Pegem Journal of Education and Instruction, Vol. 13, No. 4 , 2023 (pp. 786-799)

RESEARCH ARTICLE WWW.PEGEGOG.NET

Enhancing Efficiency and Security in Joint Cloud Storage through Data Deduplication

Dr.B.Narendra Kumar[1] Manisha Manikchand[2], Shreya Samboju[3], Ms. Hima Bindu Challa[4]

[1] Professor, Department of Information Technology, Sridevi Women's Engineering College, Hyderabad

swecnarendra@gmail.com

[2, 3, 4] Department of Information Technology, Sridevi Women's Engineering College, Hyderabad

Abstract: Data deduplication is an effective tool for reducing user bandwidth requirements and removing data redundancies in cloud storage. A trustworthy key server (KS) has been the backbone of many earlier methods, but they have their limitations and vulnerabilities, such as information leakage, attack resistance, computational complexity, etc. For example, in a single-point-of-failure scenario, the whole system would cease functioning if the trusted KS were to fail. Within the context of a Joint Cloud storage system that offers worldwide services via cooperation with different clouds, we provide a Secure and Efficient data Deduplication scheme (SED) in this article. Without the trusted KS, SED also allows for dynamic data updates and sharing. Also, unlike the traditional cloud storage solution, SED doesn't rely on a single point of failure. Theoretical evaluations show that our SED guards against attacks like brute-force and collusion with ease, and it also guarantees semantic security in the random oracle model. In addition, SED efficiently removes data duplication with little computing complexity, communication overhead, and storage requirements. The client-side usability is enhanced by SED's efficiency and functionality. Lastly, the results of the comparison demonstrate that our system outperforms the previous methods.

Keywords: Computational complexity, Trustworthy key server, brute-force, Cloud storage

1.Introduction

The term "CLOUD storage" refers to a platform that allows users to "pay as you go" for access to large-scale data storage and services. But there's a lot of duplicate data on the cloud, which is wasting space and making things worse. Data

How to cite this article: Dr.B.Narendra Kumar1 Manisha Manikchand2, Shreya Samboju3, Ms. Hima Bindu Challa4. Enhancing Efficiency and Security in Joint Cloud Storage through Data Deduplication.Pegem Journal of Education and Instruction, Vol. 13, No. 4, 2023, 786-799

Source of support: Nil **Conflicts of Interest:** None.

DOI: 10.48047/pegegog.13.04.85

Received: 12.10.2023

Accepted: 22.11.2023 **Published:** 24.12.2023

deduplication is a powerful tool for finding and removing duplicate records. Then, after that, the data is uploaded and stored in a single copy. So, data deduplication technology may lessen client-side bandwidth needs while simultaneously increasing server-side space usage efficiency. There are a lot of cloud computing services that utilize it right now to make things easier for users and reduce storage needs. Traditional data deduplication schemes and their variations rely on a trusted key server (KS) to provide security in a system that includes users, cloud storage providers (CSPs), and the data itself. The worst part is that these traditional systems could have problems with

"platform lock-in" and single-point-of-failure. There will be no way to deploy data outsourcing procedures or use the cloud storage system if the trustworthy KS fails. To address these concerns, a new cloud computing paradigm has emerged: the Joint-Cloud computing system. Interconnected service providers (CSPs) and their customers make up Joint Cloud's network architecture. Users are free to connect to any of these clouds in order to get computational services, and they work together without the trusted KS. By facilitating multilateral cooperation across different clouds, Joint Cloud is clearly able to meet the needs of worldwide cooperative cloud services and provide effective cross-cloud services. Also, the decentralized system is suitable for its construction. Academics and businesses alike have taken an interest in joint cloud computing. Data breaches that compromise billions of records are all too typical. Consequently, in order to guarantee data secrecy in cloud storage systems, it is common practice to request encryption for the outsourced data. But in the ciphertext domain, it's not easy to find and remove the duplicates that have been made. Since many users encrypting the same plaintext using conventional techniques produce distinct ciphertexts. Convergent encryption (CE) and its variations have been suggested as a means to execute encrypted

deduplication. These methods encrypt data by using keys that are extracted from the data itself. In other words, the technique is susceptible to vulnerability and the secret key is deterministic. In particular, there are a number of security flaws, such as the following:

1) the tag provides the plaintext hash value, making it susceptible to the chosen-plaintext attack; 2) the ciphertext fails to meet the semantic security requirements; 3) the predictable plaintext is not resistant to brute-force attacks; 4) users are burdened with a heavy computational load when trying to protect their data from malicious attackers, etc. Following is an explanation of the limits of the two main types of remedies that have been proposed so far to address the aforementioned problems. By restricting the ways in which data may be accessed and made available in the traditional system paradigm, one solution type enhances the security of data that is outsourced. To clarify, they safeguard the security of outsourced data access and manage encryption keys using key management and access control mechanisms. However, due to the high communication cost and computational complexity caused by exchanging several keys and keeping auxiliary information, these systems are not always the most practical. The other approach involves developing brand-new models of the system with the goal of enhancing its security and usefulness. For deduplication to work, users must communicate with a network of trustworthy key servers (KSs). Users are more likely to be attacked because to the clear rise in the number of contacts between the clientside and the server-side. The single-point-of-failure problem has persisted despite the best efforts of previous methods. In addition, the Joint Cloud system allows users to choose between storing their data in a single cloud or across many clouds. It is not possible to use the traditional deduplication methods on the Joint Cloud platform. Our goal in developing the Joint Cloud system model-based Secure and Efficient Data Deduplication scheme (SED) is to address the aforementioned concerns. Efficiency, practicality, and safety are the guiding principles of our SED's design. Data confidentiality, integrity, and attack prevention should be guaranteed by our SED's secure data outsourcing and deduplication without the trusted
key server. .

2) From a functional standpoint, our SED has to provide for renewable, shared, and accesscontrolled data storage in various CSPs. In other words, data owners have complete control over their outsourced data and may make changes and distribute it from any location

they want, whereas non-owners have very limited access to the ability to change or alter any of the
data. .

3) Our SED should make efficient use of computing resources by implementing tasks like uploading, deduplicating, updating, and sharing.

2.Literature Review

The article "A Survey of Indexing Techniques for Scalable Record Linkage and Duplication" was published in the IEEE Transactions on Knowledge and Data Engineering, volume 23, by P. Christopher. "Record linkage" refers to the practice of comparing and contrasting data from many databases that pertain to the same entity. The term "deduplication" describes how this procedure is used on a single database. The ability of matched data to include information that would be either too expensive to collect or unavailable otherwise is making them more useful in many application areas. One of the most important parts of data cleaning is finding and removing duplicate entries from databases. Duplicates may have a major impact on the results of data processing and mining that follows. One of the main obstacles to record linking and deduplication is the matching process's complexity, which is becoming more of a problem as databases grow in size. Record linking and deduplication indexing algorithms have evolved significantly in the last several years. Their goal is to keep matching quality high while minimizing the amount of record pairings that need to be compared throughout the process by eliminating pairs that are clearly not matching. Twelve different iterations of six different indexing methods are covered in this work. They are tested in an experimental setting with both simulated and actual data sets to determine their scalability, performance, and complexity. Despite the growing interest in methods that facilitate the efficient processing, analysis, and mining of enormous databases, no comprehensive study has been released to date. This is despite the fact that numerous organizations, including businesses, government agencies, and research projects, gather ever-increasing amounts of data. Finding many databases that have information about the same things is a challenge that is becoming more important in many application fields. To enhance data quality or enrich data for more in-depth analysis, it is

sometimes necessary to integrate and mix information from several sources. Clients, consumers, patients, workers, taxpayers, students, and tourists are examples of entities to whom the matching documents often pertain. Improving data quality and integrity, enabling reuse of current data sources for new research, and reducing costs and efforts in data collecting are increasingly prevalent uses for record linking. For instance, matched data might include health policy improvement information that has hitherto been gathered via costly and time-consuming survey approaches [2, 3]. By enriching data used for the identification of suspicious patterns, such as communicable disease outbreaks, linked data may also aid health monitoring systems. For decades, record linking has been a standard procedure for statistical organizations to connect census data for further study [4]. In order to increase data quality, build mailing lists, or match data across enterprises (for joint marketing or e-Commerce initiatives, for instance), many firms utilize record linking and deduplication methods to deduplicate their databases. Record linking is being used more and more by many government agencies to track down individuals who have applied for help more than once, or who work while collecting unemployment benefits, for instance, inside and between taxation offices and social security departments. Fraud and criminal detection, along with national security, are other areas of great importance when it comes to record linking [5]. The capacity to swiftly access information pertaining to a person being investigated or to compare records from different databases is becoming more and more important for security agencies and investigators in the fight against crime and terrorism. Databases containing information on individuals aren't the only ones where the issue of locating records pertaining to the same entity arises. In some cases, it is necessary to match different kinds of entities, such as data pertaining to companies, consumer goods, publications, web pages, search engine results, or genetic sequences. One use of record linkage methods in bioinformatics is the discovery of novel, undiscovered genomic sequences in massive data sets that have similarities with them. When working with search engine results, digital library collections, or automated text indexing systems, it is crucial to eliminate duplicate documents (such as web pages and bibliographic citations) [6, 7].

Locating and comparing consumer goods from various internet sites is another application that is gaining popularity. It becomes difficult to compare products since descriptions are typically somewhat different [8]. When all the databases that need to be connected have unique IDs for entities, the issue of matching entries at the entity level becomes easy: all that's needed is a

simple database join. Unfortunately, not all databases use the same unique IDs, therefore more advanced linking methods are usually necessary. Deterministic, probabilistic, and learning-based methods are the main categories into which these techniques fall [4, 9, 10]. The process of matching records is known by various names in the computer science and database communities. For example, data or field matching is used in this paper

[11]. Other names include data integration [13], data scrubbing or cleaning [14], data cleansing [16], duplicate detection [17], information integration [19], entity resolution [20], reference reconciliation [23], and the merge/purge problem [24]. As a part of ETL (extraction, transformation, and loading) technologies, record linkage is often considered in commercial processing of customer databases and business mailing lists. Overviews of record linking and deduplication strategies and problems have been published in two recent studies

[4].

The main bottlenecks in virtualization environments are limited main memory size and memory interference," the article states (Vol. 2, pages 357–368). Memory deduplication finds duplicate pages and merges them into one copy to decrease memory needs; memory partition eliminates memory interference across virtual machines by assigning distinct colors to each VM based on page color; and finally, memory partition improves performance. To improve virtualization performance while simultaneously reducing memory demand and interference, we provide CMDP, a coordinate memory deduplication and partition technique. To further decrease unnecessary page comparison cost and rapidly locate page sharing possibilities, CMDP utilizes BMD, a lightweight page behavior-based memory deduplication technique. Additionally, VMMP, a memory partition based on virtual machines, is integrated into CMDP to lessen interference between virtual machines. Application, virtual machine, and hypervisor specific page colors are assigned by VMMP based on page color. In addition to effectively improving performance (by about 15.8%), the testing findings demonstrate that CMDP can support more virtual machines simultaneously. Cloud computing is quickly becoming a popular and cost-effective computing paradigm because to its scalability and its ability to provide consumers a simple pay-as-you-go business model. Many major corporations have already made the switch from on-premises servers to Amazon Web

Services' Elastic Compute Cloud (EC2), including Foursquare and Netflix [2]. company income from cloud computing will reach $1.1 trillion by 2015 [4], according to the International Data Corporation (IDC), therefore it's safe to assume that more consumers and organizations will use the cloud to keep or grow their company on a budget. Cloud computing allows for the autonomous operation of several virtual machines (VMs) that are collocated on a single physical server using virtualization technology. This allows for improved security isolation, migration of services, and flexible allocation of resources. A software layer known as a hypervisor (or Virtual Machine Monitor, VMM) manages physical resources (such main memory) in a virtualization system. Its major purpose is to allow several virtual machines to share these resources efficiently with each other [7]. Although virtualization has increased the capacity and independence of memory systems, it has also increased the interference between virtual machines, which is becoming more problematic as the number of virtual machines running on a single physical server continues to rise (in a desktop cloud environment, there can be up to eight virtual machines running on a single physical core). Both the amount of memory and the reduction of interference are two of the main obstacles to improving the server's overall performance, as the need for memory capacity is directly proportional to the rising speed. To decrease memory demands, memory deduplication finds and cuts down on duplicate pages; to improve performance, memory partition splits memory resources across threads or virtual machines to lessen interference. Both methods have shown promising results in enhancing memory function. Running invisibly at the hypervisor layer, Kernel Samepage Merging (KSM) [8]—a memory deduplication solution implemented by the Linux kernel—needs no changes to guest OS systems. At regular intervals, KSM checks the guest VMs' memory pages for duplicate material. The hypervisor may make use of the extra physical pages because all guest VMs with the same pages share a single physical page. This is how KSM effectively lowers the VMs' memory requirements. Memory deduplication may cut memory use by 50% across virtual machines (VMs), according to Difference Engine [9], and by around 40% according to VMware [10]. There are two issues with memory deduplication, despite the fact that it may conserve memory. The first is the prohibitive expense of the overhead associated with page comparison. Second, memory deduplication isn't a panacea for memory contention. Virtual machines (VMs) compete for shared memory, particularly RAM, when they operate in tandem in a virtualization environment. System unfairness and overall performance loss might result from VM being slowed down compared to when it operates alone and completely controls the memory

system. In addition, the interleaving and interference of memory streams at DRAM memory across distinct virtual machines (VMs) eliminates their
original spatial locality and bank level parallelism, significantly lowering system performance [11–14]. The severity of contention and interference will increase in direct proportion to the growth of the server's number of cores and virtual machines (VMs). Divide memory resources across virtual machines and threads using the memory partition, which includes the channels, rank, and bank partitions, and thus reduces interference. To avoid interference from memory-intensive threads, MCP maps them to separate channels and speeds up non-intensive threads. When it comes to physically aggravating intense threads conflict, however, system injustice is much more significant. Bank partitioning separates cores into their own memory banks, separates threads' memory access streams, and therefore eliminates interference [11]. While memory partitioning does eliminate interference, it does nothing to enhance system performance or fairness. To improve virtualization performance while simultaneously reducing memory demand and interference, we provide CMDP, a coordinate memory deduplication and partition technique. With CMDP, you may take use of memory deduplication and memory partitioning at the same time, each with its own set of benefits and drawbacks. As a result, CMDP may streamline operations while lowering interference levels. In order to further decrease interference across virtual machines, CMDP incorporates a memory partition based on virtual machines (VMMP). VMMP dynamically assigns virtual machines (VMs), hypervisors, and the programs that execute on them to separate memory banks. Independent memory banks are used by the hypervisor, various virtual machines (VMs), and programs operating on those VMs in order to eradicate interference. In addition, CMDP quickly detects page sharing possibilities while reducing unnecessary page comparison cost by using BMD, a lightweight page behavior-based memory deduplication technique. Pages in BMD are categorized according to their behavior and the banks to which they belong. Pages stored in virtual machines' memory banks are more likely to contain the same material, and pages that exhibit similar behavior, particularly in terms of access behavior, are also more likely to contain the same content. Thus, pages that originate from virtual machines' memory banks and exhibit comparable behavior are categorized together. As a result, you can only compare pages within the same categorization; doing so will result in several pointless comparisons. To improve the efficiency of our CMDP, BMD may lower the overhead of unnecessary comparisons in this

manner. The following are some of the main points that the article hopes to convey via its CMDP proposal: (1) Improve performance and decrease interference concurrently by coordinating memory deduplication and partition. (2) CMDP makes use of VMMP, which configures virtual machines (VMs), hypervisors, and the programs that execute on them to use separate memory banks. Independent memory banks are used by the hypervisor, various virtual machines (VMs), and applications running on those VMs in order to prevent interference (3). This is due to the fact that pages belonging to different VMs' memory banks are more likely to contain the same content, and pages with similar behavior, particularly access behavior, are also more likely to contain the same content. Consequently, pages belonging to different VMs' memory banks and those with similar behavior are grouped together for the purpose of reducing the comparison range. In order to cut down on unnecessary comparison cost, our

suggested BMD limits page comparisons to the same categorization. . WITHIN "AND Y. "The Design of Fast Content-Defined Chunking for Data Duplication Base Storage Systems," EIE Transactions on Parallel and Distributed Systems, Vol. For data deduplication systems, Content-Defined Chunking's (CDC) strong redundancy detection ability has been crucial as of late (Vol. 31, No. 9, Pages 2017–2031, 2020). Since the chunk cut-points are declared by calculating and evaluating the rolling hashes of the data stream byte by byte, current CDCbased methods impose significant CPU overhead. This article presents FastCDC, a ContentDefined Chunking method for data deduplication-based storage systems that is both fast and efficient. The core concept of FastCDC is a combination of five important techniques: a gearbased fast rolling hash, an improved and simplified Gear hash judgment, a method to skip sub-minimum chunk cut-points, a method to normalize the chunk-size distribution in a small specified region to fix the decreased deduplication ratio caused by the cut-point skipping, and finally, rolling two bytes each time to speed up CDC even more. Using all five techniques, our evaluation results demonstrate that FastCDC outperforms state-of-the-art CDC approaches by a factor of 3-12, while achieving a deduplication ratio comparable to or higher than the classic Rabin-based CDC. Furthermore, our study on the deduplication throughput of FastCDC-based Destor (an open source deduplication project) shows that FastCDC contributes to a throughput that is 1.2-3.0X higher than Destor based on state-of-the-art chunkers. As the amount of digital data continues to rise at an exponential rate, large-scale storage systems are starting to pay more and

more attention to data deduplication as a viable solution for data reduction. It finds duplicate contents using their cryptographically secure hash signatures (e.g., SHA1 fingerprint) and removes unnecessary material at the file or chunk level. Deduplication research by Microsoft [1], [2] and EMC [3], [4] indicates that deduplication technology may eliminate almost 85% of the redundant data in EMC's secondary storage system and half of the data in Microsoft's main storage system, respectively. Because it finds and eliminates redundancy at a finer resolution, chunk-level deduplication is often more common than file-level deduplication. The most basic method of chunking for chunk-level deduplication is Fixed-Size Chunking (FSC), which entails dividing the file or data stream into equal, fixed-size pieces [5]. To fix the boundary-shift issue that the FSC method has, methods based on Content-Defined Chunking (CDC) have been suggested [6]. In contrast to FSC, which uses the byte offset to designate chunk boundaries, CDC uses the bytes themselves to determine whether data is redundant, making it easier to identify instances of deduplication. Recent research[1,2,7,8] indicates that CDC-based deduplication methods may find an additional 10–20% redundancy compared to the FSC method. At the moment, Rabin-based CDC [6, 9, 10] is the most widely used CDC method for determining chunk boundaries. This method uses the content's Rabin fingerprints. Because it calculates and analyzes (against a condition value) the data stream's Rabin fingerprints byte by byte, Rabin-based CDC is both extremely effective and time-consuming in duplication detection [11]. Other hash algorithms, including SampeByte [12],etc., have been suggested to substitute the Rabin algorithm for CDC in an effort to expedite the process. At the same time, CDC acceleration has made use of the numerous computational capabilities provided by GPU processors or multi and many core processors .

3.Problem Statement:

Data deduplication is a powerful tool for improving cloud computing storage and bandwidth by effectively removing duplicate data. Though often used, data encryption makes deduplication more difficult since several ciphertexts might be used for the same plaintext, making it harder to discover duplicates. By putting security in the hands of a single, trusted Key Server (KS), traditional methods expose themselves to the risks of platform lock-in and single-point failure. Data procedures are put at risk when cloud storage systems come to a standstill because to a KS failure. Tackling these difficulties with encrypted deduplication calls

for creative solutions that provide security even in the absence of a trustworthy key server (KS). This improves the user experience with cloud storage services by increasing data security and system resilience, which are critical in reducing the risks of data breaches.

3.Existing System

To safeguard outsourced data from unscrupulous or untrustworthy CSPs, convergent encryption is a key component of data security in deduplication. When it came to messagelocked encryption (MLE), Bellare et al. established a basic technique. Following this, a number of variations were suggested drawing on Bellare's research. The problem with these MLE-based methods is that they are vulnerable to a lot of threats as the encryption keys are extracted from the files. Based on the non-degenerate efficiently computable bilinear map, Abadi et al. developed a deterministic encrypted system and a complete randomized strategy for constrained message distributions. A method for dependable key management in deduplication was proposed by Li et al. Then, using the randomized tag, Jiang et al. demonstrated a safe deduplication strategy. The needs for data updates, however, were left out. To accomplish deduplication, each client must keep the keys situated on the path of a binary tree including all keys; later on, Hur et al. examined dynamic ownership management in the context of safe deduplication. The authors Li et al. have out a method for safe data duplication using secret sharing systems for key management. Subsequently, a decentralized server-aided encryption for deduplication was devised by Shin et al. However, it required several user-KS interactions, which provided attackers with chances to get valuable information from the conversation. A secure method of multi-server-aided data deduplication was suggested by Miao et al. Xia et al. developed a content-defined chunking method that is both fast and efficient in order to accomplish fine-grained data deduplication. In order to save space and cut down on layer restoration cost, Zhao et al. suggested a deduplication approach that is based on the Docker registry architecture. Beside, academics have lately concentrated on new technologies like blockchain that are used for deduplication in a variety of applications; these technologies address frequent security concerns with contemporary deduplication strategies.

4.Disadvantages

Intra-deduplication does not exist; it only takes into account the fact that the data owner has outsourced their data to the same KS. When compared to the current system, the backup

system is more efficient. Inter-deduplication techniques that take into account data outsourced by several data owners via numerous KSs do not yet exist.

5.Proposed System

Without the trusted KS in the Joint Cloud storage system, we provide an efficient and safe data deduplication approach called SED in this study. The completely randomized tag generation technique, which aids in duplication identification and safeguards the outsourced data from collusion assaults, served as an inspiration for several of the sub-algorithms in our SED. In contrast to other deduplication techniques, our SED guarantees that both the ciphertext and the tag may meet semantic security requirements. The tag and ciphertext are completely indecipherable by any opponent. Not only that, our SED is the pioneering system that enables safe data transfer and updates. The encryption method that enables data deduplication, updates, and sharing is designed in our SED. If the data owner decides to let authorized people access their outsourced data, SED is the first system that we are aware of that takes this into account. In particular, the participating CSPs work together to establish a master encryption key. The security and adaptability of key generation are guaranteed. Data updating and sharing processes are made easier by the authentication-based data access control of SED. To further enhance data deduplication performance, SED integrates intra-and inter-deduplication methods to eradicate duplicates in the Joint Cloud system. After that, theoretical evaluations show that SED performs better than competing systems in terms of functionality, attack and collusion resistance, data secrecy, and data integrity. Implementing and simulating SED in Ubuntu using the Crypto++, GNU, and PBC libraries allows for experimental complexity evaluation. According to the test findings, SED has a little computational overhead and works well.

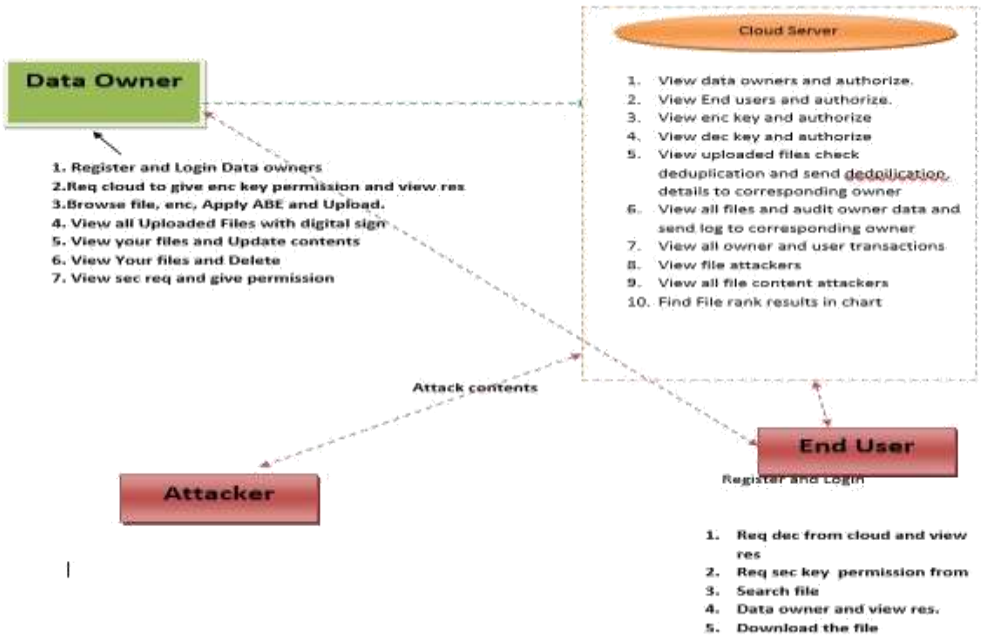
6.Advantages

- The suggested SED's tag generation method and encryption technique provide semantic security. Also, common assaults like brute-force, manipulation, and collusion are no match for SED.

- Secure deduplication is implemented by the SED independently of the trusted key server. It also facilitates the exchange and updating of data across different clouds. In addition to making the standard deduplication approach more scalable, SED eliminates the single-point-of-failure problem.

We validate that our SED is efficient and has a minimal computational cost by conducting trials of the SED method using cryptography libraries. Specifically, we model intra- and inter-deduplication, and our findings show that these techniques may boost the effectiveness of duplicates identification.

7.System Architecture



8.Modules

The data owner uploads their data to the cloud server in this module. Before storing data in the cloud, the owner ensures its security by encrypting both the file and the index name. The data encryptor may have the ability to remove a particular file. In addition, the data he uploaded to the cloud allow him to observe the transactions.

Users provide their login credentials (username and password) to access the Data User module. Once logged in, users may request search control to the cloud, which will then search for files using index keywords and the search score before downloading them. In

addition to seeing the top k documents, users may see the file search ratio. Online Server

An online data storage service is overseen by a server in the cloud. In order to facilitate secure data sharing with Remote Users, data owners encrypt their files before storing them in the cloud. Data consumers get encrypted data files from the cloud, access the ones they want, and then decrypt them. The data owner and data user are authorized by the cloud server, which also gives the results of user-sent search queries. The interest search model and the customized search model are both shown in this section. Can see who has attacked the file.

9.Conclusion

In order to do data deduplication without relying on the reliable KS, we have developed a safe and effective technique called SED. Based on the CDH issue in the Joint Cloud storage system, the suggested SED has enhanced efficiency by reducing client-side communication and computation overhead. Concise encryption methods and tag generation techniques both meet semantic security and tag consistency (validity and security included), respectively. Additionally, SED resolves the conventional cloud storage system's single-point-of-failure issue and enhances scalability. SED is very resilient to common assaults including bruteforce attacks and coordinated efforts by hostile CSPs and unauthorized users. Additional features that enhance usefulness and usability include dynamic support for data operations in SED, such as deletion, modification, and sharing. To the best of our knowledge, SED is the first scheme that takes into account the scenario when the data owner distributes their outsourced data with

authorized users. Both theoretical and practical evaluations have shown that our SED is safe and requires little in the way of resources to communicate, store, and compute. When compared to the prior method, our SED offers superior functionality, efficiency, and security.

10.References

[1] P. Christen, ―A survey of indexing techniques for scalable record linkage and deduplication,‖ IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 9, pp.

1537–1555, 2012.

[2] G. Jia, G. Han, J. J. P. C. Rodrigues, J. Lloret, andW. Li, ―Coordinate memory deduplication and partition for improving performance in cloud computing,‖ IEEE

Transactions on Cloud Computing, vol. 7, no. 2, pp. 357–368, 2019.

[3] W. Xia, X. Zou, H. Jiang, Y. Zhou, C. Liu, D. Feng, Y. Hua, Y. Hu, and Y. Zhang, ―The design of fast content-defined chunking for data de duplication based storage systems,‖ IEEE

Transactions on Parallel and Distributed Systems, vol. 31, no. 9, pp. 2017–2031, 2020.

[4] J. Li, J. Li, D. Xie, and Z. Cai, ―Secure auditing and deduplicating data in cloud,‖ IEEE Transactions on Computers, vol. 65, no. 8, pp. 2386–2396, 2016.

[5] L. Liu, Y. Zhang, and X. Li, ―Keyd: Secure key-deduplication with identity-based broadcast encryption,‖ IEEE Transactions on Cloud Computing, pp. 1–1, 2018.

[6] J. Ni, K. Zhang, Y. Yu, X. Lin, and X. S. Shen, ―Providing task allocation and secure deduplication for mobile crowdsensing via fog computing,‖ IEEE Transactions on

Dependable and Secure Computing, pp. 1–1, 2018.

[7] Y. Zheng, X. Yuan, X.Wang, J. Jiang, C.Wang, and X. Gui, ―Toward encrypted cloud media center with secure deduplication,‖ IEEE Transactions on Multimedia, vol. 19, no. 2,

pp. 251–265, 2017.

[8] H. Wang, P. Shi, and Y. Zhang, ―Jointcloud: A cross-cloud cooperation architecture for integrated internet service customization,‖ in 2017 IEEE 37th International Conference on

Distributed Computing Systems, 2017, pp. 1846–1855.

[9] K. Huang, X. Zhang, Y. Mu, F. Rezaeibagha, X. Wang, J. Li, Q. Xia, and J. Qin, ―Eva: Efficient versatile auditing scheme for iot-based datamarket in jointcloud,‖ IEEE Internet of Things Journal, vol. 7, no. 2, pp. 882–892, 2020.

[10] M. Bellare, S. Keelveedhi, and T. Ristenpart, ―Message-locked encryption and secure deduplication,‖ in International Conference on the Theory and Applications of Cryptographic Techniques, 2013, pp. 296–312

[11] Y. Tang, P. P. Lee, J. C. Lui and R. Perlman, "Secure overlay cloud storage with access control and assured deletion", *IEEE Trans. Dependable Secure Comput.*, vol. 9, no. 6, pp. 903-916, Nov.-Dec. 2012.

[12] M. Bellare, S. Keelveedhi and T. Ristenpart, "DupLESS: server aided encryption for deduplicated storage", *Proc. 22nd USENIX Conf. Secur.*, pp. 179-194, 2013.