FINANCIAL DATA ANALYTICS FOR FRAUD DETECTION

Divya Samragni Nadakuditi

dnadakud@my.bridgeport.edu

ABSTRACT

Big Data is one of the evolving streams these days. It has both structured and unstructured data. Handling this data has been researched and several models have been formed. In this paper, we are going to get the working experience with these models by performing Big Data analysis. For this analytical work we choose a dataset and find insights from it. We will work on the selected dataset and get results by implementing Hadoop Mapreduce, Hive, Spark and Big Query. We will discuss clearly the procedures we used and how we have performed the analysis using these methods.

1. INTRODUCTION.

Big data is large volumes of structured, unstructured and semi-structured data. Dealing with this data is important for making decisions strategically as well as financially. In this project, we considered a dataset from kaggle called Synthetic data from a financial payment system. This dataset is used by us to perform Financial Data analysis for fraud detection.

Why is Fraud detection an important aspect of research ? Everything in this pace of life happens with a click of a button. Everything has been changed to online and all procedures are made simple for transactions to happen. All these made life easy as well as increased the risk of fraud. The effect of fraud transactions in any field effect the economy of the country. This affects many depending companies. Hence there is a need for us to know more details about fraud. Research on this can help in finding Why there is more number of fraud, what are the ways to reduce it .

Research on this topic is quite difficult as there are very few data sets available on which one can rely upon. Also, as the data is related to financial details there are security issues which the researchers have to deal with . In the paper we will discuss our research on fraud including data collection, the methods used to draw insights regarding the data in the conclusions drawn.

2. DATASET.

This dataset has been taken from Kaggle with the name 'synthetic data from a financial payment system'. The following link takes you to the dataset https://www.kaggle.com/ntnu-testimon/banksim1 . We are using this data to research about fraud. This data is generated using a simulator called BankSim which uses the data from a bank in Spain and generates the data of financial transactions. As we all know that the bank transactions are very huge in number and considering Spain alone would lead to trillions of data. So we have considered the regions from Madrid to Barcelona. Also our data does not consists of any personal information of neither the customers nor the merchants so there are no security issues using this dataset. The data was generated keeping in mind the statistical and social network analysis of the relations between the merchants and customers.

Our data consists of around 595K rows and it has fields like customer ID and merchant ID , age group of customers, zip code of customers and merchants to know their location, the type of transactions, gender, amount spent and if it a fraud transaction or not.

3. PROBLEM STATEMENT.

Fraud detection and fraud are very important aspects in any field. Using our dataset we will analyze the fraud scenarios and draw conclusions that would help in narrowing our path to the fraud detection and reasons on how the fraud has occurred. We will be analyzing the top five categories of fraud occurrence, which age group has committed the maximum fraud likewise which gender, if there are any fraud committed from same account to the same merchant, which merchant has the maximum number of fraud transactions.

Drawing these insights will help us in finding out the reasons for maximum fraud occurring in that category or to the merchant . Like if a merchant has more number of fraud then we can further research on the reasons for fraud in that category - Due to online transactions or Cash transactions or Card swiped . This research helps us knowing if the card was theft or if the details of the card are been stolen. We can use the other insights of our research in finding out if there is any particular customers who are committing fraud frequently and can help the bank in investigating more on these kind of customers and give conclusions regarding fraud.

4. METHODOLOGIES AND THEIR IMPLEMENTATION .

We have proceeded with our research using various Big Data models like Hadoop, MapReduce, Hive and Sqoop, Spark.

We use Hadoop framework to store our data files we load our data into HDFS. We will now discuss various techniques and their implementation with our data.

MapReduce programming model works as key value pairs. In this model the input is taken as key value pairs which are further processed and the results are aggregated as key-values such that each key has zero or one output. For our dataset, we consider the key value pairs as age - fraud , category - fraud , gender - fraud viceversa. These key-values are taken and distributed among a number of clusters where they are processed and all the outputs are aggregated and a single output is produced. Below are the screenshots of the code used for MapReduce.

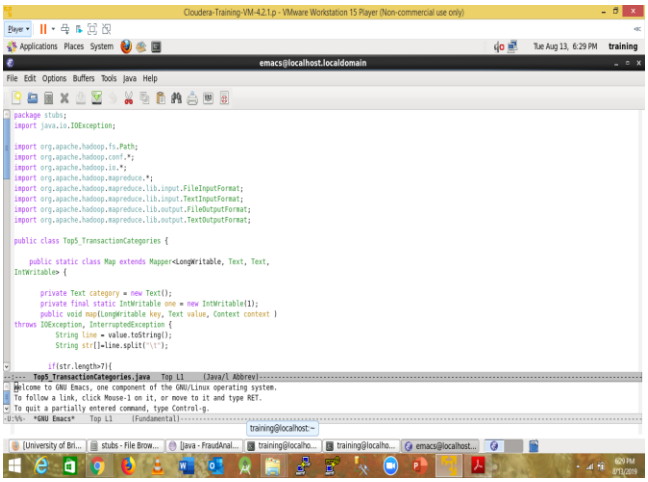

Figure 1: Code for Mapper Phase

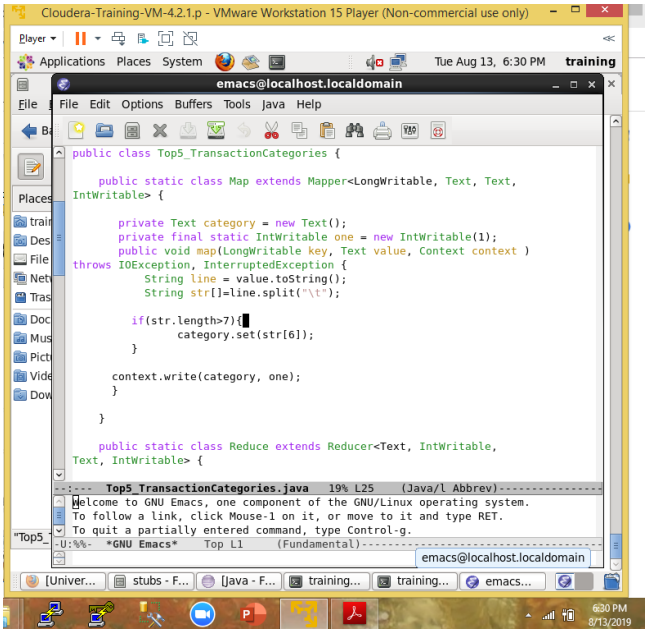

Figure 2: Code for Mapper Phase

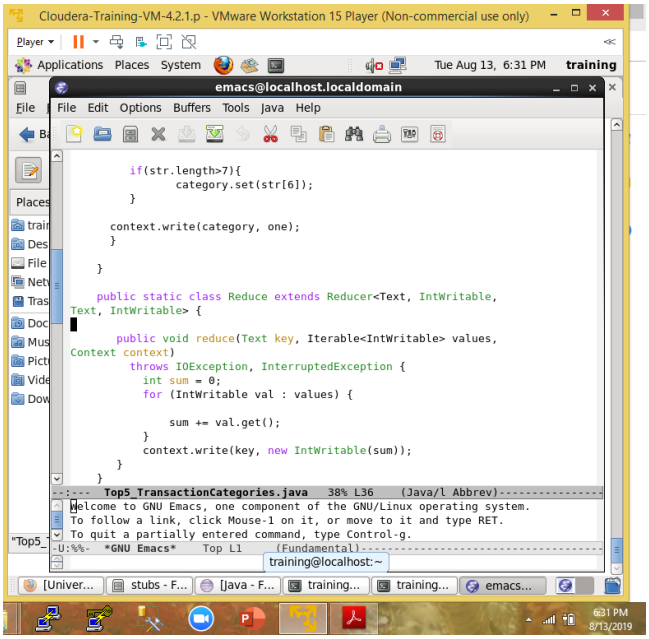

Figure 3: Code for Reducer Phase

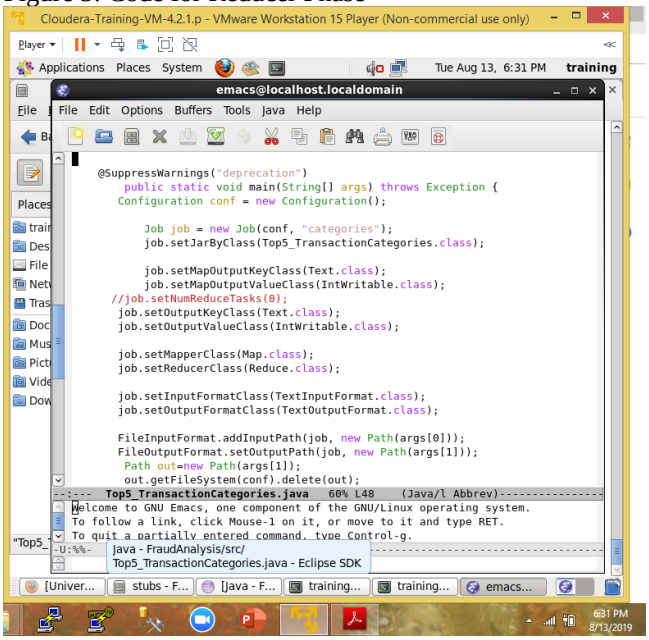

Figure 4: Code for Input and Output

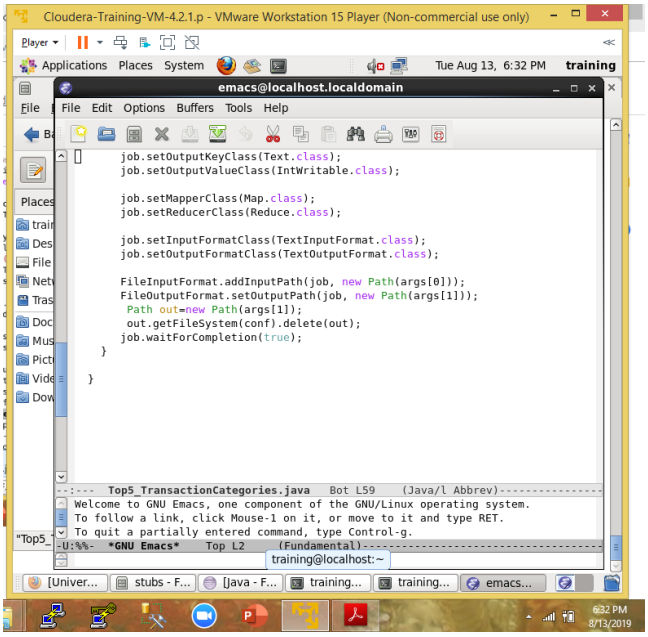

Figure 5: Code for Input and Output

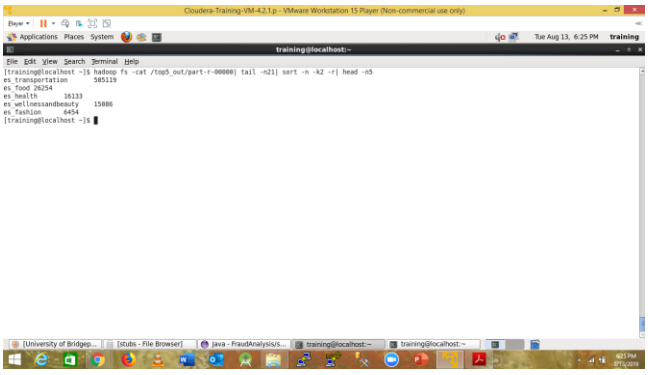

Figure 6 :Output of the Mapreduce program.

The above figures 1 - 6 gives you the MapReduce program and its output after running successfully.

Using Hive's SQL feature, we create a table of the required dataset. We run queries in Hive which actually runs a MapReduce as a query. In Hive we create an external table of our dataset and run an SQL query to find the maximum number of fraud depending on the age category, gender category, category of transactions.

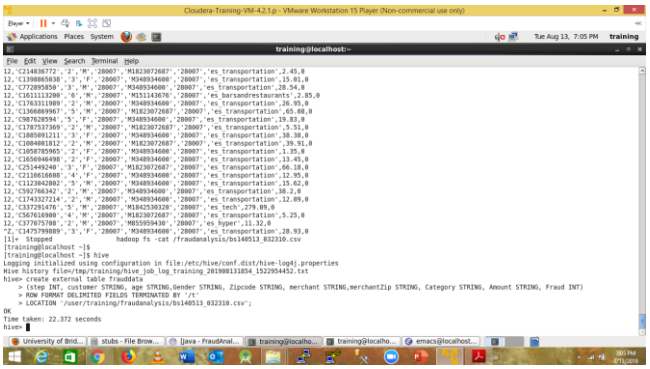

Figure 7 : Creating an external table of dataset

Select gender, count(gender) from frauddata where fraud =1 Group by gender;

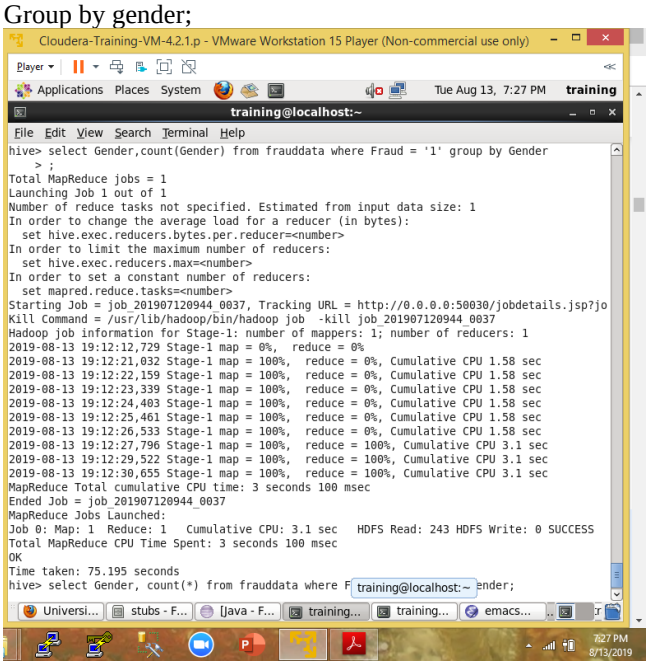

Figure 8 :Running a query for gender related fraud.

For Spark we use Scala or PySpark which python as backend. Before analysis we load the data into HDFS create the schema for our dataset and proceed with our analysis. Using Spark we get the information about the maximum fraud category.

```
val textFile = sc.textFile("hdfs://localhost:9000/frauddata.txt")
val counts = textFile.map(line=>{var Category = ""; val temp=line.split("\t"); ;if(temp.length >= 5) {Category=temp(5)};Category})
val test=counts.map ( x => (x,1) )

val res=test.reduceByKey(_+_).map(item => item.swap).sortByKey(false).take(5)
```

Using our research we found out that we can see that a number of fraud occurred more in the field of transportation, followed by Food, health, wellness and Beauty, Fashion.

6. CONCLUSION.

BigData is one of the top most sectors in which the world is working these days. This helps in analysing the data for the business and statistical decisions. In this project, we have analysed the financial data for details related to fraud.This research helps us in finding information on fraud categories which said that transportation has more number of fraud. Likely, we drew several conclusions regarding the fraud. By analysing this dataset, we became familiar and got a very good hands on on the different Big Data analytics techniques like Hadoop mapreduce, Hive, Spark. Working on this project helped me in enhancing my understanding towards the methods for Big Data Analytics and understanding the pros and corns of each method. This kind of projects helps us gain practical experience and widens our way of looking into a problem and drawing our insights.

9.REFERENCES.

[1] https://www.kaggle.com/ntnu-testimon/banksim1

[2] https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm