Real Time Scam Recognition and Prevention of Online Transactions

Mr.Vivek Pandey[1] , Mr.Kalpesh Jadhav[2] ,Mr.Roshan Pandey[3]

[1]Assistant Professor of Computer Engineering, A.R.M.I.E.T, Thane, India.

[2, 3] UG Scholar of Computer Engineering, A.R.M.I.E.T, Thane, India.

***Abstract*— Today's Technology use is rapidly increases. The use of credit card is not new in the online market. The internet becomes most popular mode of shopping using online transactions. It provides us facilities like e-commerce using credit card, debit card and internet banking. In modern shopping environment credit card use is increased rapidly. But the Frauds of credit card, Telecommunication fraud, Intrusion fraud, etc., also increase. In this paper, we focus on constructing real time scam detection and prevention of online transaction in order to detect and prevent the fraud in real time by using various machine learning algorithms. The ability to process large amount of data can also be achieved through big data technology like Hadoop.**

***Keywords*— Fraud Recognition and Prevention; e-Commerce; Hadoop; Machine Learning**

I. INTRODUCTION

There has been a tremendous increase in electronic transactions during the last decades, due to the popularization of the World Wide Web and e-commerce [1]. The number of card issuers, card users and the online merchants has increased [2]. This is mainly due to the various online retailers like eBay, Flipkart, Amazon, Walmart to name a few. Individuals have changed their mode of payment significantly with the growth of modern technology. Most of them make use of online payment modes while shopping online or at the market. The fraudulent activity on a card affects the cardholder, the merchant, the acquiring bank and the issuer. With regard to the cost of fraud, the most affected participant is the merchant, because the cost of fraud is greater than the cost of goods sold. Cyber-crime is a crime committed over internet. Fraud is generally defined as a criminal activity committed by the criminal in order to obtain financial/personal gain [2] [3]. Fraud can be mainly divided into two types: Offline Fraud and Online fraud.

Offline fraud is the one which involves some physical activity such as stealing purse/wallet which contains valuables like credit card, ID proofs etc. and using the crucial information within them.

Online Fraud occurs when the fraudster uses an electronic medium or creates a website and present their customer accounts. Other ways through which the fraudsters collect/steal personal information are Hacking, Phishing, Spoofing, Spyware, Shoulder Surfing, Dumpster Diving etc.
[3].

The online transactions take place within a fraction of seconds. Since a large number of people are associated with such e-commerce transactions, the dataset associated is also large. There exists a need for developing fast and efficient algorithms to process such large datasets and to search for fraudulent and deceitful transactions.

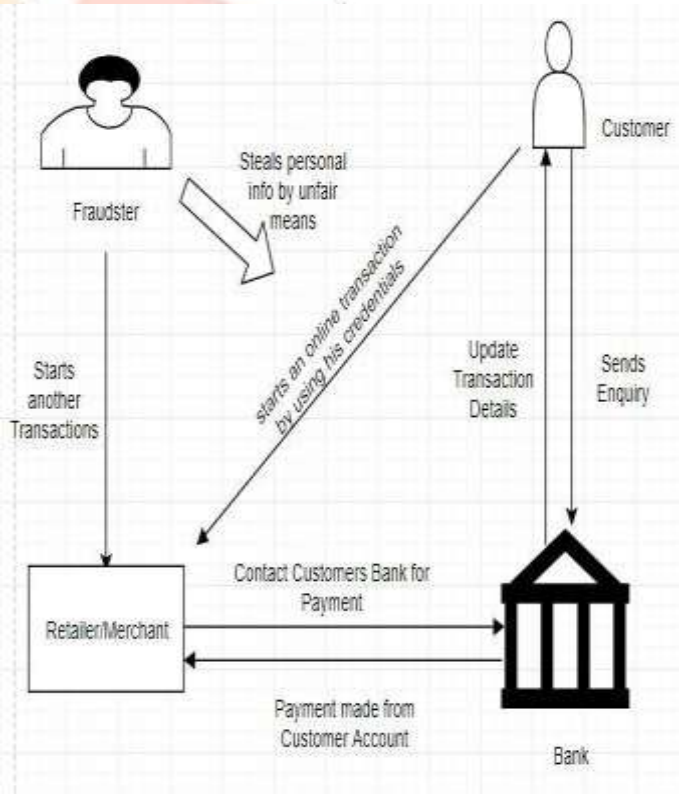

Fig. 1. Example of Transaction fraud

In the Fig. 1, it can be seen that, the fraudster gains the personal credential information of the customer via some unfair means and uses the same for online shopping. The cardholder realizes that fraud has occurred and starts an enquiry for the same at a later stage after the transaction is complete and the money is lost. Although

there are several fraud detection techniques based on Data Mining, Knowledge

Discovery and Expert System etc., they are not capable enough to detect and prevent the ongoing fraud.

Also, the growing number of users and payment transactions has brought heavy workloads to these systems. The speed of new transactions coming into the system can reach millions per second while the size of stored historical transactions can reach several PBs or even EBs. In this case, processing detection tasks and model training on so many incoming transactions with a low delay is very hard for most traditional systems. According to the recent trends, Big Data technology seems to be the key of solving the challenge of computational capacity.

The major challenge of fraud detection is the very limited time span in which acceptance or rejection is to be done. Also another peculiarity is the large number of transactions that has to be processed at a given time. Detecting fraud at the earliest when the transaction is being processed is a major concern in ecommerce systems. Another major concern of the research is to avoid rejecting the genuine customers. The fraudulent activities impose considerable financial losses to merchants and therefore fraud detection becomes a necessity.

Also, the most important thing to remember is that the fraudsters are constantly updated, so there arises a need for systems which keeps on adapting themselves as and when newer instances of fraud occurs. In this paper, a system to detect fraudulent transactions with increased accuracy, the ability to process large amount of data, the ability to do the detection in real time and adaptation to newer instances of fraud has been proposed.

The paper is organized as follows: Section II describes the Background and related work in the area of fraud detection in electronic transactions. Section III describes the fundamentals. Section IV describes the detailed system working and finally Section V presents conclusions and future enhancement.

II. BACKGROUND AND RELATED WORK

Electronic or credit card fraud detection has drawn a lot of attention in the last few decades. Some of the works that are related to fraud detection in electronic transactions or credit card operations are described in this section.

In [1][4] the authors have used a Neural Network based approach which uses MLP. In Neural Network [2], the interconnection weights between different nodes are learned during the process of training and the processing ability of the network is computed by the learnt weights. The features of the transaction were given as inputs and the inputs are weighted. This weight shows the intensity of how a particular input influences the output value. The network computes the difference between the actual and the desired output and propagates it backward to adjust the weights assigned to the inputs. Thus, it can be inferred that, fraud detection using neural network is based on Pattern Recognition, i.e. when a fraudulent transaction is detected; the weights of the inputs related to that transaction pattern are updated. The disadvantage of this approach is that every time a new pattern of fraud occurs the entire network has to be re-trained.

In [1], the authors have also used a Bayesian learning approach. Bayesian Networks were also used in different comparative studies for detecting fraud in electronic transactions especially in credit card transactions. Bayesian Networks represents dependencies between variables of a probabilistic model, where each node represents a random variable and the arcs represent the relationship of a dependencies between variables. In the fraud detection problem, the variables are the features or attributes that influence the transaction. These features were given as inputs. In the fraud detection problem, initially the network is unknown. To construct the Bayesian Network, the data has to be learned. Later from the graph that is constructed, the set of dependent variables to happen fraud is calculated. Bayesian Networks are more prominently used for classification problems. The network provides easy and fast training but is impacted when applied to newer instances.

In [5], in order to detect fraudulent transactions, a model based on Hidden Markov Model (HMM) has been proposed by the authors. Initially, the model is trained with the normal behavior or spending pattern of the cardholder. To identify the spending behavior of the cardholder or customer, K-means clustering algorithm is used. The spending profile is characterized as low, medium and high. The HMM is used to find out any deviation or variance in the spending patterns. HMM works just by remembering the customer spending behavior. Therefore, if the HMM rejects an incoming transaction with sufficiently high probability, then the transaction is considered fake or fraudulent. The model will generate an alarm to stop the transaction if any deviation or variance is observed from the spending patterns of the cardholder.

In [6], the authors have described a fusion approach based on Dempster-Shafer theory and Bayesian learning. The fraud detection system has four components: Rule-based filter, Dempster-Shafer adder, transaction history database and Bayesian Learner. Here, the rule-based filter calculates the suspicion level of each of the transactions with respect to its deviation from the genuine or good pattern. The Dempster-Shafer adder combines several such evidences and computes or calculates an initial belief. The initial beliefs are then combined to obtain an overall belief by applying the Dempster-Shafer theory. The transaction is judged as fraud or genuine based on the initial belief. Once the transaction is found to be suspicious, the belief is further enhanced or weakened by checking its similarity with fraud or genuine using Bayesian learning. The system was efficient but was highly expensive and processing power was low.

III. FUNDAMENTALS

This section describes the techniques we apply and evaluate in this work: MapReduce (Section III-A), HDFS (Section III-B), Detection Algorithms (Section III-C).

A. MAPREDUCE WORKFLOW

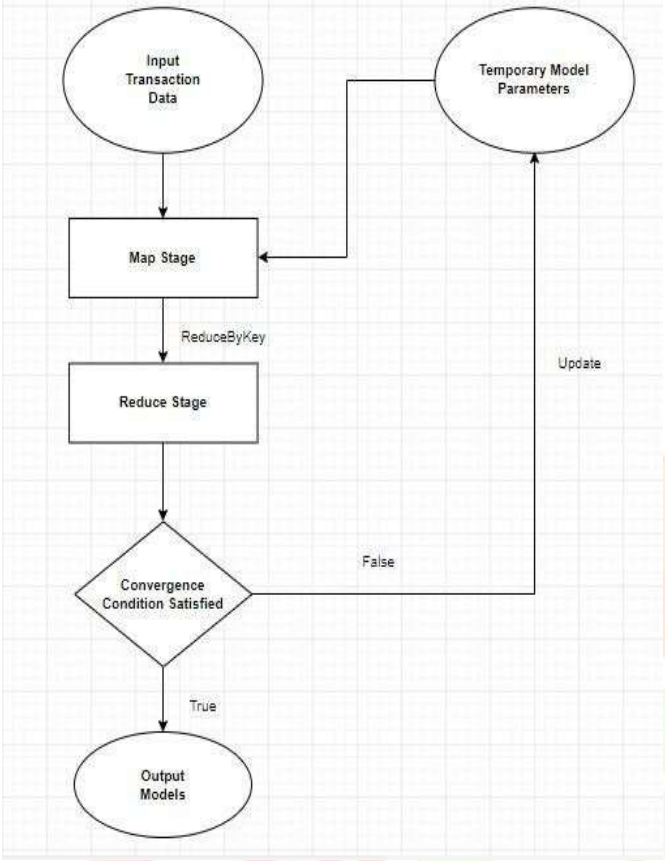

Fig 2: MapReduce workflow for Input Transaction data

Figure 2 shows the workflow of MapReduce, which also consists of two stages:

- Map Stage: for all input transactions data from the distributed storage layer, they will be partitioned and each Map Stage task will handle a part of the data. Then, each Map Stage task will also update a part of model parameters in this stage and they will be shuffled to Reduce Stage.
- Reduce Stage: this stage will aggregate all parts of model parameters and merge them into one. Then, it will check whether the convergence condition is satisfied. If the condition is not satisfied, the temporary model parameters will be updated and Map Stage will be started again. But when the condition is satisfied, the aggregated parameters will be stored in the key- value sharing layer.

The implementation of the two stages on Hadoop is very straightforward, as shown in Algorithm 1 and 2.

Algorithm 1 Mapper function for key aggregation on Hadoop

Input: t_i^j //the jth transaction of Card i;

Output: < *i*, t_i^j > //card ID and its transaction;

i = *splitCardId*(t_i^j)

< *i*, t_i^j >= *generateKV Pair*(*i*, t_i^j)

Emit < *i*, t_i^j >

Algorithm 2 Reducer function for model training on Hadoop

Input: *list*(< *i*, t_i^j >) //list of all transactions of Card i;

Output: < *i*,$model_i$ > //trained model of Card i;

for < *i*, t_i^j > in *list*(< *i*, t_i^j >) **do**

listi.add(t_i^j)

end for

$model_i$ = *train*($list_i$)

< *i*,$model_i$ >= *generateKV Pair*(*i*, $model_i$)

Emit < *i*,$model_i$ >

B. HDFS

We evaluate the throughput of Distributed Storage Layer when the number of a transaction's attributes vary from 20 to 160 to simulate possible transactions with different length. We use 10 clients to write transactions into HDFS and use 70 clients to read data from HDFS in parallel. The HDFS cluster consists of 11 nodes. However, the system is able to write up to a million of transactions per second and read up to 100 million transactions per second.

C. Detection Algorithms

Bayesian Networks

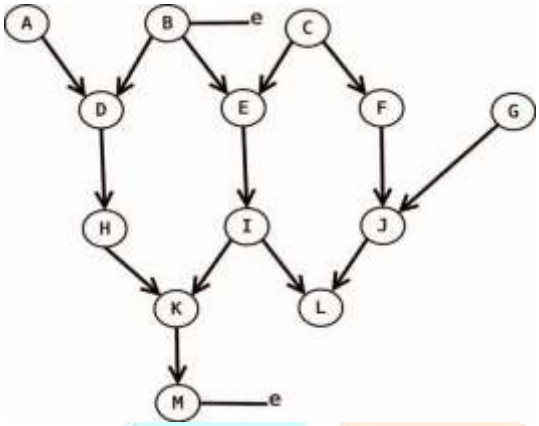

Fig. 3. Example of Bayesian Networks

The mathematical definition for BN is derived of Bayes theorem, which shows that conditional probability of an event A_i given an event B, can be calculated by Equation 1, where $P(A_i|B)$ is the probability of A when B occurs.

$$P(A_i|B) = \frac{(P(|_i)P(_i)}{P()} \quad (1)$$

In fraud detection problem the BN is unknown, therefore to build the BN graph it is need to learn it from the data. From the BN graph, we can calculate the set of dependent variables to happen a fraud (conditional probability), using Equation 1. Before calculating the conditional probability, we can find the probability of fraud applying Equation 2[23].

$$P(x_i, \ldots, x_n) = \prod_{i=0}^{n} P(x_i|Paen(x_i)), \quad (2)$$

Where, Parents(Xi) are determined.

Logistic Regression

Logistic Regression (LR) is a statistical technique that produces, from set of explanatory variables, a model that can predict values taken by a categorical dependent variable.

Bayesian Networks (BN) are directed acyclic graphs that represent dependencies between the variables of a probabilistic model, where each node in the graph represents a random variable and the arcs represents the relationships between these variables [22], as showed by Figure 1, where the event A affects directly the event D that if affected directly by event B, and so on. And e is an independent event.

- A systematic component, which corresponds to a linear function between the independent variables.
- A link function, that is responsible for describing the mathematical relationship between the systematic component and random component.

The binary LR model is a special case of the GLM model with the logit function. This function is used to get the estimation of coefficients [26]. Then, we apply these coefficients in Equation 3 that result in our fraud probability.

Decision Tree

There are two categories of decision trees, classification trees and regression trees. The decision tree learning is the construction of a decision tree from class-labeled training tuples. A decision tree consists of nodes that forms a tree structure; the topmost node is called the root node. Each non- leaf node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf node holds a class label. Leaf nodes represent classes that are returned if reached as the final prediction by the model. As Zaki&Meira (2014) elaborated, given an instance with its features' values, the model is able to classify the instance by traversing the decision tree. There are several decision tree algorithms including: ID3 (Iterative Dichotomiser 3), C4.5 (successor of ID3) and CART (Classification and Regression Tree).

To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

a) Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

b) Entropy using the frequency table of two attributes:

$$E(T, X) = \sum_{c=x} P(c)E(c)$$

Thus, a regression model is used to calculate the probability of an event, through the link function described by the following Equation:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_i x_i)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_i x_i)}}$$

where π(x) is the probability of success when the value of the predictive variable is x. β0 is a constant used for adjustment and βi are the coefficients of the predictive variables [24].

In order understand LR, it is important to explain the concept of Generalized Linear Models (GLM). This consists of three components [25]:

• A random component, which contains the probability distribution of the dependent variable (Y).

IV. SYSTEM WORKING

Figure shows the proposed system.

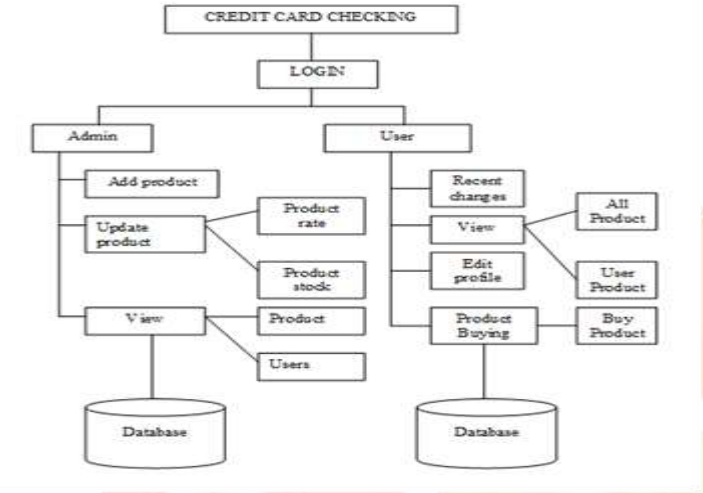

Fig. 4. E-commerce process

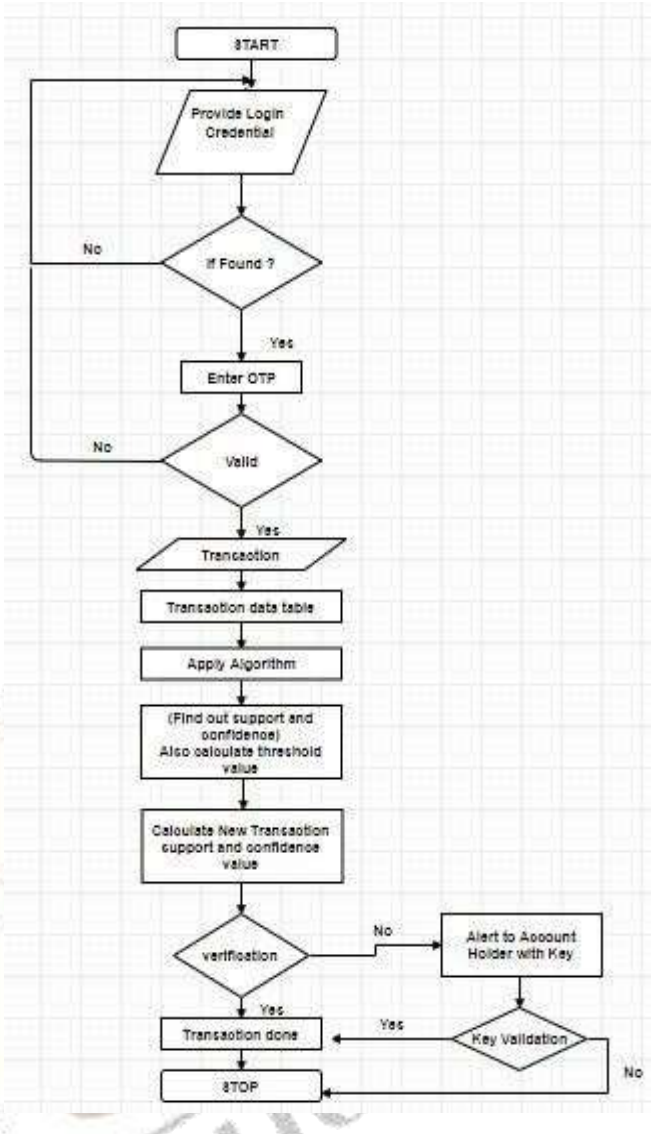

Fig. 5. Working Principle

When a customer try to buy product from the e-commerce website using his credit card as shown in figure 4. It uses credit card for online payment. In online trading the credit cards are usually used as virtual card. To prevent fraudulent transaction on internet we have introduce this system which is capable of to detect frauds and also prevent financial loss. The main challenge of the system is to improve detection accuracy, the computational capacity have become important. To solve this problem Big Data Technology seem to be key. We implement the framework with Hadoop which are able to handle the burst amount of data and build a scalable and reliable system.

The architecture of the system is nearly real-time system. The job of the system is to receive events and reply as fast as possible. It does very little processing and depends mostly on pattern matching.

In this, the system has a two module the admin will add product, update product and view which is connected to database server. The user is the buyer. They will be able to view product, can create profiles and also buy product

which will be connected to database server. When customer enter on online website, and buy product, the website will ask login credentials, if it is correct, then it will ask OTP which will be sent to customer mobile, if valid, then transaction will process, otherwise no transaction. The system constantly monitor the behavior history of transaction as shown in the figure 5.

V. CONCLUSIONS AND FUTURE ENHANCEMENT

Hence, Online Scam detection in E-commerce is becoming challenging due to processing of large amount of data and analyzing various fraud patterns that are happening everyday. In this paper, we build different fraud detection models to predict fraud in online transactions, more specifically credit card operations in real time in order to minimize the loss of users on major basis. We have applied machine learning algorithms to detect fraud and Big data technologies for the smooth functioning of the system.

Although, the proposed system gives good results with large number of inputs, future work will concentrate on preparing an application with consistent Fraud Detection with new techniques and modules, develop a sophisticated module like calculating Fraud Timings, capturing the photo of the Fraud and many more modules can be developed.

REFERENCES

[1] E. Caldeira, G. Brandao and A. C. M. Pereira, "Fraud Analysis and Prevention in e-Commerce Transactions", 9th Latin American Web Congress, Ouro Preto, pp. 42-49, 2014.

[2] E. Caldeira, G. Brandão, H. Campos and A. Pereira, "Characterizing and Evaluating Fraud in Electronic Transactions", Eighth Latin American Web Congress, Cartagena de Indias, pp. 115-122, 2012.

[3] S. Parvinder and M. Singh, "Fraud Detection by Monitoring Customer Behavior and Activities", International Journal of Computer Applications, vol. 111, no. 11, pp. 23-32, 2015.

[4] K.K. Tripathi and R. Lata, "Hybrid Approach for Credit Card Fraud Detection", International Journal of Soft Computing and Engineering, vol. 3, no. 4, pp. 8-11, 2013.

[5] S. Sorournejad, Z. Zahra, R. E. Atani, and A. H. Monadjemi, "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective", https://arxiv.org/abs/1611.06439, 2016.

[6] Lichman, Moshe, "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml, 2013.

[7]https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data).

[8]Weka Machine Learning Project, http://www.cs.waikato.ac.nz/~ml/weka.

[9]https://risnews.com/secure-file/18318.

[10]http://losspreventionmedia.com/insider/retail-fraud/e-commerce-credit-card-fraud-the-rapidly-growing-challenge-for-retail-investigations/.[11]https://www.marutitech.com/machine-learning-fraud-detection/. [12]https://www.tutorialspoint.com/articles/machine-learning-the intelligent-machine.

[13]https://www.tutorialspoint.com/big_data_analytics/machine_learning_data_analysis.htm.