Big Data Analytics Techniques and Approaches to detect Cyber Crime

S Kavitha
Department of Computer Science Engineering
Baba Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh, India
Email- kavithasanaka@gmail.com

K S N Murthy
Department of Computer Science Engineering
Baba Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh, India
Email- ksnm1925@bitsvizag.com

Abstract- This paper is about big data systems used to battle crime in cyber speculations. The fundamental difficulties in crime examination have been tended to and the kind of crimesters included have been contemplated in the light of various circumstances. The two principle kinds of data mining strategies i.e. administered and unsupervised learning techniques utilized in crime investigation have been seriously inspected with the assistance of existing writing. This is expected to help in perceiving which data mining systems help the most in crime examination in particular spaces. Banks are an indispensable piece of a nation's economy, adding to the two individuals and governments. In later past a great deal of crime exercises have been accounted for in banks due to individuals with personal stakes. This paper tosses light on normal insider crimes happening in banks and furthermore attempts to order them into various sorts. Here we give a point of view of definition, factors identified with such classes of crime and furthermore the difficulties one faces in identifying crimes. It is critical to identify such crime exercises consequently before it is past the point of no return and bring the general population or gathering of individuals into terms. Data mining system comes convenient as it recognizes irregular patters in given data set. This paper focusses on various nonexclusive data mining systems and in explicit, the procedures utilized for identifying insider crimes. We additionally drill down best accessible systems in acknowledging insider crimes with applicable outlines. As it is extremely apparent in this day and age, Big Data challenges are springing up all day every day, we have felt free to exhibited as a future improvement a couple of enormous data challenges while taking care of banking data.

Keywords – Data mining, fraud pattern detection, clustering, supervised learning, Cyber crime

I. INTRODUCTION

Cybercrime is any sort of wrongdoing that should be possible in, with, or against systems and PC frameworks [10] [3]. Cybercrime is getting expanded with the expanding dangers because of online crimes and exploitative hacking. With both data and digital security risk expanding, associations must be prepared to furnish themselves with foreseeing and forestalling cybercrime. Cybercrime specialists are utilizing huge data apparatuses to recognize the potential dangers and identify cybercrime occurrences like charge card crimes. Huge data investigation is empowering organizations to examine voluminous measure of data they assemble amid money related exchanges, any district explicit data and others too. Battling digital wrongdoing is of most extreme significance today because of expanded danger of digital robbery. Huge data instruments are being utilized to battle digital assaults. Huge Data examination can help distinguish crime and recognize burglary and can encourage advanced legal investigation. In this paper, a short overview is made about different systems utilized in examining huge data to recognize the crimes identified with charge cards by breaking down extensive arrangement of data. One point of this examination is to distinguish the client show that best recognizes crime cases. Along these lines, data mining is an essential system for every single movement of the charge card process. For instance, it tends to be utilized for ordering great clients or terrible clients which is completely founded on their application data and, likewise, distinguishing an abuse of a charge card dependent on buy data of a client. The effectiveness of anticipating the integrity or disagreeableness of a candidate can diminish credit danger of a Mastercard suppliers. While, if the supplier settle on any wrong choice by giving Visas to terrible clients, it will result in huge

loss of income and liquidity. This credit hazard issue can prompts the money related emergency of the world economy. Because of a gigantic measure of accessible data, process examination in the charge card action need to depend on data mining strategies for its viability and proficiency. Fundamentally, data mining is procedure of extricating the examples from the data. It joins the strategy which is utilized to measurable, machine learning and database so as to extricate and recognize helpful data from a great deal of database [8], [2], and [5]. As of late, there have been different works which serves to looking into the uses of data mining system in the banking division. In [8] the creators have contemplated data mining utilized in different exercises in the banking area, i.e., client relationship the executives, crime location, promoting and hazard the board. This work did not explore correctly on the Visa procedure and, along these lines, data mining techniques utilized in such a way isn't self-evident. Another study of data mining applications in banking has been introduced in [4] in which the idea of learning disclosure in database strategy (KDD) was likewise talked about. Though, they didn't show the Mastercard procedure. In extra, there are reviewing works worried for specific regions of the charge card strategy, for example, crime location nd credit scoring for client's application [6]. As for be progressively explicit study which can be for Mastercard suppliers and scientists around there, we have explored investigate on data mining applications in principle exercises of the charge card techniques.

II. Related Work

Reilly and Ghosh in 1994 [3] introduced a three layer, feedforward, range constrained recognition organize. This mode was utilized to distinguish the algorithm. In this technique, the exploratory data are perused just twice. In the outer layer, a numerical esteem is made as exchange rank. It is lower than the limit, that exchange will be recognized as a crimeulent exchange. Hawk crime the board framework that is a useful asset to keep the movement of crimesters in the abuse of charge cards utilizes the algorithms of neural systems. This framework predicts the likelihood of crime on a record by looking at the present exchanges and the past exercises of every holder (Hassibi 2000). [4] In 2004 [5], a structure was exhibited on the base of security frameworks [6] and Case based thinking [7] for crime discovery. Initial, a lot of ordinary and crime cases are produced using named data. At that point, the essential indicators are made with irregular or hereditary algorithms. At that point, negative determination and clonal choice activities are connected on essential indicators so as to get a lot of finders with various algorithms that can recognize an assortment of crimes. The introduced by Bhattacharyya in 2010 is about crime identification in charge cards of budgetary establishments and in this 50 million exchanges identified with one million Mastercards were utilized. Since the proportion of proper exchanges to exchanges with crime in this examination have been equivalent to 0/05%, the creators utilized the accompanying inspecting strategy so as to make a lot of data with various proportions from crimeulent records in the dataset (2%, 5%, 10% and 15%). [8] One of the most up to date distributed s in the field of crime location in Visas is the investigation of Dal Pozzdo et al [9]. In this investigation, the creators concentrated on 3 imperative issues of data irregularity, irregularity, and evaluation of techniques. They considered the exchanges of charge cards as data stream, that the crime identification framework must have the capacity to distinguish the crimeulent exchanges promptly. Sasirekha in his investigation in 2012 [10] expressed that numerous crime discovery frameworks that have been exhibited up until this point, have utilized data mining and neural system approaches. While no crime discovery framework with the blend of abnormality identification, abuse recognition and basic leadership framework have been utilized so far for crime location in Visas. At that point, a framework was suggested that utilized Hidden Markov Model to recognize the crimeulent exchanges. Lago in 2008 [7] utilized 5 techniques for classification for crime location: ،Naive Bayes ،Bayesian Network ، Artificial Immune System and Decision Tree. To actualize these techniques, Weka device was utilized with the exception of Artificial Immune System strategy that has a different program. Every one of these techniques was assessed in two methods of touchy to cost and straightforward. It implies that in the main mode, the cost identified with typical cases that are distinguished as crimeulent accidentally vary from the costs identified with crimeulent cases that are recognized as expected unintentionally. Additionally, the parametric strategies were assessed once by Weka parameters and some other time by advanced parameters. The consequences of looking at these two modes demonstrate that in any of these strategies, Weka parameters were not ideal. To streamline the parameters, cross breed highlight determination and hereditary algorithm were utilized. Akhilomen in 2013 [18] exhibited a mode by utilizing cross breed highlight choice and peculiarity discovery algorithm so as to recognize crime in charge cards. In this investigation, the general population who perform crimeulent exercises in the field of Visas were characterized in 3 gatherings of 1. The purchasers of charge cards data. 2. Dark cap programmers. Also, 3. The Thieve of Mastercards. The creators have noticed that crime location on the web must be done on the web and promptly. Since the utilization of charge card via

card holders pursues a settled example, this settled example can be removed from a standard lawful action of card holders in 1 or 2 years .along these lines, this example is contrasted with the utilization of procedure of card holder and in the event of non-likeness in the example, the action is viewed as unlawful. It ought to be noticed that the neural systems were utilized to show the examples location in the model in this consideration.

III. METHODOLOGY

Predictive Modeling

Predictive modeling in data mining refers to predicting a particular pattern which will be formed by collecting specific data and which will give additional data or information about the existing database. Hence it is helpful in detecting frauds as, in bank database the relevant information of frauds which has been done in past records are collected and a similar pattern or statistic model is developed which will predict the future fraud, hence bank can prevent it by analyzing the report generated by predictive model. Predictive modeling can be done in various techniques as Decision Tree Algorithm or Artificial Neural Network Algorithm or Naïve Bayes Algorithm [1].

Clustering/Segmentation

Clustering refers to making a cluster of data from the provided database. It is required when we want to find the same group of data or recognize the same pattern for analysis. Clustering can be performed with number of algorithms such as K-means, K-Medoids, etc. This algorithm will make group of data which is similar to the structure or characteristics called as clusters [2]. Clustering will help in bank database for finding same preference of customers or same type of transactional account holder; hence it can help to draw similar kind of fraud techniques used in fraud detection.

Segmentation is performed to produce finer data patterns. Segmentation can be performed using three different algorithms that is, Sliding window approach, topdown approach, or bottom-up approach. Hence segmentation is performed because it produces finer and clearer clusters than clustering algorithm [3].

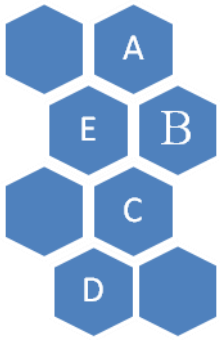

Fig Clustering/Segmentation of database

Visualization

Visualization technique in data mining is introduced for more effective presentation of formed data. Hence study reviews that human brain is more functional to remember visualizing the image rather than remembering information in data format [8]. Visualization converts the any raw data of characters or numbers to the image; image is in the form of static graph or any kind of graphical representation. Visualization includes techniques such as tree map, scatter plot matrix, parallel coordinates, and spatial visualization [4]. Banking data base is very large and there is possibility of conflicting of data to do the study of banking database visualization technique is very effective. It can draw the tree map of or plot matrix to encounter in which area the actual fraud has been done or the type of customer or employee which have done the fraud also branch location where it has been done. Hence it is useful in locating the fraud area and to prevent it.

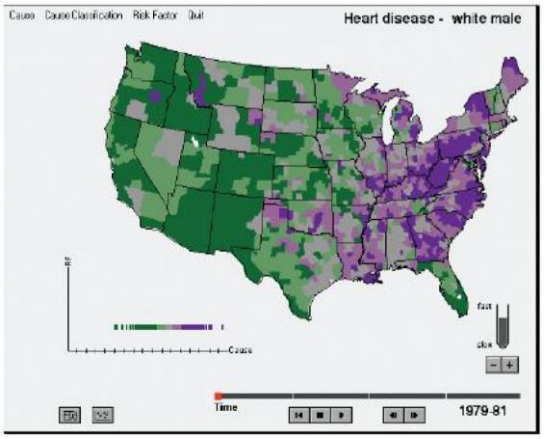

Fig E.g. Special graph visualization

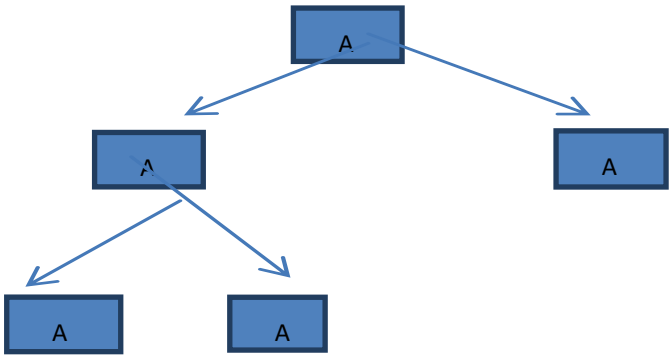

Fig E.g. Tree Diagram

Link analysis

Link analysis is one of the most important operations of data mining. It works very efficiently to find out the related data to each other. Link analysis is found out the related data i.e. one part of database is linked with other part and that connection is established by link analysis. In banking database customer linked with account then account linked with transactions further it link with type of transactions and this will continued.

Link analysis is based on part of mathematics called as graph theory where edges are connected to each other by some vectors to find out the correct path or some particular pattern of given data. It helps in fraud detection in many ways as bank employee can link the fraud detection area with each other and can summarize the data to find out the exact problem and their solution.

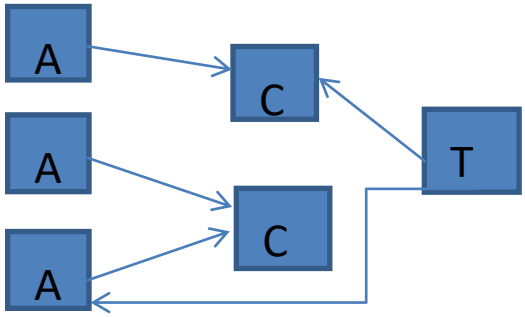

Fig E.g. Link analysis A-Account C-Customer T-Transaction

Deviation detection

Deviation detection is refers to identifying the errors or noise or exception which occurs in database. Clustering algorithm will only cluster the pattern of deviated data and non- deviated data, clustering algorithm focuses on eliminating the exception from the database where as deviation detection algorithm focuses on isolating those exception for better performance and error free data. Deviation detection is helpful while testing the many application of bank like credit card fraud can be isolates or rectify by deviation detection algorithms. Deviation detection algorithm considers the sequential exception problem to work on large databases; hence it discovers the all possible exceptions on every single dataset [5].

Data summarization

Data summarization is the major part of mining techniques at this will give the final outline to the data which we gathered from the previous operations i.e. relations between the clusters and the dependency of the subsets of data are clarify to take the important decisions. Hence data summarization will give idea about which product should be brought together and which should be avoided for better performance also to avoid the conflicts of data and to prevent the data from fraud and provide security to bank database. Data summarization is process of generating the better and more informative version of original database [7].

IV. COMMON TYPES OF CYBER CRIMES

Based on the person who committed the fraud, we can broadly classify them into two categories. When the fraud is committed by an outside person of an organization – say the customer – then it is called an Outsider Fraud. And when the fraud is committed by a person associated with organization – say the employee – then it is called an insider fraud.

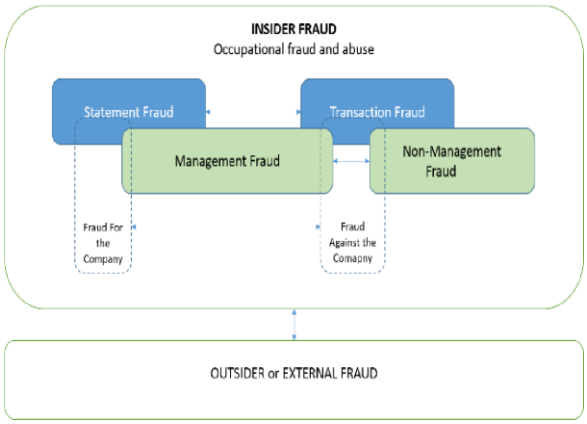

Figure Relations between the fraud taxonomies

There may be ways in which an insider commits the fraud. Following are the some common areas in which most of the inside frauds are reported:

1. Billing- Frauds related to procurement and disposal of assets / Account Payable Fraud:

These are some kind of frauds which are common in every industry and having high commit rate. Banking is not an exception to these. Some of the types are Fictious Vendors, Altered invoices, Fixed Bidding, Goods not Received, Duplicated Invoices, Inflated Prices, Excess Quantities Purchased and Duplicate Payments and Duplicate Serial Numbers, Payroll fraud, Account Payables[8].

2. Corruption- Frauds related to Lending & financial transactions

These are frauds which can happen only in a financial institution like financial institutes. Banks basically being the custodian of people‟s money, and their main source of income is through leading the money to needy. To do this they have been authorized to exercise some financial powers. But some official miss-use these power. Recent incident of CMD of Syndicate Bank, bribe-for-loan case is a good example of for the depth of corruption in the banks. There are many other ways of coming frauds, by exercising financial powers. For instance, BGL frauds, TDS frauds, misuse of unclaimed deposits, misuse of dormant accounts, ATM Related frauds etc. An interesting point to be noted here is, we might think that, in the days CBS, these kind of frauds are committed by the individuals having good system knowledge. But, the fact is these

frauds are committed by ordinary officers. But due to lack of efficient employees or mechanism to identify such frauds, has led these frauds escape un-noticed.

3. Financial Statement Frauds

Financial Statement frauds or Window dressing is one of the major problems in banks. Because each branch acts as a separate entity and each branch has to achieve their targets. Some branches adopt wrong ways to meet the target, which intern causes wrong Financial Statement at the head office level. Some ways of doing that are, transferring funds from unutilized CC accounts to SB/CA‟s to show positive growths in deposits and advances, transferring some amount to an overdrawn CC account, to make it below the limit, intern avoiding the account becoming NPA, Raising funds for short term, such that at the end of the Financial period the figures show positivity, Postponing the remittance to government account etc.

4. Expense Reimbursement - Frauds in availing facilities such as Entertainment, travels expense, out of pocket expense, Leave travel concession:

Banks provide many facilities for their employees in view of smooth functioning of banks. For example: Entertainment allowance, which can be utilized to make expense on offering of Soft drinks, coffee, tea, snacks to valuable customers, LTC(Leave travel concession), which is provided to go for a tour with the family. In the case of LTC (Annual Report – CVC – 2013) the modus operandi adopted includes use of forged/fake Air India tickets & boarding passes, claiming irregular reimbursements and in many cases officials have not travelled at all. The officials indulge in irregular claims like travelling by flexi/easy fare tickets by Air India and receiving cash discounts for the difference between LTC fare and flexi/easy fare from travel agents. In PSEs and banks, the LTC facility is allegedly used by officials for visiting abroad in collusion with certain airlines and travel agents. Commission also noticed instances where officers of Public Sector Banks visited foreign destinations and thereafter visit the designated place in India using a circuitous route on flexi/easy fare and claims were settled on the basis of full fare of entitled class to the designated placed in India. The guidelines and interpretation of „circuitous route‟ were being abused in many cases. Instances of receipt of cash discounts from travel agents have also been observed. In case entertainment allowance or out of pocket expenses, we can see examples of Over Limits, unusual or inappropriate expenses, miscellaneous/sundry expenses, split or duplicate expenses.

5. Insurance related frauds - Cross-Selling or Miss-selling?

Cross-Selling means encouraging a customer who buy product to buy a related or complementary product, with a view to expand banking business, reduce the per customer cost of operation and provide more satisfaction and value to the customer.

What is mis-selling? As per the IRDA, mis-selling can be defined as:

„By definition, mis-selling means selling a product by giving a wrong picture of a product, it may include, giving wrong information, giving unrealistic information, not giving full information about the product. You must have heard an insured, saying – but this was not I asked for. And, your agent accusing, but then I did mentioned all the details upfront, didn't I? Insurance is a business of selling commitments and here is a case where this was broken. Unfortunately the product was mis-sold. Mis-selling is not unique to insurance and happens in various lines of businesses (loans, credit cards, investment products, pharmacy, hospitality etc.), but Insurance being an intangible service – the principle of

Caveat emptor prevails in insurance‟. The **table 2** shows the complaints received by IRDA on „UNFAIR BUSINESS PRACTICES‟. We can observe from the statistics the complaints are increasing and which implies the mis-selling is also increasing.

As rightly said by Mr.P.Chidambaram, in an interview : "The reason why insurance is stumbling in India is because of the mis-selling of products and complex products. Also, If you want to sell insurance to India, you must sell simple products and must make it absolutely clear to the agents and to other officers that they should not mis-sell." LIVE MINT, FEB 12, 2013

6. Money Laundering:

Out of 140 countries India is ranked 93rd, 70th & 88th in 2012, 2013 & 2014 respectively compared to Norway, which has ranked 1st in Anti Money Laundering (AML) Basel Index[9]. This clearly show that India, in the present scenario is very vulnerable to money laundering activities and is a high risk zone.

Money laundering refers to conversion of illegal or black money in such a way that it appears that it is obtained from a legitimate source. In India it's done through a system called "Hawala" which means transfer of money. It's done in 3 stages called, Placement, Layering and Integration. Banks are used intensively in the 2nd stage, i.e., layering where fund collected is channelized through different instruments. Usually they use many fictitious account created. Even though strict implantation of KYC (Know your customer) is mandatory as per the guidelines of RBI, its difficult implement the same in rural areas. Fraudsters, with the help of corrupt bankers exploit these loop holes to run their show.

7. Identity theft :

Identity theft is committed by accessing personally identifiable information of individuals/ entities without their permission with the objective to misuse this information and gain undeserving benefits. One of the primary reasons for growth in identity theft has been the proliferation of the Internet. In India, the number of internet users has grown from 7 million in 2001 to over 98 million by 2011 and is expected to reach the 300 million mark by 2015. However, controls and regulation aimed at protecting privacy have not kept pace with this growth. Fraudsters are making use of these gaps in controls to target individuals and organizations and misuse confidential data. According to the Norton Cybercrime Report 2011, four out of five online adults in India were victims of identity theft in 2011. Considering that many employees in the corporate work force use their office laptop/computer for personal transactions (such as online banking, shopping, payment of bills etc), identity theft can compromise not just their private information, but also the companies they work for. Banks can also be a source of identify, as banks will have all the personal/confidential information of an individual. It's not only customers‟ data, we have seen the incidents of employees' identity theft. It's more dangerous than identity theft of customer data. As the thief can use that for fraudulent transactions and can easily escape. Hence it is very much essential for banks to identify & prevent the chances of thefts.

V. PROPOSED METHOD

The proposed method in this section of the study, the new solution is introduced. The proposed solution was developed by using the reinforcement learning in the neural network. Artificial neural network is a practical method for learning different functions like functions with real values, functions with discrete values and functions with vector values. A neuron alone can be used only for the detection of functions that are linearly separated. Since the functions are not linearly separable in real problems, a network of neurons is needed. A variety of neural networks is used for solving different learning problems with monitoring, learning without monitoring and reinforcement learning. Neural networks are divided into two groups of FNN (feed-forward neural networks) [12] and RNN (recurrent networks) [13] based on a variety of connections. FNN(s) are the most regular types of neural networks that are used in different functions. The first layer is called the input layer and the last layer is called the output layer and each number of layer among these two layers is called middle or hidden layer, because we are only involved with the inputs and outputs of the neural network. Neural network works as a black box and direct access to middle layers is not possible. Recurrent neural networks have oriented cycles in their graphs structure. In other words, we can return to previous and early nodes by tracking the connections between nodes. RNN(s) have a complex dynamic according to their structure and makes it difficult to teach these networks. Also, FNN networks are biologically closer to reality. FNN networks with more than one hidden layer are called MLP (multi-layer perception) and FNN network with one hidden layer are called SLP in which the output of neurons in each layer is a nonlinear function of outputs in previous layer. The number of neurons in the input and output layer is constant and the number of neurons in the input layer is equal to the space of features and the number of neurons in the output layer is specified according to the number of classes. In MLP, the nodes (neurons) are usually ordered in some layers of neural network. Each node only receives the inputs from the previous layer and provides a function of inputs.

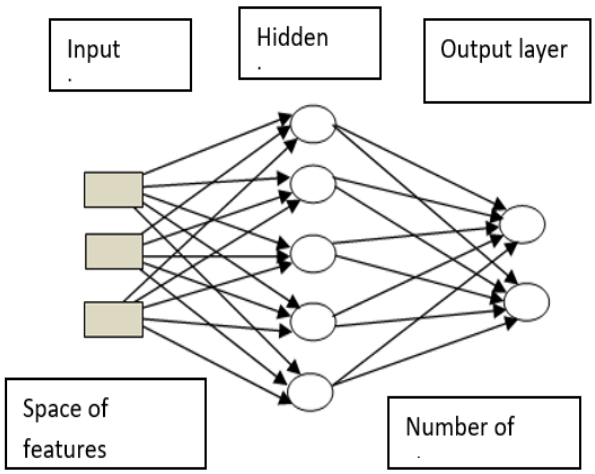

Fig. Proposed Architecture diagram

Evaluation of Big Data Fraud Detection Techniques

This section extends the analysis of the techniques proposed previously for the big data forensic investigation. There are different factors that can influence the selection and performance of the forensic techniques. These factors should be analyzed closely before carrying out the implementation. The following table has been populated based on the available resources and the sensitivity of the data to be used for investigation.

Table .1: Analysis of Techniques and its Factors

Factors / Techniques	Processing Speed	Latency	Fault Tolerance	Performance	Scalability
Hadoop	Medium	High	High	Slow	Medium
MapReduce	Slow	High	High	Slow	Medium
Spark	Fast	Low	High	Fast	High
Flink	Fast	Low	High	Fast	High

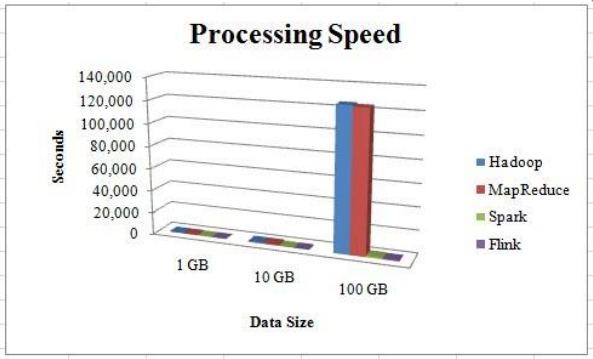

Figure : Processing Speed

In Figure , it can be seen that the processing speed of Apache Spark and Flink remains the same even though the data size increases. But Hadoop and MapReduce increases its processing time if the data size is too big to handle.

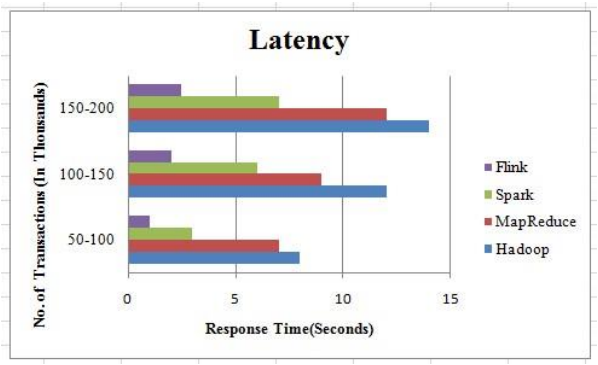

Figure: Latency

The response time should be at minimum while executing the data/transactions. Among the four techniques, Apache Flink has low latency while processing Big Data.

While analyzing the Fault tolerance factor, Spark system replicates the input data in memory which is a most useful solution for handling faults in between the execution of transactions. Data lost due to failure can be recomputed from replicated input data.

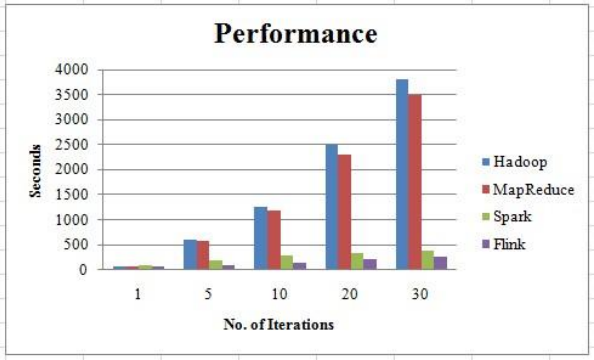

Figure: Performance

From the Figure, it is noted that Spark has small variability in the execution time when compared to other techniques.

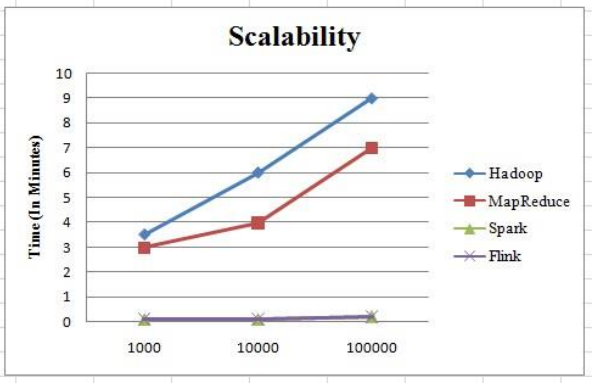

Figure : Scalability

With scalability factor, Spark and Flink processes the data smoothly if there is need to increase nodes for processing Big Data. Although the number of node increases, Spark performs better with Big Data.

VI. CONCLUSION

In this paper, the techniques such as processing speed, Cyber Crime, latency, fault tolerance, performance and scalability that were previously being used in a different context have been analyzed that can facilitate the forensic investigator in performing big data forensic. In this work, all the four techniques have been evaluated on the basis of factors such as processing speed, latency, fault tolerance, performance and scalability and spark is suggested as a better performing technique among others. Combining the results,

Apache Spark is considered as an efficient technique that helps to implement credit card fraud detection system.

References

[1] Aditya B. Patel, Manashvi Birla, Ushma Nair , "Addressing Big Data Problem Using Hadoop and Map Reduce", International Conference on Engineering, 2012.
[2] Alessandro Guarino, "Digital Forensic as a Big Data Challange", ISSE Securing Electronic Business Processes, 2013.
[3] Alexander, Rico Bergmann, Stephan, "The Stratosphere Platform for Big Data Analytics", the VLDB Journal, 2014.
[4] Abbott, D., Matkovsky, P. & Elder, J. (1998)." An Evaluation of High-End Data Mining Tools for Fraud Detection." Proc. of IEEE SMC98.
[5] Barse, E., Kvarnstrom, H. & Jonsson, E. (2003). "Synthesizing Test Data for Fraud Detection Systems". Proc. of the 19th Annual Computer Security Applications Conference, 384-395
[6] R.Anbuvizhi, V.Balakumar, "Credit / Debit Card Transaction Survey Using Map Reduce in HDFS and Implementing Syferlock to Prevent Fraudulent",International Journal of Computer Science and Network Security, 2016.
[7] Anushree Naik, Kalyani Phulmamdikar, Shreya Pradhan, Sayali Thorat, Prof. Sachin V. Dhande, "Real time Credit card transaction analysis", International Engineering Research Journal (IERJ), Vol 1, Issue 11, 2016.
[8] "Blog: Apache Flink" [Online], Available: www.odbms.org/blog/2015/06/on-apache-flinkinterview-with-volker-mark1/
[9] Brian Ye, Anders Ye, "Exploring the Efficiency of Big Data Processing with Hadoop MapReduce", School of Computer Science and Communication (CSC), Royal Institute of Technology KTH, Stockholm, Sweden.
[10] Cameron S.D. Brown, "Investigating and Prosecuting Cyber Crime: Forensic Dependencies and Barriers to Justice", International journal of Cyber Criminology, 2015.