Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

A Review on Security Framework and Risks across the Big Data Life Cycle

Dr. N. Kumaresh

Associate Professor in MCA

RV Institute of Technology and Management

Chaithanya Layout, 8th Phase,JP Nagar , Bengaluru – 570076

Abstract

Big data, a burgeoning concept, pertains to the management of vast volumes of data originating from diverse sources such as databases, log files, and social media posts. This data, spanning text, numbers, images, etc., manifests in structured, semi-structured, and unstructured forms. Characteristics like velocity, volume, variety, value, and complexity provide additional dimensions to the understanding of big data. As the field of big data technology evolves, it brings forth various security concerns and challenges. This paper introduces a comprehensive framework for the big data lifecycle, which encompasses four key phases: data collection, data storage, data analytics, and knowledge creation. Each phase is briefly outlined, and proceeds to delineate the associated security threats and potential attacks. By integrating the big data lifecycle with security considerations, a unified security threat model is proposed. This model serves as a foundation for conducting further research in the realm of big data security, with the ultimate goal of fortifying the infrastructure supporting big data.

Keywords - Big data, threats, threat model, big data lifecycle, security considerations

Introduction

In recent years, the concept of big data has emerged to address the challenges posed by the escalating volume of information. Big data, broadly defined, involves the management of vast quantities of data originating from diverse sources, such as database management systems, log files, social media posts, and sensor data (Bajaj et al., 2014). At first glance, the term 'big data' conjures images of the substantial data sets that need storage and processing. Indeed, the sheer volume of data is a defining characteristic of big data, often surpassing the scale of an Exabyte (10^{18}). This necessitates specialized storage solutions, high-performance data processing, and advanced analytical capabilities (Kaisler et al., 2013).

Big data comprises intricate datasets encompassing various data types, including text, numbers, images, and videos, exceeding the capacities of traditional database management systems (Govindarajan et al., 2014). Notably, three key attributes—volume, velocity, and variety—define big data. Additionally, other attributes, such as value and complexity, contribute to the multifaceted nature of big data (Kaisler et al., 2013; Katal et al., 2013).

The volume attribute refers to the vast amount of data, surpassing conventional storage solutions. According to Bajaj et al. (2014), an astounding 90 percent of the world's current data has been generated in the past two years, with an average of 2.5 quintillion data bytes created daily. The velocity attribute pertains to the speed at which data is generated and processed, necessitating advanced processing capabilities beyond traditional systems.

Additionally, velocity encompasses the high-speed movement of data between storage locations over networks (Bajaj et al., 2014).

Another notable attribute of big data is its variety, which encompasses the diverse sources generating data in different types and formats (Bajaj et al., 2014; Govindarajan et al., 2014; Kaisler et al., 2013; Katal et al., 2013). Data sources range from digital pictures and videos to social media, sensor data, healthcare records, text, log files, tweets, and purchase transaction records. Essentially, big data comprises various data formats, including structured, unstructured, and semi-structured.

Two other attributes associated with big data are value and complexity (Kaisler et al., 2013). The value attribute in big data pertains to the significance of information (knowledge) derived from processing and analyzing big data, which proves instrumental in the decision-making process (Katal et al., 2013). The complexity attribute refers to the intricacies of relationships and complex links within the structure of big data. In this context, envisioning the complexity arises when minor changes in big data can result in a multitude of significant alterations (Katal et al., 2013).

Considering the perspectives of security and privacy, the approach to big data security must encompass safeguarding the data itself, the processes involved, and the outputs generated. Kim et al. (2013) assert that security in big data revolves around three key aspects: data security, access control, and information security. Additionally, Xu et al. (2014) introduce a big data security model that considers user roles in security across different phases of the big data process. However, many previous studies on big data security overlook the threats and attacks specific to the big data environment, with limited attention to the big data lifecycle and its correlation with threats and attacks based on the lifecycle model.

Securing the big data environment requires a comprehensive understanding of the threats and attacks occurring throughout its lifecycle. Identifying these threats and attacks aids the security community in developing effective defenses. To our knowledge, there have been no previous studies addressing the lifecycle of big data along with its associated threats and attacks. Consequently, our paper presents a big data lifecycle model comprising four phases: data collection, data storage, data processing, and knowledge creation. Furthermore, we provide a summary of security threats and attacks in each phase of the lifecycle. The integration of the presented big data lifecycle with security threats and attacks yields a security threat model that can serve as a foundation for researching and enhancing the security of big data infrastructure.

Literature Review

In recent years, the advent of big data has been driven by the need to address the escalating volume of data. However, there is a notable gap in the existing literature concerning the intersection of big data and security. Some articles delve into defining, characterizing, and elucidating the challenges associated with big data. For instance, Katal et al. (2013) present big data as a novel technology and explore various issues and challenges pertaining to its utilization. Sagiroglu and Sinanc (2013) provide an overview, discussing the content, characteristics, scope, advantages, challenges, and privacy concerns of big data. Other research papers engage in debates regarding the novel technology of big data, discussing its challenges and potential scopes (Bakshi, 2012; Demchenko et al., 2012; Singh & Singh, 2012).

In the realm of security and privacy, certain research articles tackle the privacy issues and security challenges associated with the big data era. Smith et al. (2012), for instance, delve into personal privacy on social networks and strategies for users to control their privacy. Kim et al. (2013) explore big data security, offering a security methodology for effectively fortifying and protecting big data through safeguarding selected attributes. Jensen (2013) addresses privacy challenges in big data and explores strategies for controlling big data processes in compliance with privacy standards. Additionally, big data analytics is acknowledged as a tool to enhance security by collecting and analyzing both structured and unstructured enterprise data (Cardenas et al., 2013).

The emergence of big data technologies prompts researchers to consider the protection of this novel data framework. Information security maintenance is deemed necessary for an organization's information systems

infrastructure that incorporates big data technologies (Miloslavskaya et al., 2014). Marchal et al. (2014) propose a security model for analyzing large datasets from a security perspective, enabling the monitoring of local enterprise networks, detection and prevention of network intrusions, and conducting forensics analysis. Another line of research focuses on the user role in securing big data infrastructure through a proposed security model (Xu et al., 2014). They identify four types of user roles in the big data environment: data provider, data collector, data miner, and decision maker.

Conversely, various threats and attacks pose risks to big data technology. Dev et al. (2012) discuss a data mining-based threat that exploits techniques to extract sensitive data and valuable information. Data privacy threats, such as re-identification and incorrect results, are highlighted by Jensen (2013). Wu and Guo (2013) emphasize the significant concern of privacy and information assurance in the big data environment, where the extraction of personal or sensitive data can harm individuals and organizations, leading to various business problems. To our knowledge, no study comprehensively addresses threats and attacks within a single model concerning the big data lifecycle. Therefore, the study serves as a foundational step for future security research aiming to secure the big data environment.

Big Data Security Life Cycle

In this section, the big data security lifecycle model along with the fundamental components of any big data framework is proposed. Building upon the work of Xu et al. (2014), who examine big data from a user role perspective, identifying four roles: data provider, data collector, data miner, and decision maker, our model specifically focuses on the stages of the big data lifecycle. Our model comprises four distinct phases within the big data framework: the data collection phase, data storage phase, data processing and analysis, and knowledge creation. Figure 1 illustrates the key elements in the big data lifecycle.

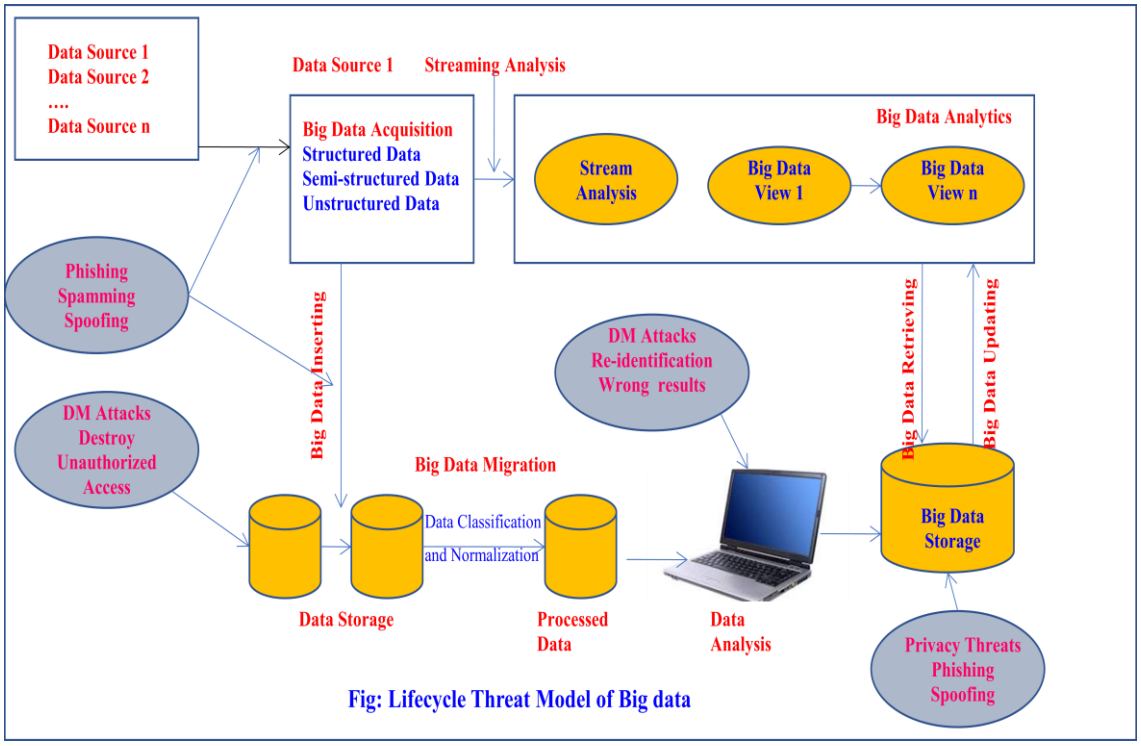

Fig: Lifecycle Threat Model of Big data

Data Collection Phase

During the data collection phase, information is sourced from diverse outlets and arrives in various formats, including structured, semi-structured, and unstructured. From a security standpoint, the safeguarding of big data

technology should commence with the initial phase of the lifecycle. It is crucial to gather data from trusted sources and ensure the security and protection of this phase. Indeed, implementing security measures is essential to prevent the unauthorized release of data. Some security precautions applicable in this phase involve enforcing limited access control, particularly for those receiving data from the provider, and encrypting specific data fields, such as personal information identifiers

DATA STORAGE PHASE

During the data storage phase, the accumulated data is stored and readied for utilization in the subsequent phase, namely the data analytics phase. Given that the stored data may include sensitive information, it is imperative to exercise ample precautions during the storage process. To ensure the security of the stored data, various protective measures can be employed, such as the implementation of data anonymization techniques, permutation methods, and data partitioning-whether vertically or horizontally.

DATA ANALYTICS PHASE

Following the collection and secure storage of data, the data processing analysis phase is initiated to derive meaningful insights. During this stage, data mining methods such as clustering, classification, and association rule mining are employed. Ensuring a secure processing environment is paramount, as data miners utilize potent algorithms capable of extracting sensitive information, thereby posing a risk of security breaches. Consequently, it is imperative to safeguard the data mining process and its outputs against potential attacks rooted in data mining activities, while also ensuring that only authorized personnel are involved in this phase.

KNOWLEDGE CREATION PHASE

Ultimately, the analytics phase yields fresh insights and valuable knowledge for use by decision-makers. The generated knowledge is deemed sensitive information, particularly in competitive environments. Organizations are vigilant in safeguarding this sensitive information to maintain a strategic distance from competitors. Additionally, they exercise caution to prevent the public release of sensitive data, such as client personal information.

RISKS IN BIG DATA LIFE CYCLE

Big data technology faces numerous security threats and attacks, primarily stemming from its reliance on data analytics techniques, including data mining algorithms. Attackers can exploit these methods to identify sensitive data, leading to potential breaches. This paper categorizes big data threats and attacks based on the four phases of the big data lifecycle. Table 1 provides an elucidation of the threats and attacks associated with each phase of big data life cycle.

Phases	**Threats and Attacks**	**Description**	**Suggested defense**
Data Collection	Phishing	These attacks involve infiltrating the data provider and collector systems to gain unauthorized access to the data during the collection phase.	Security Awareness Programs
	Spamming		
	Spoofing		
Data Storage	Attacks based on mining	Focused on specific datasets for the extraction of knowledge (Dev et al., 2012).	Divide datasets (vertically and horizontally) and non-central data storage framework.
	Attacks on	Stealing hard disks or make images of	Physical security measures

	storage devices	them	non- central data storage framework.
	Unauthorized access	Illegal access of data by people	Access Control
Data Analytics	Attacks based on mining	Employing data mining techniques to extract confidential information.	Divide datasets (vertically and horizontally) and use access control.
	Re-identification threat	Identification threats of Personal information (Jensen 2013).	Core Attribute Encryption
	Wrong result threat	Employing an inaccurate analysis process that results in erroneous findings (Jensen, 2013).	Adhere to accurate analysis procedures and document, audit, and review the entire process.
Knowledge Creation	Privacy threats	Disseminating the acquired knowledge (e.g., potential rival competitors).	Implement encryption for the obtained knowledge and adopt an access control strategy.
	Phishing and Spoofing	Decision Makers are targeted	Security Awareness Programs

Table 1: Treats Model

In the provided table, it is categorized that, the threats and attacks based on different phases of the big data lifecycle. Each phase possesses unique characteristics and is assigned distinct tasks, making them susceptible to varied threats and attacks. Specifically, the data collection phase is vulnerable to various attacks, such as phishing and spoofing, targeting individuals involved in gathering and providing data for the big data framework. Enhancing security in this phase involves implementing security awareness programs for data collection staff, educating them on compliance with security policies and procedures.

After gathering data into storage devices, awareness of potential threats and attacks becomes crucial. Hackers gaining access to stored data might employ data mining techniques to illicitly extract sensitive information, resulting in data mining-based attacks. Mitigating these attacks involves vertical or horizontal dataset division to minimize their impact and adopting a non-centralized data storage framework. Additional threats in the data storage phase include attacks on storage devices (e.g., stealing hard disks) and unauthorized access. Countermeasures for these threats include establishing physical security measures and implementing access control protocols.

In the data analytics phase, threats and attacks may target the release of sensitive data or compromise the data processing. Data mining-based attacks may uncover and release sensitive information, while correlation techniques could be employed to re-identify personal data, impacting individuals' data privacy. Protective measures for the big data framework involve dividing datasets (horizontally or vertically) and applying data encryption to core attributes with high significance. Another threat in this phase is obtaining incorrect results from the data analytics process, emphasizing the importance of following a correct analytics process and documenting it adequately.

Lastly, considering threats and attacks in the knowledge creation phase is essential for safeguarding sensitive information generated from the big data process. This knowledge is deemed confidential and should not be released publicly, especially to rival companies in a business context. Privacy threats and security attacks may target decision-makers and those with access to the final outcomes. Addressing these concerns involves

developing security policies, implementing access control procedures, and establishing security awareness programs to prevent and mitigate the impact of potential threats.

Conclusion And Future Work

Understanding the security threats and attacks associated with big data is crucial. Big data, in this context, refers to the management of vast amounts of data, and its growing utilization presents various challenges, particularly security concerns that can affect data privacy. This paper introduces a security threat model for big data, elucidating the security threats and attacks within the big data lifecycle. The big data lifecycle comprises four phases: data collection, data storage, data analysis, and knowledge creation.

During the data collection phase, where data is gathered from diverse sources, it is imperative to obtain data from trustworthy sources. The subsequent data storage phase necessitates securing collected data using reliable data storage solutions. The third phase involves data processing, demanding a focus on maintaining information assurance throughout the processing stage. Finally, the big data process produces valuable knowledge crucial for decision-makers. Organizations view this knowledge as sensitive data that must be secured and kept confidential, especially from rival entities. Given that big data heavily relies on data mining methods, potential attackers can exploit data mining to extract sensitive information. Therefore, our future efforts aim to mitigate the impact of data mining-based attacks by implementing robust security measures, such as separating storage and encrypting selected attributes of datasets.

References

[1] Bajaj, R. H., and Ramteke, P. P. L. (2014). "Big Data–The New Era of Data," International Journal of Computer Science and Information Technologies, 5(2), 1875–1885.

[2] Bakshi, K. (2012). "Considerations for big data: Architecture and approach," in Proceedings of IEEE Aerospace Conference, pp. 1–7. doi: 10.1109/AERO.2012.6187357.

[3] Cardenas, A. A., Manadhata, P. K., and Rajan, S. P. (2013). "Big Data Analytics for Security," IEEE Security & Privacy, 11(6), 74–76. doi: 10.1109/MSP.2013.138.

[4] Demchenko, Y., Zhao, Z., Grosso, P., Wibisono, A., and Laat, C. De. (2012). "Addressing Big Data Challenges for Scientific Data Infrastructure," in Proceedings of 4th International Conference on Cloud Computing Technology and Science, pp. 614–617. doi: 10.1109/CloudCom.2012.6427494.

[5] Dev, H., Sen, T., Basak, M., and Ali, M. E. (2012). "An Approach to Protect the Privacy of Cloud Data from Data Mining Based Attacks," in Proceedings of High Performance Computing, Networking Storage and Analysis, IEEE, November, pp. 1106–1115. doi: 10.1109/SC.Companion.2012.133.

[6] Govindarajan, P., and Panneerselvam, S. (2014). "Issues and challenges in big data," in Proceedings of 2nd International Conference on Science, Engineering and Management, pp. 265–272. Available at http://www.ijaert.org/wp-content/uploads/2014/04/42.pdf.

[7] Jensen, M. (2013). "Challenges of Privacy Protection in Big Data Analytics," in Proceedings of the International Congress on Big Data IEEE, June, pp. 235–238. doi: 10.1109/BigData.Congress.2013.39.

[8] Kaisler, S., Armour, F., Espinosa, J. A., and Money, W. (2013). "Big Data: Issues and Challenges Moving Forward," in Proceedings of the 46th Hawaii International Conference on System Sciences, IEEE, January, pp. 995–1004. doi: 10.1109/HICSS.2013.645.

[9] Katal, A., Wazid, M., and Goudar, R. (2013). "Big data: Issues, challenges, tools and Good practices," in Proceedings of the Sixth International Conference on Contemporary Computing (IC3), IEEE, pp. 404–409. Available at http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6612229.

[10] Kim, S.-H., Eom, J.-H., and Chung, T.-M. (2013). "Big Data Security Hardening Methodology Using Attributes Relationship," in 2013 International Conference on Information Science and Applications (ICISA), IEEE, June, pp. 1–2. doi: 10.1109/ICISA.2013.6579427.

[11] Kim, S.-H., Kim, N.-U., and Chung, T.-M. (2013). "Attribute Relationship Evaluation Methodology for Big Data Security," 2013 International Conference on IT Convergence and Security (ICITCS), IEEE, pp. 1–4. doi: 10.1109/ICITCS.2013.6717808.

[12] Marchal, S., Jiang, X., State, R., and Engel, T. (2014). "A Big Data Architecture for Large Scale Security Monitoring," in Proceedings of the International Congress on Big Data IEEE, June, pp. 56–63. doi: 10.1109/BigData.Congress.2014.18.

[13] Miloslavskaya, N., Senatorov, M., Tolstoy, A., and Zapechnikov, S. (2014). "Big Data Information Security Maintenance," Proceedings of the 7th International Conference on Security of Information and Networks - SIN '14, New York, New York, USA: ACM Press, pp. 89–94. doi: 10.1145/2659651.2659655.

[14] Sagiroglu, S., and Sinanc, D. (2013). "Big data: A review," in Proceedings of the International Conference on Collaboration Technologies and Systems (CTS), pp. 42–47. doi: 10.1109/CTS.2013.6567202.

[15] Singh, S., and Singh, N. (2012). "Big Data analytics," in Proceedings of the International Conference on Communication, Information & Computing Technology (ICCICT), pp. 1–4. doi: 10.1109/ICCICT.2012.6398180.

[16] Smith, M., Szongott, C., Henne, B., and Voigt, G. Von. (2012). "Big Data Privacy Issues in Public Social Media," in Proceedings of the Digital Ecosystems Technologies (DEST), IEEE, pp. 1–6.

[17] Wu, C., and Guo, Y. (2013). "Enhanced user data privacy with pay-by-data model," in Proceedings of the International Conference of Big Data, IEEE, October, pp. 53–57. doi: 10.1109/BigData.2013.6691688.

[18] Xu, L., Jiang, C., Wang, J., Yuan, J., and Ren, Y. (2014). "Information Security in Big Data: Privacy and Data Mining," The Journal for rapid open access publishing, 2, 1149–1176. doi: 10.1109/ACCESS.2014.2362522.