**Data Analysis on Consumer Complaints in Financial Sector (2011 – 2016)**

**Using Hive in HDInsight Cluster**

1Karan Kaura, 1 Satyam Singh, 1Jobandeep Singh Mahal, 1Daniel Razmjou, 2Dr. Jongwook Woo

1Graduate Student, Department of Computer Information Systems

2Prof, Department of Computer Information Systems

E-mail: 1[kkaura@calstatela.edu](mailto:kkaura@calstatela.edu), 1[ssingh44@calstatela.edu](mailto:ssingh44@calstatela.edu), 1[jmahal2@calstatela.edu](mailto:jmahal2@calstatela.edu), 1[drazmjo@calstatela.edu](mailto:drazmjo@calstatela.edu), 2[jwoo5@calstatela.edu](mailto:jwoo5@calstatela.edu)

**Abstract**

The dataset we have chosen is **Consumer Complaints for Financial Sector, USA** from data.gov. The dataset has a start date from December 2011 to September 2016 and the size is 260 MB. In this paper, we have used Hive on Hadoop, Tableau, Microsoft excel to analyze and visualize big data for consumer complaints in financial sector.

Financial companies are a big part of everyone’s daily life whether it is related to a checking account, a mortgage, or many other services they provide. Like any line of work errors are made daily and disputes are filled to get a resolution, by submitting a complaint, financial companies can hear consumers, get help with their issues, and help others avoid similar ones. Every complaint provides insight into problems that people are experiencing, helping companies identify inappropriate practices and allowing them to stop them before they become major issues.

**Keywords:** *Hadoop, Hive, HDInsight Cluster.*

1. **Introduction**

The dataset that we have used is from Consumer Financial Protection Bureau, which is a third party official website of US government, that addresses financial related complaints by consumers who don’t get a satisfying resolution by referring to the financial company rendering the services. Consumers can submit the complaint for loans i.e. mortgage, vehicle loan, consumer loan, student loan etc., and raise complaints about other product and services such as bank account, credit card, debt collection etc. The agency helps the customer to get resolution of their complaints in timely manner. The bureau keeps the record of the complaints that they receive and how the respective companies responded. With the help of this data consumers can chose the relevant companies which will be a potential match to their needs. Consequently, having the efficient and accurate data will help the customers to have a better overview to make a well-informed decision as to which financial company to choose who are rendering their product or services in the market. Also, consumers will have access to information to make their financial decisions based on the service and customer satisfaction level of the company. The information provided could be used by the industry leaders to examine the areas of services that lacks proper customer satisfaction or faces many complaints due to errors. As the dataset demonstrates in this case, the mortgage industry faces the highest number of unresolved complaints by the clientele. Financial companies such as Bank of America, Capital One, JP Morgan & Chase, and Wells Fargo could use this data to internally analyze and find out which areas of the business needs more attention or better training of employees to reduce the number of initial errors that trigger these complaints. Financial industry leaders will also have access to times of the year that the volumes of complaints is higher which indicates the higher demand of services requested. Every year, companies spend millions of dollars to resolve these complaints which lowers the overall profitability of a financial institution where it can generate higher revenue and maintain its competitive advantage among the competitors. [1]

1. **HADOOP**Hadoop supports processing of large data sets in distributing computing environment and consists of MapReduce, Hadoop distributed file system (HDFS) and contains many related projects such as Apache Hive, and HBase. Analysis on consumer complaints dataset was facilitated by hadoop as it provides cluster for storing and analyzing large amount of unstructured data, which run on low-cost commodity computers. Cluster is a group of loosely or tightly connected computers that works together as a single system. One of the core functions of Hadoop is that it offers to create multiple data blocks with replicas and then distributes data across various computer nodes to enable reliable and rapid computations.
2. **SYSTEM REQUIREMENT**

* Number of worker nodes used is 4.
* Number of cores used is 4.
* RAM used is 4GB

1. **FLOW**  
   We downloaded the data of consumer complaints from data.gov website in .csv format. After this, we uploaded the .csv file in Ambari. After uploading the file, we have used hive and written queries to create the tables and dumped the data into the created table. We have then created queries to filter the data from main dataset to analyze in tableau, we did this as tableau can’t handle this large dataset. Finally, by using tableau version 10.1, we have created the bar charts, bubble packed chart, filled maps, forecast, pie-chart etc. The flow of the data was more clearly shown in the picture below.

Raw Data

Login to Hadoopcluster

Export results in MS Excel

Query Data in Hive

Create Tables in Hive

Load Data in HDFS

Create Visuals using Tableau

Figure 1: Flow diagram

1. **HIVE**

Hive is a querying language, which offers similarities to SQL. Hive’s infrastructure is built on top of Hadoop. Hive was developed at Facebook and it is open-sourced for free download. The primary purpose of Hive is to query and analyze large datasets across various nodes. Hive is a high-level interface, which allows us to execute MapReduce jobs across a Hadoop cluster. MapReduce is the system used to process data in Hadoop cluster. One of the limitations of RDBMS is its lack of access and analytical capabilities towards unstructured data. Hive closes the gap on this limitation. [2]

1. **QUERY TO ANALYZE AND VISUALIZE CONSUMER COMPLAINTS DATA**

DROP TABLE IF EXISTS COMPLAINTS;

CREATE TABLE COMPLAINTS AS

(

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 ',‘ ;

We have used the above query to create the table.

LOAD DATA INPATH '/user/jmahal2/complaints.csv' INTO TABLE COMPLAINTS;

The above query is used to dump the csv file into the table, delimited by “,”.

1. **VISUALIZATION TOOL: TABLEAU**

We have used tableau to show the analysis, big data is nothing without the analysis tool. We all know that big data is powerful, however without visualization a lot of the knowledge may go unrecognized without making much logical sense. Often, data visualization allows data sets/data objects to show its true potential.

**Tableau**: Tableau allows users to connect to external data sources and quickly convert the data into interactive BI reports and dashboards. We have plotted a bubble chart-using tableau for number of complaints per product. A bubble chart is a stylized way of visually representing occurrences of words used to describe tags. The biggest circle indicates highest count followed by smaller circles displaying count in descending order.

Query used: select product, count(complaint\_id) from complaints order by complaint\_id;

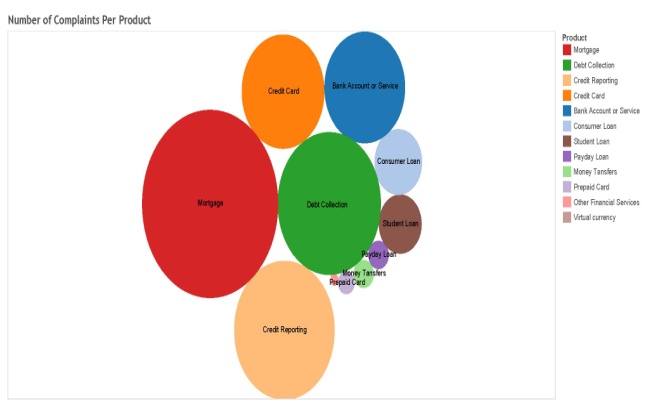


Figure 2: Bubble chart for number of complaints per product

We have plotted geographical map which shows the number of complaints in each state of the United States of America. It is observed that California has maximum number of complaints followed by Florida and Texas.

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

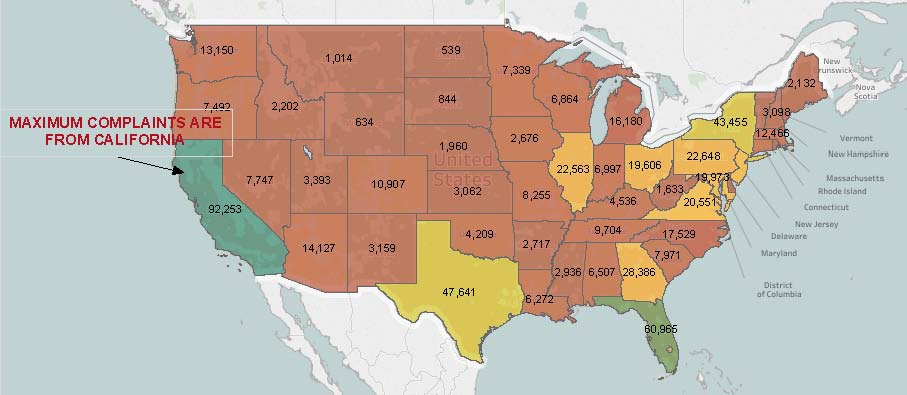


Figure 3: Filled map showing state wise complaints

We have made trend line chart using Tableau for analyzing the trend of complaints and forecast the complaints for the coming year. The trend shows that the number of complaint are increasing every year and the forecast shows that the complaints will rise to 225,493 year 2017.

Query: select count(complaint\_id), date\_recieved from complaints;

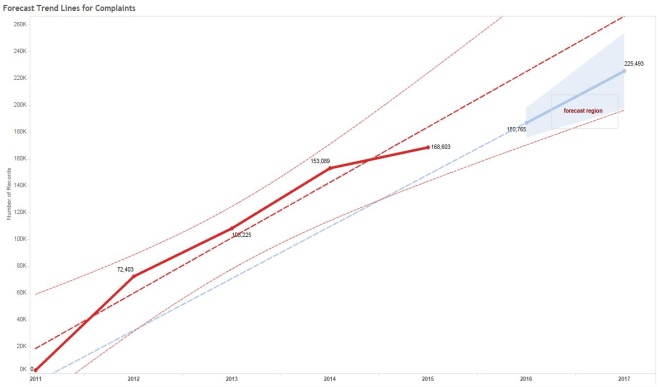


Figure 4: Trendline showing trends of complaints each year (2011-2017)

In the analyses, we figured out the mode through which people used to complain about the product or services. The maximum number of complaints were through web, followed by referrals and phone.

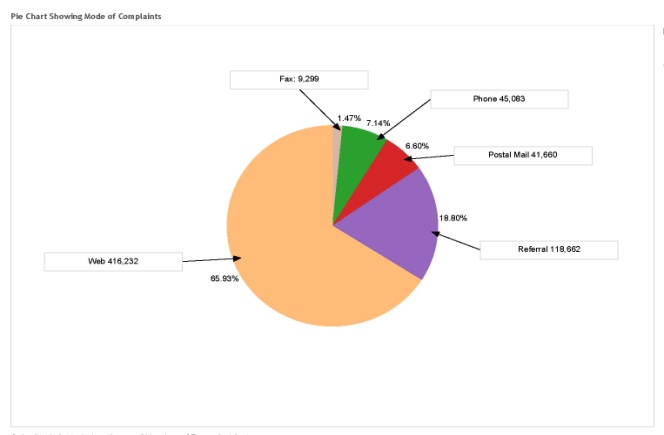
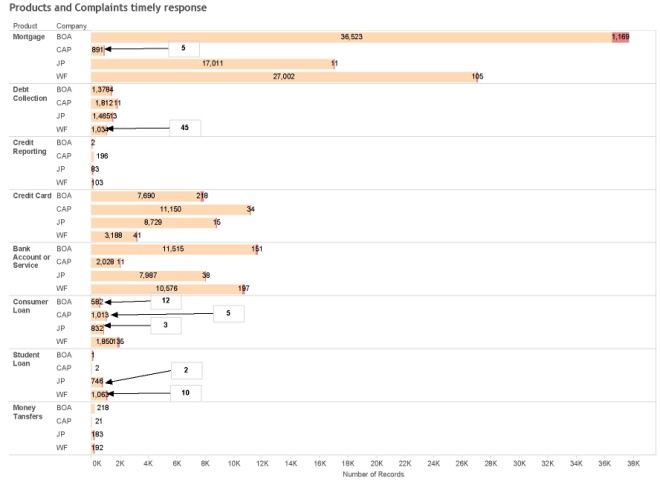


Figure 5: Pie chart showing mode of complaints

We made a bar graph, in order to analyze that in how many cases the companies were able to give timely response to their clients and how in how many cases there were delay in response.

Query: select product,company,count(complaint\_id), timely\_response from complaints group by product;

Figure 6: Timely/Untimely response to complaints

In one of the interesting analysis we found out that the financial companies pay monetary relief to thousands of customers each year to close the customer complaints. These are the complaints which could probably be addressed at the first place if given priority and proper attention. Companies still have around 3700 complaints which are in process for year 2016. Around 1150 complaints were not responded timely in year 2015 and 972 complaints are already under untimely response in year 2016.

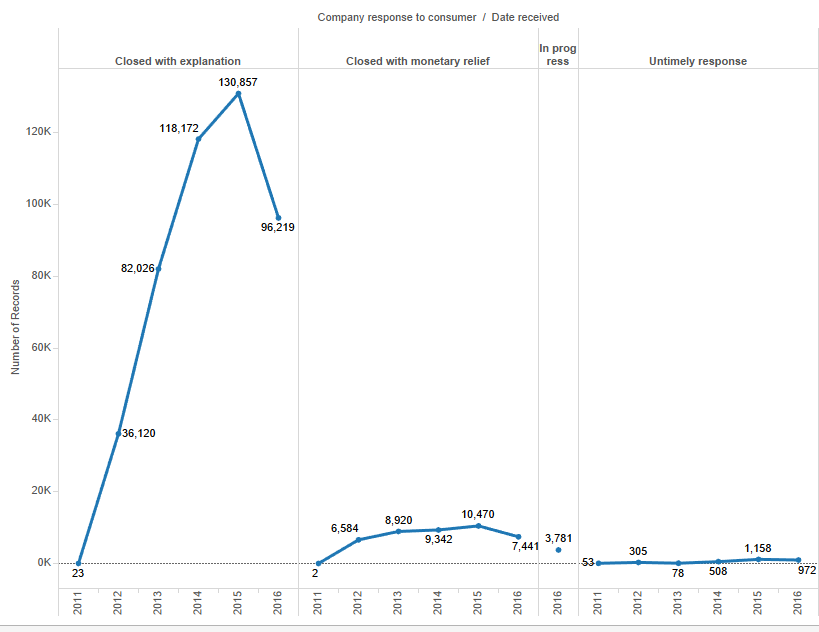


Figure 7: Companies response to customers

1. **GITHUB**

GitHub is a version control repository hosting service. It offers all the distributed revision control and source code management (SCM). The entire team had individual access to GitHub as contributors and the code developed has been uploaded on GitHub along with the tutorials/steps that could be followed to implement our exact project.

1. **CONCLUSIONS AND LEARNINGS**

We transformed big set of raw Customer complain data into information, and information into insight by generating rich visuals, graphs and charts using Hive and visualization tools. After analyzing five years of customer complain data, we came up with following results:

1. Highest number of complaints were of mortgage,

debt collection, credit reporting and credit cards.

2. The maximum number of complaints were from

California, Florida, New York and Texas

3. A lot of customer complaints are not addressed in

time.

4. The trend of customer complaint are increasing each year. Forecast shows 225,493 complaints in year 2017.

The customer will be able to decide as to which company to choose for loans or other financial services from the list of companies. The customer will be able to analyze which companies have minimum number of complaints product wise and which companies address customer complaints with utmost concern.

The companies in financial sector should prioritize the complaints in order of importance. Companies should make sure the complaints are resolved within the turnaround time in order to avoid further problems.

The companies should deploy their human resources in that states where the number of complaints are increasing day by day. Satisfied customers will lead to smooth functioning and help increase the profitability of a financial institution where they can generate more business via references from existing client base and have competitive edge in the market.

**REFERENCES**

1. Consumer Financial Protection Bureau <http://www.consumerfinance.gov/complaint/>
2. Data.gov

<https://catalog.data.gov/dataset/consumer-complaint-database>