International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 12 Issue 6, June, 2023, Impact Factor: 7.751

Advanced Techniques for Secure and Efficient Cloud Data Deduplication with Re-Encryption

Patnana Sayesu

REVA Academy for Corporate Excellence, REVA University, Bengaluru, India

ABSTRACT

Commercial cloud storage providers have mostly adopted data deduplication technology, which is necessary to handle the exponential rise of data. A variety of secure data deduplication techniques have been developed and implemented in a variety of settings to better safeguard the security of users' sensitive data in the outsourced storage mode. Numerous instruments have been developed to assist dynamic ownership management, and these schemes have attracted a lot of scholarly interest. They are particularly interested in strategies for deduplicating encrypted data that are effective and secure. The re-encryption deduplication storage system is the primary subject of this study, and we demonstrate how the recently developed lightweight rekeying-aware encrypted deduplication scheme (REED) is susceptible to an attack we refer to as the stub-reserved attack. Furthermore, we provide a secure data deduplication method based on the convergent all-or-nothing transform (CAONT) and arbitrary Bloom filter bits. Re-encryption functions nicely with this approach. Our method can withstand a stub-reserved attack and ensure the privacy of the owners' sensitive data thanks to a built-in characteristic of the one-way hash function. Additionally, the CAONT only requires a portion of the package to be re-encrypted, rather than the complete package, which lessens the computational load on the system. Our technique is safe and effective at re-encryption, according to security analysis and experimental findings.

Index Terms—Re-encryption, Data deduplication, User joining, User revocation, Cloud security, Cyber security, Data security.

INTRODUCTION

As cloud storage technology advances swiftly, more and more people and businesses are choosing to outsource their sensitive data to remote cloud service providers on a pay-per-use basis [1], [2], [3], [4], and [5]. The size of the digital world is expected to double every two years, reaching 44 zettabytes (ZB) or 44 trillion gigabytes (GB) in 2020 (more than 5200 GB for every man, woman, and child), according to a study from Internet Data Center (IDC) funded by Dell EMC [6]. Cloud service providers are under a lot of strain due to the rise of data, though. To address it, a simple solution is to demand that cloud service providers continually increase their storage space capacity in order to satisfy users' needs for high-quality storage services. However, cloud service providers might hold a lot of data that is repetitive and voluminous (such movies, music, and genetic data), necessitating a lot of redundant storage and backup space and, as a result, requiring a lot of computing and management overhead over the course of the data's entire life cycle. Bolosky et al.'s early solution to this issue, the data deduplication technique [7], reduces redundant storage space and bandwidth by removing numerous copies and storing only one duplicate of each item. By adopting data deduplication techniques, cloud service providers like Dropbox [8], Google Drive [9], and Memopal [10] have made significant strides in the modern day.

According to studies [11, 12], data deduplication can be used to significantly reduce the amount of genomic data (about 83%) and disk space used by business applications (about 90%). Although data deduplication provides many benefits, there are still some security issues that need to be resolved. Because they might try to extrapolate and analyze data that has been outsourced, cloud service providers in particular are generally seen as being unreliable [13], [14], [15], [16]. Cloud customers typically encrypt information before outsourcing it to cloud service providers to preserve the privacy of their sensitive data. It is challenging to carry out the data deduplication function because various users encrypt the same data using their private keys, producing different ciphertexts from the same data. The first workable technique for ciphertext deduplication and data security was published by Douceur et al. [17]. However, the cloud user encrypts sensitive data using an unchanged convergent key that is derived from the

hash value of the data. It will enable the cloud user whose access has been revoked to access private data by utilizing the reserved convergent key.

In the context of the cloud, user revocation is a big security problem. To exemplify this idea, we use the context of genetic research. Because there are so many genome databases, genomic researchers usually store their data in the cloud [18]. Platforms specifically designed for managing and interpreting genetic data have been made available by Google Genomics [19] and Amazon [20]. However, certain delicate genetic data generated by studies for disease sequencing needs to be secured. For instance, after leaving the genome project, a researcher won't be able to access the datasets for the genome. This problem has been resolved using methods like group key distribution and re-encryption [21], [22], [23]. These systems enable user joining and user revocation by re-encrypting sensitive data using symmetric encryption (such as AES-128 or AES-256) and distributing group keys for group users. The re-encryption technique will still need too much processing overhead even though it encrypts the entire message with a new encryption key to maintain the security of the data. According to William et al.'s findings [24], it is obvious that even a little amount of policy dynamism will result in prohibitive costs for the cryptographic enforcement of access limitations.

REVIEW OF LITERATURE

Introduction

Give a thorough introduction and introduce the idea of anomaly detection systems as a possible way to find and stop security breaches. Online data storage and access are now possible because to the development of cloud computing. Security and cloud computing use are major concerns right now, especially when it comes to storing sensitive data and maintaining vital infrastructure. It is crucial to establish efficient security solutions to reduce these risks because cloud security breaches have increased in frequency in recent years. It has been suggested that anomaly detection systems are a useful tool for locating and reducing security risks in cloud computing fields.

This chapter's goal is to evaluate the research on cloud security measures and anomaly detection systems for cloud storage applications. This chapter analyzes current research qualitatively and outlines the advantages and disadvantages of current techniques to anomaly detection and cloud security. The chapter is divided into four sections: an introduction to cloud computing and its security issues; a discussion of anomaly detection and its use in cloud networks; an examination of cloud security measures in cloud storage applications; and a conclusion that summarizes the findings and suggests directions for further research.

The emergence of cloud computing technologies has altered how people and organizations store, manage, and access data. The quantity and sophistication of cybersecurity threats have increased in tandem with the significantly rising use of cloud storage (Subashini & Kavitha, 2011). As a result, there is a growing demand for efficient security measures to ensure the accessibility, reliability, and privacy of data in cloud networks. A critical feature of cloud security measures has emerged in particular: anomaly detection systems. These technologies are able to recognize and alert users to potentially dangerous patterns of network traffic activity. Examining the current level of research in anomaly detection systems in cloud security and cloud networks quantifies in cloud storage services is the primary objective of this study of the literature. Our goal is to specifically perform a qualitative analysis of the literature to identify recurring themes and patterns in the application of anomaly detection systems and cloud security measures (Ahmad et al., 2022). We intend to gain a comprehensive knowledge of the present opportunities and problems in this sector, as well as possible directions for future study, by examining the available literature.

This paper provides an overview of cloud computing, as well as its security and storage concerns. It also discusses the function of anomaly discovery systems in cloud security, and it reviews several anomaly detection techniques. The use of anomaly detection systems and cloud security measures in cloud networks is examined in this literature review, which also includes a qualitative analysis of the literature to highlight common themes and trends. Different cloud security measures that are frequently used in cloud storage applications are also examined. Overall, this evaluation will broaden people's understanding of cloud security and anomaly detection technologies. In order to improve the security and dependability of cloud storage applications, we want to uncover common themes and trends in the literature. By doing so, we expect to offer insights that help guide the creation of efficient cloud security measures and anomaly detection systems.

METHODOLOGY

Ciphertext-Policy Attribute-Based Encryption

The cryptographic encryption technique known as ciphertext-policy attribute-based encryption (CP-ABE) allows data owners to declare an access policy over user characteristics and encrypt data in accordance with the access policy using the associated public key components [30]. The appropriate ciphertext can only be decoded by the user whose attributes meet the relevant access policy. Each policy in CP-ABE is expressed as an access tree, where each

leaf node represents an attribute of a user property (such as age, gender, department, etc.), and each nonleaf node a Boolean gate (such as AND or OR). Each user is given a private key that is connected to a certain set of characteristics. His private key can only be used to decipher the ciphertext when the user's qualities match those in the access tree. Fig. 2 depicts the CP-ABE access tree. In this inquiry, each property is seen as a distinct cloud user identifier. Each cloud user is given a CP-ABE private key that is connected to their identify. The IDs of all permitted cloud users are linked with the policy of each file by the OR gate, resulting in the creation of an access tree. Any authorized cloud user can therefore get the original message by decrypting the ciphertext.

Bloom filter

The Bloom filter, a straightforward and small data structure for determining if an element belongs to a set, is widely used in real-world applications [31]. An array of bits defined by a bloom filter has the following definitions: Hi: [1, n] [0, 1], I: [1, k], which sets k different hash functions. These hash routines convert each element into a uniform random number between 1 and n. When the array is initialized, all of its bits are set to 0. It calculates k array locations using k hash functions to add an element x to the set and sets all of the hi(x) locations to 1. After reconstructing all of the array places, a user only needs to compute hi(x) to determine whether element x is present in the set S. As a result, we may be sure that the element x is not included in the set S if one of the array locations is 0. Despite the fact that all locations of hi(x) are 1, it is unknown if the element is in the set because hi(x) = hi(y) (i.e., a collision of the hash function) exists for some y = x.

Review of REED Scheme

The REED encrypted deduplication storage system, which is based on the CAONT method, was proposed by Li et al. The CAONT feature makes it computationally difficult to reconstruct the original message without access to the entire ciphertext. As a result, the computational overhead of the system is significantly reduced because only a small piece of the package needs to be reencrypted through the CAONT.

Enhanced encryption of REED scheme:

The authors suggested a modernized strategy to increase the security of the basic plan. Fig. 3 depicts the modified scheme's workflow. Here is a recap of the improved plan. A user must first encrypt a message M using the message-locked encryption (MLE) key KM and the standard MLE procedure to create the ciphertext C1 = E(KM, M), where E() is the encryption function, in order to upload it to the cloud. The cloud user then uses the original CAONT to translate the KM||C1, where "||" stands for concatenation. In contrast to the basic system, the improved method feeds the pseudo-random mask G with the hash key h = H(C1||KM). G(h) = E(h, S), where S is a publicly known block with the same size as C1||KM, is calculated by the cloud user. C2 = (C1||KM) G(h) is the package head. The cloud user first breaks C2 into a number of fixed-sized components, each of which is the same size as h, in order to generate the package tail t.

The cloud user will then join each item with h using the XOR procedure to create the package tail t. Without understanding the entirety of message C2, it is challenging to predict how self-XOR would behave. The cloud user trims the last few bytes (for example, 64 bytes) from (C2, t), leaving the remaining part of the package as the "trimmed package." The data owner uses the file key to re-encrypt the stub package and create the stub′ package to stop the revoked cloud user from being able to recover the original message. The CP-ABE mechanism is used to disseminate the file key. The cloud user for the group can get the file key. To create the stub′ package, the data owner merely needs to re-encrypt the stub package. Finally, the cloud user just sends the cloud service provider the trimmed and stub packages.

The following is the message reconstruction process. The cloud user must first use the file key to decrypt the stub′ package in order to receive the stub package. The cloud user separates C2 into fixed-size chunks and gets h by performing an XOR operation on each piece and t in order to reconstruct (C1, KM) from the trimmed and stub packages. Then, by comparing C1||KM = C2 G(h), the package C1||KM is created, and its integrity is verified by comparing H(C1||KM) and h. Before calculating M, the cloud user determines D(KM, C1), where D() is the decryption function.

System Model

In this research, we suggest an effective re-encryption method for a secure cloud data deduplication strategy. Our approach is intended for businesses or user groups where a significant portion of members choose to outsource their data to a distant cloud service provider. Deduplication on ciphertexts can be done by the cloud service provider, saving a lot of storage space. Three components make up the system used by our method: a cloud user, a key server, and a cloud service provider (CSP). The system model of our strategy is shown in Fig. 1.

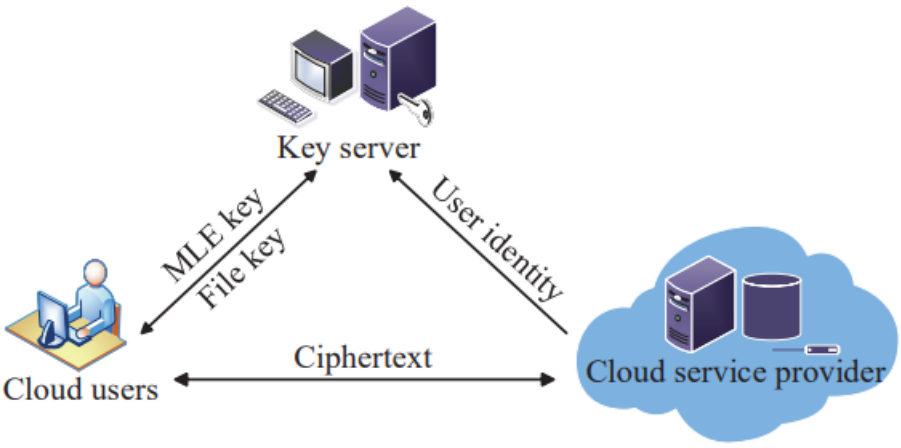

Fig. 1. The cloud storage model.

Key server: The cloud user is helped by a CP-ABE key server when distributing the file key. The MLE encryption key is also generated by the key server. Through an unknown PRF protocol, cloud users can obtain the encryption key from the key server. We presum that the authentication procedure is secure, the communication route is encrypted, and the key server is reliable.

CSP: The CSP is an organization that deals with storage and is in charge of eliminating redundant data and only keeping one duplicate of each item. The CSP is believed to be trustworthy and enquiring. In other words, it will honestly adhere to the protocol but it also requests information on the nature of the stored data. The plaintext shouldn't have access to the CSP. In particular, we make the same assumption made in [32] that the CSP won't save every version of the gathered package and leftover package1. Our plan offers the CSP powerful financial incentives to hold onto the most recent gathered and leftover packages: A bad CSP would have to increase the cost of storage if it kept all the previous iterations of the gathered package and the remaining package.

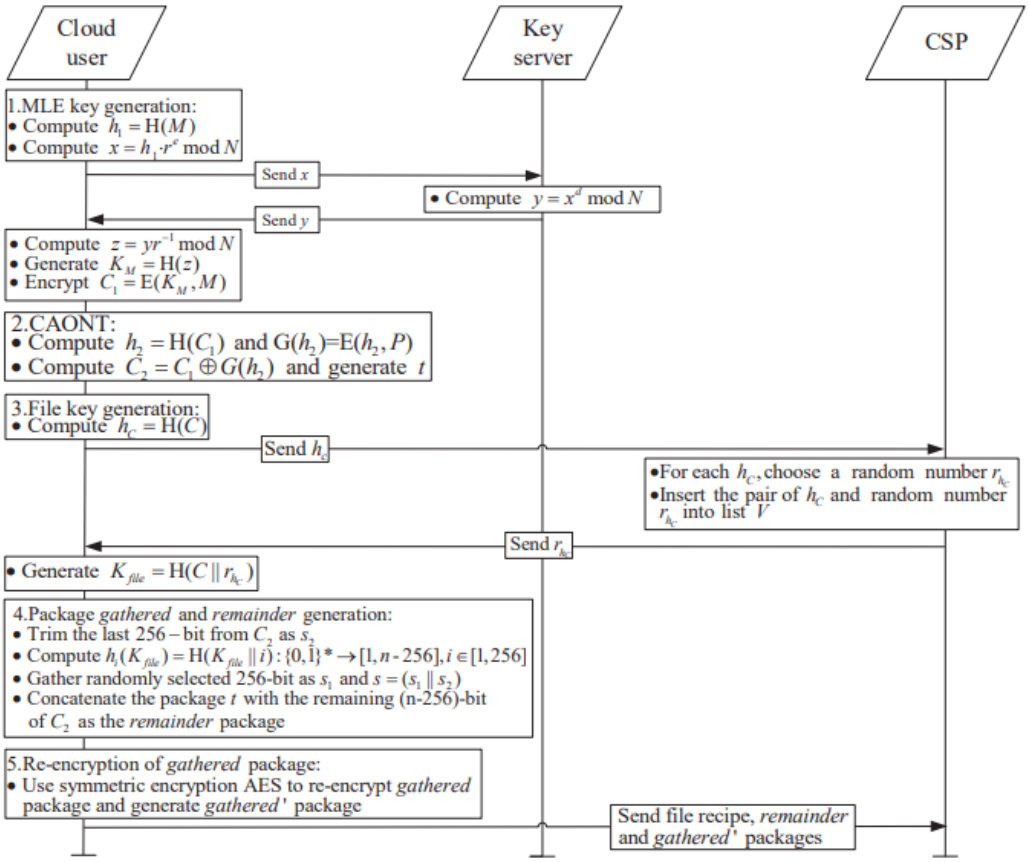

Fig. 2. The procedure of message upload operation.

Security Analysis

We investigate the security of our approach from three perspectives: integrity, secrecy, and stub-reserved attack resistance. We rely on the message-locked encryption scheme, symmetric encryption scheme, convergent allor-nothing transform strategy, bloom filter scheme, and CPABE scheme in order to have secure underpinning tools. The security of our system is ensured by these assumptions.

Resistance to Stub-reserved Attack: After deleting a cloud user from the group, the data owner re-encrypts the packages as follows: 1) The data owner receives the gathered package after it has been re-decrypted. The data owner then restores the packages t and C2 using the acquired and trimmed packages. 2) The file key, K′ file, is selected by the data owner. The data owner then creates the random 256 places of C2 using the new file key K′ f file. 3) To create a new-gathered package, the data owner mixes the newly chosen 256-bit with the last 256-bit of C2. The new-remainder package is then created by joining the remaining piece of C2 with the package t. 4) The package is created and re-encrypted by the data owner for the freshly acquired data. The C2 package is chosen at random for the newly gathered package to prevent the revoked cloud user from predicting which element of the package would be reencrypted. The suspended cloud user must retain all gathered and unopened packets in order to recover the original message. In this case, it is better to adhere to the original message. Therefore, our strategy can thwart the stub-reserved attack.

CONCLUSION

We provide a secure data deduplication method with effective re-encryption in this work, along with a site selection method based on Bloom filters. Due to the intrinsic property of the one-way hash function, our solution is resistant to the stub-reserved attack and protects the privacy of the sensitive information of the data owners. Additionally, by only needing to re-encrypt a tiny portion of the package using the CAONT as opposed to the full package, data owners avoid a substantial amount of calculation overhead. Additionally, we demonstrate the viability of our plan and offer thorough simulation tests. The outcomes of the experiments demonstrate how effective our system is at re-encryption.

REFERENCES

[1]. [1] X. Chen, J. Li, J. Weng, J. Ma, and W. Lou, “Verifiable computation over large database with incremental updates,” IEEE Trans. Computers, vol. 65, no. 10, pp. 3184–3195, 2016.
[2]. [2] M. Gerla, J. Weng, and G. Pau, “Pics-on-wheels: Photo surveillance in the vehicular cloud,” International Conference on Computing, Networking and Communications, pp. 1123–1127, 2013.
[3]. [3] X. Chen, J. Li, J. Ma, Q. Tang, and W. Lou, “New algorithms for secure outsourcing of modular exponentiations,” IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 9, pp. 2386–2396, 2014.
[4]. [4] H. Yuan, X. Chen, T. Jiang, X. Zhang, Z. Yan, and Y. Xiang, “Dedupdum: Secure and scalable data deduplication with dynamic user management,” Inf. Sci., vol. 456, pp. 159–173, 2018.
[5]. [5] H. Huang, X. Chen, Q. Wu, X. Huang, and J. Shen, “Bitcoinbased fair payments for outsourcing computations of fog devices,” Future Generation Comp. Syst., vol. 78, pp. 850–858, 2018.
[6]. [6] IDC. (2014) The digital universe of opportunities : Rich data and the increasing value of the internet of things. [Online]. Available: https://www.emc.com/leadership/digitaluniverse/2014iview/index.htm
[7]. [7] W. J. Bolosky, S. Corbin, D. Goebel, and J. R. Douceur, “Single instance storage in windows 2000,” in Conference on Usenix Windows Systems Symposium, 2000.
[8]. [8] Dropbox. (2007). [Online]. Available: http://www.dropbox.com
[9]. [9] GoogleDrive. (2012). [Online]. Available: http://drive.google.com
[10]. [10] Memopal. (2018). [Online]. Available: http://www.memopal.com
[11]. [11] Netapp. (2008) Netapp deduplication helps duke institute for genome sciences and policy reduce storage requirements for genomic information by 83 percent. [Online]. Available: http://www.netapp.com
[12]. [12] M. Dutch, “Understanding data deduplication ratios,” in SNIA Data Management Forum, 2008, pp. 1–13.
[13]. [13] T. Jiang, X. Chen, J. Li, D. S. Wong, J. Ma, and J. K. Liu, “TIMER: secure and reliable cloud storage against data re-outsourcing,” Information Security Practice and Experience - 10th International Conference, pp. 346–358, 2014.
[14]. [14] X. Chen, B. Lee, and K. Kim, “Receipt-free electronic auction schemes using homomorphic encryption,” Information Security and Cryptology - ICISC 2003, 6th International Conference, Seoul, Korea, November 27-28, 2003, Revised Papers, pp. 259–273, 2003.
[15]. [15] J. Wang, X. Chen, J. Li, K. Kluczniak, and M. Kutylowski, “Trdup: enhancing secure data deduplication with user traceability in cloud computing,” IJWGS, vol. 13, no. 3, pp. 270–289, 2017.