International Journal of Advanced Innovative Technology in Engineering (IJAITE),
Vol.6, No.5, September 2021
E-ISSN: 2455-6491
Available online at www.ijaite.co.in

Intrusion Detection System Over Big Data

[1]Ashish Kokate, [2]Prof. M. K. Nichat

ABSTRACT: Network security is a paramount concern for the organization. To secure the network, we have traditional network intrusion detection systems and firewalls but they have limitations like size of training data sets .With the inception of Hadoop technology, in industry, recently researchers have started using this new technology with traditional machine learning algorithms which generally uses pattern matching, to design and develop network intrusion detection system based on streaming of big-data using Hadoop that checks for intrusions in massive amount of data that flows in and out. In this paper, we are presenting a study and analysis of various Hadoop based network intrusion systems. Here the parameters used for comparison are detection rate and false-positive alarm rate.

Keywords: Big data, Hadoop, Network Security, Intrusion Detection System, Machine Learning

I. INTRODUCTION

Security is a major concern now-a-days as technology has reached new heights. On personal level security measures includes only installation of anti-virus and firewall. But when an organization is concerned, the solution cannot be simple. There should be a dedicated security system as the risks are many in a business organization It is essential to have a safe and secure network for following reasons:

To protect organization's assets

To adhere with regulatory requirements and ethical responsibilities

A. For Competitive Edge

An Intrusion Detection System (IDS) is a security software that constantly monitors the network to look for suspicious or malicious activities and automatically alert the administrator. In other words, it is equivalent to a Burglar Alarm. The types of intrusion detection systems are: A host-based intrusion detection and a network-based intrusion detection.

Host based intrusion detection system: Host computers are installed with software and are used to analyze system activities, log files and all network-traffic received by host computers. It can discover whether an attack is successful and also record what the attacker has performed on the host. Network based intrusion detection system: Sensors are installed throughout the network to analyze all traffic on target network in real time. These sensors describe a network-interface that receives all network-traffic and matches the defined pattern. If the pattern matches the system alerts the administrator.
IDS Triggers: The main motive of IDS is to send an alert if an intrusion is detected, it works just like a burglar alarm.

IDS have two triggering mechanisms:

1. Misuse Detection (Signature based)

2. Anomaly Detection (profile based)

Signature based detection: This type of mechanism requires signature-based files i.e., known patterns of attacks. If a pattern is matched, there is a high probability of an attack, and therefore an alarm will be triggered. It has low false-positive alarm rate as matches are based on known patterns. Signature detection fails to detect attacks which are not known or variations in the known attacks.

Anomaly based detection: This type analyzes computer and network activities and looks for an anomaly, if an anomaly is found alarm is triggered. Anomaly is abnormal behavior of network or deviation from common rule for attacks. It has high false-positive alarm rate.

The rise in use of technology has led to increase in amount of network traffic data. The traffic data is expected to be in the range of Zettabytes and is significantly increasing from past few years. This data having high Volume, Velocity and Variety is often termed as Big-data. In this generation of big-data the IDS should be good enough to process huge volume of data in real time. To overcome this challenge and the need of IDS of high accuracy and efficiency while running in parallel environment has led to introduction of IDS using Hadoop that reduces the extra overhead and does not reduce the performance. The rest of the paper has following sections. Section 2 discusses existing work done in Hadoop based IDS Section 3 provides analysis of various IDS Algorithms. Section 4 provides a conclusion of the study with comparisons.

Manuscript Received August 5, 2021; Revised 25 August, 2021 and Published on September 2, 2021

Ashish Kokate, *PG, Scholar, Department of Computer Science & Engineering, Dr. Sau. Kamalatai Gawai institute of Engineering & Technology, Darapur, Maharashtra, India,*

Prof. M. K. Nichat, *Assistant Professor, Department of Computer Science & Engineering, Dr. Sau. Kamalatai Gawai institute of Engineering & Technology, Darapur, Maharashtra, India.*

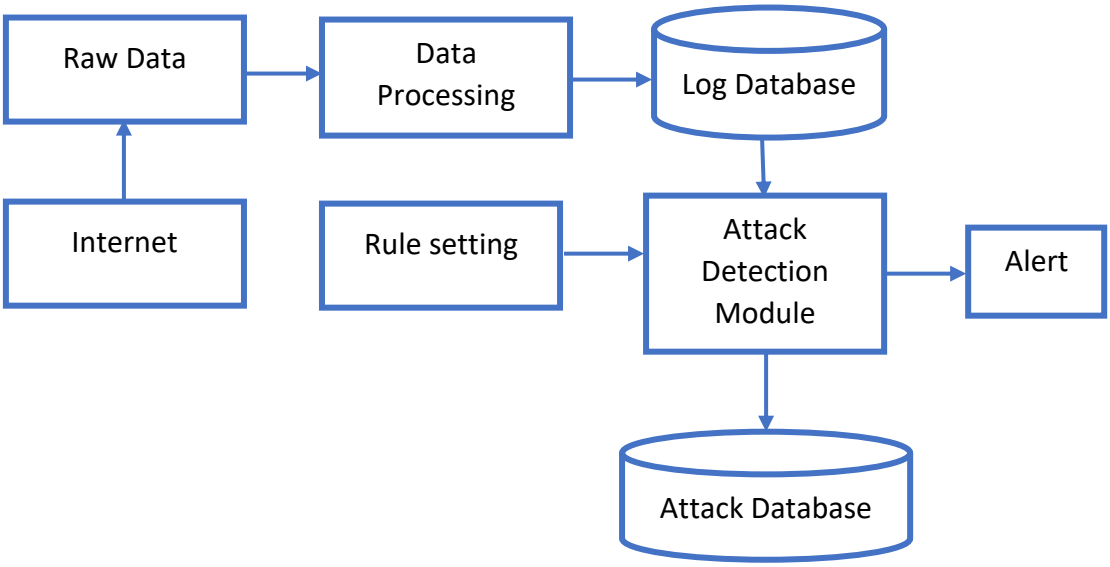

Figure 1: Working of Network Intrusion Detection System [9]

II. LITERATURE REVIEW

The work proposed by Zhiguo Shi et.al. [1] presented the system for Hadoop cluster along with improved k means algorithm, as the defined K means algorithm is only used for detecting single intrusion and not for multiple intrusion detection which results in forming a distributed IDS for manipulating big-data. They proposed the use of Map reduce to select the stable centres of the k means clusters and impose it to the Hadoop clusters to form distributed IDS. The limitation of their work is that the improved K means algorithm gives accuracy of 95% which is higher than traditional K means algorithm but is still lower than decision tree which has accuracy of 96%. Also, the use of Map reduce adds burden and can't be used for streaming data.

The Work proposed by M. Mazhar Rathore et.al. [2] presented model of Hadoop based real time intrusion detection for high-speed networks the main challenging problem is to identify intruders in high-speed networks. So, to solve this, they proposed a high-speed IDS which is capable of working in big-data. This paper makes use of KDD 99 dataset from which all parameters were selected, further using Forward Selection Ranking (FSR) and backward elimination ranking (BER) only 9 best features were used for achieving faster results. Among the widely used classification algorithms, REPTree and J48 Machine Learning classifiers were selected for classifying intrusion. The proposed work makes use of Map reduce programming using single node Hadoop architecture to accurately and efficiently detect threats and to get real time efficiency Apache Spark is used. Traffic is captured using capturing device like RF-RING and TNAPI which assures that no packet remains un captured. Filtration and load balancing server (FLBS) are used for searching and comparing in intruder's database. Main problem is with the data set i.e., KDD99 which is not advisable by authors itself to use it, as well as architecture uses single node so it becomes a single point of failure if that node fails the system may go down.

The work proposed by Manish Kumar et.al. [3] implemented the IDS log analysis using the cloud architecture by taking the help of Hadoop and Map reduce. They created the distributed system in order to extract data efficiently with the help of HDFS. Using Map Reduce they merged the two alerts and store in log files. The essential information from log files produced by IDS is extracted by the log parser which is mapped to different nodes depending on the extracted Meta information. Then the information is reduced and stored in log files for the further detection. Limitation of their work is that Map Reduce increases the burden and is not adaptive and their proposed work detects only signature-based intrusion.

The work Proposed by Sanraj Rajendra Bandre et.al. [4] presented the method which deals with NIDS based on Hadoop framework and GPGPU. Big-data from organisation is applied to Hadoop framework while

GPGPU is used for intrusion detection. Among the various Hadoop ecosystem like PIG, HIVE and HBASE, FLUME is used to pull collect real time streaming data. In propose system a parallel failure-less AHO-Corasik (PFAC) algorithm is used for multi string pattern matching algorithm and it removes failure transition from state transition machine. The problem with the proposed system is to keep on updating network security with modern technologies.

The work proposed by Basappa Kodada et.al. [5] implements a method to overcome hurdle of finding malicious activities in zettabytes of network packets. This paper evaluates the security architecture of huge amount of data that flows in and out. To process big-data, a Hadoop Map Reduce framework is used. An open-source packet analyzer called Wireshark is used for network trouble shooting, as well as it allows user to put network interface which supports promiscuous mode. Big-data analytics can be used to emphasize financial transactions, log files and network traffic to detect suspicious activities as well as identify anomalies. In the proposed work results show that analysis and computation is much faster than other systems. The limitation of their work is that Map Reduce degrades the performance of the system as it is complex and time consuming in nature. Increase in the number of nodes in their proposed system can achieve better performance and results.

The work proposed by Gurpreet Kaur Jangla et.al. [6] suggested the design to divide the entire IDS system into 4 parts. Firstly, the data is collected from various sources, then the data is processed and analyzed. If attack is occurred then the gets alarm. They propose the use of HDFS and Map reduce by mapping the (key, value) pair to process the data. Then the data is analyzed by using classification algorithm and prediction is done on the data. The limitation of their proposed work is although the data is processed and analyzed but Map reduce can't be used for streaming data and increases the complexity. Also, it can only predict the signature-based attack and still requires advancement to find advanced threats.

The work proposed by Sonali Ashok Hajare [7] presented the method for processing of unstructured data from socio-networking sites and converting them into structured data. Map reduce is applied on this data for faster retrieval. Pre-processing of web log data is done through SVM and c means clustering. If the pattern is matched then intrusion is detected. Map Reduce extract the features from the structured data set and the key, value pair is matched with the predefined signature already available. If matched with the pattern then the intrusion is detected. If any suspicious IP packet is detected then it gets cleaned using c-means clustering. The limitation of this work is Map reduce is a huge overhead on the system as it degrades the performance of the system. More efficient machine learning algorithms other than SVM and c-means clustering can be used to improve the system further.

The work done by Shaik Akbar et.al. [8] proposed an architecture in which KDD cup dataset is used to gather data which are not similar and these data are separated into learning and detection phase. Big-data along with its 5 V's Velocity, Volume, Variety, Value and Veracity and briefly discussed with their usage. A feature selection process is applied on KDD cup data set to select and extract main features. In learning phase known attacks like DOS, pros are detected while the attacks which are not been identified in learning phase, they are detected in detection phase. In detection phase enhanced C4.5 and enhanced genetically algorithm are used. These two phases store the data in large database forming an integrated hybrid large database. These hybrid techniques will enhance the detection rate by identifying the category of attack. The limitation of their work is due to advancement in technology the need to address real time data is essential.

Table 1: Comparison of various classification Algorithms [1], [2], [10].

Classification Algorithm	Accuracy (%)
K-Means	90
Improved K-Means	95
Rep Tree	72
J48	70
Support Vector Machine (SVM)	84
Decision Tree	96

III. ANALYSIS

In this section, we have analyzed various IDS algorithms with respect to detection rate, false alarm rate and data type of respective algorithm.

Table 2: Comparison of various IDS Algorithms with Detection Rate, False Alarm Rate and Data Type [11] [12] [13]

IDS Algorithm	Detection Rate	False Alarm Rate	Data Type
PCA, SVM, PSO	97.7	3.4	Conventional
Ada-Boost	90.8	0.1	Conventional
K-Mean, DT	99.5	0.3-1.7	Conventional

AdaBoost	90.9	1.4	Conventional
Hyper Spherical Cluster	85.4	13.2	Conventional
PCA	82.8	1.1	Conventional
Adaptive PCA	97.8	0.5	Conventional
Radial SVM	97.8	13	Conventional

IV. CHALLENGES

Based on the related literature, we found that the following issues have not been sufficiently solved. These are gaps in the reviewed work that would prove to be directions for future works.

A. Providing real Cloud IDS implementation

In identifying the performance and workability of a proposed Cloud IDS, prototypes need to be developed and tested in a physical Cloud computing environment. Most of the research provides solution and results based on simulation and static data analysis. Even some research that claim to be real implementation still provide analysis results based on static dataset and this is not really represents real Cloud IDS implementation. The real Cloud IDS must be developed based on a working prototype that monitoring Cloud network traffic or system calls or both which is termed as hybrid. Other research especially that implemented open-source technology in developing Cloud IDS, more focused on making use of traditional methods in protecting Cloud environment. This framework did not really represent the state-of-the-art Cloud IDS technology because implementing traditional method in Cloud computing environment will introduce another risk to the environment.

B. Cloud subscribers' data privacy

Within the process of detecting intrusions involve monitoring enormous data flow and most of the data belong to subscribers and may contains sensitive or private information. Protecting data privacy is the major challenges in Cloud computing and pooling user's data in Cloud introduces a new risk in data privacy breach.

C. Detecting co-residency attack

In [5] describe co- residency attack as an internal attack conducted by a tenant targeting another tenant residing within the same Cloud infrastructure. This type of attack is unique to Cloud and exist because of the nature of Cloud computing. Although this issue is important, little research has been carried out covering this challenge. Although multiple reviewed frameworks stressed on the importance for protecting Cloud from co-residency attack, very few of them implemented it. Other frameworks more focused on entry point security and endpoint security while neglecting the importance of monitoring the risk of co-residency attack.

V. CONCLUSION

It has been studied and analyzed, here that Hadoop based IDS are quite suitable for handling large training datasets. In this work, we compared various Hadoop based IDS, and also conclude that SVM, k means clustering, REP Tree and J48 classification algorithm have accuracy less than 95%. Map reduce is used to map the data based on (key, value) pair which is suitable for static data but due to advancement in technology, the need to work on streaming data is essential. The use of decision tree which has accuracy of 96% can further improve the system.

REFERENCES

[1] Z. Shi and J. An, "An Intrusion Detection System Based on Hadoop," 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), 2015, pp. 826-830, doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.162.

[2] M. M. Rathore, A. Paul, A. Ahmad, S. Rho, M. Imran and M. Guizani, "Hadoop Based Real-Time Intrusion Detection for High-Speed Networks," 2016 IEEE Global Communications Conference (GLOBECOM), 2016, pp. 1-6, doi: 10.1109/GLOCOM.2016.7841864.

[3] M. Kumar and M. Hanumanthappa, "Scalable intrusion detection systems log analysis using cloud computing infrastructure," 2013 IEEE International Conference on Computational Intelligence and Computing Research, 2013, pp. 1-4, doi: 10.1109/ICCIC.2013.6724158.

[4] S. R. Bandre and J. N. Nandimath, "Design consideration of Network Intrusion detection system using Hadoop and GPGPU," 2015 International Conference on Pervasive Computing (ICPC), 2015, pp. 1-6, doi: 10.1109/PERVASIVE.2015.7087201.

[5] Pai, Swathi & Kodada, Basappa. (2015). Big Data Analytics based Security Architecture to detect Intrusions.

[6] Gurpreet Kaur Jangla, D. (2015)." Development of an intrusion detection system on big-data for detecting unknown attacks".

IJARCCE vol.4.

[7] Sonali Ashok Hajare (2016). "Detection network attacks using Big-data analytics" IJRITCC vol .4.

[8] Shaik Akbar, T. R. (2016). "A hybrid Scheme based on big-data analytics using intrusion detection system" Indian Journal of Science and Technology, Volume: 9, Issue: 33, Pages: 1-4 DOI: 10.17485/ijst/2016/v9i33/97037

[9]https://www.researchgate.net/profile/Premchand_Ambhore/publication/267390056/figure/fig1/AS:392221510651908@1 470524303359/Fig-1-Block-Diagram-of-Intrusion-Detection-System-The-figure-shows-the-block-diagram-of.ppm

[10] B. Cui and S. He, "Anomaly Detection Model Based on Hadoop Platform and Weka Interface," 2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2016, pp. 84-89, doi: 10.1109/IMIS.2016.50.

[11] Riyaz A Jamadar, Prof. Ms. Mousami S Vanjale (2015)."Enhanced detection rate through PCA and radial SVM in Wireless Sensor Networks. "IERJ vol.9.

[12] K. Kato and V. Klyuev, "Development of a network intrusion detection system using Apache Hadoop and Spark," 2017 IEEE Conference on Dependable and Secure Computing, 2017, pp. 416-423, doi: 10.1109/DESEC.2017.8073860.

[13] Wei Hu and Weiming Hu, "Network-based intrusion detection using Adaboost algorithm," The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), 2005, pp. 712-717, doi: 10.1109/WI.2005.107.