# Hive Hadoop based approach to analyze the lending club loan data

Lalit pathak [x18110088]
School of Computing
National College of Ireland
Dublin, Ireland

**Abstract – Lending club is USA based company which provides the platform for the investor as well as the borrower. From the perspective of the borrower it is important to understand the loan process and other Parmenter such as interest rate and from the investor risk in investment is important. Lending club works as mediator and provide the services to borrower as well as investor. This paper discussed interest analysis, geographical analysis to understand about the business as well as effect of geographical area on the business. This analysis is useful for the business as well as for the loan applicant.**

***Keywords – Lending club, loan data analysis***

## *I.* INTRODUCTION

The 'Lending Club' is US based peer to peer lending company, which is traded on NYSE and offers lender with securities and also offer loan trading on a secondary market. This platform enables borrowers to create unsecured personal loans. In addition to this it offers business loans and the company provides quality, responsible options for people looking to finance elective medical procedures [1]. As the name suggests, P2P loans are generally personal loans. However, small business owners often rely on their personal and business finances so as overall P2P lending grew, so P2P borrowing for small business purposes [2].

Loan purpose is not one of the criteria considered when evaluating loan applications, we find that loans intended for small business purposes were possibility of been funded than loans for other purposes. We then look at the interest rate paid on those loans that did get funded [2]. The platform brings the investors and borrowers on single platform and transformed the way of using the credits. LendingClub enables borrowers to create loan listings on its website by supplying details about themselves and the loans that they would like to request. All loans are unsecured personal loans and can be between $1,000 - $40,000. On the basis of the borrower's credit score, credit history, desired loan amount and the borrower's debt-to-income ratio, LendingClub determines whether the borrower is credit worthy and assigns to its approved loans a credit grade that determines payable interest rate and fees. The standard loan period is three years; a five-year period is available at a higher interest rate and additional fees. The loans can be repaid at any time without penalty [3].

"The lending marketplace uses technology to operate an online credit marketplace - with no branch infrastructure - at a lower cost than traditional banks, passing the savings on to borrowers in the form of lower rates and better returns for investors, between 8% to 25%. Borrowers that apply for loans (up to $35,000) need to provide details about themselves (income, credit score) that Lending Club uses to calculate a credit "grade" to determine interests (ranging from about 8% to 25%) and fees. Once approved, the loan is listed on the site to prospective investors who can then fund it in increments of just $25" [4, p.1].

## II. RELATED WORK

A study by Serrano-Cinco and others – [5] discusses p2p lending and the things that result in the loan default. The study states that on platforms like lending club individual lenders are always at high risk. The financial firms are at advantage in this case. Considering the manpower that they can use. P2P sites assign grades. These grades on the other hand need improvement. The study shows the repayment percentage has decreased. But it can be seen that, as the comparison periods are different, correct grounds to do this can not be easily found.

Similar study has been performed by Bachmann and others. In this study a detailed discussion about the relation between characteristics of the borrower and possibility of the borrower getting loan is done. This study also comments about the unknown facts about the comparison between per to peer lending and the traditional system of credit and lending. This study also emphasizes that along with the characteristics of borrower, the characteristics of lender also play important role in success or failure of lending proposal.

In a study performed by Chen and others – [7], focused on the peer to peer lending in china. The study tried to find the impact of trust in the peer to peer lending system. The information provided by the borrower plays an important role. With good quality information, the lenders can show more trust. This can have a positive impact on the possibility of the borrower getting loan. A study byMingfeng and others – [8] on the analysis of friends' network in which the lender resides. The study suggests that the good network structure increases chances of funding at lower interest rates.

***Research question :***

*1) Can we analyze the relation between amoun requested by borower and amount sanction to the borower?*

## III. METHODOLOGY

Borrowers and Investors connected to each other by one of the largest online marketplace Lending Club. This online platform makes a loan more affordable for the borrowers and rewarding for the investors. Lending club is basically a bridge between the borrowers and investors.

Dataset of the lending club has been downloaded from the Kaggle [9]. This dataset contains complete loan data from the

year 2007-2018. Size of the dataset after extracting the zip file is 1.1GB with 145 columns. This dataset is useful to build the hive and Hadoop based framework, for the project.
Raw Dataset Creation. For analysis, few modifications have been carried out with the actual dataset.

*A. Technology used in the Project*

HIVE:

Hive is a data warehousing package on the top of Hadoop.
Hive query language (HQL) is used in the hive.
To handle the large number of the dataset for the processing hive tool is used with Hadoop. In this project, the hive is used to analyze the processed data from the MapReduce. Hive queries are created for the analysis.
Version: Hive 1.1.0-cdh5.13.0

MySQL:

Now a day's vast amount of unstructured and semi-structured data is available.
To store and process the structured data MySQL database is most popular.
which provide fast query results. Customer dataset is stored in the MySQL database.
This dataset is small in size. Performance of the MySql is better on the small structured dataset.
Version: MySQL 5.1.73

Sqoop:

Sqoop is the tool design for importing the data between relational databases such as MySQL, Oracle into the Hadoop HDFS and exporting the data from the Hadoop file system to the relational database. In this Project customer data which is present in the MySQL is loaded into the HDFS and hive tables are created on this data.
Version: Sqoop 1.4.6-cdh5.13.0

Shell Scripting:

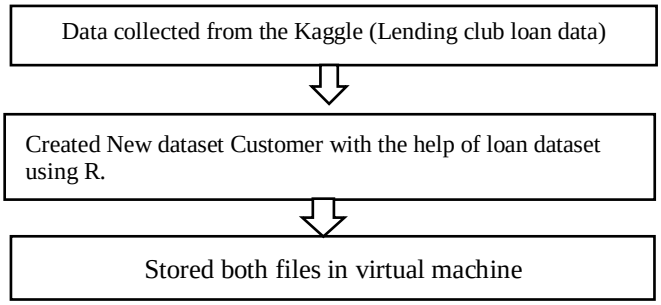
A shell script is the program design for Unix or Linux. Shell scripting can be used for operations such as manipulation, program execution, and printing text.
Scripting in this project is used to automate the program execution for the different technologies

HBase: HBase is used for Real-time random read-write access for the big data. In this project, this database is used to store the results of comparatively less size hive queries from HDFS.
HBase 1.2.0-cdh5.13.0

Tableau:
For the visualization and to get insightful information from the dataset tableau is used. Version:2018.02

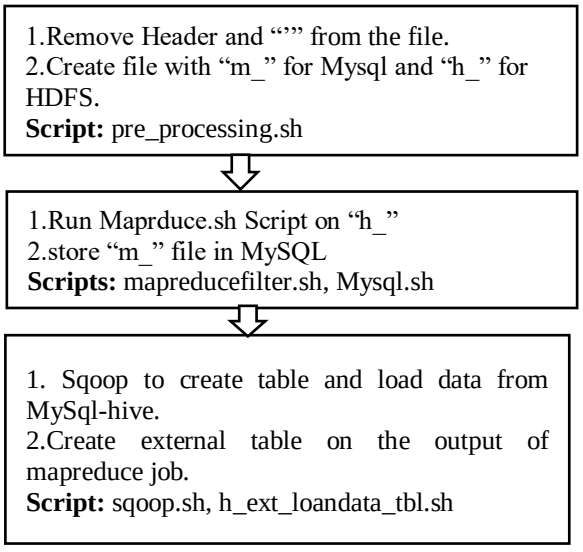
**Data Acquisition and Preprocessing:**

Processing of the data is divided into two parts Pre-Processing and Post-Processing
Pre-Processing is divided into six different parts




1) Creation of raw file.
   - Member id is generated as dataset doesn't contain values for this column.
   - Dataset customer is created, with column name zip code, address state, Member id from the actual dataset.
   - Column Gender and Birth date has been added to customer dataset.
   - loan dataset has been created by selecting 73 columns and by removing address state, zip code and other columns from the actual dataset.
   - Issue_d column from the loan dataset is split into two columns namely month and date.
   - These Raw datasets have been manually placed into the Cloudera virtual Machine.
   - Name and address have not generated for the customer dataset because of the GDPR.

**Data Pre-Processing**



2) Bash commands to clean the header and other file.

   Bash command is used to remove the "" from both the raw files. Header is also removed from both the files.
   File which will be stored on the MySQL is saved with the 'm_' pre-fix and HDFS file is saved with 'h_' on the local file system.

3) Filtering the data through MapReduce
   File with the pre-fix 'h_' is used as an input file for the map reduce program. Jar file for the Filtering the dataset is created through eclipse. This jar file is used in the mapreduce.sh. This script load data from input location on the HDFS and then apply the jar file on to input data, directory with processed file has been created on the HDFS.

4) MySQL table creation
   customer.sql and customer_data.sql has been created. These two files used in the Mysql.sh to create and load the data from the local file system with 'm_' pre-fix into the MYSQL table.

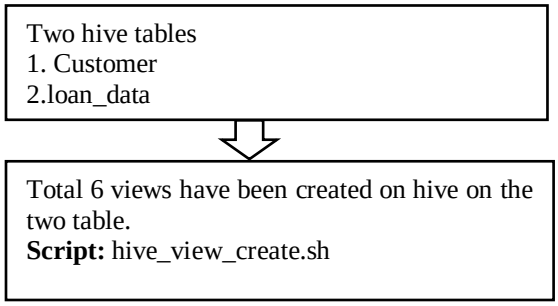
5) SQOOP
   Script sqoop.sh has been created to load the data from MySQL table into the hive table.
   hive-import command first loaded data from MySQL database into the HDFS temporary location and then executed the hive table create statement and finally loaded the data from HDFS file location into the hive table.

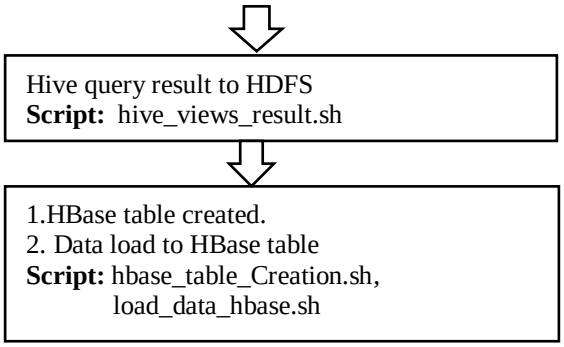
6) External hive table creation.
   Output of the map reduce file is stored int the processed directory on the HDFS. An external hive table has been created through the script h_ext_loandata_tbl.sh.

Hive queries has been created on this table.

**Data Post-Processing**

Post processing have been done with the hive and HBase.
Queries have been created in the to analyze the data.
Result of this analyze data is stored into the Hbase database.
Scripting is used to automate the process of view result generation, HBase table creation and data loading.




**Hive Views:**
To analyze the data 6 views have been created through hive_views_create.sh script.
Views has been created on the two processed table Customer and Loan data. Below is the description of the views.

lon_amt :

This view is created on the loon_amt field of loan_data table. This view is used to identify how much loan applied by the borrowers.

funded_amnt :

This view is created on the funded_amnt filed of the loan_Data. This view is used to identify the amount funded by Lenders.

avg_int_rate :

View is created on the loan_Data table and used to calculate the average interest rate against the term.
Term column contain two values 36 and 60.

loan_purpose :

This view is again created on the loan data table and used to understand the purpose of the loan. and in which categories most of the loan has been applied to the lending club

annual_inc_by_state :

Two Hive tables are customer and loan_data.
This hive tables contains a same column and this column have been used to join both the table.
This view is based on the two tables Customer and loan_data. Purpose of this View is to understand the average annual income by the state.

avg_int_rate_by_state :

This query is again based on the columns from the two table int_rate from loan_Data and state from the customer. Join condition is on Member_Id from both tables. Purpose of this view is to understand the average interest rate distribution in each state.

To visualize the data through Tableau and to analyze the data in the future these query result has been stored on the HBase database.

***hive_views_result.sh*** Script have been created to load the result of this queries into the HDFS location.

**Storing the Queries Result**

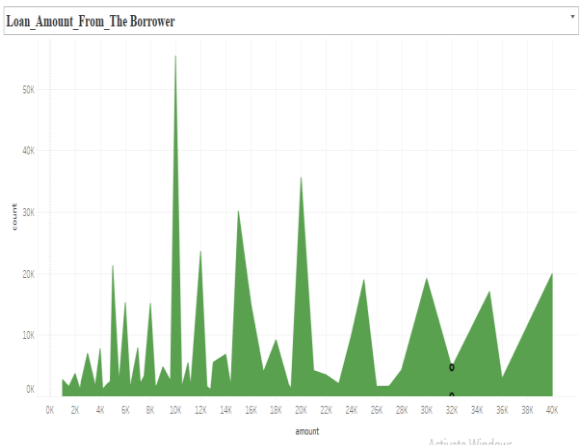
***hbase_table_Creation.sh*** script has been created to create the respective tables into the HBase database. After this it is important to load the data from the HDFS directory into the HBase table

***load_data_hbase.sh*** script is used to load the data from HDFS location into the HBase Table.

## IV. RESULT

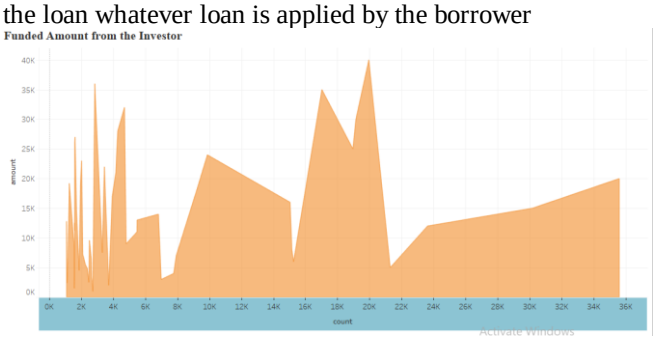
1.Maximum loan amount requirement from the borrower is in the range of 20000- 40000 USD.

.

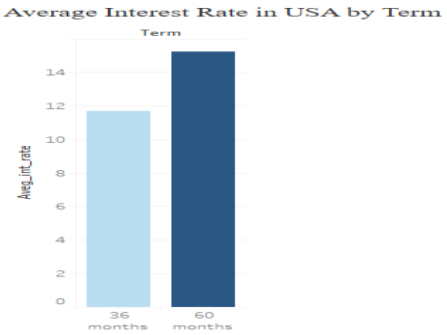



2.

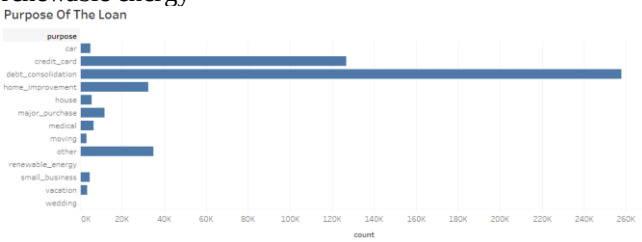
Maximum invest from the Investor is in the range of 20000 USD to 40000 USD. This means that potential borrower gets the loan whatever loan is applied by the borrower



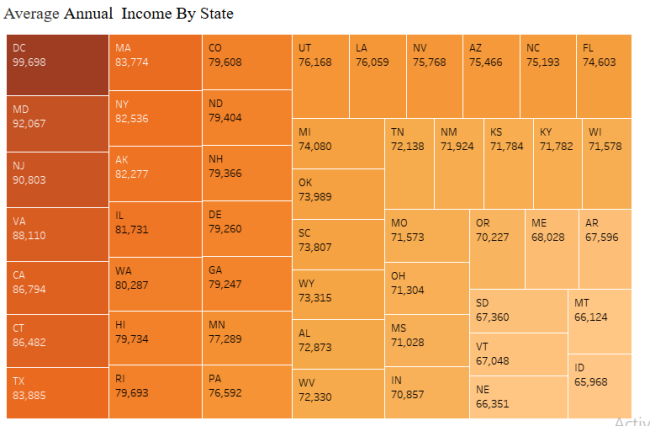

3. average interest rate is important factor while purchasing the loan. in the dataset two column are present. Which are int rate and term. int rate define how much interest has been applied to the loan. and term is period for which loan is taken. From the analysis it is clear that for 36-month period average interest rate is minimum and for loan period which is 60 interest rates is 15.20363



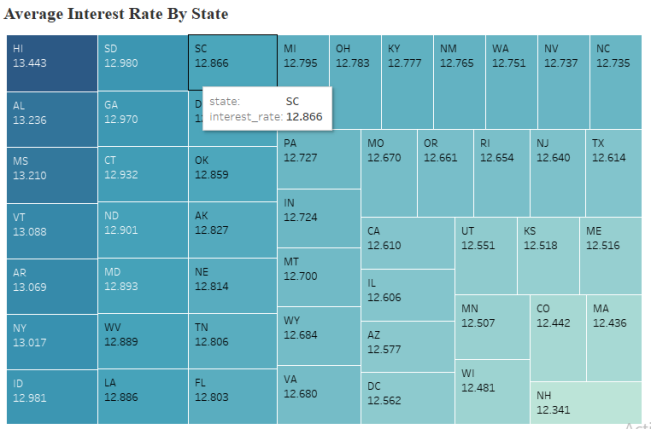

4. loan purpose column in the dataset contains the category of the loan purpose. From the analysis it is clear that most of the time reason of application is debt_consolidation. After that maximum application has been created for the credit card. Lowest loan application is for the wedding and renewable energy




5. Highest annual average income in the state Washington d.c . Idaho is the state where there is less amount of the annual average income. This analysis is important in business perspective.as this filed with other filed defines the average interest rate for the loan application from these state




6. Average interest rate is analyzed according to the state of the us. From the analysis it is clear that annual income affect the interest rate. by consider the low annual income state in USA ID, AR, VT the Annual interest rate for this state is high.




## V. Conclusions and Future Work

. In this project general analysis of the lending club business have been carried out . For the analysis it is found that maximum customer applied for the loan from 20000USD to 40000USD.This result also showed that potential borrower gets the applied loan amount. Interest rate is one of the parameters which is important while borrowing the loan. So from the borrowers perspective this analysis has been done and found that for the loan period 36 month interest rate is minimum and as the term i.e period is increases interest rate also increased. Purpose of the maximum loan applicant is debt consolidation. From the business perspective it is important to know the annual income of the people within state. Analysis showed data average annual income in the state DC is maximum and for the state ID is minimum. average interest rate. average interest rate have been calculated against the state. And the final analysis showed that there is effect of the annual income on the interest rate.
For the purpose of analysis data from the year 2018 is selected. Analysis can be extended by selecting the data from all year. Various machine learning model can be used to predict the amount of the loan application as well as to predict the interest rate. This analysis can be extend by calculating the loan good and bad loan also the risk analysis can be performed.

## References :


[1] Lending Club, *Corporate Spotlight*, 2016. Available: https://implantpracticeus.com/wp-content/uploads/2015/02/Corp-Spot-SpringStone.pdf. Accessed on: Apr. 27,2019.

[2] Mach, T., Carter, C., & Slattery, C. R. "Peer-to-Peer Lending to Small Businesses". *SSRN Electronic Journal.*, 2014. doi: 10.2139/ssrn.2390886.

[3] Reuters , "LendingClub Shares Tumble to a Record Low". *Fortune*., December 7, 2017). Accessed on: Apr. 27,2019.

[4] Jean Baptiste Su, "CEO Tech Talk: Lending Club Plans Expansion Into Car Loans, Mortgages", Forbes., Apr. 2015.

[5] C. Serrano-Cinca, B. Gutiérrez-Nieto, and L. López-Palacios, "Determinants of Default in P2P Lending," *PLOS ONE*, vol. 10, no. 10, p. e0139427, Oct. 2015.

[6] A. Bachmann *et al.*, "Online Peer-to-Peer Lending – A Literature Review," vol. 16, p. 19, 2011.

[7] D. Chen, F. Lai, and Z. Lin, "A trust model for online peer-to-peer lending: a lender's perspective," *Inf Technol Manag*, vol. 15, no. 4, pp. 239–254, Dec. 2014.

[8] M. Lin, N. R. Prabhala, and S. Viswanathan, "Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending," *Management Science*, vol. 59, no. 1, pp. 17–35, Sep. 2012.

[9] "Lending Club Loan Data." [Online]. Available: https://kaggle.com/wendykan/lending-club-loan-data. [Accessed: 28-Apr-2019].