International Journal of Internet, Broadcasting and Communication Vol.13 No.2 60-66 (2021)
http://dx.doi.org/10.7236/IJIBC.2021.13.2.60

IJIBC 21-2-9

A Study on Efficient Data De-Identification Method for Blockchain DID

Youn-A Min

Professor, Applied Software Engineering, Hanyang Cyber University, Korea
yah0612@hycu.ac.kr

Abstract

Blockchain is a technology that enables trust-based consensus and verification based on a decentralized network. Distributed ID (DID) is based on a decentralized structure, and users have the right to manage their own ID. Recently, interest in self-sovereign identity authentication is increasing. In this paper, as a method for transparent and safe sovereignty management of data, among data pseudonymization techniques for blockchain use, various methods for data encryption processing are examined. The public key technique (homomorphic encryption) has high flexibility and security because different algorithms are applied to the entire sentence for encryption and decryption. As a result, the computational efficiency decreases. The hash function method (MD5) can maintain flexibility and is higher than the security-related two-way encryption method, but there is a threat of collision. Zero-knowledge proof is based on public key encryption based on a mutual proof method, and complex formulas are applied to processes such as personal identification, key distribution, and digital signature. It requires consensus and verification process, so the operation efficiency is lowered to the level of $O\ (\log_e N) \sim O(N^2)$. In this paper, data encryption processing for blockchain DID, based on zero-knowledge proof, was proposed and a one-way encryption method considering data use range and frequency of use was proposed. Based on the content presented in the thesis, it is possible to process corrected zero-knowledge proof and to process data efficiently.

Keywords: *Blockchain, DID (Decentralized Identity), Data De-Identification, pseudonymization*

1. Introduction

With the rise of the 4th Industrial Revolution, there have been increased interests regarding infrastructures that lead in reliability, with the blockchain being the core infrastructure in the matter [1].

Blockchain is a distributed ledger management technology based on a decentralized P2P reliability network to share distributed ledgers to verify and consent to a transaction record [2]. This technology was first introduced by Satoshi Nakamoto in 2008 and was used as the core technology in Bitcoin in 2009 [2]. The applications have expanded in range and researches regarding the application possibilities have become much more diverse. Blockchain technology can be classified by generation. The period between the advent of the Bitcoin and the introduction of the Ethereum technology in 2015 is classified as the first generation, while the applications of smart contracts based on the Ethereum technology is identified as the second

Manuscript Received: February. 11, 2021 / Revised: February. 15, 2021 / Accepted: February. 22, 2021
Corresponding Author: yah0612@hycu.ac.kr
Tel: +82-2-2290-0872, Fax: +82-2-2290-0872
Professor, Applied Software Engineering, Hanyang Cyber University, Korea

generation [1, 2].

DID (Decentralized Identity) refers to the technology based on a decentralized structure for users to manage their own ID [3]. Blockchain technology-based researches have recently been carried out continuously for flexibility of ID management and security of personal information [1,2].

This paper compares various cryptographic processing methods and presents a single method to strengthen the efficiency and security of blockchain DID.

2. Related work

2.1 DID (Decentralized Identity)

DID is a method of managing ID for identity authentication.

Authentication can be classified into two types: face-to-face and non-face-to-face as methods to authenticate user identification [4]. Figure. 1 is a conceptual diagram of the processing process of DID and shows the relationship between issuer, holder, and verifier claim request.

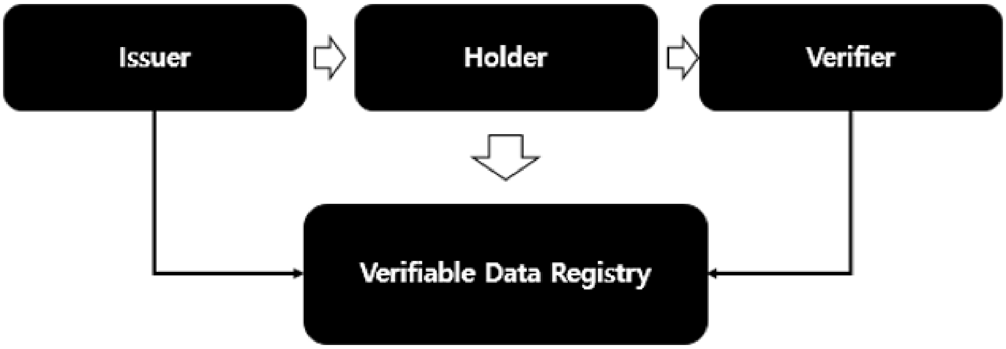

Figure1. Verifiable Credential Processing Process [4]

Face-to-face authentication refers to identifying oneself through the use of certification printed onto a plastic card or piece of paper, while non-face-to-face is a method of accessing a preferred service company's server through an online server [4]. The non-face-to-face authentication is an online-authentication method through a registered service company [5, 6]. In the case of online identification, the user goes through the registration process and receives an ID. This ID is the means to identify oneself and is comprised of a total of four elements [5, 7]. The ID has values for the identifier, attribute, authentication method, and issuer as the elements to identify the user [5, 7]. The users provide the service company private information in order to receive the service through their own IDs, then receive the necessary information to be able to login to the service company they had registered in. The non-face-to-face online identification process does come with convenience and speed. However, there is a need to consider the problems associated with the flexibility of data identification application and privacy security and management.

2.2. Data Pseudonymization and Encryption

Data Pseudonymization refers to the process that controls the connection between the series of identifying data and the subject data of an individual. Currently, there are various data non-identification processes existing to protect the identity of the data subject [7, 8].

The major methods of data pseudonymization are Heuristic Pseudonymization and Encryption, wherein this encryption can be divided into bidirectional and one-way encryptions. Among the heuristic pseudonymization, k-anonymity uses a technology that prevents leaking of private information by maintaining at least k number of identical property values as a single data set. For example, if a 100-data set processes 5-anonymity, it gives at least 5 data to have the same properties, therefore expressing 20 representative data. This notation expands the range of anonymization as k-increases and increases the security probability. On the other hand, the data set may have decreased data values as k- increases [4, 5-10]. L-diversity is the process of diversifying the data values in to process of applying anonymity in k-anonymity, while t-closeness is one which sets a certain data section that has a different pattern than the original data set to make it impossible to identify the sensitive information.

Data Encryption is defined as the replacement of personal information through an encryption algorithm that has a rigid structure during information processing. Encryption can largely be divided into bidirectional and one-way encryptions. Bidirectional encryption has two methods, a symmetric key method that uses the same algorithm during encryption and decryption, and a public key method that uses a different algorithms for each. Using the same algorithm during encryption and decryption may make the process much quicker but has the danger of decryption if the algorithm during encryption is leaked, while using a different algorithm costs much more in terms of calculation [4, 5-10].

A one-way encryption uses a salt value or key value of at least 32 bytes to the hash function to prepare for a random brute-force attack by the hacker. The one-way encryption normally uses a security-verified hash algorithm, which makes it impossible to restore the original from the code [7, 8-10]. Among the bidirectional encryption, the symmetric key processes include SEED, DES (Data Encryption Standard), and AES(Advanced Encryption Standard), among which the DES would be explained. DES uses a method to put a plain text through a 64-bit block encryption to make a 56-bit cryptogram and uses the following equation to process the calculation of round I for left and right halves of the text [7, 8-10]. As for the public key methods, they include RSA (Rivest, Shamir and Adleman), ElGamal. ECC (Elliptic Curve Cryptosystem), and Digital signatures, wherein the HE (Homomorphic Encryption) would be explained.

HE allows for calculations for encrypted data as in their encrypted state. HE gives out a new cryptogram according to a private key shown by the equation above and gradually reduces noise of the cryptogram through continuous multiplication. By using HE, the calculation of $\mathrm{Enc}(\chi^1),\mathrm{Enc}(\chi^2)...\mathrm{Enc}(\chi^n) \rightarrow \mathrm{Enc}(f(\chi^1.....\chi^n))$ becomes possible[9,10-13].

Hash functions represent the one-way encryptions, which include MD5, SHA256, etc., among which MD5 is an algorithm that outputs a 128-bit hash value from a plain text.

ZKP (Zero Knowledge Proof) is a representative process used for data encryption in blockchain DID [9, 10-13]. The definition of ZKP is as follows. Zero Knowledge Proofs is an encryption technology that proves that it knows the private key (w) regarding function f(x) of an input/output value x even when the key has not been shared [14]. ZKP consists of a prover and verifier that proves the accuracy of the input and output, where both receive a sharing key called CRS (Common Reference String) [14, 15-16]. CRS can be divided into RRS (Random Re-fence String), which creates a key with a random criteria, and SRS (Structured Reference String), which does so with a structured criteria, where both go through interactive proof by the prover and verifier[11,12-14]. To utilize the ZKP, the blockchain would establish a interaction formula and function between the original data and the committed information to create a proof, which would sequentially be delivered to the verifier who would verify the results [12, 14-17]. In order to efficiently use the ZKP, the data would be saved both off-chain and on-chain, where the verification of the existing data saved off-chain would be done by ZKP of the information on-chain. Through the verification process of the ZKP, the non-identification processing of data is available and verification regarding the actual data can be done quickly. For the ZKP to take place, an ALU that transforms the equation with addition and multiplication for a given function is created and the proof is sequentially formed [12, 14-17].

The methods of proof in ZKP can be shown as the circuit specialization method and the polynomial creation method. The circuit specialization proof method formulates the proof through a circuit expression method of the QAP (Quadratic Arithmetic Program) [15, 16-17]. Figure. 2 shows the process of proving circuit specificity through the circuit expression method.

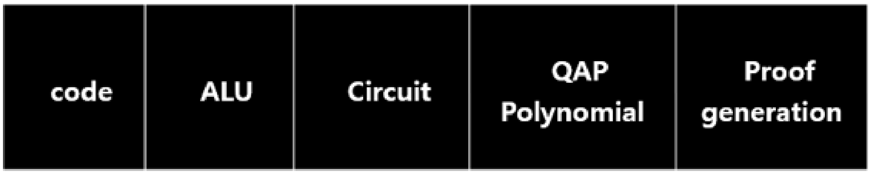

Figure 2. QAP Circuit expression method [15,16-17]

The polynomial creation method is used to minimize the interaction formula between units and is a method that expresses the relationship among the values in the circuit as a permutation-form polynomial [15, 16-17]. This can show the input and output connection among circuit gates and uses Arithmetic expression [11, 13-17]

3. Comparison and Suggestions of Data Encryption Methods

3.1. Comparison of Data Encryption Methods

This paper focuses on the encryption method used in blockchain DID and compares the utilization flexibility, security and self-sovereignty efficiency, and encryption calculation efficiency for the symmetric keys, public keys, one-way encryption, and ZKP.

Table 1. Polynomial generation processing method

Items	Content
Flexibility of Utilization	Need of flexibility depending on the property nodule range and objective of using the ID ID can be managed not only in confined areas but in various regions and places Need of self-management of information according to the ID property nodule range and objective
Security	Need of data encryption of the various information saved in a user ID
Self-sovereignty management	Need for a user to verify the value of the information given through the ID and protect/manage the information
Calculation efficiency	Need to take into account the calculation speed once the formula is applied for the data encryption

In Table.1, the comparison factors for the data encryption processing method, the reasons for considering the factors, and points to be considered are listed.

This paper compares four pre-mentioned methods--private keys, public keys, hash method of the bi-directional encryption, and ZKP of the blockchain data encryption--based on the formula above.

The utilization flexibility and self-sovereignty management items were discussed by the relationship of polynomial utilization and the information security was measured by the complexity of calculation. The calculation efficiency was measured by the speed of calculation.

- Symmetric key: Analysis through DES (Data Encryption Standard), a representative method of symmetric keys, it is shown that the usage of the same algorithm in encryption and decryption would give a high flexibility but low security. Because it uses the same algorithm for both encryption and decryption, the calculation efficiency is regarded to be high.
- Public key: Analysis through Homomorphic Encryption showed that using different algorithms for encryption and decryption gave high flexibility and security. However, the usage of different algorithms in encryption and decryption lowered the calculation efficiency.
- One-way: Analysis based on MD5, a representative hash function, the 128-bit hash value gives sufficient flexibility and a higher security strength compared to bidirectional encryption. There recently was even a collision that caused a threat in security. In the case of calculation efficiency, the hash function and salt is usually *O(N) or O(1),* which gives a high calculation efficiency.
- Homomorphic Encryption: Based on a public key encryption system, the algorithm that takes into account the Scalability for various users gives a high flexibility and security. In the calculation efficiency sense, however, the various arithmetic and logical calculations possible for the system requires a multitude of proofs, which lowers the efficiency by $O(\log_e N) \sim O(N^2)$.
- ZKP: From a public key encryption based on interactive proof, the flexibility and security are high due to complex formula applied in identification, key distribution, and digital signature, but the repetitive agreements and verification required between the prover and verifier lowers the calculation efficiency by $O(\log_e N) \sim O(N^2$.
  The summary of the comparison is as follows.

Table.2 summarizes the description of the above performance comparison in a table. Table.2 is a table that compares various cryptographic processing techniques based on the factors presented in Table.1, and is organized into High (good), medium (normal), and Low (bad) classifications.

Table 2. Cryptographic processing method bar bridge and performance comparison analysis

Method	Flexibility of Utilization	Self-sovereignty	Security	Calculation efficiency (Considering Formula complexity)
Symmetric	Medium	Medium	Low	High
Public Key	High	High	High	Medium
Hash	Medium	Medium	Medium	High
ZKP	High	High	High	Low

As blockchain DID is a technology that ensures a self-sovereign identification, there is a need for flexibility of information presentation, security, and efficiency. The comparison above shows that the public key and ZKP did lead in flexibility, self-sovereignty management and security, but the symmetric key and one-way encryption, specifically the hash method, were shown to be better in efficiency terms. Therefore, there is a need for an encryption based in public key that partially applies the one-way encryption depending on the utilization range and data usage frequency in order to increase the calculation efficiency.

4. Discussion

Among blockchain technologies, DID is one that ensures anonymity of data and strengthens security.

This paper examined the non-identification methods of blockchain-based DID and the homomorphic encryption and ZKP for data encryption. Blockchain DID aims for a self-sovereign identification and ensures transparent and secure data. Therefore, there is a need to examine various existing encryption methods and combine them efficiently. This paper suggests a data encryption system based on ZKP that also uses a one-way encryption system that considers the data usage range and frequency.

There seems to be a need for research on the suggested method regarding various error situations, which would be discussed in future papers.

References

[1] Melanie Swan, Blockchain, O'Reilly Media, Inc, pp. 10-25, 2015.

[2] S. Nakamoto, Bitcoin: APeer-to-Peer Electronic Cash System.https://bitcoin.org/bitcoin.pdf.

[3] D. Reed et al., Decentralized Identifiers (DIDs) v1.0, Core Data Model and Syntaxes,https://www.w3.org/TR/did-core/.

[4] W3C Credentials Community Group:https://www.w3.org/community/credentials/

[5] Y.A. Min, "A Study on Application of Blockchain Distributed ID Technology for Management of Welfare Dead Zone," Journal of the Institute of Internet, Broadcasting and Communication (JIIBC), Vol.20, No.6, pp. 145-150, Dec 2020.
DOI : https://doi.org/10.7236/JIIBC.2020.20.6.145

[6] J.K. Hong et al., "Blockchain Watchdog: Real-time Blockchain Surveillance System Connecting Smart Contract Code and Distributed Storage," Journal of the Korean Institute of Internet, Broadcasting and Communication, Vol.20, No.4, PP.115-121, Oct 2020.
DOI : https://doi.org/10.7236/JIIBC.2020.20.4.115

[7] M.H.Joo et al.,"De-identification policy and risk distribution framework for securing personal information," The International Journal of Government & Democracy in the Information Age, Vol. 23, No. 2, p195-219, 2018. DOI : 10.3233/IP-170057

[8] Liu, Kai-Cheng et al., "Optimized Data de-Identification Using Multidimensional k-Anonymity," in Proc. 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pp.1610-1614 Aug.1-3, 2018.
DOI : 10.1109/Trust Com/BigDataSE.2018.00235

[9] H.J.Lee et al., "De-identification and Privacy Issues on Bigdata Transformation," in Proc. IEEE International Conference on Big Data and Smart Computing (BigComp), pp.514-519, Feb.19-22, 2020.
DOI :10.1109/BigComp48618.2020.00-14

[10] Zhang. Zeyu, Lu. Zhiyang, Tian. Youliang, "Data Privacy Quantification and De-identification Model Based on Information Theory," in Proc. International Conference on Networking and Network Applications (NaNA) International Conference on, pp.213-222, Oct.10-13, 2019.
DOI:10.1109/NaNA.2019.00046

[11] D. Abouakil, J. Heurix and T. Neubauer, "Data models for the pseudonymization of dicom data," in Proc. 44th Hawaii International Conference on System Sciencesm pp. 1-11, Feb 22. 2011.
DOI : 10.1109/HICSS.2011.136

[12] Kaplun, Dmitriy I et al.,"Research and implementation of the algorithm for data de-identification for Internet of Things," in Proc. IEEE II International Conference on Control in Technical Systems (CTS), pp.363-366, Oct. 25-27, 2017.
DOI : 10.1109/CTSYS.2017.8109568

[13] J.H. Lee et al., "Personal Information Management System with Blockchain Using zk-SNARK", The Journal of the Society for Information Security, Vol. 29, No. 2, pp. 299-308, April 2019.
DOI : 10.13089/JKIISC.2019.29.2.299

[14] M.S. Kim and B.R.Kang, "Generalization of Zero-Knowledge Proof of Polynomial Equality", The Journal of the Korean Institute of Communication Sciences, Vol. 40, No. 5, pp.833-840, May 2015.
UCI : G704-A00600.2015.40.5.012

[15] S.O. Kim, "Balance points for safe processing and rational use of pseudonym information-Combined with constitutional evaluation of the 3rd Data Act, " Korea Public Law Research. Vol. 49, No. 2, pp.371-407, Dec 2020.
DOI : 10.38176/PublicLaw.2020.12.49.2.371

[16] J.Y. Chun, G.T. No, "Suggestions for Applications of Anonymous Data under the Revised Data Privacy Acts", Journal of the Korea Information Security Society, Vol. 30, No. 3, pp.503-512, Jun 2020.
DOI : 10.13089/JKIISC.2020.30.3.503

[17] Young. Jeffrey A et al., "A Methodological Framework for Validating ZKP Authentication Process," in Proc. IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI , pp.37-43 Dec.14-16, 2020.
DOI: 10.1109/HONET50430.2020.9322828