PROCEEDINGS OF SCIENCE

Cyber security monitoring and data analysis at IHEP

Tian Yan[*,ab] **Hao Hu,**[a] **Dehai An,**[a] **Fazhi Qi,**[a] **and Chen Jiang** [c†]

[a] *Computing Center, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, P.R.China*

[b] *Key Laboratory of Network Assessment Technology, CAS, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, P.R.China*

[c] *Legendsec Information Technology (Beijing) Inc., Beijing 100085, P.R.China*

E-mail: yant@ihep.ac.cn, huhao@ihep.ac.cn, adh@ihep.ac.cn, qfz@ihep.ac.cn, jiangchen@qianxin.com

Recently, cyber security threats becomes a noticeable challenge for academic institutes. In this paper, we present the security risk control model and the cyber security detection and monitoring system designed and deployed at Institute of High Energy Physics (IHEP) in China. Security data collection and analysis plays the central role in this framework. In addition to the open-source solution like Zeek, MISP and ELK stack, we also deployed a commercial Security Operation Center (SOC) as a supplement and cross-check solution.

International Symposium on Grids & Clouds 2019, ISGC2019
31st March - 5th April, 2019
Academia Sinica, Taipei, Taiwan

[*]Speaker.
[†]Corresponding author.

PoS(ISGC2019)011

1. Introduction

In recent years, along with the rapid development of large scientific facilities and e-science worldwide, various cyber security threats has becoming a noticeable challenge in many data centers for scientific research, such as DDoS attack, ransomware, crypto currency mining, data leak, etc.

Intrusion and abnormality detection by collecting and analyzing security data is an important measure for enhancing the sensitivity of security status perception, level of security protection, and agility of security incident response. However, as the scale of data center growing, it's difficult to use a single security box to process the large volume of various data generated by network traffic, device and host logs, threat intelligence, and so on.

In high energy physics (HEP) community, people are trying to establish a security operation center (SOC) for handle this problem. There is a SOC working group for the Worldwide LHC Computing Grid (WLCG). With help of this working group, we are building a cyber security monitoring and analysis framework at Institute of High Energy Physics (IHEP), Chinese Academy of Sciences. At IHEP, we have 4x10Gbps IPv4 and IPv6 dual-stacked internet connection, and 2x80Gbps inner data center network. There are also hundreds of web information servers, thousands of PC clients and thousands of computing nodes. It's a real in such an environment to handle the security related data generated by such a set of information assets.

In this framework, the Malware Information Sharing Platform (MISP)[2] is deployed for threat intelligence exchange with collaborating HEP institutes and universities. Network traffic is collected from switches and firewalls by a 10Gbps network shunt, and then flows to a Zeek (formerly called Bro) instance for traffic analysis[3]. Zeek logs and hosts/web logs, security device logs, along with vulnerability scanning results and assets detection results, etc., are defined as cyber security data. All of these data are collected by Flume/Logstash/Syslog to a data pipeline based on the Kafka[4] cluster principles. In this cluster, there are some Spark jobs running for stream processing, which are aimed at rapid intrusion and abnormality detection as well as data correlation and enrichment. Then all the processed data are written to Elasticsearch[5], MySQL and InfluxDB[6], and then visualized by Kibana[7] and Grafana[8]. At the same time, the processed data can be written to local storage, HDFS[9], or tap storage for backup and long-term analysis.

2. Cyber security threats and the risk control model

The web application systems and computing facilities faces various cyber security threats. In recent years, the main cyber security threats we face are: intrusion, crypto-currency mining, Ransomware, Phishing, DDoS. The historical statistics show that we have about a dozen of such security events every year.

To protect our IT assets from these cyber security threats, we followed a cyber security risk control model shown in Figure 1. It consists of four quadrants: prevention, protection, detection and response. For prevention, we follow the regulations and compliance, perform user security awareness and training, security policy review, security assessment and audit. For protection, we deploy firewall, Intrusion Detection and Protection System (IDS/IPS), Virtual Private Network (VPN) devices and Web Application Firewall (WAF) at the edge of our network. We also scan the web and host vulnerabilities regularly and fix up them as soon as possible, and block malicious

IPs, domains and URLs. For detection, we analyze the traffic and system/web logs with help of threat intelligence sharing by deploying both an open source and a commercial SOC. And finally, we have emergency response procedure and team for security incident response.

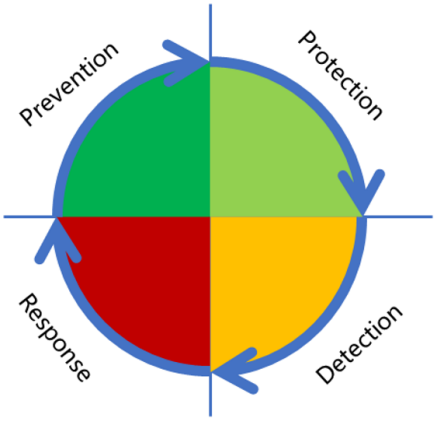

Figure 1: The cyber security risk control model

3. Cyber security detection and monitoring system

As the amounts of IT assets and cyber attacks grows rapidly these years, it's impossible to handle the security data generated everyday at our institute by a single IDS shipped in one machine. Therefore, we are trying to design and deploying a scalable system which can collect, store and analyze such a dataset.

3.1 Architecture

The work of WLCG SOC group [1] is a good start of our work. We take WLCG SOC architecture as a reference and design a simple data process architecture which is shown in Figure 2

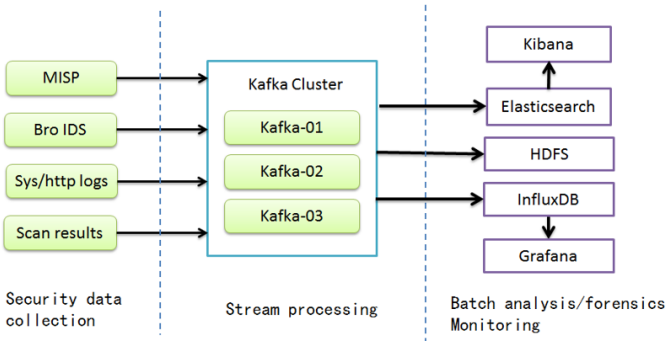

Figure 2: The architecture of the security detection and monitoring system

It can be divided into three parts. The left is the security data collection part, where different kinds of security data are collected from different sources. The middle is the data pipeline as well as the stream processing engine. The right side acts as data store and visualization.

3.2 Data sources

In the left part in Figure 2, as the security data sources, we have threat intelligence, traffic analysis logs, system logs and vulnerability scan results, etc. The threat intelligence is imported from the open source tool MISP [2]. By using MISP, we can share the intelligence with the HEP community as well as cooperating security companies. The network traffic of subnetworks are mirrored by a network shunt or SDN switch. An open source intrusion detection system named Zeek[3] is deployed to analysis the mirrored traffic. It is a flexible and powerful tool for traffic analysis and abnormality detection. Zeek can integrate the threat intelligence data from MISP as an input. The output of Zeek is a set of logs which records various aspects of the network activities. System logs of hosts and web applications are collected by a lightweight log collector called filebeat[10]. We also deployed Honeypot VMs to collect attack informations. The vulnerability scan results are collected from an open source scanning tool OpenVAS[11] as well as a commercial scanner provided by NSFOCUS, which is a domestic security provider.

3.3 Data analysis

The middle part of Figure 2, shows a Kafka [4] cluster built upon three physical machines each of which has 20 CPU cores, 128 GB memory, 8 TB disk storage and 10 Gbps network interface cards. The Kafka version 2.1.0 was installed for the real-time data pipeline and streaming processing. Some Spark[12] jobs are running on this cluster to perform the streaming processing, including data enrichment and rapid abnormality detection. The assets databases and security policy databases are stored in a relational database and act as input.

The right part of Figure 2 is data storage and visualization. The time series data are stored in InfluxDB[6] and displayed by Grafana[8]. Data of the most recent three months are stored in Elasticsearch[5] and are quickly searchable and can be statistically analyzed by Kibana[7], as shown in Figure 3. For long term data store and offline analysis, we use HDFS[9] and Spark.

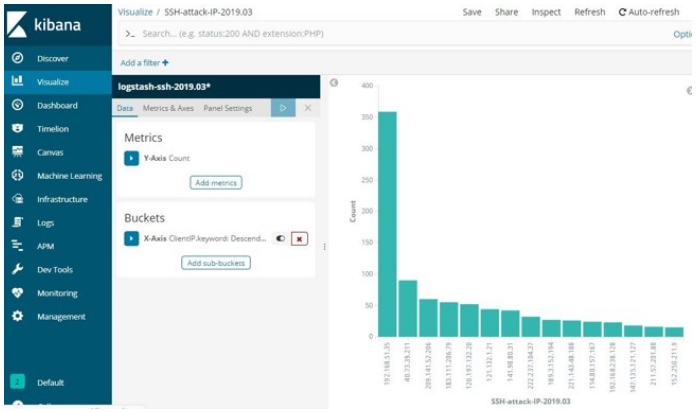

Figure 3: SSH log analysis with Kibana

The result of this data analysis is used in the visualization platform for security status monitoring, as well as used as input data for security operation. The security operations center is maintained by the security team: only that team may change the assets and policy databases.

3.4 Working with commercial SOC

We can benefit from a cross check of the results of an open-source SOC and a commercial SOC. NGSOC (Next Generation SOC) [13] is a commercial solution of SOC we chose to test, it produced by Qi An Xin Group, which is a domestic security company in China. Its major advantage is threat intelligence, as he Qi An Xin Group has about 1300 PB of reference and learning data sets on security. We start deploying and testing since Aug. 2018. All the inbound/outbound traffic are taken as input data source. It has already detected crypto-currency mining malware and web-shell in our servers.

The architecture of NGSOC is shown in Figure 4. At the center of this system is the analysis platform. It get input from the traffic sensor, log collect probe, malware sandbox, and correlation analysis engine. Each of these component is deployed in a seperated machine for scalability. The analysis platform can also communicate with the threat intelligence cloud which stores the intelligence obtained from their reference data.

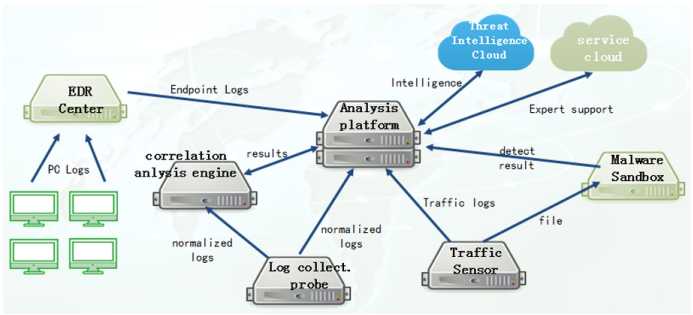

Figure 4: The architecture of NGSOC

The dashboard of NGSOC is shown in figure 5.

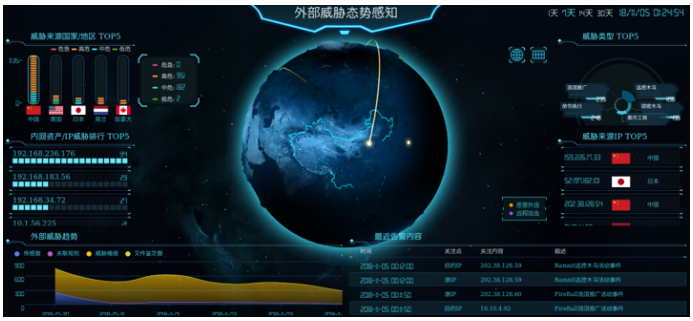

Figure 5: The dashboard of NGSOC

In this dashboard, an earth is located at the center for illustrating the security threat and the malicious connections between hosts in LAN and hosts in the internet. Around this earth is several kinds of statistics of the data collected.

3.5 Monitoring center

To visualize the security threats and the security status we get from the statistics of the collected data, we deployed a large display in a dedicated monitoring room. This monitoring center consists of 3×2 55 inch displays, as shown in Figure 6. These 6 displays can be grouped in several ways to present different monitoring demands. In this figure, we illustrated the dashboard of NGOSC and the network traffic statistics given by Grafana.

Figure 6: The IHEP security monitoring center

3.6 Comparision between the open source SOC and commercial SOC

According to our experience of operating NGSOC and the open source MISP/Zeek based SOC, we found both of them have merits and demerits. The commercial SOC is easier to setup, configure and maintain. It has fantastic monitoring dashboard and easy-to-use web user interfaces. So it is more friendly for non-expert users and these institutes which lack of manpower and experts on security. The second advantage of commercial SOC is that it has more comprehensive and up-to-date threat intelligence. On the other hand, the major advantage of open source SOC is flexibility. The system can be customized to fit the special needs of different application scenarios. Especially, Zeek has its only script language which can be used to writing new detecting partterns. The second advantage of open source SOC is that we can store and handle the security data in our data center, we don't worry about the data leak.

4. Conclusion

In this paper we described the cyber security monitoring and security data processing framework. With help of this security data collection and analysis framework, it is possible for us to handle the large amount of security data generated at IHEP, and it's also very flexible and scalable for even larger amounts of and different kinds of data in future.

Acknowledgments

This work is supported by an open project of CAS Key Laboratory of Network Assessment Technology and National Natural Science Foundation of China under grant no. 11675199 and 11775246.

References

[1] D. Crooks, et al, *Harnessing the Power of Threat Intelligence in Grids and Clouds: WLCG SOC Working Group*, in proceedings of *ISGC 2018 & FCDD*, PoS(ISGC 2018 & FCDD)012 (2018).

[2] The MISP project, https://www.misp-project.org/

[3] The Zeek project, https://www.zeek.org/

[4] The Apache Kafka project, http://kafka.apache.org/

[5] Elasticsearch B. V., "Elasticsearch", https://www.elastic.co/downloads/elasticsearch

[6] InfluxData, Inc., "InfluxDB", https://portal.influxdata.com/

[7] Elasticsearch B. V., "Kibana", https://www.elastic.co/downloads/kibana 3]

[8] Grafana Labs, "Grafana", https://grafana.com/

[9] The Apache Hadoop project, "HDFS", https://hadoop.apache.org/

[10] Elasticsearch B. V., "Filebeat", https://www.elastic.co/downloads/beats/filebeat

[11] The OpenVAS project, http://www.openvas.org/

[12] The Apache Spark project, https://spark.apache.org/

[13] The NGSOC product introduction, https://www.qianxin.com/product/ngsoc (in Chinese)