ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2014

Big Data Analytics Security Issues in Data Driven Information System

M. Chithik Raja[1], Munir Ahmed Rabbani[2]

Research Scholar, Department of Information Technology AMET University, Chennai, India[1]

B.S. Abdurrahman University, Chennai, India[2]

ABSTRACT: Enterprises routinely collect terabytes of security-relevant data (for instance, network events, software application events, and people's action events) for regulatory compliance and post hoc forensic analysis. Large enterprises generate an estimated 10 to 100 billion events per day, depending on size. These numbers will only grow as enterprises enable event logging in more sources, hire more employees, deploy more devices, and run more software. Unfortunately, this volume and variety of data quickly become overwhelming. Existing analytical techniques don't work well at large scales and typically produce so many false positives that their efficacy is undermined. The problem becomes worse as enterprises move to cloud architectures and collect much more data.

KEYWORDS: CSA, SIEM, Hadoop , APT

I.INTRODUCTION

Big data analytics is the large-scale analysis and processing of information in active use in several fields and, in recent years, has attracted the interest of the security community for its promised ability to analyze and correlate security-related data efficiently and at unprecedented scale. Differentiating between traditional data analysis and big data analytics for security is, however, not straightforward. After all, the information security community has been leveraging the analysis of network traffic, system logs, and other information sources to identify threats and detect malicious activities for more than a decade, and it's not clear how these conventional approaches differ from big data. "Big Data Analytics for Security Intelligence", focuses on big data's role in security. The report details how the security analytics landscape is changing with the introduction and widespread use of new tools to leverage large quantities of structured and unstructured data. It also outlines some of the fundamental differences from traditional analytics and highlights possible research directions. We summarize some of the report's key points.

II. RELATED WORKS

Data-driven information security dates back to bank fraud detection and anomaly-based intrusion detection systems (IDSs). Although analyzing logs, network flows, and system events for forensics and intrusion detection has been a problem in the information security community for decades, conventional technologies aren't always adequate to support long-term, large-scale analytics for several reasons: first, retaining large quantities of data wasn't economically feasible before. As a result, in traditional infrastructures, most event logs and other recorded computer activities were deleted after a fixed retention period (for instance, 60 days). Second, performing analytics and complex queries on large, unstructured datasets with incomplete and noisy features was inefficient. For example, several popular security information and event management (SIEM) tools weren't designed to analyze and manage unstructured data and were rigidly bound to predefined schemas. However, new big data applications are starting to become part of security management software because they can help clean, prepare, and query data in heterogeneous, incomplete, and noisy formats efficiently. Finally, the management of large data warehouses has traditionally been expensive, and their deployment usually requires strong business cases. The Hadoop framework and other big data tools are now commoditizing the deployment of large-scale, reliable clusters and therefore are enabling new opportunities to process and analyze data.

III. FRAUD DETECTION USING BIG DATA TECHNOLOGIES

Fraud detection is one of the most visible uses for big data analytics: credit card and phone companies have conducted large-scale fraud detection for decades; however, the custom-built infrastructure necessary to mine big data for fraud detection wasn't economical enough to have wide-scale adoption. One of the main impacts from big data technologies is that they're facilitating a wide variety of industries to build affordable infrastructures for security monitoring.

In particular, new big data technologies such as the Hadoop ecosystem (including Pig, Hive, Mahout, and RHadoop), stream mining, complex-event processing, and NoSQL databases are enabling the analysis of large-scale, heterogeneous datasets at unprecedented scales and speeds. These technologies are transforming security analytics by facilitating the storage, maintenance, and analysis of security information. For instance, the WINE platform[1] and Bot-Cloud[2] allow the use of MapReduce to efficiently process data for security analysis. We can identify some of these trends by looking at how reactive security tools have changed in the past decade. When the market for IDS sensors grew, network monitoring sensors and logging tools were deployed in enterprise networks; however, managing the alerts from these diverse data sources became a challenging task. As a result, security vendors started the development of SIEMs, which aimed to aggregate and correlate alarms and other network statistics and present all this information through a dashboard to security analysts. Now big data tools are improving the information available to security analysts by correlating, consolidating, and contextualizing even more diverse data sources for longer periods of time.
We can see specific benefits from big data tools from a recent case study presented by Zions Bancorporation. Its study found that the data quantities it had to deal with and the number of events it had to analyze were too much for traditional SIEM systems (it took between 20 minutes to an hour to search among a month's load of data). In its new Hadoop system running queries with Hive, it gets the same results in approximately one minute.[3] The security data warehouse driving this implementation lets users mine meaningful security information from not only firewalls and security devices but also website traffic, business processes, and other day to-day transactions. This incorporation of unstructured data and multiple disparate datasets into a single analysis framework is one of big data's promising features. Big data tools are also particularly suited to become fundamental for advanced persistent threat (APT) detection and forensics.[4,5]APTs operate in a low-and-slow mode (that is, with a low profile and long-term execution); as such, they can occur over an extended period of time while the victim remains oblivious to the intrusion. To detect these attacks, we need to collect and correlate large quantities of diverse data (including internal data sources and external shared intelligence data) and perform long-term historical correlation to incorporate a posteriori information of an attack in the network's history.

IV. SECURITY ISSUES IN BIG DATA ANALYTICS

Although the application of big data analytics to security problems has significant promise, we must address several challenges to realize its true potential. Privacy is particularly relevant as new calls for sharing data among industry sectors and with law enforcement go against the privacy principle of avoiding data reuse that is, using data only for the purposes that it was collected. Until recently, privacy relied largely on 75 technological limitations on the ability to extract, analyze, and correlate potentially sensitive datasets. However, advances in big data analytics have given us tools to extract and correlate this data, making privacy violations easier. Therefore, we must develop big data applications with an understanding of privacy principles and recommendations. Although privacy regulation exists in some sectors for instance, in the US, the Federal Communications Commission works with telecommunications companies, the Health Insurance Portability and Accountability Act addresses healthcare data, Public Utility Commissions in several states restrict the use of smart grid data, and the Federal Trade Commission is developing guidelines for Web activity all this activity has been broad in system coverage and open to interpretation in most cases. Even with privacy regulations in place, we need to understand that large-scale collection and storage of data make these data stores attractive to many parties, including industry (who will use our information for marketing and advertising), government (who will argue that this data is necessary for national security or law enforcement), and criminals (who would like to steal our identities). Therefore, our role as big data application architects and designers is to be proactive in creating safeguards to prevent abuse of these big data stores.

Another challenge is the data provenance problem. Because big data lets us expand the data sources we use for processing, it's hard to be certain that each data source meets the trustworthiness that our analysis algorithms require to produce accurate results. Therefore, we need to reconsider the authenticity and integrity of data used in our tools. We can explore ideas from adversarial machine learning and robust statistics to identify and mitigate the effects of maliciously inserted data.

This particular CSA report focuses on the use of big data analytics for security, but the other side of the coin is the use of security to protect big data. As big data tools continue to be deployed in enterprise systems, we need to improve systems security by not only leveraging conventional security mechanisms (for example, integrating Transport Layer Security within Hadoop) but also introducing new tools, such as Apache’s Accumulo, to deal with the unique security problems in big data management.

Finally, another area that the report didn’t cover but that needs further development is human computer interaction and, in particular, how visual analytics can help security analysts interpret query results. Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces. Compared to technical mechanisms developed for efficient computation and storage, human-computer interaction in big data has received less attention but is nonetheless one of the fundamental tools to achieve the “promise” of big data analytics, because its goal is to convey information to a human via the most effective representation. Big data is changing the landscape of security technologies for network monitoring, SIEM, and forensics. However, in the eternal arms race of attack and defense, big data is not a panacea, and security researchers must keep exploring novel ways to contain sophisticated attackers. Big data can also create a world where maintaining control over the revelation of our personal information is constantly challenged. Therefore, we need to increase our efforts to educate a new generation of computer scientists and engineers on the value of privacy and work with them to develop the tools for designing big data systems that follow commonly agreed privacy guidelines.

V. Conclusion and Future Work

Big data is undoubtedly the most popular technique in IT industry while meanwhile with plenty of controversy and concerns due to its immaturity and potential security issues. This proposed engine aims to improve the accuracy of IDS reporting and to enhance the efficiency and effectiveness of the detection on those malicious intrusion and attack. In conclusion, although there are still a lot of works to be done to defend cloud system, we believe that we have walked through the most difficult first step and we are heading to the right direction. In near future, we will come up with an effective and efficient model in protecting Big data environment.

REFERENCES

[1] A. Rowstron et al. Nobody ever got fired for using hadoop on a cluster. In HotCDP, 2012.
[2]. E. Chickowski, “A Case Study in Security Big Data Analysis,” Dark Reading, 9 Mar. 2012.
[3] E. Ryvkina et al. Revision processing in a stream processing engine: A high-level design. In ICDE, 2006.
[4]. J. François et al., “BotCloud: Detecting Botnets Using MapReduce,” Proc. Workshop Information Forensics and Security, IEEE, 2011, pp. 1–6.
[5]. N. Pansare, V. R. Borkar, C. Jermaine, and T. Condie. Online aggregation for large mapreduce jobs. PVLDB, 2011.
[6] P. Upadhyaya, Y. Kwon, and M. Balazinska. A latency andfault-tolerance optimizer for online parallel query plans. InSIGMOD, 2011.
[7]. P. Giura and W. Wang, “Using Large Scale Distributed Computing to Unveil Advanced Persistent Threats,” Science J., vol. 1, no. 3, 2012, pp. 93–105.
[8]. T. Dumitras and D. Shou, “Toward a Standard Benchmark for Computer Security Research: The Worldwide Intelligence Network Environment (WINE),” Proc. EuroSys BADGERS Workshop, ACM, 2011, pp. 89–96.
[9]. T.-F. Yen et al., “Beehive: Large-Scale Log Analysis for Detecting Suspicious Activity in Enterprise Networks,” to be published in Proc. Ann. Computer Security Applications Conference (ACSAC 13), ACM, Dec. 2013.
[10] T. White. Hadoop: The Definitive Guide. 2009.
[11]. V. Raman, B. Raman, and J. M. Hellerstein. Online dynamic reordering for interactive data processing. VLDB ’99.

BIOGRAPHY

M.Chithik Raja was born in Periyakulam , Tamilnadu , India in 1978. He received the B.Sc. degree in Computer Science from the University of Madurai Kamaraj University, India, in 1999, and the M.Sc.in 2001, M.E. from Anna university ,India , in 2007 and pursuing his Research in AMET University , India In 2001, he joined the Department Computer Science, Jayaraj Chelladurai College, as a Lecturer, and in 2007 became a Senior Lecturer at Vickram college of Engineering. Since June 2009 , he has been with the Department of Computer Science and Engineering at Surya College of Engineering and Technology, where he was an Assistant Professor. After this he has been two years with Ethiopian Institute of Technology, Ethiopia. Now he is working at Salalah College of Technology, Sultanate of Oman. His current research interests include Big Data Analytic, Security issues, Wireless sensor networks, Cloud Computing, Network Security, and Data Mining. He has publishes more than 13 International journals . He has published three Computer books with LAMPERT Publication in Germany. He fonts of conducting workshop and seminar in the recent IT technology.