A Hadoop based Framework to support Cybersecurity Situational Awareness

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]

***Resumo* — Cybersecurity Situational Awareness (CSA) involves the perception of current situations of different cybersecurity elements, the comprehension of the meaning of the current cybersecurity situation regarding the impacts on the organization assets, and the projection of future status in order to select a better positioning of defence mechanisms. Current techniques to CSA are limited by the high speed of events generation, the large volume of information from multiple sensors, and the complexity of interactions of highly automated services that shape the Cyberspace. In order to support CSA activities, a Hadoop Based Framework was developed. This framework offers a faster correlation of event logs (from different security mechanisms and over a large time period) so that is possible to identify a persistent and complex attack before it causes a great damage.**

***Palavras-Chave* — Defesa Cibernética, Consciência Situacional Cibernética, Hadoop.**

I. INTRODUCTION

Cyber-attacks are increasing and evolving. Each cyber-attack may take multiple steps to gain privileges and penetrate deep inside in a well protected network. To advance each step, an attacker must successfully explore vulnerabilities without being detected by current defense mechanisms. The top priority for cybersecurity is protecting against exfiltration and manipulation of information. To attend this priority a robust defense mechanism is required, with detailed knowledge of what is happening in the different levels of the enterprise networks and components.

Situational Awareness (SA) is formally defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" [1]. SA involves *Perception* of critical factors in the environment, *Understanding* what those factors mean, and *Projection* of what will happen with the system in the near future. Extending Endsley SA definition, Cyber Situational Awareness (CSA) can be defined as the knowledge of current status of security of all network assets, the understanding of this status regarding current threats, and forecasting the future status in the near future. From a military perspective, CSA means knowing the level of threat and the current status of all network assets to support the military operations [2]. This includes awareness of availability, confidentiality, integrity status of the C2, logistics, communications, IT systems, etc. CSA requires a proactive approach that enables cyber defenders to react to events in real-time, instead of being constrained to analyzing occurrences after the fact.

The nature of attacks in Cyberspace is changing. The attackers are avoiding attacking directly systems and services that are normally well protected. They are directing the attacks to specific users of a targeted organization. Activities of selection of targets and then launching attacks on such selected aims are known as Targeted attacks.[3]. Targeted attacks are superset of many other types of attacks that focus on specific aim. One such category is Advanced Persistent Threat or APT, which has been recently publicised by security industry as defined by a research paper from Lockheed Martin Corporation [4], they are "well-resourced and trained adversaries that conduct multi-year intrusion campaigns targeting highly sensitive economic, proprietary, or national security information". These attackers generally use many sophisticated and zero day vulnerabilities of the target system to escape from most of the security mechanisms, once inside the system, they try to maintain their conquests inside the target environment.

In order to succeed in this new scenario, CSA requires the capability to quickly collect and synthesize enormous volumes of data that comes from multiple sensors over large time periods, recognize patterns and dynamically adapt at machine speed. To attend this requirement, we developed a framework using Apache Hadoop to support the collection and synthesis of large volumes of logs. But a large scale data collection and analysis capability is not enough to deal with this challenge. It is necessary a conceptual model to clarify the conduct of cyber attacks. CSA involves far more than simply perceiving information in the cyberspace. It includes comprehending the meaning of that information to identify threats or potential threats, and forecasting future states to identify threats real impacts and motives. We adopted the Intrusion Kill Chain Model [4] as a conceptual model to support CSA activities. Intrusion Kill Chain is a series of phases that an attacker inescapably follows to carry out an intrusion. Our developed framework supports Perception by collecting security logs and identifying a potential Intrusion Kill Chain. It supports Understanding by a process of Intrusion Reconstruction identifying signs that a Kill Chain is confidently occurred or is ongoing. It supports Projection by a process of Intrusion Synthesis that identifies the current attack status and possible future actions.

In this paper, we are going to discuss our developed framework to support CSA activities. In the following sections we present Hadoop framework in section II, Intrusion Kill Chain Model in section III, our developed framework in section IV, current implementation in section V, related works in section VI and conclusions and proposals for future works are in section VII.

II. HADOOP AND RELATED TECHNOLOGIES

Apache Hadoop is an open source framework that allows distributed processing of large collection of data using cluster of computers each having local computation and storage. 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. [5]

Two main modules that Hadoop provides are HDFS and MapReduce. HDFS is Hadoop distributed file system, which distributes the files across the cluster to provide high-throughput & fault tolerant access. MapReduce is a programming model for distributed data processing. MapReduce can take the advantage of locality of data, processing data on or near the storage assets to decrease transmission of data. In the "Map" step, the master node takes an input, divides it into smaller sub-problems and distributes them to worker nodes. In the "Reduce" step, the master node collects the answers to all the sub-problems and combines them in some way to form the output. Hadoop offers scalability, cost effectiveness, flexibility and reliability to support the analysis of large sets of unstructured data. Scalability because new nodes can be added as needed, without changing data formats. Hadoop is cost effective because it can use commodity servers. It is flexile because it is schema-less, and can process any type of data, structured or not, from any number of sources. And it is reliable because when a node is lost, the system redirects work to another location of data to continue the processing.

Hadoop is supplemented by othe Apache projects, such as Pig, Flume and Hive[6]. Pig is a platform for analyzing large data sets. It is made up of a programming language and a run time environment. Flume is a distributed, and reliable service for collecting and aggregating large amounts of log data. Hive allows SQL developers to write Hive Query Language (HQL) statements that are similar to SQL statements. HQL statements are translated by Hive into MapReduce jobs and executed in a Hadoop Cluster.

III. INTRUSION KILL CHAIN

Intrusion Kill Chain model as described in [4] 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:

- Information Gathering – Selection of targets, collecting information about the target, for example emails, technologies the target uses, people on which their target trusts.
- Weaponization – coupling of malicious code with unsuspected deliverable files like pdfs, docs, ppts and etc.
- Delivery- Transferring the weaponized file to the target environment.
- Exploitation - Use of vulnerability of target system to execute the delivered malicious code.
- Installation - Remote Access Trojan's (RAT) are generally installed which allows adversary to maintain its persistence in the targeted environment.
- Command and control (C2) - Adversary requires a communication channel to control its malware and continue their actions therefore the malware needs to get connected to its C2 server.
- Actions - it's the last phase of the kill chain in which adversary achieves its objectives by performing actions like data exfiltration.

Defenders can be confident that adversary can achieve its goal after passing through all these phases. To develop a proper analysis of kill chains the following methods can be used:

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 [4]. 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. The intrusion reconstruction contributes to clarify if an event is a false positive. If a reconstruction is successfully done it is a strong argument that an attack occurred or is ongoing.

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 [4]. If defenders are able to collect more and more information about the kill chain, they can maintain an advantage over their adversary.

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 [4].

IV DEVELOPED FRAMEWORK

Our developed framework aims to provide practical implementation to kill chain reconstruction, synthesis and campaign analysis as explained in this paper using a Hadoop Cluster and Malware Analysis Lab. This framework can be helpful for management of intrusions and detection 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.

The framework is structured into 5 modules namely, Logging Module, Log Management Module, Malware Analysis Module, Intelligence Module and Control Module. Using a CSA perspective, Logging and Log Management Modules support Perception Activities; Malware Analysis and Intelligence Module support Understanding activities; Intelligence Module and Control Module support Projection Activities and Intrusion Campaign Process.

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 logs with semi structured or unstructured in property. The rules for the IDS are written by the System administrators who has access to control module, these rule set will vary according to the situation and be learned from the Analysis of kill chains.

Log Management Module

This module consists of Hadoop distributed file system. The logs data sets collected from 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.

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 [7].

As explained earlier that whenever a certain trigger event is detected to be as a phase of kill chain, then one can be sure that prior events have 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 is reconstructed, 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 intrusion synthesis and campaign analysis as explained before.

In case, there are some kill chains which could not be completed then, either they are false positives or need a detailed analysis; such cases should be reported to the 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 is a "Self Feeding framework" as it extracts information from the given resources and from the extracted information it extracts more information to produce more precise results.

Malware Analysis Modules

Malware analysis module consist of a malware analysis virtualized Lab Environment with detection tools. One of the most important threat component of targeted attack is malware [8] 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 [8]. 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 and declare it, if malicious. In case found malicious, detailed analysis should be performed and corresponding log information should be returned to the intelligence module to complete the kill chain synthesis.

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 in 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 to maintain the investigation in the right direction and the Malware analysis module for examination of the suspicious samples

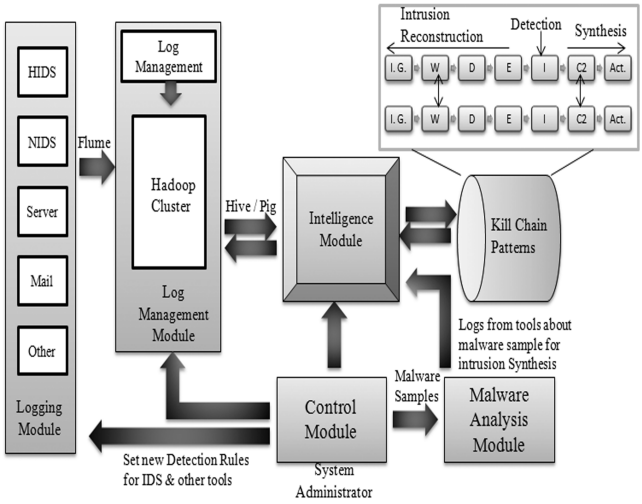

Fig. 1 Complete Model

The model can be the central structure for a CSA solution. It supports Perception by collecting security logs and identifying potential attacks. This activity is performed using Apache Flume that collects security events from different security solutions of the Logging Module and transfers these data to the Log Management Module. It supports Understanding by a process of intrusion reconstruction identifying signs that a kill chain is confidently occurred or is ongoing. This process is supported by the Intelligence Module and the Malware Analysis Module. It supports Projection by a process of intrusion synthesis that identifies the current attack status and possible future actions. This process is again supported by the Intelligence Module and interactions of the administrator provided by the Control Module.

V. IMPLEMENTATION AND RESULTS

In this section we discuss our experiments and results about kill chain reconstruction based on search/correlation algorithm programmed in Java to work with Pig and Hive. 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 and Campaign Analysis.

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 SIGE 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. 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. ”sige2013.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 [4]. Finally it searches in mail logs and finds that sige2013.pdf was an email attachment to two employees of the university. 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 given 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.

VI RELATED WORKS

Howes, Solderitsch, Chen & Craighead [9] 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.

Therdphapiyanak and Piromsopa [7] used Hadoop map reduce model to analyze 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 for log analysis and doesn`t use a conceptual model like the Kill Chain Model to detect targeted attacks.

VII CONCLUSIONS AND FUTURE WORKS PROPOSALS

The developed framework is part of an ongoing research project for the development of architectures with CSA capabilities. It can support Perception, Understanding and Projection activities with Scalability, Cost-Effectiveness, Flexibility, and Reliability provided by the Hadoop Framework. The adoption of the Intrusion Kill Chain Method to analyse attacks provided the following benefits:

- It can support the detection and analysis of targeted attacks, even the cases where the attackers explores Zero-day vulnerabilities, because it doesn't require the use of known signatures.
- The Intrusion Reconstruction process can identify false positives (A major limitation of IDS solutions).
- Quickly identifying an ongoing Kill Chain can potentially block the attack before it proceeds to the last phase.
- Kill chain construction can help the administrators to build IDS rules to strengthen their posture of defence.

Although, this framework is greatly promising and well structured for dealing with targeted threats, it still contains the following limitations:

- 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, 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.
- Although the framework provides support for Intrusion Reconstruction and Synthesis, this process is still highly dependent on the capabilities of a good administrator.

In future we plan to implement some more features to this framework such as detection of Distributed Denial of Service Attacks, using Mahout Clustering [10] for finding out anomaly cluster that could fed into the current framework as trigger events. Another research direction is the development of Visualization Environments that can improve the CSA capabilities that is still highly dependent on well trained administrators.

REFERENCES

[1] Endsley, M. R. Desing and Evaluation for Situation Awareness Enhancement. In Proceedings of The Human Factors Society 32 st Annual Meeting. 1988.

[2] Gregoire, M. Bedoin, L. Visualization for Network Situational Awareness in Computer Network Defence, NATO OTAN RTO-MP-IST-043, paper 20, 2005.

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

[4] Hutchins Eric M., Cloppert Michael J., Amin Rohan M,(2011) "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains" ICIW2011

[5] White, T., Hadoop: The Definitive Guide, Third Edition, 2012 O'Reilly

[6] Capriolo E., Wampler D. , Rutherglen J. , Programming Hive, 2012,First Edition, O'Reilly

[7] Therdphapiyanak J., Piromsopa K., "Applying Hadoop for log analysis toward distributed IDS" ACM ICUIMC 2013, Article No. 3

[8] Frankie Li, Atlasis A, " 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

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

[10] Owen, S. etalli, Mahout in Action, Manning Publications, Oct. 2011.

[11] OSSEC " www.ossec.net"