Academic Journal of Nawroz University (AJNU), Vol.11, No.3, 2022

https://doi.org/10.25007/ajnu.v11n3a1446

A Review on Big Data Analytics between Security and Privacy Issue

Renas Rajab Asaad[1], Nisreen Luqman Abdulnabi[2]
[1]Department of Computer Science, Nawroz University, Duhok, Kurdistan Region - Iraq
[2]Technical Collage of Administration, Duhok Polytechnic University, KRG - Iraq

ABSTRACT

Because of technology, a large number of portable devices, and the great use of networks and the need for them in education, industry, and politics. This widespread has led to the emergence of a huge amount of data, and securing this data has become a very important matter. Big data refers to a very large number of data collected from different sources, this huge amount of data gives us deep insight and the benefit of this is to improve the decision-making process and thus lead to better customer service. Securing this huge amount of data (big data) is very difficult. Because of the large number of data and the continuous flow of this data that's come from different sources. Big data is a term that refers to a large, diverse, and complex amount of data. Big data analytics is a sophisticated and advanced technology used to analyze large, different, and diverse data in order to understand this type of data. However, there are many issues when using and analyzing big data, including the problem of security and privacy. This paper focuses on security and privacy problems and focuses on the biggest privacy issue Big data. The aim of this paper is to study the aspects of security and privacy and to make a comparison of the most important data security methods and the biggest privacy issues.

Keywords: *Big Data, Privacy Issue, Cyber Security, Data Science, Analysis.*

1. Introduction

Big Data is a new concept that is used to manage datasets that are too large for standard software tools to capture, handle, and analyze in a timely manner. Every two years, the amount of data to be analyzed is predicted to double.[1] All of this information comes from a variety of places, including social media, sensors, scientific applications, surveillance, video and image archives, Internet search indexing, medical records, corporate transactions, and system logs. Big data is gaining traction as the number of devices connected to the so-called "Internet of Things" (IoT) continues to grow at unfathomable rates, generating massive volumes of data that must be turned into useful information. It is also extremely useful. Furthermore, buying on-demand additional computing power and storage from public cloud providers to undertake demanding data-parallel processing is increasingly popular. The volume, variety, and wide-area deployment of system infrastructure to serve Big Data applications can possibly exacerbate security and privacy concerns.[2]

Traditional security solutions designed to private computer infrastructures, confined to a well-defined security perimeter, such as firewalls and demilitarized zones (DMZs), are no longer viable as Big Data expands with the help of public clouds. Security functions using Big Data must work across a heterogeneous mix of hardware, operating systems, and network domains. The abstraction capacity of Software-Defined Networking (SDN) appears to be a very crucial property in this puzzle-type computing environment, as it can enable the effective deployment of Big Data secure services on top of the heterogeneous infrastructure. Because SDN isolates the control (higher) plane from the underlying system infrastructure being supervised and regulated, it introduces abstraction. In contrast to traditional networks, where administrators must codify functionality in terms of low-level device configuration, separating a network's control logic from the underlying physical routers and switches that forward traffic allows system administrators to write high-level control programs that specify the behavior of an entire network. SDN allows for the intelligent management of security operations in a logically centralized controller, simplifying aspects such as security rule implementation, system (re)configuration, and system development. A hierarchy of controllers and/or the use of redundant controllers helps minimize the robustness downside of a centralized SDN solution, at least for the most important system functions to be handled.[3]

2. Literature Review

2. Securing Big Data

the huge number of data needs security. because this data is for companies or individuals or whatever, so it needs security to preserve. Since the traditional methods of preserving this data are no longer sufficient. Therefore, new methods and new programs were developed and used, and a focus was placed on its infrastructure to preserve this huge amount of data. When it comes to big data security, the most important security problems must be mentioned. Therefore, following the most important security issues were mentioned.[4]

2.1. Secure Computations

One of the benefits of using big data technologies is to process a large amount of data by distributed programming frameworks. Security protection is not good for these distributed frameworks like Map Reduce. processed and split the data in Map Reduce, by a mapper and allocated storage. The mappers do not have an extra layer of security which affect the data that

are in the process if the settings were not changed to overcome this challenge. Obviously, that will make it hard to ensure the security of the data as well as to identify the untrusted mappers. Due to the importance of data, it must be preserved by securing the arithmetic operations that are processed in distributed programming frameworks, see bellow diagram.[5]

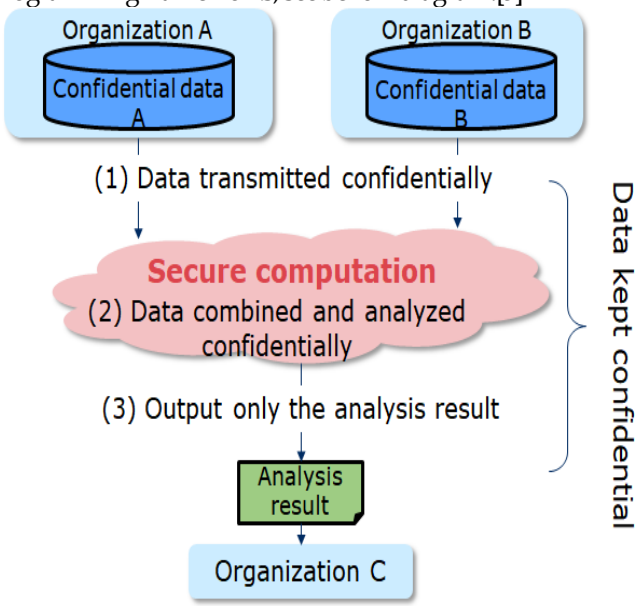


Figure 1: Securing Computation

2.2. Protecting Data and Transaction Logs

Anonymized data need special treatment prior to using them, to eliminate the security risk of accessing the anonymized data. For an instant, when anonymized data are mixed for a certain process with other separate data that are not really gone under security test, it will make it easier to access and identify the individuals of the anonymized data. This means that having in place safely established rules for using and combining the anonymized data will add another layer of security for the individuals, see bellow diagram.[6]

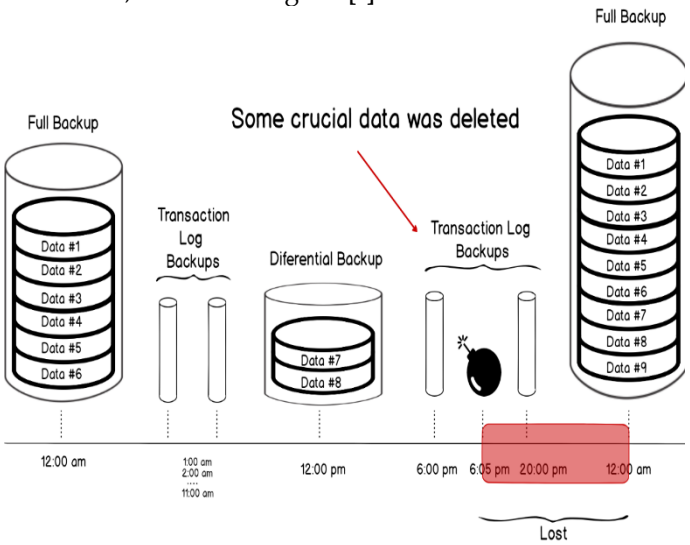


Figure 2: Transaction Logs

2.3. Validation of Inputs from Endpoints

Big data is collected from different sources or inputs, including endpoints. The inputs may come from untrusted endpoints, so it will affect the results or outputs. Therefore, all inputs must be verified as coming from trusted sources to get correct outputs.[7,8]

2.4. Secure Non-Relational Data Stores

When big data uses insecure and weak data stores and there is no credibility between client and server such as Non-Relational data stores like NoSQL, it will cause threats to privacy and pose a great risk to data, as shown in bellow table.[9,10]

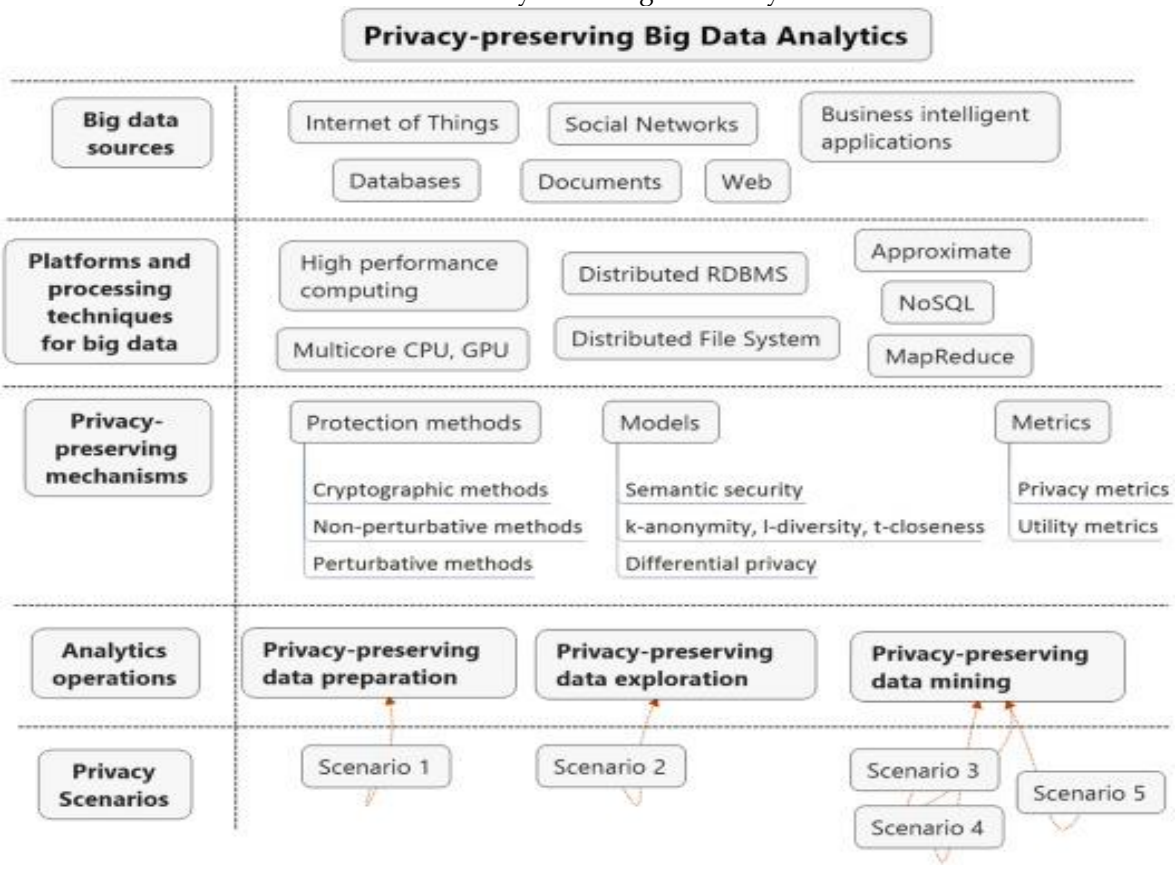
Table 1: Relation vs Non-Relation Data Store

		Relational		Non-Relational
Analytics	Proprietary Storage	Amazon Redshift, EMC Greenplum, HP Vertica, IBM Netezza, Oracle, Teradata — MPP		
	Hadoop Storage	Cloudera Impala, Presto, Hive — SQL-on-Hadoop		MapReduce
Operational		Traditional SQL	NewSQL	NoSQL
	Proprietary Storage	Oracle, DB2, SQL Server, MySQL	User-Sharded MySQL, NuoDB, Clustrix — On-Disk; MemSQL, VoltDB — In-Memory	Key Value: Aerospike, Riak; Column Family: Cassandra; Document: MongoDB; Graph: Neo4j, InfiniteGraph
	Hadoop Storage		Splice Machine — On-Hadoop	Column Family: HBase

2.5. Privacy-Preserving Data Analytics

When using big data analysis techniques, privacy must be taken into account because it is a very important issue. Due to a large amount of data, the collection and analysis of this amount of data may lead to a violation of user privacy. The basic goal of organizations using big data analysis is to improve customer service, but they need to ensure user protection while doing so, as shown in bellow table.[11]

Table 2: Privacy-Preserving Data Analytics



2.6. Access Control

Big data deals with a variety of data, including sensitive personal data. To protect this data, must implement an access control policy, which is means there is no one can access such data without permission to access it. This is to maintain personal data and privacy.[12]

2.7. Real-Time Security Monitoring

Data security monitoring is very essential especially when the monitoring is in real-time to maintain the data infrastructure you are dealing with. It is difficult and there are challenges due to the alerts generated which contain a large number of false positives. This is the reason why companies struggle to monitor data security in actual time.[13]

3. Big Data Privacy Issue Concerns

Every day, big data analytics are being used for an ever-increasing range of purposes. these different new methods of utilizing analytics can bring a good improvement for the business. Retailers, for example, are successfully employing big data analytics to predict the hottest things each season and the geographic areas with the highest demand, to mention a few applications.

Big data has strong analytics, which means that, in addition to all the wonderful commercial opportunities, it also creates a slew of new privacy problems. Here are ten of the most serious privacy threats, explained in figure 3.[14,15]

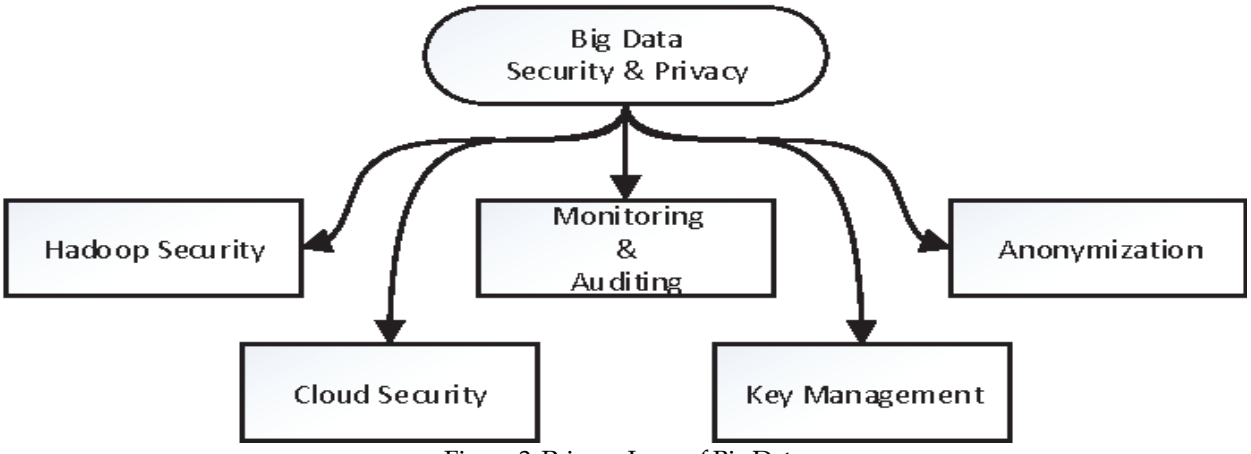


Figure 3: Privacy Issue of Big Data

3.1. Privacy Breaches and Embarrassments

The actions taken by means of groups and different businesses as an end result of big data analytics might also breach the privateness of those involved, and lead to embarrassment and even they are losing their jobs. Consider that some retail businesses have used big data evaluation to predict such intimate private important points as the due dates of pregnant shoppers. In such instances subsequent advertising things to do resulted in having individuals of the family find out a household member used to be pregnant earlier than she had told anyone, ensuing in an uncomfortable and detrimental household situation. Retailers, and different kinds of businesses, need to now not take moves that end result in such conditions, see bellow chart. [16]

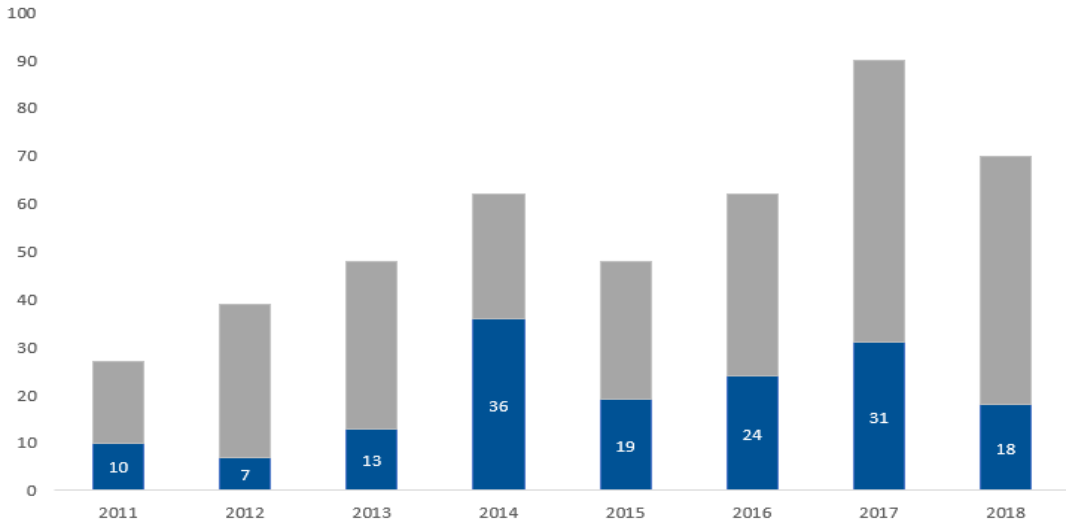


Figure 4: Privacy Breaches Affecting Public Companies

3.2. Anonymization

With so lot of data, and with strong analytics, removing the potential to identify an individual becomes impossible if there are no regulations established for the use of anonymized data files. For instance, the individuals could be re-identified in case anonymized data sets are mixed with another totally separate database, except first determining if any other data items should be removed prior to combining to shield anonymity. the rules and policies for how to combine anonymized data and used it together are very important and necessary keys to should know.[17]

3.3. Data Masking

the data masking is not used suitable, it will be easy to reveal the actual individual by big data analysis whose data has been masked. Organizations need to set up powerful policies, tactics, and approaches for the use of information protection to make sure privacy is preserved. most of the organizations don't have an idea about the risks of big data analytics because big data analytics are new so they use data masking in methods that might breach privacy. Many sources are available, such as these from IBM, to offer steerage in data masking for big data analytics, details in bellow diagram.[18]

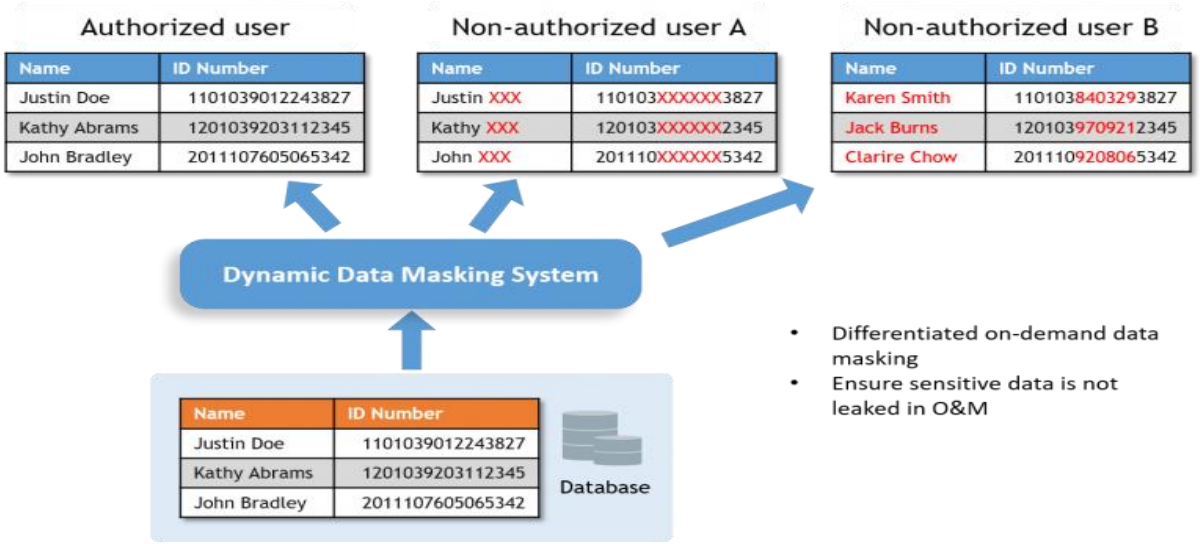


Figure 5: Data Masking System

3.4. Unethical Actions Based on Interpretations

big data analytics can be applied to try and affect behaviors. There are my moral troubles with using behavior. Just due to the fact you CAN do something doesn`t imply you should. For example, if the kill or injuring people was cheaper than fixing the faulty equipment in the vehicles. companies or organizations can use big data analytics to take commercial decisions without taking into account people lives. The capacity to show non-public information as it isn't always illegal, but can harm the lives of people, must be considered.[19]

3.5. Big Data Analytics Accuracy

the problem with big data analysis is that it is not always accurate, although it is very powerful, its results are not always accurate because not all individuals' data is correct. The results of big data not always are good, depending on the validity of the data collected and analyzed, so when using incorrect data or algorithms will affect the result clearly. This problem increases when using complex data analysis models or adding more data. This negatively affects the organization's decisions and takes inappropriate and damaging actions. [20]

The data about individuals when it is wrong and incorrect, the decision-making process is also incorrect, and the incorrect decision negatively affects individuals and they may be deprived of services and their rights, bellow figure mentioned the major contents of big data[21,23]. There are some important characteristics of big data such as speed, variety, volume....., etc. These characteristics help us understand big data, how it is measured, and how different these data are. [22] The below figure mentioned the major characteristics of big data.

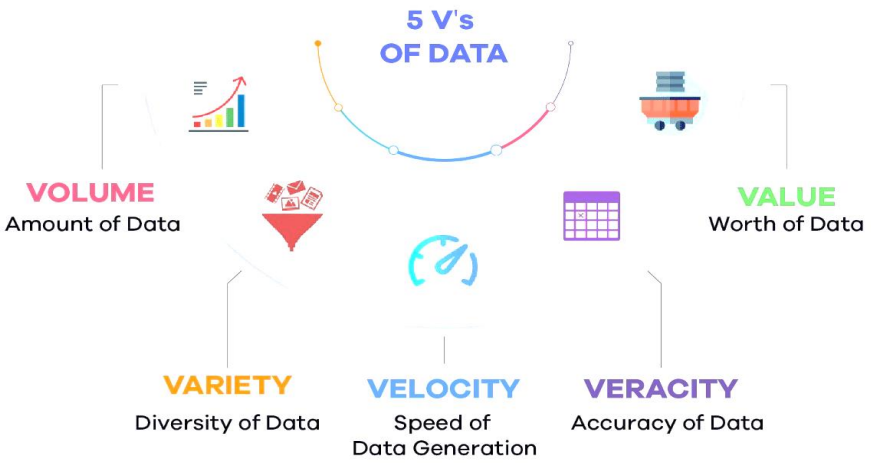


Figure 6: 5 v's of Big Data

3.6. Discrimination

big data analytics is an amazing tool to use for trying to choose job candidates, give promotions, etc. however; If the analyzes are inaccurate, it will give completely opposite and incorrect results. discrimination is the biggest problem in our life, and this problem increased with big data analytics which makes it more prevalent. For example, it is difficult for any financial institution or bank to know the race or gender of the applicant through a credit application (because it is illegal for the credit decision to be based on gender), but with the technology of big data analytics, he can get more information about the credit applicant, including race or gender, and after Obtaining the information rejects the individual's request.[24,28]

3.7. Legal Protections

There must be legal requirements to protect privacy while using big data analytics. Therefore, the US White House and the Federal Trade Commission have expressed their concerns about privacy threats in the context of using big data analytics.[25]

3.8. Big Data Existence

After reading many articles and studies, there is no indication that big data will be deleted from repositories. All studies indicate that the data is preserved and used. The larger the data, the better.[26]

3.9. E-Discovery

big data analytics creates many problems one of them is the e-discovery problem.[27]

3.10. Patents and Copyrights

Big data has a significant impact on patents submitted due to a large number of data, which makes it difficult to obtain patents and also difficult for patent offices to verify whether the submitted patent is unique or not.

Big data analytics brings unlimited benefits to individuals, as well as improves all sectors of organizations. However, organizations must verify privacy and security before they choose to use big data analytics.[29]

- There are ten privacy risks you should consider during the planning to use big data analytics.
- Responsibility and policies should be defined with all actions for big data analytics.

- Privacy and security controls must be incorporated into relevant processes before they are put into commercial use.

4. Previous Study

In [30] This paper discussed the problems of security and privacy in brief. It focused on the ways that solve the problem of security and privacy. And this paper mentioned that the most important way to solve the problem of security is the law. But due to the difference in laws between countries, security technology and other methods are necessary to protect data well. In [31] the methodology of systematic mapping study was followed to solve the security problem in big data. The most important results in this research were, that the reason behind the insecurity is the characteristics of the big data system, as well as the lack of attention to security when using big data from the beginning. Also, the Hadoop program is a concern for security. A comparison was made between big data and the Internet of things(IoT) in terms of its role in the development of technology. In [32] Big data and its protection, data security analysis, and proposing the most important strategies to solve the security problem are among the most important issues discussed in this paper. In [33,34] discusses the concerns of big data as well as the ecosystem for this data. This paper also deals with a comparison between security and privacy in terms of data, infrastructure, and applications. This paper focused on security and privacy issues in terms of literature.

5. Conclusion

The most important thing that this large data need is to increase privacy requirements and give more attention to privacy and security when collecting, storing, analyzing, and transmitting big data. In this research paper, we examined a set of previous studies related to the privacy and security of big data. because a lot of problems that we face when we analyze this type of data. Therefore, privacy is one of the biggest problems that we face when dealing with big data. So, it is necessary to discuss it more in the future, as well as develop technologies and find better solutions in order to improve the interaction between humans and computers in order to give more accurate results. The aim of this research paper is to help and understand how to secure big data as well as understand privacy problems in order to avoid them and take them into consideration and to develop techniques and find solutions to such problems not only for today but also for the future.

5. References

[1] Khan, M., & Ansari, M. D. (2019). Security and privacy issue of big data over the cloud computing: a comprehensive analysis. IJRTE-Scopus Indexed, 7(6s), 413-417.

[2] Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. Journal of Big Data, 3(1), 1-25.

[3] Asaad, R. R. (2021). Penetration Testing: Wireless Network Attacks Method on Kali Linux OS. Academic Journal of Nawroz University, 10(1), 7–12. https://doi.org/10.25007/ajnu.v10n1a998

[4] Mishra, A. D., & Singh, Y. B. (2016, April). Big data analytics for security and privacy challenges. In 2016 International Conference on Computing, Communication and Automation (ICCCA) (pp. 50-53). IEEE.

[5] Salinas, S., Chen, X., Ji, J., & Li, P. (2016). A tutorial on secure outsourcing of large-scale computations for big data. IEEE Access, 4, 1406-1416.

[6] Mishra, A. D., & Singh, Y. B. (2016, April). Big data analytics for security and privacy challenges. In 2016 International Conference on Computing, Communication and Automation (ICCCA) (pp. 50-53). IEEE.

[7] Wibowo, S., & Sumari, A. D. W. (2020). The Utilization of Blockchain for Enhancing Big Data Security and Veracity. In Combating Security Challenges in the Age of Big Data (pp. 157-187). Springer, Cham.

[8] Zaki, A. K. (2014). NoSQL databases: new millennium database for big data, big users, cloud computing and its security challenges. International Journal of Research in Engineering and Technology (IJRET), 3(15), 403-409.

[9] Tran, H. Y., & Hu, J. (2019). Privacy-preserving big data analytics a comprehensive survey. Journal of Parallel and Distributed Computing, 134, 207-218.

[10] Asaad, R. R. (2019). Güler and Linaro et al Model in an Investigation of the Neuronal Dynamics using noise Comparative Study. Academic Journal of Nawroz University, 8(3), 10–16. https://doi.org/10.25007/ajnu.v8n3a360

[11] Zeng, W., Yang, Y., & Luo, B. (2013, October). Access control for big data using data content. In 2013 IEEE International Conference on Big Data (pp. 45-47). IEEE.

[12] Marchal, S., Jiang, X., State, R., & Engel, T. (2014, June). A big data architecture for large scale security monitoring. In 2014 IEEE International Congress on Big Data (pp. 56-63). IEEE.

[13] Yu, S. (2016). Big privacy: Challenges and opportunities of privacy study in the age of big data. IEEE access, 4, 2751-2763.

[14] Rajab Asaad, R., & Masoud Abdulhakim, R. (2021). The Concept of Data Mining and Knowledge Extraction Techniques. Qubahan Academic Journal, 1(2), 17–20. https://doi.org/10.48161/qaj.v1n2a43

[15] Ram Mohan Rao, P., Murali Krishna, S., & Siva Kumar, A. P. (2018). Privacy preservation techniques in big data analytics: a survey. Journal of Big Data, 5(1), 1-12.

[16] Sedayao, J., Bhardwaj, R., & Gorade, N. (2014, June). Making big data, privacy, and anonymization work together in the enterprise: experiences and issues. In 2014 IEEE International Congress on Big Data (pp. 601-607). IEEE.

[17] Asaad, R. R., Ahmad, H. B., & Ali, R. I. (2020). A Review: Big Data Technologies with Hadoop Distributed Filesystem and Implementing M/R. Academic Journal of Nawroz University, 9(1), 25–33. https://doi.org/10.25007/ajnu.v9n1a530

[18] Cui, B., Zhang, B., & Wang, K. (2017, July). A data masking scheme for sensitive big data based on format-preserving encryption. In 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC) (Vol. 1, pp. 518-524). IEEE.

[19] Asaad, R. R., Abdurahman, S. M., & Hani, A. A. (2017). Partial Image Encryption using RC4 Stream Cipher Approach and Embedded in an Image. Academic Journal of

Nawroz University, 6(3), 40–45. https://doi.org/10.25007/ajnu.v6n3a76

[20] Devins, C., Felin, T., Kauffman, S., & Koppl, R. (2017). The law and big data. Cornell JL & Public Policy, 27, 357.

[21] Laptev, N., Zeng, K., & Zaniolo, C. (2013, April). Very fast estimation for result and accuracy of big data analytics: The EARL system. In 2013 IEEE 29th International Conference on Data Engineering (ICDE) (pp. 1296-1299). IEEE.

[22] Rajab Asaad, R. (2021). Review on Deep Learning and Neural Network Implementation for Emotions Recognition . Qubahan Academic Journal, 1(1), 1–4. https://doi.org/10.48161/qaj.v1n1a25

[23] Gillis, T. B., & Spiess, J. L. (2019). Big data and discrimination. The University of Chicago Law Review, 86(2), 459-488.

[24] Kalyvas, J. R., & Overly, M. R. (2014). Big Data: A business and legal guide. CRC Press.

[25] Abdulfattah, G. M., Ahmad, M. N., & Asaad, R. R. A RELIABLE BINARIZATION METHOD FOR OFFLINE SIGNATURE SYSTEM BASED ON UNIQUE SIGNER'S PROFILE.

[26] Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of big data privacy. IEEE access, 4, 1821-1834.

[27] Scholtes, J. C., & van den Herik, H. J. (2021). Big data analytics for e-discovery. In Research Handbook on Big Data Law. Edward Elgar Publishing.

[28] Price, W., & Nicholson, I. I. (2015). Big data, patents, and the future of medicine. Cardozo L. Rev., 37, 1401.

[29] https://bigdataldn.com/news/big-data-the-3-vs-explained/

[30] Matturdi, B., Zhou, X., Li, S., & Lin, F. (2014). Big Data security and privacy: A review. China Communications, 11(14), 135-145.

[31] Moreno, J., Serrano, M. A., & Fernández-Medina, E. (2016). Main issues in big data security. Future Internet, 8(3), 44.

[32] Zhang, D. (2018, October). Big data security and privacy protection. In 8th international conference on management and computer science (ICMCS 2018) (Vol. 77, pp. 275-278). Atlantis Press.

[33] Asaad, R. R., Abdulrahman, S. M., & Hani, A. A. (2017). Advanced Encryption Standard Enhancement with Output Feedback Block Mode Operation. Academic Journal of Nawroz University, 6(3), 1–10. https://doi.org/10.25007/ajnu.v6n3a70

[34] Terzi, D. S., Terzi, R., & Sagiroglu, S. (2015, December). A survey on security and privacy issues in big data. In 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST) (pp. 202-207). IEEE.