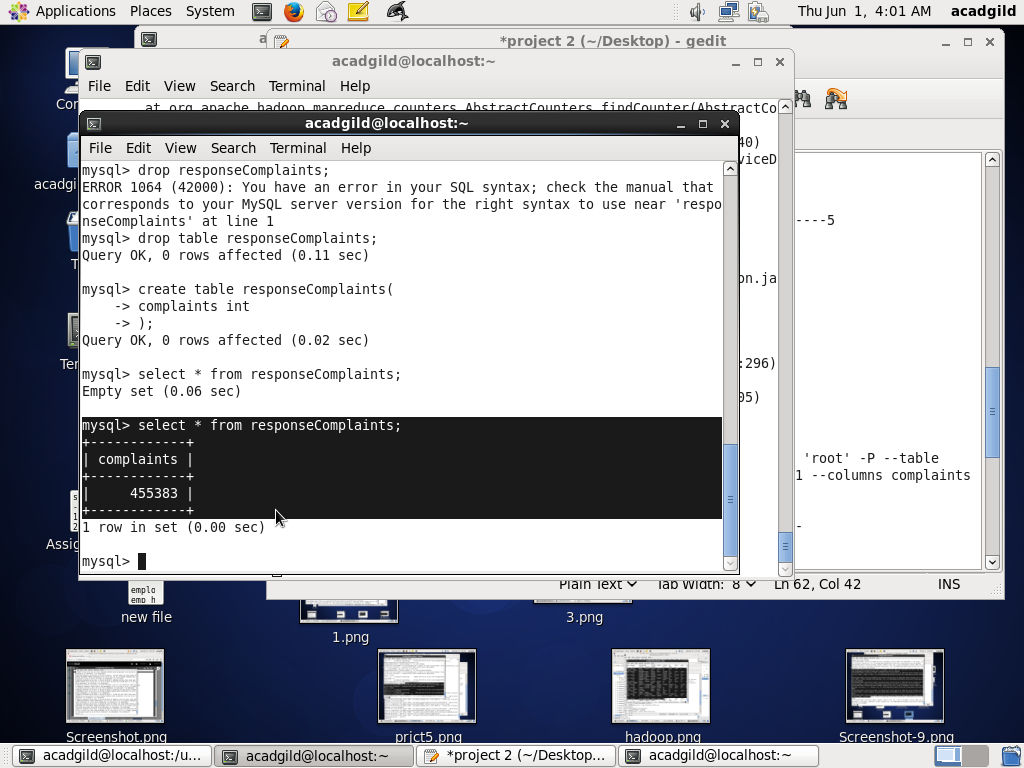
);

**Exporting data from HDFS into mysql:**

sqoop export --connect jdbc:mysql://localhost/project2 --username 'root' -P --table 'responseComplaints' --export-dir '/user/acadgild/hadoop/ps1' -m 1 --columns complaints

**Checking output in Mysql:**

select \* from responseComplaints;

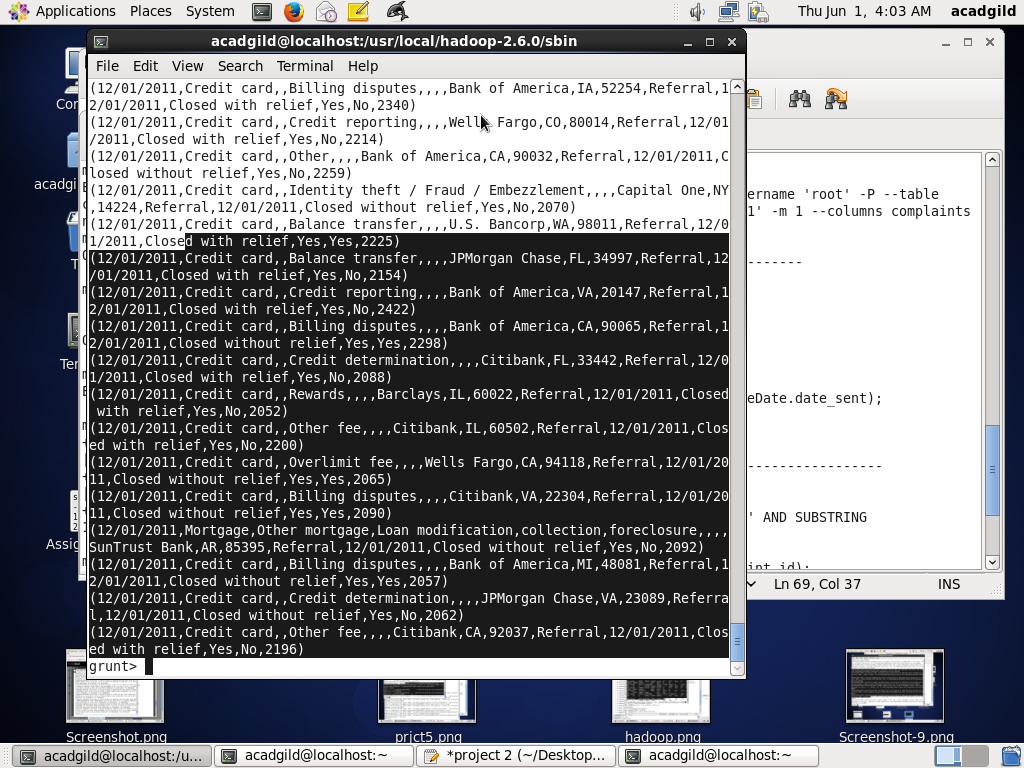


**Problem Statement 2:** **Write a pig script to find no of complaints where consumer forum forwarded the complaint same day they received to respective company**

**Filtering data by date:**

filterSameDate= FILTER consumerData BY date\_rec==date\_sent;

**DUMP filterSameDate;**

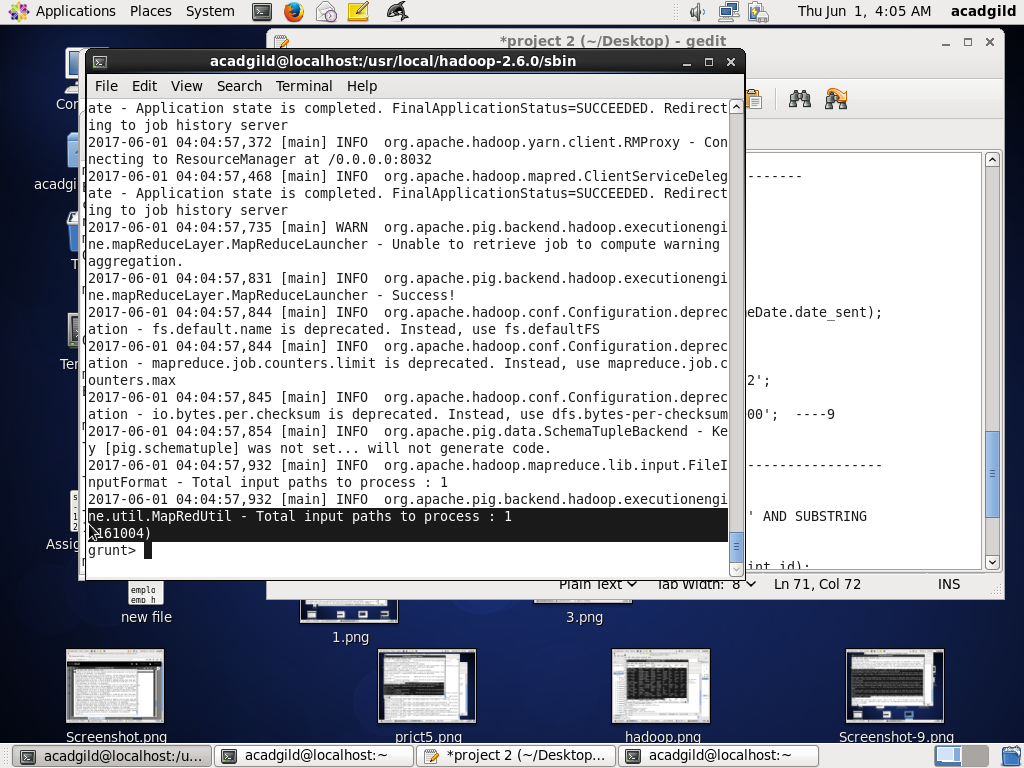


**Grouping:**

**groupDate**= GROUP filterSameDate all;

**finalComplaints**= FOREACH groupDate GENERATE COUNT(filterSameDate.date\_sent);

**DUMP finalComplaints;**

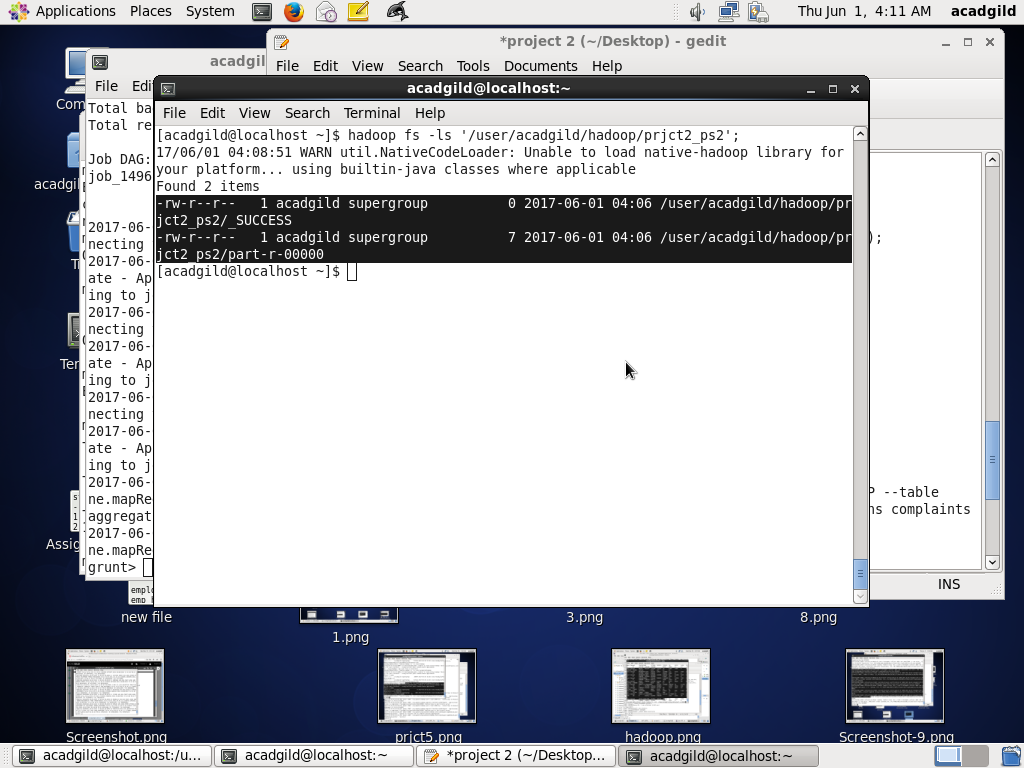


**Storing output into HDFS:**

**STORE** finalComplaints INTO '/user/acadgild/hadoop/prjct2\_ps2';

**Checking in HDFS location:**

hadoop fs -cat '/user/acadgild/hadoop/prjct2\_ps2/part-r-00000';



**Use same database project2:**

mysql> **use project2;**

Database changed

**Create table:**

create table forwardComplaints(

complaints int

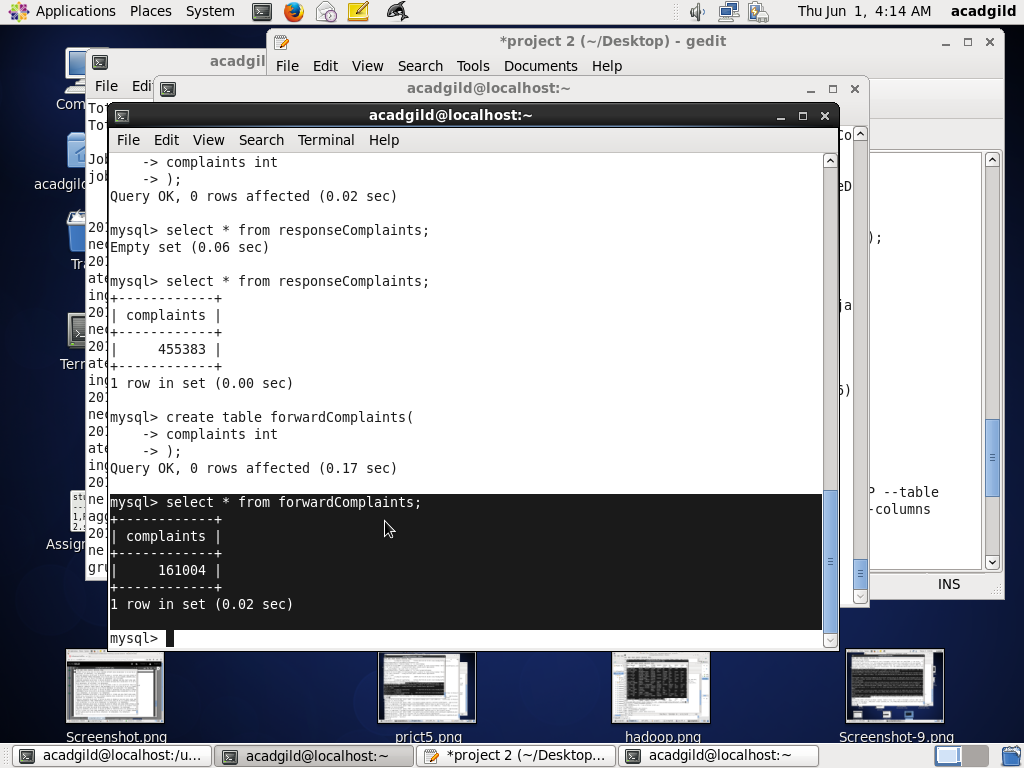
);

**Export data from HDFS to Mysql:**

sqoop export --connect jdbc:mysql://localhost/project2 --username 'root' -P --table 'forwardComplaints' --export-dir '/user/acadgild/hadoop/prjct2\_ps2' -m 1 --columns complaints

**Checking output in Mysql:**

select \* from responseComplaints;



**Problem Statement 3**: **Write a pig script to find list of companies toping in complaint chart (companies with maximum number of complaints)**

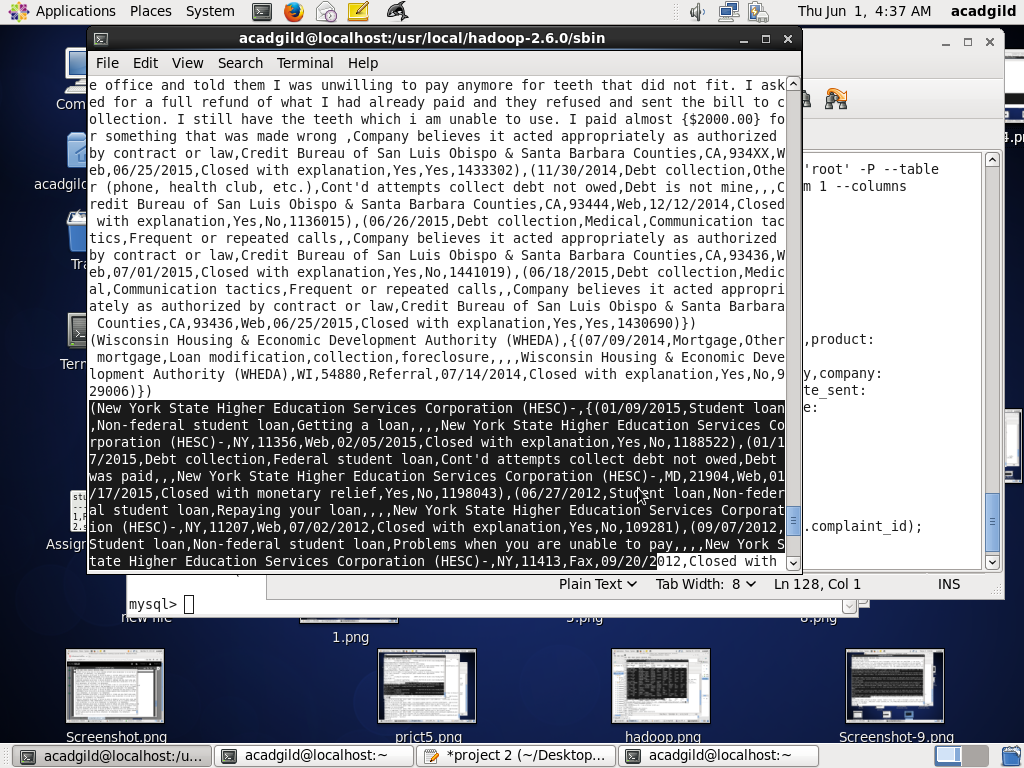
**Grouping data based on company:**

**groupedData**= GROUP consumerData BY company;

**DESCRIBE** groupedData;

groupedData: {group: chararray,consumerData: {(date\_rec: chararray,product: chararray,subproduct: chararray,issue: chararray,subissue: chararray,consumer\_complaint: chararray,company\_response: chararray,company: chararray,state: chararray,zip: chararray,submit\_via: chararray,date\_sent: chararray,company\_res: chararray,timely: chararray,consumer\_dispute: chararray,complaint\_id: int)}}

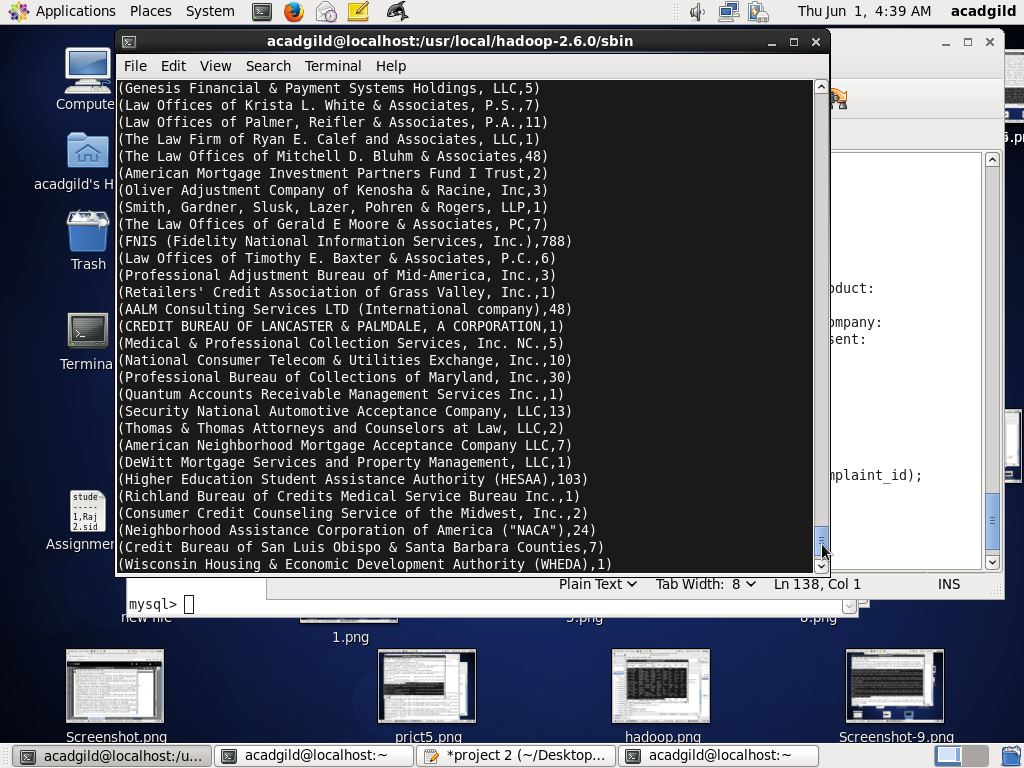
**DUMP groupedData;**



**Counting number of complaints:**

**countedData**= FOREACH groupedData GENERATE group,COUNT(consumerData.complaint\_id);

**DUMP countedData;**



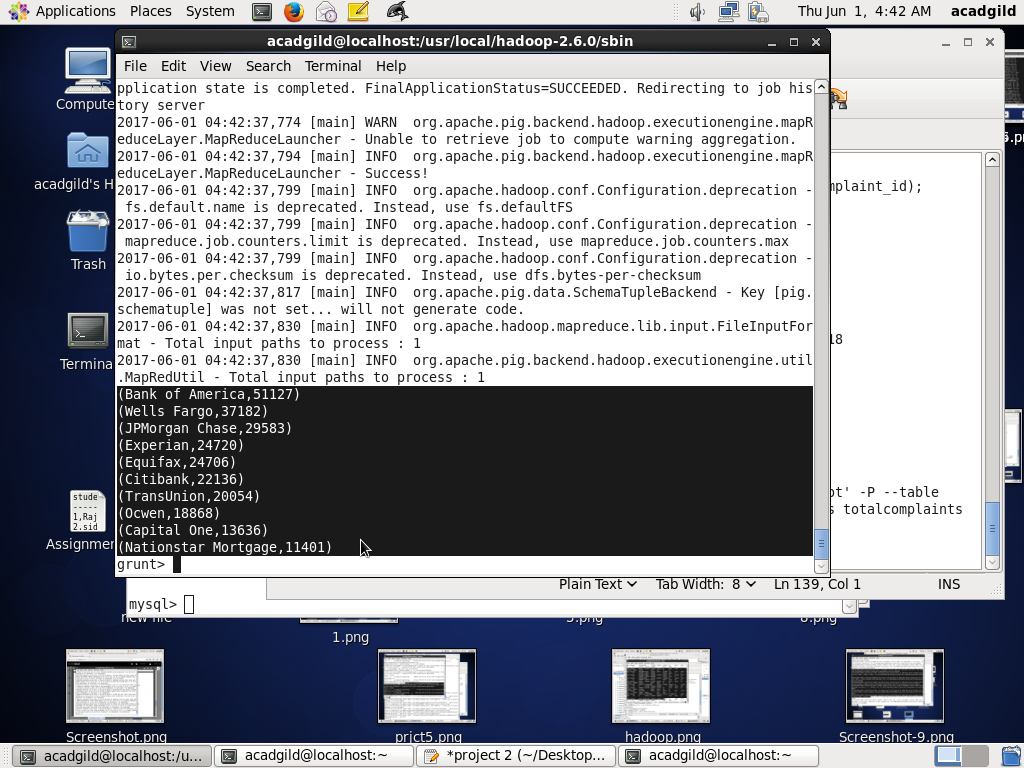
**Ordering complaints on descending order:**

**orderedData**= ORDER countedData BY $1 DESC;

**Top 10 List of companies toping in complaint chart:**

**finalCount**= LIMIT orderedData 10;

**DUMP finalCount;**

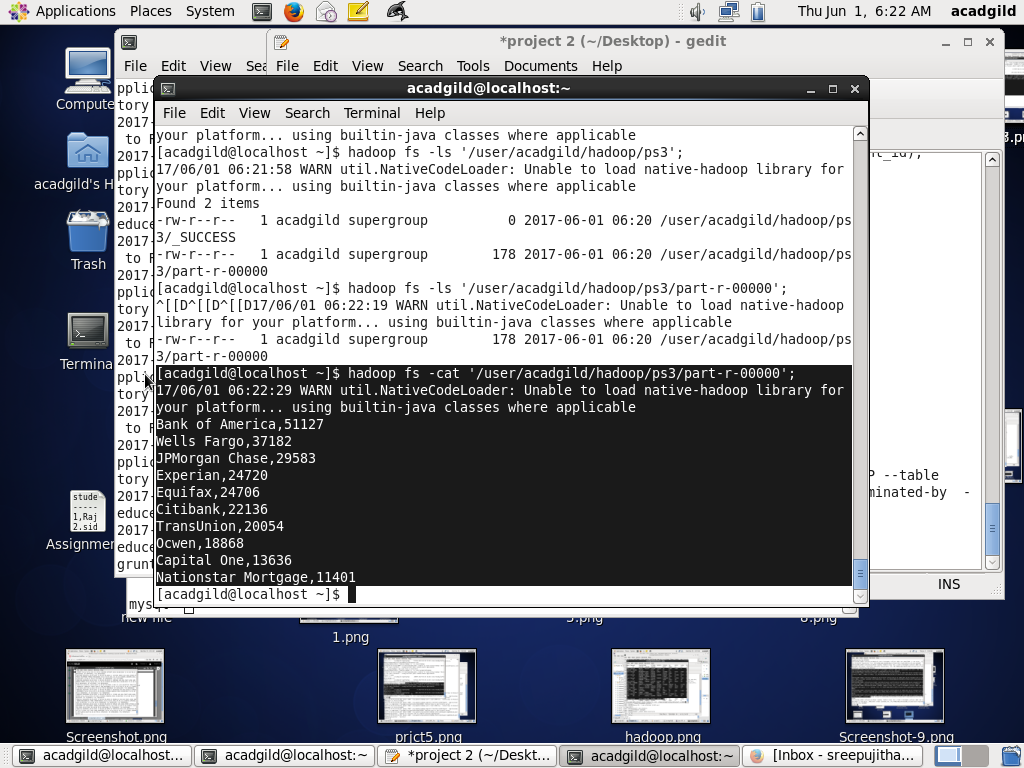


**Storing output into HDFS:**

STORE finalCount INTO '/user/acadgild/hadoop/ps3' USING PigStorage(',');

**Checking in HDFS location:**

hadoop fs -cat '/user/acadgild/hadoop/prjct2\_ps4/part-r-00000';



**Creating table:**

create table **topCompanies**(

company varchar(20),

complaints int

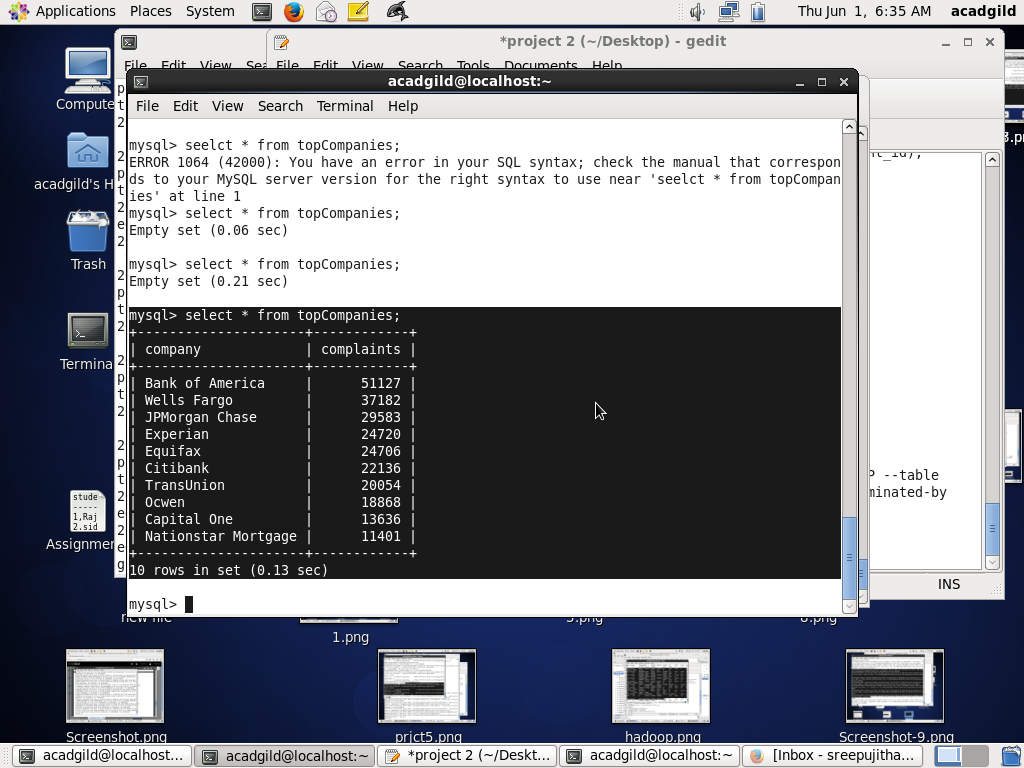
);

**Exporting data from HDFS to Mysql table-- ‘topCompanies’:**

sqoop export --connect jdbc:mysql://localhost/project2 --username 'root' -P --table 'topCompanies' --export-dir '/user/acadgild/hadoop/ps3' --input-fields-terminated-by ',' -m 1 --columns company,complaints

**Checking output in Mysql:**

select \* from topCompanies;

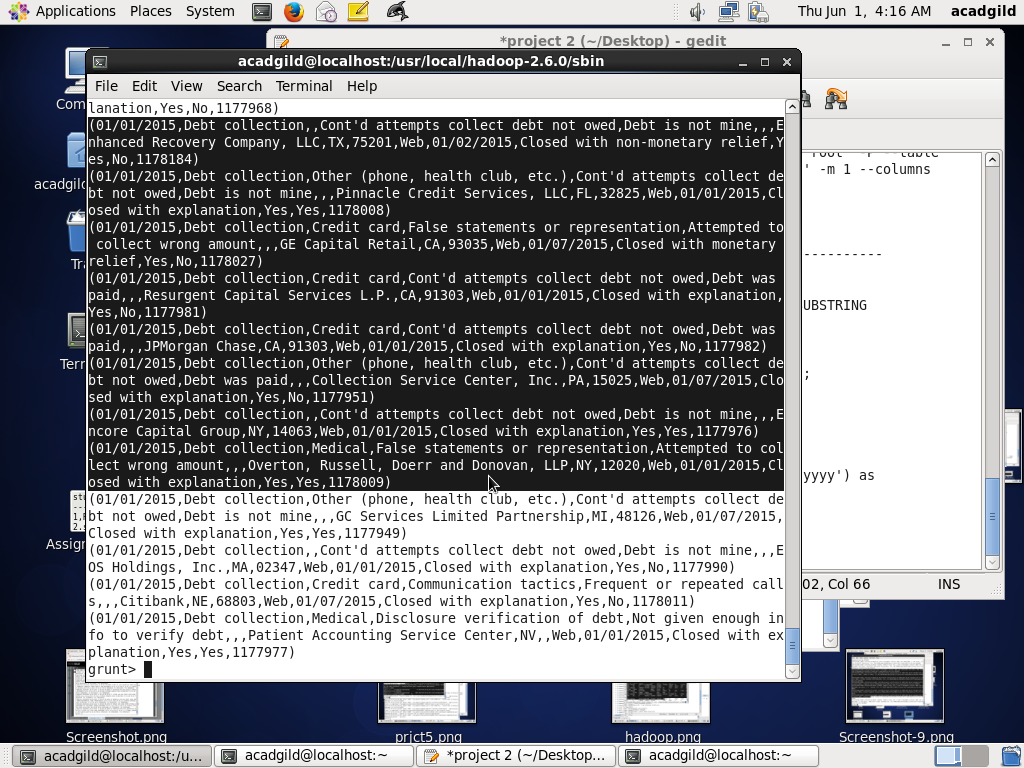


**Problem Statement 4:** **Write a pig script to find no of complaints filed with product type has “ Debt collection” for the year 2015**

**Filtering dataset by product and date:**

**yearData**= FILTER consumerData BY (product=='Debt collection' AND SUBSTRING(date\_rec,6,10)=='2015');

**DUMP yearData;**



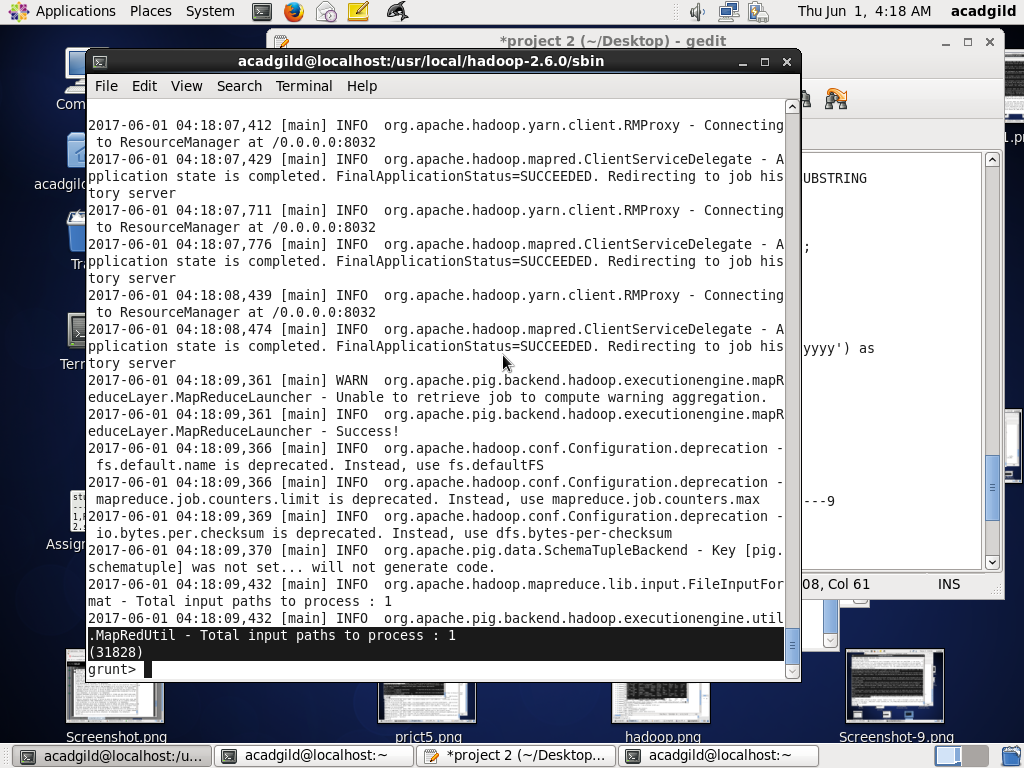
**Grouping data:**

**groupData**= GROUP yearData all;

**Counting number of complaints:**

**countData**= FOREACH groupData GENERATE COUNT(yearData.complaint\_id);

**DUMP countData;**



**ANOTHER APPROACH:**

**Using GetYear method:**

productData= FILTER consumerData BY product=='Debt collection';

year= FOREACH productData GENERATE product,ToDate(date\_rec,'mm/dd/yyyy') as (date:DateTime);

yearData= FILTER year BY GetYear(date)==2015;

groupData= GROUP yearData all;

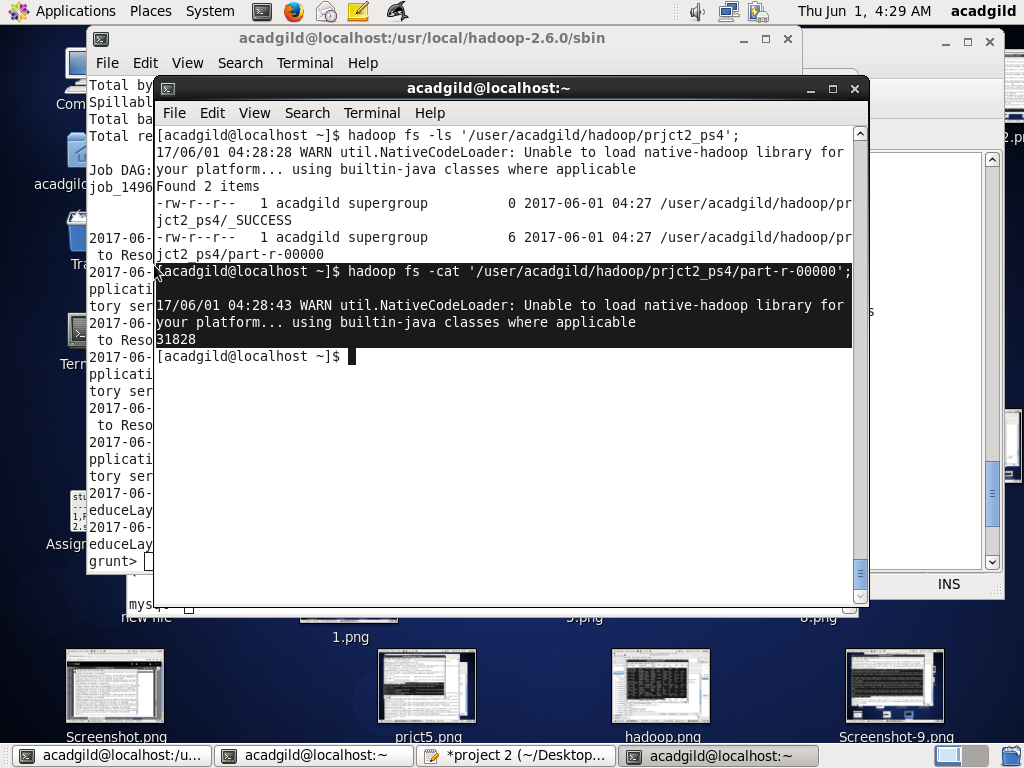
countData= FOREACH groupData GENERATE COUNT(yearData.product);

**Storing output into HDFS:**

STORE countData INTO '/user/acadgild/hadoop/prjct2\_ps4';

**Checking in HDFS Location:**

hadoop fs -cat '/user/acadgild/hadoop/prjct2\_ps4/part-r-00000';



**Creating table:**

create table Complaints2015(

complaints int

);

**Exporting output from HDFS to Mysql table – ‘Complaints2015’**

sqoop export --connect jdbc:mysql://localhost/project2 --username 'root' -P --table 'Complaints2015' --export-dir '/user/acadgild/hadoop/prjct2\_ps4' -m 1 --columns complaints

**Checking output in Mysql:**

seelct \* from Complaints2015;

