ISSN 2395-1621

A Survey on Anomaly based Intrusion Detection using K-means Classifier in Hadoop

[#1]Neha Dhanawade, [#2]Meghana Mohite, [#3]Dhiraj Bahakar, [#4]Pratibha Gitte

[1]dhanawadeneha27@gmail.com,
[2]mmohite333@gmail.com,
[3]dhirajbahakar@gmail.com,
[4]pratibhagitte95@gmail.com

[#1234]Deparment of Computer Engineering
JSPM's, ICOER, Wagholi, Pune.

ABSTRACT

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices. An intrusion detection system (IDS) monitors network traffic and monitors for suspicious activity and alerts the system or network administrator. It identifies unauthorized use, misuse, and abuse of computer systems by both system insiders and external penetrators. Intrusion detection systems (IDS) are essential components in a secure network environment, allowing for early detection of malicious activities and attacks. By employing information provided by IDS, it is possible to apply appropriate countermeasures and mitigate attacks that would otherwise seriously undermine network security. In this project We have analyzed Machine learning techniques to detect intrusion which can scale up to build such systems. There are many algorithms one can opt for depending upon the need of system. This paper deals with Naïve Bayes and K-mean in Map-Reduce framework and also check their performance Our preliminary analysis. In this system, we have mainly focused on Anomaly Detection technique based on Machine Learning algorithms for Intrusion Detection in Big Data environment.

Keywords: Intrusion Detection System, Hadoop File System, MapReduce. Classification algorithms

ARTICLE INFO

Article History

Received: 2nd December 2016

Received in revised form :

2nd December 2016

Accepted: 5th December 2016

Published online :

5th December 2016

I. INTRODUCTION

Sophisticated hacking attacks are continuously increasing in the cyber space. Hacking in the past leaked personal information or were done for just fame, but recent hacking targets companies, government agencies. This kind of attack is commonly called APT (Advanced Persistent Threat). APT targets a specific system and analyses vulnerabilities of the system for a long time. Therefore it is hard to prevent and detect APT than traditional attacks and could result massive damage. Up to today, detection and protection systems for defending against cyber-attacks were firewalls, intrusion detection systems, intrusion prevention systems, anti-viruses solutions, database encryption.

In this paper, we propose a new model based on big data analysis technology to prevent and detect previously unknown attacks. Moreover, integrated monitoring technologies for managing system logs were used. These security solutions are developed based on signatures and blacklist. We compared previous researches which are based on data mining technology for predicting or analysing correlation between attack behaviours and explained its limits. Furthermore we list various sources and their details that can be collected and explain attack predictions earned from applying big data technologies such as classification, text mining, clustering, and association rules. Finally, we develop an Intrusion Detection System model based on big data technologies and evaluate the model.

We expect this research to be the basis for future implementation of APT attack detection and prevention systems based on big data analysis technologies. Detection systems and intrusion prevention systems are not capable of protecting systems against APT attacks because there are no

signatures. Therefore to overcome this issue, security communities are beginning to apply data mining technologies to detect previously unknown attacks.

II. LITERATURE SURVEY

[1] P. S. Rachana Sharma, "A Novel Approach towards Big Data Challenges," , 2014, The focus of this paper is on giving a holistic view of Big Data, its challenges, how present technologies are dealing with these challenges and what is more to be explored as a solution to Big Data. Also to look over technologies like Hadoop, MapReduce, BigQuery and Apache Sparks.

[2] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large Clusters", 2008, implementation of MapReduce runs on a large cluster of commodity machines and is highly scalable: a typical MapReduce computation processes many terabytes of data on thousands of machines.

[3] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, et al., "Apache Hadoop YARN: yet another resource negotiator,", 2013, in this paper, we summarize the design, development, and current state of deployment of the next generation of Hadoop's compute platform: YARN. The new architecture we introduced decouples the programming model from the resource management infrastructure, and delegates many scheduling functions.

[4] C. Zhang, E.-C. Chang, and R. H. Yap, "Tagged-mapreduce: A general framework for secure computing with mixed-sensitivity data on hybrid clouds," 2014, in this paper present a general security framework for analyzing MapReduce computations in the hybrid cloud which captures how dataflow can leak information through execution. Experiments on Amazon EC2 with our prototype in Hadoop show that we are able to obtain security while effectively outsourcing computation to
the public cloud and reducing inter-cloud communication.

[5] S. N. Srirama, P. Jakovits, and E. Vainikko, "Adapting scientific computing problems to clouds using MapReduce," 2012, this work shows how to adapt algorithms from each class into the MapReduce model, what affects the efficiency and scalability of algorithms in each class and allows us to judge which framework is more efficient for each of them, by mapping the advantages and disadvantages of the two frameworks.

III. INTRUSION DETECTION SYSTEM

A. Firewall:

A firewall is a network security system, either hardware or software-based, that controls incoming and outgoing network traffic based on a set of rules. Acting as a barrier between a trusted network and other untrusted networks such as the Internet or less trusted networks such as a retail merchant's network outside of a cardholder data environment a firewall controls access to the resources of a network through a positive control model. This means that the only traffic allowed onto the network defined in the firewall policy is; all other traffic is denied.

B. Intrusion Detection System:

An intrusion detection system (IDS) is a device or software application that monitors network or syste activities for malicious activities or policy violations an produces reports to a management station An intrusion detection system (IDS) inspects all inbound and outbound network activity and identifies suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system. There are several ways to categorize IDS:

Misuse detection vs. Anomaly detection:

In misuse detection, the IDS analyze the information it gathers an compares it to large databases of attack signatures Essentially, the IDS looks for a specific attack that ha already been documented. Like a virus detection system misuse detection software is only as good as the database of attack signatures that it uses to compare packets against In anomaly detection, the system administrator defines the baseline, or normal, state of the networks traffic load, breakdown, protocol, and typical packet size. The anomaly detector monitors network segments to compare their state to the normal baseline and look for anomalies.

IV. PROPOSED SYSTEM

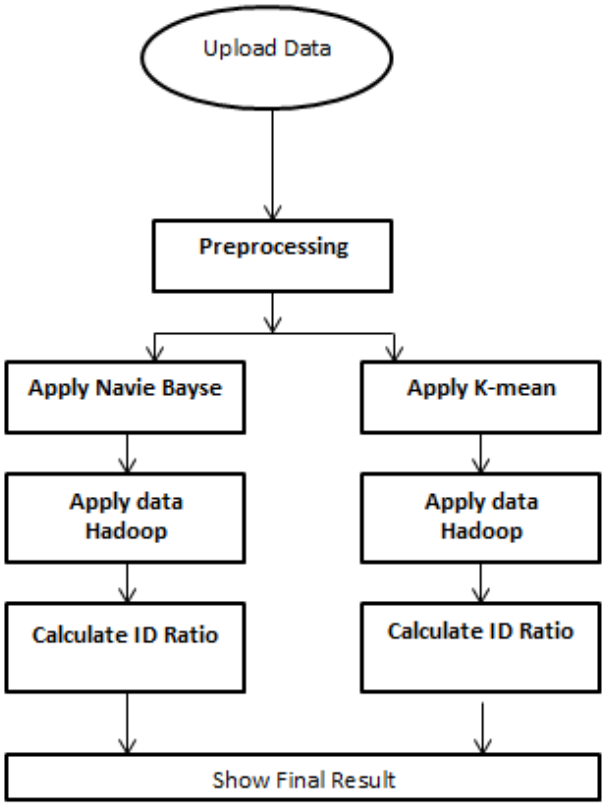

Fig 1. System architecture

In a Hadoop cluster, data is distributed to all the nodes of the cluster present on which data is stored as shown in fig. 2.Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others. In short, Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

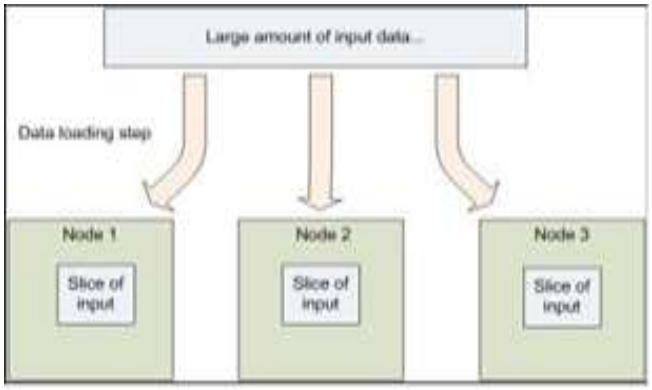

Fig 2. Hadoop Cluster

V. METHODOLOGY

Introduction of K-means clustering

K-Means Clustering is a method used to classify semi structured or unstructured data sets. This is one of the most commonly and effective methods to classify data because of its simplicity and ability to handle voluminous data sets. It accepts the number of clusters and the initial set of centroids as parameters. The distance of each item in the data set is calculated with each of the centroids of the respective cluster.

Introduction of Naïve Bayes Classifier

Naïve Bayes classifier is one of the supervised learning classification algorithms that can be programmed in form of MapReduce. In our study, we build a Naïve Bayes MapReduce model and evaluate the classifier on datasets based on the prediction accuracy. Also, a scalability analysis is conducted to see the speedup of the data processing time with the increasing number of nodes in the cluster.

VI. CONCLUSION

In this paper, we have focused Intrusion Detection in Big Data environment and successfully implemented Naïve Bayes and K-mean classification algorithms on the authentic network dataset. We have revamped the standard algorithms for map and reduce phases and made little changes for the sake of performance.

REFERENCES

[1] P. S. Rachana Sharma, "A Novel Approach towards Big Data Challenges," International Journal of Innovative Computer Science & Engineering, vol. 1, pp. 28 - 34, 2014.

[2] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, pp. 107-113, 2008.

[3] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, et al., "Apache Hadoop YARN: yet another resource negotiator," presented at the Proceedings of the 4th annual Symposium on Cloud Computing, Santa Clara, California, 2013.

[4] C. Zhang, E.-C. Chang, and R. H. Yap, "Tagged-mapreduce: A general framework for secure computing with mixed-sensitivity data on hybrid clouds," in Cluster, Cloud and Grid Computing (CCGrid), 2014 14th IEEE/ACM International Symposium on, Chicago, IL 2014, pp. 31-40.

[5] S. N. Srirama, P. Jakovits, and E. Vainikko, "Adapting scientific computing problems to clouds using MapReduce," Future Generation Computer Systems, vol. 28, pp. 184-192, 2012.

[12] W. Dai and W. Ji, "A mapreduce implementation of C4. 5 decision tree algorithm," International Journal of Database Theory and Application, vol. 7, pp. 49-60, 2014 2014.

[13] C. Zhang, F. Li, and J. Jestes, "Efficient parallel kNN joins for large data in MapReduce," in Proceedings of the 15th International Conference on Extending Database Technology, New York, NY, USA, 2012, pp. 38-49.

[14] H. Jingwei, Z. Kalbarczyk, and D. M. Nicol, "Knowledge Discovery from Big Data for Intrusion Detection Using LDA," inIEEE International Congress on Big Data (BigData Congress), Anchorage, AK 2014, pp. 760-761.

[15] J. Xiang, M. Westerlund, D. Sovilj, and G. Pulkkis, "Using extreme learning machine for intrusion detection in a big data environment," in Proceedings of Workshop on Artificial Intelligent and Security Workshop, Scottsdale, Arizona, USA, 2014, pp. 73-82.

[16] B. Panda, J. S. Herbach, S. Basu, and R. J. Bayardo, "Planet: massively parallel learning of tree ensembles with mapreduce," Proceedings of the VLDB Endowment ACM, vol. 2, pp. 1426-1437, 2009.

[17] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, et al., "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, pp. 1-37, 2008.

