2016 IEEE 2nd International Conference on Big Data Security on Cloud, IEEE International Conference on High Performance and Smart Computing, IEEE International Conference on Intelligent Data and Security

OwlSight: Platform for Real-time Detection and Visualization of Cyber Threats

Vasco Samuel Carvalho
Globinnova, CIICESI/ESTGF/IPP
Felgueiras, Portugal 4610-108
Email: vasco.carvalho@globinnova.com

Maria João Polidoro
CIICESI/ESTGF/IPP
Felgueiras, Portugal 4610-156
Email: mjp@estgf.ipp.pt

João Paulo Magalhães
GCC, CIICESI/ESTGF/IPP
Felgueiras, Portugal 4610-156
Email: jpm@estgf.ipp.pt

***Abstract*—Security reports published by leading companies reveal the growing number of cyber attacks. Thefts of money or sensitive data, harm the reputation of organizations and sabotage of national critical infrastructures are some of the motivations behind these attacks. The sophistication of these attacks is very high, creating major challenges to the detection and mitigation in useful time. In this context the development of systems to provide situational awareness, to detect cyber threats and alert them in real-time are very important to mitigate the impact of the attacks.**

In this paper we present a cyber threat platform targeted for real-time detection and visualization of cyber threats. The platform is composed by several building blocks and it is able to collect huge amounts of data from multiple sources, prepare and analyze the data and present the findings through a set of insightful dashboards. A version of the platform is already available and used in a real-context. It collects more than 107 million of malware events daily from different data sources and provides visualization and alerts in real-time for more than 2.7 million of infected unique IPs spread around the world.

***Index Terms*—cybersecurity; threat intelligence; big data security; big data visualization; malware**

I. Introduction

Last years have witnessed a steady increase in the number, depth and breadth of incidents related to cyber attacks, both in government and private-sector organizations around the world. Examples include theft of sensitive data from defense companies and the military, attacks to media and broadcasting organizations [1], theft of millions of customer records, and huge losses in financial services companies due to online fraud and breaches in payment networks [2]. The motivations behind these attacks include state-sponsored espionage, financial gain and politically-motivated activism. The impacts have ranged from theft of strategic and highly valued intellectual property and direct financial loss to significant damage to brand and customer trust.

The growing number of incidents is a clear indication of the limitations of the traditional strategies for protecting information assets. The traditional approach to cyber security is built around an outdated 'fortress mentality', where organizations work to define a trusted environment for their data and networks, in which everything inside the environment is trusted; everything outside the environment is not. The fortress mentality implies that the way to stay secure is by striving to identify and fix all vulnerabilities before the attacker can find and exploit them. In today's complex systems and ever-changing threat scenario, this no longer holds true. Nowadays attacks can target organizations at anytime and any place. In this context, devise proactive systems able to detect cyber attacks and remediate them quickly as possible is very important.

In this paper we present a proactive cyber threat platform targeted for real-time detection and visualization of cyber threats. The platform is named OwlSight and it was conceived with the following goals in mind:

- data source agnostic;
- able to cope with the volume, velocity and variety of security related events;
- provide real-time detection of cyber threats with low false alarms;
- provide insightful visualization and analysis techniques;
- contribute to reduce the mean time to remediate.

OwlSight is a big data security platform composed by several building blocks. It gathers different types of security events from multiple sources, prepares the data, enriches the data and processes the data by means of big data analysis. It relies on elastic NoSQL databases and big data engines to collect millions of events per second of different formats and perform clustering analysis, correlations and summary statistics of the data to reveal security threats, that otherwise would be difficult and time consuming to detect in a useful time.

The security threats are presented by means of real-time alerts and insightful and contextual dashboards that promote the fast detection and mitigation of the incidents. We show through use cases how the platform can be used to discover, follow and provide real-time detection of malware communications affecting organizations worldwide. We also present results of a real utilization scenario.

The remainder of the paper is organized as follows: section 2 describes some related work, section 3 describes the OwlSight platform, section 4 presents some use cases of application and the results of a real scenario and section 5 concludes the paper and points some future work.

II. Related Work

The volume and complexity of cyber attacks is requiring the development of advanced solutions able to detect and

978-1-5090-2403-2/16 $31.00 © 2016 IEEE
DOI 10.1109/BigDataSecurity-HPSC-IDS.2016.73

stop these attacks in time. Both academia and industry are focused and actively proposing solutions aimed at addressing the problem in the best way.

Gartner has published recently a report in [3] with a list of representative vendors developing threat intelligence platforms. Between the vendors we find: the platform provided by ThreatConnect [4] allows government agencies and large enterprises to aggregate all available threat data, analyze it rapidly, automate actions, and then produce tactical, operational and strategic threat intelligence all in one place. This data can be access through an API; the ThreatStream [5] is another threat intelligence platform that aims to help security teams to sort through to find hidden threats that can threaten the business, customers, intellectual property, and reputation of organizations. ThreatStream can be used in the cloud or on premise and like ThreatConnect it provides an API that allows the integration between the analysis and the existing security solutions (e.g. SIEMs); the LookingGlass ScoutVision [6] is another threat analysis platform. It relies on multiple and different data sources to detect threats and present the findings by means of graph-based views; Codenomicon [7] is another company providing threat intelligence solutions to enable effective security response. These solutions rely on multiple data sources to detect threats and provides high-level and drilled-down visualizations with detailed information to allow teams, particularly governments, CERTs and cyber authorities, to investigate further.

In [8] the authors argues that, despite all the efforts to develop cyber visualization technologies these are not capturing attention. Isolated solutions and pretty picture visualizations developed mainly to impact users are, according to the authors, falling short. Clearly understanding the users' needs and addressing their requirements is pointed as a critical factor to successfully develop a platform and insightful visualizations. The set of challenges identified in this work were an important basis for the discussion that we have with experts in the field, to identify important design considerations and devise appropriate dashboards. Lee et al. [9] show that the use of visualization speeds up the analysis process. Their work focus malware analysis and provides a good case for visualization, which is needed to recognize and extract unseen malware patterns. In [10] authors propose an online collaborative and explorative analysis tool, named OCEANS to help network administrators and security analysts to analyze network flow and log data. OCEANS provided multi-level visualization with temporal overview about IP connections and allows participants to collaborate on finding events and targeting attacks. Another interesting work on visual analytics is presented in [11]. It is a system to analyze data streams allowing the analysts to interact with the system and steer the clustering process to reduce the size of data streams to meaningful segments. The system includes big data analytics and combines different types of data to gain gain situational awareness and enhance the network security.

To our knowledge, there are only few platforms targeted for reducing the mean time to recover from cyber threats. From the platforms presented above [7] and [6] are focused on a type of organization and provide very limited visualizations. The works proposed in [9], [10] and [11] focus only some types of threats. OwlSight aims to provide visualization dashboards according the user needs and focus different types of cyber threats to provide an integrated vision around the threat.

III. The OwlSight Platform

Cyber attacks have always been like a cat-mouse game. As current attacks are found and handled by the cyber threat defenders, cyber threat actors are finding new ways to escape from the traps. Because of this, organizations are conscious about the difficulty of preventing a cyber attack and in this context proactive monitoring is seen as very import to early detect and stop cyber attacks before severe damages occur.

A. OwlSight: design considerations

OwlSight aims to provide real-time detection of cyber threats affecting different types of organizations and provide means to allow organizations to quickly respond to the attacks. Its conception involved the collection of feedback gathered from experts in the field and an extensive discussion of requirements. From this resulted design considerations considered crucial for the development and success of the platform.

- DC1 - Multiple data sources and types: External and internal sinkholing techniques, vulnerability analysis, sandbox analysis, honeypots, social networks, network and system logs are examples do data sources containing data useful to detect cyber attacks. Nowadays the number of connected devices and the number and variety of applications/services used is increasingly high. This leads to a high volume and heterogeneous log events that need to be conveniently stored and prepared for analysis. In this context, adopting big data principles and best practices of scalable real-time data systems to store and analyze such amount of apparently uncorrelated data is fundamental to build the basis of a cyber threat platform and provide real-time visualization and support for fast incident response.
- DC2 - Real-time and historic analysis: New attacks are most of the time an evolution from previous attacks. Analyze historic data is important to understand the tactics, techniques and procedures used by the attackers. This knowledge is useful to understand current attacks and to find the best approaches to respond to them. Despite the usefulness of combining historic with real-time attacks it brings a new challenge: analyze a huge amount of data to find patterns without induce a significant delay in the process of detecting and respond to the attacks. To address this challenge, the platform should follow high performance computing principles and include efficient algorithms able to extract useful data in a useful time.
- DC3 - Real-time visualization and alerts: Data visualization enables a deeper understanding of what is happening. By combining historic and real-time data it allows to gain situational awareness about cyber threats and to uncover

hidden patterns of data, identify emerging threats and support the remediation with efficient countermeasures. The variety of organizations requiring threat analysis and visualization premised that one visualization type does not fit all. Devise easy-to-use and easy-to-adopt dashboards, provided with insightful visualization and analysis is mandatory in a threat intelligence platform like OwlSight.

- DC4 - Low mean time to remediate: The growing and continuous sophistication of cyber attacks difficulties its prevention and so reduce the mean time to remediate takes a central point in the question. To accomplish this, a solution needs to detect a cyber attack in its early stage and provide mechanisms to pinpoint its route cause and quickly respond to the incident. A huge amount of real-time data consumed in real-time and analyzed towards the detection of cyber threats affecting the organizations imposes challenges to achieve low detection and remediation time. To address this challenge it is fundamental to correctly choose the data sources and devise a platform grounded by the cloud-computing and big-data principles able to collect, store and analyze very quickly continuously a huge amount of different types of data.
- DC5 - High accuracy: It is very difficult to quantify the coverage provided by any cyber security solution available in the market. Combine multiple data sources to increase the amount of cyber threats detected is a practice in this area. These sources can operate in different forms (e.g. domain sinkholing, URL and file sandboxing, black lists) and have more or less accuracy. Choose data sources that maximize the data accuracy, like domain sinkholing, and devise mechanisms like clustering analysis or voting are required to improve the threat detection accuracy.
- DC6 - On-premise versus Cloud: The use of public clouds is frequently prohibited in highly-regulated industries, enterprises with conservative views and requirements on proprietary control. To promote the widely adoption of a cyber threat platform it must be available to be used as a service in the cloud, or as an appliance kept and managed inside the organizations. OwlSight should attend this requirement by providing both a entirely public cloud solution ready to be used to detect cyber threats and a virtual appliance ready to run on the most widely virtualization platforms. Tools like vagrant [12] and docker [13] are considered useful to automate and simplify the management of these environments.

B. OwlSight: building blocks

The OwlSight platform is composed by several building blocks. It is illustrated in Figure 1.

As illustrated in Figure 1, OwlSight allows to collect data from different data sources. These data sources, also referred as data feeds, are divided into external data and internal data. By external data we mean data captured in a non intrusive way, i.e. outside the network and without requiring

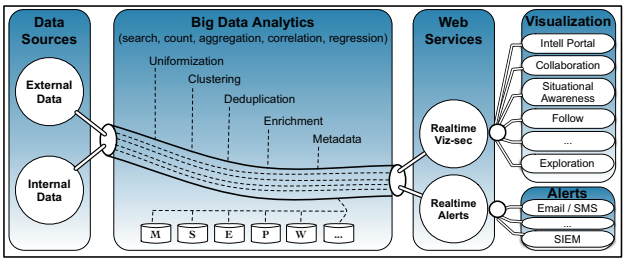

Fig. 1. OwlSight Threat Intelligence Platform

the involvement of organizations under monitoring. External sinkholing, passive DNS or social media data are examples of external sources. By internal data we mean all the network flow, logs and analysis outputs (e.g. DNS traffic, Web traffic, Email traffic, URL and file analysis) captured inside the network and that can be used to detect eminent or ongoing threats. Considering the platform's goal is given priority to the sources that provide real-time data. The platform has an agnostic collector allowing to quickly and easily consume data from different data sources.

As the data is consumed it is submitted to the data preparation, analysis and storage process. Since data is collected by multiple data sources, a data type uniformization and timestamp synchronization is primarily done. A clustering analysis is then performed to find similar events. Data deduplication is done to eliminate duplicate copies of repeating data. The resulting data is enriched, i.e. it is complemented with geolocation, WHOIS, DNS and reverse DNS lookups, hashing, autonomous systems name and number, SKIM and SPF records and file and URL analysis data. This data is important to get context around the threat event under analysis. All of this data is then analyzed and stored on different databases systems (e.g, Malware DB, Social Media Database, Email Database, Phishing Database). These databases contain historic data useful to perform both real-time and historical analysis. The analysis process is supported by a big data analysis engine. This engine is used to perform search, count, aggregation, correlation and regression analysis operations. The platform is also planned to detect new attacks based upon the trained recognition of malicious behavior patterns (e.g. recognize communication type, volume and content, match tactics, techniques and procedures across attacks).

The threat intelligence platform is provided as a Software-as-a-Service (SaaS) model, allowing users to register themselves and define the list of networks/companies to monitor. The communication between users and the platform is made via a set of RESTful Web services. These Web services provide the integration between a layer of visualization and real-time alerts. The real-time alerts can be triggered by email, SMS or to a Security Information and Event Management (SIEM). Per alert, they allow users to know: the IP address from where the malicious communications are leaving; the timestamps of the malicious communication attempts; which

malware is behind the malicious communications; what are the indicators of compromise associated to the malware. The visualization layer provides insightful visualization and analysis over the data. It combines real-time with historic analysis and allows organizations to: gain situational awareness; discover organizations requiring remediation services; follow a list of selected organizations; share a dashboard with partners for a predefined amount of time (collaboration); have access to the events details, understand the indicators of compromise, methods of infection, pinpoint the internal root cause for the problem and follow the remediation steps.

IV. OwlSight: real usage scenarios

In this section we describe the current state of OwlSight and present some use cases about the platform usefulness. An evaluation of platform, regarding the volume of data stored and processed, its performance and the mean time to detect is also presented.

A. Production Environment

Currently OwlSight is consuming data from two external data sources, i.e. the data is collected in a non-intrusive way. One of the feeds is based on external domain sinkholing and the other is provided by passive DNS systems installed in main Internet Service Providers. These feeds provide data in real-time for more than one hundred malware families with activity all over the world. By adopting domain sinkholing strategies and a well defined list of command and control server domains these feeds intercept malicious communications with a low false negative rate. From the data collected is possible to identify networks with machines participating in botnet activities.

This platform is hosted in five virtual machines running on AWS (Amazon Web Services). Three virtual machines are responsible for the Big Data Analytics process (BDA1, BDA2, BDA3) and the other two act as Web servers (WS1, WS2).

BDA1, BDA2 and BDA3 are c3.2xlarge instances with 8vCPUs, 15GiB of memory and 4TB of SSD storage each. These VMs run three types of databases: a cluster of Elasticsearch [14], a cluster of Cassandra [15] and three standalone installations of MySQL. A 90 days of historic data is kept in the database. Apache Spark [16] and the Spark SQL component is also installed in these machines providing the API and engine for big data analytics. A layer of code composed by Perl, Python and PHP programs is used to prepare, deduplicate, enrich and process the data. A set of RESTful Web Services is also provided allowing the integration with the visualization and alerts layer.

WS1 and WS2 are c3.large instances with 2vCPU and 4GiB of memory each. Each VM runs Apache 2.0 and PHP modules. These servers are responsible for providing the front-end interface (visualization layer) to the users. This interface combines HTML5, JavaScript and Ajax technologies. It also includes javaScript libraries like jQuery and d3js [17] to provide rich and interactive data visualizations. AWS load balancing is used to distribute client requests across the Web servers.

B. Use case 1: Cyber defense

In this subsection we describe how OwlSight can be used by cyber defense organizations (national/governmental CERTs and CSIRTs) to gain situational awareness and respond to incidents. For this type of organizations the malware data collected is prepared, enriched, analyzed and presented in a situational awareness interface and in an interface containing detailed information about the events observed.

A situational awareness interface is illustrated in Figure 2.

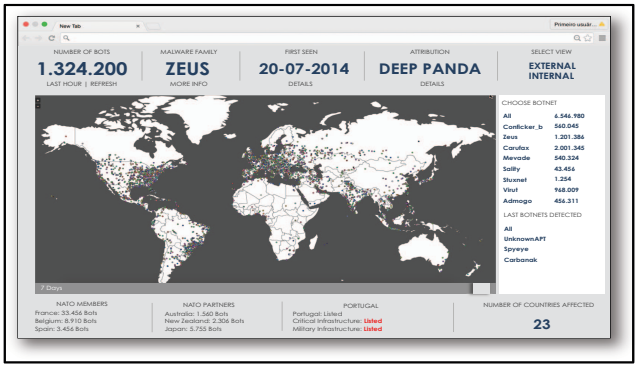

Fig. 2. OwlSight Cyber Defense Operations Console

The interface illustrated in Figure 2 allows to know in real-time the capacity that can suddenly be used by the attackers to initiate an attack against a militar or critical infrastructure. The attack capacity is represented in the world map by the amount and dispersion of bots identified by a color. Through the dashboard is also possible to navigate over the last 7 days and observe how the number of bots changes in a daily basis. The number of bots observed in predefined networks is also show in the bottom, together with indications of the number of bots observed in the national militar or critical infrastructures under monitoring. On the top is show the number of active bots in the last hour, the malware family clicked in the map, the first occurrence of that malware in the database and when possible an attribution field determined by the malware forensics process.

From the dashboard illustrated in Figure 2 is possible to jump to a second dashboard. It is the event details dashboard and it is illustrated in Figure 3.

The event details dashboard, as illustrated in Figure 3, provide detailed information about the malware, the last events detected with information about the source IP and port used for the communication and the command and control server contacted. It also provides information for incident response, including the detection rate achieved by different anti virus technologies, details regarding the static and dynamic analysis of the malware and the steps necessary to repair the problem. Another feature included in this dashboard is the possibility of share information with other partners, by providing and an access to the dashboard by a predefined period of time.

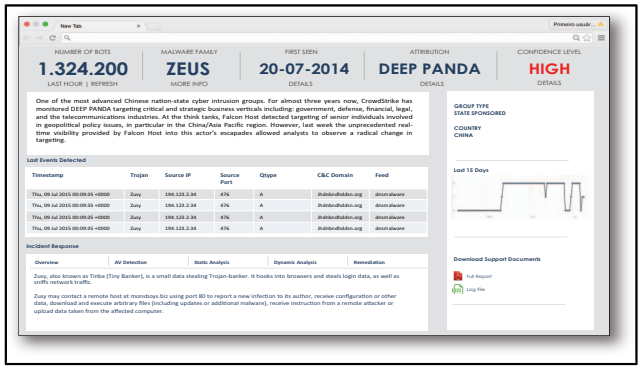

Fig. 3. OwlSight Cyber Defense Event Details

C. Use case 2: MSSPs/SOCs/Organization

In this subsection we describe how OwlSight can be used by Managed Service Security Providers (MSSPs), Security Operation Centers (SOCs) and Organizations in general. Depending on the type and interests of the organization the platform can be used for market prospection, allowing MSSP and SOC to identify potential clients, or to follow a set of companies and provide real-time detection, pinpointing and remediation services.

Figure 4 and Figure 5 shows how a MSSP can use the platform to identify potential clients. For privacy and security issues the data included in the figures is obfuscated.

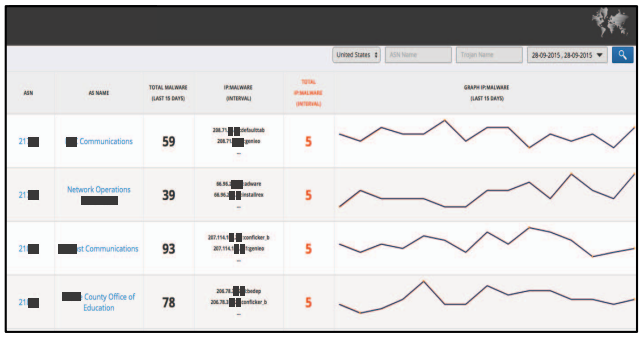

Fig. 4. OwlSight - Visualizing malware by country

Figure 4 shows a list of malware occurrences for a given country. It includes the autonomous system number (ASN) and name (ASName), the network name associated with the ASN, the total number unique combinations of IP addresses and malware observed in the last 15 days, a short list with the IP and malware detected and finally the total number of unique combinations of IP addresses observed in the interval selected by the user and a graphic showing the number of unique combinations of IP addresses during the last 15 days. By clicking on the ASN it is presented a list as illustrated in Figure 5. This list reveals the sub networks inside the ASN and allows to identify the organizations infected with malware. From this point a MSSP can decide to follow a organizations more closely. The data included in the exploration dashboard

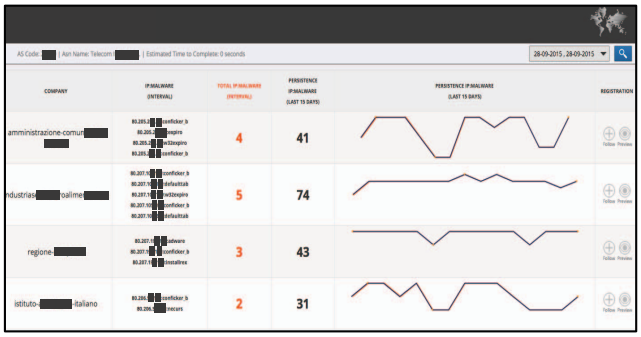

Fig. 5. OwlSight - Automatic discovery of infected organizations

is automatically discovered using the enrichment algorithms included in the platform.

MSSPs can follow a set of organizations through a partner dashboard. The organizations can be registered automatically or manually by indicating the range of IPs to monitor. Emails addresses, phone numbers and SIEM integration can be also set up for real-time alerts. The partner dashboard is illustrated in Figure 6.

Fig. 6. OwlSight - Partner dashboard to follow organizations more closely

From the dashboard illustrated in Figure 6 it is possible to manage the latest occurrences of malware detections in enterprise customers, understand its severity, observe the last 15 days malicious communications pattern, status, and enable sharing of dashboards with full detailed information of the security event detected to help to mitigate the incident as fast as possible. From this dashboard is also possible to access the organization dashboard for a closer look.

The organization dashboard is illustrated in Figure 7. It includes the total number of unique IP-malware combinations observed during the last 90 days worldwide, the total number of of unique IP-malware observed in the organization, the number of unique IP-malware observed in the last 24 hours and the number of consecutive days observing malware. The graph shows the variation of malware in a daily basis and the world map allows to quickly identify the region where the malware communication was observed (source). The tables in the bottom show the malware family observed, the severity of the malware and include a link to the intelligence portal.

This portal provides detailed information about the malware, allows the analysis of internal logs (e.g. DNS and Web Proxy logs) to pinpoint the compromised devices and provide steps for remediation.

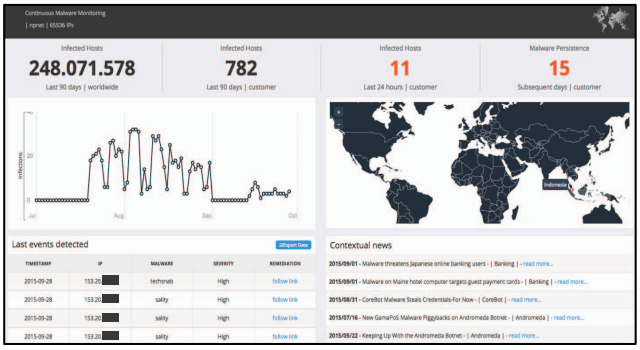

Fig. 7. OwlSight - Organization dashboard for a closer look

D. Production environment: fast overview

The work presented in this article was born of an academic challenge, which quickly extended to the industry, becoming a usable product. As said previously, the platform currently collects real-time data from two sources of information related to malware. It collects, prepares, enriches and analyses the data allowing to detect malware occurrences affecting organizations and enabling the pinpointing and fast response to the incidents. The platform is already used by real customers and the feedback until now is very interesting. It is commonly referred the ability of the platform to centralize a huge amount of logs and provide useful information by means of real-time alerts and insightful visualization.

In average the platform is storing more than 107 millions of malware events per day. These events are being aggregated in one minute intervals. The peak number events observed in the last 15 days was 6123 events per second. Per day is observed more than 2.7 million of unique IP-malware communications. The platforms contains historic data for the last 90 days, totaling more than 9600 million events and 4.8 terabytes of uncompressed data. The average time between the event consumption and its presentation through the dashboards is about 32 seconds and this is something that we are improving as a result of the query and analysis processes optimization conducted continuously.

V. Conclusions

Achieve a low mean time to remediate is the best chance to reduce the impact of cyber threats. This ability comes from combining threat avoidance and threat response capabilities into a strategic approach. These capabilities must be built on effective controls that are appropriate for the organization. Detection, pinpointing and response capabilities can be mapped directly to the threat impact. As fast as these tasks occur lower will be the damage.

In this paper we presented a platform targeted for real-time detection and visualization of cyber threats. It is a modular platform able to consume a huge amount and different types of data. The data is prepared, stored, analyzed by means of a big data analytics engine and presented through insightful and easy-to-use dashboards. These dashboards can be used by different types of organizations to gain situational awareness and to promptly detect threats, pinpoint its origin and support the incident response actions. An alpha version of the platform is already used in real scenarios contributing to the fast detection and incident response. The feedback collected from its utilization is very positive: OwlSight reduces the complexity of collecting, preparing and analyzing security data; accelerates the analysis; contributes to the decision making. In the short term we will integrate more data sources. We have performed some experiments on three new data sources and according to the results the coverage of malicious activities related with malware will increase up to 5 times.

References

[1] N. Perlroth, "Researchers find clues in malware," The New York Times, May 2012.

[2] VERDASYS, "Cyber attack defense: A kill chain strategy," http://www.idgconnect.com/view_abstract/ 14109/cyber-attack-defense-a-kill-chain-strategy, Mar 2013.

[3] C. Lawson and R. McMillan, "Technology overview for threat intelligence platforms," online, Dec 2014. [Online]. Available: https://www.gartner.com/doc/2941522/technology-overview-threat-intelligence-platforms

[4] "ThreatConnect - threat intelligence platform," www.threatconnect.com, Sep 2015. [Online]. Available: www.threatconnect.com

[5] "ThreatStream - threat intelligence platform," www.threatstream.com, Sep 2015.

[6] "LookingGlass scoutvision - threat intelligence analysis and management," www.lgscout.com, Sep 2015.

[7] "AbuseSA - threat intelligence platforms," www.codenomicon.com, Sep 2015.

[8] D. M. Best, A. Endert, and D. Kidwell, "7 key challenges for visualization in cyber network defense," in *Proceedings of the Eleventh Workshop on Visualization for Cyber Security*. New York, NY, USA: ACM, 2014, pp. 33–40.

[9] D. Lee, I. S. Song, K. J. Kim, and J. hyeon Jeong, "A study on malicious codes pattern analysis using visualization," *2014 International Conference on Information Science & Applications*, vol. 0, pp. 1–5, 2011.

[10] S. Chen, C. Guo, X. Yuan, F. Merkle, H. Schaefer, and T. Ertl, "Oceans: Online collaborative explorative analysis on network security," in *Proceedings of the Eleventh Workshop on Visualization for Cyber Security*, New York, NY, USA, 2014, pp. 1–8.

[11] F. Fischer and D. A. Keim, "Nstreamaware: real-time visual analytics for data streams to enhance situational awareness," in *Proceedings of the Eleventh Workshop on Visualization for Cyber Security, Paris, France, November 10, 2014*, 2014, pp. 65–72.

[12] M. Peacock, *Creating Development Environments with Vagrant*. Packt Publishing, 2013.

[13] J. Turnbull, *The Docker Book*. James Turnbull, 2014. [Online]. Available: https://books.google.pt/books?id=CtMEBwAAQBAJ

[14] C. Gormley and Z. Tong, *Elasticsearch: The Definitive Guide*. O'Reilly Media, 2015.

[15] A. Lakshman and P. Malik, "Cassandra: A decentralized structured storage system," *SIGOPS Oper. Syst. Rev.*, vol. 44, no. 2, pp. 35–40, Apr. 2010. [Online]. Available: http://doi.acm.org/10.1145/1773912.1773922

[16] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, 2010, pp. 10–10.

[17] M. Bostock. (2012) D3.js - data-driven documents. [Online]. Available: http://d3js.org/