Lending Club Loan Data Analysis

Group B
David Sayad, Chris Wallace, Riki Chang, Carolina Rivera,
California State University Los Angeles
E-mail: dsayad2@calstatela.edu, cwallac9@calstatela.edu, criver47@calstatela.edu, rchang12@calstatela.edu

Abstract: Our aim is to take this loan data from the Leading Club Company and create a geospatial analysis of each state in the United States of America. This will include state by state information on Home Ownership, Loan Status, and Purpose of Load for each state.

1. Introduction

We took data from kaggle.com, a database of public datasets, tutorials, and machine learning job assets service.

This dataset offers 887379 entries of data totalling at 392 MB. Each line includes a great amount of data that is not used for our aim such as: loan amount, term, interest rate, installment payment, grade, annual income, payment installations, etc. [1]

Narrowing down this information into a form which can be properly analyzed will bring much insight. This will be changed into be in terms of total loans given in a month or state-by-state allocation.

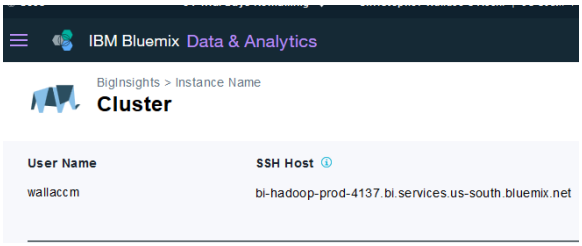
2. General Instructions

2.1 BigInsights

You need to open up your BigInsights account or create one on IBM's website if you do not have one [2]. We have included two links below the first being IBM's Bluemix and the second being the link to our dataset. You can download the data file Lending Club data from Kaggle which is what we did. [1]

2.2 Hadoop

Once you have logged into your BigInsights copy your ssh link.



Start the logging in process in putty or terminal. Once in terminal do as follows:

$ ssh username@yourlink.com

You will be prompted to enter your password next.(it will be invisible when you type.. Do not be alarmed it is recording your keystrokes)

2.3 Import Data

After logging in enter the next command to create a directory inside HDFS that will be used to store the data.

$ hdfs dfs -mkdir tmp/project;
$ hdfs dfs -mkdir tmp/project/tables;

Next we will manually upload the data through Ambari. Navigate to the same page where you grabbed your login link and click on the link entitled Ambari. Use the same username and password that you used to log into HDFS. After logging into Ambari do as follows:

FileBrowser > User > (username) > tmp > project > tables > Upload > *local place you stored Data*

2.4 Hive

Start up your hive shell by entering the following command.

$ hive

In the hive shell we will upload the data into a table stored in the tables directory. Next we will create a table that will extract the columns of the data that we are looking for.

```
hive>CREATE EXTERNAL TABLE IF NOT Exists project( json_responce STRING) STORED AS TEXTFILE LOCATION "/tmp/project/tables";

hive>CREATE TABLE IF NOT EXISTS project( id BIGINT, loan_amnt BIGINT, home_ownership STRING, issue_d STRING, loan_status STRING, purpose STRING, addr_state STRING);
INSERT OVERWRITE TABLE project SELECT id, loan_amnt, home_ownership, issue_d, loan_status, purpose, addr_state FROM credit_card_dataset_CSV.csv WHERE id = id
```

To be sure that both command work type in the following command to look at all the tables in your hive shell.

hive> show tables;

Once it is verified that the tables have been created you can execute a number of commands to query the table further.

Example:

hive> Select * From project limit 50;

Which translates to select the first fifty rows from the project table.

3. Visualization of Data

3.1 Power View

To visualize the data you must download the csv file from Ambari File Manager. From then you open this file in excel and save it as an excel file(xlsx). From there navigate to the **Insert** tab and select **Power View** to begin a Power View Report. From here you select your **Power View Fields** that you would like to visualize(date, loan amount). Make sure to order and format the date in order to output the correct graph. Next from the **Design** Tab select **Other Charts (Bar)** to display a bar graph of Total Loan Amounts given by the Leading Club (Appendix 1).

3.2 3D Map

Visualizing this data in a state by state basis brings much insight to the needs of the state population. To do this you must have Data Analysis add-in enabled in excel. After that navigate to the Insert Tab and select **3D Map.** From here you must create different layers to show different attributes of data. Select **Add Layer** and select **Bubble** visualization. From here the **Location** of the data will always be *addr_state* and you must select **State/Province.** Then for **Category** select the attribute you would like to to visualize. For our project we selected to have it in terms of Home Ownership (Appendix 2), Loan Status (Appendix 3), and Purpose (Appendix 4).

4. Analysis

4.1 Temporal Analysis

The temporal analysis can show a great amount of knowledge to the loan usage of the general public of the United States. As one can see there are major spikes of total loan amounts in the months of March, July, and October (Appendix 1). They are tapered off greatly after these spikes also until the next month in this cycle is approached. The end of Winter and the beginning of Spring is

typically the time when people buy homes and this could explain the spike in March. Colleges start towards the end of August and many loans for students need to be taken out, this could explain the spike in July. October is the highest total loan amount and one could rationalize this to the holiday season and the expenses involved with them.

4.2 Geospatial Analysis

The geospatial analysis was also showed many trends in the United States with loan receivers: home ownership status, loan status, and purpose of the loan. It is very interesting that in most states about half of the people within this dataset own their house with the notable exception of California and New York (Appendix 2). This can be explained by the well-known cost of living in these states and the amount of capital is needed to own a home. In terms of Loan Status per State there does not seem to be much variability between the states (Appendix 3). The only notable exception being that of California having a slightly larger percentage of Fully Paid loans that being 27.8%. Lastly in terms of Geospatial Analysis, we have the Loan Status per State (Appendix 4). There does not seem to be a distinguishable trend from state to state everything is very similar.

6. Conclusion

After loading the data into excel we gained insights into how all of this lending data can be visually observed. By looking at the maps provided we can see the total loan amounts(Appendix 1), the loan status(Appendix 3), the purpose of the loan(Appendix 4), and proportionally what kind of loans people are taking out for home ownership(Appendix 2). For example, in the Geospatial Analysis we found out that Californians tend to take out more loans for rent where Texans they tend to take out loans for mortgages. We expect that this due to the fact that the price of owning a house is very high in California as opposed to Texas where it is much lower. In conclusion, there are a number of ways this data can be interpreted and we will leave it up to you to continue to explore and find helpful insights into the world of Big Data.

5. Key Terms

Ambari: The Apache Ambari project is aimed at making Hadoop management simpler by developing software for provisioning, managing, and monitoring Apache Hadoop clusters. [5]

Analytics: The process of collecting, processing and analyzing data to generate insights that inform fact-based decision-making. In many cases it involves software-based analysis using algorithms. [3]

Apache Pig: a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. [4]

BigInsights: It's a software platform designed to help firms discover and analyze business insights hidden in large volumes of a diverse range of data—data that's often ignored or discarded because it's too impractical or difficult to process using traditional means. [7]

Flume: A distributed and reliable way to collect, group, and transfer large amounts of data from many sources to a central data store.[3]

Hadoop: Apache Hadoop is one of the most widely used software frameworks in big data. It is a collection of programs which allow storage, retrieval and analysis of very large data sets using distributed hardware (allowing the data to be spread across many smaller storage devices rather than one very large one). [3]

HCatalog: Makes metadata (metastore) for Hive and merges it with what Pig does.[3]

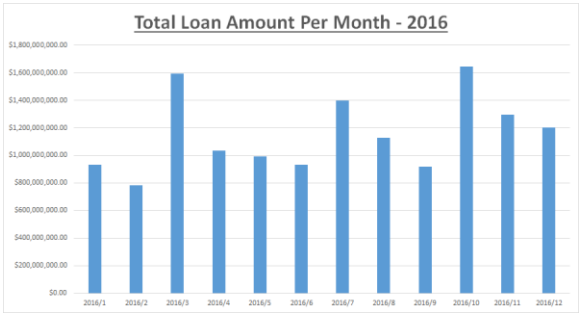
HDFS: Hadoop Distributed File System; the way that Hadoop structures its files. [3]

Hive: A higher level language that uses HQL, which is similar to SQL (Structured Query Language) in its syntax. [3]
MapReduce: Refers to the software procedure of breaking up an analysis into pieces that can be distributed across different computers in different locations. It first distributes the analysis (map) and then collects the results back into one report (reduce). [3]
Sqoop: Apache Sqoop(TM) is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases. [6]

References

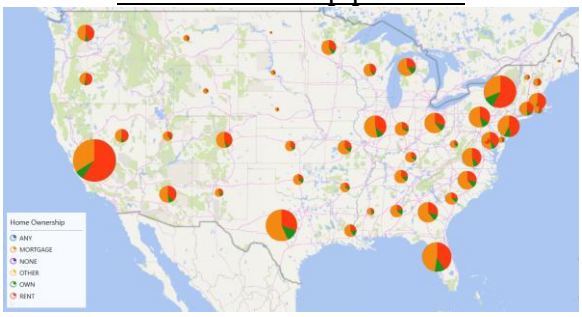
[1] Kan, W. "Lending Club Loan Data." Retrieved from https://www.kaggle.com/wendykan/lending-club-loan-data
[2]https://www.ibm.com/cloud-computing/bluemix/
[3]"Big Data: The Key Vocabulary Everyone Should Understand"(n.d.). Retrieved from https://www.linkedin.com/pulse/20141203075716-64875646-big-data-the-key-vocabulary-everyone-should-understand
[4]"Welcome to Apache Pig!" Retrived from https://pig.apache.org/
[5]"Apache Ambari" Retrieved from http://ambari.apache.org/
[6] "Apache Sqoop" Retrieved from http://sqoop.apache.org/
[7] "Understanding InfoSphere BigInsights" Retrieved from http://www.ibm.com/developerworks/data/library/techarticle/dm-1110biginsightsintro/index.html

Appendix 1



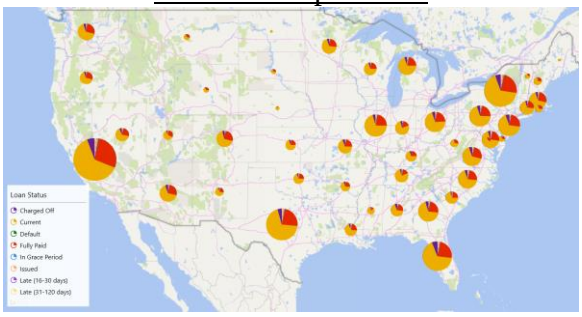
Appendix 2

Home Ownership per State



Appendix 3

Loan Status per State



Appendix 4

Purpose of Loan per State