International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 1, pp: 754-758 www.ijaem.net ISSN: 2395-5252

An Intrusion Detection System Based On Big Data for Detecting Unknown Attacks

Ms. Sneha Sakhare, Mr. HirendraHajare
MTech Student, Assit. Professor
Department of CSE, Ballarpur Institute of Technology (BIT), Ballarpur.

Date of Submission: 21-06-2020 Date of Acceptance: 07-07-2020

ABSTRACT
Nowadays, Cyber-attacks are increasing because the existing security technologies are not capable of detectingit.Previous cyber-attacks were having simple motive of hacking and damaging the system.But today the motive haschanged from attacking the system or network to attacking the large-scale systems such as organizations or nationalagencies. In other words, existing security technologies to counter these attacks are based on pattern matching methodswhich are very limited. Because of this fact, in the presence of new and previously unknown attacks, detection ratebecomes very low and false negative increases. For this reason, a new model has been proposed based on Big Data fordetecting unknown attacks. Big Data can extract information from variety of sources to detect future attacks. We expectour model to be the basis of the future Advanced Persistent Threat (APT) detection and prevention systemimplementations.

KEYWORDS: Intrusion detection, Data mining, Hadoop, Map Reduce, Targeted Attacks.

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 aspecial kind of attack that use social engineering, zero-day vulnerabilities and other techniques to penetrate into thetarget system and persistently collect valuable information. It can give massive damage to national agencies orenterprises.

An advanced persistent threat (APT) uses multiple phases to break into a network, avoid detection, and harvestvaluable information over the long term. This info-graphic details the attack phases, methods, and motivations thatdifferentiate APTs from other targeted attacks. Till today, security systems for detection and protection systems againstcyber-attacks are firewalls, intrusion detection systems, intrusion prevention systems, antiviruses' solutions, databaseencryption, 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, intrusiondetection systems and intrusion prevention systems are not capable of defending against APT attacks because there areno signatures. Therefore, to overcome this issue, security experts are beginning to apply data mining technologies todetect previously targeted attacks. we propose a new model based on big data analysis technology to prevent and detectpreviously unknown APT attacks.

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

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 1, pp: 754-758 www.ijaem.net ISSN: 2395-5252

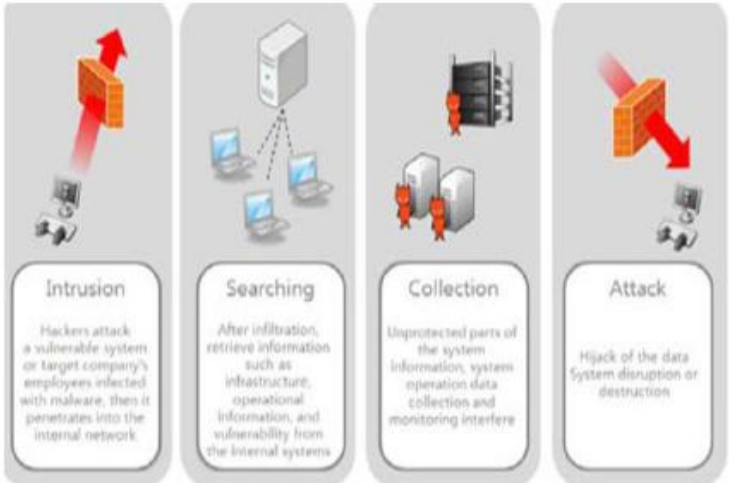


Fig: The sequence of APT attacks

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 thetechniques were Firewall, DS, Web Application Filter. Previously unknown attacks such as APT are evolving to bypassexisting security measures. These attacks are impossible to detect or prevent with current technologies. Therefore, security events constantly occur using state-of-the-art attack technologies. New security measures to react to theseattacks are needed. The new paradigm requires big data analysis techniques as a core of defense technologies, centralsecurity management, incident prediction technologies.We plan to develop a big data-based system for detecting attackswhich are unknown to the existing system. This is done using previous learning about the attacks on which the systemis trained previously and finding patterns about these attacks. Once the pattern learning process is done, we wouldapply these learned patterns to the new input stream in order to detect any unknown attacks. Hadoop will be used toprocess the data, it will first map the input dataset into code understandable patterns, and then reduce these patterns toget information about the intrusion type. Big Details a collection of large datasets that cannot be processed usingtraditional computing techniques. It is not a single technique or a tool, rather it involves many areas of business andtechnology. 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 dataanalytics.

III. SECURITY TECHNOLOGIES

Researchers developed various cyber security technologies to protect the system from threats and attacks. Some of the techniques to maintain cyber security:

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

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 1, pp: 754-758 www.ijaem.net ISSN: 2395-5252

system activities for malicious activities or policy violations and 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:

1. Misuse detection vs. Anomaly detection: In misuse detection, the IDS analyze the information it gathers and compares it to large databases of attack signatures. Essentially, the IDS look for a specific attack that has 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 thebaseline, 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.
2. Network-based vs. Host-based systems: In a network-based system, or NIDS, the individual packets flowing through a network are analysed. The NIDS can detect malicious packets that are designed to be overlooked by a firewalls simplistic filtering rule. In a host-based system, the IDS examine at the activity on each individual computer or host.
3. Passive system vs. Reactive system: In a passive system, the IDS detect a potential security breach, logs the information and signals an alert. In a reactive system, the IDS respond to the suspicious activity by logging off a user or by reprogramming the firewall to block network traffic from the suspected malicious source.

Though they both relate to network security, an IDS differs from a firewall in that a firewall looks out for intrusions in order to stop them from happening. The firewall limits the access between networks in order to prevent intrusion and does not signal an attack from inside the network. An IDS evaluates a suspected intrusion once it has taken place and signals an alarm. An IDS also watches for attacks that originate from within a system.

IV. PROPOSED ALGORITHM

SVM (Support Vector Machine) for Classification: "Support Vector Machine" (SVM) is a supervised machinelearning algorithm which can be used for both classification or regression challenges. However, it is mostly used inclassification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is numberof features you have) with the value of each feature being the value of a particular coordinate. Then, we performclassification 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 whichbest segregates the two classes (hyper-plane/ line).There are many linear classifiers (hyper planes) that separate thedata. However, only one of these achieves maximum separation. The reason we need it is because if we use a hyperplane to classify, it might end up closer to one set of datasets compared to others and we do not want this to happenand thus we see that the concept of maximum margin classifier or hyper plane as an apparent solution.

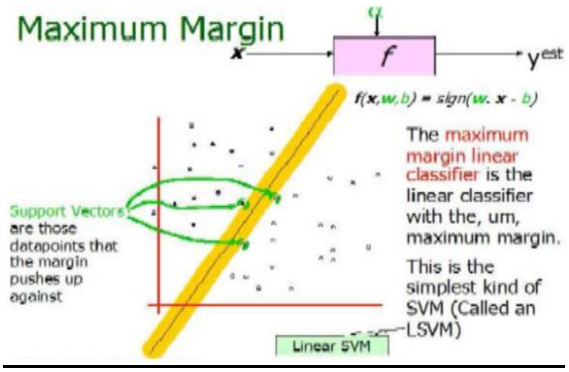


Fig: Illustration of Linear SVM

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

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

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

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

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

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 1, pp: 754-758 www.ijaem.net ISSN: 2395-5252

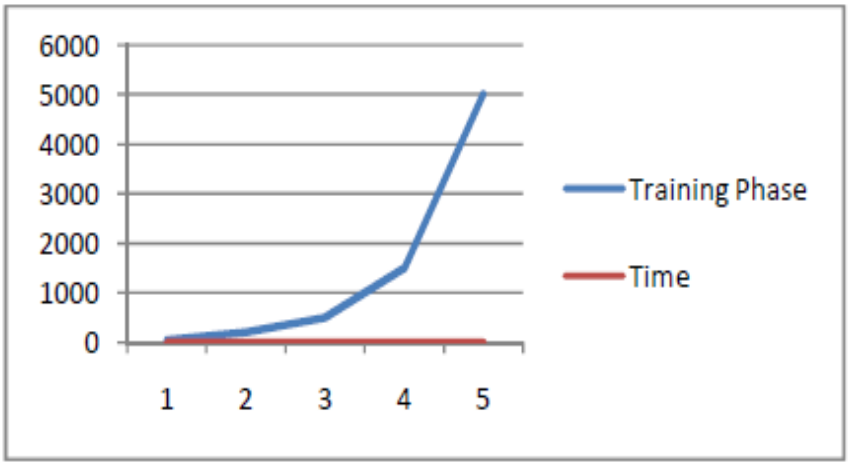
V. RESULTS

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 1. 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.

VI. CONCLUSION

Recent unknown attacks easily bypass existing security solutions by using encryption and obfuscation. Therefore, there is a need to develop a new detection method 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 activity. Events will be filtered down and summarized view will be available to each individual employee.

REFERENCES

[1]. Tai-Myoung Chung Sung-Hwan Ahn, Nam-Uk Kim. "`Big data analysis system concept for detecting unknown attacks"'. Technical Report, February 2014
[2]. R. D. Pietro and L. V. Mancini, Intrusion detection systems, in: S.Jajodia (Series editor), Handbook of Advances in Information Security, Springer, 2008.
[3]. P. Chapman. et al, "CRISP-DM 1.0 – Step-by-step data mining guide",http://www.crisp-dm.org (2000).
[4]. "Advanced Persistent Threat: A Decade in Review", Command Five Pty Ltd, June, 2011.
[5]. Dr. Kiran Jyoti, 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.
[6]. R. Magoulas and B. Lorica, Introduction to Big Data, Release 2.0 (Sebastopol Reilly Media, February 2009.
[7]. N. Srinivasan and V. Vaidehi. "Timed Coloured Petri Net Model for Misuse Intrusion Detection" First International Conference on Industrial and Information Systems, 8-11 Aug. 2006.
[8]. "Hadoop Tutorial from Yahoo!", Module7: Managing a Hadoop Cluster.
[9]. Larry Barrett, "Big data analytics: the enterprises next great security weapon?" February 2014.

International Journal of Advances in Engineering and Management
ISSN: 2395-5252

IJAEM

Volume: 02 Issue: 01 DOI: 10.35629/5252

www.ijaem.net
Email id: ijaem.paper@gmail.com