A Review of Evidence Extraction Techniques in Big Data Environment

Siti Hawa Mokhtar[1], Gopinath Muruti[2]
Department of System and Networking
College of Computer Science and Information Technology
University Tenaga Nasional
Selangor, Malaysia
ct.hawa10@yahoo.com,
gopinathmuruti@gmail.com

Zul-Azri Ibrahim[3], Fiza Abdul Rahim[4]
Department of System and Networking
Institute of Informatics and Computing in Energy
Universiti Tenaga Nasional
Selangor Malaysia
zulazri@uniten.edu.my,
fiza@uniten.edu.my

Hairoladenan Kasim
Department of Information System
Institute of Informatics and Computing in Energy
Universiti Tenaga Nasional
Selangor Malaysia
hairol@uniten.edu.my

Abstract*—Today, information era where data is being generated at high in volume, variety, and velocity, a new technology is needed to cope with such data. Companies are no longer depends on the traditional tools and techniques to cater and handle data. Not only ending on how to store and process the data, they also wanted to gain insight of the data to optimize business process and gain a larger profit. To satisfy these requirements, a good analytic method must be applied to big data in order to extract value and knowledge from these data sets. While computer engineers are working on that part, this valuable data is also being eyed somewhere else. New attacks and attempts to taint the security, privacy, and integrity of the data are being developed somewhere without we knowing. This paper aims to analyze different analytics methods and tools, which can be applied in big data environment, in actionable time while at the same time extract evidence of intrusion in order for the results to be presented in a court of law fitting a digital forensic process.

***Keywords*—Big Data; Big Data Analytic; Digital Forensic**

I. INTRODUCTION

Our world is moving at fast pace, companies and organizations are competing to get the upper hands in their industries by leveraging technology development. Internet of Things (IoT) has also been a common thing among society as well as corporate bodies. This competition of technology is even tougher in areas of service and utilities providers. In order to optimized their resources and fulfils their business needs, billions of sensors and networked computers are utilized to gather ,monitor and process data automatically .We take for example, utilities provider; The analog way of recording data would no longer suffice for to fulfil their business needs. This automation however, would surely enhance data recording and capture more data than before. The larger the volume, the more materials in hand for the companies to be creative with and gain better understanding of consumer behavior and consumption thus, optimizing production and distribution. In relation to this, the community is also becoming more and more obsessed with mobile gadgets, with an average of 3 digital devices owned per person according to Global Web Index [1]. It is no brainer that large data is generated every day. According to IBM, 2.5 quintillion bytes of data created every day and about 90% of data in the world today has been created in the last two years alone [2].

II. BIG DATA

Big Data can be defined as the explosion of high-frequency digital data with behavior of high in "Velocity, Volume, Variety, Veracity, and Value" that is causing difficulty for it to be efficiently processed, stored and analyzed by traditional methods and technologies [3]. This data is too big, moves too fast and, as a result, exceeds the processing capacity of conventional data processing systems.

Generally, much of this data is unstructured as it comes in form of images, documents and other undefined data model that does not reside in conventional databases and is not readily searchable by direct algorithms or machine analysis causing business today drowned with data. To make the matter worst, even if they could collect broad and various input, the process of exploring and inquiring the output insight of it would be a very time-consuming and costly process. However, rapid data analytic tool has helped a lot with this matter and since then, most organization have opted for big data projects as a game changer.

For example, The Security Exchange Commission (SEC) is using big data to monitor financial market activity using network analytics and natural language processor that helps in detecting fraud and illegal trading activity [4]. Both the health sectors and policy makers have also been investing in data mining projects to unearth trends and inclination in the population with a mission to provide the right healthcare at the right times and to help provide better support for public safety. Not ending on that, utility providers such as telecommunication and energy companies are also gearing up their Big Data Analytic game to offer tailored service to fit the customer satisfaction. Data analytics helps by segmenting the market according to the demographic, preference and usage as well as providing accurate solution service to user.

Despite opening doors for cyber challenges, Big Data offers many interesting patterns and behaviors secluded deep within the massive collections of information and of course, this does not just caught the interest of white bodies but also includes cyber-attackers. Various approaches are being used by both inside and outside attacker to gain personal benefit by gaining and make use of the sensitive and confidential information through passive attack or attempt to modify or paralyze the operation of a company using an active attack.

For example, Ukrainian electrical failure that come about in December 2015 had left around 230,000 people in the west of the country without power for hours, after a Supervisory Control and Data Acquisition (SCADA) cyber- attack using spear-phishing emails with infected email attachments [5]. There are also attack where hackers gain access to the core command-and-control system by using a cellular modem to imitate mobile tower allowing the fake tower to send a command to the end devices, such attack happened in Rye Brook, New York Dam Attack [5].

The system vulnerabilities itself can also open up the door for attackers. In 2015 and 2016, the SWIFT global banking system was used to initiate a string of cyber-attacks, which resulted in millions of dollars being stolen. The attacker succeeded in gaining legitimate SWIFT credentials after they have gained access to the systems of banks due to vulnerabilities of the system [5]. These cases show that how in Big Data environment, it is important not only on collecting the information but also securing it and make use of it using data analytics to both gain insight and detect suspicious activity.

III. DATA SCIENCE

In 2009, Google's chief economist, Hal Varian said that 'Data science is the ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it [6]. It is a field of study about methods, process, and system to gain insight from both structured and unstructured data [7]. Variety of techniques and methods from a various range of knowledge from mathematics to computer science employed in order to drive out knowledge on the data are a subpart of this field. It can be concluded that data analysis techniques and methods are among the area in data science.

Among these data analysis techniques are machine learning and data mining methods. Data mining is a technique of analyzing data using statistics and another programming method to mine what had happened in a data. While machine learning uses data mining technique and another algorithm to form a model to automatically program the machine to observe and recognize the pattern of data and further act on the data. Machine learning is also one of the ways to implement artificial intelligent.

Additionally, a data scientist is recognized as one of the position business bodies require leveraging big data effectively. Other three includes business analysts, big data developers, and big data engineers.

IV. BIG DATA ANALYTIC

Big data analytics is the process of examining the big volume of data such as big data environment to uncover hidden patterns, unknown correlations, and other insights. It is a form of advanced analytics, which involves complex applications with elements such as predictive models, statistical algorithms and what-if analyses powered by high-performance analytics systems [8].

Driven by specialized analytics systems and software, data analytics technologies and techniques efficiently analyze data sets, as well as coming up with timely insights about them to help organizations make informed business decisions, contributing to various business benefit. These benefits include new revenue opportunities, more effective marketing, better customer service, improved operational efficiency and competitive advantages over rivals [8].

V. DATA SCIENCE, DATA ANALYTICS AND BIG DATA

Data science is a field of study just like other branches of science. While biology focuses on living things, data science is about studies of data; its attributes, the way it moves, the size, the behavior and techniques to be applied on it in order to understands it.

Back when big data is still far from being discovered, this studies main branch would be data structure, which focuses on tabulation and modelling of data to organize and store the data. Nowadays, aligned with how voluptuous and fast data has been generated, this branch of studies has spread wider to various other kind of fields. Data need not only have to be in cycle of 'produce-kept-taken out', but many companies and organizations have sought to be creative with it in order to satisfies industrial needs. Hence, the field evolve to satisfy this needs.

This requires these data scientist in this field to find ways to fine tune the data, to able data to explain itself, be relatable and have other underlying meanings within its pattern and even forecast yet exist data. Among techniques used-then-become specific fields are data extraction, data analysis, data visualization and many more. While this branch-studies covers varies of methods and processes to scan and filter through big chuck of data, they also come forward with new terms of the specific 'kind' of data that they acted on such as metadata and big data through these programs and practices.

In conclusion, big data is term to describe a type and behavior of data that is of big volume and velocity and data analysis is used to scan and uncover hidden knowledge within this vast data in field of studies known as data science.

Fig.1 shows the illustration of the differences between the terms Data Science, Data Analytics and Big Data.

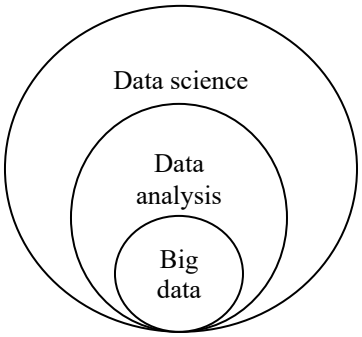

Fig. 1: Differences between Data Science, Data Analytics and Big Data

VI. DATA ANALYSIS TECHNIQUE

Normally, an organization's big data analytics need to include solving issues of insufficient memory to hold a

large volume of data and not having enough processing power causing latency in achieving the result. Hence, most of the organization would turn to distributed NoSQL databases as well as high-end data processing tools such as MapReduce and Spark.

The most famous data big data implementation framework would be Hadoop, where it offers Hadoop Distributed File System (HDFS) for storage and Hadoop MapReduce for processing [9]. Another pillar for Hadoop framework would be Yet Another Resources Negotiator (YARN). The fundamental idea of YARN is to split-up the functionalities of resources management and job scheduling/monitoring into separate daemons [9]. This allows multiple and batch data processing to handle data stored in a single platform interactively in real-time manner. This feature is added in second generation Hadoop as it is so well accepted even for large-scale, massively parallel and distributed data processing, that a whole ecosystem of software got to build around or can work with it such as Pig, Hive, HBase, Flume and many more for specific need of application [9]. Processing engine and big data tools like Storm and Spark that are well-known in streaming and real analysis can also integrate with Hadoop via YARN [9].

The organized, configured and partitioned data would then be analyzed with the custom-made program or off-the-shelf tools and software commonly used to do advanced analytics processes. While there are a lots of jargons used in data analytics such as data mining, machine learning, deep learning and many more; they are actually made up off statistical, logic, and algorithms equation that collectively function to analyze large data sets and uncover patterns and relationships which could later be used to either classify, forecast or predict next occurrence and way to counter it. There is also some analyzing tool that can be integrated with Hadoop such as Weka.

For a better accurate and customized ETL (Extract, Transform and Load) and analytics applications, queries can be written in batch-mode MapReduce or programming languages, such as R, Python, and Scala as well as SQL, the standard language for relational databases that are supported via SQL-on-Hadoop technologies. Algorithm and method used in data mining /machine learning techniques can be classified into two broad categories that are supervised learning and unsupervised learning.

Supervised learning is where the machine is required to observe a set of training data and compares, groups or identifies the new data according to the training data sets. Classification come under supervised learning. As an example, the C4.5 method constructs a classifier in the form of a decision tree. Suppose we have a group of students with related attributes such as marks for every subjects, interest, and skills and we want to predict which students will pass or fail the test. In order to classify students membership in these two class, C4.5 does not learn by itself, but we must tell it what each student class is. C4.5 will then proceeds to generate decision tree using attributes of students' of the corresponding class. Hence, the new students' class will be evaluated using the decision tree.

Meanwhile, unsupervised learning is when the machine is not given any training data to analyze with; so it teaches itself how to identify and manage the related data sets. For example, this method can compare data attributes with conditions repeatedly until it manages to develop a pattern and suitable data set. Clustering also comes under unsupervised learning, such as K-means clustering/grouping algorithm. Suppose we have a group of student with various vectors representing attributes such as age, address, marks for every subjects, interest, and skills. K-means will interpret this list as number coordinates in multi-dimensional space. Each vector will represent dimension and K-means will take the number of the cluster that the user wants and pick points in multi-dimensional space to represent each of the K clusters. These are called centroids. The students will cluster around their closest centroids to form K clusters. K means then find the center of each K cluster using the students' vector. As a result of changing the centroids, the student will also move and become the member of another cluster. This process is repeated until no other changing of membership occurs.

To make it simpler, data analytics, data mining and machine learning actually come down to statistical mathematics like C4.5, K-means, Support vector machines, *k*-NN, Naive Bayes and others as a basis, and how it is used to eventually coming out with analysis report for further action.

VII. DESCRIPTIVE, PREDICTIVE AND PRESCRIPTIVE ANALYTICS

Algorithm or combinations of algorithms and ways they are used to analyze data defines what type of data analytic in a data analysis model. There is three types of data analysis that will be further discussed in this paper; i.e. descriptive analytics, predictive analytics, and prescriptive analytics.

Descriptive analytics works as events counter that give overview of the data and what data by surface can describe about itself that changes with every occurrence of new data. With this technique, past and current events are described. It answers question such as "What has happened?" and "What is currently happening?" [10]. The example of this kind of analytic is to categorize utility customer by their usage or to select a user with highest consumption.

Predictive analytics did not really 'predict' the future event but more of forecasting the data that you do not have by using in-hands data. With this technique, probability of something happening or recurring in the future are predicted. It answers question such as "What will happen?" and "Why it will happen?" [11]. The example of this kind of analytic is a system to predict next utility bill for a specific user.

On the other hand, prescriptive analytics itself is a predictive analytic model but with two additional component that are actionable data and a feedback system

that tracks the outcome produced by the action taken. With this technique it suggest ways to cater the predicted events and stimulates it outputs. It answers question such as "What shall we do?" and "Why shall we do it?" [12] . The example of this kind of analytic is to predict when the high and low peak utility usage is and make preparation to fulfill the demand.

VIII. DIGITAL FORENSICS EVIDENCE EXTRACTION IN BIG DATA

A digital forensic investigation is an inquiry into the unfamiliar or questionable activities in the Cyberspace or digital world [7]. Computer forensic resolve around the literal meaning of the compound word that is the application of computer science technique and method to investigate the crime. With the increasing popularity of the usability of big data brings a new challenge in the forensic investigation as well as other information technology security fields [13].

This paper will try to analyst the past literature on how techniques and algorithms used in big data analysis can be adapted to extract evidence in digital forensic context. The main security domain will be highlighted for each literature. The literature was selected based on topics of intrusion detection related to big data.

Four security domain will be used to categorize each of the past extraction methods in big data extraction to get clear direction and will be used in the future development of the big data evidence extraction in the digital forensic framework.

A. Performance

Performance in evidence extraction is one of the big challenges in the big data environment. Theoretically [14] has stated that with the use of the computer in analyzing digital evidence will help forensic investigator in producing a faster result. However, the reality is the time to investigate a single hard drive to find relevance evidence can have the same duration in investigating an archive room that loaded with file and document because of human think time required by the investigator to analyze the whole evidence. It is important to have an extraction method that integrates the human thinking ability which can increase the ability to find evidence in the big data environment.

While it may be hard to reach out to all author of the papers that have been reviewed in order get the program and test it on same data to test performances of all the models, performance can be compared by reviewing the results and tools that have been used. Moreover, if the models are being compared with number of results produced in compare with times taken to produce it to determine which is of higher performance.

B. Accuracy

The accuracy of evidence produce in big data environment is one of the biggest concern in the digital forensic investigation as the false evidence after an investigation conducted can mislead the direction of crime investigation and false accusation.

In big data environment, the evidence can come from much different data formats and the ability of even a beginner criminal to alter, manipulate and destroyed the evidence in a manner that even forensic analysis tool cannot be taken lightly. The process of digital investigation that includes preservation phases makes thing more challenging as the integrity of the evidence need to be preserved for a certain duration before it will be presented in the court of law [15]. Hence, accuracy would best be compared by accuracy result of individual model against the number of test data used.

C. Speed

According to [15] the current forensic model still needs improvement and many researchers have proposed new process model which can cope with the demand of the new technologies. The increase popularity cloud technologies and big data have opened a requirement in digital forensic investigation process as the increased volume of data storage will require techniques that can allow the forensic investigator to meet the investigation time.

However the increase in volume of data storage is posing a new hurdle in increasing the analysis speed in which the processing load has become too high for the central processing units (CPUs) to handle, [16] have proposed the use of FPGA boards, parallel or pipeline processing in handling huge chunks of data to increase the analysis speed in a timely manner. Therefore, speed comparison can be used to compare only the time taken to get the result against data tested.

D. Cost

The cost has a broader aspect from another parameter. According to [17] cost in technology means the "Expenditure associated with acquisition or development, implementation, deployment, and maintenance of technology assets, including depreciation of R&D equipment and amortization of know how".

This definition precisely explain that the term cost of a technology application is not only about the financial cost of hardware and software but also cover execution cost of the system, whether it uses the resources (memory, processor, battery) efficiently, time spent developing the new technology / upgrading current technology, its reusability, its future cost such in maintenance and many more.

The same goes to the field of digital forensic, even more specifically, because in this field, the cost in the context of efficient resource usage is very crucial due to some large volume of sacred volatile evidence to analyze. Moreover in big data environment such as AMI where a high volume of data generated at a time. It would also smart to analyze cost as in financial expenditure. This is because it would be a foolish act to spend more to capture the evidence and catch the perpetrator than the actual damage amount. Hence, the cost is a very important factor indeed.

In a big data environment of a digital forensic investigation, the cost can be minimized with the use of open source framework such as Hadoop. Taking advantage of existing BI (Business Intelligence) tool against Hadoop,

and compressing data to the most granular level which not only reduce storage requirements but also drives down the number of nodes and simplifies the data analysis infrastructure.

While cost is very subjective, we choose model which is implemented in database and not needing extra requirements such as special machine or software license to be of the best cost.

IX. LITERATURE REVIEWS OF TECHNIQUES USED FOR EVIDENCE EXTRACTION IN BIG DATA

While there are a variety of other tools and technique developed to extract evidence in big data by combining the algorithm and methods, we review here some of them: Sung-Hwan Ahn [18] uses Big Data Analysis Model where the entire system is divided into 4 steps. 'Data collection' step, 'Data Processing' step, 'Data Analysis' step and 'Result' step. Collected data from anti-virus, database, network device, and the system will be processed using modern distributed database framework like NoSQL and Hadoop MapReduce to allow parallel processing. MapReduce consists of two jobs that are 'Map' job and 'Reduce' job. 'Map' job will take the dataset and modify it into a pair of key and value. After that, 'Reduce' job will group the output into sets of the pair with the same key.

Data from the previous step will then be analyzed using either or combination of 4 techniques of data mining that are a prediction, classification, relation rule, atypical data-mining to detect unknown new attacks. Regressing analysis technique is used to predict allied routines from collected attack logs. Logistic regression analysis and SVM (Support Vector Machine) are used in the classification of data according to the context of intrusion detection. Moreover, relation rule is applied to determine abnormal behavior by analyzing user or process behaviors. Lastly, typical data-mining techniques such as text-mining, web-mining, and social-mining are used analyses data that cannot be expressed in numbers [18].

From these model, we have deduced that it focuses on "performance" and "accuracy", as it performs continuous monitoring and detects previously unknown attacks that cannot be detected or mitigated using existing pattern matching methods such as signature, rule, and blacklist based solutions. But is never applied to the actual environment due to lack of professional software and distributed system. The downside of this model is that not an effective parallel processing algorithm for real-time analysis.

Shengyi Pan [19] uses the common paths mining-based IDS that provides stateful monitoring of an electric transmission distance protection system by leveraging a fusion of synchrophasor data and information from the relay, network security logs, and EMS logs to classify power system faults and cyber-attacks. Common path mining algorithm builds common paths from captured data logs. The critical states that describe a specific cyber-attack power system event are ordered and listed as a common path. From reading this paper, the strength of this technique is on "speed" and "cost" as it suits the scale of a power system by having the ability to process data as a stream, therefore minimizing the cost and time needed to analyze high volume data [19].

Priyanka Salunkhe [20] suggest the use of Decision Tree to works as a classifier to analyze where an attack happens and also the type of an attack using browser history as input. Decision tree represents a specific set of association rules regarding each node which will help to classify the dataset. The benefit of using decision tree is that it can cater any amount of data as the tree size is expandable and have the capacity to handle a huge amount of data [20]. Hence, we conclude that this model pillar parameters are "efficiency", "cost" and "speed" as Decision Tree can allow the system to the quick, easy and inexpensive analysis of log data. The system proposed in this paper finds out victim nodes or recognizes the pattern of an attack after collecting log files as an input and analyzing collected data. The evidence is stored in the database when a crime occurred and similar attacks will be identified using mining technique like k-means to separate attacker and normal user as well as a decision tree to match attacker data pattern with a pattern in training dataset to detect a type of an attack.

Rachana Sharma [3] proposed exercising two Machine Learning classification algorithms, Naïve Bayes and K Nearest Neighbor for classifying attacks which do not match the patterns of normal activities that are not offered in pattern-matching 'Signature-based IDS'. It was built on a Hadoop cluster using the Apache MapReduce programming model.

MapReduce based Naive Bayes algorithm means for binary or multi-class classification. It is simple yet useful for the large data set. 'Bayes Law' that describe the probability of something based on prior knowledge of conditions used as a base for this algorithm. To make it clearer, let say cancer is assumed to be related to age, hence the probability of an aged person to get cancer would be high.

MapReduce based K Nearest Neighbor is a classification algorithm which classifies an object on the bases of similarity matrices with objects of the known class. The objective of this algorithm is the minimization of the mean squared Euclidean distance of the data from the center of their clusters. MapReduce Programming model is also being compared with WEKA with both are using the same datasets (NSL-KDD) where it allows processing of large dataset in parallel and distributed manner [3], hence faster processing speed.

From the readings, we come out with the conclusion that these model parameters are "speed" and "performance" where C4.5 algorithm accelerates the construction of decision trees and also ensures the accuracy of classification. This is then further improved with MapReduce Hadoop parallel and distribution processing power.

R. Kumari [21] have a better approach of using Apache Spark with K-Means model to detect intrusion of the network by newer attacks that are not detected by supervised learning technique as it does not follow patterns of the previous ones. In this paper, it was used to detect anomalous network connection. KDD dataset are fed into Apache HDFS before network activities of similar types are group

into the good cluster and a bad cluster contains activities of dissimilar types. This is done by kmeans () function from spark's MLib [21].

After reviewing this paper, we come out with the conclusion that this K-Means Model can be used as classification technique to detect anomalies and its parameters are "efficiency", "speed" and "performance". This is proven as this model is further combined with spark making it of higher performance than of that other prototype and possible to detect anomalies in data arriving in real time thus building an actual deployable network intrusion detection application. Not ending on that, In comparison to Hadoop, spark uses multi-staged in-memory processing technique which obtains up to 100 times faster results [21].

Saad Mohamed Ali Mohamed Gadal [22] proposed a hybrid approach of k-means clustering algorithm and Sequential Minimal Optimization (SMO) implemented in WEKA to test the classification performance in detecting online network anomaly. The parameters that we draw from reading this paper is "accuracy" and this is proved in his paper by accuracy test, that when K-mean + SMO algorithm is applied using WEKA for NLS-KDD dataset with 22 attributes approach outperforms application of k-mean algorithm only and SMO algorithm only by a positive detection rate (94.48%) and reduce the false alarm rate (1.2%) and high accuracy (97.3695%) [22].

TABLE I. SUMMARY OF PAST LITERATURE FOR EVIDENCE EXTRACTION

No.	Author(s)	Technique(s)	Tool(s)	Parameter
1	Sung-Hwan Ahn, Nam-Uk Kim, Tai-Myoung Chung	SVM (Support Vector Machine), relation rule, atypical data-mining	Hadoop ,MapReduce	Performance & Accuracy
2	Shengyi Pan , Thomas Morris,Uttam Adhikari	Common path mining	real time digital simulator (RTDS) script	Speed & Cost
3	Ms.Priyanka Salunkhe, Mrs. Smita Bharne , Mrs.Puja Padiya	k-means and Decision Tree data mining	browser history (input),Database ,script	Efficiency, Cost & Speed
4	Rachana Sharma , Priyanka Sharma ,Preeti Mishra & Emmanuel S. Pilli	Naïve Bayes and K-Nearest Neighbor classifier	MapReduce framework and comparison with WEKA	Speed & Performance
5	R. Kumari, Sheetanshu ,M. K. Singh,R. Jha,N.K. Singh	K-Means model	Apache Spark , Spark MLib	Efficiency , Speed & Performance
6	Saad Mohamed Ali Mohamed Gadal ,Rania A. Mokhtar	k-means clustering algorithm and Sequential Minimal Optimization (SMO)	WEKA (using JAVA language)	Accuracy
7	Nguyen Thanh Van, Tran Ngoc Thinh, Le Thanh Sachearning	Deep learning using Restricted Boltzmann Machines (RBM) and AutoEncoder	Database ,script	Efficiency
8	Chetan R & Ashoka D.V.	k-means, support vector machine and association rule mining algorithm	Oracle database	Performance, Efficiency & Cost
9	Reza Adinehnia, Nur Izura Udzir, Lilly Suriani Affendey, Iskandar Ishak, Zurina Mohd. Hanapi	apriori algorithm, sequence mining	Java code, Relational Database	Performance & Accuracy

Nguyen Thanh Van [4] proposed Network Intrusion Detection system with the implementation of deep learning using Stacked Restricted Boltzmann Machines (RBM) and Stacked AutoEncoder to reduce feature or feature extraction in pre-training phase. Deep learning consists of multi-layer neural network that could be of "feed-forward neural networks" or "recurrent neural networks". Their biggest difference is that the first one with "feedback" from the outputs of the neurons towards the inputs throughout the network and the latter is not [23].Instead of classifying the object after a feature extraction, it learns feature based on locality and does classification together making it not only that can also be implemented supervised way, also an unsupervised way. In this paper, the strategy used is of layer-wise unsupervised training that preceded supervised fine-tuning allows efficient training of deep networks and gives promising results for many challenging learning problems, substantially improving upon the current [4]. Hence, the parameter that we conclude from this model is "efficiency".

Chetan R [24] propose DAID (Database-centric Architecture for Intrusion Detection) which is an evolution of relational database management system (RDBMS) that optimize the usage of database from being a centralized data repository, to place where all major operations such build, manage, deploy, score and analyses data mining-based intrusion detection models take place .This architecture is also said to be easier in generating analysis result and report as well as can be used for all types of databases. In this paper, a prototype was build that utilizes .NET to instrument the network activity reporting and analysis mechanism. The KDD'99 dataset is used and k-means, support vector machine and association rule are applied producing higher performance detection rate than other algorithm IDS such K-means + Naive Bayes [24].

As the model itself targets to the optimized existing database and leveraging it by building a high-performance IDS model that is easy and that can be used for all types of database, we deduced that these model parameters are, "performance", "efficiency" and "cost".

Reza Adinehnia [25] studied comparison of Apriori and algorithm and sequence mining that can be further used in the field of database intrusion detection systems. Apriori algorithm used to extract association rules which are divided into two parts, "if" and "then" continued with coming up of "support" and "confidence" ,which are derived from frequency that items appear in database and number of items the if/then has been found to be true to uncover precise patterns of relationships between unrelated data [25]. In

comparison, sequence mining is mining of the frequently ordered events or subsequences elements that come with or without timestamp with focus is on the order of accesses during the mining procedure. The result from the experiments conducted in this paper indicated that the parameters of this paper are "performance" and "accuracy" by showing that Apriority algorithm can provide more precise patterns, leading to higher detection rate [25]. Table 1 shows the summary of literature reviews of techniques used for extraction evidence in big data analytics.

X. DISCUSSION AND RESULT

Through this literature review of papers on intrusion detection and forensic investigation in the big data environment, few techniques and algorithms used in the big data analysis have been evaluated and parameters of each reviewed papers are identified.

While it is important to implement a technique that can produce an accurate result, there are also other studies that focus on other parameters such as developing a cost-saving technique or efficiently optimizing existing technology.

Hence, identifying the parameters leads to knowing a way to tackle problems and specified need in building our future work. From this study as well it has been clear that among combination that has been applied, a hybrid approach combining both supervised and unsupervised would be the better choice to be used in extracting evidence in Big Data. urthermore, Apache Spark is by far the most robust computing tool to implement the machine learning algorithm as it runs applications up to 100 times faster in memory and 10 times faster on disk than Hadoop. The same problem that keeps on recurring in the literature is the difficulties capturing logs for training data. Another is to figure out the type of input to be used. There are also not much of prototype being developed with some of the models even having difficulties to be implemented or having short-coming such as real-time analysis or cannot be run in distributed or parallel manner.

XI. CONCLUSION AND FUTURE WORK

In future, a frame of big data evidence extraction will be developed base on the parameter from Table 1, real-time Intrusion Detection system for the energy utility smart meter system will be developed using algorithms reviewed with focusing on accuracy and cost parameter.

ACKNOWLEDGEMENT

Work presented in this paper forms part of the research on Formulation of Evidence Source Extraction Framework for Big Data Digital Forensics Analysis in Advanced Metering Infrastructure, which was funded by Universiti Tenaga Nasional Start-Up Grant 2017.

REFERENCES

[1] C. Buckle, "Digital consumers own 3.64 connected devices," *GlobalWebIndex*, 2016. .

[2] R. Jacobson, "2.5 quintillion bytes of data created every day. How does CPG & Retail manage it?," 2013. .

[3] R. Sharma, P. Sharma, P. Mishra, and E. S. Pilli, "Towards MapReduce Based Classification approaches for Intrusion Detection," pp. 361–367, 2016.

[4] N. T. Van, T. N. Thinh, and L. T. Sach, "An anomaly-based network intrusion detection system using Deep learning," *2017 Int. Conf. Syst. Sci. Eng.*, pp. 210–214, 2017.

[5] T. Ball, "Top 5 critical infrastructure cyber attacks," *Comput. Bus. Rev.*, 2017.

[6] A. Guarino, "Digital Forensics as a Big Data Challenge," *ISSE 2013 Secur. Electron. Bus. Process.*, pp. 197–203, 2013.

[7] K. K. Sindhu, "Digital Forensics and Cyber Crime Datamining," *J. Inf. Secur.*, vol. 03, no. 03, pp. 196–201, 2012.

[8] M. Rouse, "big data analytics," *WhatIs.com*, 2017. .

[9] The Apache Software, "Welcome to Apache™ Hadoop™!," *Innovation*, 2012. .

[10] M. J. Mortenson and S. Robinson, "Operational research from taylorism to terabytes: a research agenda for the analytics age," *Eur. J. Oper. Res.*, vol. 241, no. 3, pp. 583–595, 2014.

[11] D. Delen and H. Dermikran, "Data, infomation and analytics as services," *Decis. Support Syst.*, vol. 55, no. 1, pp. 359–363, 2013.

[12] J. R. Evans and C. H. Linder, "Business analytics: the next frontier for decision sciences," *Decis. Line*, pp. 4–6, 2012.

[13] A. Irons and H. Lallie, "Digital Forensics to Intelligent Forensics," *Futur. Internet*, vol. 6, no. 3, pp. 584–596, 2014.

[14] V. Roussev and G. G. Richard III, "Breaking the Performance Wall: The Case for Distributed Digital Forensics," *Digit. Forensics Res. Work.*, no. November, pp. 1–16, 2004.

[15] C. Hosmer, "Proving the integrity of digital evidence with time," *Int. J. Digit. Evid.*, vol. 1, no. 1, pp. 1–7, 2002.

[16] "Accelerating Data Analytics Speed: New Technology Developed with Customers," *Ltd. Hitachi*, 2017. .

[17] BusinessDictionary.com, "technology cost," *technology cost. BusinessDictionary.com. WebFinance, Inc.* .

[18] S. H. Ahn, N. U. Kim, and T. M. Chung, "Big data analysis system concept for detecting unknown attacks," *Int. Conf. Adv. Commun. Technol. ICACT*, pp. 269–272, 2014.

[19] S. Pan, T. Morris, S. Member, U. Adhikari, and S. Member, "Developing a Hybrid Intrusion Detection System Using Data Mining for Power Systems," vol. 6, no. 6, pp. 3104–3113, 2015.

[20] M. P. Salunkhe, "Data Analysis of File Forensic Investigation," pp. 372–375, 2016.

[21] R. Kumari, "Anomaly Detection in Network Traffic using K- mean clustering," *2016 3rd Int. Conf. Recent Adv. Inf. Technol.*, pp. 387–393, 2016.

[22] S. Mohamed, A. Mohamed, and R. A. Mokhtar, "Anomaly Detection Approach using Hybrid Algorithm of Data Mining Technique," *2017 Int. Conf. Commun. Control. Comput. Electron. Eng.*, 2017.

[23] M. H. Sazli, "A brief review of feed-forward neural networks," no. January 2006, pp. 10–17, 2014.

[24] R. Chetan and D. V Ashoka, "Data mining based network intrusion detection system: A database centric approach," *2012 Int. Conf. Comput. Commun. Informatics*, pp. 1–6, 2012.

[25] R. Adinehnia, N. I. Udzir, L. S. Affendey, I. Ishak, and Z. M. Hanapi, "Effective Mining on Large Databases for Intrusion Detection," *Int. Symp. Biometrics Secur. Technol.*, vol. 3, no. 7, pp. 204–207, 2014.