Deduplication on Encrypted Data in Cloud Computing

Aditya Tryambak Sambare[1]; Prathamesh Hanmant Shingate[2]; Amol Kishor Shelke[3]
Mansi Ranjit Thakur[4]; G Nazia Sulthana[5] (Professor)
Department of Computer Engineering Mahatma Gandhi Mission’s College of Engineering and Technology,
Navi Mumbai, Maharashtra

Abstract:- Cloud storage is a crucial component of cloud computing, allowing users to expand their storage without upgrading their equipment and overcome resource constraints. Cloud users' data is always encrypted before being outsourced to ensure their security and privacy. However, encrypted data may result in a significant waste of cloud resources. Storage complicates data sharing among authorized users. We continue to face issues with encrypted data storage and deduplication. Traditional deduplication strategies are designed for certain application settings where data owners or cloud servers have full control over the process. They cannot meet data owners' varying requests based on data sensitivity. This study proposes a flexible data storage management method that combines deduplication and access control across various Clouds. Service Providers (CSP). We assess its performance through security analyses, comparisons, and implementations. The results demonstrate security, effectiveness, and efficiency for actual applications.

Keywords:- *Data Deduplication, Cloud Computing, Access Control, Storage Management.*

I. INTRODUCTION

Cloud computing provides centralized data storage and online access to computer services or resources. This new approach to IT services reorganizes resources and tailors them to meet user needs. Cloud computing offers numerous benefits, including scalability, elasticity, fault tolerance, and pay-per-use. Cloud storage allows users to store large amounts of data without the need for gadget upgrades and access it anytime, anywhere. Cloud data storage provided by Cloud Service Providers (CSPs) is not without issues. Data stored in the cloud may require varying levels of protection based on its sensitivity. The cloud stores sensitive personal information, publicly shared data, and group-shared data. Important data should be securely stored in the cloud to avoid illegal access. Unimportant data may not be subject to such requirements. Outsourced data may contain sensitive information, so data owners may prefer to control it themselves or delegate control to a third party if they are unavailable or unsure how to do so. Adapting cloud data access control to varied scenarios and user needs is a practical issue. Access control for encrypted data has been extensively researched in the literature. Few cloud data protection solutions can meet diverse needs uniformly, particularly when it comes to cheap deduplication. Flexible cloud data deduplication with access control remains an open subject. Duplicated data can be encrypted and stored in the cloud by multiple users across different CSPs. Data deduplication and access control are supposed to be compatible. The same data, whether encrypted or not, is stored once in the cloud and can be accessed by multiple people based on the policies of data owners or holders. Duplicate data in cloud storage can waste network resources, burn energy, increase prices, and complicate data management. Economic storage benefits both CSPs and cloud consumers by lowering operating expenses and service prices. Cloud data deduplication is crucial for storing and managing large amounts of data. However, there are few research on flexible cloud data deduplication across several CSPs. Existing solutions lack flexibility and uniformity in supporting both deduplication and access control in the cloud. This work proposes a heterogeneous data storage management method to address the issues mentioned above. The proposed approach is compatible with the access control scheme proposed previously. It allows for flexible cloud storage management, including data deduplication and access control, which can be managed by the data owner, a trusted third party, or neither. The suggested technique addresses data security concerns while also saving storage space through deduplication across many CSPs. Thus, it can be used in a variety of data storage applications. Our scheme is unique and distinct from prior work. This study proposes using encryption and deduplication to conserve cloud storage across several CSPs while maintaining data security and privacy in different scenarios. Our proposed heterogeneous data management scheme supports deduplication and access control based on data owners' needs, adapting to various application scenarios. Our method allows for flexible data exchange among eligible users, governed by data owners, trusted parties, or both. The suggested scheme's performance is validated by security analysis, comparison to current work, and implementation-based evaluation. The results demonstrate security, benefits, efficiency, and possible use.

II. EXISTING SYSTEM

Yang et al. presented the Provable Ownership of the File (POF) approach, which enables users to establish their ownership of a file without uploading the complete file to the server. Data ownership evidence is an important part of data deduplication, particularly for encrypted data. However, this technique does not provide for flexible deduplication control across many CSPs.

Yan et al. presented a PRE-based deduplication strategy that relied solely on authorized parties to govern data deduplication. It is unable to adapt to many conditions, particularly the data access regulated by the data proprietors. Another sentence from our earlier work.

- *Disadvantage-*

Disadvantages of the current method include little research on flexible cloud data deduplication across several CSPs. Existing solutions lack flexibility and uniformity in supporting deduplication and access control in the cloud.

III. LITERATUE SURVEY

The SRRS system was presented by the authors in [1]. It uses a role re-encryption algorithm to effectively accomplish approved data deduplication and a convergent algorithm to maintain data confidentiality. To manage keys and user roles, a management center is introduced. On the client side, computational cost and overhead are decreased with the addition of the management center to the system. The SRRS system decreases bandwidth usage and storage space requirements by performing data deduplication.

A unique attribute-based storage system that facilitates safe and effective deduplication has been proposed by the authors in [2]. Additionally, it discussed the flaw in the common attribute-based encryption method, which is its inability to provide secure deduplication. The system operates in a hybrid cloud setting, with the public cloud handling storage and the private cloud handling the identification of identical copies. There are two main benefits to the system:

- Data sharing is done while maintaining data confidentiality by defining an access policy.
- Here, high standard theory is used to achieve the concept of data security, while others were unable to carry things out in accordance with this philosophy

The author of [3] described the ABE (Attribute Based Encryption) technology, which is utilized to effectively transfer data and minimize storage space. In this method, the user is granted the ability to calculate and decode the encrypted data if their attributes match.

Convergent encryption is a mechanism that authors [4] devised to secure data throughout the deduplication process. The data that is outsourced is transformed into encrypted text prior to deduplication. Additionally, the authors have given the users access to various rights.

The authors of [5] presented (MLE), which offers secure deduplication. Large files work best with this method because it requires server-side schema maintenance. Large files require better upkeep; thus this plan works well. Both file-level and block-level deduplication are supported by this method.

Updatable block-level deduplication, which allows for simple data updating and deduplication on encrypted data, was presented by the authors in [6]. Here, the problem with file-level deduplication of efficient data updating is resolved. MLE handles certain obstacles, while UBLDE protocol efficiently handles others. The difficulty of dynamic ownership management is met here.

The authors of [7] propose a method to lower the expense of data updates. The user cannot update encrypted data in an efficient or secure manner using the current MLE solution. A single piece of data update comes at a hefty expense. Thus, the authors have presented message-locked encryption that is updateable at the block level. method that seeks to lower the logarithm of computing cost to file size. Additionally, it now requires confirmation of ownership for users to access files.

In order to enable allowed data duplication, the author of [8] presents a strategy that makes use of the symmetric encryption algorithm, hashing technique, convergent encryption algorithm, and token generation scheme. Here, the security and confidentiality of user data are upheld. Both passive and active attacks are prevented on the data.

To facilitate dynamic ownership management, writers in [9] presented PoW (Proof-of-ownership) with data deduplication. Data deduplication at the file, user, and block levels is supported by this system. This plan successfully protects data confidentiality and performs secure deduplication. uniformity. It also lessens the need for storage space and key management. The author of [10] has reviewed numerous approaches and technological advancements for putting data deduplication into practice. They've also included a comparison of different technology. The study illustrates how conducting data deduplication compromises data confidentiality to varying degrees.

To facilitate dynamic ownership management, PoW (Proof-of-ownership) with data deduplication has been presented by the authors in [11]. Block-level, cross-user, and file-level data deduplication are all supported by this system. This plan successfully maintains data and does secure deduplication, secrecy and regularity. It also lessens the workload for storage and key management.

IV. PROBLEM STATEMENT

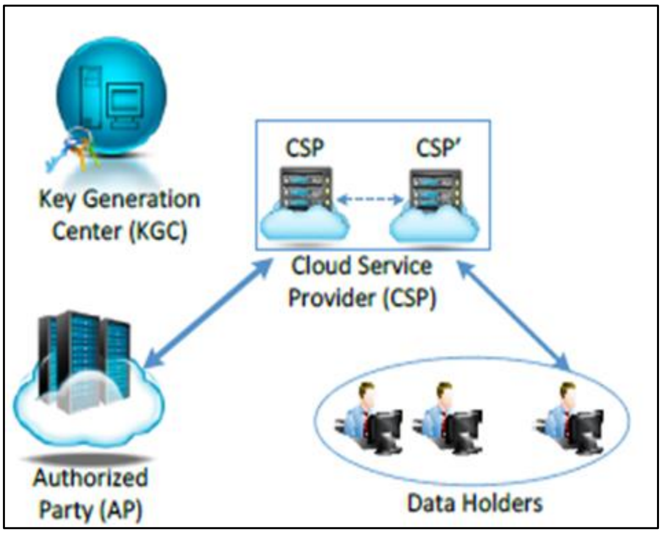

Fig 1 System and Security

➢ *System and Security –*

Model Figure depicts the system where the proposed strategy can be applied. It includes four sorts of entities:

- A trusted Key Generation Center (KGC) is responsible for generating system parameters and issuing certificates.
- The Cloud Service Provider (CSP) provides data storage services. There could be multiple CSPs in the system. Cloud users can select one of these options to manage their uploaded data and access advanced features. CSPs might collaborate in a business agreement to save storage space through deduplication.
- Data is stored at cloud service providers. Various CSPs may service the data holders. Multiple data holders or a single cloud user can store encrypted or plain data at one or multiple CSPs.
- The Authorized Party (AP) manages data access as a delegate for data owners and supports deduplication. In this arrangement, all entities trust AP. CSPs cannot be completely trusted. They are fascinated about cloud users' raw data, but rigorously adhere to system design and protocols. We assume that the AP will never collude with CSPs because to differing financial incentives and objectives collusion could harm the reputation of CSPs, ultimately resulting in loss of business.

➢ *Notations-*

PKu u's public key for ABE, along with their unique user ID. The key for user attribute verification generates a personalized secret attribute key for u. SKu The secret key for decrypting ABE. The public key of the Public-Key Cryptosystem (PKC) is used for encryption and signature verification. $SK'u$ is the secret key for PKC decryption and signature generation. The symmetric key of u is used to encrypt user data. $DE\boldsymbol{W}1,u$ is the partial key 1 of $\boldsymbol{PQW}u$. $\boldsymbol{PQW}2,u$ is the partial key 2 of $\boldsymbol{PQW}u$. The public key of u for attribute $\boldsymbol{UP}$ is used to encrypt $\boldsymbol{PQ}K2,u$.

V. PROPOSED SYSTEM

This study proposes a holistic and heterogeneous data storage management scheme to address the issues mentioned above. Our goal is to save cloud storage across many CSPs while maintaining data security and privacy using encrypted storage with deduplication in diverse scenarios. Our proposed heterogeneous data management approach allows for deduplication and access control based on data owners' needs, adapting to various application scenarios. Our method allows for flexible data exchange among eligible users, managed by either data owners, trusted parties, or both. ϖ Our scheme is unique and distinct from prior work. It is a general approach for realizing encrypted cloud data, deduplication with access control facilitates cooperation across multiple CSPs.

➢ *Advantages-*

Advantages of the proposed system include compatibility with access control schemes and flexible cloud storage management with data deduplication and access control, which can be controlled by the data owner, a trusted third party, or neither.

The suggested technique addresses data security concerns while also saving storage space through deduplication across many CSPs. Thus, it can be used in a variety of data storage applications.

We justify the proposed scheme's performance by doing a security analysis, comparing it to prior work, and evaluating its implementation. The findings demonstrate its security, benefits, efficiency, and possible use.

VI. ALGORITHM

A. Algorithm-

This sub-section introduces key algorithms for the proposed scheme-

➢ *System Setup*

- *Initiate Systems-*

This algorithm is conducted at KGC. It generates basic system settings for ABE and PRE, including generators and universal properties. Cloud user u generates key pairs based on system parameters, including ABE master key pair $P\boldsymbol{W}u$ and $S\boldsymbol{W}u$ for encryption and user decryption, and PKC key pair $P\boldsymbol{W}'u$ and $S\boldsymbol{W}u$ for PKC key issuance.

- *Setup Node (u)-*

With node identity u and public keys as input, this algorithm executed at KGC outputs a number of user credentials, $Cert(PKu)$, $Cert(PK'u)$ and $Cert(pku)$, which may be confirmed by CSPs and their users. The AP initiates itself by generating $pk\boldsymbol{M}P$ and $sk\boldsymbol{M}P$. $pkAP$ is transmitted to users of CSPs.

➢ *ABE Key Generation*

- $\boldsymbol{CreateIDPK}(\boldsymbol{ID}, \boldsymbol{SKu})$. This algorithm examines ID's policies and returns $pk\boldsymbol{UP}$, u for user u to regulate data deduplication and access.
- $\boldsymbol{IssueIDSK}(\boldsymbol{ID}, \boldsymbol{SKu}, \boldsymbol{PK}u')$. This algorithm is conducted by u to issue $skID$, u, u' to u' if the eligibility checks for u' is positive.

Otherwise, it returns $NULL$. User u verifies the properties of u'. If the policy is met, u will provide a secret key to u' for sharing duplicated data storage and future access. Otherwise, it will refuse the request.

To simplify the presentation, we utilize user identification as an example attribute, rather than complex attributes.

Access control based on user identity requires practice, as most data access in the cloud relies on user identity.

➢ *Data Encryption and Decryption-*

$\boldsymbol{Encrypt}$ $(\boldsymbol{DEKu}, \boldsymbol{M})$ encrypts M with $DE\boldsymbol{W}u$ and outputs ciphertext $\boldsymbol{O}Tu$ to secure M stored at CSP. $\boldsymbol{Decrypt}$ $(\boldsymbol{DEKu}, \boldsymbol{CTu})$ decrypts CTu with $\boldsymbol{P}E\boldsymbol{W}u$ and returns M. The process allows data holders to retrieve the plain content of CTu stored at CSP.

- *Symmetric Key Management-*

This approach generates partial keys (e.g., $\boldsymbol{PEW}1$, u and $\boldsymbol{PQW}2$, u) from input $\boldsymbol{PEWu}$ using random separation. If necessary, $DEKu$ can be separated into various pieces.

$\boldsymbol{CombineKey}(\boldsymbol{DEK1},\boldsymbol{u},\ \boldsymbol{DEK2},\boldsymbol{u})$. This algorithm combines partial keys of $\boldsymbol{PEKu}$, such as $\boldsymbol{PEW}1$, u and $\boldsymbol{PQW}2$, u, to produce the full key $\boldsymbol{PQWu}$.

- *Partial Key Control based on ABE Operated by the Data Owner –*

$\boldsymbol{EncryptKey}$ $(\boldsymbol{DEK2},\ \boldsymbol{u},\ \boldsymbol{\lambda},\ \boldsymbol{pkID},\ \boldsymbol{u})$ encrypts $DEK2$, u with policy λ and outputs cipher-key **X**. This algorithm is executed at u.

$\boldsymbol{DecryptKey}$ $(\boldsymbol{CK2},\ \boldsymbol{u},\ \boldsymbol{\lambda},\ \boldsymbol{SKu'},\ \boldsymbol{skID},\ \boldsymbol{u},\ \boldsymbol{u'})$ decrypts cipher key $CK2$, u and outputs $DEK2$, u. The algorithm is executed at u'.

Partial Key Control with PRE Operated by AP. We use PRE to enable AP to re-encrypt $C1$. During cipher text re-encryption, CSP does not learn about $DEK1$. The PRE algorithms are represented as follows: The function $\boldsymbol{E}$ $(\boldsymbol{pkAP},\ \boldsymbol{DEK1},\ \boldsymbol{u})$ generates $CW1 = E\ (pkMP,\ \boldsymbol{PQW}1,\ u)$ by taking $pkMP$ and $\boldsymbol{PEW}1$, u as input. $\boldsymbol{RG}$ $(\boldsymbol{pkAP},\ \boldsymbol{skAP},\ \boldsymbol{pku'})$ outputs re-encryption key $rkAP{\rightarrow}u'$ for the proxy CSP by taking $pkAP$, $skAP$, and pku' as input. $\boldsymbol{R}$ $(\boldsymbol{rkAP{\rightarrow}u'},\ \boldsymbol{CK1})$ takes input $rkAP{\rightarrow}u'$ and $CK1$, and outputs $R\ (rkAP{\rightarrow}u',\ CK1) = E\ (pku',\ DEK1,\ u) = CK'1$, which can be decrypted with sku'. The function $\boldsymbol{D}$ $(\boldsymbol{sku},\ \boldsymbol{CK'1})$ generates $\boldsymbol{PEW}1$, u from the inputs sku and $\boldsymbol{OW}'1$.

VII. CONCLUSION

Data deduplication plays a crucial role in cloud storage, particularly for huge data. management. This work proposes a heterogeneous data storage management method with customizable cloud data deduplication and access control. Our scheme provides cost-effective big data storage across numerous CSPs, adapting to different application scenarios and demands. It supports data deduplication and access control with varying security needs. Our security analysis, comparison to prior work, and performance evaluation demonstrated that our scheme is secure, sophisticated, and efficient. Our approach protects user privacy by storing encrypted data on the cloud. Using pseudonyms can help protect identify privacy. The Key Generation Center (KGC) verifies and certifies the relationship between a genuine identity and a pseudonym. Our future effort is to strengthen user privacy and improve our system for actual deployment. We will analyze the suggested method using game theory to ensure its security and rationality.

REFERENCES

[1]. R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina, "Controlling data in the cloud: outsourcing computation without outsourcing control," in Proc. 2009 ACM Workshop Cloud Comput. Secur., pp. 85-90, 2009.

[2]. S. Kamara, and K. Lauter, "Cryptographic cloud storage," Financ. Crypto. Data Secur., pp. 136-149, Springer, 2010.

[3]. Q. Liu, C. C. Tan, J. Wu, and G. Wang, "Efficient information retrieval for ranked queries in cost-effective cloud environments," in Proc. 2012 IEEE INFOCOM, pp. 2581-2585, 2012.

[4]. M. Kallahalla, E. Riedel, R. Swaminathan, Q. Wang, and K. Fu, "Plutus: scalable secure file sharing on untrusted storage," in Proc. USENIX Conf. File Storage Technol., pp. 29–42, 2003.

[5]. E.-J. Goh, H. Shacham, N. Modadugu, and D. Boneh, "SiRiUS: securing remote untrusted storage," in Proc. Netw. Distrib. Syst. Secur. Symp., pp. 131-145, 2003.

[6]. J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attribute-based encryption," in Proc. of IEEE Symp. Secur. Privacy (SP'07), pp. 321-334, 2007.

[7]. V. Goyal, O. Pandey, A. Sahai, and B. Waters, “Attribute-based encryption for fine-grained access control of encrypted data”, in Proc. of 13th ACM Comput. Commun. Secur., pp. 89–98, 2006.

[8]. S. Muller, S. Katzenbeisser, and C. Eckert, “Distributed attribute-based encryption,” in Proc. of 11th Annual Int. Conf. Inf. Secur. Crypto., pp. 20–36, 2008.

[9]. A. Sahai, and B. Waters, “Fuzzy identity-based encryption,” in Proc. of 24th Int. Conf. Theory App. Cryptographic Tech., pp. 457– 473, 2005.

[10]. S. C. Yu, C. Wang, K. Ren, and W. J. Lou, “Achieving secure, scalable, and fine-grained data access control in cloud computing,” in Proc. of IEEE INFOCOM, pp. 534–542, 2010.

[11]. G. J. Wang, Q. Liu, J. Wu, and M. Y. Guo, “Hierarchical attribute-based encryption and scalable user revocation for sharing data in cloud servers,” Comput. Secur., vol. 30, no. 5, pp. 320–331, 2011.