ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 11,November 2021

A NOVEL DEDUPLICATION TECHNIQUE ON ENCRYPTED BIG DATA IN CLOUD

S. Seethalakshmi
Research Scholar,
Manonmaniam Sundaranar University, Tirunelveli.
Tamilnadu, India
Email: seethasri.lakshmi@gmail.com

Dr. B. Balakumar
Assistant Professor,
Manonmaniam Sundaranar University, Tirunelveli.
Tamilnadu, India
Email: balakumarcite@msuniv.ac.in

Abstract — Cloud computing offers customers with apparently limitless "virtual computing" resources as services across the whole Internet, while concealing platform and implementation specifics. Cloud service companies today provide both highly accessible storage and massively parallel computing capabilities at reasonable cost. As cloud computing grows more popular, an increasing quantity of data is kept on the cloud and shared by users with certain privileges that specify the access rights to the stored data. The handling of the ever-increasing volume of data is a major problem for cloud storage providers. De duplication is a well-known strategy for making data management scalable in cloud computing, and it has recently gotten a lot of attention. Data deduplication is a technique for removing duplicate copies of data that has become popular in cloud storage to minimise storage space and upload bandwidth. However, even if a file is held by a large number of people, there is only one copy for each file saved in the cloud. As a result, the deduplication system increases storage use while decreasing dependability. Furthermore, when users outsource sensitive data to the cloud, the issue of privacy emerges. This work is the first attempt to codify the concept of distributed reliable deduplication system in order to handle the aforementioned security problems.

Keywords — Cloud Computing; DeDuplication; Cloud Storage

I. INTRODUCTION

Data deduplication is a specific data dimensionality reduction technique that removes multiple copies of repeated data from storage. The approach is used to enhance storage usage and may also be used to minimise the amount of bytes that must be delivered during network data transfers. Deduplication reduces redundant data by retaining just one physical copy and directing additional redundant data to that copy, as contrast to storing numerous data copies with the same information. Deduplication can occur at either the file or block level. It removes multiple copies of the same file for file level deduplication. Deduplication can also occur at the block level, removing duplicate data blocks that appear in non-identical files. Although data deduplication has many advantages, it raises security and privacy problems since users' sensitive data is vulnerable to both insider and outsider assaults. While standard encryption ensures data secrecy, it is incompatible with data deduplication.

Standard encryption, in general, requires distinct users to encrypt their data with their own keys. As a consequence, identical data copies from various users produce different cypher texts, rendering deduplication impractical.

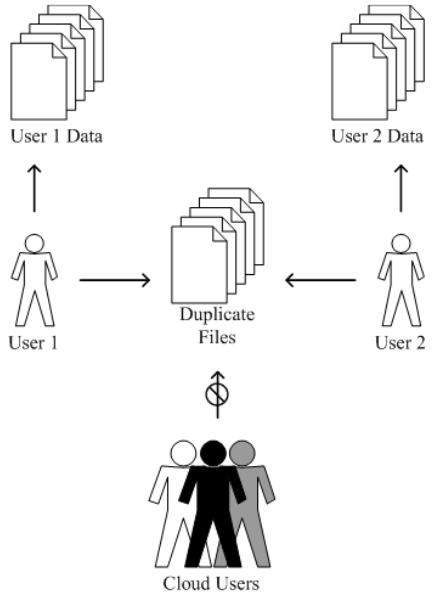


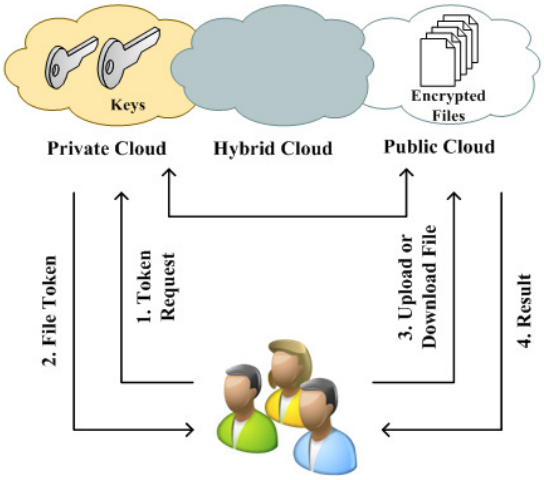
Fig 1.1: Data Depuplication Working Model

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 11,November 2021

Convex encryption has been presented as a means of maintaining data secrecy while allowing for deduplication. It encrypts or decrypts a data copy using a convergent key generated by calculating the cryptographic hash value of the data copy's information. Users keep the keys after key creation and data encryption and submit the cypher text to the cloud. Because the encryption method is Deterministic and derived from the data content, identical data copies will produce the same convergent key and hence the same cypher text.

A secure proof of ownership protocol is also required to offer verification that the user truly owns the same file when a duplicate is discovered in order to prevent unwanted access. Following the proof, following users with the same file will receive a reference from the server, eliminating the requirement to upload the identical file. The encrypted file containing the pointer can be downloaded from the server by a user, but it can only be decrypted by the relevant data owners with their convergent keys.



Thus, convergent encryption enables the cloud to conduct cypher text deduplication, while proof of ownership prevents unwanted users from accessing the file. Earlier deduplication systems, on the other hand, were unable to enable differential authorization duplicate check, which is critical in many scenarios. During the system's startup in such an approved deduplication system, each users are assigned a set of privileges. Every file uploaded to the cloud is further limited by a set of rights that define which types of users are permitted to execute the duplication check and access the files. The user must take this file and his own privileges as inputs before making his duplication check request for any file. If and only if there is a copy of this file and a matched privilege saved in the cloud, the user can discover a duplicate for this file. Employees at a firm, for example, will be given a variety of advantages.

To save cost and handle data more efficiently, the data will be migrated to a Storage Cloud Service Provider (SCSP) in the public cloud with certain permissions, and the deduplication technique will be used to keep only one copy of the same file. Because of privacy concerns, some files will be encrypted and repeated checks will be permitted by staff with specific rights to achieve access control. Traditional de duplication solutions based on convergent encryption, although giving some secrecy, do not provide the duplicate check with differential privileges.

SCOPE

Data deduplication methods are frequently used to backup data and reduce network and storage overhead by finding and deleting data redundancy. In other words, no differential privileges were taken into account in the convergent encryption-based deduplication. If we wish to perform both deduplication and differential authorization duplicate check at the same time, it appears that we are contradicting ourselves.

To save cost and handle data more efficiently, the data will be migrated to a Storage Cloud Service Provider (SCSP) in the public cloud with certain permissions, and the deduplication technique will be used to keep only one copy of the same file. Because of privacy concerns, some files will be encrypted and repeated checks will be permitted by staff with specific rights to achieve access control. Conventional de duplication solutions based on convergent encryption, although giving some secrecy, do not provide the duplicate check with differential privileges.

II. LITERATURE SURVEY

In [1], Yang Tang et. al. designs and implements FADE, a secure overlay cloud storage system that offers fine-grained, policy-based access control and file guaranteed destruction, in this work. It connects outsourced files with file access policies and reliably deletes files to make them unrecoverable to anybody when file access policies are revoked. FADE is based on a set of cryptographic key operations that are self-maintained by a quorum of key managers that are

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 11,November 2021

independent of third-party clouds in order to meet such security requirements. FADE, in particular, functions as an overlay system that works in tandem with today's cloud storage systems. Author builds a proof-of-concept FADE prototype using Amazon S3, one of today's cloud storage providers. The author also conducts comprehensive empirical investigations to demonstrate that FADE provides security protection for outsourced data while incurring relatively minor performance and monetary costs. The author's work shows how to include value-added security elements into today's cloud storage services.

Advantages

- A secure overlay cloud storage solution that offers fine-grained, policy-based access control and file ensured deletion.
- FADE is based on a set of cryptographic key operations that are self-maintained by a quorum of key managers that are independent of third-party clouds in order to meet such security requirements.

Drawbacks

- While users may now outsource data backup to third-party cloud storage providers to decrease data management expenses, there are security issues about protecting the privacy and integrity of outsourced data.

In [2], Huiqi Xu et. al. proposes the random space perturbation (RASP) data perturbation method for providing safe and efficient range query and kNN query services for protected data in the cloud. The RASP data perturbation approach combines order-preserving encryption, dimensionality expansion, random noise injection, and random projection to give significant resilience to attacks on perturbed data and queries. It also retains multidimensional ranges, allowing current indexing techniques to be used to speed up range query processing. To process the kNN queries, the kNN-R algorithm collaborates with the RASP range query method.

Advantages

- The widespread deployment of public cloud computing infrastructures, which use clouds to host data query services, has become an enticing alternative because to the benefits of scalability and cost-saving.
- It also retains multidimensional ranges, allowing current indexing techniques to be used to accelerate range query processing.

Drawbacks

- To maintain data security and query privacy, new ways are required. The effectiveness of query services, as well as the benefits of employing clouds, must also be retained.

In [3], Ming li et al. provide a cross-layer optimization approach for hybrid crowdsourcing in CPSs to ease heavy-duty computing. We specifically define an offline finite-queue-aware CPS service maximisation problem to crowdsource nodes' computing workloads in a CPS by combining computing resource management, routing, and link scheduling. Then, we establish both lower and upper bounds on the best solution to the issue. Furthermore, the lower limit solution is demonstrated to be a realistic approach that assures all queues in the network are finite, i.e., network strong stability. Extensive simulations were run to validate the performance of the suggested methods.

Advantages

- It determines both the lower and upper limits on the best solution to the issue. Furthermore, the lower limit conclusion is demonstrated to be a viable result, ensuring that all queues in the network are finite, i.e. network strong stability.

Drawbacks

- In general, facilitating collaborations in a CPS involves two issues: computer resource management and network architecture.
- It develops an offline finite-queue-aware CPS service maximisation problem to crowdsource the computing duties of nodes in a CPS.

In [4], Jin Li et al. propose Dekey, a novel architecture in which users do not need to keep their own keys but instead safely distribute convergent key shares across several servers. Dekey is secure according to the definitions stated in the proposed security model, according to security analysis. As a proof of concept, we use the Ramp secret sharing mechanism to develop Dekey and show that it has a low overhead in practical contexts.

Advantages

- In secure deduplication, it achieves efficient and dependable key management.
- Furthermore, with a growing number of users, such a basic key management technique creates a tremendous number of keys and needs users to safeguard the master keys diligently.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 11,November 2021

Drawbacks

- As promising as it sounds, doing safe deduplication in cloud storage is an emerging difficulty.
- Although convergent encryption has been widely used for safe deduplication, one significant difficulty in making convergent encryption feasible is managing a large number of convergent keys effectively and reliably.

In [5], Chaoling Li et al. proposed novel technique, data blocks are encrypted using a two-level encryption approach in which control keys are produced using a key derivation tree, encrypted using an All-Or-Nothing algorithm, and then distributed into a DHT network after being partitioned via secret sharing. This ensures that only authorised users may retrieve the control keys and decrypt the outsourced data throughout the data lifespan determined by the owner. Aside from secrecy, data dynamics and deduplication are performed independently by key derivation tree and convergent encryption adjustments. The analysis and practical results suggest that our approach can meet its security aim while also providing ensured deletion at a cheap cost.

Advantages

- This ensures that only authorised users may retrieve the control keys and decrypt the outsourced data throughout the data lifespan determined by the owner.
- Its security purpose is to achieve guaranteed deletion at a reasonable cost. However, with a growing number of users, such a basic key management technique creates a tremendous number of keys and needs users to safeguard the master keys diligently.

Drawbacks

- Resistance to hopping and sniffer assaults, data dynamics, and deduplication all occur.
- Although convergent encryption has been widely used for safe deduplication, one significant difficulty in making convergent encryption feasible is managing a large number of convergent keys effectively and reliably.

III. SYSTEM MODEL

In this project, we improve the security of our system. In particular, we describe an improved approach for enhancing security by encrypting the file with differential privilege keys. As a result, users who lack the necessary rights are unable to run the duplication check. Furthermore, even if such unauthorised users collude with the S-CSP, they cannot decipher the encrypted text. Our system is secure according to the definitions stated in the proposed security model, according to security analysis.

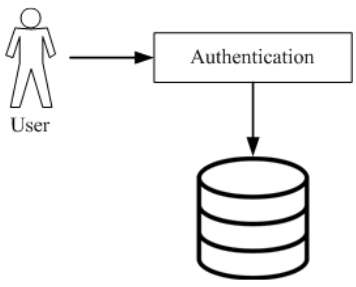
Convex encryption has been presented as a means of maintaining data secrecy while allowing for deduplication. It encrypts/decrypts a data copy using a convergent key generated by computing the cryptographic hash value of the data copy's content. Users keep the keys after key creation and data encryption and submit the cypher text to the cloud. Because the encryption method is predictable and derived from the data content, identical data copies produce the same convergent key and hence the same cypher text. A secure proof of ownership protocol is also required to offer verification that the user truly owns the same file when a duplicate is discovered in order to prevent unwanted access.

Modules

- User Module
- Server Startup and Upload File Module
- Secure DeDuplication Module
- Download Module

1. User Module

In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first. At the very least, you need to provide an email address, username, password, display name, and whatever profile fields you have set to required. The display name is what will be used when the system needs to display the proper name of the user.

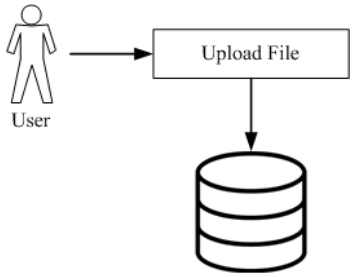


ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 11,November 2021

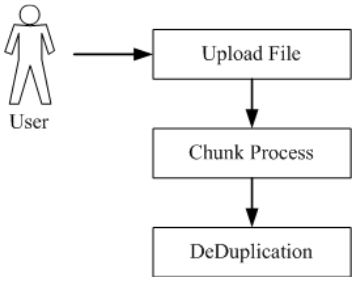
2. Server Startup and Upload File Module

The user can start up the server after cloud environment is opened. Then the user can upload the file to the cloud.



3. Secure DeDuplication Module

The tag of a file will be decided by the file and the privilege to facilitate approved de duplication. To distinguish it from typical tag notation, we refer to it as a file token instead. A secret key KP will be coupled with a privilege p to build a file Token to support allowed access. De duplication takes use of similar material, whereas encryption tries to make all information look random; the same content encrypted with two distinct keys yields radically different cypher text. As a result, integrating the space economy of de duplication with the secrecy elements of encryption is difficult.



4. Download Module

During cloud storage, the user can download the file using the key or token. Once the key request has been received, the sender can either send the key or deny it. The recipient can decrypt the message using this key and the request id that was produced when the key request was sent.

VI. CONCLUSION

In this proposed work, the concept of approved data de duplication was developed to preserve data security by adding differential user privileges in the duplicate check. We also showed many innovative de duplication constructions that provide approved duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are created by the private cloud server using private keys. Security research shows that our techniques are secure against the insider and outer threats defined in the proposed security architecture. We implemented a prototype of our suggested approved duplicate check technique as a proof of concept and ran test bed tests on it. We demonstrated that, as compared to convergent encryption and network transmission, our permitted duplication check technique had a low overhead.

In future, cloud data storage security is still fraught with difficulties and critical, and that many research issues remain unresolved.

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: (2013). Serveraided encryption for deduplicated storage. In USENIX Security Symposium.

[2] M. Bellare, S. Keelveedhi, and T. Ristenpart (2013.). Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312

[3] M. Bellare, C. Namprempre, and G. Neven(2009). Security proofs for identity-based identification and signature schemes. J. Cryptology, 22(1):1–61.

[4] M. Bellare and A. Palacio(2002). Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162–177.

[5] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider.(2011) Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011).

[6] P. Anderson and L. Zhang(2010). Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA.

[7] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless(2013): Serveraided encryption for

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 11,November 2021

deduplicated storage. In USENIX Security Symposium.

[8] M. Bellare, S. Keelveedhi, and T. Ristenpart(2013). Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296– 312.

[9] M. Bellare, C. Namprempre, and G. Neven(2009). Security proofs for identity-based identification and signature schemes. J. Cryptology, 22(1):1–61.

[10] M. Bellare and A. Palacio(2002.). Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162–177.

[11] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider(2011). Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011).

[12] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer(2002). Reclaiming space from duplicate files in a serverless distributed file system. In ICDCS, pages 617–624.

[13] D.Ferraiolo and R. Kuhn(1992). Role-based access controls. In 15th NIST-NCSC National Computer Security Conf.

[14] S.Halevi, D. Harnik, B. Pinkas, and A. ShulmanPeleg(2011). Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM.

[15] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou(2013). Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems.