Recent Trends in Intensive Computing
M. Rajesh et al. (Eds.)

doi:10.3233/APC210190

Data Security and Privacy-Preserving Framework Using Machine Learning and Blockchain in Big-Data to Data Middle Platform in the Era of IR 4.0

CHEN CHUQIAO [a, b,1] , S. B. GOYAL [a]
[a] *City University, Petaling Jaya, Malaysia*
[b] *City College of Huizhou, China P.R.C*

Abstract. The modem data is collected by using IoT, stored in distributed cloud storage, and issued for data mining or training artificial intelligence. These new digital technologies integrate into the data middle platform have facilitated the progress of industry, promoted the fourth industrial revolution. And it also has caused challenges in security and privacy-preventing. The privacy data breach can happen in any phase of the Big-Data life cycle, and the Data Middle Platform also faces similar situations. How to make the privacy avoid leakage is exigency. The traditional privacy-preventing model is not enough, we need the help of Machine-Learning and the Blockchain. In this research, the researcher reviews the security and privacy-preventing in Big-Data, Machine Learning, Blockchain, and other related works at first. And then finding some gaps between the theory and the actual work. Based on these gaps, trying to create a suitable framework to guide the industry to protect their privacy when the organization contribute and operate their data middle platform. No only academicians, but also industry practitioners especially SMEs will get the benefit from this research.

Keywords. Data Security, Privacy-preserving, Machine Learning, Blockchain, Big-Data, Data Middle Platform, IR 4.0

1. Introduction

Modern data is nothing new to any enterprises as well as smaller and medium-sized firms due to multiple benefits like cost-cutting and increase efficiency and effectiveness in the data management system in the circumstance of Big-Data.

According to calculation, there are over 2.5 Eb of the data (2.5×10^{18} bytes) has been created in pre-day, and the number of data creating is still increasing [11]. Data is collected by using internet of things technology, transferred via the Internet, stored in the distributed cloud storage, released and using in the industry such as data mining, training artificial intelligence, or business decision making. To effectively manage and utilize these data, the Big-Data theory has been come up. In the engineering field, the Alibaba Group first create a new concept called the Data Middle Platform which is

[1] CHEN CHUQIAO, Ph.D. candidate in City University, Malaysia, working in City College of Huizhou
E-mail: james1989vip@163.com.

based on the Big-Data theory [26]. Sometimes Middle Platform also can be seen as a strategy to guide the Big-Data construction for enterprises. Because different organizations have different situations, the pathway of implement the Data Middle Platform is a few different [25]. The construction goal of the Data Middle Platform can be simply summarized as providing tools, processes, and methodology to realize the abstraction, reuse, and sharing of data capabilities, empower business departments and improve the efficiency of realizing data value. Alibaba creates this concept wants to address the problems of information island and repeated development and to highlight the concepts of data sharing and reuse. That is the main difference between the Data Middle Platform and the existing big data platforms.

New information technologies, that is represented by Artificial Intelligence, Blockchain, Cloud Computing, Internet of Things, and Big-Data, are integrated and have facilitated the progress of industry, promoted the fourth industrial revolution. But it also has caused threats and a challenge for data security and privacy protection. When private data gets in the wrong hands, it causes the interests of people or organizations are harmed. For example, the government`s privacy breach can put confidential information in the hands of the enemy power. A breach in an organization can put asset data in the hands of a competitor. Educational institutions` data breaches could put students' personally identifiable information in the hands of criminals who could launch internet fraud against students` parents and students themselves. A breach at the medical institutions can put the Protected Health Information in the hands of those who sell bogus medicine and then cause the interests of patients damaged. Yet, new data security and privacy challenges are being exposed as the data security and privacy protection framework based on new information technologies missing.

2. Related Work / Literature Survey

Data Security and Privacy Protection is an interdisciplinary complex problem. Data security refers to the use of a set of methods and technologies to protect an organization's data avoid unauthorized access, destroy, or theft from malicious users throughout its whole lifecycle [21]. But the Privacy concerns the ability of personal or organizations to seclude themselves or information about themselves and thereby express themselves selectively [23].

There is both difference and relation between the data security and Privacy-preventing. Privacy is concern about the use and governance of individual sensitive data — like setting up policies to guarantee the student's personal information is being collected correctly, shared with the right users, and utilized appropriately. Different from privacy, security concentrates more on how to protect data avoid malicious attacks, and the misuse of stolen data for profit. It means that security is essential for protecting data, but not enough to handle privacy [9].

The privacy data breach can happen in any phase of the Big-Data life cycle. But data breaches easily occur in data storage, data transmission, and data release. To solve the privacy leakage issue, predecessors have developed different solutions according to the characteristics of privacy leakage in different stages of the big data life cycle. Privacy-preserving technologies can be classified into the following categories [16].

Privacy-preserving technology based on data distortion

Data distortion-based privacy-preserving technology refers to adding some noise into the original data and then make sensitive data distorted but keeping data properties unchanged. But the statistical characteristics of these distorted sensitive data will not be changed using the privacy-preserving technology based on data distortion.

By extension, there are three kinds of data distortion-based privacy-preserving technologies. The first is randomization. It is a simple way that put some stochastic noise into the raw data and then publishes the disturbing data. The second one is the blocking and cohesion method. Blocking refers to not releasing certain specific data when the data is released, and Cohesion refers to grouping and storing the original data, and then combining them together during statistics, to fulfill the effect of privacy protection. The third one is differential privacy.

Differential privacy, as be short as DP, is a new kind of privacy-preventing model [6-7]. This method is able to unriddle two major shortcomings of the general privacy-preventing model: First of all, it gives a fairly precise attack model. In the model, the researcher doesn`t need to care about the background knowledge of malicious users, even if the malicious users have a good command of all record information except for a certain record, the privacy of the record cannot be uncovered. Secondly, there is a rigorous definition and a quantitative evaluation method has been given for the grade of privacy protection. Due to differential privacy`s advantages, the traditional privacy protection models are quickly replaced. Now, differential privacy is widely discussed in the privacy research area and has attracted the attention of many fields. Not only include computer science, but also contain database, data mining, and machine learning.

3. Privacy-preserving technology based on data encryption

Data encryption-based privacy-preserving technology means the method of using encryption technology to hide privacy in the data mining process. Two representative data encryption-based privacy-preserving technologies are the security multi-party computation and the homomorphic encryption.

The security multi-party computation belongs to a subdiscipline of cryptography. It is also called secure computation, multi-party computation (MPC), or privacy-preserving computation. The aim of security multi-party computation is to create methods to help parties involved in the computation to complete the computation while keeping the data that input by each participant private. It is different from traditional cryptographic tasks. The traditional cryptographic task is using passwords to make sure the data is secure and integrity in communication and storage. It is only useful for the user who outside the system. For the user who involves in, the traditional cryptographic task can do anything to help. But the security multi-party computation gives an idea to solve this situation. The security multi-party computation will protect each party's privacy only be known by itself [22]. In order to ensure data integrity, the blockchain will be used in the process of data transmission and storage.

The homomorphic encryption is an encryption form created by Gentry in 2009 [1]. He puts forward a feasible method which is called "fully homomorphic encryption" in mathematic. That means the encrypted data can be operated without decryption, and the

result which operates by the encrypted data is the same as the result which operates by the encrypted data after decryption [1].

4. Privacy-preserving technology based on restricted release

The privacy-preserving technology based on restricted release is to realize privacy protection by controlling the release of original data. For example, people can release the filtered data or sensitive data with low precision and then make privacy protection.

Normally the research on restricted release-based privacy-preserving technology focuses on two aspects. One is data generalization, the other one is data anonymization. The aim of them is the same is to ensure that the risk of leakage of sensitive data and privacy is within a permissible range.

In general, data generalization has using a process to summarize data by replacing relatively low-level values with higher-level concepts, or by reducing the number of dimensions so that the data can use fewer dimensions to cover. For example, in educational institutes, when the engineer designs a sheet of the database, he can use mark grade from the letter A to the letter E instead of numeric values for an attribute student`s mark. Or, removing birth date and telephone number when summarizing the behavior of a group of students. Given the large amount of data stored in databases, it is useful to be able to protect the specific privacy value at generalized levels of abstraction [8].

Currently, there are three kinds of technologies which is K-anonymity, L-diversity, and T-closeness in data anonymization.

The earliest widely accepted privacy protection model in Privacy-preserving technology based on restricted release is K-anonymity, which has been defined in 2002. In order to dispose of the de-anonymization attacks, each data record that is released by k-anonymity must be difficult to distinguish from no less than k-1 other records (called an equivalence class). Even though the hacker gets the data that is deal with by using the k-anonymous model, he will get the records of at the minimum k different people, and thus cannot make an accurate judgment. The parameter k signifies the strength of privacy-preventing. The larger number the K gets, the stronger strength of privacy-preventing you get. But it also means the lower availability of the data and the more information will be lost [18-19].

In 2006, Machanavajjhala et al. [14] who is working at the Cornell University noticed that the k-anonymity model has a weak point. Even though there is nothing be restricted on sensitive attributes. Hackers also can confirm the relationship between sensitive data and individuals by using background knowledge attacks, re-identification attacks, and consistency attacks. For example, the attacker obtains the k-anonymized data. if the equivalence class of the k-anonymized data is all AIDS patients, then the attacker can easily make the judgment which one in the k-anonymized data definitely has AIDS. To prohibit consistency attacks, the new privacy protection model l-diversity improves k-anonymity to ensure that the sensitive attributes in any equivalence class have at least l different values. Based on l-diversity, t-Closeness requires the distribution of sensitive attributes in all equivalence classes to be as close as possible to the global distribution of the attribute [13]. (a, k)-anonymity principle, on the basis of k-anonymity, further ensure that the percentage of records related to any sensitive attribute value in each equivalence class is not higher than a [17, 20, 24].

However, the privacy-preventing model above is still flawed and needs to be upgraded continuously [2, 3, 5, 10, 15]. Fundamentally, no single privacy protection model can effectively protect privacy. Only by using various privacy protection technologies comprehensively to form a privacy protection technical framework can privacy data be protected effectively.

5. Problem Statement

This research aims to build a privacy protection framework based for Data Middle Platform as follows:

- To list out the effect of security leak & methods in the modern data
- To list out the effect of privacy leak & methods in the modern data
- To compare big-data and Data middle platform
- To identify the importance of Big-Data and Data middle platform in the era of IR 4.0
- To design a framework to handle the security and privacy in the modern data
- To utilize the machine learning techniques in the proposed framework
- To apply the blockchain techniques in the proposed framework
- To test the proposed framework in real-life data

6. Solution Approach

The proposed research needs the machine learning techniques concept to handle the security and privacy in the large volume of data in addition to the blockchain due to its core characteristics like a consensus, smart contract, public, and private key notion. So, the researcher will apply machine learning and blockchain techniques to ensure no data accessibility to an unauthorized person and no one can do the unwanted operation without the accessibility rights.

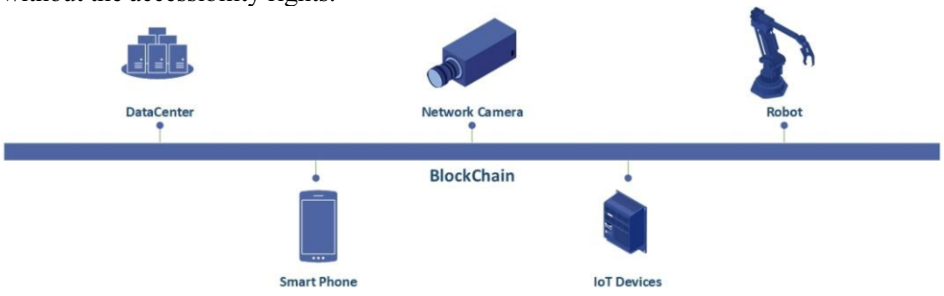

Figure 1. Different Privacy Sources Formed Blockchain

Figure 1 shows what is in the Blockchain. In IR 4.0, privacy comes from many pathways. Some privacy comes from the existing database located in the data center. But others are coming from devices such as IoT devices, mobiles, network cameras, and industrial robots. That privacy will be packed and then put in the Blockchain.

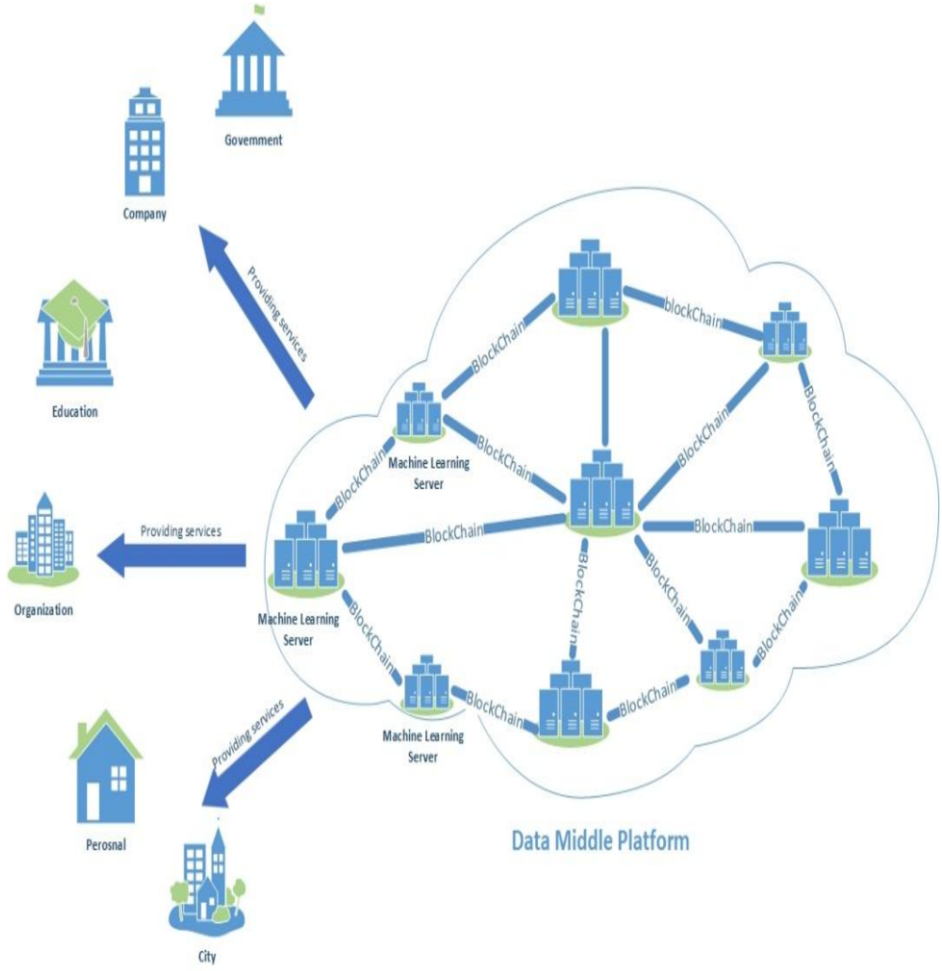

Figure 2. Framework of Privacy Blockchain in Data Middle Platform

The Blockchain will flow in the Data Middle Platform as shown in the figure 2. It will be analyzed by using Machine Learning theory. The user will get the result but doesn`t know the specific data. In this way, the privacy will be protected and then avoid different kinds of malicious attacks.

There are five steps which lists in below that will be confirmed in the research.

- A preliminary literature review study will be performed. This encompasses the background theory and pertinent topics including data security technology, privacy protection, machine learning, blockchain, Big-Data, Data Middle Platform, and the Industry Revolution 4.0.
- The literature review will be followed by the state-of-the-art in Data security and privacy-preventing for Big-Data in the Era of the Industry Revolution 4.0. This includes classifying the technologies of data security and privacy protection for identifying the relevant literature. We will use the results to guide to identify gaps in the current research on privacy-preservation in order to suggest areas for further investigation.

- Data security and privacy-preventing framework using Machine Learning and Blockchain in Big-Data to Data Middle Platform which is consisted of conceptual and logic will be provided.
- we will try to use a use case to verify the framework and then try to identify any potential defects. In this section there are several open-source architectures will be used such as Apache Hadoop, Docker, MongoDB, MySQL, Python.
- Papers will be presented at publications and conferences of proceedings for reviewed and discussion.

7. Expected Impact

This research work will benefit academicians, industry practitioners, and researchers to open the new dimension in the middle data platform and SME to Enterprises will not hesitate to adopt the middle data platform and enhance the socio-economy aspect of the enterprises.

8. Conclusion

With the rapid growth of the number of data, data security and privacy-preserving technology are not adequate in the era of the fourth industrial revolution. This is because the development of the data middle platform makes the data security and privacy-preserving technology delay. The old technology may not ensure the user`s privacy because of operational and efficiency problems. So, updating the data security and privacy-preserving framework using new technology such as machine learning and blockchain is of great urgency. Not only to help the industry protect the privacy in Data Middle Platform but also will provide clear guidance to those people who will be involved in the data governance of new initiatives related to data security and privacy-preserving.

References

[1] Gentry, Craig. A fully homomorphic encryption scheme. Vol. 20, no. 9. Stanford: Stanford university, 2009.
[2] Campan A, Truta TM, Cooper N. 2010. User-controlled generalization boundaries forp-sensitivek-anonymity. In: Proceedings of the 2010 ACM Symposium on Applied Computing - SAC '10.
[3] CAO M, Zhang L, Bi X, Zhao K. 2017. Personalized (α, l) -diversity k-anonymity Model for Privacy Preservation. Computer Science. 45(11):180–186. doi:https://doi.org/10.11896/j.issn.1002-137X.2018.11.028.
[4] CEH v11 Note Summary Module 1 (A). 2020 Nov 27. blogcsdnnet. https://blog.csdn.net/taof211/article/details/110219994.
[5] Chen XB, Han B, Huang S. 2018. Research on fuzzy t-intimacy protection method. Computer Applications and Software. 35(9).
[6] Dwork C. 2006. Differential Privacy. In: Proceedings of the 33rd International Colloquium on Automata, Languages and Programming. Venice, Italy,. p. 1–12.
[7] Dwork C. 2008. Differential privacy: A survey of results. In: International conference on theory and applications of models of computation. Berlin, Heidelberg: Springer. p. 1–19.
[8] Han, J. K. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann.
[9] Jain P, Gyanchandani M, Khare N. 2016. Big data privacy: a technological perspective and review. Journal of Big Data. 3(1). doi:10.1186/s40537-016-0059-y.

[10] Jurczyk P, Xiong L. 2009. Distributed anonymization: Achieving privacy for both data subjects and data providers. In: IFIP Annual Conference on Data and Applications Security and Privacy. Springer. p. 191–207.
[11] Karki D. 2020 Nov 9. Can you guess how much data is generated every day? Takeo. [accessed 2021 Feb 16]. https://www.takeo.ai/can-you-guess-how-much-data-is-generated-every-day/.
[12] LI B. 2018. Research on privacy protection data release algorithm based on different anonymous requirements. [Jilin University].
[13] Li N, Li T, Venkatasubramanian S. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In: 2007 IEEE 23rd International Conference on Data Engineering.
[14] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. 2007. L-diversity. ACM Transactions on Knowledge Discovery from Data. 1(1):3-es. doi:10.1145/1217299.1217302.
[15] Martinelli F, SheikhAlishahi M. 2019. Distributed Data Anonymization. In: 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech).
[16] Privacy protection technology in big data environment. 2015 Jun 1. Cyberspace Administration of China. [accessed 2021 Jan 22]. http://www.cac.gov.cn/2015-06/01/c_1115473995.htm.
[17] SUN J. 2019. Research and Implementation of k-anonymity based on sensitivity classification [Beijing University of Posts and Telecommunications].
[18] SWEENEY L. 2002a. ACHIEVING k-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION. In: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. Vol. 10. p. 571–588.
[19] SWEENEY L. 2002b. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. In: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. Vol. 10. p. 557–570. [accessed 2019 May 10]. https://epic.org/privacy/reidentification/Sweeney_Article.pdf.
[20] Wang J, Li H, Guo F, Zhang W, Cui Y. 2019. D2D Big Data Privacy-Preserving Framework Based on (a, k)-Anonymity Model. Mathematical Problems in Engineering. 2019:1–11. doi:10.1155/2019/2076542.
[21] What is data security? Definition, solutions and how to secure data. 2021. www.ibm.com. [accessed 2021 Feb 16]. https://www.ibm.com/topics/data-security.
[22] Wikipedia Contributors. 2019 Dec 6. Secure multi-party computation. Wikipedia. https://en.wikipedia.org/wiki/Secure_multi-party_computation.
[23] Wikipedia contributors. (2021b, March 23). Privacy. Wikipedia. https://en.wikipedia.org/wiki/Privacy
[24] WU S. 2015. Research on the t-closeness privacy protection model supported by rough sets and clustering. [Shanxi Normal University].
[25] Zhang C, Hou L. 2020. Data middle platform construction: The strategy and practice of National Bureau of Statistics of China. In: Statistical Journal of the IAOS. Vol. 36. p. 979–986.
[26] Zhong H. 2017. The transformation of enterprise IT architecture: Alibaba's strategic thinking and structure practice of middle platform. China Machine Press.