IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Development of an Intrusion Detection System based on Big Data for Detecting Unknown Attacks

Miss Gurpreet Kaur Jangla[1], Mrs Deepa.A.Amne[2]

Student, Computer Science &Engineering Department, BIT, Ballarpur, India [1]

Assistant Professor, Computer Science &Engineering Department, BIT, Ballarpur, India [2]

Abstract: Today Cyber-attacks are increasing because the existing security technologies are not capable of detecting it.Previous cyber-attacks were having simple motive of hacking and damaging the system.But today the motive has changed from attacking the system or network to attacking the large scale systems such as organizations or national agencies.. In other words, existing security technologies to counter these attacks are based on pattern matching methods which are very limited. Because of this fact, in the presence of new and previously unknown attacks, detection rate becomes very low and false negative increases.For this reason,a new model has been proposed based on Big Data for detecting unknown attacks.Big Data can extract information from variety of sources to detect future attacks.We expect our model to be the basis of the future Advanced Persistent Threat (APT) detection and prevention system implementations.

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

I. INTRODUCTION

Sophisticated hacking attacks are continuously increasing in the cyber space. 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 targets a specific system and analyses vulnerabilities of the system for a long time. Therefore it is hard to prevent and detect APT than traditional attacks and could result massive damage. Up to today, detection and protection systems for defending against cyber-attacks were firewalls, intrusion detection systems, intrusion prevention systems, anti-viruses solutions, database encryption.

In this paper, we propose a new model based on big data analysis technology to prevent and detect previously unknown attacks. Moreover, integrated monitoring technologies for managing system logs were used. These security solutions are developed based on signatures and blacklist. We compared previous researches which are based on data mining technology for predicting or analysing correlation between attack behaviours and explained its limits. Furthermore we list various sources and their details that can be collected and explain attack predictions earned from applying big data technologies suchas classification, text mining, clustering, and association rules. Finally, we develop an Intrusion Detection System model based on big data technologies and evaluate the model.

We expect this research to be the basis for future implementation of APT attack detection and prevention systems based on big data analysis technologies. Detection systems and intrusion prevention systems are not capable of protecting systems against APTattacks becausethere are no signatures. Therefore to overcome this issue, security communities are beginning to apply data mining technologies to detect previously unknown attacks.

II. MOTIVATION

1. Advanced Persistent Threats

Advanced Persistent Threats (APTs) are a well-resourced, highly capable and relentless class ofhacker increasingly referred to in the media, by ITsecurity companies, victims, and law enforcement.Most hackers target indiscriminately and instead ofpersisting with a particular target draw their focusto more vulnerable targets. APTs on the other handare not only well resourced and capable butpersistent in their covert attempts to accesssensitive information, such as intellectual property,negotiation strategies or political dynamite, fromtheir chosen targets.

APTs often target unpublicized vulnerabilities in computer programs or operating systems using ‘zero day’ exploits. Typically only well-resourcedhackers develop such exploits as they are expensive2, timeconsuming, and the vulnerabilities they target may be patched prior to deployment affecting the value of the investment. In addition, zero day exploits are exposed the first time they are used and, if detected, may be less effective in future Attacks.

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

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

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

various attack techniques such as phishing, spoofing etc.

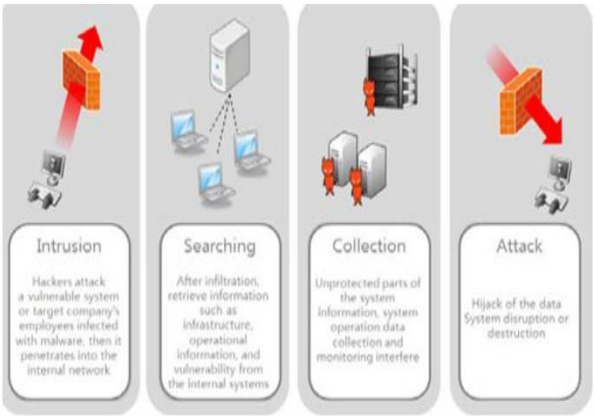

Fig1**.** The sequence of APT attacks

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.

2. Existing Security Technologies

Researchers developed various cyber security technologies to protect the system from threats and attacks.
Following are 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 theInternet -- or less-trusted networks -- such as a retail merchant's network outside of a cardholder data environment -- a firewall controls accessto 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 othertraffic 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 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 looks 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 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.

2. Network-based vs. Host-based systems: In a network-based system, or NIDS, the individual packets flowing through a network are analyzed. The NIDS can detect malicious packets that are designed to be overlooked by a firewalls simplistic filtering rules. In a host-based system, the IDS examines at the activity on each individual computer or host.

3. Passive system vs. Reactive system: In a passive system, the IDS detects 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.

III. BIG DATA SYSTEM MODEL

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 tothese attacks are needed. The new paradigm requires big data analysis techniques as a core ofdefense technologies,central security management, incident prediction technologies. We propose a system model that uses big data analysis technology for extracting data from various sources to react to previously unknown attacks.

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. The data in it will be of three types.

- **Structured data**: Relational data.
- **Semi Structured data**: XML data.
- **Unstructured data**: Word, PDF, Text, Media Logs.

DOI 10.17148/IJARCCE.2015.41252

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Big data analysis uses various existing analysis techniques such as machine-learning, artificial-intelligence, data-mining and etc.

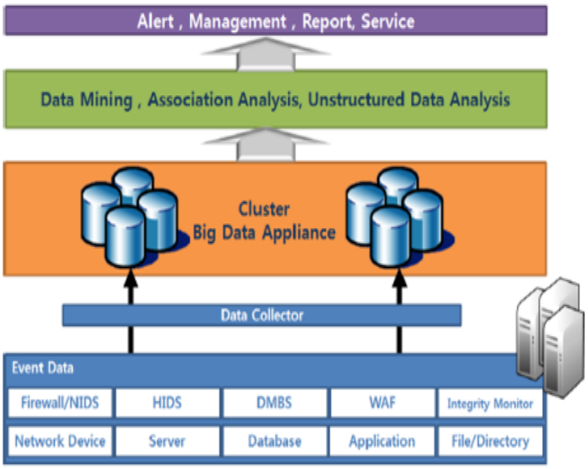

Fig 2.Big Data Analysis System Architecture

The entire system is divided into 4 steps.

• Data Collection: Data collection step collects event data from firewalls and log, behavior, status information (date, time, inbound/outbound packet, daemon log, user behavior, process information etc.) from anti-virus, database, network device and system. Collected data is saved in big data appliance

• Data Processing: This step validates whether collected data satisfies certain requirements. Then key value pair is created and classified using No-SQL, Hadoop and Map reduce and etc. It is known that approximately 80 per cent of time required for collecting and processing data using data mining isneeded. For faster processing, we introduce cloud or distributed system.

• Data Analysis: Pre-processed data from previous step is analyzed using prediction, classification, association analysis, and unstructured data analysisto decide user behavior,system status, packet integrity and misuse of file or system.

• Result: If attack or abnormal behaviors are detected, it alarms the administrator and terminates. Moreover, we provide dashboard, management tools to monitor results in real time. Prediction information of analyzed system is summarized and reported to the manager.

IV. HADOOP – TOOL FOR ANALYSIS

Hadoop is a software framework for storing and processing Big Data and work under Big Data Analytics. It is an open-source framework build on java platform and aimed at to improve the performance in terms of data processing on Big Data.

Features of Hadoop:

- Hadoop has two core components: HDFS and Map Reduce. HDFS is used for storing huge data sets while Map Reduce is used for processing these huge data sets.
- Hadoop consists of multiple concepts like COMBINER, PARTITIONER HBASE, PIG, HIVE, SQOOP to perform the easy and fast processing of huge data sets.
- Hadoop is different from Relational databases and can process the high volume, high velocity and high variety of data i.e. Big Data to produce result.

In a Hadoop cluster, data is distributed to all the nodes of the cluster present on which data is stored as shown in fig. 4.Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others. In short, Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

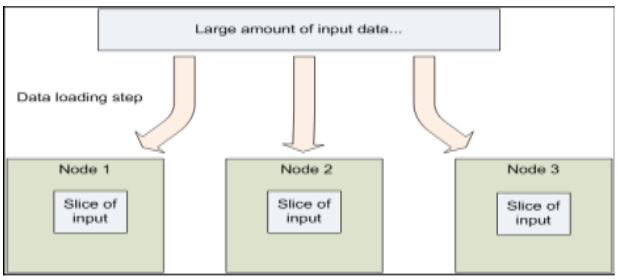

Fig 3.Hadoop Cluster

The two core components of Hadoop are:
A) HDFS
B) Map Reduce

A)HDFS:
The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities withexisting distributed file systems.However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets.

The Hadoop Distributed File System allows individual servers in a cluster to fail without aborting the computation process by ensuring data is replicated with redundancy across the cluster. There are no limits on the data that HDFS stores as it can be unstructured and schema-less.

B) Map Reduce:
MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodityhardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework.

Hadoop minimizes the processing time as the files are distributed across different nodes in the cluster and these nodes work parallel thereby minimizing the processing time and increasing the performance. Programs must be written to a particular programming model, named "Map Reduce." In Map Reduce, records are processed in separately by isolated tasks called Mappers. The output from the Mappers is then moved together into a next set of tasks called Reducers, where results from different

mappers can be merged together after shuffling and sorting as shown in fig. 5.

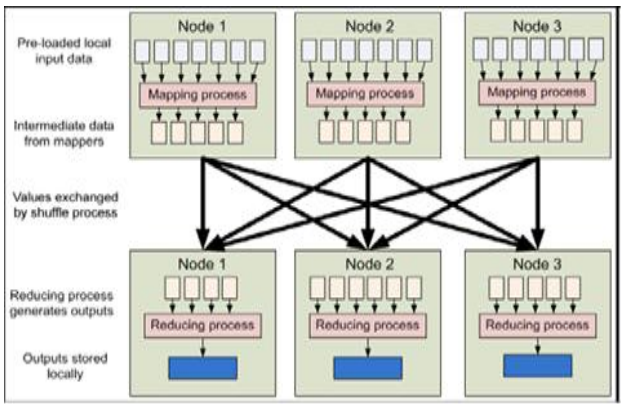

Fig.5. Mapping of records using Hadoop Distributed File System

V. ISSUES

Although a lot of research is going on big data but still many concepts are still to be researched. Researchers would try to enhance security platform to improve ability of software to find advanced threats, react accordingly and would developpreventive measures for future. Researchers would try to improve quality and reliability of security system. Some researchers are planning to taken up data collection, pre-treatment, integration, Map Reduce and analysis using machine learning techniques. They would use the results for securing and implementing preventive measures from threats to enterprise data.

VI. CONCLUSION

In this paper, we propose the use of Big Data Analytics for developing an Intrusion Detection System for detecting unknown attacks. We discussed a framework based on Hadoop for dealing the targeted attacks using Big Data Security Analytics. We can manage the Big Data characteristics of large volumes of enterprise data. Recent unknown attacks easily bypass existing security solutions by using encryption and obfuscation. Therefore there is a need to develop 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.

ACKNOWLEDGMENT

This work was supported by **Prof.DeepaA.Amne**, Assistant Professor, Department of Computer Science &Engg, BIT, Ballarpur and for being a constant source of inspiration.

REFERENCES

[1] "Advanced Persistent Threat: A Decade in Review", Command Five Pty Ltd, June, 2011.
[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] Tai-Myoung Chung Sung-Hwan Ahn, Nam-Uk Kim. "`Big data analysis system concept for detecting unknown attacks"'. Technical Report, February 2014
[5] 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.
[6] R. Magoulas and B. Lorica, Introduction to Big Data, Release 2.0 (Sebastopol Reilly Media, February 2009