ISSN 2278-3091
Volume 9, No.1.4, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse6391.42020.pdf
https://doi.org/10.30534/ijatcse/2020/6391.42020

Big Data Analytic of Intrusion Detection System

Noor Suhana Sulaiman[1], Nur Sukinah Aziz[2], Nooraida Samsudin[3], Wan Ainul Alyani Wan Mohamed[4]
[1,2,3,4] Faculty of Computer, Media and Technology Management, TATI University College, 24000 Kemaman, Terengganu, Malaysia

ABSTRACT

Analysing network flows, logs, and system events has been used for intrusion detection. Network flows, logs, and system events, etc. generate Big Data. Big Data analytics can correlate multiple information sources into a coherent view, identify anomalies and suspicious activities, and finally achieve effective and efficient intrusion detection. This paper presents methods and subsequent evaluation criteria for network intrusion detection, stream data characteristics and stream processing systems, feature extraction and data reduction, conventional data mining and machine learning, deep learning, and Big Data analytics in network intrusion detection. Current challenges of these methods in intrusion detection are also introduced.

Key words : Data Analytic, Big Data, Intrusion Detection System.

1. INTRODUCTION

Big Data enables the analysis of data through a new generation of high speed, data capture, storage, and analysis technologies and architectures. Big Data requires enormous amount of storage space. As the collection of data continues to grow, companies are building large databases to store, compile and derive value from their data. Big Data supposed to be processed through a new generation of high speed, data capture, storage and analysis of technology and architecture. Big Data requires enormous amount of storage space. As the collection of data continues to grow, companies are building large databases to store, compile, and derive value from their data. Privacy, security, intellectual property and liability must be concerned and handled in Big Data processing. Big Data repositories allow companies to save money, increase revenue and accomplish many other goals by creating new applications, improving efficiency and reducing the cost of existing applications

However, due to their scale, range, and speed properties, data traffic through the network leads to Big Data problems. The use of large-scale cloud infrastructures, with the variety of software platforms distributed through large computer networks, further increases the entire system's attack surface. The major contribution of this article includes intrusion detection approaches, and especially the analytic approach to Big Data, various security threats, and Big Data techniques [1].

Section 2 gives an overview and efficiency metrics of IDS and Big Data and information on type and method of IDS implementation approach. Section 3 discusses source of IDS real data and overview of IDS benchmark dataset. In section 4 discusses the issues of IDS dataset in Big Data analytic approach due to many threats, have to analyze and process a massive amount of data. Recent intrusion detection techniques in Big Data are given in Section 5 in which discuss contribution regarding the method employed.

2. INTRUSION DETECTION SYSTEM

In computer security perspective, the detecting malicious activities or intrusions are crucial. An intrusion is different from the system's normal behavior, and thus anomaly detection techniques are important in the field of intrusion detection. There are three types of methods for Intrusion Detection System (IDS) including Signature-Based Detection (also called Misuse Detection), Anomaly-Based Detection, and Hybrid Intrusion Detection. The most widely used and reliable of these is intrusion detection based on signature [2]. After a new attack is initiated, the pattern or signature of the attack is specified which may be targeted resources during an attack, the way the resources are targeted in, or a name (in characters) within the attack code body. After the attack signature is identified, network security specialists will build a response against a new assault. The IDS is modified accordingly by the proposed defense to identify and respond to the new pattern of attack.

All method of IDS can be developed into three different types which are Network-Based Intrusion Detection System (NIDS), a Host-Based Intrusion Detection System (HIDS), and a Hybrid-Based Intrusion Detection System (Hybrid IDS). An HIDS detects malicious activity on a single computer, while an NIDS identifies intrusions by multiple hosts monitoring and network traffic analysis. In an NIDS, sensors are located at the network's choke points to track, often in the Demilitarized Zone (DMZ) or on network borders, and capture all network traffic. Hybrid IDSs detect intrusions by examining program logs, device calls, changes to the file-system (password files, binaries, access control lists, and capability databases) and other host states and activities [3]. IDS technologies such as HIDS, NIDS, Network

Behavior Anomaly Detection (NBAD) and Wireless Local Area Network (WLAN) IDS are used in tandem with each system to compare data and determine what these IDSs track. IDS technologies such as HIDS, NIDS, Network Behavior Anomaly Detection (NBAD) and Wireless Local Area Network (WLAN) IDS are used in tandem with each system to compare data and determine IDS's monitored.

In particular, the effectiveness of an IDS can be determined by the number of false positives and false negatives, discern in following cases:

i) True negative is a normal activity which correctly considered as such by the IDS.
ii) False positive is a normal activity which is wrongly considered an IDS attack.
iii) False negative is an undetected attack, which wrongly considered as a normal activity by the IDS.
iv) True positive tests an attack which correctly considered by the IDS.
v) Accuracy is a measurement of the percentage of failure and correct detection and the number of false alarms produced from the IDS.

True negatives and true positive ones actually correspond to the desired behaviors. An IDS is usually incomplete, however, and results in the presence of two other undesirable habits. The multiple IDS usually suffer from imperfections resulting in these undesirable behaviors detection. Other metrics and analyzes are more or less used in the context of IDS, such as; the ability to detect new attacks, the ability to detect an attack's success, and resistance to attacks, the ability to work continuously with minimal human interference, the ability to restore the previous state of failure and, ultimately, the interoperability with other computer security systems and instruments.

An anomaly detection program has a profile of the defense system's usual behavior patterns. A sequence of coming data is known as an attack because it's different from regular pattern. The anomaly detection approach can use unsupervised learning techniques to identify new emerging attacks without the need for marking patterns [4]. Anomaly detection techniques show good accuracy in detecting attacks at network level such as SYN surge, teardrop, and denial of service (DOS) but not in recognizing application level exploits such as Remote to Local (R2L) and User-to-Root (U2R). Anomaly detection schemes consider only the fields of the packet headers such as flags, port numbers and IP addresses and etc. Therefore, they operate well if an attack includes only the relevant fields at the network level. In next section, the source and data types of IDS are discussed.

3. INTRUSION DETECTION SYSTEM DATA SOURCE

HIDS inspect host and audit source data such as operating system, window server logs, firewall logs, computer system audits, or database logs. HIDS can detect attacks by insiders not involving network traffic [5]. NIDS tracks network traffic that is extracted from a network through packet capture, NetFlow, and other data sources on the network. It is possible to use network-based IDS to control several computers that are linked to a network. NIDS is capable of detecting any malicious activities which could be triggered at an earlier stage from an external threat, before the threats spread to another computer system. On the other hand, because of the volume of data flowing through modern high-speed communication networks, NIDSs have limited ability to examine all data in a high-bandwidth network [6]. Together with HIDS and firewalls, NIDS deployed at a number of positions within a specific network topology can provide a solid, resilient and multi-tier security from threats on both the outside and the inside.

KDD Cup 99 is a benchmark dataset for IDS. The KDD Cup 99 informational collection was utilized as a part of KDD CUP 99 Classifier Learning Competition. The first raw information preparing of 4 gigabytes of compacted parallel tcpdump information acquired from the initial 7 weeks of system movement at MIT. Information were preprocessed with the element development structure Mining Audit information for computerized models for Intrusion Detection (MADAM ID) to deliver around 4,898,431 and 311,029 records in set of preparing and record of testing, including 42 features as in Table 1. Dataset of highlights from arrange packet characterized into non attack and 4 attack classes. Information are named as attack or normal, and besides are named with attack sort assembled to 4 general classes of attack. Database contained a huge military network environment of intrusions simulation. The dataset consist 19.48% normal and 80.52% attack connections, consisted of 311,030 records, among which 60, 593 (19.48%) were 'normal', 229,853 (73.90%) DOS, 4,167 (1.34%) Probe, 16,347 (5.26%) R2L and 70 (0.02%) U2R attacks, show up in Table 2 below. The IDS data issue relating to Big Data is explained.

Table 1: KDD Cup 99 Features

No	Features	No	Features
1	duration	22	is_guest_login
2	protocol_type	23	count
3	service	24	srv_count
4	flag	25	serror_rate
5	src_bytes	26	srv_serror_rate
6	dst_bytes	27	rerror_rate
7	land	28	srv_rerror_rate
8	wrong_fragment	29	same_srv_rate
9	urgent	30	diff_srv_rate
10	hot	31	same_srv_rate
11	num_failed_logins	32	dst_host_count
12	logged_in	33	dst_host_srv_count
13	num_compromised	34	dst_host_same_srv_rate
14	root_shell	35	dst_host_diff_srv_rate
15	su_attempted	36	dst_host_same_src_port_rate
16	num_root	37	dst_host_srv_diff_host_rate
17	num_file_creations	38	dst_host_serror_rate
18	num_shells	39	dst_host_srv_serror_rate
19	num_access_files	40	dst_host_rerror_rate
20	num_outbound_cmds	41	dst_host_srv_rerror_rate
21	is_host_login	42	normal or attack

Table 2: Instance Types in Corrected KDD Dataset

Class	Number of instances	% of occurrence
Normal	60,593	19.48
Dos	2,29,853	73.90
Probe	4,167	1.34
U2R	70	0.02
R2L	16,347	5.26
Total	**3,11,030**	**100**

4. BIG DATA: INTRUSION DETECTION SYSTEM ISSUE

Many measures and analyzes are more or less used in the sense of IDS, such as: the ability to detect new attacks, the ability to detect an attack's performance, the degree of impermeability and resistance to attacks, the ability to work continuously and with minimal human interference, the ability to restore the pre-failure state and eventually the interoperability with other computer security systems and devices. The key challenge for detecting anomalies in this domain is the immense amount of data. To accommodate these large sized inputs, the anomaly detection techniques must be computationally efficient. In addition, the data typically comes in streaming format, which includes online analysis. Another issue that arises is the false alarm rate because of the large size input. Since the data amounts to millions of data items, for a researcher, a few per cent of false alarms will make research challenging. Another concern is false alarms because of the large input number. Labeling data is often available which correlates to normal behavior, but labels for intrusions often don't happen. Because input\data can contain millions of data objects, a small percentage of false alarms will overwhelm the analysis. Big Data technologies offer the ability to collect, store, process, and analyze data at large scales and at high speeds. The use of large data to detect intrusion has a lot of potential, but fundamental weaknesses in the detection of intrusion are the ability to enhance detection results by reducing the rate of false negatives and false positives.

This is why Big Data Applications are becoming part of the security management software and the Big Data Analytics for intrusion detection has gained more and more interest as it facilitates the analysis of vast volumes of data with different formats from heterogeneous sources and detects anomaly. Suit to this context, models can be generated to accurately profile data online, thus helping to predict and detect intrusions and attacks in real time through [7];

i) Capturing large scale data from numerous internal and external sources such as vulnerability databases
ii) Conducting deep analytics on the data
iii) Achieving real-time analysis of stream data
iv) Presenting an integrated view of security related information.

5. BIG DATA ANALYTIC IN INTRUSION DETECTION SYSTEM

Big Data defines the large volume of structured and/or unstructured data from a multitude of sources that can not be processed efficiently with the conventional applications that exists and that need more real time analysis [8]. However, with such information, it is not only the amount of data that is important but also what the organizations would do. The interpretation of Big Data is becoming more important nowadays and not limited to a single approach or instrument [1]. Big Data Analytics actually refers to a set of procedures and statistical models for extracting data from a wide variety of data sets. Moreover, Big Data also brings new ways to discover new values, helps us gain a thorough understanding of the secret values and also creates new challenges. However, redundant attributes and records make intrusion detection a particularly complicated and challenging task in Big Data analytics. Dimension reduction is highly important for both performance and efficiency, also helps to reduce anomaly detection computational complexity and improves better classifier performance [9].

Data mining methods such as clustering, classification, and association rule mining are often used by analyzing network data to obtain valuable information on network intrusion. Clustering can be used in both detection of misuse and detection of anomalies while classification is used primarily for detection of anomalies and is a supervised method of learning. Classification-based IDS can classify all network traffic into either malicious or normal. Association rule mining searches for frequently occurring items from a large dataset and identifies correlation or association rules between large dataset data items [10]. In particular, there is a difference between classification and clustering applications within IDS [11];

i) Clustering: In intrusion detection, clustering is useful because malicious activity should cluster together, separating itself from normal activity. Clustering benefit over classification approach is that it does not need to use classified data set.
ii) Classification: Classification-based IDS aims to identify the entire traffic as either malicious or natural. The question is how the number of false negatives and false positives can be minimized. Five general intrusion detection techniques including support vector machine (SVM), fuzzy logic, inductive rule generation, neural networks, and genetic algorithms were used to perform classification.

Big Data analytics for intrusion detection and prevention of network security issues has been rapidly attracting attention as it facilitates the analysis of large volumes of complex and fragmented data with different formats from heterogeneous sources, detects anomalies and battles cyber attacks. Ultra-high-dimensional data models can be created to accurately profile online stream data, thus helping to predict

and detect intrusion and attacks in real time [12]. Big Data technologies such as the Hadoop ecosystem and stream processing can store and analyze large heterogeneous data sets at high speeds, transforming safety analytics through;

i) Capture large-scale data from many internal and external sources, including vulnerability databases.
ii) Carrying out in-depth data analysis
iii) Realizing real-time stream data analysis
iv) Presenting an integrated view of information related to security.

In factor of Big Data analytics tools need to be properly configured, system analysts and architects need an intimate knowledge of their systems [13].

6. CONCLUSION

In a security perspective, Big Data's main concerns about outsourced data are safety, fairness, availability and confidentiality. As the use of Big Data has increased, the protection is therefore very critical, consideration is given to the intrusion which is an important feature for the implementation of Big Data environment detection systems. The field of machine learning involves building up a model from data by using an algorithm. At best this model is generalized to represent or approximate the data. This helps to predict unknown ones and to better understand existing ones according to the data given to it in the input. The application field of machine learning is very varied: financial value prediction, computer security intrusion detection, user profile-influenced search engine, data theft detection, anti-virus implementation, and cryptanalysis.

ACKNOWLEDGEMENT

This paper is based based on funded research by Short Term Grant Phase 1/2018, 9001-1801 under University College TATI (UC TATI).

REFERENCES

[1] M. Britel, “Big Data Analytic for Intrusion Detection System,” in *International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 2018.

[2] L. Wang and R. Jones, “Big Data Analytics for Network Intrusion Detection: A Survey,” *Int. J. Networks Commun.*, vol. 7, no. 1, pp. 24–31, 2017.

[3] B. M. Beigh and M. A. Peer, “Intrusion Detection and Prevention System: Classification and Quick Review, ARPN,” *J. Sci. Technol.*, vol. 2, no. 7, pp. 661–675, 2012.

[4] J. F. Nieves and Y. C. Jiao, “Data clustering for anomaly detection in network intrusion detection,” *Res. Alliance Math Sci.*, pp. 1–12, 2009.

[5] C. G and H. J, “A semantic approach to host-based intrusion detection systems using Contiguousand Discontiguous system call patterns,” *IEEE Trans. Comput.*, vol. 63, no. 4, pp. 807–819, 2014.

[6] B. MH, B. DK, and K. JK, “Network anomaly detection: methods, systems and tools.,” *IEEE Commun. Surv. Tutorials*, vol. 16, no. 1, pp. 303–336, 2014.

[7] “No Title.” [Online]. Available: http://article.sapub.org/10.5923.j.ijnc.20170701.03.html.

[8] C.-W. Tsai, C.-F. Lai, H. Chao, and A. V. Vasilakos, “Big Data Analytics: A Survey,” *Journal of Big Data*, 2015. https://doi.org/10.1186/s40537-015-0030-3

[9] B. Kicanaoglu, “Unsupervised Anomaly Detection in Unstructured Log-Data for Root-Cause-Analysis.,” Tampere University of Technology, 2015.

[10] M. K. Patond and P. Deshmukh, “Survey on Data Mining Techniques for Intrusion Detection System,” *Int. J. Res. Stud. Sci. Eng. Technol.*, vol. 1, no. 1, pp. 93–97, 2014.

[11] T. Lappas and K. Pelechrinis, “Data Mining Techniques for (Network) Intrusion Detection Systems,” 2010.

[12] L. Cui, R. F. Yu, and Q. Yan, “When Big Data Meets Software-Defined Networking: SDN for Big Data and Big Data for SDN,” *IEEE Netw.*, pp. 58–65, 2016.

[13] A. A. Cárdenas, P. K. Manadhata, and S. Rajan, “Big Data Analytics for Security Intelligence,” 2013.