Security Challenges in Big Data

G.L.Anand Babu, Dr. K.S. Reddy, G.Sekhar Reddy

Department of Information Technology, CVSR School of Engineering, Anurag Group of Institutions, Hyderabad, Telangana, India

***Abstract*—** Big data has recently received considerable attention and many research efforts have been directed to big data processing due to its high volume, velocity and variety challenges (referred as "3V"). However, in addition to the 3V challenges, the flourishing of big data also turning point on fully understanding and managing newly arising security and privacy challenges. If data are not authentic, new mined knowledge will be unconvincing, while if security is not well addressed, people may be reluctant to share their data. Because security has been investigated as a new dimension, "veracity" in big data, in this paper, we aim to highlight new challenges of big data in terms of security and privacy. These challenges will motivate increased focus on fortifying big data infrastructures.

Keywords— Big Data, Security, Privacy.

I. INTRODUCTION

Big data is a collection of data sets so large (in size) and complex (in real-time, streaming, or non-structured form) that it is difficult to use traditional data management and analysis tools to efficiently gain useful information.

The term "Big Data" refers to the massive amounts of digital information companies and governments collect about human beings and our environment. The amount of data generated is expected to double every two years, from 2500 exabytes in 2012 to 40,000 exabytes in 2020 [1].

Security and privacy issues are magnified by the volume, variety, and velocity of Big Data. It is not merely the existence of large amounts of data that is creating new security challenges. Big Data has been collected and utilized by many organizations for several decades. The current use of Big Data is novel because organizations of all sizes now have access to Big Data and the means to employ it. In the past, Big Data was limited to very large organizations such as governments and large enterprises that could afford to create and own the infrastructure necessary for hosting and mining large amounts of data. These infrastructures were typically. Proprietary and were isolated from general networks.

II. SECURITY REQUIREMENTS

Big data is usually characterized by "3V"- volume, velocity and variety. Volume captures the huge amount of data generated by organizations or individuals. Velocity shows the high speed at which data is generated, captured and shared. Variety means a proliferation of new data types from social networks, machine devices and mobile sources that are integrated into traditional transactional data types. In the general architecture of big data analytics, both distributed big data storing and parallel big data processing are driven by the big data 3V challenges [2]. In addition to the 3V challenges, big data also faces new security and privacy challenges.

While big data creates enormous values for economic growth and technical innovation, we are already aware that the deluge of data also raises new privacy concerns. Thus, privacy requirements in big data architecture should be identified as deeply as possible to balance the benefits of big data and individual privacy preservation.

If big data is not authentic, newly mined knowledge becomes useless. Recently, a new dimension, veracity, has been advocated to address the security challenges in big data. However, the study of privacy in big data is still in its early stage. Therefore, we focus ourselves on big data privacy in this article and identify the privacy requirements of big data analytics as follows.

Privacy requirements in big data collection: As big data collection takes place pervasively, eavesdropping is possible, and the data could be incidentally leaked. Therefore, if the collected data is personal and sensitive, we must resort to physical protection methods as well as information security techniques to ensure data privacy before it is securely stored.

Privacy requirements in big data storage: Compared to eavesdropping an individual's data during the big data collection phase, compromising a big data storage system is more harmful. It can disclose more individual personal information once it is successful. Therefore, we need to ensure the confidentiality of stored data in both physical and cyber ways.

Privacy requirements in big data processing: The key component of big data analytics is big data processing, as it indeed mines new knowledge for economic growth and technical innovation. Because big data processing efficiency is an important measure for the success of big data, the privacy requirements of big data processing become more challenging. We never sacrifice big efficiency for big privacy, and should not only protect individual privacy but also ensure efficiency at the same time. In addition, since inter big data processing runs over multiple organizations' data, big data sharing is essential, and ensuring privacy in big data sharing becomes one of the most challenging issues in big data processing.

Therefore, it is desirable to design efficient and privacy-preserving algorithms for big data sharing and processing.

In recent years, we have witnessed plenty of privacy-preserving techniques being proposed. However, as they are tailored for privacy requirements in traditional analytics, they are not sufficient to satisfy the privacy requirements in big data analytics [3].

How to tackle big data from a security point of view

Big data is an immensely popular talking point, but what are we really discussing? From a security perspective, there are two distinct issues: securing the organization and its customers' information in a Big Data context; and using Big Data techniques to analyze, and even predict, security incidents.

III. SECURING YOUR BIG DATA

Many businesses already use Big Data for marketing and research, yet may not have the fundamentals right – particularly from a security perspective. As with all new technologies, security seems to be an afterthought at best. Big Data breaches will be big too, with the potential for even more serious reputational damage and legal repercussions than at present. A growing number of companies are using the technology to store and analyze petabytes of data including web logs, click stream data and social media content to gain better insights about their customers and their business. As a result, information classification becomes even more critical and information ownership must be addressed to facilitate any reasonable classification. Most organizations already struggle with implementing these concepts, making this a significant challenge. We will need to identify owners for the outputs of Big Data processes, as well as the raw data. Thus data ownership will be distinct from information ownership – perhaps with IT owning the raw data and business units taking responsibility for the outputs.

Very few organizations are likely to build a Big Data environment in-house, so cloud and Big Data will be inextricably linked. As many businesses are aware, storing data in the cloud does not remove their responsibility for protecting it from both a regulatory and a commercial perspective [4]. Techniques such as attribute based encryption may be necessary to protect sensitive data and apply access controls (being attributes of the data itself, rather than the environment in which it is stored). Many of these concepts are foreign to businesses today.

IV. DEPLOYING BIG DATA FOR SECURITY

The deployment of Big Data for fraud detection, and in place of security incident and event management (SIEM) systems, is attractive to many organizations. The overheads of managing the output of traditional SIEM and logging systems are proving too much for most IT departments and Big Data is seen as a potential saviour. There are commercial replacements available for existing log management systems, or the technology can be deployed to provide a single data store for security event management and enrichment. Taking the idea a step further, the challenge of detecting and preventing advanced persistent threats may be answered by using Big Data style analysis. These techniques could play a key role in helping detect threats at an early stage, using more sophisticated pattern analysis, and combining and analyzing multiple data sources. There is also the potential for anomaly identification using feature extraction.

Today logs are often ignored unless an incident occurs. Big Data provides the opportunity to consolidate and analyze logs automatically from multiple sources rather than in isolation. This could provide insight that individual logs cannot, and potentially enhance intrusion detection systems (IDS) and intrusion prevention systems (IPS) through continual adjustment and effectively learning "good" and "bad" behaviours.

Integrating information from physical security systems, such as building access controls and even CCTV, could also significantly enhance IDS and IPS to a point where insider attacks and social engineering are factored in to the detection process. This presents the possibility of significantly more advanced detection of fraud and criminal activities.

V. BIG DATA SECURITY

Hadoop Security for the Enterprise

Organizations use Hadoop big data systems to store and process an ever-growing volume of enterprise data. The growth of big data has created a pressing need to secure data in order to avoid data breaches and to comply with regulations such as the Payment Card Industry Data Security Standard (PCI DSS), Sarbanes Oxley, HIPAA, HITECH and many state and federal data privacy laws. When developing a big data strategy, organizations need to consider a comprehensive solution for data security and data governance for their enterprise Hadoop implementation. Data security and data governance can be achieved by an optimum combination of appropriate security tools with customized configuration, clear policy definition and adherence to best practices.

The MetaScale Big Data Security Solution

MetaScale leverage our experience of implementing big data solutions within highly regulated industries such as retail, healthcare and finance to provide a holistic approach to Hadoop security. Our approach to creating an ideal data security platform enables enterprise customers to secure their data and comply with regulatory requirements by encrypting data that is stored and processed by Hadoop systems, centralizing key management, enforcing access control policies and gathering security intelligence on data access. MetaScale offers Big Data Security Assessment and Customized Hadoop Security Solutions to help customers apply the right combination of security measures to achieve an ideal data security platform for their specific requirements. To achieve an ideal data security platform for your big data implementation, MetaScale analyzes all security stages for gaps and develops solutions to augment the standard configurations of your Hadoop distribution with customized plugins and domain specific best practices.

Securing Big Data Environments with Vormetric

Vormetric solutions for big data security enable organizations to maximize the benefits of big data analytics—while maximizing the security of their sensitive data and addressing the requirements of their compliance office. The Vormetric Data Security Platform offers the granular controls, robust encryption, and comprehensive coverage that organizations need to secure sensitive data across their big data environments—including big data sources, big data infrastructure, and big data analytic results. By delivering a single security solution that offers coverage of these areas, Vormetric enables security teams to leverage centralized controls that optimize efficiency and compliance adherence. The Vormetric Data Security Platform offers capabilities for big data encryption, key management, and access control—featuring several product offerings that share a common, extensible infrastructure. Further, the solution generates security intelligence on data access by users, processes, and applications.

VI. CONCLUSION

In our fast-paced and connected world where Big Data is king, it is critical to understand the importance of security as we process and analyze massive amounts of data. This starts with understanding our data and associated security policies, and it also revolves around understanding the security policies in our organizations and how they need to be enforced.

VII. REFERENCES

[1]. http://www.emc.com/leadership/digital-universe/iview/big-data-2020.html

[2]. X. Wu et al., “Data Mining with Big Data,” IEEE Trans. Knowledge Data Eng., vol. 26, no. 1, 2014, pp. 97–107.

[3]. A. Cavoukian and J. Jonas, “Privacy by Design in the Age of Big Data,” Office of the Information and Privacy Commissioner, 2012.

[4]. M. Li et al., “Toward Privacy-Assured and Searchable Cloud Data Storage Services,” IEEE Network, vol. 27, no. 4, 2013, pp. 1–10.