Big Data LifeCycle: Threats and Security Model

Emergent Research Forum papers

Yazan Alshboul
Dakota State University
yaalshboul@pluto.dsu.edu

Yong Wang
Dakota State University
Yong.wang@dsu.edu

Raj Kumar Nepali
Dakota State University
rknepali@pluto.dsu.edu

Abstract

Big data is an emerging term referring to the process of managing huge amount of data from different sources, such as, DBMS, log files, postings of social media. Big data (text, number, images… etc.) could be divided into different forms: structured, semi-structured, and unstructured. Big data could be further described by some attributes like velocity, volume, variety, value, and complexity. The emerging big data technologies also raise many security concerns and challenges. In this paper, we present big data lifecycle framework. The lifecycle includes four phases, i.e., data collection, data storage, data analytics, and knowledge creation. We briefly introduce each phase. We further summarize the security threats and attacks for each phase. The big data lifecycle integrated with security threats and attacks to propose a security threat model to conduct research in big data security. Our work could be further used towards securing big data infrastructure.

Keywords

Big data, big data lifecycle, threats and attacks, threat model.

Introduction

Big data emerged in the last few years to meet the requirements and challenges of the growing size of data. Big data as a term refers to the process of managing huge amount of data that comes from several sources like DBMS, log files, posting to social media, and sensor data (Bajaj et al. 2014). At the first glance when we hear big data, we think about the huge amount of data that we should store and process. In fact, huge volume of data is a big data pertained attribute where it exceed Exabyte (10^{18}) which needs special storage solutions, high performance data processing, and special analytics capability (Kaisler et al. 2013). Big data is a pool of complex datasets that may include different data types (text, number, images, and videos) of large volume that beyond the traditional database management systems (Govindarajan et al. 2014).

In particular, big data has three main attributes namely: volume, velocity, and variety. Beside the three Vs, other attributes pertained to big data like value and complexity attributes (Kaisler et al., 2013; Katal et al. 2013). Volume attribute refers to the amount of data. Generally speaking, big data has large volume of data that is beyond the traditional storage solutions. According to Bajaj et al. (2014), 90 percent of the current data in the world has been formed in the last two years with average of 2.5 quintillions of data bytes being created daily. The velocity attribute of big data refers to the speed rate of generating and processing data (Bajaj et al. 2014). Currently, data and information are generated and processed at high-speed rate producing huge volume of knowledge added to the knowledge base. This velocity rate of big data requires higher processing capabilities than traditional systems. Furthermore, the term velocity refers to high speed of data movements between data storage over networks (Bajaj et al. 2014).

Another big data attribute is the variety. Variety in big data refers to the variety of resources that generate data of different types and different formats (Bajaj et al. 2014; Govindarajan et al. 2014; Kaisler et al. 2013; Katal et al. 2013). Data resources could be from digital pictures and videos, social media, sensors data, healthcare data records, text, log files, tweets, and purchase transaction records. In other words, big data consists of different data formats: structured, unstructured, and semi-structured.

Other two attributes related to big data are value and complexity (Kaisler et al. 2013). The value attribute in big data refers to the worthy of information (the knowledge) that could be generated from processing and analyzing big data. This created knowledge is helpful and supportive for decision-making process (Katal et al. 2013). The complexity attribute refers to the complication of relationships and complexity links in big data structure. In this regards, we can imagine how complex it is when few changes occur in big data that may lead to large number of changes (Katal et al. 2013).

In terms of security and privacy perspective, we should look at big data security from different angles and perspective. We should think about how to protect data itself, the process of the big data, and the output of big data process. In this regards, Kim et al. (2013) argue that security in big data refers to three matters: data security, access control, and information security. Furthermore, Xu et al. (2014) present a big data security model considering the user role of security in different phases of big data process. However, most of the previous big data security studies do not focus on the threats and attacks that face big data environment. Furthermore, there is less focus on the big data lifecycle and how to correlate the threats and attacks based on the lifecycle model.

In order to secure big data environment, it is important to identify the threats and attacks of the big data within its lifecycle. Threats and attacks identification helps security community to develop security defenses against these threats. Up to our knowledge, there is no previous studies address the lifecycle of the big data and the associated threats and attacks. Therefore, our paper presents a big data lifecycle model where it consists of four phases: data collection, data storage, data processing, and Knowledge creation. Furthermore, we summarize the security threats and attacks in each phase of the lifecycle. The presented big data lifecycle integrated with security threats and attacks provides a security thread model to conduct research on big data security. The threat model could be further used towards securing big data infrastructure and provide a clear foundation for further research to secure big data.

Literature Review

Big data emerged in the last few years to meet the requirements and challenges of the growing size of data. There is not much work about big data and security in the literature. Some articles in literature tackle the definition, characteristics, and challenges of big data. Katal et al. (2013) introduces the big data as a new technology and discuss some issues and challenges for using big data technologies. Sagiroglu and Sinanc (2013) present a big data overview paper discussing the content, characteristics, scope, advantages, challenges, and the privacy of big data. Other research papers provide a debate about the new technology of big data and its challenges and scopes (Bakshi, 2012; Demchenko et al, 2012; Singh & Singh, 2012).

In terms of security and privacy, some research articles argue the privacy issues and security challenges of the big data era. In their research, Smith et al. (2012) discuss personal privacy on social network and how social web users can control their privacy. Kim et al. (2013) discuss big data security and provide a security methodology to hardening and protecting big data security effectively through protecting the selected attributes. Jensen (2013) discusses privacy challenges in big data and how to control big data process under privacy complaint. Big data analytics can be used to improve the security through collecting and analyzing enterprise’s data (structured and unstructured) (Cardenas et al. 2013).

The emergent of big data technologies attracts researchers to think about protecting the new data framework. In this regards, information security maintenance is required for any organization’s information systems infrastructure considering big data technologies (Miloslavskaya et al. 2014). Marchal et al. (2014) propose a security model to analyze large amount of data from security perspective to monitor local enterprise network, perform network intrusions detections and preventions tasks, and perform forensics analysis. Furthermore, another research focuses on the user role in securing big data infrastructure through the proposed security model (Xu et al. 2014). They argue four types of users' role in big data environment: data provider, data collector, data miner, and decision maker (Xu et al. 2014).

On the other side, several threats and attacks are threatening big data technology. A data mining based threat is discussed by (Dev et al. 2012). According to them, this type of threat exploits data mining techniques and methods to extract sensitive data and valuable information. Another threat to big data is data privacy and releasing sensitive data that may harm persons or organizations like re-identification threat and wrong results threats (Jensen 2013). Wu and Guo (2013) argue that privacy and information assurance is a big concern in big data environment where extracting personal or sensitive data can harm persons and organizations and lead to several business problem. Up to our knowledge, there is no study addresses the threats and attacks in one model in terms of big data lifecycle. Therefore, our study can be used as a foundation for future security research to secure big data environment.

Big Data Security Lifecycle

In this section, we present big data security lifecycle model and the main components of any big data framework. We extend our model from (Xu et al. 2014). They address big data from user role perspective where they argue four types of users' role in big data environment: data provider, data collector, data miner, and decision maker. However, our model addresses the phases of the big data lifecycle. Our model consists of four phases in big data framework consists of data collection phase, data storage phase, data processing and analysis, and knowledge creation. Figure 1 presents the main elements in big data lifecycle.

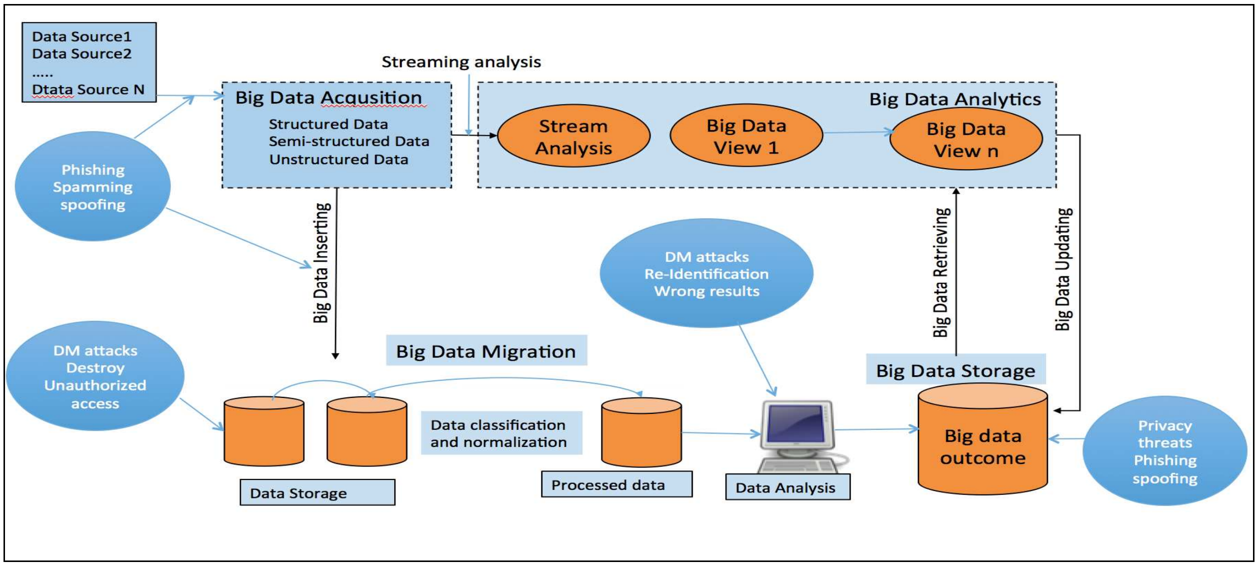

Figure 1: Big Data Lifecycle Threat Model

Data Collection Phase

In data collection phase, data from different sources comes with different formats: structured, semi-structured, and unstructured. From a security perspective, securing big data technology should start from the first phase of the lifecycle. It is important to gather data from trusted sources and make sure that this phase is secured and protected. In fact, we need to take some security measures in order to keep data from being released. Some security measures can be used in this phase like limited access control (for those who receive data from data provider) and encrypting some data fields (personal information identifier).

Data Storage Phase

In data storage phase, the collected data is stored and prepared for being used in the next phase (data analytics phase). As the collected data may contain of sensitive information, it is essential to take sufficient precautions during data storing. In order to guarantee the safety of the collected data, some security measures can be used like data anonymization approach, permutation, and data partitioning (vertically or horizontally).

Data Analytics Phase

After collecting data and storing it in secured storage solutions, data processing analysis is performed to generate useful knowledge. In this phase, data mining methods such as clustering, classification, and association rule mining are used. It is crucial to provide secure processing environment. In fact, data miners use powerful data mining algorithms that can extract sensitive data. Thus, a security breach may happen. Therefore, data mining process and it's output must be protected against data mining based attacks and make sure that only authorized staff work in this phase.

Knowledge Creation Phase

Finally, the analytics phase comes up with new information and valued knowledge to be used by decision makers. The created knowledge is considered as sensitive information especially in a competition environment. Organizations take care of their sensitive information to be far away from their rivals. Further, they aware of their sensitive data (e.g. client personal data) not to be publicly released.

Big Data Threats and Attacks

Big data technology is exposed to many security threats and attacks. Big data threats and attacks are mainly derived from the characteristics of big data technology that rely on data analytics techniques including data mining algorithms. In fact, attackers can also use data mining methods and procedures to find out sensitive data and release it to public and thus data breach happens. In this paper, we classify threats and attacks of big data in terms of the four phases of big data lifecycle. Table 1 explains that the threats and attacks at each phase.

Phases	Threats and attacks	Description	Suggested defense
Data Collection	Phishing	These attacks are hacking data provider and collector to get an access to the data in the collection phase.	Security awareness program
	Spamming		
	Spoofing		
Data storage	Data mining based attacks	Targeted datasets to extract knowledge (Dev et al. 2012).	Divide datasets (vertically and horizontally) and non-central data storage framework.
	Attacks on data storage devices	Stealing hard disks or make images of them	Physical security measures non-central data storage framework.
	Unauthorized data access	People access data illegally	Access control
Data analytics	Data mining based attacks	Using data mining methods to extract sensitive knowledge.	Divide datasets (vertically and horizontally) and use access control.
	Re-identification	Identification threats of personal information (Jensen	Core attribute encryption.

	threat	2013).	
	Wrong result threat	Using incorrect analysis process, which lead to incorrect results (Jensen 2013).	Follow correct analysis procedures and document, audit, and review the process.
Knowledge creation	Privacy threats	Releasing the resulted knowledge (ex. Rival competitors)	Adopt encrypt the resulted knowledge and adopting access control strategy.
	Phishing and spoofing	Decision makers are targeted	Security awareness programs

Table 1: Threats Model.

In the above table, we classify the threats and attacks based on the big data lifecycle phases. Each phase has special characteristics and assigned different tasks, thus each phase is vulnerable to different threats and attacks. In this regards, data collection phase is vulnerable to several attacks like phishing and spoofing attacks. These kinds of attacks are targeting people who work in gathering and providing data to big data framework. One way to improve security in this phase is to provide security awareness programs to data collection staff and teach them how to comply with security policies and procedures.

After we the gathered data in data storage devices, we have to be aware of some threats and attacks. In this regards, hackers who get access to data in storage devices may use data mining techniques to extract sensitive data and use it illegally. This kind of attacks called data mining based attacks. In order to deal with kind of attacks, we may divide the datasets vertically or horizontally to reduce the impact of this attacks and adopt non-central data storage framework. There are other threats related to data storage phase like attacks on data storage devices (ex. Stealing hard disks) and unauthorized access attacks. In this regards, we can build a physical security guard and develop access control protocols.

In data analytics phase, some threats attacks may happen to release sensitive data or harm the data process. Data mining based attacks may occur to discover and release sensitive information or correlation techniques could be used to re-identify personal information which impact people data privacy. In order to protect big data framework from this kind of attacks, we may adopt some defense procedures like dividing datasets into several parts (horizontally or vertically) and perform data encryption to the core attributes (attributes with high weight). Other threat in this phase is getting incorrect results from the data analytics process. Therefore, it is important to follow correct analytics process and document it.

Finally, we have to consider the threats and attacks in the knowledge creation phase and how to protect it. In fact, the created knowledge from big data process is considered sensitive information that needs not to be released to the public and especially to rival companies in the business context. Some privacy threats and security attacks that may target the decision makers and those who have access to the final outcome of the big data process. Therefore, we need to develop security policy and follow access control procedures besides developing security awareness programs to prevent and mitigate the impact of any threat.

Conclusion and Future Work

It is essential to be aware of security threats and attacks of the big data. Big data as a term refers to the process of managing huge amount of data. With the increasing usage of big data, several challenges are raised and especially security challenges that impact data privacy. In this paper, we present a security thread model for big data and explain the security threats and attacks of big data in terms of the big data lifecycle. Big data lifecycle consists of four phases: data collection, data storage, data analysis, and knowledge creation. Data collection phase consists of collecting data from different sources. In this phase, it is important to collect data from trusted data sources. In data storage phase, the collected data need to be stored in secure data storage solutions. The third phase of our model is data processing. We need to make sure how to keep information assurance during data processing. Finally, we have the output of the big data process. In fact, big data results represent significant knowledge that is important for decision makers. Decision makers and organizations consider the generated knowledge of the big data process as a sensitive data that need to be secured and unreleased to the public especially the rival organizations. Big data basically rely on data mining methods and attackers can use data mining to extract sensitive data. Therefore, our future work is to reduce the impact of the data mining based attacks by developing effective security measures like storage separation and encrypt selected attributes of the datasets.

REFERENCES

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), pp. 1875–1885.

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

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

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

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 *Proceeding of High Performance Computing, Networking Storage and Analysis*, Ieee, November, pp. 1106–1115 (doi: 10.1109/SC.Companion.2012.133).

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).

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

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

Katal, A., Wazid, M., and Goudar, R. 2013. "Big data: Issues, challenges, tools and Good practices," in *Proceeding 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).

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).

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).

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

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).

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

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

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

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

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), pp. 1149–1176 (doi: 10.1109/ACCESS.2014.2362522).