**Group-F Lab Tutorial**

Karan Kaura

Daniel Razmjou

Satyam Singh

Jobandeep Singh Mahal

Fall 2016 (12/10/2016)

**Analysis of Consumer Complaints about Financial Services**

**Objectives:**

* Step 1: Raw data
* Step 2: Create an IBM Bluemix cluster
* Step 3: Load data in HDFS
* Step 4: Query data in Hive and Pig
* Step 5: Create tables in Hive
* Step 6: Analyze data and create visuals in Tableau

Step 1: Raw data

The data for our project was downloaded from data.gov website which has the dataset for consumer complaints regarding financial departments. The original data can also be found on cfpb website. The links to download the dataset are:

<https://catalog.data.gov/dataset/consumer-complaint-database#topic=consumer_navigation>

and

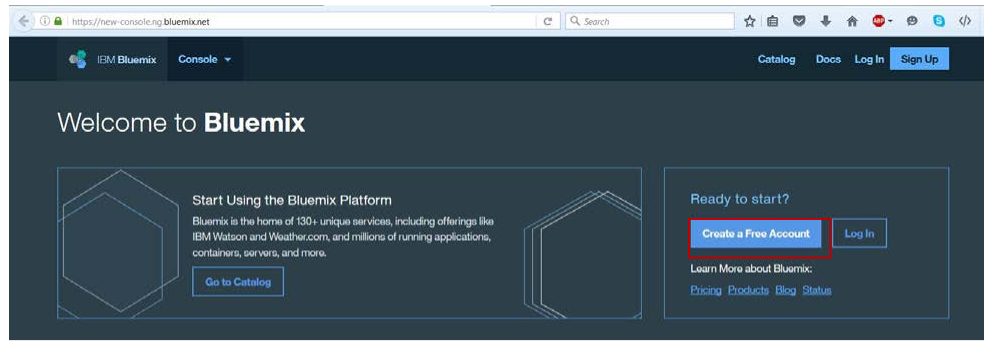
<http://www.consumerfinance.gov/data-research/consumer-complaints/>

After the data was downloaded, which is the .csv file is filtered and unwanted columns and missing values were removed so that we can have the proper data to analyze.

Step 2: Create an IBM Bluemix cluster

Sign up Bluemix

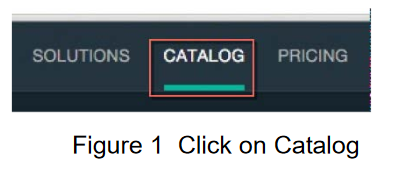
You need to open a web browser and type in url address https://console.ng.bluemix.net. Then, you need to sign it up by selecting “Create a Free Account” as follows:



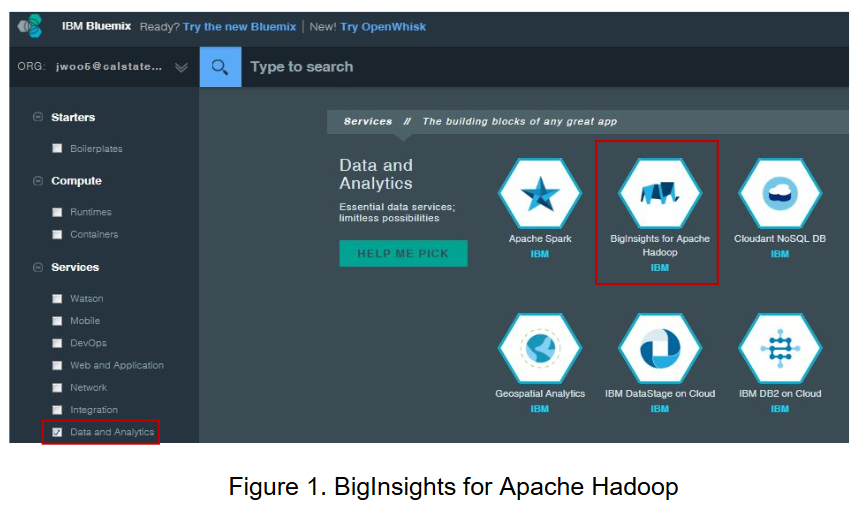
In the next page to sign it up, you will see that y our 30-day trial is free, with no credit card required. You get access to 2 GB of runtime and container memory to run apps, unlimited IBM services and APIs, and complimentary support.

Launch BigInsights from Bluemix

1. After you successfully log in, in the top right of the Bluemix web site, click on Catalog on the top navigation bar on Figure 1



Now you need to mark “Data and Analytics” and then click on a tile “BigInsights for Apache Hadoop ” on Figure 1.



3. You need to set up “BigInsights for Apache Hadoop” on Figure 2. You could leave “Space”, “Service Name” and “selected plan” to the default value. You could modify “Space”, “Service Name” but not “selected plan” as it charges more as an enterprise edition. Then click on Create button

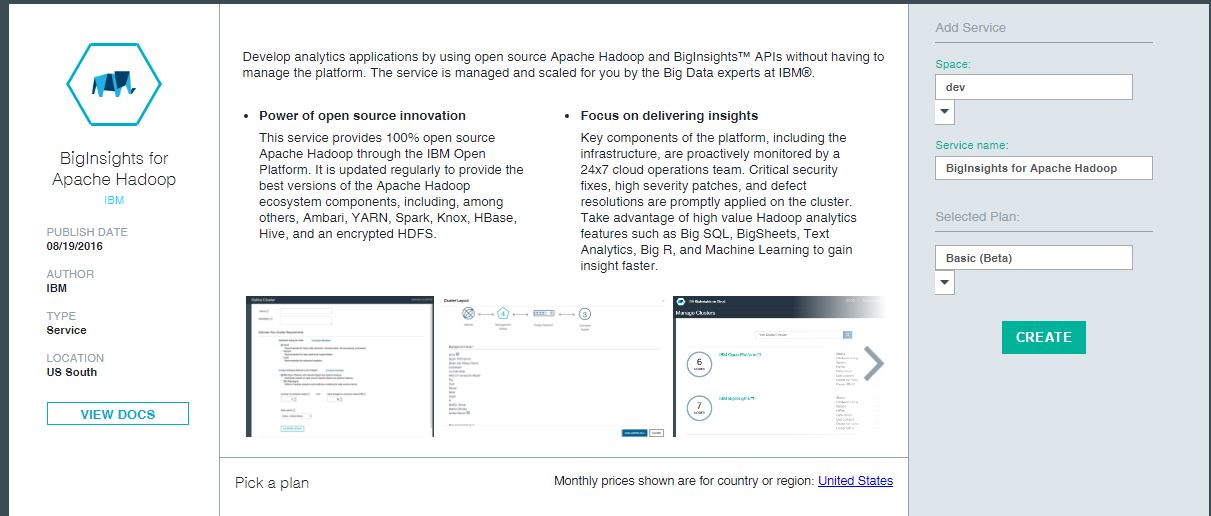


Figure 2. Set up BigInsights for Apache Hadoop

4. You need to click “Open” to see the BigInsights service in Figure 3.

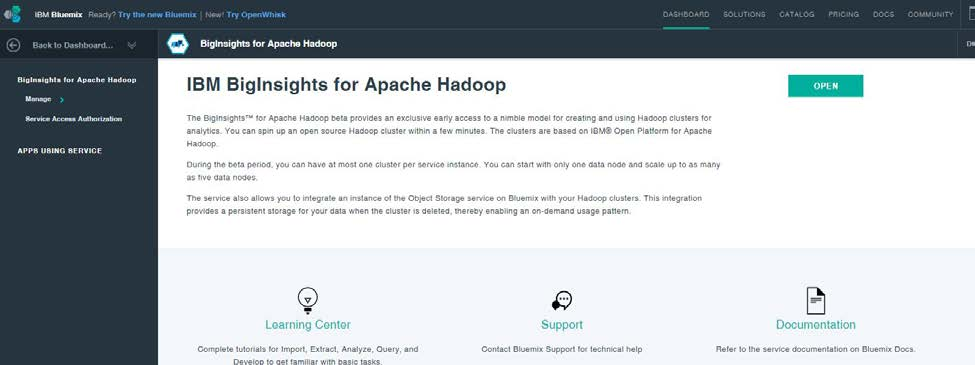
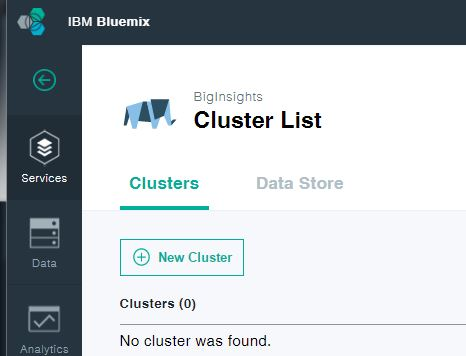


Figure 3. BigInsights for Apache Hadoop service

5. Now you have to create New Cluster. Cluster is of servers that run BigInsights service. Click on “New Cluster”



6. You need to set up the cluster name, username, password, which runs BigInsights service. The password must be at least 12 characters long and contain only letters or numbers. It must contain at least one letter and at least one number.

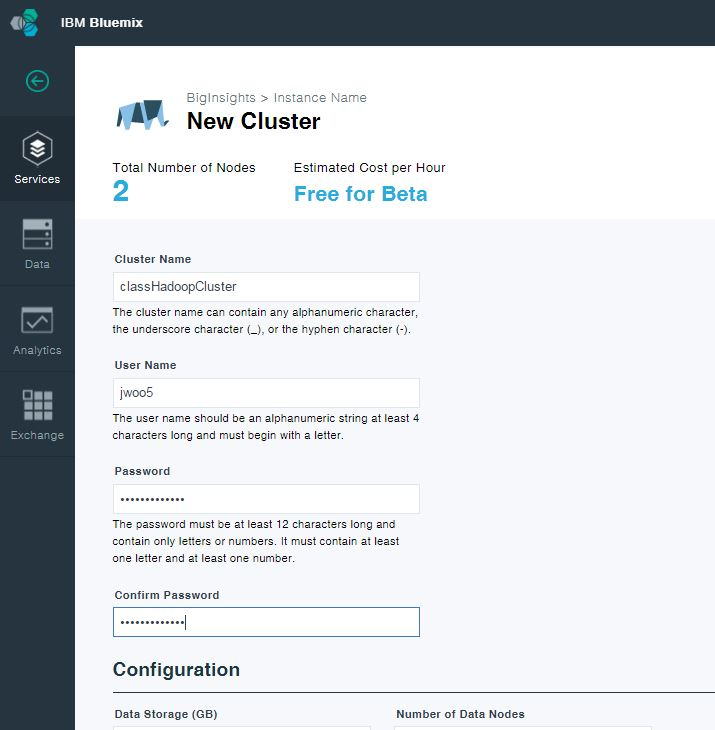
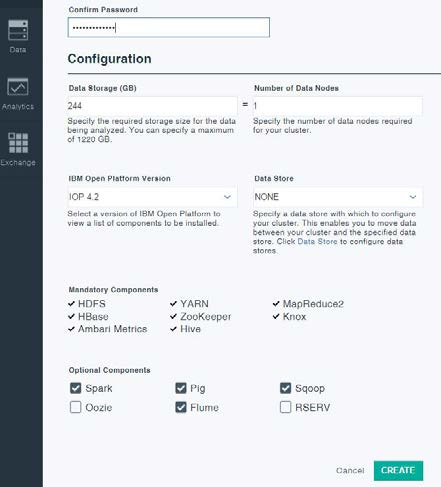


Figure 4. Add cluster of BigInsights service

7. You also need to mark Spark, Hive, Pig, Flume, which are popular Hadoop ecosystems – you may not use them all in the course but in other courses. Now, click on “Create”



8. Now you need to wait for several minutes till “Pending” status becomes “Preparing”

then finally “Active”

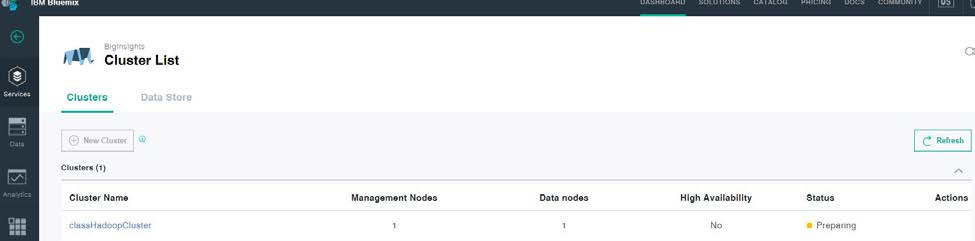
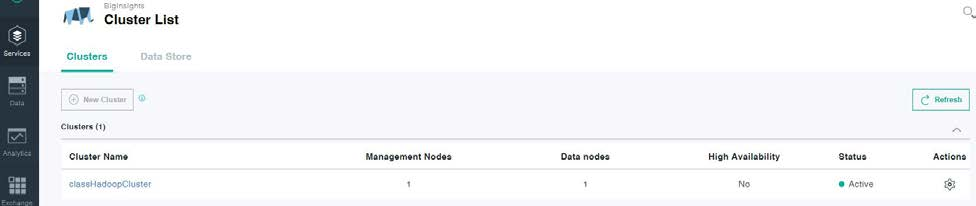


Figure 5. cluster of BigInsights service created

**3. Connect to BigInsights using *ssh***

Click on the cluster name to look at its properties when the status is “Active”



You can see the ssh property in order to remotely connect to the cluster using “ssh” shell commands with your user name and password for your cluster.

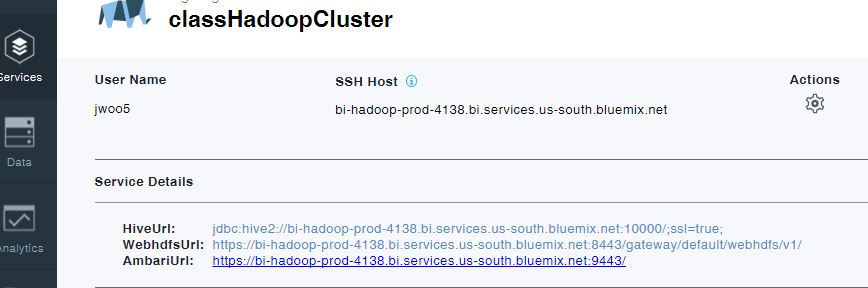


Figure 6. SSH Host of the cluster

The figure below is the ibm bluemix dashboard where you can see the active services you are using.

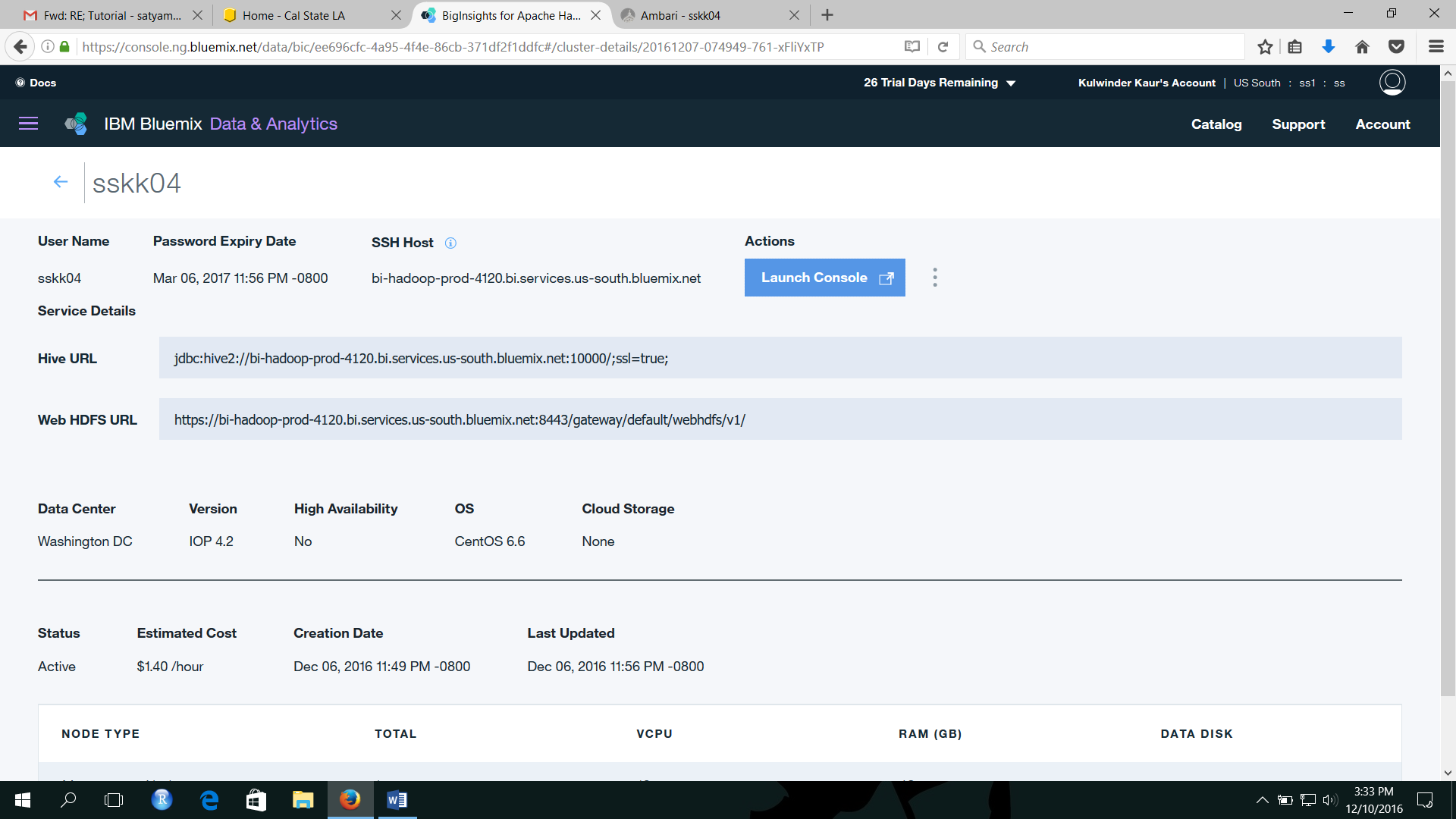


Figure 7. IBM bluemix dashboard

Step 3: Load data in HDFS

Upload .csv file in Ambari using the browser tab. Launch console in the above figure will open the ambary, enter the username and password to log in. The figure 8 below shows the ambari file system.

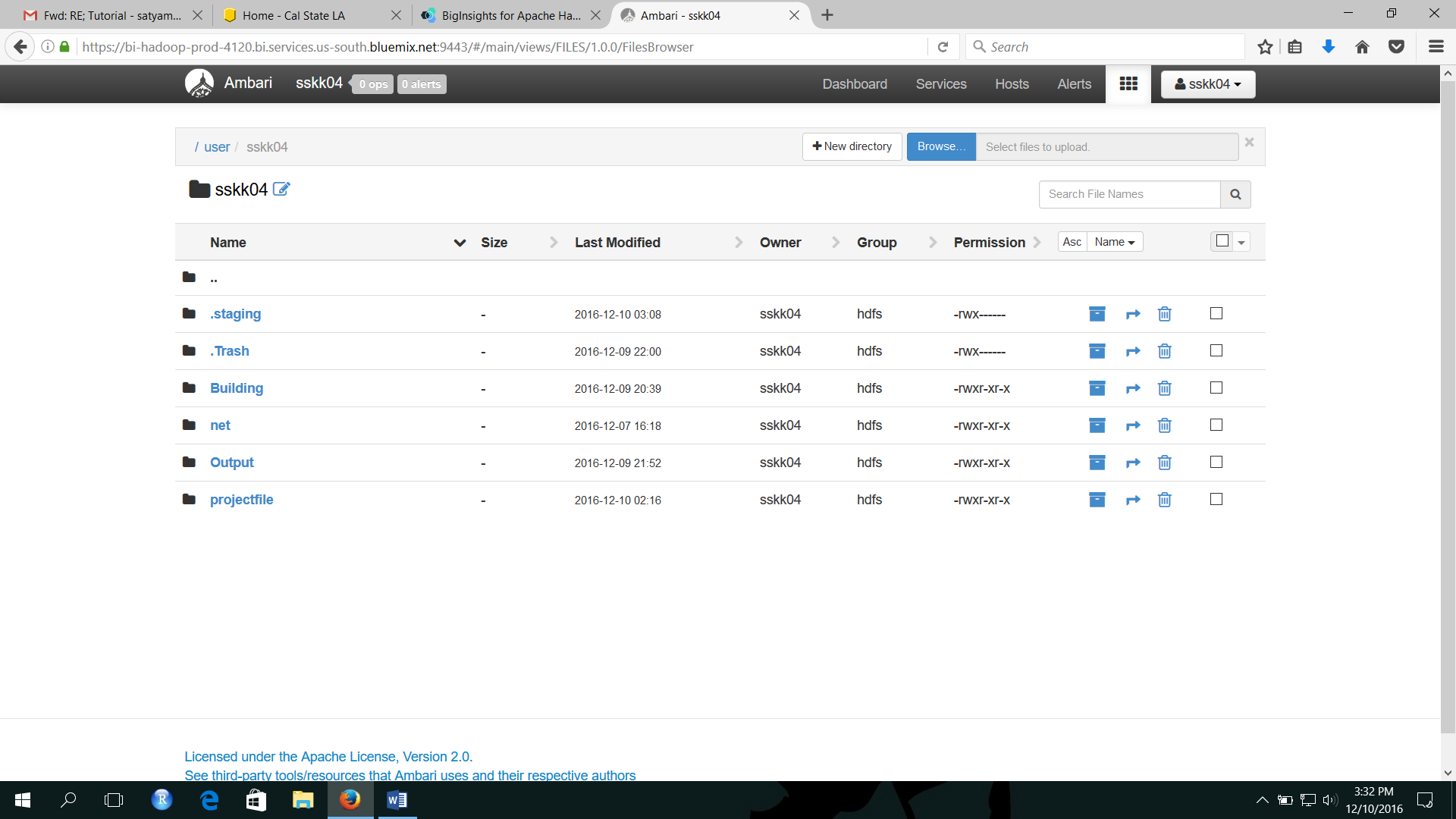
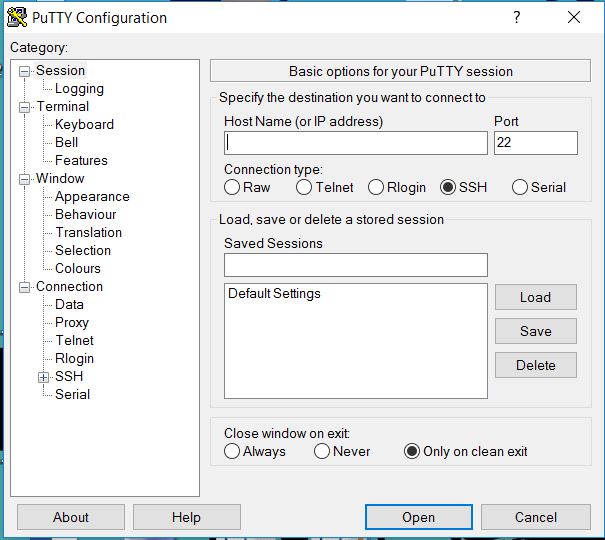


Figure 8. Upload file in Ambari.

Step 4: Query data in Hive and Pig

Using your ssh host url like, bi-hadoop-prod-4120.bi.services.us-south.bluemix.net login into putty as shown in figure below. Furthermore, the other ssh clients can also be used like Minty. If using the mac operating system then terminal can be used as ssh client to connect to your bluemix HD big insight.



Write your ssh host url here.

Figure 9 Putty

1. After opening putty login using your username and password you created when making bluemix account.
2. Open hive.
3. Using hive query, create table as per the columns your .csv file have.
4. The below is the query used to create the table.

create table complaints

(

date\_recieved string, product string, sub\_product string, issue string, sub\_issue string,

complaint\_narrative string, public\_response string, company string, state string, zip int, tags string, consent\_provided string, submitted\_via string, date\_sent string, response\_to\_consumer string, timely\_response string, consumer\_disputed string, complaint\_id bigint) row format delimited fields terminated by ',';

1. As we have created the table. Now, we should dump the data into the table we just created. Below is the command used to dump the data into table.

LOAD DATA INPATH '/user/username/complaints.csv' INTO TABLE COMPLAINTS;

Step 5: Create tables in Hive

The dataset we used can crash the tableau software if we upload the whole .csv file. Therefore, we must filter the data and create different tables to analyze the data using tableau. Now, create each table using hive. Below are the hive queries used to analyze the data.

HIVE>select product, count(complaint\_id) from complaints order by complaint\_id;

HIVE>select state, sum(complaint\_id) from complaints group by state order by state desc;

HIVE>select count(complaint\_id), date\_recieved from complaints;

HIVE>Select product,company,count(complaint\_id) from complaints group by product;

Each of the queries above can be compiled easily in hive and the table can be created using the create table query in hive.

Create the 4 tables using which includes the above query and set path of the table to your file location in hdfs.

For e.g.,

CREATE EXTERNAL TABLE IF NOT EXISTS barchart(product string,complaints\_id)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ' '

STORED AS TEXTFILE LOCATION '/user/satyam24/barchart/'

TBLPROPERTIES 'skip.header.line.count'='2');

In the same way we have to create the tables and then connect with tableau using Hadoop file system and the analyze the data in tableau.

Step 6: Analyze and visualize data with Tableau

1. Open data set in Tableau

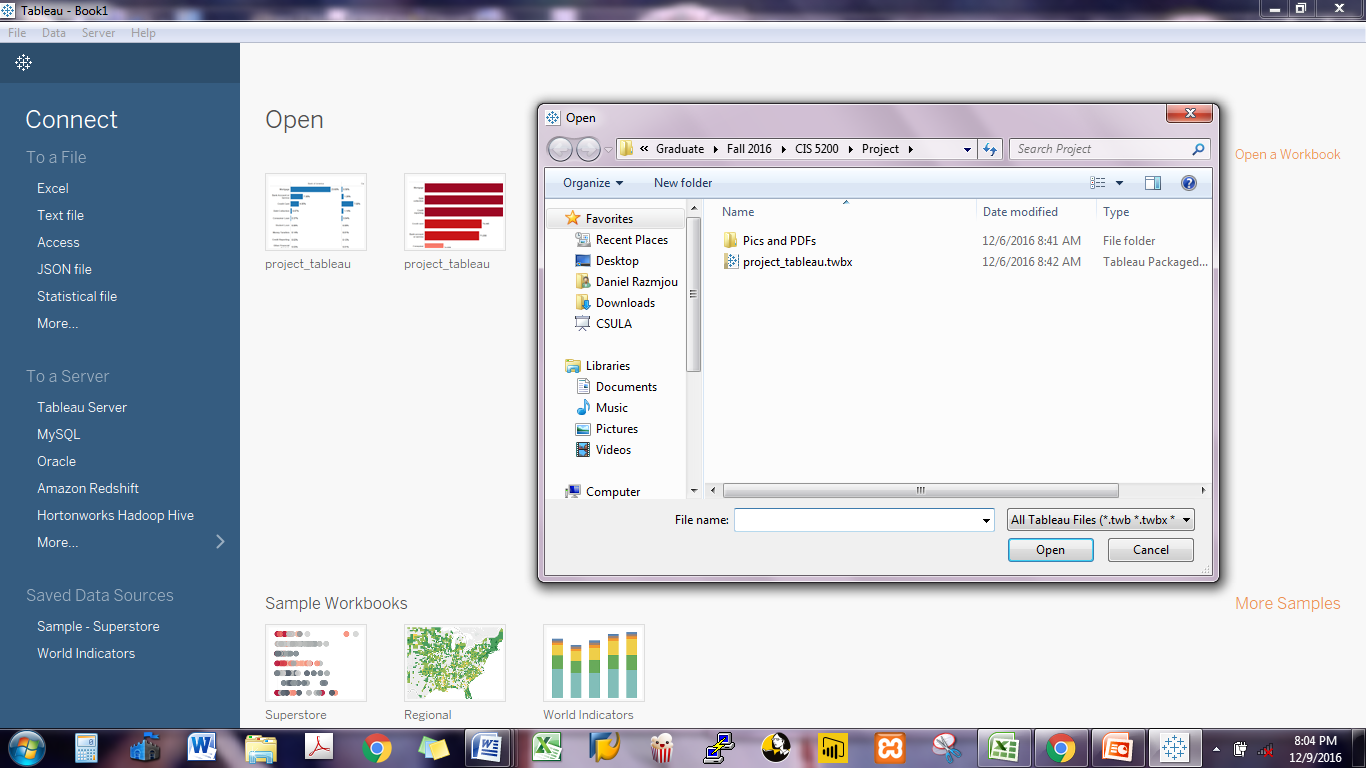


Figure 10. Tableau Interface

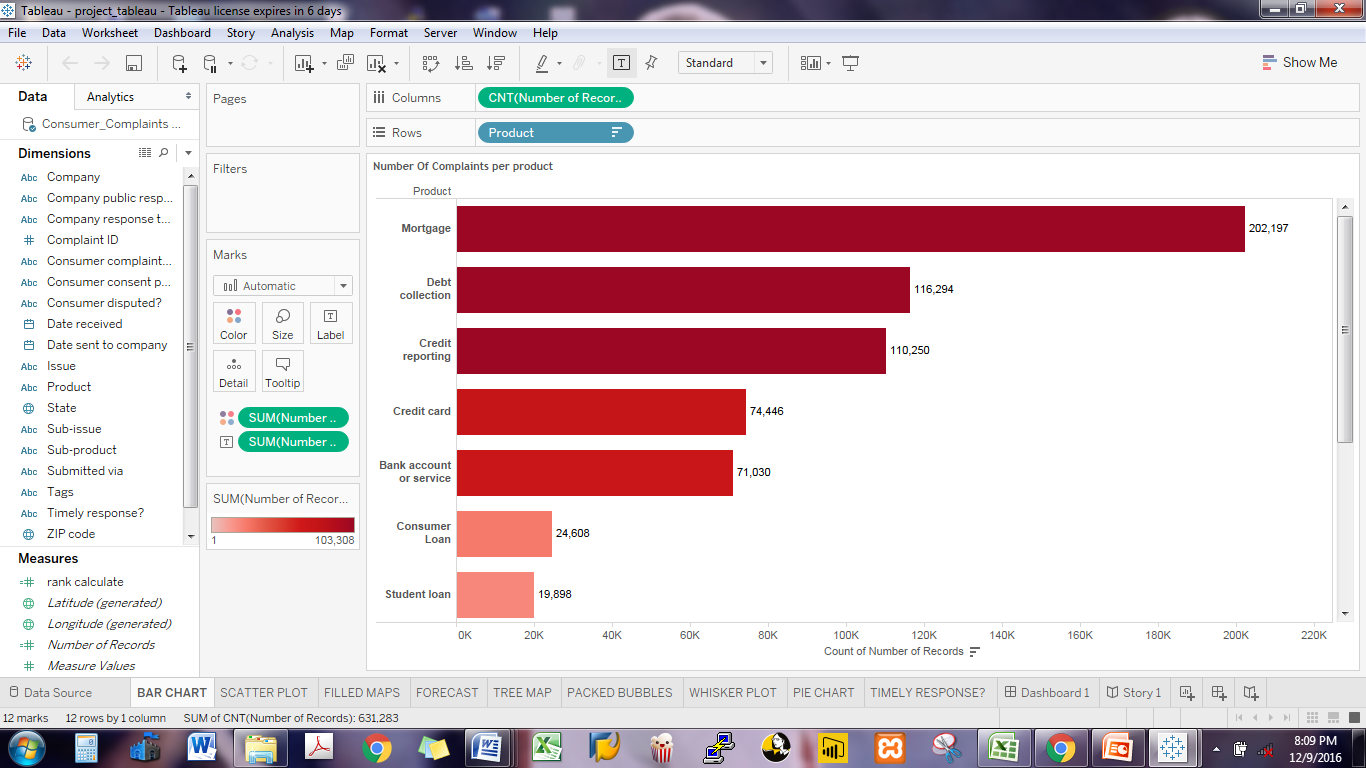
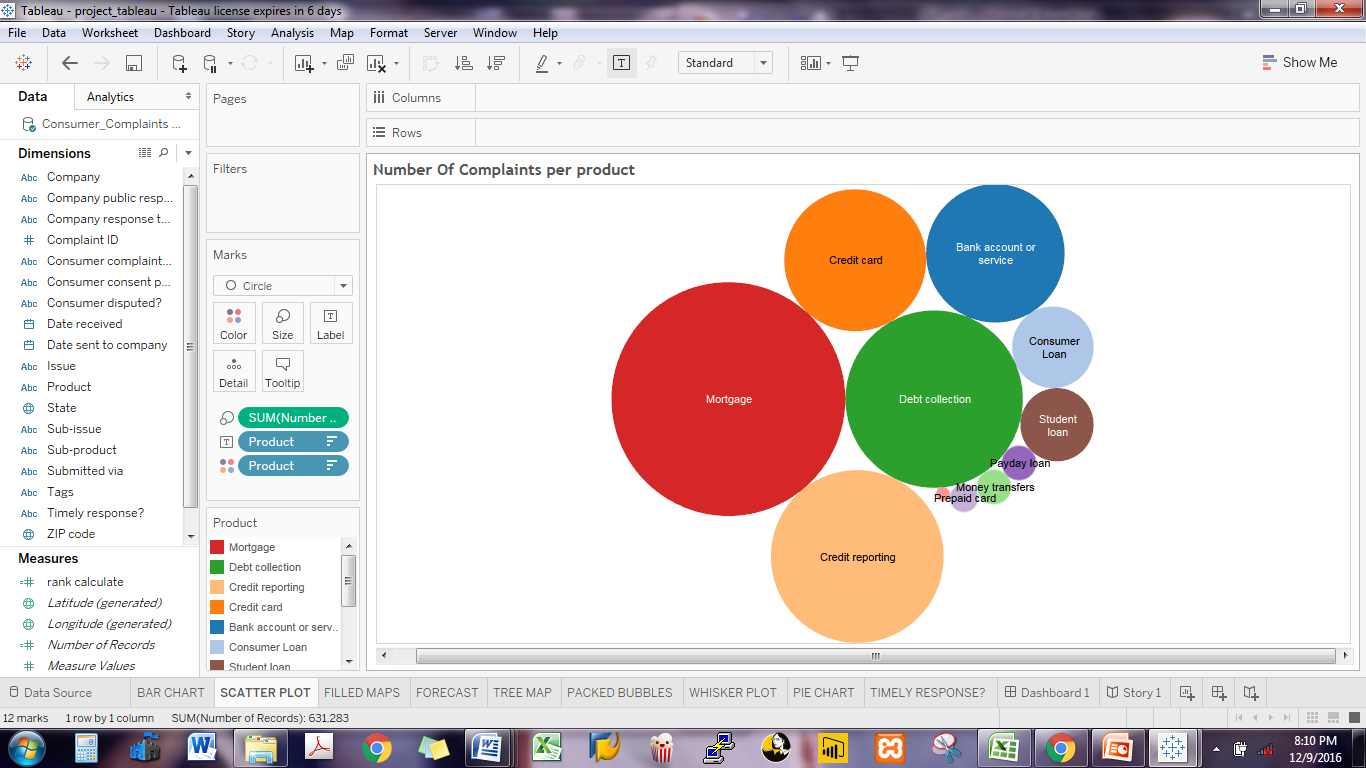
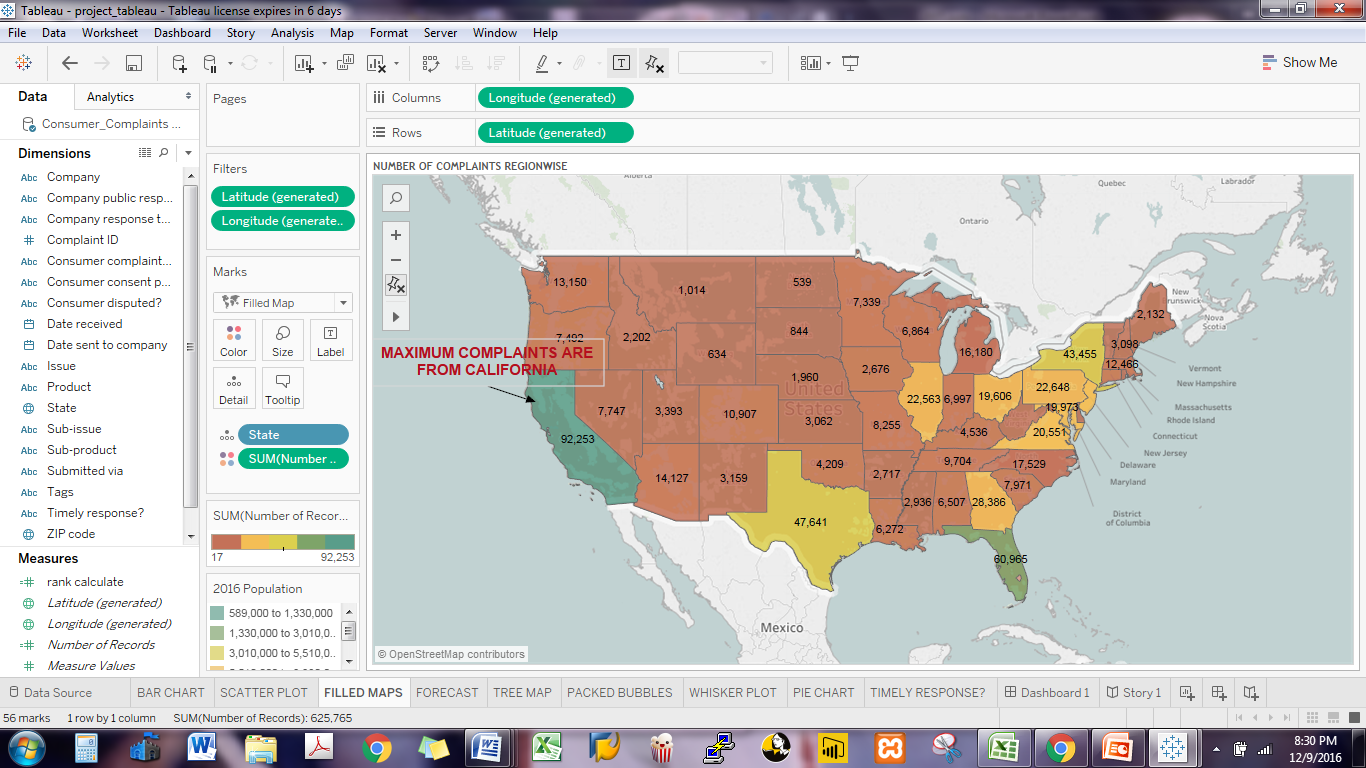
2. First demonstrate all categories in a bar chart by having the total number of records and product categories. This also could be represented in percentages. 

Figure 11. Bar Graph Showing records per product

3. By using the scatter plot the proportional sizes of each category is demonstrated

Figure 12. Bubble packed chart showing records per product

4. With the field maps a demonstration of each state and its relate number of complaints is presented by plugging in the longitude and latitude values. This also could be represented in percentages.

 Figure 12. Filled Map showing areawide records of complaints

5. Also a demonstration of the number of complaints and the forecasted number in the future is presented by using the total number of the complaints and year filters.

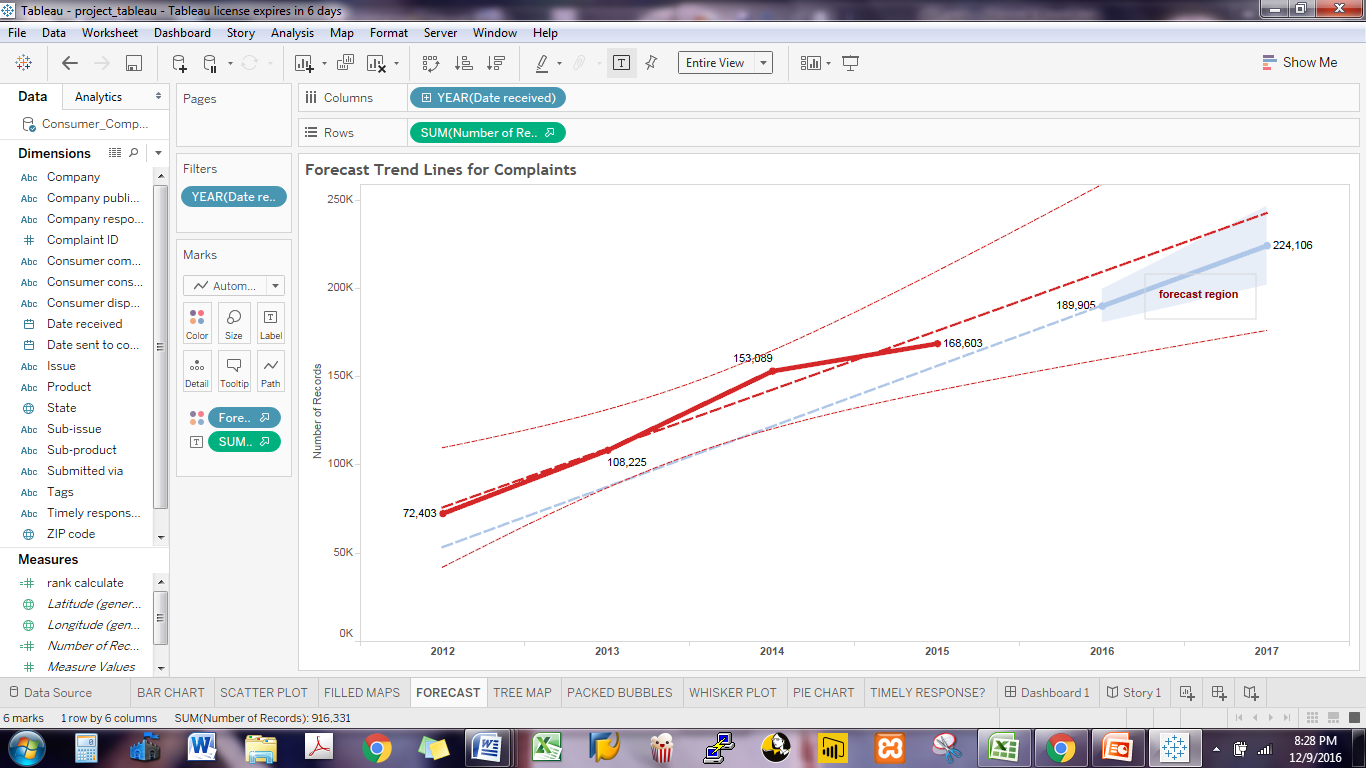


Figure 13. Forecast for next two years

6. A pie chart is used to show the different portions of complaint initiation process by the consumers.

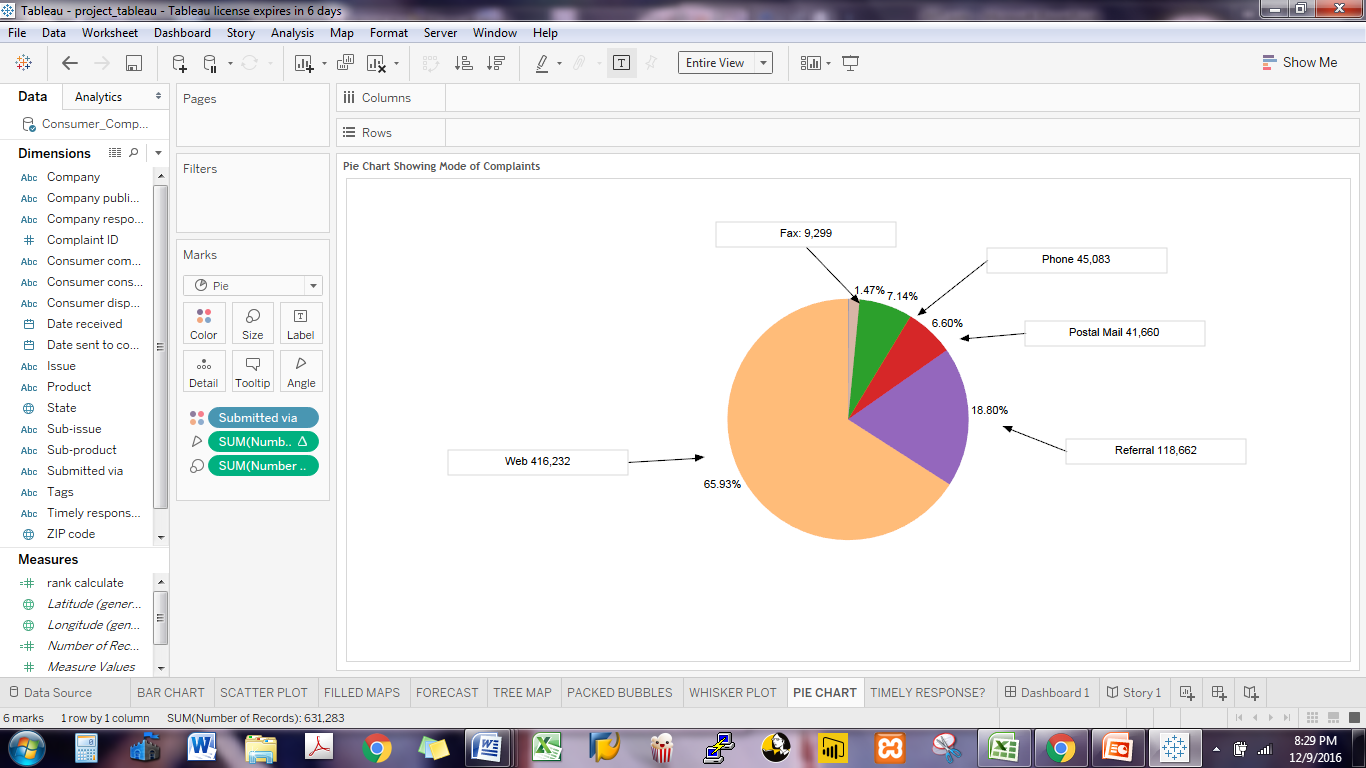


Figure 14. Forecast for next two years

7. A breakdown of response times by the top four banks is shown based on each category of the financial service by using the total number of complaints, products, and company names.

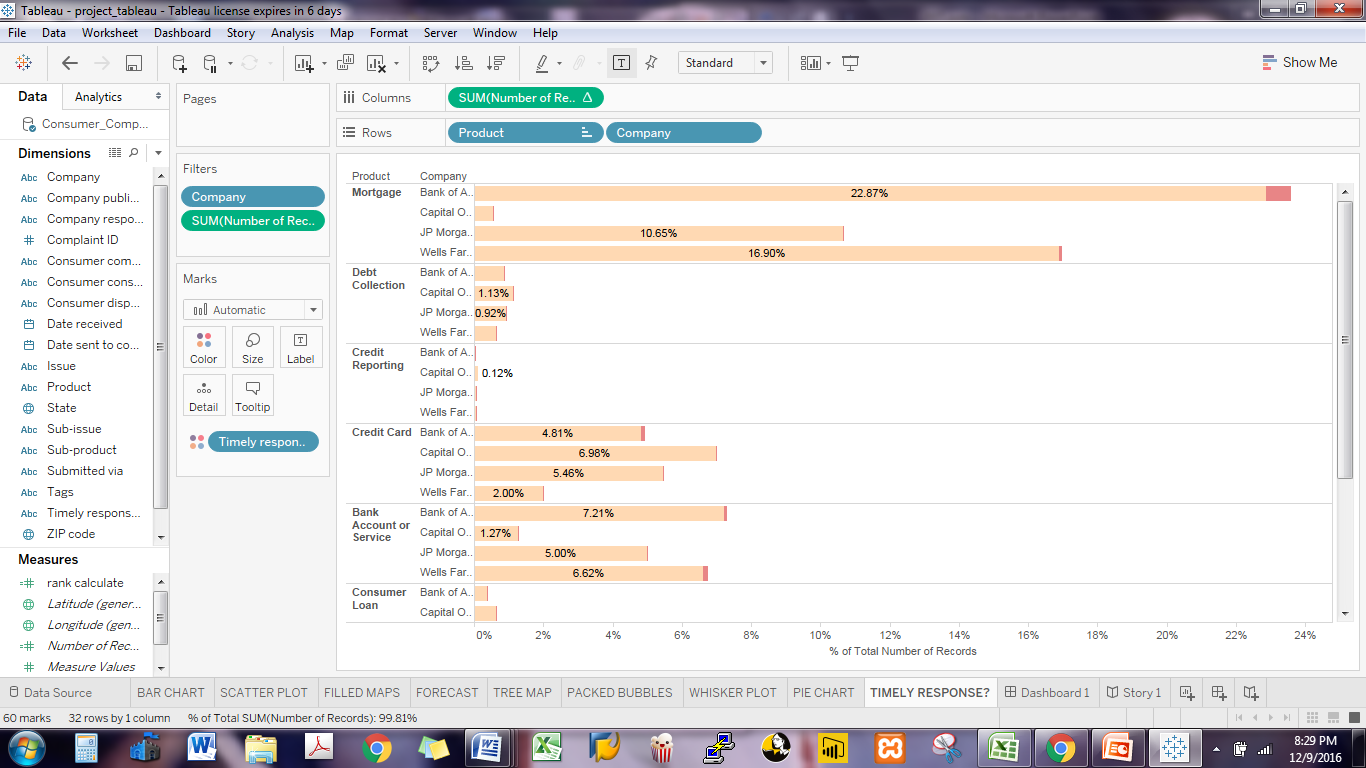


Figure 15. Responses of Consumer Companies towards Complaints.