2017 IEEE International Conference on Software Quality, Reliability and Security (Companion Volume)

Security Analytics in the Big Data Era

Dušan Mondek, Rudolf B. Blažek, and Tomáš Zahradnický
Boxtrap Research, s.r.o., Prague, Czech Republic
{ dusan.mondek | rudolf.blazek | tomas.zahradnicky }@boxtrap.net

***Abstract*—This paper discusses the reality of the state-of-the-art of existing information security systems that often provide senseless functions based on "buzz-words". It points out real-life requirements that these systems tend to ignore. It proposes that dynamically changing understandable use cases should be created in collaboration with corporate management to address objectives that are essential for the businesses. Big data analytic methodologies should be utilized to assist the design of efficient implementations of these use cases. Big data approaches should also be used for heuristic detection of unknown attacks and anomalies, for data enrichment, and for post-hoc forensic analysis. The main goal for information security systems that are usable in real-life should be that corporate decision makers are only provided trustworthy and actionable insights that are relevant to the business and provided services.**

Keywords—information security systems, buzz-words, real-life requirements, business objectives, research collaboration.

I. Introduction

Today's information security environments consist of highly heterogeneous systems that are distributed both horizontally and vertically among many geographical locations in physical and virtual domains. There are hundreds of nodes, user endpoints, servers, and other active network devices of all possible kinds, with their availability being the primary objective. In modern corporate IT environments, availability highly depends on many different factors. One of the most discussed factors is information security – a topic with direct influence on service availability and continuity, and consequently also on client satisfaction. At the end, it influences the most important objective in business – the revenue. Security teams will thus play a significant role in the ongoing information-age enterprise transformation. The success of the security teams will, however, depend on their ability to integrate controls and processes, and simultaneously develop new practices for the increasingly important fields of analytics, data intelligence and incident response [2].

This paper discusses the relationship between information security and availability, and their respective business-related consequences. The paper is organized as follows: Section I. introduces the reader to the topic. Section II. presents state-of-the-art methodologies used by modern enterprises in todays reality when dealing with big-data and security issues. Section III. describes commonly used approaches to security analytics on top of big-data that satisfy technical and compliance requirements. Summary, concluding remarks, and prediction of future trends are presented in Section IV.

II. Real-World State of the Art

The hostility of the Internet environment requires proactive detection of attacks. The size of the Internet and the scope of the ongoing intrusion attempts require the ability to store security-related data for complex long-term analysis.

Once a device is connected to the Internet, it is only a matter of time until it gets under attack and a security incident occurs. It is necessary to be able to detect that such a problem has occurred, decide whether it is still in progress, and to learn from the attacker's behavior. For this purpose security-related data must be stored, analyzed, and evaluated. However, one can hardly know ahead of time which data will be relevant to each possible attack vector, especially for unknown attacks. In addition, different kinds of data may be useful for real-time analysis, and for detailed post-hoc forensic analysis. Therefore the approach that is usually taken in real life can be described as "the more data stored the better". As a result, the amounts of stored data are constantly growing, frequently to gigantic extents. Managing daily loads of hundreds of Gigabytes of logs is not uncommon for todays corporations. Even smaller companies, with virtual environments containing a small number of virtual machines and a few mid-size firewalls, face the problem of increasing amounts of security-related data that must be stored, analyzed, and evaluated. The analyses include real-time or near-real-time attack detection, heuristic detection of unknown attacks, and computer-assisted forensic analysis. Analyzing such data volumes in such a complex manner moves the problem into the domain of big data processing.

The desired scope of security data analysis also enforces requirements on the richness of the analyzed data. The data that is processed by existing security systems comes from various heterogenous sources, and in many different forms. Including data related to system events, network traffic, or the results of analysis performed by other monitoring, detection, and analytics systems. Therefore nowadays it is necessary to deal with different delivery standards like syslog messages (based on a few different RFCs), SNMP, JDBC, ODBC, OBSEC/ LEA, NetFlow, IPFIX, and many more. The data may be generated in various text encodings, at various rates, and may or may not require active polling. Some data sources generate hundreds of records per second, some just a few messages per day. Moreover, data production often highly depends on external factors and on the condition of the surrounding infrastructure, e.g. producing more or fewer data at specific times (daytime versus night, working days versus weekends), or during specific occasions (an ongoing attack, misconfigured or purely configured services or backups, software or service malfunction, etc.). Does it sound familiar? Of course, it does.

978-1-5386-2072-4/17 $31.00 © 2017 IEEE
DOI 10.1109/QRS-C.2017.136

In real life, companies must deal with terabytes of security-related data which is always in motion, consumes significant bandwidth within the companies' LANs, and requires highly-available storage systems capable of complying with strict retention policies. In addition, this storage should be secure, and allow for advanced analytic solutions that need to access the stored data in distributed manner while performing various on-line and off-line data processing jobs. Security related analytics in the infrastructure of todays companies represents a typical big-data problem that requires big-data solutions.

III. SECURITY ANALYTICS ON TOP OF BIG DATA

Defense of information systems and networks is an integral part of information assurance and infrastructure protection not only in business and enterprise, but in all walks of life. Including public and emergency services provided by governments and local authorities, and in cyberspace warfare. Information assurance is one of the key requirements to guarantee that a whole system of essential services is not destroyed or disabled by an enemy [1].

The main prerequisite for efficient cybersecurity defense is the capability to collect relevant data. However, data gathering, its storage, and retention is just the tip of the iceberg. The problem is way more massive, and in real life it often starts at the point where people attempt to find and rapidly apply simple solutions to a highly complex problem. Users have basically two options: (a) accept the rules of the game and start designing highly complex solutions; (b) try to change the rules and simplify the problem as much as possible.

If companies invest more resources into initial analysis, it will result in significant decrease of the complexity of selected utmost important problems, and consequently improve the efficiency and capabilities of security defense solutions to achieve the desired level of availability and trustworthiness of the protected information systems and infrastructure. Today's IT market offers many types of scalable and highly available database systems like MongoDB, ElasticSearch, Apache Hadoop, etc. These systems are capable to store enormous amounts of heterogeneous data in a distributed manner for complex distributed off-line analysis. At the some time, however, they allow sensible integration of smaller (possibly relational) databases for selected real-time or near-real-time analytic jobs that must be performed on-line. Depending on the type of the protected environment, and after more extensive analysis of the problems, it is often possible to shrink the amounts of data needed for on-line analysis or detection, while providing more rapid analysis results.

An integral part of this design process in real life is the creation of understandable use cases in collaboration with the management to address objectives that are essential for the business and provided services. In order to to implement these use cases, the analysts must select specific data, decide what relevant characteristics are to be extracted from the data, and design various data transformations. This data selection process should ideally be computer assisted by employing big data analytics on the huge amounts of available data.

Having too little or too much data is often a problem, therefore efficient selection of relevant data is an essential part of the design process, especially for on-line analysis tasks. In this case more data gathered does not necessarily mean more value. Scalable, distributed storage systems may be used to gather all other available data that may be relevant later, for forensic analysis, or anywhere where analysis is not required to be performed on-line. Off-line jobs may run in the background and enrich the results of on-line analysis, ideally in an automated fashion.

There is a significant number of information security solutions available that claim to provide superior security-related analytic functions. However, in real-life the majority of them do not even meet the most elementary requirements of corporate information security teams. In the light of newly introduced cyber-security legislation, and with constantly evolving threats in mind, there is urgent need for systems that provide constantly evolving analytic output in "ready to use" form that is trustworthy and understandable for both technical experts and higher corporate management. Only then, decision makers receive enough relevant and understandable information that allows them to make correct and fast decisions. And to articulate the decisions clearly and efficiently back to the technical teams so that they can be capable of taking immediate action when it is needed.

Information security market is, however, saturated with solutions that use trivial detection mechanisms often inaccurately called "behavior analysis" or "anomaly detection". These solution profit from the lack of deeper expertise in enterprises, and from the absence of standardized and widely available test platforms for assessing the efficiency and trustworthiness of the analytic output of the tested solutions. These high-level facts and misunderstandings have significant influence on low-level architectural and technical aspects of corporate information security. At that point the real security capability of a corporation does not comply with its information security strategy and goals, leaving thus significant gaps in company cyber-defense.

IV. CONCLUSIONS

This paper provides readers with elementary introduction to real-world security analytics needs, and its necessity for today's corporation IT environments. Existing information security solutions, in vast majority, do not benefit from scientific research and real-life experiences. Security software vendors rather present senseless functions including "buzz-words" with very little value that would allow cyber-security teams to successfully perform every day security operations and ensure availability and reliability of critical systems.

This negative trend, on the other hand, brings new opportunities for innovative cyber-security startups, and security-related research projects in collaboration with the scientific and academic communities around the globe. We can expect completely new ways how to design information security strategies in highly dynamic environments, how to detect both simple and comprehensive attack campaigns, and how to ensure that decision makers are only provided with trustworthy and actionable insights relevant to the business.

REFERENCES

1. Yaoqi Li, Jianjing Shen, Application of Big Data in Cyberspace Warfare, IEEE Xplore, ISBN: 978-1-4673-9714-8, Yinchuan, China, 2016
2. Erik van Ommeren, Martin Borret, Marinus Kuivenhoven, Staying Ahead in the Cyber Security Game, 1st Edition, April 2014, LINE UP boek en media bv, Groningen, the Netherlands, 1892, pp. 8-9.