Implementation of Anomaly based Intrusion Detection using Mining Algorithm using Hadoop

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

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

#1,2,3,4Department of Computer Engineering
JSPM’s, ICOER, Wagholi, Pune.

ABSTRACT

IDS is a device or software that monitors a network or system from malicious activity. The most common classification of IDS is anomaly based detection (to detect unknown attacks, this approach enables the detection of previously unknown attacks) and signature based it refers to the detection of attacks by looking for specific patterns, such as byte sequence in network traffic or known malicious instruction sequence used by malware. In this project we have analyzed machine learning techniques to detect intrusion which can scale up to build such system. This project deals with Naive Bayes and K-means in Hadoop and check their performance are preliminary analysis. In this system we have mainly focused on anomaly based on machine learning algorithms for IDS in Hadoop environment.

Keyword: IDS, Hadoop file system, classification algorithm, and clustering algorithm.

I. INTRODUCTION

Bigdata is term for the large data sets that are so large and complex which is not able to be handled by conventional applications such as databases and many others. There are many challenges occurs in our day to day lives that includes capturing of this data, sorting, analysing data, data curation, searching, sharing, transforming, visualization, querying, updating on data and provide data privacy. The term bigdata also refers to use of predictive analysis, user behaviour analytics or certain other data analytics methods that extract information from data. Scientist, Business Executives, Practitioners of medicine, Advertising and Governments alike regularly meet difficulties with this large amount of which is sharing on internet, finance, urban information and business informatics.

Security and Privacy issues are magnified by velocity, verity, volume of data such as a large scale cloud infrastructure diversity of data sources and formats, streaming nature of data acquisition and high volume inter cloud migration. Bigdata has many challenges to deal with secure computations in attributed programming framework, security best practices for non relational data store, security data storage and transaction logs, real time security compliance monitoring, scalable and composable privacy preserving data mining and analytics, cryptographically enforced access control and secure communication, granular access control, granualar audits, data provinance. There are many attacks that harms data. There are many more security breaches happened attacks like Utah's Health Departments breach and Twitter breach are some examples. Major of this attacks are classified in following categories DOS(Denial Of Service), R2L(Remote To Local), U2R(User to Root) and Probing attack. Cyber criminal or attacker attempts to gain access to information by staging the attacks.

As amount data being collected continues to grow more and more companies are building bigdata repositories to store, aggregate and extract meaning from their data unfortunately large set of consolidated data are a tempting target for cyber attackers breaching and bigdata repository can provide criminal groups with bigger pay offs and more recognition from single attack. As per network concern intrusion detection is very important, intrusion detection is type of security management system for computers and networks and intrusion detection system gathers and analyses information from various areas within a network to identify possible security breaches which include both intrusions(attacks from outside organization) and misuse (attacks from within organization).
Intrusion detection uses vulnerability assessment which technology develop to assess the security of computer network in big data environment. Intrusion detection function includes monitor and analyze network activities, analyzing network configuration and vulnerabilities, assessing system and file integrity, recognize patterns of attacks, analysis of abnormal activity pattern, tracking use policy violation. Machine learning is mostly used for building IDS needs less expert knowledge and gives accurate result even for zero day attacks. We used classification algorithm.

We mainly focus on anomaly based intrusion detection technique based on machine learning algorithm for intrusion detection in big data environment using Hadoop framework. Motivation for writing this paper is to utilize a programming model for scalable , fast and accurate detection of anomaly behavior and attacks using mapreduce framework. Hadoop is most widely used bigdata security solution as it provide proper authentication of users who access hadoop, insures that authorized hadoop users can only access the data that they entitled to access , ensuring that data access histories for all users are recorded in accordance with compliance regulation and for other important purpose, insuring protection of data both at rest and transit through enterprise-grade encryption. Hadoop also provide scalability, fault tolerance, highly available and distributed architecture. We implemented Navie Bayes Classifier(supervised classifier) and K-means Classifier(usupervised classifier) for intrusion detection in Hadoop version 1 framework for better anomaly detection.

II. 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 system 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.

III. PROPOSED SYSTEM

In our project we are implemented anomaly based intrusion detection system using k-means classifier and naive bayes in Hadoop framework. In this project, we first upload the KDD dataset then this data is pre-processed. This data is write on Hadoop for further processing, Hadoop process this data and after that k-mean and naive bayes classifier separately read this data from Hadoop and perform it’s classification on operation and give the further data to IDS for intrusion detection then final result is calculated is shown figure.

IV. METHODOLOGY

1 Anomaly based IDS:

An anomaly based intrusion detection system, is an intrusion detection system for detecting both network and computer intrusion and misuse by monitoring system activity and classification it as either normal or abnormal. The two phases of a majority of anomaly detection system consist of the training phase and testing phase, whereas in training phase a profile of normal behavior is built and in testing phase current traffic is compared with the profile created in the training phase.

2 Signature based intrusion detection system:
Signature based IDS refers to the detection of attacks by looking for specific pattern such as known malicious intrusion sequence used by malware. Signature based IDS can easily detect known attacks, it is impossible to detect new attacks for which no pattern is available.

HADOOP:
Hadoop is an open source, java based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. Hadoop having an a two component such as mapreduce and HDFS.

MAPREDUCE:
Hadoop mapreduce is a software frames for easily writing application which process big access of data in parallel an large clusters.

HDFS(Hadoop Distributed File System)
Hdfs provides a distributed file system that is design to run on large clusters (Thousands of computer) of small computer machine.

MACHINE LEARNING AND CLASSIFICATION USING MAPREDUCE:

Our paper studies of these algorithms that is Naive Bayes and k-means.

1) Naive Bayes Algorithm Formula: -

Naive Bayes classifier is one of the supervised learning classification algorithms that can be programmed in form of MapReduce. In our study, we build a Naive 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. The Naive Bayes algorithm is based on mathematical Bayes Theorem.

\[
P(H/D) = \frac{P(D/H) P(H)}{P(D)}
\]

Where,

\(P(h)=\) Independent probability of \(h\): prior probability.

\(P(D) = \) Independent probability of \(D\).

\(P(D/h) = \) Conditional probability of \(D\) given \(h\): like hood.

\(P(h/D) = \) Conditional probability of \(h\) given \(D\): posterior probability.

2) K-means Algorithm Formula :-

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.

\[
j = \sum_{j=1}^{k} \sum_{i=1}^{n} |x_i - \mu_j|^2
\]

Where,

\((j) = A\) data point

\(C_j = The\ cluster\ centre\)

\(N = \) Number of data points

\(K = \) Number of clusters

\(\| X_i - C_j\| = distance\ measure\ between\ a\ data\ point\ and\ the\ cluster\ centre\ C_j\).

V. SYSTEM ANALYSIS

We have created system in java. Data is processing on hadoop environment windows system. We have created a java application with local server.java application that communicates with local server and Trustee Server using REST API. We have uploaded text document on hadoop. We have evaluated time, speedup, automatically processing required for uploaded dataset. Here we also calculate the Anomaly based Intrusion Detection using Mining Algorithm and analysis result shown fig 3.

VI. RESULT
VII. CONCLUSION

We have focused Intrusion Detection in Big Data environment and successfully implemented Naive Bayes and K-means classification algorithms on the authentic network dataset.

We have also implemented the supervised and unsupervised algorithms and compared results of both algorithms. MapReduce prove to be more efficient with larger datasets. Our future work will include implementation of more Machine Learning classification algorithm on Spark and its comparison to MapReduce. Our future work will be dedicated to implementation of Machine Learning algorithms such as C5.0, SGD, C4.5 and others on Spark, MapReduce and other YARN platforms.

ACKNOWLEDGMENT

I am extremely thankful to Head, Department of Computer Engineering for providing me the necessary facilities during the course of this programmer. I would also like to say thanks to guide for his advice and valuable time during the course of my project work. Constant encouragement and inspiration during each and every step of my project work throughout the semester.

REFERENCES


