International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 11
ISSN: 2454-4248
295 – 297

Network Intrusion Detection System using Spark's Scalable Machine Learning Library

Pradeep Laxkar
PhD Scholar , CSE Department
SPSU Udaipur
Udaipur, India
pradeep.laxkar@gmail.com

Prof. Prasun Chakrabarti
HOD , CSE Department
SPSU Udaipur
Udaipur, India
prasun.chakrabarti@spsu.ac.in

Abstract—In this paper, considering that the serious network security situation we are facing and the problem of an increasing amount of data generated by the network, we proposed an Intrusion Detection System based on Spark's scalable machine learning library,In this paper we are showing that performance of Intrusion Detection system using sparks machine learning library is high in compare to hadoop. Fro IDS we will use K-Means algorithm.

Keywords—*K-means algorithm,* Hadoop*, KDD'99, Security, Intrusion Detection System, performance.*

******

I. INTRODUCTION

An Intrusion Detection System (IDS) monitors network traffic and monitors for suspicious activity and alerts the system or network administrator. In some cases the IDS may also respond to anomalous or malicious traffic by taking action such as blocking the user or source IP Addresses from accessing the network.

IDS come in a variety of "flavors" and approach the goal of detecting suspicious traffic in different ways.

There are network based (NIDS) and host based (HIDS) intrusion detection systems. There are IDS that detect based on looking for specific signatures of known threats- similar to the way antivirus software typically detects and protects against malware- and there are IDS that detect based on comparing traffic patterns against a baseline and looking for anomalies.

There are IDS that simply monitor and alert and there are IDS that perform an action or actions in response to a detected threat. We'll cover each of these briefly.

Network Intrusion Detection System

Network Intrusion Detection Systems are placed at a strategic point or points within the network to monitor traffic to and from all devices on the network. Ideally you would scan all inbound and outbound traffic, however doing so might create a bottleneck that would impair the overall speed of the network.

Host Intrusion Detection System

Host Intrusion Detection System are run on individual hosts or devices on the network. A HIDS monitors the inbound and outbound packets from the device only and will alert the user or administrator of suspicious activity is detected.

Signature Based

A signature based IDS will monitor packets on the network and compare them against a database of signatures or attributes from known malicious threats.This is similar to the way most antivirus softwares detects malware. The issue is that there will be a lag between a new threat being discovered in the wild and the signature for detecting that threat being applied to your IDS. During that lag time your IDS would be unable to detect the new threat.

Anomaly Based

An IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline will identify what is "normal" for that network- what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other- and alert the administrator or user when traffic is detected which is anomalous, or significantly different, than the baseline.

Passive IDS

A passive IDS simply detects and alerts. When suspicious or malicious traffic is detected an alert is generated and sent to the administrator or user and it is up to them to take action to block the activity or respond in some way.

Reactive IDS

A reactive IDS will not only detect suspicious or malicious traffic and alert the administrator, but will take pre-defined proactive actions to respond to the threat. Typically this means blocking any further network traffic from the source IP address or user. One of the most well known and widely used intrusion detection system is the open source, freely available Snort. It is available for a number of

 ISSN: 2454-4248
295 – 297

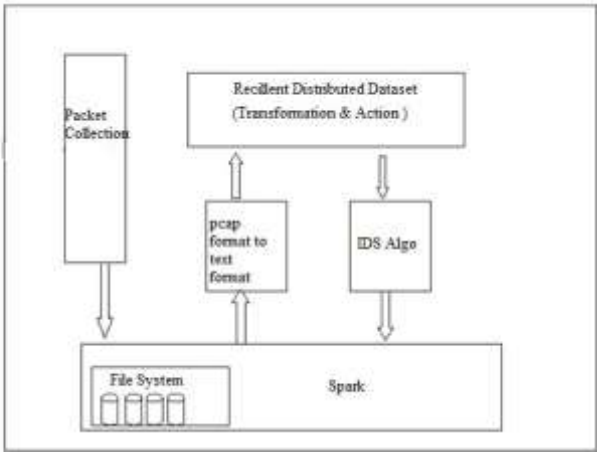
platforms and operating system including both Linux and windows. Snort has a large and loyal following and there are many resources available on the Internet where you can acquire signatures to implement to detect the latest threats.

An IDS can be a great tool for proactively monitoring and protecting your network from malicious activity, however they are also prone to false alarms. With just about any IDS solution you implement you will need to "tune it" once it is first installed. You need the IDS to be properly configured to recognize what is normal traffic on your network vs. what might be malicious traffic and you, or the administrators responsible for responding to IDS alerts, need to understand what the alerts mean and how to effectively respond. Parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. LITERATURE REVIEW

1. Shi (2015)[8] considered that the serious network security situation people are facing and the problem of an increasing amount of data generated by the network, author proposed an Intrusion Detection System based on Hadoop, due to the lack of the traditional K-Means algorithm exists at we propose an improved K-Means algorithm, we analyze the performance of the K-Means algorithm and the improved K-Means algorithm with KDD '99 data sets by using the Intrusion Detection System based on Hadoop.
2. Zamani et. al.(2015) [8]reviewed several influential algorithms for intrusion detection based on various machine learning techniques. Characteristics of ML techniques makes it possible to design IDS that have high detection rates and low false positive rates while the system quickly adapts itself to changing malicious behaviors. They divided these algorithms into two types of ML-based schemes: Artificial Intelligence (AI) and Computational Intelligence (CI). Although these two categories of algorithms share many similarities, several features of CI-based techniques, such as adaptation, fault tolerance, high computational speed and error resilience in the face of noisy information, conform the requirement of building efficient intrusion detection systems.
3. Kumari et. al.(2016) [5]worked on K-Means clustring algorithm and conclude that K-Means model itself can be used to demonstrate an essence to detect anomalies which can form the core for detecting any anomalies in a network traffic based on the features obtained from the data. This can be combined with spark streaming to detect anomalies and hence in turn possible intrusions from data arriving in real time thus building an actual deployable network intrusion detection application from detecting anomalies. Machine learning library also includes a variation called Streaming KMeans, which can update a clustering incrementally as new data arrives in Streaming KMeans Model. This could be used to continue to learn, approximately, how new data affects the clustering, and not just assess the given data for anomalies. Besides we used the simplistic Euclidean distance based approach since it was readily available in the Machine learning library.
4. Mukund et. al. (2016) [6]described an account of the existing algorithms for Intrusion Detection using Machine Learning, along with certain new ideas for improving the same. They discussed that employing the Decision Tree mechanism for Intrusion Detection and improve it with the distributed file system, Hadoop. Initially a method that used a dirty-flags to check the consistency of the Decision Tree, which changes with every wrong classification of the system was employed. The wrong classification was identified by a certain user who informs the system about the same and helps it learn.

III. PROPOSED METHOD



i) Packet collector

Packet collector capture incoming packet with a fix time window out every 1000 packets and generate packet trace file . Then information is extracted from six fields of IP headerthat accommodate source IP address, destination IP address, time interval and length of the packet trace file that would be stored to RDD.

ii) Spark RDD

Spark RDD is used to store big data . Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is

 ISSN: 2454-4248
295 – 297

an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster. RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes.

Activity 2:

i) Format Converter

The common format of spark is a test file. But the packet trace file (.pcap) is a binary format. So there is a need to convert the packet to trace file to text file, and split by each line to Data processor.

ii) Data Processor

This phase is used to resilient distributed dataset (RDD) to compute some necessary features of packets. Grouping of the packet and calculation of the features which have the same source IP address and destination IP address have to be done. The data processor will compile by R language .

Implementation:

The implementation part will be carried out using java programming.

IV. CONCLUSION

By implementing Intrusion Detection System using Spark's scalable machine learning library we can get more faster IDS.For implementation we will use python/Java.In next paper we will use experimental result and will show how spark is better than hadoop for intrusion detection system.

REFERENCES

[1] Bhatnagar R. ,Shrivastava A.K. and Sharma A, "An Implementation Approach for Intrusion Detection System in Wireless sensor Network", International Journal on Computer Science and Engineering 5(2) pp.303-308 (2010).

[2] Bhatti R, LaSalle R, Bird R and Grance T, E "Emerging trends around big data analytics and security Panel" , In Proc. 17th ACM Symposium on Access Control Models and Technologies. SACMAT '12. ACM, New York, NY, USA, pp. 106-112 (2012).

[3] Cheon J and Choe T-Y, "Distributed processing of snort alert log using hadoop",International Journal Engineering & Technology, 5(3), pp. 2685–2690(2013)

[4] Frank J, "Artificial intelligence and intrusion detection: current and future directions", In Proc. 17th national computer security conference,10,Citeseer, Baltimore, MD, USA , pp.194-198(1994).

[5] Kumari R. , Sheetanshu , Singh M. K. , Jha R and Singh N.K. , "Anomaly Detection in Network Traffic using K-mean clustering , In proc. 3rd Int'l Conf. on Recent Advances in Information Technology, pp. 78-82(2016) .

[6] Mukund , Nayak Y. R. S. and Chandrasekaran K. , "Improving False Alarm Rate in Intrusion Detection Systems Using Hadoop ",In Proc. International Conference on Advances in Computing, Communications and Informatics (ICACCI), India, pp. 21-24(2016).

[7] Shi Z. "An Intrusion Detection System Based on Hadoop" In Proc. UIC-ATC-ScalCom- CBDCom-IoP, pp.98-103(2015).

[8] Zamani M. and Movahedi Ma., "Machine Learning Techniques for Intrusion Detection "Journal of arXiv:1312.2177 v2 9th May , pp. 20-27(2015).