Fraud Find: Financial Fraud Detection By Analyzing Human Behavior

Tejashree M [1], Thejashreee J [2], Varshini S[3], Ms. Sahana M [4]
[1, 2, 3] Dept of Information Science
[4]Asst. Professor, Dept of Information Science
[1, 2, 3, 4] East West Institute of Technology, Bengalure, India.

Abstract- *Financial fraud is commonly represented by the use of illegal practices where they can intervene from senior managers until payroll employees, becoming a crime punishable by law. There are many techniques developed to analyze, detect and prevent this behavior, being the most important the fraud triangle theory associated with the classic financial audit model. In order to perform this research, a survey of the related works in the existing literature was carried out, with the purpose of establishing our own framework. In this context, this paper presents Fraud Find, a conceptual framework that allows to identify and outline a group of people inside an banking organization who commit fraud, supported by the fraud triangle theory. Fraud Find works in the approach of continuous audit that will be in charge of collecting information of agents installed in user's equipment. It is based on semantic techniques applied through the collection of phrases typed by the users under study for later being transferred to a repository for later analysis. This proposal encourages to contribute with the field of cybersecurity, in the reduction of cases of financial fraud.*

Keywords- bank fraud; triangle of fraud; human factor; human behavior

I. INTRODUCTION

Fraud is a worldwide phenomenon that affects public and private organizations, covering a wide variety of illegal practices and acts that involve intentional deception or misrepresentation. According to the Association of Certified Fraud Examiners (ACFE) [1] fraud includes any intentional or deliberate act of depriving another of property or money by cunning, deception or other unfair acts.

The 2016 PwC Global Economic Crime Survey report describes that more than a third of organizations worldwide have been victims of some kind of economic crime such as asset misappropriation, bribery, cybercrime, fraud and money laundering. Approximately 22% of respondents experienced losses of between one hundred thousand and one million, 14% suffered losses of more than one million and 1% of those surveyed suffered losses of one hundred million dollars. These high loss rates represent a rising trend in costs caused by fraud. In organizations, 56% of cases are related to internal fraud and 40% to external, this difference is since any individual related to accounting and financial activities is considered a potential risk factor for fraud [2]. When observing the behavior of people in the scope ofbusiness processes, it can be concluded that the human factor is closely linked and related to the fraud triangle theory of the Donald R. Cressey [3], where three basic concepts: pressure, opportunity and rationalization; are needed.

Nowadays, there are different solutions in the commercial field [4], [5] as well as the academic field, where some works in progress had been identified [6], [7] aimed at detecting financial fraud. In both cases, these solutions are focused on the use of different tools that perform statistical and parametric analysis, as well as behavioral analysis, based on data mining techniques and Big Data; but none of them solve the problem of detection financial fraud in real time. FraudFind, unlike other proposals, detects reports and stores fraudulent activities in real time through the periodic analysis of the information generated by users for further analysis and treatment. In real time, based on the theory of the fraud triangle.

For the design of the FraudFind framework, some software components related to the processing of information were analyzed, among them, RabbitMQ, Logstash and Elastic Search. In addition, the computerization of the triangle of fraud and the use of semantic techniques will allow finding possible bank delinquents with a lower false positive rate. In our paper, we apply multiple binary classification approaches- Logistic regression, Linear SVM and SVM with RBF kernel on a labeled dataset that consists of payment transaction. Our goal is to build binary classifiers which are able to separate fraud transaction from non-fraud transaction. We compare the effectiveness of these approaches in detecting fraud transaction.

The rest of the document is structured as follows. Section 2 presents related works found in the literature. Section 3 presents the theoretical framework on the definition

of Fraud and the concept of the fraud triangle. Section 4 details the architecture of the model and the prototype to be implemented as future work. Section 5 continues with the discussion and section 6 concludes with the conclusions and future work

II. RELATED WORK

This study aims to design an architecture model adapted to the fraud triangle factors, complemented with the humanfactor and analyzing suspicious behavior to identify possible cases of fraud. For a future work to carry out its imple- mentation. In this context, several studies were found in the literature, which contribute to this topic.

Most of the documents address the issue of financial fraud and the different circumstances surrounding it. Nevertheless, identifying people who might be involved in fraudulent activities is a determining factor. The incursion into the behavioral analysis is quoted to [6], whose authors introduce an automatic text mining process by e-mail for the detection of different types of patterns in messages. While in [7] a generic architectural model is proposed that supports the factors of the fraud triangle. In addition, it performs the classic quantitative analysis of commercial transactions that are already applied as part of the fraud detection audit. The identification and classification of possible fraud by suspicious individuals is a central element of the internal threat prediction model [8]. A key aspect is to classify individuals by focusing on reducing the internal risk of fraud through a descriptive mining strategy [9].

Besides, the experience of auditors plays an important role in the fight against financial fraud. Some work is proposed which points to the creation of new frameworks that provide systematic processes to help auditors to discover financial fraud within an organization by analyzing existing information and data mining techniques using their own experience and skills [10]. Accordingly, another proposal creates generic frameworks for the detection of financial fraud FFD, to evaluate the different characteristics of FFD algorithms according to a variety of evaluation criteria [11].

New approaches detect a typical values by studying and modifying clustering algorithms such as K-Means, with the purpose of improving the performance and accuracy in the detection of unusual values in a data set [12]. Capturing unusual patterns related to fraudulent activity involves the analysis of the number of variables that can be examined simultaneously, the same as with the technological advance have increased considerably and can be addressed by the use of more sophisticated neural networks increasing the number of neurons and / or layers at the expense of a higher computational cost [13]. An important factor to mention is how expensive it is to detect potential fraudulent transactions manually. For this reason the FFD is vital for the prevention of the destructive consequences of financial fraud by making a complete comparison of data mining techniques in order to use the best one [14].

Reviewing the literature, it can be concluded that related work does not cover the anticipated detection of fraud, since they perform an analysis after the incident occurred. This paper aims to reduce this gap by conducting an online fraud audit by developing a model that will allow the timely identification of suspicious behavior patterns taking into account the human factor supported by the fraud triangle theory. This prototype is a tool that will allow individuals to be analyzed inside a corporation in order to identify possible cases of financial fraud.

III. FRAUD AND THE FRAUD TRIANGLE THEORY

In general, there is not an scientific definition of fraud. Nevertheless, it is considered as a subset of internal threats such as corruption, misappropriation of assets, fraudulent statements, among others [15]. According to ACFE, fraud is defined as [1] *"the use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resources or assets".* How- ever, due to the scope of this paper, only financial fraud will be considered within a banking environment. In financial fraud there are two types of fraud: internal and external [16]. Internal fraud encompasses a series of irregularities and illegal acts characterized by the intentional deception of fraudsters leading to the misappropriation of money and other important resources of the company. In the case of external fraud, this is commonly done in the financial statements, which are falsely presented in reports. Most of the known anomalies are due to the weakness of the internal control mechanisms and in such situations the fraudsters commit acts of fraud exploiting these weaknesses.

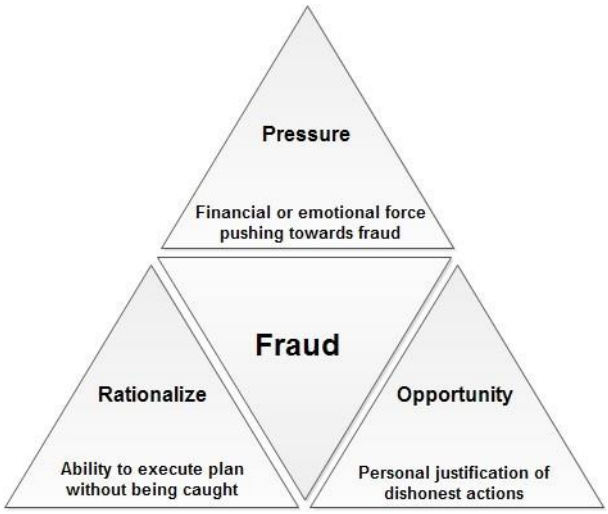

Figure 1. Triangle of Fraud

The occurrence of fraud is best explained with the help of The Fraud Triangle Theory, illustrated in Figure 1, proposed by Donald R. Cressey, a leading expert in crime sociology who wrote a series of books on crime prevention. Cressey investigates the reasons behind the question of ”why do people commit fraud?” and determines the response in the following three critical elements: perceived pressure, perceived opportunity and rationalization. Cressey's theory implies that the three elements must be consecutively present to provoke the desire to commit fraud. The first necessary condition in the fraud triangle is the idea of perceived pressure related to the motivation and impulse behind the fraudulent actions of an individual. This motivation often occurs frequently in people under some kind of financial stress. [17]. The second element is the perceived opportunity, it is the action behind the crime and the ability to commit fraud. Finally, the third component relates to the idea that the individual can rationalize his dishonest actions, making his illegal choices seem justified and acceptable [18]. The risk of committing fraud increases exponentially when there is an increase in the connection between pressure, opportunity and rationalization.

IV. FRAUDFIND FRAMEWORK

The proposed framework operates in the continuous auditing approach to discover financial fraud within an organization belonging to the banking sector which will be our main study environment and also focused on the fraud triangle theory with the human factor considered as an essential element. Fraud Find is proposed with the objective of analyzing large amounts of data from different sources of information for later processing and registration, using the ELK stack. ELK is a scalable open source platform used for real-time data analysis composed by Elastic Search, Logstash and Kibana [19] [20] applications, which will be explained below.

1) ElasticSearch is an open source search engine developed in Java, which is a distributed, scalable document warehouse and works in real time. Designed mainly to organize data in order to be easily accessible [21].
2) Logstash is an open source tool used for event management, by centralizing and analyzing a large number of structured and unstructured data types [22].
3) The Kibana web interface is an adjustable board that can be altered and changed to suit our environment. It allows the creation of tables and diagrams, in addition to complex representations [20].

In Figure 2 we can observe the different modules that compose the framework: Agent, QoS, Collect & Transform, Search & Analyze, and View & Manage.

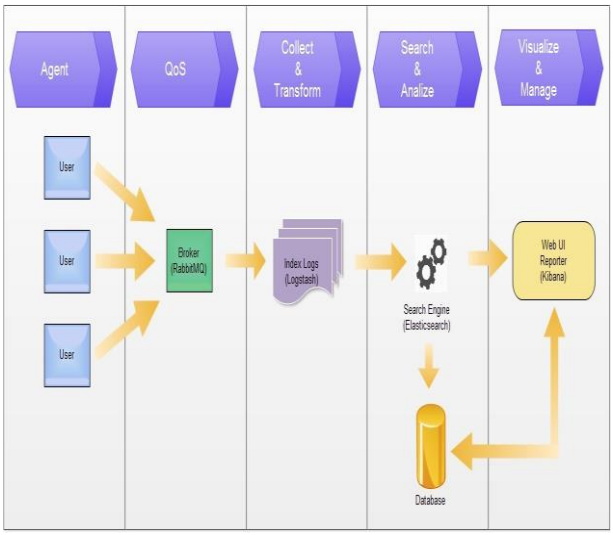

Figure 2: Fraudfind Framework

A. *Agent*

The agent is an application installed in the workstations of the users (endpoints), in order to extract the data that they generate from the different sources of information that reside on their equipments. This application is responsible for sending the data entered by the user into RabbitMQ for ordering and classification. Later this organized information is received by Logstash for its treatment.

B. *QoS*

The integration between several systems or components suggests the need to receive or send information, so these communications must be reliable, safe, fast and above all be permanently available. Due to that the volume of

information generated by the agents is considerable and recurrent, this module will ensure its delivery in an orderly and reliable way to Logstash. For this, an intermediary component was introduced, RabbitMQ, to organize and properly distribute the data to for further processing. RabbitMQ is an open source platform that operates as a message broker, where third-party applications can send and receive messages offer- ing persistence, confirmation of sending-receiving and high availability. The cluster of RabbitMQ servers can form a logical broker allowing the implementation of features such as load balancing and fault tolerance. By default RabbitMQ sends the messages using the Round - Robin algorithm. After being delivered it is removed from the queue [23]. Figure 3 shows the operation of RabbitMQ.

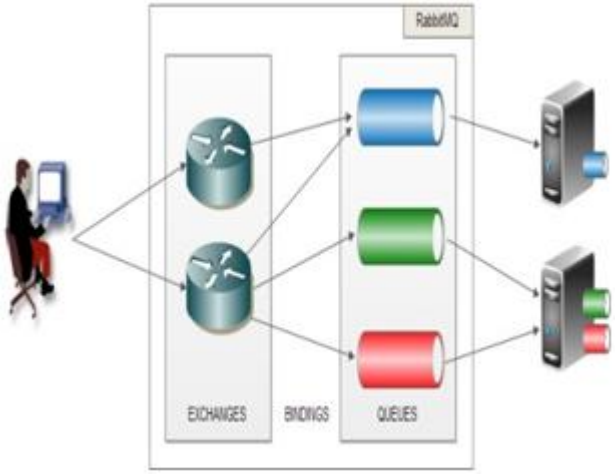

Figure3.RabbitMQ

C. Collect and Transform

This module is responsible for processing the data sent by the agents. As seen in Figure 2, after ordering the input data of the agents in the QoS module, they are recorded in a temporary file that has raw information that Logstash does not understand and does not know how to handle it. To inter- pret this information, Logstash has tools called codec's and filters, which perform operations and transformations on the collected data, allowing this information to be converted into a compressible format. Once processed, the information is sent to Elastic Search for storage. The operation of Logstash is presented in Figure 4.

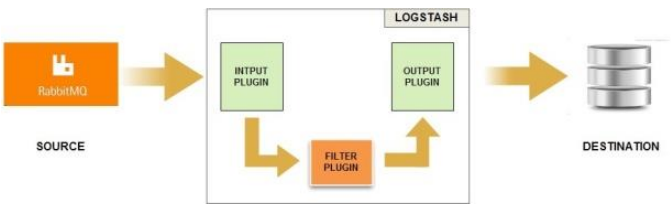

Figure 4. Logstash

D. Search and Analyze

This module has all the information processed by Logstash, which is stored immediately after it is received, being able to perform searches efficiently. ElasticSearch is a tool designed with the clustering approach, based on the premise of no fault tolerance hardware. With this property, the information is protected and replicated so that if the physical infrastructure collapses the data will not be com- promised. Figure 5 shows the architecture of ElasticSearch and its component.

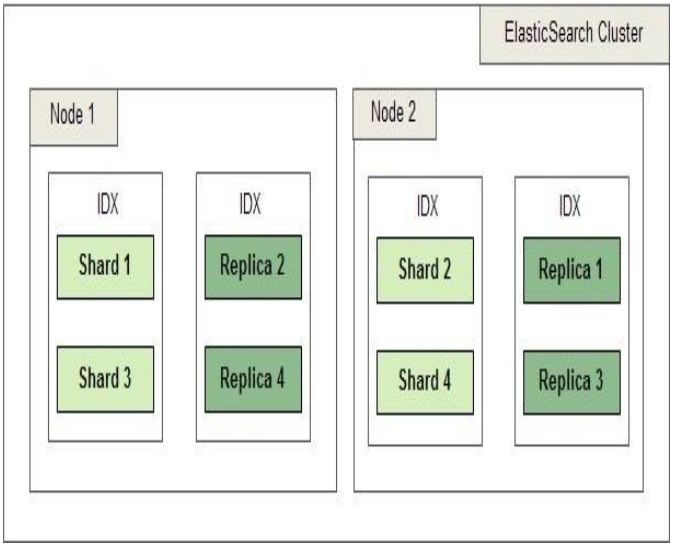

Figure 5. Elastic Search

E. Visualize and Manage

Finally, in this module the presentation of the data con- tained in Elastic search is performed, using for this purpose Kibana. This tool has been designed to work with Elas- tic Search that allows the visualization and search of the information in a customizable way, using histograms, pie charts, and metrics among others. This tool provides information analysis in real time.

V. FRAMEWORK IMPLEMENTATION

In this section, we describe a prototype for the automatic detection of financial fraud, which is currently in the implementation phase. In Figure 6 we can see the diagram of the proposed framework implementation, which describes the concept of the different modules for practical implementation using free and open source platforms.

To begin, the information extracted by the agents is sent through data queues, which must be attended quickly, safely and in a reliable way. To achieve this goal,

RabbitMQ has been used, which is an open source message broker that implements the Advanced Message Queuing Protocol (AMQP) standard. First of all, several RabbitMQ servers on a local network can be grouped into a logical (distributed) broker. This allows the implementation of features such as load balancing and fault tolerance. Another important feature is the AMQP protocol that RabbitMQ uses, which accepts connections between different platforms.

The data sent by RabbitMQ is received by Logstash for its treatment (organize and categorize). Logstash is a tool that collects, processes and filters information. According to Figure 4, it is composed by 3 main plugins: input, filter and output. First, we have the input plugin that allows the collection of records in different formats, such as: files, TCP / UDP, etc. Second, we have the filter plugins that allow Logstash to execute the transformation on the input data. Finally, the output plugin allows processed and transformed data to be written in a variety of formats that go to ElasticSearch [24].

The data sent by Logstash is received by ElasticSearch which indexes and analyzes this information. ElasticSearch is a search and storage engine that can handle lots of data in real time, providing speed and reliability, along with Kibana, as a visualization tool.

Periodically, a task that do the alert tracking, checks the information entered and compares it with a fraud triangle library to determine if there is a relation in order to generate an alert that will be stored in the database. The library of the fraud triangle is just a dictionary that contains three definitions: pressure, opportunity and justification. Under these parameters, the sentences and words associated with these behaviors are composed.

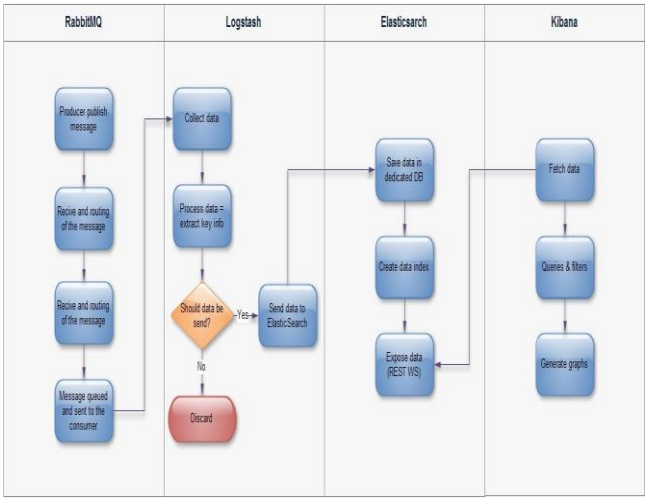

Figure 6. Framework Implementation

VI. ANALYSIS AND DISCUSSION

Performance analysis

FraudFind consists of the extraction of data from different sources of information through agents installed in workstations, which collect behavioral data and send these information in an organized way, reporting its activity to the central server. The typed words are sent to Rabbit MQ, an application that manages message queues, which delivers fast, secure and reliable information to Logstash, a tool used to collect, analyze data from monitoring heterogeneous sources and finally to Elastic Search that performs indexing. All this is aimed at ensuring the security in the transactions generated by the users trying to identify possible acts of fraud through the analysis of human behavior and the treatment of the results. Unusual behavior does not guarantee the intentionally of committing fraud, so it should take into consideration the analysis of risk factors associated with this behavior, which should be measurable and weighted in accordance with security policies in an organization.

When there are different sources of information, we find inconsistency in the logs, given that the formats are different. This represents a problem since administrators require access to this information for analysis and there is the difficulty for searching in different formats. When Logs are distributed among the different analysis teams, they are decentralized and each of them has a different format and different routes to find them, complicating their administration and analysis. ELK solves these problems because it collects all this information with the aim of processing it, storing it in a distributed manner and using treatment techniques such as big data to obtain accurate results.

Additionally, the study of human behavior plays an im- portant role in this work because through this analysis it is possible to discover transactions that are part of a pattern
not identified in the data traffic and that would have stopped discovering using traditional means.

Technical analysis

The ELK (Elastic Search, Logstash and Kibana) platform provides versatile and functional records management when searching and analyzing information from a source. To evaluate the efficiency ADABOOST and voting methods are used. Centralized data logging can be useful for identifying unusual traffic patterns, allowing you to search for all stored records that quickly execute the necessary event correlation.

Security analysis

The possible violation of privacy is a factor that should be considered when implanting this solution within a company. Legal data protection regulations should be considered in a given region. The legal regulations for data protection in a given region should be considered. The level of monitoring will depend on the internal policies in an organization and the laws that are governed in each country and should be determined taking into account the advice of the legal part of the institution or company.

VII. CONCLUSIONS

The present work proposes FraudFind, a conceptual framework to detect financial fraud supported by the fraud triangle factors which, compared to the classic audit analysis, makes a significant contribution to the early detection of fraud within an organization. Taking into account human behavior factors, it is possible to detect unusual transactions that would have not been considered using traditional audit methods. These patterns of behavior can be found in the information that users generate when using the different ap- plications on a workstation. The collected data is examined using data mining techniques to obtain patterns of suspicious behavior evidencing possible fraudulent behavior. Neverthe- less, the legal framework and the different regulations that are applied in public and private institutions of a particular region represent a high risk for the non-implementation of this architecture as an alternative solution. Future work will have as its main objective the implementation and evaluation of the framework as a tool for continuous auditing within an organization.

REFERENCES

[1] "ACFE Association of Certified Fraud Examiners," (Date last accessed 15-July-2014). [Online]. Available: http://www.acfe.com/uploadedfiles/acfewebsite/ content/documents/rttn-2010.pdf

[2] "PwC," (Date last accessed 15-July-2014). [Online]. Avail- able: https://www.pwc.com/gx/en/economic-crime-survey/ pdf/GlobalEconomicCrimeSurvey2016.pdf

[3] N. B. Omar and H. F. M. Din, "Fraud diamond risk indicator: An assessment of its importance and usage," in 2010 Inter- national Conference on Science and Social Research (CSSR 2010). IEEE, dec 2010.

[4] "Lynx," (Date last accessed 15-July-2014). [Online]. Available: http://www.iic.uam.es/soluciones/banca/lynx/

[5] "Ibm," (Date last accessed 15-July-2014). [Online]. Available: https://www.ibm.com/developerworks/ssa/local/ analytics/prevencion-de-fraude/index.html

[6] C. Holton, "Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem," Decision Support Systems, vol. 46, no. 4, pp. 853–864, mar 2009.

[7] S. Hoyer, H. Zakhariya, T. Sandner, and M. H. Breitner, "Fraud prediction and the human factor: An approach to include human behavior in an automated fraud audit," in 2012 45th Hawaii International Conference on System Sciences. IEEE, jan 2012.

[8] M. Kandias, A. Mylonas, N. Virvilis, M. Theoharidou, and Gritzalis, "An insider threat prediction model," in Trust, Privacy and Security in Digital Business. Springer Berlin Heidelberg, 2010, pp. 26–37.

[9] M. Jans, N. Lybaert, and K. Vanhoof, "Internal fraud risk reduction: Results of a data mining case study," International Journal of Accounting Information Systems, vol. 11, no. 1, pp. 17–41, mar 2010.

[10] P. K. Panigrahi, "A framework for discovering internal finan- cial fraud using analytics," in 2011 International Conference on Communication Systems and Network Technologies, June 2011, pp. 323–327.

[11] D. Yue, X. Wu, Y. Wang, Y. Li, and C. H. Chu, "A review of data mining-based financial fraud detection research," in 2007 International Conference on Wireless Communications, Networking and Mobile Computing, Sept 2007, pp. 5519– 5522.

[12] M. Ahmed and A. N. Mahmood, "A novel approach for outlier detection and clustering improvement," in 2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA*)*, June 2013, pp. 577–582.

[13] A. Vikram, S. Chennuru, H. R. Rao, and S. Upadhyaya, "A solution architecture for financial institutions to handle illegal activities: a neural networks approach," in 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the, Jan 2004, pp. 181–190.

[14] H. Li and M. L. Wong, "Financial fraud detection by using grammar-based multi-objective genetic programming with ensemble learning," in 2015 IEEE Congress on Evolutionary Computation (CEC), May 2015, pp. 1113–1120.

[15] D. Cappelli, A. Moore, R. Trzeciak, and T. J. Shimeall, "Common sense guide to prevention and detection of in- sider threats 3rd edition–version 3.1," Published by CERT, Software Engineering Institute, Carnegie Mellon University, http://www. cert. org, 2009.

[16] P. K. Panigrahi, "A framework for discovering internal fi- nancial fraud using analytics," in *2011 International Confer- ence on Communication Systems and Network Technologies*. IEEE, jun 2011.

[17] G. Mui and J. Mailley, "A tale of two triangles: comparing the fraud triangle with criminology's crime

triangle," Accounting Research Journal, vol. 28, no. 1, pp. 45–58, jul 2015.

[18] D. Al-Jumeily, A. Hussain, MacDermott, H. Tawfik, Seeckts, and J. Lunn, "The development of fraud detection systems for detection of potentially fraudulent applications," in 2015 International Conference on Developments of E- Systems Engineering (DeSE), Dec 2015, pp. 7–13.

[19] S. GVK and S. R. Dasari, "Big spectrum data analysis in dsa enabled lte-a networks: A system architecture," in 2016 IEEE 6th International Conference on Advanced Computing (IACC), Feb 2016, pp. 655–660.

[20] T. Prakash, M. Kakkar, and K. Patel, "Geo-identification of web users through logs using elk stack," in 2016 6th Interna- tional Conference - Cloud System and Big Data Engineering (Confluence), Jan 2016, pp. 606–610.

[21] U. Thacker, M. Pandey, and S. S. Rautaray, "Performance of elasticsearch in cloud environment with ngram and non- ngram indexing," in 2016 International Conference on Elec- trical, Electronics, and Optimization Techniques (ICEEOT), March 2016, pp. 3624–3628.

[22] D. N. Doan and G. Iuhasz, "Tuning logstash garbage col- lection for high throughput in a monitoring platform," in Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2016 18th International Symposium on. IEEE, 2016, pp. 359–365.

[23] V. M. Ionescu, "The analysis of the performance of rabbitmq and activemq," in 2015 14th RoEduNet International Con- ference - Networking in Education and Research (RoEduNet NER), Sept 2015, pp. 132–137.

[24] D. N. Doan and G. Iuhasz, "Tuning logstash garbage col- lection for high throughput in a monitoring platform," in 2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Sept 2016, pp. 359–365.