International Information and Engineering Technology Association
Advancing the World of Information and Engineering

International Journal of Safety and Security Engineering
Vol. 14, No. 2, April, 2024, pp. 435-446
Journal homepage: http://iieta.org/journals/ijsse

Security Concerns and Data Breaches for Data Deduplication Techniques in Cloud Storage: A Brief Meta-Analysis

Anjuli Goel[1], Chander Prabha[1*], Meenakshi Malik[2], Preeti Sharma[1]

[1] Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab 140401, India
[2] Department of Computer Science, BML Munjal University, Gurgaon 122413, India

Corresponding Author Email: prabhanice@gmail.com

https://doi.org/10.18280/ijsse.140211

Received: 1 February 2024
Revised: 7 March 2024
Accepted: 12 March 2024
Available online: 26 April 2024

Keywords:
storage, cloud, deduplication, encryption, security, information, data

ABSTRACT

Over the last decade, data has exploded on cloud storage, and outsourcing data to cloud storage has become an appealing trend, which is not a fully reliable service. Data growth has made cloud storage management difficult. To provide more efficient and secure data storage on the cloud, cloud service providers employed a deduplication technique. Cloud data deduplication has developed as a popular research subject to boost the efficiency of cloud storage and minimize network communication traffic. It is not easy to provide security while increasing the huge volume of data on the cloud. Users are more concerned about the security of their data, as the data on the cloud is not secure and safe even after encryption. The various encryption and decryption available for providing the security to data. Conventional encryptions cannot be employed in data deduplication. To maintain the confidentiality and integrity of data, Convergent encryption and proof of ownership are applied. Various other approaches like proof of retrievability, Dekey, DupLESS, Identity-based encryption, Message-locked encryption, Attribute-based encryption, provable data possession, proof of storage with Deduplication, etc. are the security research topics these days. This paper presents a review of the literature on several proposed methodologies for safe deduplication techniques in cloud storage and current research trends. The primary contribution of the paper is to offer a complete understanding of the issues and solutions associated with safe data deduplication in cloud storage systems. The study dives into the various encryption solutions, security concerns, and potential challenges with data deduplication in the cloud.

1. INTRODUCTION

Cloud computing refers to data that is calculated and stored on the internet, and it offers several benefits to its consumers. Cloud computing provides multiple advantages, including the capacity to access diverse resources on demand, the availability of updated software, and the option to pay for resources as they are used, making user chores easier [1, 2]. Because of the benefits listed above, organizations that are unfamiliar with their internal mechanisms can store their data on the cloud. Cloud users may access their data from any place or device [3]. Many social networks, including Facebook, LinkedIn, Amazon, and Hotmail, as well as search engines such as Yahoo, Bing, and Google, employ cloud computing technologies to store their customer's data [4]. The use of cloud computing has various issues in terms of data privacy and security, a lack of flexibility when switching from one cloud to another, cloud performance, network reliance, and the requirement for experienced and qualified people to manage the cloud [5]. During and after the COVID pandemic, the usage of mobile, computers increased the tremendous volume of data on the cloud all over the globe. This increased data is a major concern and requires storage for the users. People transmit their data to distant storage since they have limited storage space, overhead and upkeep, and so less budgeting resources, hence cloud computing is becoming a more prevalent option [6].

According to the International Data Corporation (IDC) report, data will grow from 33 ZB in 2018 to 173 ZB in 2025. Figure 1 depicts the global data sphere, like a repository of data collected from digital data and other various sources [6, 7]. The internet is critical in enabling easy access to cloud computing resources. The internet has become a vital tool in our everyday lives, with a plethora of apps that have transformed how we interact, work, study, and enjoy ourselves. The desire for real-time and on-demand data access has spurred the growth of cloud computing, whether it be social networking platforms, video streaming services, collaboration tools, or corporate applications [8]. The internet's adaptability has made it an indispensable aspect of modern life, altering, and renewing the landscape of human contact and participation across a wide range of applications. Every minute, a lot of users use the internet to upload content on the cloud or exploit the content already there on the cloud [9, 10].

An Internet Minute infographic often emphasizes the mind-boggling quantity of online activity that takes place in only one minute. Figure 2 and Figure 3 depict the internet minute infographic for the year 2020-2021 and 2022. It can be seen how various cloud apps consume cloud storage. The infographic graph in Figure 4 also shows the growth of users on cloud surfing for different applications through the internet in the year 2023. These figures demonstrate the astounding speed and magnitude of internet activity in a single minute. The goal of such infographics is to highlight the huge volume of data and interactions taking place in the digital sphere, emphasizing the importance of the Internet in our everyday lives.

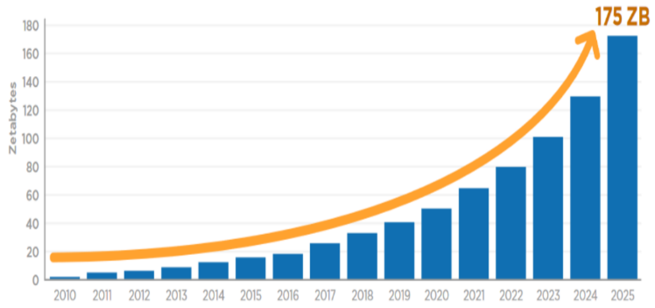

Figure 1. Annual size of global data sphere [7]

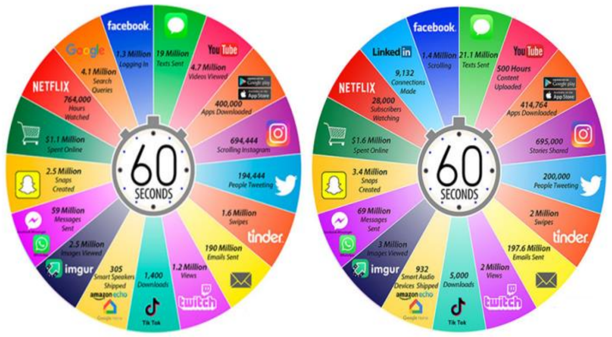

Figure 2. Internet minute infographic 2020-2021 [11]

A research study paper is important in the field of research since it serves several purposes and contributes to the expansion of knowledge and understanding in a certain area. Table 1 explains the significance of this research report and how it might contribute to the field.

Further, Section 2 provides insights into various cloud security issues. Section 3 presents a brief literature analysis of secure deduplication techniques for cloud storage. Section 4 identifies the problems that arise after a thorough analysis of the literature conducted in Section 3 along with future scope for researchers. Finally, Section 5 concludes the findings.

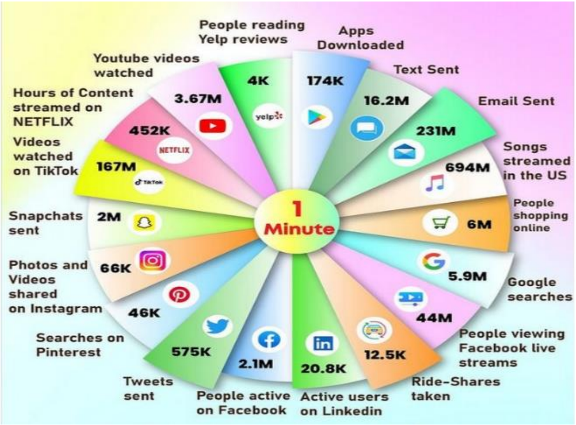

Figure 3. Internet minute infographic 2022 [12]

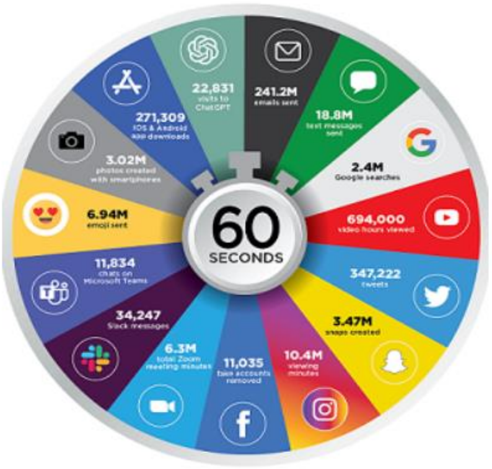
22,831 visits to ChatGPT
241.2M emails sent
18.8M text messages sent
2.4M Google searches
694,000 video hours viewed
347,222 tweets
3.47M snaps created
10.4M viewing minutes
11,035 fake accounts removed
6.3M total Zoom meeting minutes
34,247 Slack messages
11,834 chats on Microsoft Teams
6.94M emoji sent
3.02M photos created with smartphones
271,309 iOS & Android app downloads
60 SECONDS

Figure 4. Internet minute infographic 2023 [13]

Table 1. Research questions

Questions	Objectives
Why traditional encryption techniques are not compatible with data deduplication?	It helps to learn about available encryption techniques.
What are the most successful tactics for increasing data integrity and confidentiality in cloud storage by incorporating secure deduplication techniques?	It helps to learn about secure data deduplication techniques in trend.
How can dependability and proof of ownership in multi-cloud architectures be improved to reduce security concerns and data breaches in cloud storage systems?	It is very common to steal or destroy data while transferring in multi-cloud architectures. It helps in determining how to save data in multi-cloud or hybrid cloud storage systems.
What novel ways may be created to decrease communication overhead between source and target systems in cloud storage settings while ensuring data integrity and security?	It helps in determining which one is better: client-side or server-side deduplication along with integrity and security.

2. CLOUD SECURITY ISSUES: SURVEY MOTIVATION

2.1 Cloud computing security issues

Security is the main challenge in the cloud environment. A group of security precautions intended to safeguard cloud-based data, apps, and infrastructure is referred to as cloud security. Every organization is moving to the cloud to store their data and exploit the cloud services offered the bill as per use, resource pool, updated software, etc. It is becoming difficult to provide security when the data is at rest, data is moving. An unauthorized user can access the other's data for any misconduct use. Data can be attacked by internal employees as well as external people of any organization [14, 15].

Cloud security is the most important part of organizations and needs to be focused. Various types of attacks exist these days due to which data can be leaked or theft or cause many losses. To implement strong security mechanisms, important to identify the various threats and attacks that could happen. User credentials are compromised, hijacking accounts, data breaches, open interfaces and ports, loss of data, DOS, malicious attacks, ransomware, and many more [14, 16]. As per the analysis 2023 done by Verizon, 50% of attacks are social engineering, 74% are breached due to the misuse or compromised credentials, and 83% of data is breached by external activities. 24% of breaches involved ransomware, phishing, etc. [17].

As per the report provided by the UK [18]. In 2023 the total number of records were breached 8,214,886,660 and a total number of 2814 incidents took place. Table 2 gives all the details of incidents and data breached month-wise in the year 2023. As per the provided business report, 29% of businesses experience the same attack every 3 years as they have not upgraded their security system, 59% of medium businesses and 69% of large organizations got affected because their data had been compromised. 81% of attacks were email phishing and 79% of attacks were through ransomware attacks out of which 62% of businesses were affected and down more than 6 days. 82% of cyber-attacks were directly focused on public and private sectors [19]. Dark Beam with 3.5 billion breached records was the biggest record breached in the UK in 2023. Victoria Court System Data Breach is the current data breach in January 2024. The authorities reported that unauthorized parties gained access to court hearings, except that no loss happened [20].

It is the foremost principle to ensure cloud security and data deduplication for storage at the same time. According to a recent survey, it is concluded that 75% of data on the cloud is redundant and it can exceed 95% in coming years. Data owners lose their control over data after outsourcing the data on the cloud which is still a big question on data security and privacy. The owners encrypt the data before sending it to the cloud which is again a hurdle in the deduplication process because it is possible to encrypt the same data with different keys which generate different ciphertexts for the same data. So, message-locked encryption, proof of ownership, multi-set authentication, and provable data possession are some techniques provided as a solution to it [21, 22].

Table 2. Number of incidents and data breached records month-wise in the year 2023

Months	Total Incidents	Data Breached	Top Three Breaches Month-Wise
January	104	277,618,767	• Twitter • T-Mobile • JD Sports
February	106	29,582,356	• PeopleConnect • Elevel • CentraState Medical Center
March	100	41,970,182	• Latitude Financial • GoAnywhere • AT&T
April	120	4,353,257	• Shields Health Care Group • NCB Management • Kodi
May	98	98,226,877	• Luxottica • MCNA Insurance • PharMerica • MOVEit
June	79	14,353,113	• Oregon and Louisiana departments of motor vehicles • Genworth Financial • Wilton Reassurance
July	87	146,290,598	• Tigo • Indonesian Immigration Directorate General • Teachers Insurance and Annuity Association of America
August	73	79,729,271	• UK Electoral Commission • Pôle emploi • University of Minnesota
September	71	3,808,687,191	• DarkBeam • Undisclosed Restaurant Database
October	114	867,072,315	• ICMR (Indian Council of Medical Research) • 23andMe • Redcliffe Labs
November	470	519,111,354	• Kid Security • SAP Se Bulgaria • TmaxSoft
December	1,351	2,241,916,765	• Real State Wealth Network • TuneFab • Dori Media Group

Although various security measures have been taken in cloud computing, still some gaps are identified in studied literature. One of the gaps in cloud security research is the need for in-depth understanding of the growing risks and dangers associated with cloud systems. A key flaw with cloud systems is the lack of transparency and auditing tools. Organizations typically lack visibility into CSP-implemented security controls, which can inhibit effective monitoring and incident response efforts. Effective threat intelligence is critical for implementing proactive security measures. However, there remains a gap in the provision of comprehensive threat intelligence for cloud settings. Research is needed to improve threat intelligence techniques specific to cloud-based threats [23].

2.2 Data storage

Storage is the major concern to use the various applications on the cloud. It oversees managing and carrying out the activities of data storage as a service. The volume of data continues to grow, while the need for online access to information grows as well. It is difficult for the in-house computer servers to manage data that exceeds a certain limit. Cloud storage services address such challenges and enable large-scale data management [24, 25]. Users save their data and files on cloud storage through the internet. Cloud data storage is used to store various documents, videos, pictures, etc. Individuals store their data. Organizations check their mail online, from the cloud only [26]. Various uses of cloud data storage are shown in Figure 5.

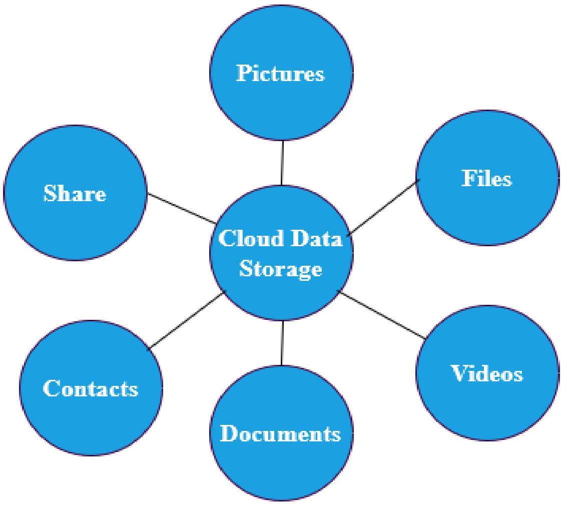

Figure 5. Cloud data storage

The Cloud Service Provider (CSP) hosting the client's data must give data access, and the data cannot be read or updated by an uncertified individual. There are various cloud storage providers are available in the market and also provide the various securitites. Each provider has its market share. Amazon web service (AWS), Microsoft Azure, Google Cloud Platform, Oracle, Alibaba, IBM, Dropbox and many more. Out of these, AWS shares the market 62%, Microsoft Azure 20%, and Google Cloud Platform 12% shares the market. Rest comes in the remaining 6% of the total market [27, 28].

2.3 Cloud storage for data deduplication

Despite the multiple possibilities for data storage, including storage provided by the cloud, data deduplication is one of the major challenges that consumers and organizations face. There is a huge duplicate data on cloud storage which is growing with the growing of data on cloud storage. Data Deduplication is an approach for reducing the multiple copies on cloud storage and saving bandwidth. It stores only a copy of data rather than storing replicas [29].

Deduplication can be achieved in a variety of methods depending on the least data size evaluated by the system for redundancy. Data Deduplication can be done at granularity: file level and block level and location-based: client side and server side as shown in Figure 6. A unique hash value is generated by using a hashing algorithm that hash is an index value where the whole file is stored in file-level deduplication. Here, the whole file is considered as one chunk and the whole file is stored on one index only. It is easy to maintain file-level deduplication due to fewer hash values and comparisons. Also, computation overhead is another benefit of it. In block-level deduplication, the file is divided into smaller pieces called chunks. These chunks are of fixed size and variable length. The file is divided into fixed smaller pieces of data called fixed-size chunks. If the data has been added into fixed-size chunks, then the problem of boundary shifting occurs. This issue is solved by variable length block-level deduplication. The fingerprint value is calculated for each chunk. Each chunk stores on unique index value. The processing overhead is increased in block-level deduplication than in file-level deduplication. Each chunk is compared with the coming chunk and stored only if it is unique otherwise chunk is deleted and the pointer is set to the original one. Hash collision is the limitation of block-level Deduplication due to the generation of the same hash number of two different chunks. Deduplication can also be implemented either at the client side or the target side. Duplicate data is removed at the client side before transmitting the data onto a server which saves transmitting bandwidth and more data can be sent to the server by using less bandwidth. On the server side, data is sent without applying deduplication. The deduplication process takes place at the target side which requires high bandwidth and extra hardware to perform deduplication [6, 30, 31].

2.4 Cloud computing security terms

There are some important terms defined by the International Standards Organization (ISO) for information security. Those also can apply and the same for cloud computing security. To protect data and increase the efficiency of the cloud, it ensures that every user and service provider should know these key terms. Many of the security threats and attacks can be avoided by gaining knowledge of these terms. The following are the key security terms [32]:

- Confidentiality: To protect user privacy, only privileged entities have access to data.
- Integrity: Integrity ensures that no data is altered or updated while being sent or stored and that only those with the necessary authority may add, modify, copy, or remove data.
- Authentication: Authentication is the process of authenticating the user's identity before providing data access. This may be accomplished by verifying the security measures of a user.
- Availability: The capacity to access data and services at any time and from any location.
- Authorisation: To ensure that users who have requested

a certain piece of information have the legal authority to access it, one must get authorization.

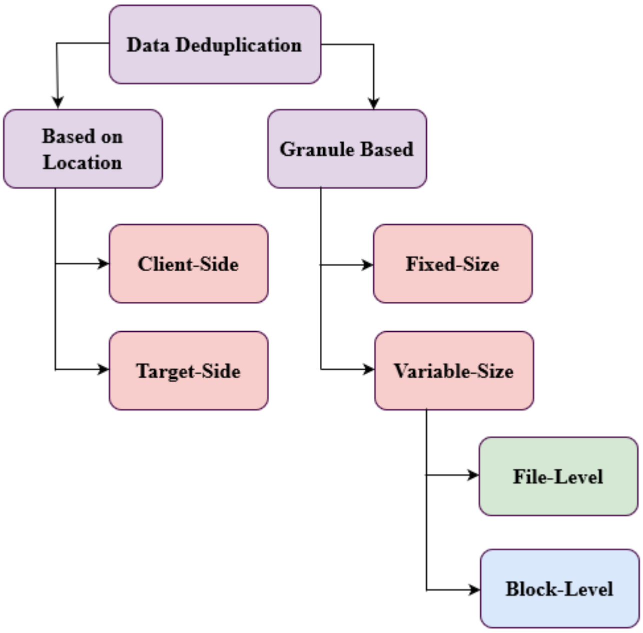

Figure 6. Data deduplication types

3. LITERATURE REVIEW

Convergent encryption is used in conjunction with deduplication technology to address the key problems associated with cloud data storage, which include data privacy, security, confidentiality, and integrity. Traditional encryption does not support deduplication since multiple users will encrypt the same data using different keys, producing distinct decrypted texts for similar content. Convergent encryption is therefore utilized as it offers data secrecy and permits deduplication. It calculates the data copy's cryptographic hash value and utilizes that value as a convergent key for data encryption. Comparable data copies result in equal convergent keys and, consequently, comparable ciphertext, enabling deduplication to be performed by the cloud storage. It is difficult to maintain and distribute so many keys to the growing number of users. Message-locked encryption (MLE) is a version of convergent encryption that came into the picture. The keys are generated from the message itself so, maintaining and disseminating keys is easy.

To solve the re-encryption, two techniques known as Duplicate-less encryption for simple storage (DupLESS) and Dekey are proposed. Authors have also researched and implemented secure deduplication on backup data and the popularity of file basis. Some authors proposed a proof of storage with deduplication (POSD) to increase the integrity and privacy of clients' data. It also integrates a performance-oriented data scheme (POD), proof of retrievability (POR) and proof of ownership (POW). Many techniques are based on POW and MLE mostly. In terms of security, deduplication is still popular in various fields such as crime investigation, based on blockchain, hybrid cloud architecture, Hadoop, and more. Table 3 provides comparable information on various secure deduplication techniques regarding methodologies, confidentiality, integrity, level of deduplication, and type of deduplication and tools used.

Various technologies based on the level of deduplication and type of deduplication are shown through graphs in Figure 7 and Figure 8. The number of different tools used by different techniques by different researchers is shown in Figure 9. These three figures give a clear picture that file-level deduplication with server-side deduplication has been mostly used by researchers. It has also been concluded from existing articles that the Advanced Encryption Standard locks the confidentiality and the Secure Hash Algorithm locks integrity. Figure 10 depicts the various types of security adopted by authors. Table 4 depicts the notations related to Figure 9 and Figure 10.

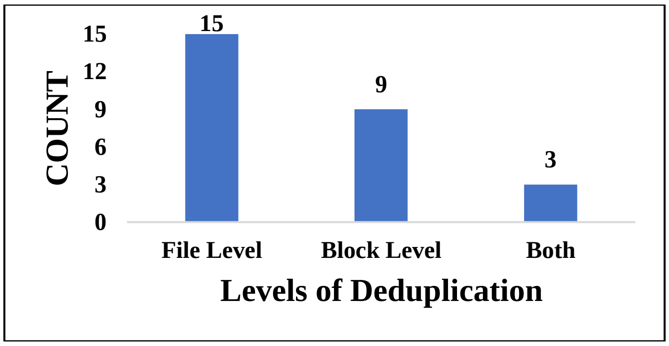

Figure 7. Levels of deduplication

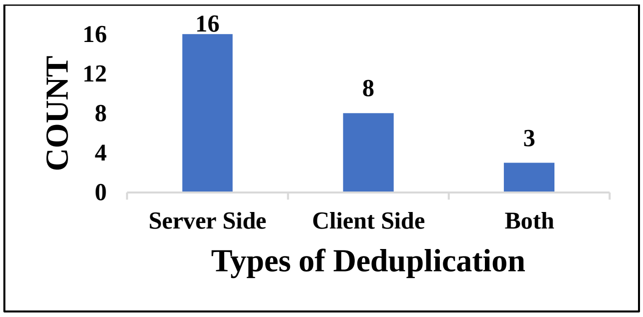

Figure 8. Levels of deduplication

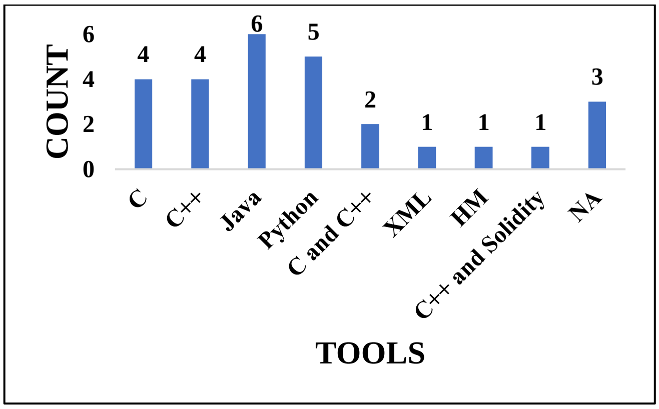

Figure 9. Implementation tools

Table 3. Comparison of various secure deduplication techniques

Tools and References	Integrity/ Confidentiality	Deduplication Level/ Deduplication Type	Methodology	Security Approaches
Java [33]	Yes/Yes	Block Level/ Server Side	• Boneh Goh Nissim Bilinear attribute-based optimal cache Oblivious Control (BGNBA-OCO) has been proposed to improve encryption and enhance the policies regarding access control with reduced communication and computation costs. • After encryption using a matching key data deduplication is checked by the optimal cache oblivious technique. • Then data is transmitted to the cloud server.	Convergent Encryption
NA [34]	Yes/Yes	Block Level/ Server Side	• To decrease the frequency of ownership changes for cloud data, a lazy update method is used. • When data is changed or removed from the cloud, the original ciphertext is not updated instantly. • A flag signals a change in ownership, which prompts re-encryption when a legitimate user requests a download.	Duplicate Less Encryption
NA [35]	Yes/Yes	Block Level/ Server Side	• It is based on data sharing on cloud protocol. • The Clients encrypt the data before uploading the file on CS. • CS is responsible for further convergent encryption and deduplication processes on the cloud. • It can also be useful when ownership is commutated very often. • It reduced the dependency on CS by applying dynamic ownership changes.	Convergent Encryption
Python [36]	Yes/Yes	Block Level/ Server Side	• Three categories of entities are present in our target system: • The data holder, owner of the data and stores it at CSP in numerous blocks. The person who uploads the data blocks to the CSP initially is designated as the data owner. • CSP that offers data holders a deduplication-enabled data storage service. • Authenticated auditors (AA), is third party to verify data ownership, grant access to data, and collaborate with the other two categories.	Convergent Encryption
Java [37]	Yes/Yes	File Level/ Client Side	• Secure data deduplication benefits leverage with users' access control policies. • It uses AES to encrypt and decrypt the stored cloud data. AES provides strong confidentiality and protects the data from unauthorized users. • It uses the MD5 hash function to provide robust integrity. • Based on the hash values of each file, the file has been detected as duplicate or new.	Convergent Encryption
C [38]	Yes/No	Block Level/ Client Side	• An MLE scheme encrypts and decrypts a message using a key obtained from the message itself. • It ensures privacy and integrity of data. • It is stratified into two parts: • Practical: analyze the available techniques with new techniques using random oracle model (ROM). • Theory: determining the standard MLE model that integrates with public key decryption and hashing functions.	Convergent Encryption
C++ [39]	Yes/Yes	File Level/ Server Side	• This technique operates in three phases: • Uploading: clients upload the files on the backup server first then transmit to the remote central server by encrypting the data and calculating the short hash values. • Batch deduplication: similar short hash files are grouped and make a batch of them. Comparison made on short hash values of a group.	Data Backup

C++ [40]	Yes/Yes	File Level/ Server Side	• File retrievable: Only one copy has been kept from a group, rest are discarded. Only the original copy can be accessed by users. • For any discrepancy, the backup server is contacted. • This scheme introduced a fusion of public and private cloud architectures. • Clients store their data on a private cloud as data owners. Re-encryption can be done in a private cloud because it acts as a proxy server. Also, manages the dynamic ownership. • The public cloud manages the storage and the public cloud server can only deliver ciphertext to cloud users that are on the ownership list.	Hybrid Cloud Architecture
Java [41]	Yes/Yes	File Level/ Server Side	• This scheme is divided into three forms: • A user is someone who wants to send private data to a Cloud Service Provider so they can access it whenever required. • Key server generates the encryption key i.e. MLE. It is responsible for disseminating the keys to cloud users. • The CSP offers storage services and deletes superfluous data, storing just one copy. CSP never access plaintext as it is.	Convergent Encryption
C [42]	Yes/Yes	File Level/ Client Side	• The SED enhance the deduplication securely without the use of a trusted key server. • Additionally, it allows for cross-cloud data exchange and updating. • SED addresses the single-point-of-failure issue and enhances the scalability of traditional deduplication schemes. • Inter- and intra-deduplication are simulated and found that improves duplicate detection performance.	Hybrid Cloud Architecture
Python [43]	No/Yes	File Level/ Server Side	• This approach is based on encryption with deduplication. • It uses AES for the encryption process with a hashing function. • It also uses a long string of salt with a hashing function to make security robust and make the process fast. • After encryption, a signature is computed which in turn is used to remove redundant files.	Convergent Encryption
C [44]	Yes/Yes	Block Level/ Server Side	• It recommended a deduplicated integrity of data auditing method for authority changes. • It ensures data integrity with dynamic access controls. • They employed safe ElGamal encryption, identity-based encryption, and Randomized Convergent Encryption.	Convergent Encryption
C++ [45]	Yes/Yes	File Level/ Server and Client Side	• Advocated for data deduplication based on popularity and encryption. • Researchers implemented bi-linear mapping and attribute-based encryption. Third-party assistance is not required for deduplication	File Popularity
Java [46]	Yes/No	File Level/ Server Side	• Ciphertext-policy Attribute-based Encryption with Equality Test (CP-ABEET) includes access policies in the encryption message only. • Only the authorized user who has these attributes set can only be decrypt this message. • The server cannot access or encrypt the message and does not affect the encrypted data deduplication.	Convergent Encryption
C++ [47]	Yes/Yes	Block Level/ Client Side	• The suggested Tunable Encrypted Deduplication method finds duplicate blocks using sketch-based frequency counting. • They employed probabilistic key generation to prevent information leaking in files by encrypting blocks using a collection of candidate keys.	Convergent Encryption
C++ and Solidity [48]	Yes/No	Block Level/ Server Side	• Suggested approach used blockchain-based deduplication to protect the privacy of CCTV video. • Masking settings have been applied to the encryption of actual CCTV video. Synchronization technique used to identify altered parts of CCTV video. This approach prevents sniffer attacks due to the usage of unique keys for each data block.	Blockchain Based

C [49]	N/Yes	File Level and Block Level/ Server and Client Side	• The suggested SM-Tree has an advantage over Merkle-Tree as SM-Tree accepts addition and removal within a certain chunk. • In this, each user holds a set of access privileges and encrypts the data semantically. • TEE does the encryption at the client side. • It enhances security and lessen the overhead cost.	Convergent Encryption
C++ and C [50]	Yes/Yes	File Level/ Client Side	• Fog computing also employed security on deduplication which decreased the overhead involved in communication between mobile consumers and stations. • A chameleon hash function is used to make it more string and undetectable.	Convergent Encryption
Java [51]	Yes/No	File Level/ Client Side	• This method has been implemented on Hadoop, manages and store huge data. • Hash value of file is computed in HBASE and comparison is done on the basis of hash values. • Same hash values file are removed and unique files are supposed to save.	Deduplicate in Hadoop
XML [52]	No/Yes	File Level/ Server Side	• The suggested system will be implemented using cloud-based or centralized data centres. • The proposed technique involves connecting the device carrying possible evidence to a workstation and calculating and comparing the hash value of each artefact to a local database. • This technique reduces duplication in data analysis and simplifies disc reconstruction. This technology offers several benefits and effectively addresses challenges related to digital forensic backlog.	Crime Investigation
Python [53]	Yes/Yes	File Level and Block Level/ Server and Client Side	• Suggested a "keyword search" for data that is already encrypted and kept in cloud storage. This paper presents two secure keyword search algorithms and performs an analysis of security to ensure their secrecy.	Deduplication Based on Keyword Searching
Python [54]	Yes/Yes	File Level/ Server Side	• The proposed design comprises improvements to the existing DupLESS system, which uses blowfish. • It uses a shared key server to generate message-derived keys, enabling deduplication and protecting against external assaults. • The blowfish method significantly decreases processing time	Duplicate Less Encryption
Java [55]	Yes/Yes	File Level/ Server Side	• Authors created a distributed technique that eliminates the requirement for Key Servers. Experiments show that a decentralised encrypted deduplication system can achieve the same level of security as DupLESS.	Duplicate Less Encryption
C++ and C [56]	Yes/Yes	File Level and Block Level/ Client Side	• Authors proposed the Dekey method in which no need to manage a huge number of keys. • Users can share their convergent key with multiple other users safely.	Dekey
Hardware Module [57]	Yes/Yes	Block Level/ Server Side	• It consists of two primary parts: • A Metadata Manager (MM) that aids in block key management in addition to deduplication operations. • A server that will offer an extra degree of protection. • The system is protected against single points of failure by employing these two extra components.	Convergent Encryption
Python [58]	Yes/Yes	File Level/ Server Side	• It uses a message-locked encryption method to avoid the double encryption process. • DupLESS approach is applied for secure deduplication and brute force attack.	Duplicate Less Encryption
NA [59]	Y/No	File Level/ Client Side	• Three participants are involved in the suggested solution: Data is uploaded to the cloud-by-Cloud Client. • Cloud Storage Server: holds deduplicated data and offers customers storage space. • An auditor is a third party that a customer may choose to confirm the accuracy of data uploaded to the cloud. A client can also be used as an auditor if they have duplicate files.	Deduplication Based on Proof of Storage

Table 4. Notations related to Figure 9 and Figure 10

Notations	Full Form
HM	Hardware Module
CE	Convergent Encryption
DupLESS	Duplicate Less Encryption
CI	Crime Investigation
HCA	Hybrid Cloud Architecture
KS	Keyword Searching
NA	Not Applicable

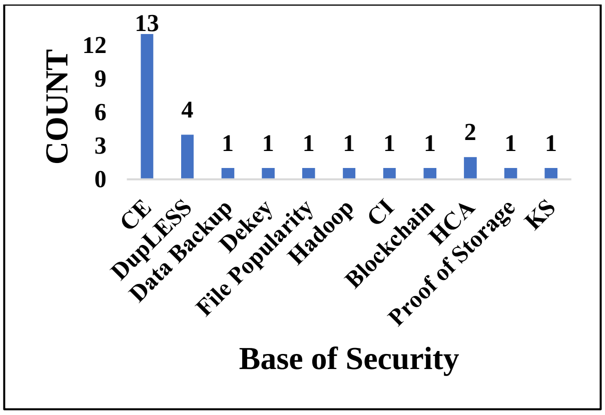

Figure 10. Base of security

4. PROBLEMS IDENTIFIED

The emphasis on centralizing data storage on cloud platforms has prompted research on information security and distributed data storage systems, such as identity-based encryption, Additive Homomorphic Encryption, and file/block-level deduplication. The significance of integrity and deduplication in data stored on cloud storage, including static and dynamic content custody, will be addressed. This research attempts to build an adjustable and transferable integrity design for cloud computing applications involving storage, enabling integrity checks and data deduplication solutions. Some problems are identified during the literature study which can be the focus of future researchers. Upcoming researchers can study and provide related solutions. The following are the problems to be considered while designing and providing the solution:

- Convergent encryption is not fully secure as it leaves the user with privacy issues. An unauthorized user can create a convergent key from the message and decode it. So, the ownership management of data is still an issue [33, 35, 36].
- Peer-to-peer (source and target side) data integrity verification should be provided [36].
- Some schemes do not focus on privacy while developing solutions for integrity and applying symmetric encryption [41].
- The reliability and POW of the multi-cloud architecture should be the focus of further research. Losing and compromising data on multi-cloud architecture is still a challenge [40, 42].
- In some of the schemes, random keys are generated for encryption from the key server only. But if the key server fails, then the keys are also corrupted. The solution to this problem should also be provided [44].
- There is a need to reduce the cost of additional hardware needed for deduplication on the client side and server side. This additional cost should be less. Less communication overhead should be provided between source and target systems [49, 53].
- Every scheme should be compatible with various operations of data such as updation, insertion, deletion etc [54, 55].
- Auditing schemes have less concentration on privacy and confidentiality. A framework with secure auditing schemes needs to be developed [34, 36, 59].
- Data integrity to be ensured on the client side when the client downloads data from the server [39, 59].

5. CONCLUSIONS

Data deduplication can reduce storage costs and save network bandwidth. Netizens exploit the internet facility to access cloud services. A massive amount of data has been stored on the cloud thus, storing multiple copies of the same data occupies a lot of storage space. Cloud storage providers are responsible for providing various services to users and security also. With the increasing amount of data on the cloud, security concerns also need to be addressed. Different attacks exist such as loss of data, leakage of data ownership, social engineering, ransomware attacks, malicious attacks, and more. The monthly number of incidents and data breaches are mentioned for the year 2023 with the top three incidents of each month. The dark beam was the biggest attack in 2023. It is found that most of the attacks are done through email phishing i.e. 81%. Several small, large, private, and public businesses are affected due to these attacks as they did not update their security systems. Most importantly current attack that took place in Jan 2024 has also given. Some authors adopted the file level deduplication and some executed block level deduplication. Better proof of storage output can be achieved by the fusion of file and block-level deduplication. To increase the efficiency of deduplication, it is very important to develop and apply strong security techniques in the cloud to save data from unauthorized users, data can be compromised if multiple copies are there in storage. Also, need to enhance the security techniques at the same time. A security study of cloud-based deduplication solutions is presented and explored. The suggested methods can be integrated with existing cloud storage providers. Existing deduplication systems are classed depending on their design decisions. The security analysis of the methods ensures the integrity and security of client data in the cloud, protecting against both internal and external attacks. Different authors adopted different tools to implement their schemes. The comparison of various secure deduplication techniques is discussed with several parameters: whether integrity and confidentiality are provided or not, implementation tools are also classified, file level or block level, server side or client side, mechanisms mentioned, and security approached. The number of bar graphs is given to show how many authors work on which tool, the level of deduplication, the type of deduplication, and the security approach. The deduplication approaches can be implemented on each source and target side, depending on the needs of the application. Client-side data deduplication can reduce storage server computation, and bandwidth, and save time for small organizations using private cloud storage. Combining security measures such as data privacy, security, confidentiality, and integrity in cloud storage with data deduplication addresses client security concerns and removes unnecessary information from cloud storage, resulting in increased efficiency of storage and decreased network traffic. To ensure data security in cloud storage, practitioners must use strong encryption and decryption technologies. Researchers may concentrate on creating improved encryption algorithms and security measures to combat emerging threats and weaknesses in cloud systems.

REFERENCES

[1] Garg, A., Krishna, C.R. (2014). An improved honey bees life scheduling algorithm for a public cloud. In IEEE International Conference on Contemporary Computing and Informatics (IC3I), Mysore, India, pp. 1140-1147. https://doi.org/10.1109/IC3I.2014.7019783

[2] Sharma, N., Prabha, C. (2021). Computing paradigms: An overview. In Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, pp. 1-6. https://doi.org/10.1109/ASIANCON51346.2021.9545007

[3] Shobana, R., Shalini, K., Leelavathy, S., Sridevi, V. (2016). De-duplication of data in cloud. International Journal of Chemical Sciences, 14(4): 2933-2938.

[4] Gupta, D., Gupta, K., Kumar, N. (2019). Emerging Technologies and Trends in Cloud Computing. An International Journal of Advanced Computer Technology, 8(6): 3146-3149. https://ijact.in/index.php/j/article/view/494

[5] 7 most common cloud computing challenges. https://www.geeksforgeeks.org/7-most-common-cloud-computing-challenges/, accessed on Oct. 22, 2023.

[6] Prajapati, P., Shas, p. (2022). A review on secure data deduplication: Cloud storage security issue. Journal of King Saud University - Computer and Information Sciences, 34(7): 3996-4007. https://doi.org/10.1016/j.jksuci.2020.10.021

[7] Reinsel, D., Gantz, J., Rydning, J. (2017). Data age 2025: The evolution of data to life-critical don't focus on big data. Framingham: IDC Analyze the Future, IDC White Paper, 1-25.

[8] Sharma, N., Prabha, C., Goyal, S.B. (2021). Resource allocation in FC environment: A review. AIP Conference Proceedings, 2555(1): 050024. https://doi.org/10.1063/5.0124592

[9] Muskan, Prabha, C., Singh, G., Singh, J. (2022). Data Visualization and its key fundamentals: A comprehensive survey. In 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, pp. 1710-1714. https://doi.org/10.1109/ICCES54183.2022.9835803

[10] Singh, J., Prabha, C., Singh, G., Muskan, Verma, A. (2023). Addressing role of ambient intelligence and pervasive computing in today's world. Proceedings of International Conference on Recent Innovations in Computing, Springer, Singapore, vol 1011, pp. 257-269. https://doi.org/10.1007/978-981-99-0601-7_20

[11] Austin, D. (2021). Here is your 2021 internet minute infographic!: eDiscovery trends. https://ediscoverytoday.com/2021/04/16/here-is-your-2021-internet-minute-infographic-ediscovery-trends/, accessed on Dec. 14, 2023.

[12] Marino, S. (2023). What happens in an internet minute: 90+ fascinating online stats. https://localiq.com/blog/what-happens-in-an-internet-minute/, accessed on Dec. 14, 2023.

[13] Austin, D. (2023). 2023 internet minute infographic, by eDiscovery today and LTMG!: eDiscovery Trends. https://ediscoverytoday.com/2023/04/20/2023-internet-minute-infographic-by-ediscovery-today-and-ltmg-ediscovery-trends/, accessed on Dec. 14, 2023.

[14] Hashizume, K., Rosado, D.G., Fernández-Medina, E., Fernandez, E.B. (2013). An analysis of security issues for cloud computing. Journal of Internet Services and Applications, 4: 5. https://doi.org/10.1186/1869-0238-4-5

[15] Tyagi, R., Bagchi, S., Kaur, G., Sharma, N. Prabha, C. Khan, M.N.A. (2023). Cyber security architecture for safe data storage and retrieval for smart city applications. In International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, pp. 580-585. https://doi.org/10.1109/CISES58720.2023.10183581

[16] Prabha, C., Sharma, N., Singh, J., Sharma, A., Mittal, A. (2023). A review of cyber security in cryptography: Services, attacks, and key approach. In Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, pp. 1300-1306. https://doi.org/10.1109/ICAIS56108.2023.10073747

[17] 2023 data breach investigations report. https://www.verizon.com/business/resources/T6e/reports/2023-data-breach-investigations-report-dbir.pdf, accessed on Dec. 20, 2023.

[18] Ford, N. (2023). List of data breaches and cyber attacks in 2023 – 8,214,886,660 records breached. https://www.itgovernance.co.uk/blog/list-of-data-breaches-and-cyber-attacks-in-2023, accessed on Dec. 22, 2023.

[19] Aenugu, J. (2023). The 2024 cyber security trends report for the UK. https://techforce.co.uk/blog/2024/the-2024-uk-cyber-security-trends-report, accessed on Dec. 28, 2023.

[20] Drapkin, A. (2024). Data breaches that have happened in 2022, 2023 and 2024 so far. https://tech.co/news/data-breaches-updated-list, accessed on Jan. 5, 2024.

[21] Peng, L., Yan, Z., Liang, X.Q., Yu, X.X. (2023). SecDedup: Secure data deduplication with dynamic auditing in the cloud. Information Sciences, 644: 119279. https://doi.org/10.1016/j.ins.2023.119279

[22] Arun Kumar, M., Ashok Kumar, K. (2022). A survey on cloud computing security threats, attacks and countermeasures: A review. International Journal of Human Computations & Intelligence, 1(3): 13-18. https://milestoneresearch.in/JOURNALS/index.php/IJHCI/article/view/34.

[23] Song, M., Hua, Z., Zheng, Y., Huang, H., Jia, X. (2023). Blockchain-based deduplication and integrity auditing over encrypted cloud storage. IEEE Transactions on Dependable and Secure Computing, 20(6): 4928-4945. https://doi.org/10.1109/TDSC.2023.3237221

[24] Chauhan, M., Shiaeles, S. (2023). An analysis of cloud security frameworks, problems and proposed solutions. Network, 3(3): 422-450. https://doi.org/10.3390/network3030018

[25] 10 best cloud storage services 2023. https://www.geeksforgeeks.org/best-cloud-storage-services/, accessed on Dec. 10, 2023.

[26] Khattar, N., Sidhu, J., Singh, J. (2019). Toward energy-efficient cloud computing: A survey of dynamic power management and heuristics-based optimization techniques. The Journal of Supercomputing, 75: 4750-4810. https://doi.org/10.1007/s11227-019-02764-2

[27] Speiser, K. (2022). How to evaluate cloud service provider security (checklist). https://sonraisecurity.com/blog/how-to-evaluate-cloud-service-provider-security-checklist/, accessed on Jan. 12, 2024.

[28] Upadhayay, A., Sharma, A., Agrawal, C. (2018). Different secure data deduplication approaches for cloud storage: A review. International Journal of Advanced Research in Computer Science, 9(3): 46-51. http://doi.org/10.26483/ijarcs.v9i3.6006

[29] Goel, A., Prabha, C. (2023). A detailed review of data deduplication approaches in the cloud and key challenges. In 4th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, pp. 1771-1779. http://doi.org/10.1109/ICOSEC58147.2023.10276004

[30] Joice, S.A., Mohamed, M.A.M. (2019). Cloud storage: A review on secure deduplication and issues. Journal of Internet Technology, 20(3): 861-876. http://doi.org/10.3966/160792642019052003019

[31] Verma, K., Bhardwaj, S., Arya, R., Bhushan, M., Kumar, A., Samant, P. (2019). Latest tools for data mining and machine learning. International Journal of Innovative Technology and Exploring Engineering, 8(9S): 18-23. http://doi.org/10.35940/ijitee.I1003.0789S19

[32] Amara, N., Qui, H.Z., Ali, A. (2017). Cloud computing security threats and attacks with their mitigation techniques. In International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Nanjing, China, pp. 244-251. http://doi.org/10.1109/CyberC.2017.37

[33] Pavithra, M., Prakash, M., Vennila, V. (2024). BGNBA-OCO based privacy preserving attribute based access control with data duplication for secure storage in cloud. Journal of Cloud Computing, 13: 8. https://doi.org/10.1186/s13677-023-00544-1

[34] Wang, M., Xu, L., Hao, R., Yang, M. (2023). Secure auditing and deduplication with efficient ownership management for cloud storage. Journal of Systems Architecture, 142: 102953. https://doi.org/10.1016/j.sysarc.2023.102953

[35] Lee, M., Seo, M. (2023). Secure and efficient deduplication for cloud storage with dynamic ownership management. Applied Sciences, 13(24): 13270. https://doi.org/10.3390/app132413270

[36] Yu, X.X., Bai, H., Yan, Z., Zhang, R. (2023). VeriDedup: A verifiable cloud datadeduplication scheme with integrity and duplication proof. IEEE Transactions on Dependable and Secure Computing, 20(1): 680-694. https://doi.org/10.1109/TDSC.2022.3141521

[37] Solanke, V., Tambe, T., Melkunde, S., Litake, P.D.S.R. (2023). Secure deduplication with user-defined access control in cloud storage. International Journal for Research in Applied Science & Engineering Technology, 11(6): 394-400. https://doi.org/10.22214/ijraset.2023.53650

[38] Bellare, M., keelveedhi, S., Ristenpart, T. (2023). Message-locked encryption and secure deduplication. In Annual International Conference on the Theory and Applications of Cryptographic Techniques, vol 7881. Springer, Berlin, Heidelberg, pp. 296-312. https://doi.org/10.1007/978-3-642-38348-9_18

[39] Zheng, H.Y., Zeng, S.K., Li, H.W.,Li, Z.J. (2024). Secure batch deduplication without dual servers in backup system. In IEEE Transactions on Dependable and Secure Computing, pp. 1-13. https://doi.org/10.1109/TDSC.2024.3362796

[40] Ma, X.W., Yang, W.Y., Zhu, Y.S., Bai, Z.Q. (2022). A secure and efficient data deduplication scheme with dynamic ownership management in cloud computing. In IEEE International Performance, Computing, and Communications Conference (IPCCC), Austin, TX, USA, pp. 1-8. https://doi.org/10.1109/IPCCC55026.2022.9894331

[41] Yuan, H.R., Chen, X.F., Li, J., Jiang, T., Wang, J.F., Deng, R.H. (2022). Secure cloud data deduplication with efficient re-encryption. IEEE Transactions on Services Computing, 15(1): 442-456. https://doi.org/10.1109/TSC.2019.2948007

[42] Zhang, D., Le, J.Q., Mu, N.K., Wu, J.H., Liao, X.F. (2023). Secure and efficient data deduplication in JointCloud storage. IEEE Transactions on Cloud Computing, 11(1): 156-167. https://doi.org/10.1109/TCC.2021.3081702

[43] Almrezeq, N., Humanyu, M., Ahmed, A.E.A., Jhanjhi, N.Z. (2021). An enhanced approach to improve the security and performance for deduplication. Turkish Journal of Computer and Mathematics Education, 12(6): 2866-2882. https://doi.org/10.17762/turcomat.v12i6.5797

[44] Bai, J.L., Yu, J., Gao, X. (2020). Secure auditing and deduplication for encrypted cloud data supporting ownership modification. Soft Computing, 24: 12197-12214. https://doi.org/10.1007/s00500-019-04661-5

[45] He, Y., Xian, H., Wang, L., Zhang, S.G. (2021). Secure encrypted data deduplication based on data popularity. Mobile Networks and Applications, 26: 1686-1695. https://doi.org/10.1007/s11036-019-01504-3

[46] Wang, Y.H., Cui, Y.Z., Huang, Q., Li, H.B., Huang, J.Y., Yang, G.M. (2020). Attribute-based equality test over encrypted data without random oracles. IEEE Access, 8: 32891-32903. https://doi.org/10.1109/ACCESS.2020.2973459

[47] Li, J.W., Yang, Z.R., Ren, Y.J., Lee, P.P.C., Zhang, X.S. (2020). Balancing storage efficiency and data confidentiality with tunable encrypted deduplication. In Proceedings of the Fifteenth European Conference on Computer Systems, pp. 1-15. https://doi.org/10.1145/3342195.3387531

[48] Lee, D., Park, N. (2020). Blockchain based privacy preserving multimedia intelligent video surveillance using secure Merkle tree. Multimedia Tools and Applications, 80(19): 34517-34534. https://10.1007/s11042-020-08776-y

[49] Fan, Y.K., Lin, X.D., Liang, W., Tan, G., Nanda, P. (2019). A secure privacy preserving deduplication scheme for cloud computing. Future Generation Computer Systems, 101: 127-135. https://doi.org/10.1016/j.future.2019.04.046

[50] Ni, J.B., Zhang, K., Yu, Y., Lin, X.D., Shen, X.S. (20). Providing task allocation and secure deduplication for mobile crowdsensing via fog computing. IEEE Transactions on Dependable and Secure Computing,

17(3): 581-594. https://doi.org/10.1109/TDSC.2018.2791432

[51] Prajapati, P., Shah, P., Ganatra, A., Patel, S. (2017). Efficient cross user client side data deduplication in Hadoop. Journal of Computers, 12(4): 362-370. https://doi.org/10.17706/jcp.12.4.362-370

[52] Scanlon, M. (2016). Battling the digital forensic backlog through data deduplication. In 2016 Sixth International Conference on Innovative Computing Technology, Dublin, Ireland, 10-14. https://doi.org/10.1109/INTECH.2016.7845139

[53] Li, J., Chen, X.F., Xhafa, F., Barolli, L. (2015). Secure deduplication storage systems supporting keyword search. Journal of Computer and System Sciences, 81(8): 1532-1541. https://doi.org/10.1016/j.jcss.2014.12.026

[54] Prajapati, P., Shah, P. (2014). Efficient cross user data deduplication in remote data storage. In International Conference for Convergence of Technology, International Conference for Convergence for Technology, Pune, India, pp. 1-5. https://doi.org/10.1109/I2CT.2014.7092019

[55] Duan, Y. (2014). Distributed key generation for encrypted deduplication: Achieving the strongest privacy. In Proceedings of the 6th edition of the ACM Workshop on Cloud Computing Security, pp. 57-68. https://doi.org/10.1145/2664168.2664169

[56] Li, J., Chen, X.F., Li, M.Q., Li, J.W., Lee, P.P.C., Lou, W.J. (2014). Secure deduplication with efficient and reliable convergent key management. IEEE Transactions On Parallel And Distributed Systems, 25(6): 1615-1625. https://doi.org/10.1109/TPDS.2013.284

[57] Puzio, P., Molva, R., Onen, M., Loureiro, S. (2013). ClouDedup: Secure deduplication with encrypted data for cloud storage. In IEEE 5th International Conference on Cloud Computing Technology and Science, Bristol, UK, pp. 363-370. https://doi.org/10.1109/CloudCom.2013.54

[58] Bellare, M., Keelveedhi, S., Ristenpart, T. (2013). DupLESS: Server-aided encryption for deduplicated storage. In SEC'13: Proceedings of the 22nd USENIX conference on Security, Washington, D.C., USA, pp. 179-194.

[59] Zheng, Q.J., Xu, S.H. (2012). Secure and efficient proof of storage with deduplication. In CODASPY '12: Proceedings of the second ACM conference on Data and Application Security and Privacy, pp. 1-12. https://doi.org/10.1145/2133601.2133603