ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Big Data Analytics and Hadoop for Detecting Targeted Attacks

GurpreetK Jangla, Deepa.A.Amne

M.Tech Student, Dept. of C.S.E., B.I.T. Ballarpur, Gondwana University, Maharashtra, India

Asst. Professor, Dept. of C.S.E., B.I.T. Ballarpur, Gondwana University, Maharashtra, India

ABSTRACT: Today, cyber threats are increasing because existing security systems are not capable of detecting them. Previously, attacks had simple aim to attack or destroy the system. However, 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. Existing defence technologies to detect these attacks are based on pattern matching methods which are very limited. Because of this fact, in the event of new and previously unknown attacks, detection rate becomes very low and false negative increases. To defend against these unknown attacks,we propose a new model based on big data analysis techniques that can extract information from a variety of sources to detect future attacks.

KEYWORDS: Cyber-crime, Data Mining, Intrusion Detection, Hadoop

I. INTRODUCTION

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 attack is a special kind of attack that use social engineering, zero day vulnerabilities and other techniques to penetrate into the target system and persistently collect valuable information. It can give massive damage to national agencies or enterprises.

An advanced persistent threat (APT) uses multiple phases to break into a network, avoid detection, and harvest valuable information over the long term. This info-graphic details the attack phases, methods, and motivations that differentiate APTs from other targeted attacks. Till today, security systems for detection and protection systems against cyber-attacks are firewalls, intrusion detection systems, intrusion prevention systems, anti-viruses solutions, database encryption, DRM solutions and etc. Moreover, integrated monitoring technologies for managing system logs are used. These security solutions are developed based on signatures and blacklist. According to various reports, intrusion detection systems and intrusion prevention systems are not capable of defending against APT attacks because there are no signatures. Therefore to overcome this issue, security experts are beginning to apply data mining technologies to detect previously targeted attacks. we propose a new model based on big data analysis technology to prevent and detect previously unknown APT attacks..

APT attack is usually done in four steps: intrusion,searching, collection and attack. Figure 1 describes the attack process in detail.

DOI: 10.15680/IJIRCCE.2016. 0406295

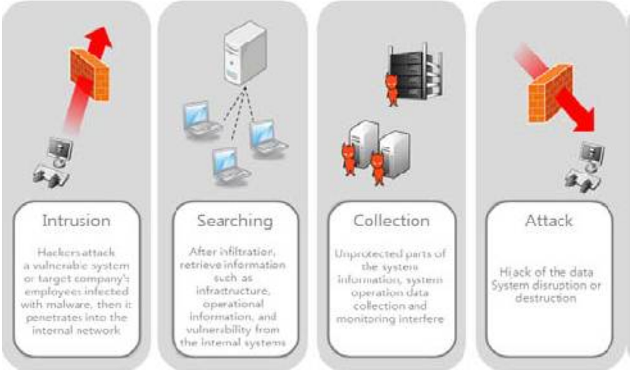

Intrusion
Hackers attack a vulnerable system or target company's employees infected with malware, then it penetrates into the internal network
Searching
After infiltration, retrieve information such as infrastructure, operational information, and vulnerability from the internal systems
Collection
Unprotected parts of the system information, system operation data collection and monitoring interfere
Attack
Hi ack of the data System disruption or destruction

Figure 1:Steps in APT Attacks

In the intrusion step, the hacker searches for information about the target system and prepares the attack. To get the access to the system, the attacker searches for users with high access privileges such as administrators and use various attack techniques such as phishing, spoofing etc.

Searching is done after the hacker gained access to the system. Hacker analyses system data such as system log for valuable information and look for security vulnerabilities than can be exploited for further malicious behaviours.

In the collection step, after the hacker has obtained valuable information in the system then the hacker installs malwares such as Trojan horse, trapdoors and backdoors to collect system data and maintain system access for the future.

In the final step, the hacker leaks data and destroys target system using the gained information.

II. Related work

Researchers developed various cyber security technologies to protect the system from threats and attacks. Some of the techniques were Firewall,IDS, Web Application Filter. Previously unknown attacks such as APT are evolving to bypass existing security measures. These attacks are impossible to detect or prevent with current technologies. Therefore security events constantly occurs using state-of-the-art attack technologies. New security measures to react to these attacks are needed. The new paradigm requires big data analysis techniques as a core of defense technologies, central security management, incident prediction technologies.We plan to develop a big data based system for detecting attacks which are unknown to the existing system. This is done using previous learning about the attacks on which the system is trained previously and finding patterns about these attacks. Once the pattern learning process is done, we would apply these learned patterns to the new input stream in order to detect any unknown attacks. Hadoop will be used to process the data, it will first map the input dataset into code understandable patterns, and then reduce these patterns to get information about the intrusion type. Big Datais a collection of large datasets that cannot be processed using traditional computing techniques. It is not a single technique or a tool, rather it involves many areas of business and technology.Thus Big Data includes huge volume, high velocity, and extensible variety of data.

In this paper we will see how practically attacks can be detected using data mining algorithms based on big data analytics.

III. Network Behaviour Analysis

Within the last few years, Network Behavior Analysis (NBA) has been one of these emerging technologies that have been sold as a security management tool to improve the current network security status. The main focus of NBA is to monitor inbound and outbound traffic associated with the network to ensure that nothing is getting into the servers, software, and application systems which helps enhance the overall security of the network at all levels. It is stated that approximately 25% of large enterprises systems will be using NBA by 2011.

First of all, the model have little proactive capability attitude toward preventing any security incidents because the architecture is built with technologies that discover most security events in progress while it misses opportunities to detect and resolve other small threats before it become major problems for the network. Firewalls and

DOI: 10.15680/IJIRCCE.2016. 0406295

intrusion detection systems are typically stationed at a network gateway, which doesn't stop laptops infected with malware or subversiveemployees from accessing the network. A typical security tactic to overcoming this problem is to deploy firewalls and intrusion detection devices throughout the internal network. This can get extremely expensive and can increase network maintenance and complexity even without addressing many of the security threats.

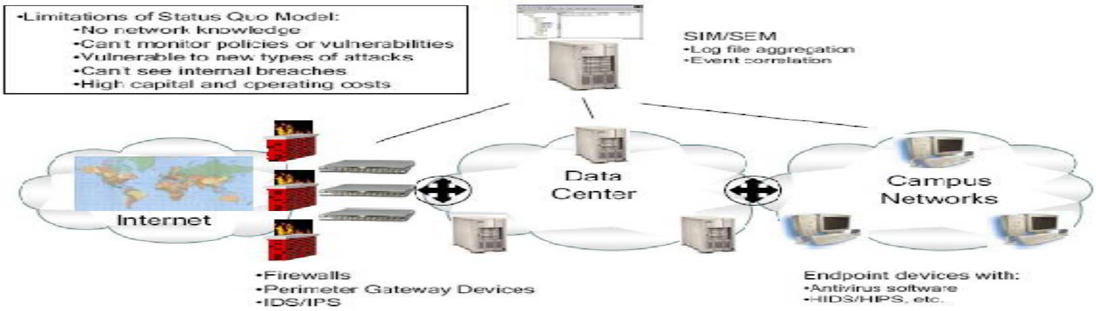

•Limitations of Status Quo Model:
•No network knowledge
•Can't monitor policies or vulnerabilities
•Vulnerable to new types of attacks
•Can't see internal breaches
•High capital and operating costs
SIM/SEM
•Log file aggregation
•Event correlation
Internet
Data Center
Campus Networks
•Firewalls
•Perimeter Gateway Devices
•IDS/IPS
Endpoint devices with:
•Antivirus software
•HIDS/HIPS, etc.

Figure. 2: Traditional Network Defense Strategy Model

IV. Proposed Algorithm

SVM (Support Vector Machine) for Classification: "Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.

A. *Description of the Proposed Algorithm:*

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).There are many linear classifiers (hyper planes) that separate the data. However, only one of these achieves maximum separation. The reason we need it is because if we use a hyper plane to classify, it might end up closer to one set of datasets compared to others and we do not want this to happen and thus we see that the concept of maximum margin classifier or hyper plane as an apparent solution.

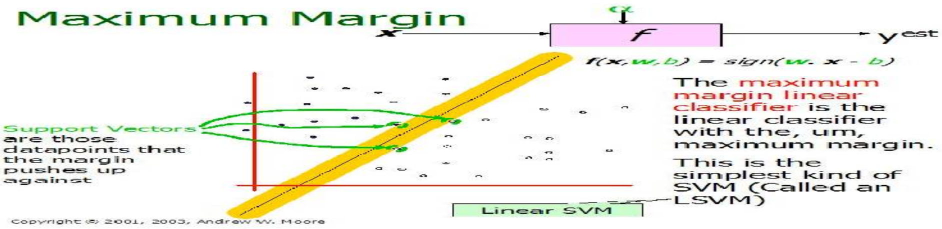

Maximum Margin
x
α
f
y^est
f(x,w,b) = sign(w. x - b)
The maximum margin linear classifier is the linear classifier with the, um, maximum margin.
This is the simplest kind of SVM (Called an LSVM)
Support Vectors are those datapoints that the margin pushes up against
Linear SVM
Copyright © 2001, 2003, Andrew W. Moore

Figure 3: Illustration of Linear SVM

A support vector machine constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Steps: 1) Correctly classify all training data

if$y_i = +1$ $\quad wx_i + b \geq 1$

if$y_i = -1$ $\quad wx_i + b \leq 1$

for all i $\quad y_i(wx_i + b) \geq 1$

2) Maximize the Margin $M = \frac{2}{|w|}$

same as minimize $\frac{1}{2} w^t w$

We can formulate a Quadratic Optimization Problem and solve for w and b

Minimize

subject to $\Phi(w) = \frac{1}{2} w^t w$

$$y_i(wx_i + b) \geq 1$$

V. RESULTS

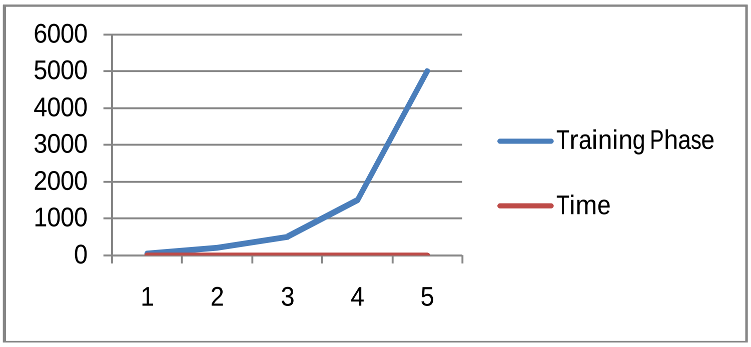
The proposed algorithm is used with data mining technique of classification with linear classifier.Detecting the unknown attacks means here we are comparing the signature of a attack with other type of signature. The data undergoes two phases i.e. training and testing phase.In training phase,we enter the number of entries to read.In training phase,we enter number of entries to train data sets.

We draw a table of both testing and training entries of data set and time required for manipulating the dataset. Below table shows the entries for testing and training datasets along with time required.

Training Phase Entries	Testing Phase Entries	Time required(ns)
50	150	1.3134
200	300	1.4639
500	650	1.6025
1500	1500	1.8771
5000	6000	2.1545

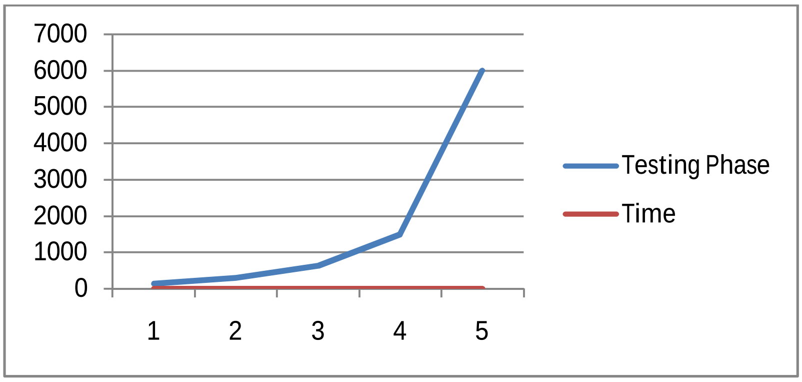
Table 2.Showing training and testing phase entries with time required

As shown in above table, as the number of entries in each phase increases the time required for manipulating that dataset also increases. The graph of each phase versus time is plotted which is shown in below two graphs.



Graph 1:Training phase vs time

As seen in graph 1,as we increase number of entries of datasets in training phase,the time required for manipulating that dataset also increases.



Graph 2:Testing phase vs time

In testing phase, the unknown attacks is being detected by manipulating the entries.Similarly, as the entries of datasets for testing phase increases, the time required for manipulating such datasets also increases shown in graph 2.

VI. Conclusion and Future Work

Recent unknown attacks easily bypass existing security solutions by using encryption and obfuscation. Therefore there is a need to develop a new detection methods for reacting to such attacks. To defend against these unknown attacks, which cannot be detected with existing technology the model is proposed.We presented a survey of the various data mining techniques that have been proposed towards the enhancement of IDSs. We have shown the ways in which data mining has been known to aid the process of Intrusion Detection and the ways in which the various techniques have been applied and evaluated by researchers. Finally, in the last section, we proposed a data mining approach that we feel can contribute significantly in the attempt to create better and more effective Intrusion Detection Systems.Enterprise data security is challenging task to implement and calls for strong support in terms of security policy formulation and mechanisms. We plan to take up developing security alerts which will provide employees with the ability to view tactivity. Events will be filtered down and summarized view will be available to each individual employee.

References

[1] Giovanni Vigna, “A Stateful Intrusion Detection System for World-Wide Web Servers” Computer Security Applications Conference, 2003. Proceedings. 19th Annual 8-12 Dec. 2003, IEEE.

[2] Baiju Shah “How to Choose Intrusion Detection Solution” SANS Institute Resources, July 24, 2001.

[3] Christos Douligeris, AikateriniMitrokotsa, ”DDoS attacks and defense mechanisms: classification and state-of-the-art” ,Computer Networks: The International Journal of Computer Telecommunications Networking,Vol. 44, Issue 5 , pp: 643 - 666, 2004.

[4] MithcellRowton, Introduction to Network SecurityIntrusionDetection,December 2005.

[5] R. Magoulas and B. Lorica, “Introduction to Big Data”, Release 2.0 (Sebastopol O’Reilly Media), Feb, 2009.

[6] Anderson. J. P. “Computer Security Threat Monitoring and Surveillance.” Technical Report, James P Anderson Co., Fort Washington, Pennsylvania, 1980.

[7] J. Feiman, “Hype Cycle for Application Security, 2012”, Gartner Group, July, 2012.

[8] “Advanced Persistent Threat: A Decade in Review”, Command Five Pty Ltd, June, 2011.

[9] Dr. KiranJyoti, Bhawna Gupta. "`Big data analytics with hadoop to analyse targeted attacks on enterprise data"'. Technical Report, International Journal of Computer Science and Information Technologies, IJCSIT, Vol 5(3) 2014.

[10] R. D. Pietro and L. V. Mancini, Intrusion detection systems, in: S.Jajodia (Series editor), Handbook of Advances in Information Security, Springer, 2008

[11]] Tai-Myoung Chung Sung-Hwan Ahn, Nam-Uk Kim. "`Big data analysis system concept for detecting unknown attacks"'. Technical Report, February 2014

[12]] “Advanced Persistent Threat: A Decade in Review”, Command Five Pty Ltd, June, 2011.