Detection of Network Attacks Using Big Data Analysis

Sonal Ashok Hajare
Computer Science and Engineering,
Department of Computer Technology
Kavikulguru Institute of Technology and Science
Ramtek, India
e-mail: sonalhajare1989@gmail.com

Abstract: Security is always an important issue especially in the case of computer network which is used to transfer personal/confidential information's, ecommerce and media sharing. Data in computer networks is growing rapidly; the analysis of these large amounts of data to discover anomaly fragments has to be done within a reasonable amount of time. Recently, threat of previously unknown cyber-attacks is increasing because existing security systems are not able to detect them. The goal of recent hacking attacks has changed from leaking information and destruction of services to attacking large-scale systems such as critical infrastructures and state agencies. To defend against these attacks, which can not detected with existing intrusion detection algorithm we propose a new model based on big data analysis. Previous intrusion detection algorithm detects predefined attacks. This kind of intrusion detection system is called as signature based intrusion detection system. Big data analysis technique can extract information from Varity of sources to detect future attack. Big data analysis framework use MapReduced intrusion detection system based on clustering algorithm.

Keywords- *Hadoop, MapReduce, Targeted attacks, Intrusion detection system, C-Means Clustering, Support Vector Machine (SVM).*

*****

I. INTRODUCTION

Big data analytics is the process of examining large data sets containing a variety of data type's i.e. big data to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Big data analytics enables organizations to analyze a mix of structured, semi-structured and unstructured data in search of valuable business information and insights. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modellers and other analytics professionals to analyze large volumes of transaction dataBig data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analysis, data mining, text mining, statistical analysis. Mainstream BI software and data visualization tools can also play a role in the analysis process. But the semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually -- for example, real-time data on the performance of mobile applications or of oil and gas pipelines. As a result, many organizations looking to collect, process and analyze big data have turned to a newer class of technologies that includes Hadoop and related tool such as MapReduced.

Due to rapid development of Internet and technology, all the machines are connected to each other either by networked system or via mobile communication. The users are producing more and more data through communication media in the unstructured form which is highly unmanageable and this management of data is the challenging job. The main focus is to gather the unstructured data from all the terminals, processed the data to convert into structured form so that accessing of the data would be easier [2].

Apache Hadoop is an open source software framework written in java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines or racks of machines) are commonplace and thus should be automatically handled in software by the framework. The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce). Hadoop splits files into large blocks and distributes them amongst the nodes in the cluster. To process the data, Hadoop MapReduce transfers packaged code for nodes to process in parallel, based on the data each node needs to process. This approach takes advantage of data locality nodes manipulating the data that they have on hand to allow the data to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are connected via high-speed networking. The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file-system written in Java for the Hadoop framework. Each node in a Hadoop instance typically has a single namenode, and a cluster of datanodes form the HDFS cluster. The situation is typical because each node does not require a datanode to be present. Each datanode serves up blocks of data over the network using a block protocol specific to HDFS. The file system uses the TCP/IP layer for communication. Clients use Remote procedure call (RPC) to communicate between each other.

As a data parallel model, MapReduce is a patented software framework introduced by Google to support distributed computing on large datasets on clusters of computers. Known as a simple parallel processing mode, Map-reduce has many advantages: such as, it is easy to do parallel computation, to distribute data to the processors and to load balance between them, and provides an interface that is independent of the backend technology. MapReduce is designed to describe the process of parallel as Map and Reduce. The user of the

MapReduce library expresses the computation as two functions: Map and Reduce. The MapReduce library groups together all intermediate values associated with the same intermediate key and passes them to the Reduce function. It merges together these values to form a possibly smaller set of values.

II. DETAIL IMPLEMENTATION

A. *Collection of data set*

The data set is collected from the transactional database of any telecommunication networks, marketing and risk management, and inventory control etc. Data collection step collect data from firewalls and logs, behavior status information from antivirus, database, network device and system. Firewall is a regulation device that controls the network traffic between separated networks and hosts. Collected data is saved in big data appliance.

B. *Preform map reduce on data*

Map, written by the user, takes an input pair and produces a set of intermediate key/value pairs. The MapReduce library groups together all intermediate values associated with the same intermediate key I and passes them to the Reduce function. The Reduce function, also written by the user, accepts an intermediate key I and a set of values for that key. It merges together these values to form a possibly smaller set of values. Typically just zero or one output value is produced per Reduce invocation. The intermediate values are supplied to the user's reduce function via iterator. This allows us to handle lists of values that are too large to fit in memory. The Map and Reduce functions are both defined with respect to data structured in (key, value) pairs. MapReduce provides an abstraction that involves the programmer defining a "mapper" and a "reducer", with the following signatures:

Map::(key1)⇒ list(key2,value2)

Reduce::(key2,list(value2) ⇒ list(value2).

C. *Pre-processing of web log data*

Pre-processing of web log data include classification. Pre-processed data from previous step is analysed using prediction, classification, association analysis, and unstructured data analysis to decide user behaviour, system status, packet integrity and misuse of file or system. Classification is a technique that predicts the group of new attack from huge data. Classification helps security administrator to decide direction of protection and analysis. Most used classification techniques are C-Means clustering and SVM (Support Vector Machine). In fuzzy clustering each data point belongs to every cluster by some membership value and the process of grouping is iterated till the change in the membership values of each data point stops changing. In many situations, fuzzy clustering is more natural than hard clustering.

In the proposed system we firstly dataset is used to generate the training vectors. Now after collecting all these parameters fuzzy C-means clustering is applied and the data with closer distances are eliminated then this data is used to train the neural network which is later used for detection of Intrusion. The algorithm can be described in detail by following steps:

Step 1: Read the given numbers of samples from dataset.

Step 2: Filter selected features from the dataset for further processing.

Step 3: Partition the data into training and testing sets.

Step 4: Cluster the training dataset using the Fuzzy C-means Clustering.

Step 5: Select the fraction of data points from each edge of each cluster.

Step 6: Now train the Probabilistic Neural Network (PNN) using these vectors (data points) with their classification group.

Step 7: Test the trained PNN by the testing dataset.

Step 8: Calculate the performance of the trained system. [10]

D. *Pattern analysis*

Pattern analysis includes classification which contain support vector machine algorithm. Support Vector Machine is clustering algorithm which is used for detection of network attacks. Support Vector Machine is used for intrusion detection system to detect which IP contain viruses and clean that IP packet. Support Vector Machine constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Support Vector Machine are based on the concept of decision planes that defines decision boundaries. A decision plane is one that separates between a set of objects having different class membership. In Support Vector Machine algorithm, data set is partition into two part training part and testing part. Most current offline intrusion detection systems are focused on unsupervised and supervised machine learning approaches. Existing model has high error rate during the attack classification using support vector machine learning algorithm. Besides, with the study of existing work, feature selection techniques are also essential to improve high efficiency and effectiveness. In this proposed system, we used SVM classifier to detect Network attacks. [11]

Improved Support Vector Machine (SVM) based IDS model is presented the method for improvement of SVM to achieve the higher accuracy. A data pre-processing and removal of similar data to reduce the training data size using k means clustering presented in which shows significant improvement in training time with maintaining accuracy. One important requirement of classification is parameter selection because some of the features may be redundant or with a little contribution to the detection process.

E. *Suspicious IP's from where attack can occur*

Data set is used to generate the training vectors. Now after collecting all these parameter C-Means clustering is applied and the data with closer distance are eliminated then this data is used for detection of intrusion. The output of C-Means clustering algorithm with Support vector machine algorithm is compared and then declared which IP is suspicious or not. If any suspicious Ip is detected then that IP packet is cleaned. If attack or abnormal behaviours are detected, it alarms the administrator and terminates. Prediction information of analysed system is summarized and reported to the manager. Also configuration update, rule manipulation and deletion, analysis pattern updates are done both automatically and passively.

III. CONCLUSION

Data set are too large, to unstructured or semi-structured or to fast changing for analysis. Due to increase in number of sophisticated targeted threats and rapid growth in data, the analysis of data become too difficult and security of that data is limited using existing security technology. Recent unknown attack easily bypasses existing security system by using encryption and obfuscation. Therefore new detection method for reacting to such 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 MapReduced based detection algorithm that are C-Means clustering algorithm and Support Vector Machine algorithm.

REFERENCES

[1] 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.
[2] Bhawna Gupta and Dr. Kiran Jyoti, "Big data analytics with hadoop to analyzse targeted attack on enterprise data," International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3867-3870.
[3] Y. Lee, W. Kang, and Y. Lee, "Detecting DDoS Attacks with Hadoop", TMA, April 2011.
[4] Holtz, Marcelo D., Bernardo David, Sousa Jr., R. T., "Building Scalable Distributed Intrusion Detection Systems Based on the MapReduce Framework", Telecomunicacoes (Santa Rita do Sapucai), v. 13, p. 22-31, 2011.
[5] Prathibha.P.G and Dileesh.E.D, "Design of a Hybrid Intrusion Detection System using Snort and Hadoop", International Journal of Computer Applications (0975 – 8887) Volume 73– No.10, July 2013.
[6] Y.Lee, W.Kanf, H.Son, "An Internet Traffic Analysis Method with MapReduce",IEEE/IFIP Network Operations and Management Symposium Workshops,2010 ,pp:357-361.
[7] J. Mirkivic and P. Reiher, " A Taxonomy of DDoS Attack and DDoS Defense Mechanisms", ACM SIGCOMM CCR, 2004.
[8] Jeong Jin Cheon and Tae-Young Choe, "Distributed Processing of Snort Alert Log using Hadoop", IJET,Vol 5, No-3, Page 2685-2690, Jun-Jul 2013,
[9] 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.
[10] Rachnakulhare, Divakar singh "Intrusion Detection System based on Fuzzy C Means Clustering and Probabilistic Neural Network", International Journal of Computer Applications (0975 – 8887) Volume 74– No.2, July 2013.
[11] Venkata Suneetha Takkellapati1 , G.V.S.N.R.V Prasad "Network Intrusion Detection system based on Feature Selection and Triangle area Support Vector Machine", International Journal of Engineering Trends and Technology- Volume3Issue4- 2012.