Data Deduplication in A Blockchain-Enabled Big Data Ecosystem: Secure and Efficient Cloud Storage

S. Seethalakshmi[1*], Dr. B. Balakumar[2]

[1*]*Research Scholar, Manonmaniam Sundaranar University, Tirunelveli. Tamil Nadu, India, seethasri.lakshmi@gmail.com*
[2]*Professor, Manonmaniam Sundaranar University, Tirunelveli. Tamil Nadu, India balakumarmsu@gmail.com*

KEYWORDS

Local Storage, Cloud Storage Services, Data Compression, Blockchain, Large Amounts Of Data, Data Security And Privacy, Ethereum

ABSTRACT

To get around the limits of local storage and ensure that data can be accessed from anywhere at any time, cloud storage services have become very important. Data compression technology helps save bandwidth and disk space. Removing extra data from cloud storage can save up to 95% of room. Despite recent improvements in cloud storage, there are still security and privacy concerns. This piece discusses the security and privacy concerns related to big data and outlines the basic requirements that future solutions should address. This paper aims to introduce a new technology strategy for handling and controlling security and privacy risks related to big data. This paper uses a design science study method. The proposed system uses Blockchain technology to securely store large amounts of data by managing its information and rules and keeping outside parties out protecting data security and privacy. This method uses very little computing power to create file encryption keys. Tests and security checks show that it keeps the keys safe and protects the data's privacy. Computational speed improves because it takes less than 2 seconds to create encryption keys, even for a 500 MB file. We show a working model of our proposed system in a real-life data store scenario using the Ethereum Blockchain. The review and research show that our proposed framework provides a useful method for safely storing data in a Big Data environment.

1. Introduction:

Cloud computing is a novel approach to computing that takes advantage of advancements such as distributed computing, virtualization, and dynamic scheduling to provide scalable computing services, flexible demand allocation, and elastic resource scheduling [1]. It benefits both businesses and customers. This successfully addresses issues such as inefficient storage and unequal resource sharing by significantly increasing the rate of computing resource utilization [2]. By adopting the cloud computing service paradigm, both people and businesses may reduce their data management expenses. Through distributed systems, heterogeneous storage devices, and cluster functions, this approach provides appropriate resource configuration and the purchase of corresponding compute and storage resources as required. Cloud storage a major technical extension of cloud computing development [3] provides remote and effective data storage services to people and enterprises by using network services and integrating with a wide variety of heterogeneous storage devices. This move away from the traditional wisdom of keeping data locally will result in a significant reduction in the exorbitant costs of data management and maintenance. Cloud storage has quickly become the industry standard for data storage due to its numerous advantages, including low cost, high efficiency, scalability, and energy consumption. The growing volume of consumer data is putting a lot of strain on cloud storage providers. According to studies from the Internet Data Centre (IDC), global data storage will approach 16ZB in 2016 and may potentially surpass that amount by 2025 [4].

To reduce the need for service space and bandwidth, the deduplication approach is extensively used in cloud storage. Deduplication ensures that data that is freshly uploaded to the cloud is not duplicated with existing data in the cloud. Rather, the client's information is associated with the same data stored in the cloud. The data stored may be accessed and retrieved at any time by clients who have joint ownership of it [5]. Data deduplication offers the same benefits as storing data multiple times while possibly saving 90% of storage space [6].

Users should be mindful of various security problems while using the service. Service interruptions or the revelation of crucial data are common concerns for corporate customers, in contrast to the concerns of individual consumers over possible breaches of personal information. Immediate encryption of client data uploads is essential to resolving these security problems. While uploading and encrypting data, customers will likely utilize several secret keys. The cipher texts will be different when using various secret keys to encrypt the same material. Because of the duplicate data check in cloud storage, it is impossible to tell whether two cipher texts are indeed generated from the same underlying plaintext. Consequently, data deduplication is out of the question. A novel approach to data deduplication encryption, convergent encryption (CE) was developed as a means to circumvent this issue [7]. Keys for encryption in CE are created automatically from the data. The data's hash value is the encryption key, guaranteeing that two sets of identical data will produce the same deciphertext. As a result, deduplicating encrypted data for many clients sharing the same data becomes viable.

Because blockchain technology provides consumers with security advantages, it has found several applications. These consist of transaction information traceability, transparency, and immutability. Intelligent transportation, supply chain management, industrial production, education, and financial services are a few of these domains. The decentralized nature of blockchain technology may help alleviate the cloud storage systems' weaknesses outlined above.

Large volumes of data are produced by websites, social media, the Internet of Things, multimedia archives, and other customized services; this is referred to as "Big Data." Another way to define big data is as a collection of very massive or intricate data sets that are challenging to manage using conventional database management systems [8]. Unstructured, semi-structured, and structured are the three primary categories into which big data may be divided [9]. One of the several systems mentioned above is often used to gather and store data. As more individuals become aware of the significance of data in decision-making, big data storage has gained popularity as a debated subject. In only the previous 20 years, the quantity of Big Data has increased significantly, from PETA-BYTE levels to ZETA-BYTE levels. In 2020, there will be over 40 terabytes of data, according to EMC and IDC predictions [10, 11]. IDC's most recent projections show that by 2025, the volume will have surpassed 75 ZETA-BYTES [12].

1.1 Contribution:

- The creation of a mobile agent and blockchain architecture for big data apps to enable safe data exchange and storage.
- For the blockchain architecture, two layers with two private blockchains are suggested; data is preserved using IPFS technology.
- To provide basic device security, a trustability check mechanism is proposed for new joint requests to the Big Data system.
- To create the suggested method to securely operate on local storage rather than cloud storage to protect data privacy;
- To develop a generic, lightweight Big Data solution for many contexts
- Requires reliable third-party key management server.
- Encrypts key security through block transactions.
- Prevents unauthorized access.
- Convergent 3DES encryption and decryption techniques hide data.

- Despite human input, brute-force attacks can't acquire original data.
- Provides vital security and data confidentiality.

2. Related Works:

Enterprise backup storage setup is studied in [13] using the secure chunk-based deduplication approach. The authors proposed generating keys at random utilizing internal aspects of the backup service. On the other hand, low chunking throughput is also a reality. It describes three fingerprint deduplication techniques: chunking, fingerprinting, and fingerprint indexing. In chunking, the boundaries between chunks are determined by the values of the divisors [14].

However, the low deduplication throughput can be due to the ZEUS process. The authors integrate Application Programming Interfaces (APIs) from many cloud-based services, such as those that provide video finding, scanning, security recovery, quantitative evaluation, and more, into their security assessment model [15]. Nevertheless, the adversary could be able to brute-force their way into accessing the data if it is in a knowable set for CE. The issue of optimizing video caching with effective cellular and bandwidths is addressed in [16] by proposing a Collaborative Edge (CE) network and the CE-D2D framework.

Every hash has a unique block identification. Blockchain has the potential to revolutionize several industries, including finance, healthcare, education, voting, and more, due to its many benefits. The most important benefits are decentralization, security, immutability, auditability, traceability, and transparency [17]. The three primary types of blockchains are consortium, private, and public [18]. A public Blockchain network may be joined by anyone. They are decentralized and neither regulated nor directed by any organization [19, 20]. Private blockchains are not accessible to the general public and are only available to institutions [21]. A consortium blockchain, which lies between the public and private blockchains, is often utilized when there are many user categories [22]. The main focus of the many Blockchain systems that have been developed has been on smart contracts and decentralized apps. These systems differ in several aspects, including public vs. permissioned networks, integrated cryptocurrency support, transaction process, pricing, privacy, and performance. The decentralized, open-source Ethereum network was introduced by the Ethereum Foundation in 2015. Ethereum's main purpose is to serve as a platform for smart contracts. Additionally, Ethereum includes Solidity, a programming language that enables anyone to create their own Blockchain [23]. The smart contract is one of the attractive aspects of Blockchain. This code becomes executable when it is deployed to the Blockchain network. It is triggered when a transaction is sent to a smart contract. Blockchain technology and smart contracts may be used to provide real-time, very secure process execution. This integration accelerates the conception, development, and deployment of solutions to real-world problems by removing the need for an intermediary [24]. Access Control (AC) is a crucial part of Big Data when it comes to safeguarding private data. Every business that works with big data should have a policy in place for access control. Access control is the policy that regulates how different nodes are connected to the system. An inadequate access control system might allow hackers to access stored data without authorization, jeopardizing security and privacy [25]. By giving nodes and devices security and privacy permissions, AC policies help secure data. Despite much research, there are still a lot of unanswered concerns about access control systems.

3. Methodology:

This study describes novel ways to data integrity in cloud storage using cloud-based data de-duplication analysis and policy-based encryption. We have performed storage and cloud-based data analysis, used policy-based encryption to guarantee data integrity, and conducted e-duplication data analysis using blockchain technology. Utilizing the below-described system model, we will augment the current threat model. We will conduct a thorough review of the risk model associated with the blockchain system model for the primary reasons outlined below.

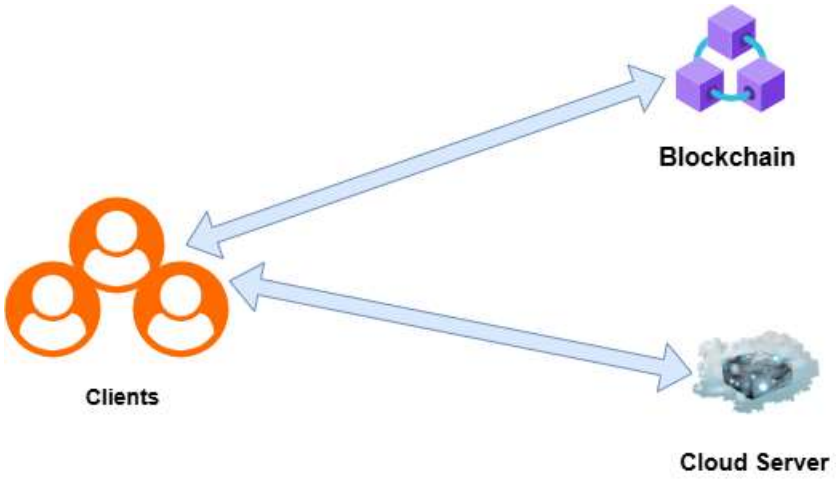

Figure 1: System Model

Figure 1 shows the proposed scheme's system model, which includes the User, CSP, and blockchain. Users outsource data to cloud storage services to preserve local storage space owing to limited availability. Users must securely encrypt data and allow the cloud storage service provider to store it on the cloud server to ensure data security. Cloud storage companies can supply ample space. In a cloud storage system that optimizes data deduplication, cloud service providers must discover duplicate user data and provide storage space to remove it. Since 2008, blockchain, a decentralized database system, has garnered popular attention. Bitcoin and Ethereum, blockchain-based apps, have grown in popularity and relevance among experts. Blockchain has progressed through three phases: blockchain 1.0, 2.0, and 3.0, each with more applications. Academics have focused on blockchain security issues. Malicious actors may utilize blockchain technologies to launch Sybil, dust, double-spend, DDoS, and mining attacks. These may disrupt blockchain systems, preventing their functioning or forcing their collapse. Blockchain design generally has six layers: application, contract, incentive, consensus, network, and data. Each layer performs specific duties to produce a decentralized distributed ledger. Blockchain, an open and decentralized distributed management system, uses encryption, distributed network architecture, and consensus procedures to secure transaction data. This paper suggests leveraging Ethereum for reliable convergence key management and creating a cloud storage key security management system.

We can take the required precautions to protect the blockchain system from potential threats by analyzing the threat model, which helps us understand and anticipate the many circumstances that might put it in danger. To avoid security breaches, the threat model may be used to find weak spots in the blockchain system's defenses and fix them before attackers can use them. By delving into the threat model, we may enhance the blockchain system's resilience to internal and external threats without compromising its functioning. One way to get a better grasp on the internal and external threats that the blockchain system encounters is to examine the threat model. In turn, this lends credence to the idea that tailored security policies and approaches are required.

A potential external threat comes from unauthorized users who do not have the proper credentials to access the system. The proposed method is seriously jeopardized by outside forces who want user data via collusion with system businesses, including suppliers of cloud storage.

One example of an internal adversary would be a sincere and inquisitive person working within the system. This kind of company will strictly adhere to the terms of service while also seeking out customer data and divulging as much of it as possible to external competitors for financial gain. Insider threats pose a bigger threat to user data security than external ones since they have access to more user data.

3.1 Proposed System Model

The system model of our proposed secure deduplication process is shown in Figure 2. Examples of entities in the system model include the following:

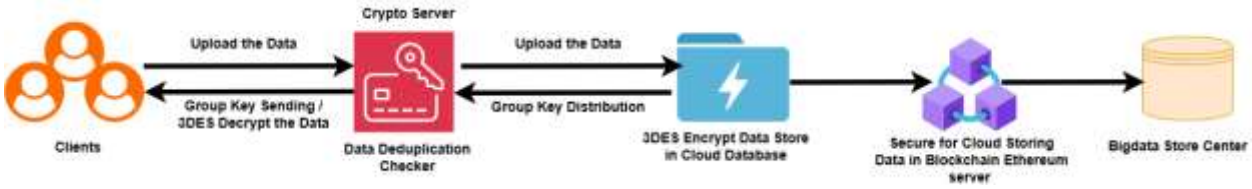

Figure 2: Proposed Architecture

Figure 2 shows a secure data storage technique that integrates the Ethereum blockchain with Triple Data Encryption Standard (3DES) encryption to safeguard 500 MB of data on a big data server in only 10 seconds. Data is sent via a Crypto Server once clients have uploaded it to the system. To make the most efficient use of storage space, the crypto server employs a data deduplication checker to eliminate duplicate data. All of the data is securely encrypted using 3DES. A robust group key is assigned to approved clients. By assuring a secure multi-layer encryption technique, this enhances protection against unauthorized access. After encryption, the data is sent to a secure cloud database for temporary storage. The next step is for the encrypted data to pass via a Secure Ethereum Blockchain Server. This server employs Ethereum's distributed ledger technology to maintain an immutable and unbreakable record of the store transactions and encryption keys. This ensures that the whole storing operation is secure and transparent. Finally, the data is securely stored in the Big Data Store Centre after receiving encryption. Using 3DES encryption, the system ensures data privacy; the blockchain from Ethereum provides credibility and security. Management of large datasets in resource-intensive environments benefits greatly from the procedure's efficiency, which permits the secure storage of 500 MB of sensitive data in only 10 seconds. Table 1 displays the pseudocode for proposed work.

Table 1: Pseudocode for Proposed Model

```
Step 1: Initialize System Variables:
  Input: Data D (500 MB), Group Key G
  Output: Encrypted Data E, Storage Location L
Step 2. Client Data Upload:
  Upload D → Crypto Server
Step 3. Deduplication Check:
  If D exists in Database:
     Skip to Step 8
  Else:
     Proceed
Step 4. Generate Group Key for 3DES:
  G = GenerateKey(ClientID, Timestamp)
Step 5. Encrypt Data Using 3DES:
  E = 3DES_Encrypt(D, G)
  E = 3DES(D, G1, G2, G3)
Step 6. Upload Encrypted Data to Cloud:
  Store E → Cloud Database
Step 7. Log Metadata on Ethereum Blockchain:
  BlockchainLog(G, FileHash(E))
Step 8. Validate Data Retrieval:
  Request → Blockchain Verification
Step 9. Decrypt Data (Optional):
  D = 3DES_Decrypt(E, G)
Step 10. Store in Big Data Center:
  FinalStorage(L, E)
```

3DES Equation:

3DES uses multiple keys in three distinct phases to encrypt data in Groups G1, G2, and G3.

$$E = E_k 3(D_k 2(E_k 1(D)))$$

Where, E_k Is encrypted for a single DES key, D_k Is decrypted for a single DES key. Using immutable records and blockchain verification guarantees that data remains private and intact.

4. Results and Discussion:

To verify the solution, the proposed system was implemented and tested. Here, we provide the details of how our proposed framework will be implemented. The prototype that was implemented consists of two parts. The client interaction is the front end, while the Blockchain-based security management is the back end. The Blockchain platform was essential to the development of the security manager. OpenChain, IBM Blockchain, and Ethereum are just a few of the several Blockchain systems that are accessible.

We used the private blockchain Ethereum as the foundation for our platform. The Big Data management system involves a lot of sensitive information. Because of this, the core Blockchain architecture of our proposed solution is based on private blockchain. The fact that Ethereum leverages the entire programming language Solidity to create smart contracts, its user-friendly interface, and its interoperability with decentralized applications are just a few of its benefits. Ethereum has a higher verified transaction rate per second compared to Bitcoin. The most important part of the prototype system is its smart contract. We used the python flask and JavaScript programming language to construct it. For the front-end web GUI, Metamask and Javascript were used to create forms that users could interact with. The machine has an Intel(R) Core(TM) i7-1035G1 CPU running at 2.8 GHz, 16 GB of RAM, and the 64-bit Ubuntu 20.04 operating system. It uses cloud services to imitate storage nodes. We tested the suggested framework using the following configurations:

Infrastructure for the private blockchain was supplied by four virtual machines running on Google Cloud. The virtual desktops came with Intel i7 processors, which had four cores and ran at 2 GHz. Hyperledger Besu v2.21.2, an open-source client for Ethereum that enables a permissioned private Blockchain, was used in the construction of the private Blockchain. Using the Clique Consensus Protocol, our P2P network had three nodes that could communicate with one another and one node that could verify their validity. For this, we rely on Hyperledger Calliper v0.3.1, a tool that assesses blockchain systems. Resource utilization, read latency, transaction throughput, transaction delay, and transaction latency are just a few of the performance characteristics included in the Calliper tool's performance report. Solidity and JavaScript are the languages that are used for programming. Figure 3 shows file size measurements for timestamps and Figure 4 displays the 3DES data security mentioned in D, G1, G2, G3, and G4.

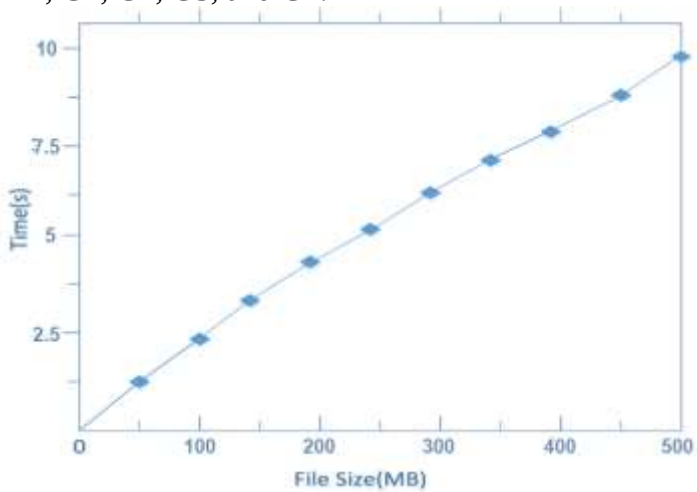

Figure 3: File Size

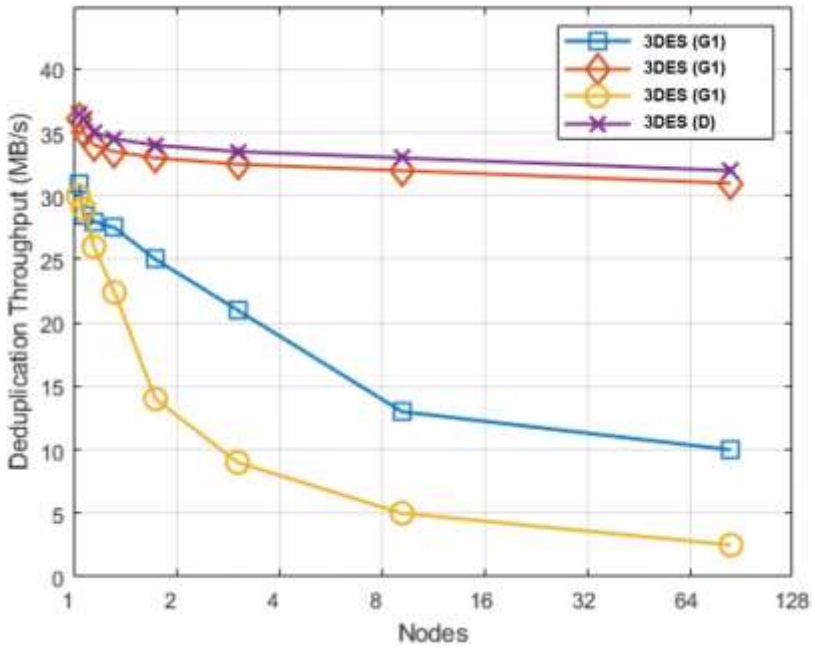

Figure 4: Blockchain Nodes in 3DES for Deduplication Throughput

5. Conclusion:

The goal of this initiative was to increase Big Data's security and privacy. Through the development of a functional prototype, we were able to identify the most critical issues with current Big Data storage systems and provide a practical fix. Our main objective was to provide a dependable and comprehensive Blockchain-based architecture for secure and effective data interchange and storage in Big Data scenarios. To test the effectiveness of the approach, we deployed an Ethereum Blockchain on Google Cloud. To achieve decentralized data storage, we also integrated the Blockchain with the decentralized 3DES secure block server storage system. We were able to greatly improve Big Data privacy and security while avoiding policy manipulation by using two private Blockchain layers for metadata storage and data access regulations for policy and metadata security. To achieve objectives such as reducing traffic overheads and showcasing the high level of intelligence and mobility of our solution, we used a mobile agent to confirm the security requirements of the new user device. We conducted a performance evaluation to ensure that our proposed framework was effective and to identify areas where it may be improved for further research. We also discussed how our suggestion differed from other concepts. According to our model, Blockchain technology might be helpful in a range of Big Data scenarios, and we think it's a step in the right direction for improved data security and privacy.

References:

1) Liang H, Li M, Chen Y, et al (2021) Architectural protection of trusted system services for Sgx enclaves in cloud computing[J]. IEEE Trans Cloud Comput 9(3):910–922
2) Huang Q, Yang Y, Yue W, et al (2021) Secure data group sharing and conditional dissemination with multi-owner in cloud computing[J]. IEEE Trans Cloud Comput 9(4):1607–1618
3) Chen N, Li J, Zhang Y, et al (2022) Efficient Cp-Abe scheme with shared decryption in cloud storage[J]. IEEE Trans Comput 71(1):175–184
4) Xia W, Jiang H, Feng D et al (2016) A comprehensive study of the past, present, and future of data deduplication[J]. Proc IEEE 104(9):1681–1710
5) Ng, W.K.; Wen, Y.; Zhu, H. Private data deduplication protocols in cloud storage. In Proceedings of the 27th Annual ACM Symposium on Applied Computing, Trento, Italy, 26–30 March 2012; pp. 441–446.
6) Dutch, M. Understanding data deduplication ratios. In Proceedings of the SNIA Data Management Forum, Orlando, FL, USA, 7 April 2008; Volume 7.

7) Douceur, J.R.; Adya, A.; Bolosky, W.J.; Simon, P.; Theimer, M. Reclaiming space from duplicate files in a serverless distributed file system. In Proceedings of the 22nd International Conference on Distributed Computing Systems, Vienna, Austria, 2–5 July 2002; pp. 617–624.
8) Baig, M.I.; Shuib, L.; Yadegaridehkordi, E. Big data adoption: State of the art and research challenges. Inf. Process. Manag. 2019, 56, 102095.
9) Samsudeen, S.N.; Haleem, A. Impacts and challenges of big data: A review. Int. J. Psychosoc. Rehabil. 2020, 7, 479–487.
10) Bao, R.; Chen, Z.; Obaidat, M.S. Challenges and techniques in Big data security and privacy: A review. Secure. Priv. 2018, 1, e13.
11) Yang, P.; Xiong, N.; Ren, J. Data security and privacy protection for cloud storage: A survey. IEEE Access 2020, 8, 131723–131740.
12) Alex, W. Global DataSphere to Hit 175 Zettabytes by 2025, IDC Says. Datanami 2018, 17, 13237–13244.
13) Wenhai Sun, Ning Zhang, Wenjing Lou & Thomas Hou, Y, "Tapping the Potential: Secure Chunk-based Deduplication of Encrypted Data for Cloud Backup", IEEE Conference on Communications and Network Security (CNS), pp.1-9, 2018.
14) Shubhanshi Singhal, AkankshaKaushik & Pooja Sharma, "A Novel Approach of Data Deduplication for Distributed Storage", International Journal of Engineering & Technology, vol.7, no.2.4, pp.46-52, 2018.
15) Aobing Sun, Guohong Gao1, Tongkai Ji & Xuping Tu, "One quantifiable security evaluation model for cloud computing platform", 2018 Sixth International Conference on Advanced Cloud and Big Data, 978-1-7281-3129-0/20/$31.00 ©2020 IEEE, pp.197-201, 2020
16) Emna Baccour, Aiman Erbad, Amr Mohamed, Mohsen Guizani & Mounir Hamdi, "CE-D2D: Collaborative and Popularity-aware Proactive Chunks Caching in Edge Networks", Ninth International Conference on Computational Intelligence and Security, pp.607-609, 2020
17) Conoscenti, M.; Vetro, A.; De Martin, J.C. Blockchain for the Internet of Things: A systematic literature review. In Proceedings of the 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), Agadir, Morocco, 29 November–2 December 2016; pp. 1–6.
18) Zahid, J.I.; Ferworn, A.; Hussain, F. Blockchain: A Technical Overview; IEEE Internet Policy Newsletter; IEEE: Piscataway, NJ, USA, 2018; pp. 1–3.
19) Dai, H.N.; Zheng, Z.; Zhang, Y. Blockchain for Internet of Things: A survey. IEEE Internet Things J. 2019, 6, 8076–8094.
20) Aujla, G.S.; Chaudhary, R.; Kumar, N.; Das, A.K.; Rodrigues, J.J. SecSVA: Secure storage, verification, and auditing of big data in the cloud environment. IEEE Commun. Mag. 2018, 56, 78–85.
21) Dinh, T.T.A.; Wang, J.; Chen, G.; Liu, R.; Ooi, B.C.; Tan, K.L. Blockbench: A framework for analyzing private blockchains. In Proceedings of the 2017 ACM International Conference on Management of Data, Chicago, IL, USA, 14–19 May 2017; pp. 1085–1100.
22) Lei, K.; Zhang, Q.; Xu, L.; Qi, Z. Reputation-based byzantine fault-tolerance for consortium blockchain. In Proceedings of the 2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS), Singapore, 11–13 December 2018; pp. 604–611.
23) Shahnaz, A.; Qamar, U.; Khalid, A. Using blockchain for electronic health records. IEEE Access 2019, 7, 147782–147795.
24) Mohanta, B.K.; Panda, S.S.; Jena, D. An overview of smart contract and use cases in blockchain technology. In Proceedings of the 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bengaluru, India, 10–12 July 2018; pp. 1–4.
25) Centonze, P. Security and privacy frameworks for access control big data systems. Comput. Mater. Continua 2019, 59, 361–374.