Computational Intelligence in Intrusion Detection System for Snort log using Hadoop

C.Seelammal
Associate Professor, Department of Information Technology,
Sethu Institute of Technology,
Tamil Nadu, India
(seelammal@sethu.ac.in)

K.Vimala Devi
Professor, Department of Computer Science and Engineering,
Velammal Engineering College, Chennai,
Tamil Nadu, India
(k.vimaladevi@gmail.com)

***Abstract*— Today Cyber-attacks are increasing and the security technologies so far we used are in capable of handling the large data sets of data. Now a day the attacks motive has changed from simple hacking, damaging to large scale network and server system. In the presence of fresh and formerly unknown activities, intrusion detection rate is inaccurate and low. Many research works are focused in this area. In all the security infrastructures, Network Intrusion Detection Systems (NIDS) have grown a standard and in separable component. For this reason, a new model has been proposed based on Big Data for detecting unknown attacks. The core idea of this paper is to analyze the activity of the network users, and to classify whether the user is normal or anomaly. For implementation of this project, we have used the snort, the tools which are used to capture the online behavior of network users. After collecting the network behavior, the dataset was analyzed with hadoop framework using C4.5 classification. The performance of the project was compared with KDD Dataset. The proposed approach improves the IDS performance and its capability is to provide quick response to various types of network attacks.**

Keywords*— *Security, Big-Data Analysis, Hadoop, Snort, Machine Learning

I. INTRODUCTION

Fashions and Developments in technology, such as Internet of Things (IoT) devices, Cloud Infrastructures, Ecommerce and healthcare are allowing for unimaginable and massive access of unstructured data. Proficient learning from data always necessitates multipart architectures including mechanisms and techniques for collection of data, processes, and potential analysis [10]. The free access data security community is prolific, with the aim of many more decisions to be made for fruitful analysis. Progressively machine learning impressions are extensively embraced in all the multidisciplinary exploration [5], the need of facilitating the learning tasks is also more important.

IDS are commonly implemented using statistical, rule, state diagram and probabilistic techniques where detection of novel attacks is low. Securing the IT infrastructure from novel attack turns in to the major domain. Recently Network intrusion detection systems have grown a predominant section in all the security infrastructures. The attacks may be the following types in network intrusion detection [18]. In misuse detection, each occurrence is trained over known patterns then it is marked as normal and in anomaly detection creates models of usual activities, and automatically detects unusual activities from the normal, declining the unusual as anomaly.

Attacks are categorized in to four types based on their characteristics [20]. They are Denial of Service (DoS) , remote to user (R2L), User to root (U2R) and Probing. For DoS attacks, attacker utilizes the computing or memory resource, in turn will create the system too busy or too full to prevent from legitimate requests, A R2L gain the local user access by sending indefinite requests using the available network and exploits machine's vulnerability, U2R gain the root access and exploits root by sending request as normal user through the network and finally in probing, the attacker monitors a network to collect valuable information or discover known vulnerabilities.

A.Big Data analytics on security

Machine learning may be more suitable if the data at high volume and deeper analysis is required to make powerful decision making. Some of the promising area where the Machine Learning [10] plays vital role across industries are Ecommerce, Healthcare, financial services, energy conservation, feedstock and utilities. The following section highlights the changes in analytical landscape of Big Data from traditional analytics.

B. Traditional analytics and Big Data analytics -highlights

The traditional data processing and Big Data analytics is highlighted in the following way. First, the technological advances are based on memory usage, data processing, and scrutiny of Big Data include

- Handling high volume data with decreasing cost of storage and CPU power
- The effective usage of storage management for flexible computation and storage and
- Distributed computing systems through flexible parallel processing with the development of new frameworks such as Hadoop

These technological changes have shaped a number of improvements from conventional analytics and Big Data analytics. Second, Big Data frameworks for instance Hadoop

ecosystem and No SQL databases for speed up the processing and efficiently handles complex queries, analytics and finally Extract, Transform, and Load (ETL) which is complex in conventional data warehouses. With Big Data tools, users may provide structured and unstructured data in different formats, handles data at high volume, speed of data, finding uncertainty and incompleteness.

Hadoop [13] is a trendy and admired technology for batch processing. The two main requirements of Hadoop system is,

- Support partial failure - Failure of a particular component should not result in the loss of data. i.e. it must not cause the failure of the entire system only the performance may be degraded based on the nature of applications
- Scalable – Increase in resources should increase the capacity of load , Increasing the load on the system will have graceful decline in performance for all jobs

The Hadoop framework with Hadoop Distributed File System (HDFS) allows the developers for keeping huge data and the Map Reduce training model for both distributed and parallelized that is adapted for often occurring data processing problems.

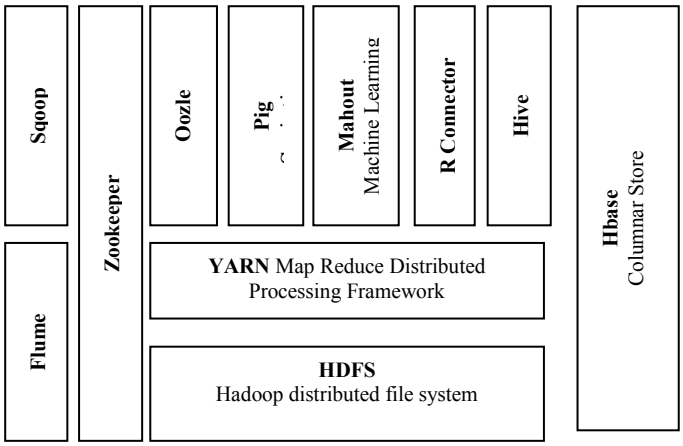

Fig 1.Hadoop Eco System

The common organization of the Hadoop ecosystem (shown in figure 1) contains three layers: storage layer, processing layer, and management layer. The storage layer is the lowermost level it includes the HDFS the core unit of the Hadoop. The processing layer is the place where the concrete analysis will be taken. The basis of this is YARN that is used to permit more processing engines to run on a Hadoop cluster. The management layer provides tools for interacting with user and complex organization. These include scheduling, monitoring, coordination, and user interface.

In research background [14], the selection of machine learning methods or precise algorithms will be based on variety of factors; typically it all depends on the requirements of the particular project and development team endorsed. The prioritization of these factors will be dependent on Scalability, Speed, Coverage, Usability and Extensibility. Machine Learning is wisdom from available occurrences or instances to formulate concrete predictions for future directions. This is fully designed on making models from data instances for powerful decision making. The machine learning models may be of these like:

- Supervised learning (To forecast an effect by training the already available data and testing data)
- Unsupervised learning (Finding useful information from hidden patterns , computing correlations and similarities from raw data)
- Semi-supervised Learning (Combining the fore said learning models for well behaved predictions)
- Reinforcement learning (How to maximize the effectiveness of analyzing by trying different actions to maximize the reward)

All machine learning projects will keep track of data capturing, cleansing, processing and analysis apart from concrete learning models and procedures used to resolve. Classic machine learning are performed the following they are featurization, training and Model Evaluation. Figure 1 below explains the progression of a machine learning solution.

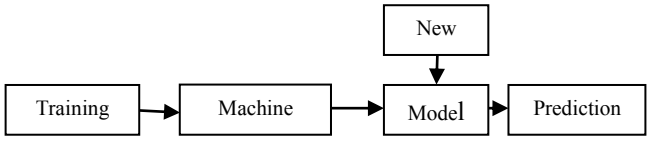

Fig 2. Typical Machine Learning Solution

A lot of tools [13] may be used by analysts to handle difficult queries and execute machine learning algorithms on Hadoop. It also includes Pig (scripting language for complex queries), Hive (SQL-friendly query language), and Mahout or RHadoop (machine learning algorithms for Hadoop). Recent frameworks like Spark 4 were designed to increase the effectiveness of machine learning algorithms that frequently reprocess a functioning of data by improving the effectiveness of complex data analytics algorithms.

C. Big data Challenges

- Automatic learning and handling from dynamic environments for feature selection
- Utilization of Big Data technologies to the challenges of Intrusion Detection.
- Premature early warnings for Intrusion systems.
- Evaluating a system for normal behavior and alerts for anomaly activities.

In this work, we have developed an IDS based on anomaly detection using big data analytics with hadoop framework. The following section of this paper is ordered as follows related works which dealt about IDS and big data analytics,

Intrusion detection data set challenges, proposed system, Experimental results and discussions. We perform experiments on snort and 99KDDcup dataset [19]. System experimental analysis was conducted using the measures execution time and speed up of the system in Hadoop framework.

II RELATED WORKS

The growth of computer networks from LAN to cloud systems always keeps Intrusion Detection Systems (IDS) as a critical factor in the field of security infrastructure. Researchers developed various cyber security technologies like firewalls and IDS to protect the system from threats and attacks. Firewall is acting as a barrier between a trusted network and other untrusted networks by controlling incoming and outgoing network traffic based on a set of rules. An intrusion detection system (IDS) monitors each and every network instances and finds any suspicious patterns to indicate that a network or system is compromised by someone.

Network-based anomaly detection systems play gradually more critical moves in recent security systems. Anomaly detection are being developed for identifying attacks in computer networks [3], malicious and misuse activities in web systems [7]. Further modern Anomaly Detection Systems are implemented using machine learning techniques similar to artificial neural-networks [20], Kohonen's SOMs [15] fuzzy rules [18] and others [14]. IDS become accepted by all the researchers because of their high detection accuracies even with low false alarms. Network security by Data mining [2] engages a variety of search algorithms, statistical and deviation analysis, rule induction, neural abduction, making associations, correlations, and clustering.

Wei- Yu Chen et al. [10] proposed a hadoop based analysis system for Snort log. They create a map reduce to improve the performance of Snort. X.Fang [3] planned to integrate artificial intelligence using Elmann Neural Network to snort. The integrated methodology flags the normal and malicious activities. B.E. Lavender [8] has tried a research work for implementing Genetic Algorithms to Snort.

From recent years, IDS have developed from LAN to High level IOT and large scale cloud to optimize their revealing capabilities. Latest environments are collecting information high pool of users like telecommunication, financial services army applications and online services and doing the decision analysis from centralized sometimes decentralized. This scenario has been addressed with a lot of limiting factors and designed only for compromised small scale that may not be applicable for large scale implementations.

Both industry and academia has been deeply engaged with intrusion Detection, Security agencies and Cyber security forums [4],[12] requires more threat analysis to be done and high accuracy alert are required to secure their systems with the trendy attacks. Big Data Analytics will extensively improve the limiting factors such as data preprocessing, processing like clustering, classification and by facilitating them to detect abuse activities which are moving in large scale security issues and trendy attacks. Although big data have considerable promise, there are tremendous and prospective challenges and can be overcome to recognize its true potential.

III INTRUSION DETECTION DATA SET CHALLENGES

Researches in intrusion detecting based on machine learning and big data always paying attention on data processing for large volume and optimization. Because of wide usage of internet, Cloud infrastructures, Wireless technologies and Internet of things, network users now create 2.5 quintillion bytes of data per day.

Mostly unobserved and skipped area of the IDS is the strong basis of preprocessing mechanisms instead of development of techniques for the analytics. The raw data collected from sources are not suitable to process for many machine learning and this may lead to inaccurate predictions. The quality of data makes the predictions more accurate otherwise abuse activities may be missed or deviated to massive impacts over the internet and security frameworks. According to the size of the data, the data is not enough the predictions may be wrong and it is too big the learning may feel hurdle to handle. The following are the some of the pinpoints based on the nature of the information

- *Mislabeled data* – From Data capturing to decision making the instances and mislabeled data is proportionally increases. When working with billions of instances it is unable to handle all the training instances are arranged meticulously and handled properly which may be efficiently handled before processing to make fruitful decision making

- *Missing values* - related to mislabeled data, in the absence of values the models may be less robust, inaccurate models. Because all unsupervised learning are based on similarity and correlation [13]

- *Noise* -Noisy data means that the data i.e. irrelevant or meaningless. These issues will take into over fitting. Even though tools and techniques were helpful to handle this models will lead to predictions inappropriate [13].

- *High dimensionality* –If the Feature to instance ratio is high and this leads to implicit nature of big data. Techniques for dimensionality reduction, feature selection and extraction better than Principal Component Analysis (PCA) may be used for resolving the high-dimensionality [10]

- *Imbalance* - In supervised and unsupervised learning, unfair training data can guide to pathetic learners. Data sampling , forced conversions of data and normalization techniques are some of the causes [10]

A numerous research has recommended with variety of learning algorithms could not manage the foresaid issues. 90% of the datasets and novel attacks have been created recently and rate of data creation has also been increased day by day. This acceleration has created a need for new technologies to analyze massive data sets.

The primary step in selecting big data frameworks whether the data is scalable and heterogeneous in nature, processing requires more computation and optimization, efficient storage management and finding the complexities decision making. Three-dimensions of data handling in big data are characterized by 3V's.

- *Volume* is the ratio between features and instances the data. The enormous usage of IoT, ecommerce healthcare, medical diagnosis, image analysis, army applications, ecommerce and financial are facing the complexities in processing and analyzing.

- *Velocity* pinpoints the speed of data is able to be receive and analyzed. Real time data may be stream or batch processing or the rate of receiving may vary from time to time.

- *Variety* identifies the dissimilar and mismatched data formats. Data may arrive from heterogeneous sources and several forms, and immediately ready in considerable amount of time for analyzing the decisions.

Big Data refers to managing the data at large and significantly enhance the data analytics of conventional data processing and in turn it will give a boom to data mining and machine learning.

IV PROPOSED SYSTEM

This section briefs the collection of dataset using snort and classification algorithm C4.5 that is used in the IDS for building decision trees. Due to increase in number of sophisticated targeted threats and fast development in data, the analysis of data become too difficult and security of that data is limited using existing security technology.
Recent unknown attack simply bypasses earlier security system by using cryptographic and expert systems. Consequently a trendy detection technique for resolving to this novel attack is needed. For defending this attack we use big data analysis for analyzing dataset.
Big data analysis framework based on hadoop for dealing the targeted attack using classification algorithm will surely enhance the detection capabilities of the anomaly detection accuracy.

Collection of data set

The raw data is collected from transactional database of any health care, ecommerce, telecommunication, marketing, economics, and inventory control etc. Data collection [4], [21] step collects data from logs, behaviors from network users, flag information from antivirus, catalog, smart devices and system. Snort is a famous tool for Intrusion Detection System (IDS), which is used to gather and analyze network packet in order to decide attacks through network. We generated packet dump file based on network packet dump files [16]. The files are composed of 35 files and total size of the dump files is about 2.22Gbytes. In order to handle the huge amount of network messages, we have formed cluster node with hadoop framework.

Anomaly Detection with C4.5 Decision Tree

Classification [17] is a technique that analyses the novel attack from huge volume of data. Classification helps security administrator to decide direction of protection and analysis. Most used classification techniques are decision tree.

C4.5 Algorithm: Machine learning algorithms are used for developing classification models, from which decision tree [3] is found to be the suitable classifier for intrusion classification. The detailed step by step procedure of the decision tree is specified below. The researcher may also refer [3], [9] to become an expertise in this field.

Algorithm: Create_ decision Tree
Input: Training instances may be discrete / continuous attributes; the set of class attributes.
Output: Decision tree

Procedure:

01. Selection of an attribute to test at each node - choosing the most useful attribute by calculating Information gain
02. Information gain measures the expected reduction in entropy, or uncertainty
03. Similarly, we can compute the information gain for other attributes.
04. At each node, choose the attribute with the largest information gain.
05. Measures(02) defines how well a given attribute separates the training examples according to their target classification
06. Measure is used to select among the candidate attributes at each step while growing the tree
07. Entropy - A measure of homogeneity of the set of examples.
08. Given a set S of positive and negative examples of some target concept (a 2-class problem), the entropy of set S relative to this binary classification
09. The entropy is 0 if the outcome is ``certain''.
10. The entropy is maximum if we have no knowledge of the system (or any outcome is equally possible).

The gain ratio, defines the proportion of useful information generated by split, i.e., that categorizes the attribute which is

helpful for classification at each level. For obtaining decision rules from the classification map reduce. The rolls of hadoop in analysis are as follows. In hadoop name node is the node where actual data to be stored, in data node the actual data processing will be done. Secondary node always maintains the meta data. Job trackers identify the input file and allocate the job to idle node to process and give the result. Task tracker tells the task which are running currently.

Map is written by the developer according to input pair and produces a set of intermediate key / value pairs. It also allows the jobs execution in parallel and this makes data processing speed of learning algorithms. Reducer collects all the intermediate results and gives meaningful set of rules with the available activities. These activities are the base for further analysis and predictions i.e. the known and unknown attack categories.

V.EXPERIMENTAL RESULTS AND DISCUSSIONS

Numerous platforms and machine learning libraries are available in Java to implement machine learning in big data. Hadoop is an open-source framework from Apache that allows manages and process big data in heterogeneous environment across data node from small training models.

Apache Mahout is an open source project that is mainly used for developing machine learning for large scale. Apache Mahout is a extremely scalable machine learning library that enables developers to use optimized algorithms. It implements popular machine learning techniques such as Recommendation, Classification and Clustering

Mahout - Classification

Classification is a machine learning technique that utilizes the existing signatures and trying to find any new findings and deviations from the usual behaviors. While classifying a given set of data, the classifier system performs the following actions:

- Initially a new data model is prepared using any of the learning algorithms.
- Then the prepared data model is tested.
- Thereafter, this data model is used to evaluate the new data and to determine its class.

The following steps are to be followed to implement Classification:

- Generate example data
- Create sequence files from data
- Convert sequence files to vectors
- Train the vectors
- Test the vectors

Snort IDS Alert Log was used for the implementation of this paper which is based on PCAP files with the IP addresses to hosts on the internal USMA network. Table I shows the improvement in execution time and speedup for various iteration in Hadoop with slave nodes .

```
[**] [129:4:1] TCP Timestamp is outside of PAWS window [**]
[Priority: 3]
11/08-09:45:54.394828 7.204.241.161:25 -> 10.1.60.203:50176
TCP TTL:64 TOS:0x0 ID:1283 IpLen:20 DgmLen:40 DF
*****R** Seq: 0xE06DEE6F  Ack: 0x0  Win: 0x0  TcpLen: 20

[**] [129:4:1] TCP Timestamp is outside of PAWS window [**]
[Priority: 3]
11/08-09:45:54.395078 7.204.241.161:25 -> 10.1.60.203:50176
TCP TTL:64 TOS:0x0 ID:1284 IpLen:20 DgmLen:40 DF
*****R** Seq: 0xE06DEE6F  Ack: 0x0  Win: 0x0  TcpLen: 20
```

TABLE I. COMPARISON OF EXECUTION TIME AND SPEED UP WITH HADOOP

Iteration	Training Instances	Execution Time(sec) without Hadoop	Execution Time(sec) with Hadoop	Speed up (%) without Hadoop	Speed up (%) based on Hadoop with single node
0	186010	1054	336	7.9	3.4
1	217860	1386	313	7.4	2.9
2	249710	1718	290	6.9	3.1
3	281560	2050	267	6.4	5.1
4	313410	2382	244	5.9	5.4
5	345260	2714	221	5.4	3.5
6	377110	3046	198	4.9	3.8
7	408960	3378	175	4.4	3.4
8	440810	3710	152	3.9	2.8
9	472660	4042	129	3.6	3.0

From the above table, it shows that Execution time increases if the no of training instances increased. When compared to hadoop framework, the analytics learns the behavior from the pattern and it gradually decreasing the execution time. Fig 1. Shows that rise of speed up is linear, i.e. that the proposed system is having the capability of data if it is highly scalable.

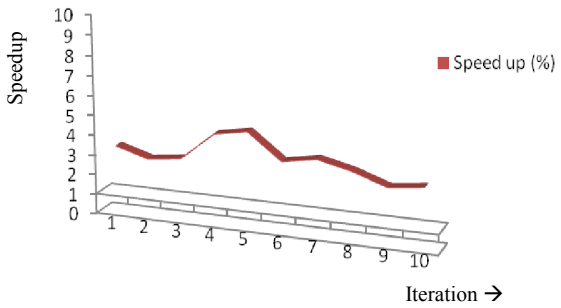

Fig 1. Improvement in Speed up with hadoop

Several efficient mechanisms [3], [5], [6], [9] are also available to detect anomalies, the advantage of the proposed system are

- The character attributes found in p data set are not converted into any numerical form instead it is used directly which reduces the data type conversion overload.
- Random sampling is used in classification from the snort log as well as in KDD dataset, where the characteristics of the high dimensional data are processed in the process.

- Missing values are effectively handled properly in classification and map reduces.
- Nearest consensus rule which is used in decision tree, which does the pruning.

VI. CONCLUSION

The recent industry approach is fully focused on real time detection and it unable to ideal feature set, individuality of a Attack detection and Prevention strategies. As elaborated, Big Data Analytics currently tackles a number of practical restrictions like selecting the number of data nodes according to the size of the data, effective writing practice of map reducer coding and advance research work is essential for building an operational solution. In our proposed work, we have developed IDS with Hadoop data platform in turn to analyze large volumes of network traffic.

We believe that big data have considerable promise, there are tremendous and prospective challenges and can be overcome to recognize its true potential.

a.

REFERENCES

[1] Chandola.V, Baneriee.A,Kumar.V ,"Anomaly Detection for Discrete Sequences: A Survey ", IEEE Transaction on Knowledge and Data Engineering, 24(2012), 823 – 839.

[2] Bando, M., Artan, N.S. , Chao, H.J.,"Scalable Look ahead Regular Expression Detection System for.Deep Packet Inspection", IEEE/ACM Transactions on Networking, 20(2012),699 - 714.

[3] X.Fang ,"Integrating Artificial Intelligence into Snort IDS", Proc of 3rd International Workshop on Intelligent Systems and Applications, May 2011,pp: 1- 4

[4] W.Chen, W.Kuo and Y.Wang, "Building IDS Log Analysis System on Novel Grid Computing Architecture" , National Center for High-Performance Computing, Taiwan,2009.

[5] Iftikhar Ahmad, Abdullah Alghamdi and Abdulaziz Alsadhan, "Multi-class DOS Attacks Classification in C4I Systems", Information an interdisciplinary journal,16(2013) 8853-8862.

[6] Snort Official Page, Online 2014, and Available: http://www.snort.org.

[7] Apache Hadoop 2.4.1, Online 2014, Available :http://hadoop.apache.org

[8] B. E.Lavender, "Implementation of Genetic Algorithm into a Network Intrusion Detection System (netGA) and Integrating to nProbe" , Thesis Work .

[9] Prathibha,"Design of a Hybrid Intrusion Detection System using Snort and Hadoop", International Journal of Computer Applications (0975 – 8887), Volume 73– No.10, July 2013.

[10] Shun-Fa Yang, Wei-Yu Chen, and Yao-Tsung Wang, "ICAS: An inter-VM IDS log cloud analysis system," IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS2011), 2011.

[11] Y. Lee, W. Kang, and Y. Lee, "Detecting DDoS Attacks with Hadoop", TMA, April 2011.

[12] Yeonhee Lee, Wonchul Kang, and Youngseok Lee, "A Hadoop-based packet trace processing tool-Traffic Monitoring and Analysis", LNCS 6613, Springer, 2011.

[13] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler, "The Hadoop Distributed File System," IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp.1-10, 2010.

[14] Sung -Hwan Ahn, Nam-Uk Kim and Tai-Myoung Chung, "Big Data Analysis for detecting unknown attack", IEEE/IFIP Network Operations and Management Symposium Workshops,2010 ,pp:357-361.

[15] Jay Beale, James C. Foster, Jeffrey Posluns, and Brian Caswell, Snort Intrusion Detection 2.0, Elsevier Inc., 2003.

[16] Toby Kohlenberg, Jay Beale, and Andrew R. Baker, Snort IDS and IPS Toolkit, Syngress Publishing, 2007.

[17] S.R.Gaddam, V.V. Phoha, and K.S. Balagani, KMeans+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods , IEEE Trans.Knowledge and Data Eng, 19(2007),345.

[18] R.P. Lippman, D.J. Fried, I. Graf, J. Haines, K. Kendall, D.McClung, D. Weber, S. Webster, D. Wyschogrod, R.K. Cunningham, and M.A. Zissman, "Evaluating Intrusion Detection Systems: The 1998 DARPA Off-Line Intrusion Detection Evaluation", Proc. DARPA Information Survivability Conf. and Exposition (DISCEX ″00), (2000),12-26.

[19] The third international knowledge discovery and data mining tools competition dataset KDD99-Cup, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999.

[20] Rong-Tai Liu, Nen-Fu Huang, Chia-Nan Kao, and Chih-Hao Chen, "A fast pattern matching algorithm for network processor-based intrusion detection system," 2004 IEEE International Conference on Performance, Computing, and Communications, pp. 271-275, 2004.

[21] Kakuru.s, "Behavior based network traffic analysis tool", IEEE International Conference on Communication Software and Networks (ICCSN), 2(2011), 649-652.

[22] Nassar M, al Bouna B, Malluhi Q: Secure outsourcing of network flow data analysis. In Big Data (BigData Congress), 2013 IEEE International Congress On. IEEE, Santa Clara, CA, USA; 2013:431–432.10.1109/BigData.Congress.2013.71

[23] Group BDW (2013) Big Data Analytics for Security Intelligence. Accessed 2015–1-10. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdfGoogle Scholar.