Big Data Security Issues in Three Perspectives: A Review

Elias Bassa Badacho
Department of Computer Science (Lecturer), Wolaita Sodo University, Ethiopia

Abstract
Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes. With regard to the definition of big data, IBM Company uses volume, velocity, variety, value and veracity as 5Vs to summarize the concept of big data. There are different types of big data, for example, structured, semi-structured and un-structured data. The contents of big data can be text data, audio data, video data and still image and it indicates that the big data may have diverse data types as well as data qualities. Big data has variety of sources such as healthcare center, commercial system, industries, social media, telecommunication, transportation, sensor machines and others. In this paper, I reviewed three the most security challenging perspectives and I studied lack of concentrations in these areas by most research works. To confirm security in the big data platforms, it is critical to ascertain the data rendering points and their security techniques to safeguard the data in this pacing digital world. Then I envisage directions for the future research. In this paper, I have reviewed the big data sources and its security issues in the three directions such as data at rest, data at communication and data in process/use.

Keywords: Big Data, Big Data Security, Big Data source, Attribute based encryption, storage path, Transport layer security, secure shell, Cloud service Provider
DOI: 10.7176/CEIS/12-3-01
Publication date: November 30th 2021

1. Introduction

The term "Big data" is believed to be originated from the Internet search companies who had to query loosely structured very large distributed data [1]. On the premise of this definition, the properties of big data are reflected by 3V's, which are, volume, velocity and variety. But later studies pointed out that the definition of 3Vs is insufficient to explain the big data we face now [3]. Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes [7]. With regard to the definition of big data, *IBM Company uses volume, velocity, variety, value* and *veracity* as 5*Vs* to summarize the concept of big data [12]. Due to recent technological development, the amount of data generated by internet, social networking sites, sensor networks, healthcare applications, and many other companies, is drastically increasing day to day and all the enormous measure of data produced from various sources in multiple formats with very high speed is referred as big data [2]. The contents of big data can be text data, audio data, video data and still image. It indicates that the big data may have diverse data types as well as data qualities. This diverse nature of big data yields security and privacy issues on the data within data life cycles such as data in use, data at transit and data at storage on the cloud computing environments. Data has great commercial value for Internet service providers, but the analysis and application of data will be more complex and difficult to manage, and personal privacy will be threatened [3]. Consequently, cloud platforms that handle big data that contain sensitive information are required to deploy technical measures and organizational safeguards to avoid data protection breakdowns that might result in enormous and costly damages [4]. With the rapid development of the Internet, people put a lot of datasets on the Internet every day and this paves the ways to collect the information on the Internet and then conduct illegal activities such as reselling, fraud, stealing, damaging, etc.., not only for people [3].

The 5Vs' concepts of Big Data

- **Volume:** It is the large amounts of data generated every second via emails, twitter messages, videos, sensor data, airbus, YouTube and other digital media.
- **Velocity:** It is the speed of data moving in and out from the data management systems in the big data environments.
- **Variety:** There are different data formats in terms of structured or unstructured in big data management system.
- **Value:** It is the insights/understandings we can reveal within the big data.
- **Veracity:** It is the trustworthiness of the big data.

This paper presents an overview of the research on security of big data in the three perspectives such as security of data at storage, at transit and in use.

2. Big Data: Source of Big Data

There are various sources for big data, such as Social media, Healthcare system, Transportation system, telecommunication, commercial system, Industries, education system and others.

2.1 Healthcare Source

In fact, digitization of health and patient data is undergoing a dramatic and fundamental shift in the clinical, operating and business models and generally in the world of economy for the foreseeable future [5]. So the healthcare organization collects, transforms, models and visualizes huge amount of sensitive data contents every day, every month and every year all over the world. To keep the security and privacy of the patients and all users in this organization, it requires most sophisticated security techniques, policies, and software and hardware tools. The healthcare industry is recording the data in electronic medical records and images, which is used for short-term health monitoring and long-term epidemiological research programs. Big Data is a collection of large and complex datasets and getting adopted in the healthcare significantly, security and privacy issues in healthcare becomes necessary to deal with [6, 7].

2.2 Social Media and Entertainment: Big Data Sources

The entertainment industry has moved to digital recording, production, and delivery in the recent times and is now collecting large amounts of rich content. At the same ways, the social networks/media such as Facebook, Twitter, Whatsapp, Viber, Imo, Telegram, Instagram, youtube, yahoo, Tango, whatscall etc., are used by hundreds of millions of Internet users all over the world to provide free services in the two service providing manners such as online and offline connections [6, 7]. There are several the biggest internet companies such as ebay, alibaba, amazon, google, Microsoft, yahoo, forcesale and others which can produces enormous big data within every Pico-seconds.

2.3 Transportation System: Big Data sources

Now day, there are diverse digital transportation systems that are generating big data in daily bases. The transportation system may include airline reservation system, train/railway reservation system, bus reservation and other sea transportation systems.

2.4 Telecommunication System: Big Data Source

In this globe, telecommunication system is the oldest and the most popular digital technology which engendering large amount data in different data formats from all over the world within fractions of minutes. There is a tremendous amount of *geospatial* (e.g., GPS) data, such as that created by cell phones, that can be used by applications like Four Square to help you know the locations of friends and to receive offers from nearby stores and restaurants [7].

2.5 Commercial system: Big Data Sources

In each and every day, the commercial system such as banks, supermarkets, credit associations and business companies carries out several transactions. Due to these reasons, this system has great potentials to produce big data in altered ways.

2.6 Industries and Education System: Big Data Sources

In this 21st century, there are so many manufacturing, mining, weather forecasting and service providing industries, these all industries yield tremendous data contents in diverse data formats at any time at anywhere. In like ways, the academia also produces the huge amount of data in the public and privates universities and colleges all over the world.

3. Big Data Security

Privacy and security are the most important aspects in the areas of Big Data and the technologies generating it. Security is the practice of defending information and information assets through the use of technology, processes and training from unauthorized access, disclosure, disruption, modification, inspection, recording, destruction, ..,etc and it concentrates more on protecting data from malicious attacks and the misuse of stolen data for profit [2]. Security is the way to keep the confidentiality of data, integrity of data and availability of data in the big data life cycles. The conventional security mechanisms to protect data can be divided into four categories. They are file level data security schemes, database level data security schemes, media level security schemes and application level encryption schemes [11].

3.1 Security issues and Three Perspectives

Actually, sensitive data can easily be leaked if there is no effective protection in its lifetime, including data collection, storage and management, transport, analysis, and data destruction [1]. Security focuses on protecting data from pernicious attacks and stealing data for profit [5]. According to the Big Data Working Group at the Cloud Security Alliance organization there are, principally, four different aspects of Big Data security includes infrastructure security, data privacy, data management, and integrity and reactive security and this division of Big Data security into four principal topics has also been used by the International Organization for Standardization in order to create a security standard for security in Big Data [8]. At present, people have serious problems with the security of big data, and think that big data is not safe. The security problems of personally-carried intelligent terminals are also very worrying [3]. Data encryption technology is an important means to protect data confidentiality, it safeguards the confidentiality of the data, but it cut down the performance of the system at the same time and the homomorphic encryption has become a research hotspot in data privacy protection [9]. There are many widespread techniques in cryptography which are used to provide the security for the data and some of them are Homomorphic encryption (HE), Verifiable Computation (VC) and Multi-Party Computation (MPC) which can be deployed on trusted, semi-trusted and untrusted clouds [10]. Big data contains a wealth of information resources, all professions and trades have great demand of the data, so we must manage access rights of big data carefully. Access control is an effective means to achieve controlled sharing of data, but in big data environment, the number of users is huge, the authority is complex, and a new technology must be adopted to realize the controlled sharing of data [9]. Virtual barriers such as firewalls, secure socket layer and transport layer security are designed to restrict access to data [10]. HIPPA (Health Insurance Portability and Accountability Act): It is the federal Health Insurance Portability and Accountability Act of 1996. The primary goal of the law is to make it easier for people to keep health insurance, protect the confidentiality and security of healthcare information and help the healthcare industry control administrative costs [6]. While healthcare organizations store, maintain and transmit huge amounts of data to support the delivery of efficient and proper care, the downsides are the lack of technical support and minimal security [5]. The three perspectives of security issues are as follows:

3.1.1 Big Data Security at Rest

Now day, the cloud technology is increasingly being used to store and process big data in different sort of cloud platforms. Data at rest refers to storing data remotely in the cloud storage platforms, on cloud Service Provider side and storing data locally in the database, on data owner side. Hence, for providing the security for the data at such cases we use the following technique. Initially the big data is divided into sequenced parts and then stores them among multiple Cloud storage service providers and in cloud computing, big data storage services represent a basic function for their tenants [10]. Storing high volume data is not a big challenge due to the advancement in data storage technologies such as the boom in cloud computing. However, securing the data is very challenging [11]. Lastly, to secure data at storage, we have to impose a number of security techniques, for instance, Attribute based encryption techniques, homomorphic encryption algorithm, storage path/file directory encryption and using hybrid cloud platforms [2]. Assuring the confidentiality of the data objects, authorizing data modifications and ensuring that resources are available when needed [4].

3.1.2 Big Data Security at Transit

However, the usage and sharing of user private data are lacking in specifications, and lack of supervision, mainly rely on the self-discipline of enterprises, which leads to the failure of client to determine the purpose of their private information [1]. A large number of cloud services require users to share private data for data analysis or mining and under such cases the privacy can be provided using a scheduling mechanism Optimized Balanced Scheduling (OBS) [10] to apply the Anonymization on the sensitive field only depending upon the scheduling. In the case of data security at communication phase, we can use different types of security techniques or security algorithms, such as application layer and Transport layer security techniques. The transport layer security (TLS) and its predecessor, secure sockets layer (SSL), are cryptographic protocols that provide security for communications over networks such as the Internet. Not only that, we can again implement IOT device level security mechanisms to keep the security of user data in Big Data Environments. Hashing techniques like SHA-256 and Kerberos mechanism based on Ticket Granting Ticket or Service Ticket can be also implemented to achieve authentication [5]. Lastly, to protect data at transmission, we have to analyze the DNS traffic, HTTP traffic, IP flow records and others and make decisions regarding data transmission and in addition to aforementioned techniques, any has to consider web security (IPsec), Secure Sockets Layer security Techniques, Transport layer Security Techniques, HTTPS and secure shell (SSH) techniques at the level of data transmission.

3.1.3 Big Data security in use

In the big data environment, data can be accessed by more than a millions of users at a time within the data management systems. Data integrity is one of the most critical elements in any information system. Generally, data integrity means protecting data from unauthorized deletion, modification, or fabrication. Managing entity's admittance and rights to specific enterprise resources ensures that valuable data and services are not abused, misappropriated, or stolen. Data security in use refers to the way how to protect the users' data during data manipulation processes in the big data platforms. At the same time [3], as the transmission of information, due to the weak supervision of data information, lack of technical support, imperfect supervision system, and the vulnerability of information loss, the use of data information is not of high value and data is reduced. The data processing phase incorporates Privacy Preserving Data Publishing (PPDP) and knowledge extraction from the data [2]. Eventually, the goal of keeping security data during data in process or use is bring the data confidentiality, data integrity and data availability to the end users in safe manner in big data environments and the authentication and authorization techniques should be employed.

4. Conclusion and Future Directions

In this paper, I reviewed big data security issues and its techniques in three directions, i.e big data security at storage, big data security at transmission and big data security in use. Again I reviewed the sources of big data, for instance, social media, healthcare, commercial system, telecom, transportation, IOT devices, and others. In this paper, I reviewed that no one has been identified and conducted research for these three specific areas and their best fitting security techniques separately. But so many research works are done regarding big data security and privacy issues in the randomly manner. In the future, anyone who is interested can conduct security researches in these three specific points separately and can propose their best suiting security techniques. And in the same way, can conduct their counterpart privacy preserving algorithms and techniques.

5. Reference

1. Challenges and techniques in big data security and privacy: A review by Rongxin Bao1 Zhikui Chen1 Mohammad S. Obaidat2, on March 2018.
2. Big data privacy: a technological perspective and review by Priyank Jain, Manasi Gyanchandani and Nilay Khare.
3. Big Data Security and Privacy Protection by Dongpo Zhang, 8th ICMCS 2018, Advances in Computer Science Research, Volume 77.
4. Big Data Security And Privacy Issues In The Cloud by Ali Gholami and Erwin Laure, IJNSA Vol.8, No.1, January 2016.
5. Big healthcare data: preserving security and privacy by Karim Abouelmehdi, Abderrahim Beni Hessane and Hayat Khaloufi.
6. Big Data Security and Privacy Issues: A Review by Nitin Kr. Agrawal and Dr. Aprna Tripathi, International Journal of Innovative Computer Science & Engineering Volume 2 Issue 4; September-October-2015; Page No.12-15.
7. Big Data Analytics: Concepts, Technologies, and Applications by Hugh J. Watson, *University of Georgia:* Communications of the Association for Information Systems *Article 65, volume 34.*
8. Main Issues in Big Data Security by Julio Moreno, Manuel A. Serrano and Eduardo Fernández-Medina.
9. Big Data and Information Security by Gang Zeng: *International Journal of Computational Engineering Research, June 2015.*
10. Security and Privacy in Big Data Analytics by P.Shobha Rani1, Vigneswari D: International Journal on Intelligent Electronic System, Volume.10, No 2, July 2016.
11. Protection of Big Data Privacy by Abid Mehmood, Iynkaran Natgunanathan, Yong Xiang, *Senior Member, IEEE*, Guang Hua, *Member, IEEE*, and Song Guo, *Senior Member, IEEE:* DOI 10.1109/ACCESS.2016.2558446, IEEE Access.
12. Security and Privacy Issues of Big Data by José Moura1, 2, Carlos Serrão1.