Analyzing The Risk on Bank Loans

By:Dawit Abay

***Abstract*—Since of today, there is a significant increase in the growth of the economy, which has resulted in a tremendous increase in the demand for personal loans from consumers, as the behavior of the borrowers is unpredictable and hazy. Credit risk is a significant issue for both lenders and borrowers, since it directly or indirectly impacts the banks' trustworthiness. The present essay has emphasized on the danger posed by lending money to consumers, as well as the risk posed by investors. The purpose of this study is to examine the credit risk and loan performance of the "Lending Club" firm, one of the largest online credit marketplaces. Analyses of bank loan and credit risk performance using a big dataset with 10 characteristics that was obtained from Lending Club between 2016 and 2017. The Hadoop technique has been employed in this article, and to implement the Hadoop methodology, we will be utilizing the Cloudera software, which is an open source platform for data analysis. It is compatible with the Hadoop environment, which is used to manage, store, and analyze huge amounts of data. We utilized Hive in this post as a data warehouse system for managing and analyzing data stored in HDFS (Hadoop Distributed File System) through HiveQL. To have a better understanding of the bank loan data's performance, we conducted several studies on the bank's gathered dataset.**

I. INTRODUCTION

For my project I combines HDFS data and MapReduce algorithms to evaluate the risk of granting a loan to a person based on their location, loan type, and average risk. The banking area is critical in a bank-dependent financial system. It offers money to households and businesses through lending loans. It is important to consider the riskiness of lending money to clients. Due to the present rapid growth of the economy, there has been a significant increase in the demand for personal loans among clients. Credit risk is a significant problem for both lenders and borrowers, and it has a direct or indirect impact on the banks' trustworthiness. The focus of this essay is on the dangers of lending money to consumers and the risk that investors face.

By evaluating the lending club's data collection, we were able to determine the credit riskiness of client usage data. Lending Club is one of the world's largest online banking markets, where borrowers may receive loans at cheaper interest rates and investors can make huge profits while investing. However, there is a danger of loans and interest not being returned. We utilized a big data technique to assess the credit risk associated with the consumer. To determine the credit risk, we utilized Hadoop to evaluate the data for several factors. The data was imported into Apache software and then injected into Hadoop HDFS, where it was stored for future processing. Following that, the Hive data warehouse technology was utilized to manage and refine the data. HiveSQL was used to conduct further analysis on the data.

II. SOFTWARE

Hadoop is an open source framework for processing huge amounts of data quickly, storing it, and analyzing it. Hadoop includes several processing and analytical tools, including Pig, Hive, Impala, and Spark. Hadoop is a highly adaptable, scalable, and interconnected platform. Hadoop is fault-tolerant software, as data is duplicated to other nodes in the cluster when it is transmitted to a certain node. It stores and maintains enormous amounts of data, and numerous analysis may be

done using various Hadoop analytical tools since it produces accurate and precise facts.

Risks may be effectively assessed with the help of Big Data Solutions. Hadoop provides a comprehensive and accurate picture of risk and impact, enabling businesses to make educated decisions based on market behavior, customer scoring, and future client scoring.

III. Big Date in Hadoop with Banking System

This is a CSV data file including collections of CustomerID, Customer Name, Loan Account Number, Sanctioned Loan Amount, Currency, Disbursed Loan Amount, Loan Status, Risk in Percentage, City, State, Country, and Reason For Taking Loan.

Customer ID	Customer Name	Loan Account Number	Sanctioned Loan Amount	Currency	Disbused Loan Amount	Loan Status	Risk in %	Location	Reason For Taking Loan

Banks and finance businesses' primary concerns are the protection, simple storage, and access to financial data. While the Hadoop Distributed File System (HDFS) enables scalable and dependable data storage across huge clusters of commodity computers, MapReduce processes each node in parallel, transmitting just the node's package code. This implies that data is kept across many clusters but with added security to give a more robust and secure data storage solution.

Although numerous financial firms have adopted Hadoop and it serves as the backbone for several applications utilizing Big Data technology, there are several reasons why Hadoop may not always be the ideal choice.

While Hadoop enables analysis, there are several tools that enable data analysis. Thus, while Hadoop may be used for analysis, adopting the framework just for analytical purposes is not a good choice. Hadoop is helpful only if several scenarios exist in which its USPs can be utilized effectively.

Hadoop is frequently used when Big Data is being implemented. However, before implementing it, one must ask the appropriate questions and consider whether it is the best solution. Any business that receives a large volume of data from a variety of sources and struggles to retain and efficiently utilize that data might benefit from Hadoop and Big Data solutions.

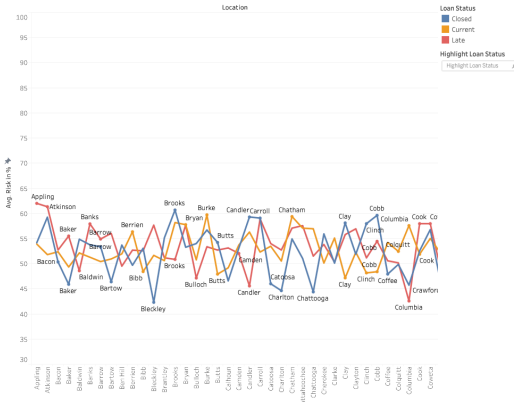

Image 1:using tableau this chart represent location and close, let and open account in the county's in Georgia

Hadoop is not a panacea. While fraud detection and risk management make use of Hadoop's features, Hadoop alone does not address these challenges. Programmers must create code with a thorough grasp of the problem in order to take use of Hadoop's strengths in order to address the business challenge. For instance, Big Data does not aid in the detection of unexpected trends. Big data just enables the parallel processing of enormous amounts of data.

IV. Hive

For Apache Hive is a data warehouse infrastructure that enables the querying and management of massive data sets stored in a distributed storage system. It is based on Hadoop and is being developed by Facebook. Hive provides a means of querying the data through the use of a SQL-like query language known as HiveQL (Hive query Language). Internally, a compiler converts HiveQL queries to MapReduce jobs, which are subsequently submitted for execution to the Hadoop framework.

```
After starting hdfs,yarn,mapreduce put input file in hdfs.
        >hadoop fs -put source_path destination_path_with_file_name.
then start hive
        >hive
To see database
        >show database;
to create database
        >create database bank;
To use a database
        >use bank;
To create table
        >create table
bankloan(cos_id,cos_name,loan_no,amount,currency,amount_paid,status,risk,lo
cal,reason)
                        >row format delimited
                        >fields termineted by ','
                        >stored as textfile;
```

Hive has a very similar appearance to a typical database with SQL access. However, because Hive is built on top of Hadoop and MapReduce, there are numerous significant distinctions. Given that Hadoop is designed for long sequential scans and Hive is based on Hadoop, you would anticipate extremely high latency for queries. This means that Hive is not suitable for applications that require extremely rapid response times, like a typical RDBMS database would. Hive is a read-only database and thus unsuitable for transaction processing, which often requires a high proportion of write operations.

V. MAP & REDUCE

Map reduction is a programming approach that is used to associate and implement huge datasets using distributed parallel algorithms on a cluster. Map Reduce is a combination of two separate terms, namely map and reduce. The first step is the map phase, which is responsible for processing the

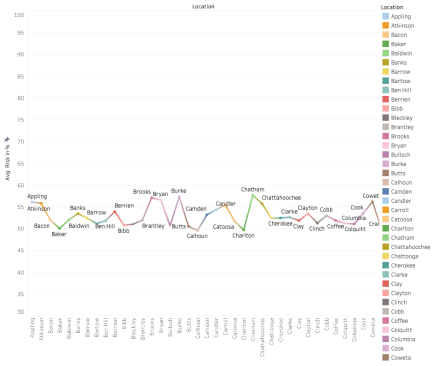

Image 2: using tableau this chart represent location and risk assessment the county's in Georgia

programming, filtering, and storing operations; the second phase is the reduction phase, which is responsible for performing the summary of activities.

Hadoop's primary components are HDFS for storage and MapReduce for processing. They are connected in that computation takes place near to the server that contains the necessary data. The map function is called in parallel, as illustrated below, and returns a pair with the key being the integer and the value being its square. Even and odd number keys are then combined and reduced to the final value by summation in successive reduction procedures.

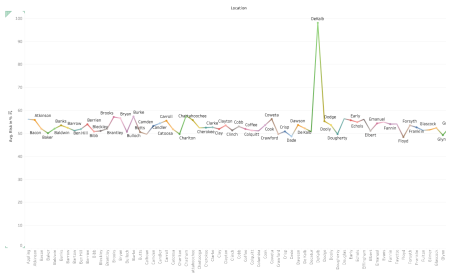

Image: using tableau this chart represent location and risk assessment the county's in Georgia

The map reductions are the fundamental functions of the algorithm and are responsible for two critical tasks.

Map, the map or mapper function will process the input data, and then the input file will be provided line by line to the mapper function. Later on, this mapper will transform the data into little sets, or small bits of data. Essentially, this map () function will conduct filtering and sorting.

```
package com.AvgRiskcat;
import java.io.IOException;
import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
public class MyMapper extends Mapper<LongWritable,Text,Text,DoubleWritable> {
public void map(LongWritable key, Text value, Context con) throws IOException, InterruptedException
{
String line = value.toString();
String[] linePart = line.split(",");
//problem -1 -avg risk
Double risk = Double.parseDouble(linePart[7]);
//problem-2 - avg risk per location
//String loc = linePart[8].toString();
String cat = linePart[2].toString().substring(1, 3);
if(cat.equals("HL"))
{
cat = "Home Loan";
}
else if(cat.equals("PL"))
{
cat = "Personal Loan";
}
else if(cat.equals("VL"))
{
cat = "Viechel Loan";
}
else
{
cat = "Retailer Loan";
}
con.write(new Text(cat), new DoubleWritable(risk));
}
}
```

Image 3: shows us how to calculate the risk of each loan

The map function is divided into the following sub-steps:

a. Splitting the input dataset - The input dataset should be selected from the source data and then divided into smaller sub-datasets.

b. Mapping - Mapping will take the sub-datasets and conduct any necessary actions or computations on them.

The key-value pair has been successfully established. Now comes the shuffling combining.

c. Merging - This step will combine all key-value pairs that share a common key. d. Sorting - Sorting will use the key to sort the key-value pair. As an output of this full map ( ) function, a sorted key pair is returned.

Reduce, the reducer will process the data that was obtained during the map phase of the mapper and will then generate a new set of output. Essentially, the reduction function summarizes the procedure.

VI. RESULTS AND CONCLUSIONS

For my project I combines HDFS data and MapReduce algorithms to evaluate the risk of granting a loan to a person based on their location, loan type, and average risk. This project examines different data processing systems, including Hadoop, MapReduce, and Hive, and gives a comparison of prominent software frameworks such as these. Some important characteristics and interesting problems surrounding Big Data in finance have been addressed in an accompanying dialogue on many platforms. The model of loan risk analysis described in this presentation was created utilizing different essential tools and characteristics of big data. However, it is a general but a more focused study has to be done on include each component in this model. The creation of a hybrid model for credit assessment would enhance credit risk decision quality and may perhaps cut down on instances of credit defaults and banking frauds. Even though banks have access to significant quantities of consumer data, various restrictions prevent the information from being used for insight. Banks that wish to remain competitive in the fiercely competitive financial services industry must adopt a data-driven strategy. There will be almost limitless possibilities for incumbents in the financial services industry due to these insights, which means Big Data will become a critical differentiator in future market share.

```
10.15
```

Image 4:is calculating the average risk overall

```
Home Loan       18.048
Personal Loan   13.968
Retailer Loan   12.261333333333333
Viechel Loan    10.15
```

Image 5: calculating the risk of each loan for home, medical retailer and vehicle

```
Alpharetta      10.485714285714286
Atlanta 10.324137931034484
Augusta-Richmond County 10.513636363636364
Columbus        10.362068965517242
Dunwoody        10.222222222222221
Macon-Bibb County       10.104651162790697
Savannah        10.15
```

Image 6: calculating the risk for each county's in state of Georgia

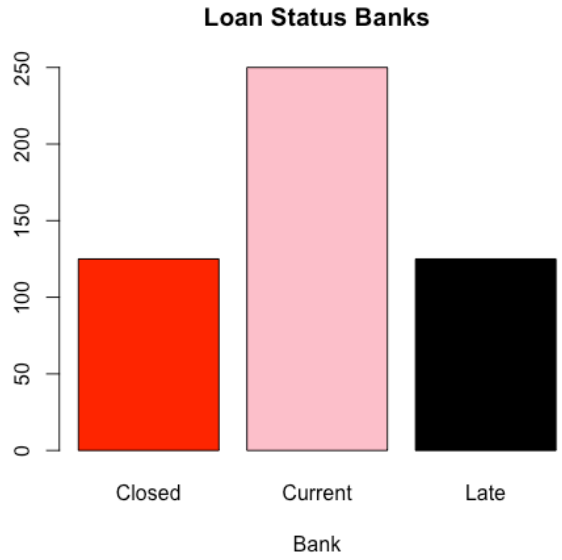

image 7:is counting the account that is open closed or late

References

[1] https://data.world/datasets/loan

[2] https://www.kaggle.com/zaurbegiev/my-dataset

[3] https://www.kaggle.com/panamby/bank-loan-status-dataset