DOI: 10.5769/C2013004 or http://dx.doi.org/10.5769/C2013004

Analyzing Targeted Attacks using Hadoop applied to Forensic Investigation

Parth Bhatt[1], Edgar Toshiro Yano[2]
Dept. of Electronics and Computer Engineering
Instituto Tecnológico de Aeronáutica
São José dos Campos,SP,Brasil
parthbhatt09@gmail.com[1], yano@ita.br[2]

Abstract—**Conventional intrusion detection and prevention technologies are mostly based to work on traditional methodologies to detect malicious events, while mining on a midsized log data. In recent years, we have seen the evolution of sophisticated targeted attacks performed by well trained adversaries exhibiting multiyear intrusions; therefore existing security toolsets have become insufficient for analysing targeted attacks with necessary speeds and agility.**

Dealing with such sophisticated attacks requires working with huge volume of multiyear security log data. Big Data technologies, such as Hadoop, enable the analysis of large and unstructured data sources, therefore, in this paper we propose our framework based on Hadoop for dealing with Intrusions performed by Targeted threat adversaries, using concept of Intrusion kill chains which will be helpful for forensics analysis.

Keywords-targeted threats ; Hadoop; intrusion kill chain;

I. Introduction

Due to rapid development of internet, all connected and unconnected machines everywhere around the world can be somehow reached and hence can become targeted for malicious purposes. Such activities of selection of targets and launching attacks directed towards a specific target are known as targeted attacks.

Although this term is not new in the area of Computer Security but due to recent advancement and increase in number of such sophisticated attacks, they have become the subject of high alert for every government, organization and industry. The threat of this class are targeted groups or individuals who are willing to spend extra effort to gain their objectives.[1]

A subset of targeted threat that is majorly famed by security industry in these recent years is Advanced Persistent Threats (APT). APT, as defined by a research paper from Lockheed Martin Corporation [2] are "well-resourced and trained adversaries that conduct multi-year intrusion campaigns targeting highly sensitive economic, proprietary, or national security information". These adversaries use sophisticated techniques to elude most of the contemporary security defence mechanisms and once being successful, aims to keep their persistency without getting detected inside their target environments.

Stuxnet was one such most complex APT, which was primarily written to target industrial control systems. The majority of infections were found in Iran specifically targeting programmable logic controllers of gas pipelines and nuclear power plants [5].Another variant of Stuxnet called duqu was recently found.

Operation Aurora aims to steal intellectual property of variety of technological, security and defence companies. A drive-by download attack was used to infect user's machine with a malware exploiting vulnerability [1]. Some other examples of such target attacks are given in case studies in [2], [4].

In this paper, we discuss a framework for dealing with Intrusions performed by Targeted threat adversaries, using kill chains. As described above, targeted attacks are getting advanced and more sophisticated, and for a typical organization dealing with critical information, there may be many adversaries attacking during a common time frame. In such a case, when finding correlation between detected suspicious events is already a challenge, it becomes really important to examine attack patterns to classify them into different groups of adversaries. This problem can be addressed by using intrusion kill chain method [2] which can help in situational awareness, intrusion correlation and intrusion prevention for future attacks.

APT adversaries generally aim to remain persistent inside their targeted environment, and they are likely to continue more intrusions, and these intrusions can be distributed in a big time frame (for example some months or years, as explained in the case study in [2]) . Thus, security log data collected from different sources in any such time period cannot be filtered or discarded, because any intrusion detected in future can be useful to find undetected malicious patterns of similar adversaries in the past. Furthermore, security log data collected from different sources (Network Intrusion detection system or NIDS, Host Intrusion detection system or HIDS , server logs ,mail logs, error logs etc) in a long time period can be classified as unstructured big data[8,9].

In our experiments we successfully tested our framework, which uses Intrusion Kill chain method and Big Data technologies (such as HDFS and map reduce for efficient log management and faster information retrieval of Big Data [3]) can be the future trend for Intrusion detection and prevention systems applied to targeted threats.

The rest of the paper is organized as follows Section II describes related works, Section III is about intrusion kill chains and analysis , Section IV is about technologies used in the framework, Section V explains proposed framework ,Section VI is implementation and results , Finally section VII is about conclusion and future works .

II. Related Work

During the last decade, due to increase in number of sophisticated targeted threats and rapid growth in volume of data traffic, the landscape of analysing log data has drastically changed, as now working with multiyear log data has entered the category of Big Data problem [9].

J. Howes, J. Solderitsch, I. Chen & J. Craighead [8] proposed an analytical security model considering the security analytics using Big Data. Their architecture is directed towards dealing with operational concerns in security organizations that aim to use existing security tools with Big Data analytics. Since their work is aimed towards operational side of security analytics therefore, it does not demonstrate any methodology of practical analysis of security threats as compared to our framework.

J. Therdphapiyanak and K. Piromsopa[3] used Hadoop map reduce model to analyse high volume of log files from server and distributed intrusion detection system and they proved that their frameworks performance was better than a standalone intrusion detection system .They were able to extract important information from the large security logs using their analysis and scalability of Hadoop, but their work was limited to use of K-means clustering algorithm of Mahout for detection of the deviated behaviour Clusters from normal behaviour Clusters. Using the proven capabilities of Hadoop for log analysis as in [3], our proposed framework is directed towards practical analysis of dealing with Targeted threats.

In [16], a blind automatic method for detecting malicious activities in honeypot data is proposed which uses RADOI to successfully identify attacks without any human intervention, while in [14] and [15], blind automatic detection for distributed honeypot data is proposed.

Authors could not find any other academic work that uses practical approaches to deal with detection of targeted threats using Big Data Technologies.

III. Intrusion Kill Chains

To understand the behaviour of targeted threat adversary is a significant problem as it generally deviates some attack vectors for each intrusion [2]. Our event logging systems can capture most of the system events but still correlation of similar group of events to identify behavioural characteristics of an adversary is of great importance, which can be addressed by Intrusion kill chain.

Intrusion kill chain model as given by Eric M. Hutchins, Michael J. Cloppert, and Rohan M. Amin [2] from Lockheed Martin Corporation is a series of phases that an attacker inescapably follows to model and carry out his intrusion. The intrusion kill chain phases are as follows:

The first phase is Information Gathering which involves selection of targets, collecting information about the target, for example searching emails, technologies the target uses, people on which their target trusts. The very next step an attacker will take is Weaponization which is coupling of malicious code with unsuspected deliverable files such as pdfs, docs, ppts and etc. Next in the third phase the attacker delivers the weaponized file to its target environment The most common delivery vectors are email, drive by download through a website link or through USB removable device. Once the malicious weaponized file gets successfully delivered in its target environment, the use of the vulnerability of the target system is taken to execute its malicious code, thus this phase is called as Exploitation

Next, the most important and crucial phase of the Kill chain is the installation of the malicious code inside the target environment .Remote Access Trojan's (RAT) are generally installed which allows adversary to maintain its persistence in the targeted environment.

The second last phase of the intrusion kill chain is Command and control (C2), in this phase the installed trojan or other malicious code generates a communication channel to control its execution and continue its actions to achieve its target

Actions is the last phase of the kill chain in which adversary achieves its objectives by performing activities like data exfiltration.[2] Defenders can be confident that adversary achieves its goal only after passing through all these phases.

Figure 1: Intrusion Kill Chain

After understanding the Intrusion kill chain phases, we need some methods to proceed towards construction or completion of kill chains once an malicious event is found. The following approaches can help us to deal with kill chain Construction.

a) Intrusion reconstruction: when a certain malicious event is detected and its phase is identified, analyst can be sure that the prior phases have been executed successfully [2]. Intrusion reconstruction is done by discovering the previous phases of the kill chain as those phases must have been taken in order to reach the detected phase. This can help defenders to mitigate the future intrusions and to understand the adversary's method of attacking.

b) Intrusion Synthesis: It is important to estimate what might have happened if defenders did not mitigate the intrusion on time. If such measure is not taken then, there is a chance that same type of attack may go undetected in future intrusions [2]. If defenders are able to collect more and more information about the kill chain, they can maintain an advantage over their adversary.

c) Campaign analysis: It consists of analysing multiple correlated intrusion kill chains expected to be from similar adversary over a long period of time (i.e. months or years of intrusion activity). Attacking persistently is an inherent disadvantage for the adversary which can be a great opportunity for defenders to identify the intrusion behaviour and improve their detection for future attacks. Re-using tools and techniques for intrusion is important for adversary to be quick in next intrusion and cost effective. Furthermore, campaign analysis can be very important to identify the adversary's target person or technology [2]

IV. Technologies used in the framework

Although explaining Hadoop and related technologies in details is out of scope of this paper but we provide a brief overview of technology terms that we use in this paper.

a) Hadoop: Apache Hadoop is a framework that allows distributed processing of large collection of data using cluster of computers each having local computation and storage [10]. Hadoop provides high availability, fault tolerance and faster processing speeds of large (structured, semi-structured or un-structured) data sets even with cheap commodity hardware.

Two main modules that Hadoop provides are HDFS and Map Reduce. HDFS is Hadoop distributed file system, which distributes the files across the cluster to provide high-throughput & fault tolerant access. Map Reduce is a programming model for distributed data processing. [10, 11]

b)Hive: It is a data warehouse system for Hadoop, it provides SQL like language HiveQL which becomes comfortable to start working, as SQL knowledge is widespread [12].

c)Pig: "Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs" [12].

d)Flume: "Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data" [7]. It helps to transfer data fault tolerantly from different applications to Apache Hadoop's HDFS.

e) OSSEC: "OSSEC is an Open Source Host-based Intrusion Detection System that performs log analysis, file integrity checking, policy monitoring, root kit detection, real-time alerting and active response [6]."

V. Proposed framework

We are going to address Targeted Attack intrusion management using kill chain approach and test its efficiency using Big Data technologies.

Our proposed framework aims to provide practical implementation to kill chain reconstruction, synthesis and campaign analysis as explained above in this paper using a Hadoop Cluster and Malware Analysis Lab. This framework can be help for management of intrusions for forensics analysis of targeted attacks.

The idea behind using a Hadoop cluster becomes clear when we aim to use large amount of security logs(semi or unstructured logs in text files) from different sources (distributed HIDS, NIDS, server ,mail and etc) collected in huge time frame (1-2 year or more). As explained above, targeted attackers persistently attack on their target environment therefore; using Hadoop cluster gives an advantage for extracting useful information from a huge log data set for campaign analysis.

To simplify the framework we divide it into 5 modules namely, Logging Module, Log Management Module, Malware Analysis Module, Intelligence Module and Control Module.

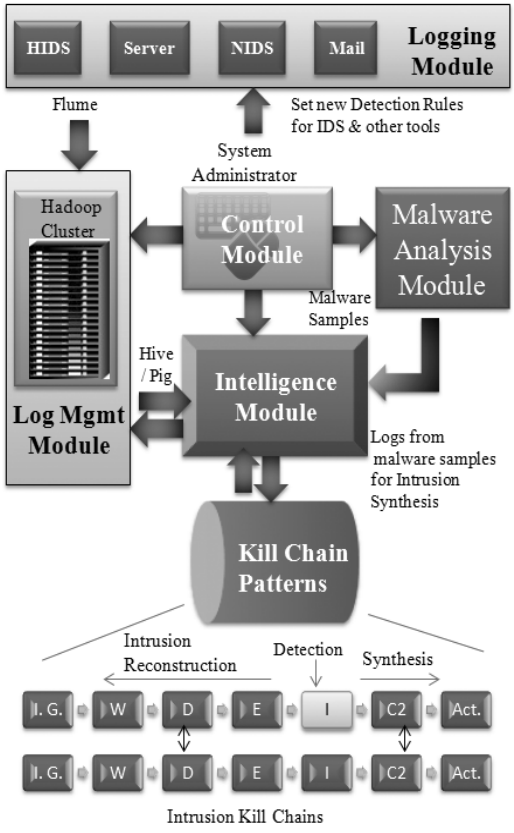

Figure2. Overview of Complete framework

A. Logging Module

This module consists of collecting logs from different sources such as distributed HIDS (Host intrusion detection system) and NIDS (Network intrusion detection system), Web Server Logs, Mail logs and etc. All these logs are generally text heavy semi structured or unstructured data. The rules for IDS s are written by the System administrators who has his hands on control module , these rule set will vary according to the situation and be learned from the Analysis of kill chains.

B. Log Management Module

This module consists of Hadoop distributed file system. The log data sets collected by the logging module are moved into HDFS and can be pre-processed here according to the needs of Intelligence module that would access Hadoop for information extraction. Services like Apache Flume can be used to move such large data into HDFS. Size of Hadoop cluster can vary according to the needs of the organization and size of data to be processed.

C. Intelligence Module

This is the main component of the framework which is responsible for construction and completion of kill chains, correlation between kill chains for campaign analysis and making new rules for Logging module. The suggested rules by Intelligence module can be further checked by administrator and then implemented.

Firstly, Intelligence module will identify the kill chain phase for each trigger event and start with construction of kill chains and as soon as it gets the detected suspicious malware information it will alert the malware analysis module for analysis of suspicious sample.

Trigger events: The trigger events are the events on which the Intelligence module will initiate the kill chain construction for the suspicious events occurred. Trigger events can be rule based or a system administrator input. Generally, a trigger could be a NIDS or HIDS high risk alert and can also be obtained for example: as the deviated behaviours from the Apache server logs after Clustering [3].

As explained earlier that whenever a certain trigger event is detected and found to be a phase of a possible kill chain, then one can be sure that prior events have been already occurred, on this basis Intelligence module can start relating events and identifying the previous phases.

Furthermore, intrusion synthesis can be done after getting log information from the malware analysis module about the malicious delivered payload.

After the intrusions kill chain in completed it can be further analysed and correlated with other kill chains to get more information about the adversary's target and attacking techniques. This process is for the campaign analysis as explained before.

In case, there are some kill chains which could not be completed then, either they are false positives or there is a need of detailed analysis; such cases should be reported to the system Administrator. System administrator can manually decide what type of actions should be taken about the situation. At any stage of kill chain construction and analysis, system administrator can correct or change any information, to improve the automated process of Intelligence module.

Another interesting property of this framework is that it a "Self Feeding framework" as it extracts information from the given resources and from the extracted information it extracts more information to produce highly precise results.

D. Malware Analysis Module

Malware analysis module consist of a malware analysis virtualized Lab Environment (virtual machines configured with different types of event logging tools). One of the most important threat component of targeted attack is malware,[4] therefore Administrator or an Analyst would need detailed knowledge about malware analysis. Although explaining malware analysis in detail is out of scope of this paper, the primary approaches for malware analysis are Code Analysis and Behavioural Analysis [4]. There are varieties of tools that help to perform such analysis of executables. Selection of tools for the analysis depends on the analyst. The primary goal of this module is to examine the suspicious sample. In case found malicious, detailed analysis should be performed and corresponding behavioural log information should be returned to the intelligence module to complete the kill chain synthesis.

E. Control Module

Using control module, the administrator governs the complete framework. The administrator is capable to drive the system in the direction he wants the system to proceed his investigation. This explains how the administrator can use this framework for digital forensic purposes against targeted attacks.

The control module is capable to control all other modules namely, Logging module for setting rules, Log management module for managing the cluster and formatting of logs, intelligence module for maintain the investigation in the right direction and the Malware analysis module for examination of the suspicious samples

VI. IMPLEMENTATION AND RESULTS

In this section we discuss our primary experiments and results about kill chain reconstruction based on search/ correlation algorithm programmed in Java to work with Hive using Hive thrift service on Hadoop.

A Hadoop Cluster was implemented in an academic environment using 5 nodes, which were powered by Intel® Core™ 2 Duo CPU E4500 @2.20Ghz×2 with 2GB of RAM, 80GB Hard Disk, 32bit machines forming a homogeneous cluster. As according to some sources it is not advisable to have heterogeneous cluster therefore, for a better performance analysis of our experiments we preferred to not increase cluster and to let it homogenous as no more same configuration machines were available. A Fast Ethernet Switch was used for the networking within the cluster nodes.

As, persistent targeted attacks are very rare in academic environments therefore, we collected Apache logs, OSSEC logs, Snort logs & mail logs generated at our university for primary experiments and these logs were further simulated according to targeted attack scenarios to perform the tests for Intrusion Reconstruction .

Simulation of a targeted attack using Targeted malicious Email.

A scenario of Targeted phishing Email was created, where attackers sent a phishing Email with attached malicious pdf to two university employees. The Email was well crafted and disguised as an authorized Email from ICoFCS 2013 Conference about invitation to Call for papers. When the pdf is downloaded and opened a benign pdf is extracted and showed to the user while another hidden malicious executable was extracted named as wp8.exe . Corresponding Log entries were simulated in the logs and are put into Hadoop. The intelligence module program was allowed to run over them. In the following we show how our program responded to such events

On June 10, OSSEC detects a malware getting installed into one of the hosts. The log entry about this event is fed into the intelligence module as a trigger event. Upon the reception of the trigger event by Intelligence module, intrusion reconstruction function is invoked that tags it to the Installation phase of kill chain and proceeds as already programmed for this kill chain phase.

Next, it searches in the logs for the location of the malware executable detected by OSSEC, after getting the location "C:\ Users\Master-Infoway\Documents\wp8.exe"; the intelligence module runs another query about timestamp and application that created this executable in the logs of past few months (considering the case that some malwares are intelligent enough to become dormant for some duration of time).

It finds out that the executable "wp8.exe" was created on May 25 upon execution of a pdf. "icofcs.pdf" located at "C:\ Users\Master-Infoway\Desktop" and created on May 25.

This time intrusion reconstruction function starts searching for delivery phase of this kill chain. Delivery is generally made by drive by download, targeted malicious email or USB [Hutchins , Amin & Cloppert 2011]. Finally it searches in mail logs and finds that icofcs.pdf was an email attachment to two employees of the university.

Table 1: Kill Chain Analysis

Info. Gathering	Mailing List ,ICoFCSWebsite
Weaponization	Malware Analysis Lab
Delivery	adm@icofcsconference.com ip : 161.xyz.pq.35 Sub: ICoFCS conference 2013 ICoFCS2013.pdf
Exploitation	0-day PDF
Installation	Malicious File detected " wp8.exe"
C2	Left for intrusion Synthesis phase
Actions	-

This completes the intrusion construction of the Kill chain using Hadoop and Hive queries accessed using our Java Program. Total number of log records fed into Hadoop were 7,049,627 and 5 Hadoop nodes with configuration mentioned above processed it completely in 2 minutes and 12 seconds .Using 5 node Hadoop cluster, we were able to process huge amount of semi-structured logs, Hive queries run the map reduce on Hadoop and the tasks are distributed across the cluster, finally quickly fetching the results.

Comparision with other Log Analysis Technologies

Our framework is completely based on hadoop platform which uses distributed processing of logs but generally all other popular log analysers/intrusion detection systems such as Snort work only on standalone machines. Thus, our framework takes an advantage of utilizing capability of Cluster of machines and bring results faster in comparison to standalone machine based systems. Additionally, our system has no limits to log data size as cluster sizes can be increased dynamically but the capacity of a standalone machine are fixed.

Snort is capable of giving an alert event from log data but our framework gives an alert on finding a pattern of events that form an Intrusion Kill chain.

In general sense our framework is only a log analyser which detects the presence of Intrusion kill chains but technologies such as Snort are intrusion detection and prevention technologies. For example our framework is not capable to detect atomic malicious events.

VII .CONCLUSION AND FUTURE WORK

In this paper, we discussed a framework based on Hadoop for dealing with Intrusions performed by Targeted threat adversaries, using concept of Intrusion kill chains. We simulated a realistic scenario of targeted attack and our framework could detect it using intrusion reconstruction through different sources of semi-structured logs.

The proposed framework has some of the major contributions such as:

- This framework can help in identification of targets, techniques and tactics of the adversary which is useful for forensics analysis of targeted attacks.
- Kill chain construction can help the administrators to build IDS rules to strengthen their posture of defence.
- Other than detecting and analysing targeted attacks, this framework can also help its administrators to identify unknown vulnerabilities (also called 0 day vulnerabilities) in their system that the attacker used.

Although, this framework is greatly promising and well structured for dealing with targeted threats but still it contains some limitation such as following:

- This framework uses Hadoop for managing the log files, while Hadoop is a perfect framework for working with unstructured and semi structured text heavy data sets but, it is not good fit for real time applications and small amount of data set therefore, this deficiency of Hadoop makes the framework slower in response for small data sets in comparison to other relational database systems.
- Classifying of kill chains from common malware to targeted malware, this framework will give some effort for administrator to differentiate a target malware or a common unsophisticated but on the other side, this can be used to analysis of common malwares also.

Using our experiments we successfully tested our framework, which uses Intrusion kill chain method and big data technologies (such as Hadoop HDFS and map reduce for efficient log management and faster information retrieval from semi-structured big data). Finally, according to our analysis, using intrusion kill chain method and Big Data technologies can be the future trend for Intrusion detection and prevention systems applied to targeted threats such as Targeted malicious Emails.

In future, we plan to implement more typical targeted threat scenarios and analyse them with bigger homogenous Hadoop cluster and evaluate its efficiency. We also intend to implement automated correlation of Kill chains for Campaign Analysis.

REFERENCES

[1] A.K. Sood, R.J. Enbody " Targeted cyber attacks: A Superset of advanced persistent threats" Security & Privacy, IEEE Volume 11 , Issue 1 ,pages 54 – 61, Jan.-Feb. 2013

[2] Eric M. Hutchins, Michael J. Cloppert, Rohan M. Amin, "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains" ,6th International Conference on Information Warfare and Security(ICIW2011) http://www.lockheedmartin.com/content/dam/lockheed/data/corporate/documents/LM-White-Paper-Intel-Driven-Defense.pdf

[3] J. Therdphapiyanak, K. Piromsopa, "Applying Hadoop for log analysis toward distributed IDS", ACM International Conference on ubiquitous Information Management and Communication (ICUIMC) 2013, Article No. 3

[4] Frankie Li, A Atlasis , " A Detailed Analysis of an Advanced Persistent Threat Malware" SANS Institute InfoSec Reading Room,2011 http://www.sans.org/reading_room/whitepapers/malicious/detailed-analysis-advanced-persistent-threat-malware_33814

[5] N. Falliere, L.O. Murchu, and E.Chien "W32.Stuxnet Dossier" Version 1.4 (February 2011)

[6] OSSEC "http://www.ossec.net/"

[7] Flume http://flume.apache.org/

[8] J. Howes, J. Solderitsch, I. Chen & J. Craighead, "Enabling trustworthy spaces via orchestrated analytical security", ACM, CSIIRW 2013, Article No. 13

[9] MacDonald, Neil, 2012, Information Security is Becoming a Big Data Analytic Problem, Gartner, (23 March 2012), DOI= http://www.gartner.com/id=1960615

[10] Apache Hadoop Project http://hadoop.apache.org/

[11] Tom White "Hadoop: The Definitive Guide", 2009, 978-0-596-52197-4

[12] Hive"http://hive.apache.org/", Apache Pig http://pig.apache.org/

[13] "Hadoop Tutorial from Yahoo!", Module 7: Managing a HadoopCluster http://developer.yahoo.com/hadoop/tutorial/module7.html#machines

[14] J. P. C. L. da Costa, E. P. de Freitas, A. M. R. Serrano, and R. T. de Sousa Jr. "Improved Parallel Approach to PCA Based Malicious Activity Detection in Distributed Honeypot Data," International Journal of Forensic Computer Science (IJoFCS), 2012

[15] B. M. David, J. P. C. L. da Costa, A. C. A. Nascimento, M. D. Holtz, D. Amaral, and R. T. de Sousa Jr., "A Parallel Approach to PCA Based Malicious Activity Detection in Distributed Honeypot Data," International Journal of Forensic Computer Science (IJoFCS), 2011

[16] B. M. David, J. P. C. L. da Costa, A. C. A. Nascimento, M. D. Holtz, D. Amaral, and R. T. de Sousa Jr., "Blind Automatic Malicious Activity Detection in Honeypot Data," The International Conference on Forensic Computer Science (ICoFCS) 2011, Florianópolis, Brazil