ISSN 2278-3091
Volume 10, No.5, September - October 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse011052021.pdf
https://doi.org/10.30534/ijatcse/2021/011052021

Development of Privacy-Preservation of Big Data with support of Hyperledger Fabric and IPFS

Alpesh Vaghela[1], Anilkumar Suthar [2]
[1]Research Scholar, Department of Computer Science, Gujarat Technological University, Gujarat, India and Programmer, GMB Polytechnic, Rajula Gujarat, India,
vaghela.a79@gmail.com
[2]Director, New L J Institute of Engineering and Technology, Gujarat Technological University, Gujarat, India,
sutharac@gmail.com

Received Date : August 05, 2021 Accepted Date : September 14, 2021 Published Date : October 06, 2021

ABSTRACT

Academics and industry researchers alike find privacy-preservation of large data to be a very intriguing field of study. Data collection, storage, and processing are the three steps of big data's life cycle. At different stages of the big data life cycle, different privacy and security solutions are used. Many health-care stakeholders are working together to develop a new pattern for safeguarding people from an unknown disease while also promoting economic prosperity. The methods of big data processing and big data analytics will be employed to discover new economic growth patterns. Because the current method of data anonymization leads to data breaches, researchers needed to develop a new way of large data mining or knowledge discovery in databases (KDD), in which numerous parties share their data to identify new patterns. This study introduces a novel way for data mining privacy protection based on Blockchain and the InterPlanetary File System (IPFS) (PPDM). The authors propose leveraging Blockchain and IPFS to create the ChainPPDM approach for preserving big data privacy. The data saved on the blockchain is immutable, transparent, and safe, and it allows for decentralized storage. IPFS is a distributed file system that stores data in a decentralized manner.

Key words: Big Data, Blockchain, Hyperledger Fabric, IPFS, Privacy Preservation, Security

1. INTRODUCTION

Big Data is a favored research subject and hot topic for industrial researchers and academicians around the world, and it is fast developing in contrast to all other topics in the research domain. Big data is one method for efficiently storing, processing, analyzing, and searching a massive Volume, velocity, and variety are the 3V problems of big data processing. Big data analytics research focuses not just on solving 3V problems, but also on balancing privacy and security concerns. If researchers are unable to develop a solution for privacy and security, big data concepts and implementation will be rendered useless. Big data, for example, may save the health-care industry up to US$450 billion if it is used in a methodical manner. Patients' information has been separated into sensitive and non-sensitive sections. When sensitive data is used in health care research, privacy issues arise. However, when big data is employed in conjunction with privacy and security, a structured programme will be used to forecast for the safety of county residents [2]. Big data is a type of data that includes both structured and unstructured information [3]. Data security and data privacy are two distinct ideas. Data security refers to the confidentiality, availability, and integrity of data, as well as ensuring that data is not accessed by unauthorized individuals.

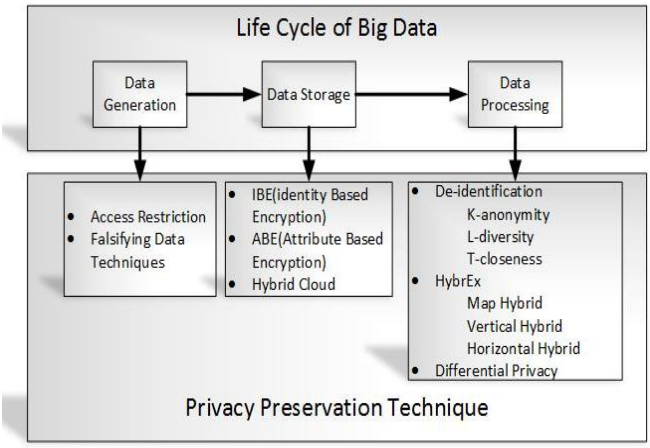

Figure 1: show the different methods for a different phase of the big data life cycle.

Data security entails gathering and protecting data from unauthorized users, as well as destroying data that is no longer needed. The right use of data is characterized as data privacy. Customers' personal data is used by businesses on the condition that no personal information is disclosed and no data is used that is not useful to customers [4]. The concerns of data abuse, such as privacy data transactions, are not properly regulated due to a lack of legal limits and immature auditing tools, and the difficulty of maintaining privacy is worsened. The keyword search feature needs to be performed over encrypted centralized data storage without disclosing any information about the search query or the retrieved document in order to ensure data privacy. A privacy-preserving keyword search is what it's called. Figure 2 shows the list of users who have their privacy protected by the centralization data mining server. The requirements for preserving users' privacy from the centralization big data mining server are listed below.

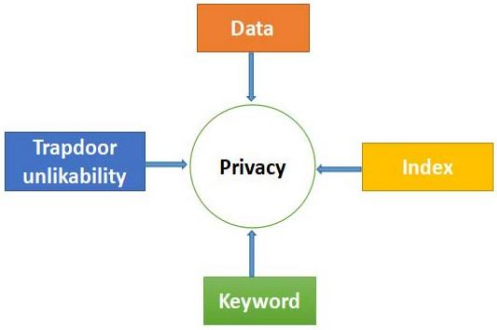

Figure.2: shows the list of privacy-protecting of users from the centralization data mining server.

- **Data privacy**- In the centralization big data mining server, outsourced patient data is secured without revealing any private or sensitive information to unknown third parties.
- **Index Privacy**- The security of the hospital's big data infrastructure must be ensured so that other hospitals cannot steal private data from centralization-based indexes.
- **Keyword privacy**- The centralization server of hospital management does not steal any particulars data collection, index, and encrypted keywords.
- **Trapdoor unlikability**- Developed a random trapdoors function to safeguard data and fire privacy, as well as ensuring that each trapdoor generated query is unique and that no one can connect the trapdoors [1].

We created novel healthcare data mining algorithms based on blockchain, which is a decentralised storage system with immutable transaction properties, and IPFS, which is a decentralised storage system for huge data. The blockchain method is depicted in figure 3 for all phases of the big data life cycle. Data Provider, Data Collector, Data Minor, and Decision Maker are the four key actors involved in the privacy-preservation of big data mining, and they all play a role in the privacy-preservation of big data in the healthcare industry. Figure 4 depicts the roles of several actors in data mining in the hospital setting.

- **Data Provider**: In the privacy-preservation of big data in the hospital area, all users of the hospital, including doctors, patients, nurses, pharmacies, and lab reports, who supply some input in the form of data are known as data providers.
- **Data Collector**: A Data Collector is a hospital that provides a server and storage area for storing data provider data. Because all data providers must rely on the data collector for their data privacy and security, the data collector is the most important actor for privacy protection.
- **Data minor**: A data minor is a technical expert who is familiar with a variety of algorithms for discovering meaningful information.
- **Decision maker**: For economic growth, the decision-maker is obliged for publishing the optimal report from data analysis and data mining.

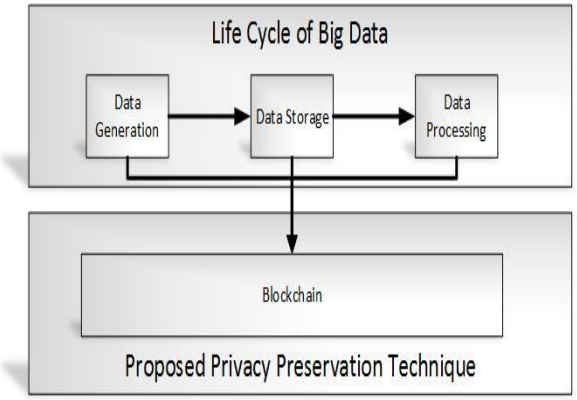

Figure 3: shows the blockchain method for all phases of the big data life cycle.

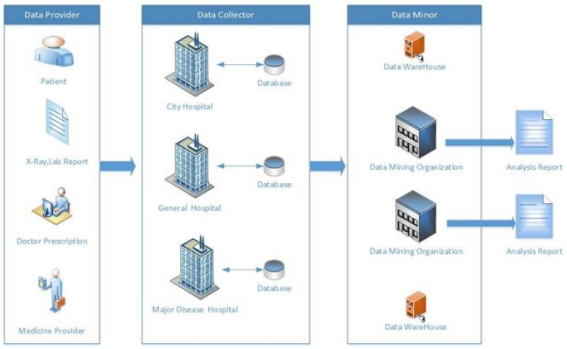

Figure 4: show role of the different actors in hospital area data mining

2. LITERATURE REVIEWS

The existing PPDM strategies have been thoroughly examined and classed based on their methods, which include data manipulation procedures, which seem to be the current technique's main downside. [5], [6].

To solve the problem of medical data sharing among network hospitals, the authors propose the MeDShare system.

MeDShare is a verifiable method for data auditing and control of medical shared data in cloud repositories for big data databases, based on blockchain technology. Cloud service providers and other data guardians can achieve data provenance and auditing while sharing medical data with entities like research and medical institutions with minimum risk to data privacy by utilizing MeDShare. Users can access data using their pre-generated private key, and if the key is legitimate, the user can access data; otherwise, the transaction is declared invalid [7]. The authors [8] suggested framework provides uniform admittance standards for these records and safe record storage based on blockchain. Because blockchain does not have enough storage capacity for huge data, the study utilizes an IPFS system to store off-chain storage. And the role-based access also benefits the system as the medical records are only available to the trusted and related individuals.

The blockchain technology creates a distributed ecosystem with decentralised and tamper-proof records, as well as a new way to secure and share electronic health record systems. The authors suggest a new electronic healthcare data sharing system based on blockchain that verifies data integrity. The author created a smart contract that allows users to access, alter, and remove healthcare records stored in the cloud. This solution employs a private key that is kept in the blockchain, and an authorized individual can use this key to access any cloud-based record using the smart contract [9].

The BlockDBM approach was presented by the authors for decentralised trust management systems for various tasks such as data gathering, data sharing, and storage for processing through permission blockchain-based smart contracts. The authors [10] used ways to improve blockchain technology's trustworthiness. For data privacy and security, blockchain technology eliminates the need for a third party. BlockDMB is a secure storage solution that uses both on-chain and off-chain storage mechanisms to keep parties' trust. Blockchain is very limited storage capacity as well as slow processing power. Blockchain as storage for traditional big data systems is not possible. The authors [11] created a blockchain-based personal management system. Using the on-chain and off-chain data storage methods, the author devised unique data-based storage to solve the problem of data redundancy and insufficient storage capacity. The authors found a solution to the problem of personal management system security in big data. The authors used a standard database-managed approach to overcome the blockchain challenge of data redundancy and storage.

This concept is expanded and improved, and a new ChainPPDM for Big Data privacy preservation is proposed. Many anonymization and encryption approaches have been developed to protect the privacy of big data while collecting relevant knowledge from it during the data mining process. At this time, all PPDM approaches are based on a centralized and distributed data-based system. Data suppliers should put their trust in a third party or a centralized system authorized person. We presented a decentralised blockchain-based method for large data privacy preservation, in which data is securely transmitted from data collector to data miner. Because blockchain's data storage capacity is restricted, we use HDFS for data mining. Private data is saved on-chain, while non-sensitive information or data required for data mining is stored off-chain. ChainPPDM is a permission-based decentralised system with excellent data mining feasibility and security.

3. BLOCKCHAIN STRUCTURE

A Blockchain is made up of a database of transactions linked by a cryptographic hash. To store enormous amounts of medical data, however, a big data storage system with a blockchain is required. [12]. The blockchain's fundamental design is depicted in Figure 5, which contains an individual block with a block header and a block body that is linked to the next to continue blocks. Each hospital has its own servers, and each doctor associated with that hospital has his or her own Blockchain network, which is connected to the others [13]. The entire network linked to the system manager will be in charge of system management.

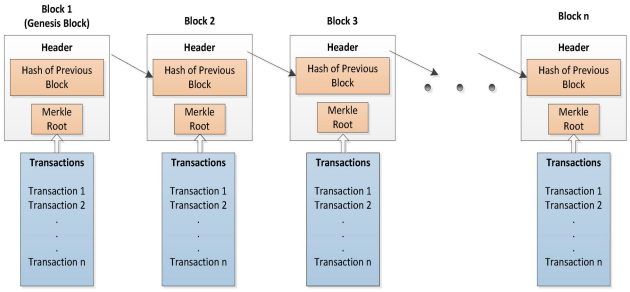

Figure 5: The basic execution structure of blockchain.

4. PRIVACY PRESERVATION OF BIG DATA-PERMISSION BC AND IPFS

The authors [14] proposed our novel schema and novel architecture, ChainPPDM, privacy preservation of big data, or privacy preservation of data mining. ChainPPDM is the core four technology used in implementation parts i.e a) Permission Blockchain - Hyperledger Fabric, b) HDFS, c) IPFS, d) SHA256 algorithm. Hyperledger Fabric can handle 3000 transactions per second. The ECDSA technique is used in the Hyperledger Fabric membership management, which is based on a conventional X.509 certificate. Each member receives a digital certificate through the PKI system. The Gossip protocol can only distribute data between nodes in the same MSP on the channel. MSP guarantees that access to the Fabric platform is secure.

ChainPPDM presents new on-chain and off-chain solutions for protecting big data privacy. We believe, however, that the big data analysis method is faster than a database system. Authors should keep bid data in mind while mining data or looking for patterns. The most serious issue is data privacy. For data analysis, ChainPPDM employed the HDFS big data system and a blockchain system for user privacy. Healthcare has information on customers who placed orders online. The Hospital's patient database contains both confidential and no sensitive information, such as the patient's address, neighborhood, and city. ChainPPDM is kept private in a Hyperledger Fabric collection, and non-sensitive data is kept in HDFS. To send data from the data collector database to the data minor warehouse, ChainPPDM employs IPFS.

4.1 Hash Value Generated and Data Transfer to Blockchain

From the entire amount of data, the data collector has constructed two independent sets of data. The patient's private data is one set, while nonsensitive data from data mining is the other.

$$w = \{p_t, s_t\} \; where \; p_t = \{p_1, p_2, p_3, \cdots p_n\} \quad (1)$$

is private data attributes of consumer

$$s_t = \{s_1, s_2, s_3 \ldots s_n\} \quad (2)$$

is nonsensitive data attribute of consumer and w is whole data set. The data collector is created data set for data mining. Before giving data set to transfer to Data Minor warehouse, Data collector is decided sensitive data attribute for data privacy. To measure hash value, we use

$$h_t = H(p_t, salt) \quad (3)$$

where p is the private data of the patient and we add some salt to generate hash because every same input of value for a hash function is generated the same value. It has a good chance of guessing a private value from similar hashes. For data of any length, the SHA256 algorithm can generate a 256-bit long hash value [15]. A hash value is a numerical representation of a piece of data that is unique and exceedingly compact. Generated has become essential for both sensitive and non-sensitive data. Now we create new data set using has value, np_t & ns_t. New data set has key and value pair. Hash value becomes key for access to private as well as non-sensitive value.

$$np_t = \{\{h_t, p_t\}\} \quad (4)$$

$$ns_t = \{\{h_t, s_t\}\} \quad (5)$$

np_i is inserted into blockchain private collection and ns_i is inserted into the HDFS system of data minor. Figure 6 shows the process of data set generation and stored in the blockchain and centralized database.

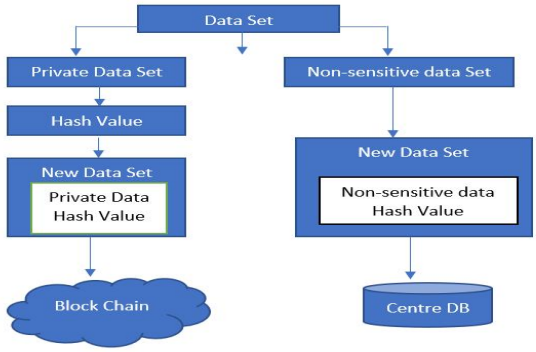

Figure 6: the process of data set generation and storage in blockchain and centralized database

5. ARCHITECTURE AND ALGORITHMS

IPFS is utilized by ChainPPDM to send data from the data collector to the data minor quickly and securely. ChainPPDM is implemented using Chaincode, MSP, statedb, IPFS, and HDFS. CouchDB is a state database that keeps track of transaction results. Only core data is stored in the blockchain network, while non-core data is stored in a central database. The MSP is a pluggable interface that consists of a set of encryption methods and protocols for issuing and verifying certificates and identities in the blockchain. The CA is used to generate certificates and secret keys, as well as to initialize the MSP, and the order node serves as a network proxy for data distribution. Figures 7 and 8 depict ChainPPDM's architecture and algorithms, respectively.

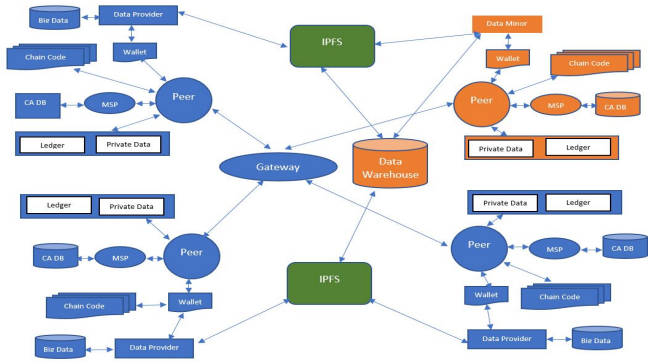

Figure 7: shows the architecture of ChainPPDM

6. EXPERIMENTAL RESULTS AND ANALYSIS

We created a prototype system based on the Virtual Machine for ChainPPDM and the Hyperledger fabric to ensure that our data separation and storage solution is genuinely effective for privacy preservation of big data using blockchain systems like data mining. The system runs on an Ubuntu 20.04 (64-bit) virtual machine,

Intel(R) Core (TM) i7-4790 CPU @ 3.60 GHz processor and 8 GB RAM, and uses the HDFS simulation central database. We observe the changes in the storage capacity after separately sensitive and non-sensitive data with the hash value of non-sensitive data. Table 1 shows compression between other method for privacy-preservation of big data with ChainPPDM. Hyperledger Caliper tool has been used for evaluation of ChainPPDM.

Algorithm 1. Insert Data into blockchain and data warehouse

Input: Data File
Output:Boolean

1. If not private data attribute and non-sensitive data attribute value exits
   Throws
2. If data duplicate
   Throws
3. Divine data in two parts
4. Create collection of private data
5. Create collection of public data
6. Appy SHA256 on private data
7. Define hash value as key
8. Create new private dataset using haskey and private
9. Create non-sensitive dataset using hashkey and private
10. Transfer new private dataset to blockchain
11. Transfer new non-sensitive dataset to data warehouse
12. Apply data mining algorithm
13. Find patterns
14. Give result to decision maker

return

Figure 8 shows the Algorithms of ChainPPDM

Table 1: shows compression between other method for privacy-preservation of big data with ChainPPDM.

	BlockBDM [8]	**PBMS [9]**	**ChainPPDM**
Public Blockchian	Ethereum	NA	NA
Permissioned Blockchain	hyperledger Fabric 1.4	hyperledger Fabric 1.4	hyperledger Fabric 2.0
Data Type	multimedia	text	text
Database before system	MySql	MySql	MySql
Nodes Number	4	4	4
On-Chain Storage	Blockchain	Blockchain	Blockchain
Off-chain Storage	IPFS	MySql	IPFS, HDFS

The average size of a single data provider record in single files is 182 bytes, and the data provider separates this data into two sections, with private data taking up 142 bytes and nonsensitive data taking up 40 kb. The storage load of non-sensitive hash data grew with byte kb of each record on blockchain as well as on centralized data minimal storage. Figure 9 depicts the Read Latency(second) plot using Hyperledger Caliper tool. BlochBDM and PBMS read latency low compare to ChainPPDM because ChainPPDM is used private state of Hyperledger Fabric. With the secrecy of data, ChainPPDM increases file size on hard disc.

Figure 10 depicts Write Latency(second) plot using Hyperledger Caliper tool. In comparison to the BlockBDM and PBMS, ChainPPDM function better because ChainPPDM write private data in private state of Hyperledger Fabric.

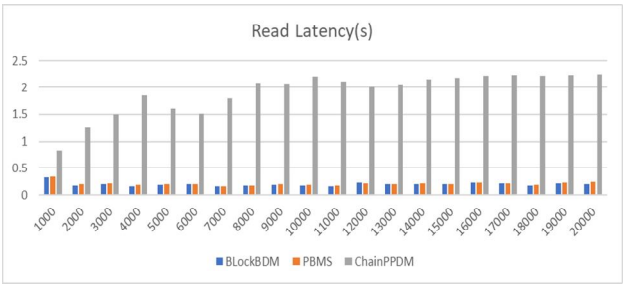

Figure 9: Read Latency(second) plot using Hyperledger Caliper tool

When data is minimal, uploading takes longer than cloud uploading, but as data grows, cloud speed suffers. ChainPPDM has a larger storage capacity, but it takes less time to do a data mining process while maintaining anonymity.

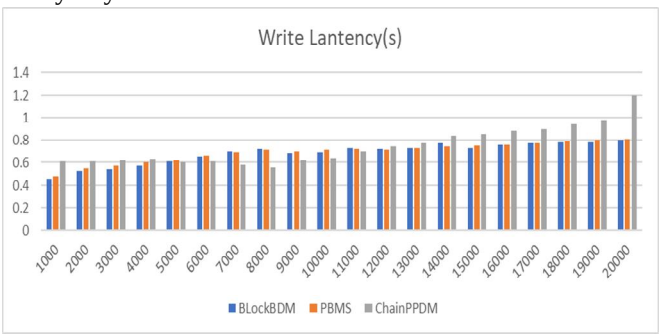

Figure 10: Write Latency(second) plot using Hyperledger Caliper tool.

7. CONCLUSION

ChainPPDM has solved the challenge of massive data privacy preservation. ChainPPDM presents a blockchain-based architecture for large data privacy preservation, as well as a prototype system to demonstrate feasibility and provide a solution for big data mining technologies. In addition, data was separated into two sections, which increased data privacy and solved the problem of greater data computation in blockchain vs traditional data mining systems. In this article, we will just utilise text big data to create the system; however, we will assess it and try to extend it to additional fields in the future.

ACKNOWLEDGEMENT

The authors are thankful to the GMB staff members, Dr. Nalin Jani, KSV and Dr. Seema Mahajan, IU for their best cooperation, support and guidance.

REFERENCES

1. P. Sreekumari, **Privacy-Preserving Keyword Search Schemes Over Encrypted Cloud Data: An Extensive Analysis**, in 2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS), 2018.
2. R. Lu, H. Zhu, X. Liu, J. K. Liu and J. Shao, **toward efficient and privacy-preserving computing in big data era**, IEEE Network, vol. 28, pp. 46-50, 2014.
3. D. Zhang, "Big Data Security and Privacy Protection," in Proceedings of the 8th International Conference on Management and Computer Science (ICMCS 2018), 2018/10.
4. A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua and S. Guo, **Protection of Big Data Privacy**, IEEE Access, vol. 4, pp. 1821-1834, 2016.
5. Q. Xia, E. B. Sifah, K. O. Asamoah, J. Gao, X. Du and M. Guizani, **MeDShare: Trust-Less Medical Data Sharing Among Cloud Service Providers via Blockchain**, IEEE Access, vol. 5, pp. 14757-14767, 2017.
6. A. Shahnaz, U. Qamar and A. Khalid, **Using Blockchain for Electronic Health Records**, IEEE Access, vol. 7, pp. 147782-147795, 2019.
7. S. Wang, D. Zhang and Y. Zhang, **Blockchain-Based Personal Health Records Sharing Scheme with Data Integrity Verifiable**, IEEE Access, vol. 7, pp. 102887-102901, 2019.
8. M. Zhaofeng, W. Lingyun, W. Xiaochang, W. Zhen and Z. Weizhe, **Blockchain-Enabled Decentralized Trust Management and Secure Usage Control of IoT Big Data**, IEEE Internet of Things Journal, vol. 7, pp. 4000-4015, 2020.
9. J. Chen, Z. Lv and H. Song, **Design of personnel big data management system based on blockchain**, Future Generation Computer Systems, vol. 101, pp. 1122-1129, 2019.
10. R. Kumar, N. Marchang and R. Tripathi, **Distributed Off-Chain Storage of Patient Diagnostic Reports in Healthcare System Using IPFS and Blockchain**, in 2020 International Conference on COMmunication Systems NETworkS (COMSNETS), 2020.
11. C. Rahalkar and D. Gujar, **Content Addressed P2P File System for the Web with Blockchain-Based Meta-Data Integrity**, in 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), 2019.
12. **The Hyperledger Project. [Online]. Available:** [Online]. Available: http://www.Hyperledger .org..
13. D. Rachmawati, J. T. Tarigan and A. B. C. Ginting, **A comparative study of Message Digest 5(MD5) and SHA256 algorithm**, Journal of Physics: Conference Series, vol. 978, p. 012116, 2018.
14. A. Vaghela and A. Suthar, **A review of Bigdata analysis using smart contract,** Juni Khyat Journal, vol.X, issue.XII, pp 109-115, 2020.
15. S. H. Begum and F. Nausheen, **A comparative analysis of differential privacy vs other privacy mechanisms for Big Data**, in 2018 2nd International Conference on Inventive Systems and Control (ICISC), 2018.