IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005267 |

Efficient Data Deduplication and Secure Auditing Data in Cloud Computing

Miss.Himani Mulay, Prof.Sareeka G Joshi

Department of Computer Engineering, Vishwabharati Academy's College of Engineering, Ahmednagar, India

ABSTRACT: Cloud storage is a critical component of cloud computing since it allows users to store and share their data with other authorised users. As a way to reduce data redundancy while still maintaining encryption, secure data deduplication has been extensively studied in cloud storage. For the most secure com-munications, server-assisted encryption approaches have been proposed, but the cost of maintaining a key server is prohibitive (KS). As opposed to the previous systems, which all use a single secret key, the new approach assumes that there is only one KS in the system. It's not just the effectiveness of deduplication that is affected by these methodologies, but also the scalability of cloud storage as the number of users grows. There are various KSs in this work, which expands server-aided encryption. A cloud storage service provider can deduplicate ciphertexts from various KSs within the same tenant or between tenants using our proposed inter-KS deduplication mechanism. KS management freedom and cross-tenant deduplication of encrypted data are therefore provided by our system. The decentralised nature of the concept avoids the need for centralised entities to coordinate or preshare secrets across KSs. So cloud storage services can deliver great deduplication efficiency and scalability while also keeping good data security. There are a lot of experiments done to show the performance study's findings on the proposed scheme work. In addition, we found that the suggested technique has all of the desired security characteristics in our research.

KEYWORDS: Cloud storage, data deduplication, Server-aided encryption, Message-locked encryption,Secure deduplication, ac-cess control.

I. INTRODUCTION

On recent years, the amount of data that has been stored in the cloud has been steadily growing. This is mostly because to the extensive usage of data outsourcing. Cloud storages make advantage of cross-user client-side data deduplication so that they can be more cost-effective and reduce the amount of bandwidth they utilise. Because of the rapid expansion of digital data, the practise of data deduplication, which is an efficient and inexpensive solution to the problem of data reduction, has gained popularity in large-scale storage systems. At the file or chunk level, it identifies and eliminates unneeded data by employing cryptographically secure hash signatures. It does this by recognising and removing duplicate information. It's possible that cloud service providers may store a significant number of redundant data, which will eventually lead to a huge quantity of redundant storage and backup space, in addition to a large amount of computational and administration overhead during the course of the cloud service's life cycle. In order to solve this problem, we first devised a method known as data deduplication. This is a process that eliminates redundant copies of data and stores only one copy of each item, so reducing the amount of redundant storage space and bandwidth. The data deduplication approach and its variants each provide safety by depending on a trustworthy key server (KS) as the backbone of the system, which also includes users and cloud storage providers (CSPs). In the event that the trustworthy KS is compromised, the cloud storage system will become inoperable, and it will be impossible to do any data outsourcing operations. Users and a variety of cloud service providers (CSPs) who offer a wide range of services together make up the Joint Cloud's network architecture. Users are free to connect to any of these clouds in order to access computing services, and the clouds work together without the need for a trusted key. The Joint Cloud has the potential to supply effective cross-cloud services and to fulfil the criteria of globalised cooperative cloud services if multilateral collaboration among various clouds is utilised. It is conceivable that it may be built via a decentralised approach. Joint cloud computing has garnered interest from both the business world and the academic community.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005267 |

As a result of the fact that encryption techniques prevent the CSP from identifying duplicates over encrypted data, data deduplication necessarily conflicts with the process of protecting the secrecy of data that has been outsourced. Convergent encryption is the first step in providing deduplication for data that has already been encrypted. It does this by computing an encryption key, known as a convergent key, from the data itself, which results in the generation of identical ciphertexts from identical plaintexts. Convergent encryption is susceptible to brute-force attacks since the entropy source of a convergent key is solely based on message distribution, which is regarded to be easily anticipated. As a result, convergent encryption is not widely used. Server-aided encryption increases the security of convergent encryption and provides higher confidentiality that is resistant to brute-force attacks. However, the maintenance of a semi-trusted key server is a cost associated with using server-aided encryption (KS). It treats a collection of clients who utilise a cloud storage service as though they were part of an existing company network. In this method, a tenant (for example, a company) instals a dedicated KS on its premises, and each user uses an interactive protocol to obtain a convergent key. This allows the tenant to prevent unauthorised access to the premises. When calculating the convergent key that is the outcome of the interaction, in addition to the data, a secret key that is associated with the KS is also used as an input. A source of entropy is used in conjunction with the randomness in the process of key generation. Convergent keys are created as a result, and as a consequence, they are statistically independent of the message distribution. This improves the safety of encrypted data while preserving the efficiency of deduplication.

A. Motivation

- Data deduplication has been widely deployed in storage systems for space savings, the fingerprint-based deduplication approaches have an inherent drawback: they often fail to detect the similar chunks that are largely identical except for a few modified bytes, because their secure hash digest will be totally different even only one byte of a data chunk was change.
- It becomes a big challenge when applying data dedu-plication to storage datasets and workloads that have frequently modified data, which demands an effective and efficient way to eliminate redundancy among frequently modified and thus similar data.

B. Objectives

- To improved integrity.
- To increase the storage utilization.
- To remove the duplicate copies of data and improve the reliability.
- To improve the security by storing unique blocks on multiple nodes.
- Apply Deduplication on every small block (e.g.4 kb, 8kb, or less) to reduce the bottleneck problem. To maximally detect and eliminate redundancy at very low overheads.

II. HISTORY & BACKGROUND

Gangyong Jia et al: The purpose of this study is to present CMDP, which stands for coordinate memory deduplication and partitioning method. Its purpose is to improve the performance of virtualization by simultaneously lowering the amount of memory requirements and interference. The CMDP is broken up into two distinct sections. To begin, the CMDP's VMMP dynamically maps the hypervisor, VMs, and applications running on VMs onto separate memory banks rather than accessing all of the memory banks. This is done to limit the amount of VM interference that occurs. Their memory requests can be isolated from those of other people, which will allow the spatial locality to be reserved. Second, in order to decrease the number of times memory is requested, the CMDP's BMD splits pages into several classes based on both the banks to which they belong and the behaviour of the pages themselves.

Yukun Zhou et al: In this work, we suggest the use
of FastCDC, a CDC strategy for data deduplication that is significantly faster than current state-of-the-art CDC approaches while yet achieving a comparable deduplication ratio. The fundamental concept behind FastCDC is to combine five important procedures, including gear-based fast rolling hashing, optimal hash judgement for chunking, subminimum chunk cut-point skipping, normalised chunking, and rolling two bytes each time.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005267 |

Ling Liu et al: This article argues that The security study presented in this paper reveals that our KeyD protects user ownership privacy while simultaneously maintaining data confidentiality and convergent key security. The results of the experiments demonstrate that the performance of our method is not negatively impacted by the addition of security. In the future, we will make every effort to discover answers to the problems that prevent our system from protecting the data owners' right to their own privacy and identify.

Jianbing Ni et al: In this research, we present a fog-assisted mobile crowdsensing architecture known as Fo-MCS. This design makes advantage of fog nodes to improve the accuracy of job allocation. We recommended a fog-assisted secure data deduplication technique in order to cut down on the amount of unnecessary communication that takes place between the fog nodes and the CS-server (FoSDD). The Fo-SDD enables fog nodes to detect and delete duplicate data in sensing reports, as well as providing a high level of protection against brute-force and "duplicate-replay" assaults. In addition, the Fo-SDD enables fog nodes to detect and delete duplicate data in sensing reports. We improved the Fo-SDD to conceal the identity of mobile users in order to prevent "duplicate-linking" breaches. This ensures that no adversary may link similar sensing reports to specific mobile users in the case that they are targeted.

Xue Yang et al: In this work, we describe an efficient and safe data deduplication system that also has access control that is user-defined. Our system does not require the addition of an additional authorised server or the utilisation of a hybrid cloud architecture in order to accomplish the approved deduplication that is desired. In our system, no one other than the CSP has the authority to restrict access rights on behalf of data owners without putting the data's secrecy at risk.

Haoran Yuan et al: In this body of work, we present both a Bloom filter-based site selection approach and a safe data deduplication mechanism with efficient re-encryption. The inherent quality of one-way hash functions makes our method impenetrable to stub-reserved attacks and guarantees the data owners' right to the privacy of their sensitive information.
This is made possible by the hashing algorithm.

YOUNGJOO SHIN et al: In this article, we offered a taxonomy of state-of-the-art deduplication algorithms. The taxonomy was based on the environment of the deduplication system as well as the security goals for a number of different threats. We examined the various known safe deduplication algorithms and classified them according to cryptographic and protocol categories, such as message-dependent encryption, proof-of-work, traffic obfuscation, and several alternatives for deterministic information dissemination.

Dongyoung Koo et al: We evaluated the proposed method's security by analysing its ability to prevent collusion, maintain the integrity of the data, and keep data secret. Because it makes use of a KS, the solution that has been described offers the highest level of data secrecy in cloud storage against any users who do not have lawful ownership of the data. This applies to both an honest but curious CSP as well as KS. In addition, the less stringent sense of secrecy is maintained in order to defend against accusations of collaborating or compromising with the KS.

Chia-Mu Yu et al:Both ZEUS and ZEUS+ are based on the architecture of zero-knowledge deduplication response and prevent an attacker from gaining existence status information through duplicate checks. These approaches are proposed as a result of this research. Although ZEUS and ZEUS+ are capable of providing a more secure privacy notion known as two-side privacy, our investigations of real-world datasets demonstrate that ZEUS and ZEUS+ result in a slightly increased number of communications.

Yinjin Fu et al: AppDedupe is an application aware scalable inline distributed deduplication framework for big data management. It makes use of application awareness, data similarity, and proximity to accomplish a trade off between scalable performance and the efficacy of distributed deduplication. In this research, we describe AppDedupe. It makes use of a two-tiered data routing scheme to route data at the super-chunk granularity in order to reduce cross-node data redundancy while maintaining good load balance and low communication overhead. Additionally, it utilises

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005267 |

application-aware similarity index-based optimization in order to improve deduplication efficiency in each node while consuming very little RAM.

III. PROPOSED SYSTEM

The important and popular cloud service is data storage service. Cloud users upload personal or confidential content data to a cloud service provider (CSP) data center and allow to keep this data. In addition, the loss of control over Personal data leads to high data security risks, especially data privacy leaks. Due to the rapid Development of data mining and other analysis technologies, the question of privacy becomes serious. Therefore, a good practice is to encrypted data in order to guarantee the security of the data and the privacy of the user.

In recent time, there are many problems of storage places in cloud. If data holder store file in cloud which is already available in cloud, so this is waste of memory. So, in this paper we remove the deduplication using hashing techniques. Also, security issues are occur during the cloud server data storage. We are solving these problems with the help of encryption technique and hash code techniques.

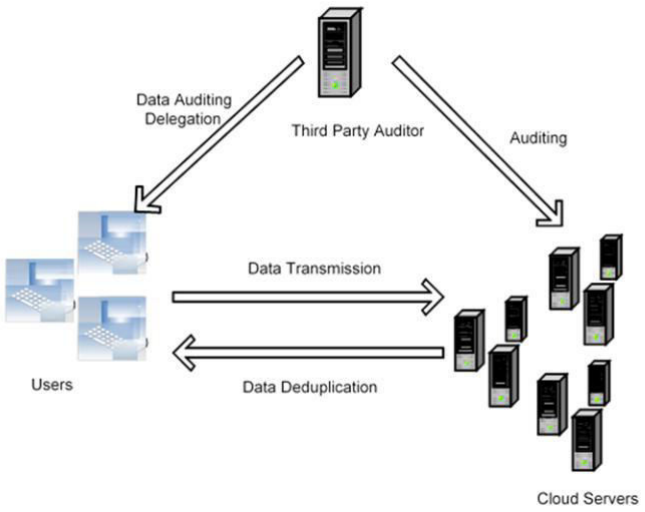

Fig. 1. Proposed System Architecture

A. Algorithms

1. AES Algorithm for Encryption.

AES (advanced encryption standard).It is symmetric algorithm. It used to convert plain text into cipher text .The need for coming with this algo is weakness in DES. The 56 bit key of des is no longer safe against attacks based on exhaustive key searches and 64-bit block also consider asweak.AES was to be used128-bit block with128-bit keys. Rijendeal was founder. In this drop we are using it to encrypt
the data owner file.
Input:
128 bit /192 bit/256 bit input (0, 1) Secret key (128 bit) +plain text (128 bit). Process:
10/12/14-rounds for-128 bit /192 bit/256 bit input
Xor state block (i/p)
Final round:10,12,14
Each round consists: sub byte, shift byte, mix columns, add
round key.
Output:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005267 |

cipher text(128 bit)

2. MD5 (Message-Digest Algorithm)

The MD5 message-digest algorithm is a widely used cryptographic hash function producing a 128-bit (16-byte) hash value, typically expressed in text format as a 32 digit hexadecimal number. MD5 has been utilized in a wide variety of cryptographic applications, and is also commonly used to verify data integrity.

Steps:

1. A message digest algorithm is a hash function that takes a bit sequence of any length and produces a bit sequence of a fixed small length.
2. The output of a message digest is considered as a digital signature of the input data.
3. MD5 is a message digest algorithm producing 128 bits of data.
4. It uses constants derived to trigonometric Sine function.
5. It loops through the original message in blocks of 512 bits, with 4 rounds of operations for each block, and 16 operations in each round.
6. Most modern programming languages provides MD5 algorithm as built-in functions.

B. System Model

We introduce the system model and threat model, as well as some evaluation criteria, to present and analyse our SED scheme completely.

1) System Model

It is an entity that possesses the resources and processing power to supply distributed cloud computing services and protocols for execution. This entity is referred to as a provider of cloud services (CSPs). the user (U) - The user could be any client who has data that they wish to store in the cloud and who also desires data computing services such as updating and sharing. Users might either be individual consumers or commercial enterprises. In addition, there are two distinct types of users that can be distinguished from one another. The owner of the data is the person or entity that possesses the original copy. In particular, the owner is referred to as the ”first uploader,” but users that upload the same material into CSPs in a subsequent time period are referred to as ”subsequent uploaders.”

2)Threat Model

Users are often reliable individuals who are able to regularly challenge clouds to ensure that the data that is outsourced is accurate. CSPs is an honest but curious entity (also known as a semi-trusted entity), which means that while they faithfully adhere to our protocol, they are also very interested in learning as much as they can about the data that is outsourced. CSPs, on the other hand, do not make changes to or delete any stored data without first obtaining the appropriate authority because of the industry’s reputation and accountability. In addition, we are operating under the assumption that the point-to-point communication channels that exist between clouds and users are trustworthy and safe. Threats posed by channels are not the focus of this particular study; nevertheless, they will be given additional attention in subsequent research.

IV. CONCLUSION

We propose SED, a secure and efficient data deduplication method that does not rely on the trusted KS in this research article. SED was developed by us. The proposed SED reduced the amount of client-side communication and computation overhead, which resulted in an increase in efficiency. This solution was based on the CDH problem in the cloud storage system. The straightforward encryption and generation of tags algorithms that it employs satisfy both semantic security and tag consistency requirements. Traditional cloud storage systems have a single point of failure in KS, which is eliminated thanks to SED, which also enhances scalability. SED is well-equipped to protect against common threats such as brute-force attacks and collaboration between rogue CSPs and unauthorised users. Other

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005267 |

examples of these types of threats include. SED is the first system that takes into account the scenario in which a data owner distributes his or her outsourced data with authorised users. This scenario can occur when a data owner uses a third-party service. According to theoretical and experimental evaluations, our SED is safe and has a low complexity level across all three domains (computing, communication, and storage).

Future Scope - This work raises one potentially fascinating issue. The creation of a revocable attribute-based encryption scheme with data integrity that satisfies the requirements of replayable chosen-ciphertext security is one of these options.

REFERENCES

1. G. Jia, G. Han, J. J. P. C. Rodrigues, J. Lloret, and W. 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.
2. 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 deduplication based storage systems," IEEE Transactions on Parallel and Distributed Systems, vol. 31, no. 9, pp. 2017–2031, 2020.
3. L. Liu, Y. Zhang, and X. Li, "Keyd: Secure key-deduplication with identity-based broadcast encryption," IEEE Transactions on Cloud Com-puting, pp. 1–1, 2018.
4. 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
5. X. Yang, R. Lu, J. Shao, X. Tang, and A. Ghorbani, "Achieving efficient secure deduplication with user-defined access control in cloud," IEEE Transactions on Dependable and Secure Computing, 2020.
6. J. H. Yuan, X. Chen, J. Li, T. Jiang, J. Wang, and R. Deng, "Secure cloud data deduplication with efficient re-encryption," IEEE Transactions on Services Computing, pp. 1–1, 2019
7. Y. Shin, D. Koo, and J. Hur, "A survey of secure data deduplication schemes for cloud storage systems," ACM Computing Surveys, vol. 49, no. 4, pp. 74:1–74:38, 2017.
8. Y. Shin, D. Koo, J. Yun, and J. Hur, "Decentralized server-aided encryption for secure deduplication in cloud storage," IEEE Transactions on Services Computing, vol. 13, no. 6, pp. 1021–1033, 2020.

[9] C. Yu, S. P. Gochhayat, M. Conti, and C. Lu, "Privacy aware data deduplication for side channel in cloud storage," IEEE Transactions on Cloud Computing, vol. 8, no. 2, pp. 597–609, 2020

[10] Y. Fu, N. Xiao, H. Jiang, G. Hu, and W. Chen, "Application-aware big data deduplication in cloud environment," IEEE Transactions on Cloud Computing, vol. 7, no. 4, pp. 921–934, 2019.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details