Security Aware High Scalable paradigm for Data Deduplication in Big Data cloud computing Environments

Anitha P¹, Dhanushram R¹, Harihara Sudhan D¹, Indhresh T R S¹
¹Computer Science and Engineering, Sri Krishna College of Engineering and Technology
anithap@skcet.ac.in

Abstract. The challenge of data deduplication has piqued the interest of the research community in order to provide effective and efficient solutions for cloud data centres to remove bottlenecks caused by duplicate data on the exponential growth of outsourced data. An imminent challenge is the presence of high redundant existence in the storage memory of the cloud environment. Traditional algorithm has been employed to mitigate the data duplication in the cloud environment but it is a daunting task since it requires an effective solution to manage the duplication in the application basis to eliminate the privacy violation on sensitive data. In order to tackle above mentioned issues, a new paradigm named as Security Aware High Scalable has been proposed on large scale computing data centres. Proposed method employed to secure the data Deduplication process using access control mechanism on data of the privileged authorities. Outsourced data will be in the form of encrypted files. Access control mechanism is provided to the secure authorities to conduct the data deduplication on the data that was outsourced. Access Control Mechanism is employed using encryption techniques. It uses randomised encryption that is convergent and stable distribution of ownership party keys to enable the cloud service provider to handle outsourced access to data even though control changes on a regular basis. This prevents data from being leaked not just to those who have had their rights removed who previously owned the data, but even to a degree trustworthy yet suspect server for cloud storage. Furthermore, the proposed technique protects integrity of data from attacks based on label inconsistencies. As a measure, the proposed technique has been modified that enhances security. The efficiency review reveals that the proposed system is nearly as successful as the existing ones, with just a small increase in computational overhead.

Keywords: Deduplication, Encryption, Cloud Storage, Access Control Mechanism, Big Data

1. Introduction

Due to advancements in cloud computing has led to great facilities to transfer their information into the cloud without revealing confidential information to third parties, they want only access to the data is restricted to clients through unique privileges. This necessitates the storage of data in a safe format with policies for access management that prevent someone other than users with unique attributes (or credentials) from decrypting the encrypted data[1]. ABE (attribute-based encryption) is a type of encryption where a message is encrypted using an access policy (or access structure) over a collection of attributes, and a user can use his or her private key to decrypt ciphertext if the set of attributes matches the access policy associated with this ciphertext.
Deduplication [2], a standardized technique for the reduction of data popular in computer storage solutions, reduces not just to data storage capacity, energy, and ventilation in data services, but also management time, technical efficiency, and the risk of human mistake. It breaks down huge data items into manageable bits, assigns fingerprints to these chunks, replaces identical bits with signatures after looking up the chunk signature index, and then transfers or stores the specific bits for better communication and storage performance. Data deduplication has proven to be efficient in a number of applications, including backup systems [3], storage in virtual devices and storage in primary form[4].

Big Data deduplication is a scalable solution distributed deduplication strategy for managing data deluges as architecture of storage shifts to get together cloud storage agreement on service standard specifications. [5] It is a daunting task since it requires an effective solution to manage the duplication in the application basis to eliminate the privacy violation on sensitive data. In order to tackle above mentioned issues, a new paradigm named as Security Aware High Scalable has been proposed on large scale computing data centres. Proposed method employed to secure the data Deduplication process using access control mechanism on data of the privileged authorities. Outsourced data will be in the form of encrypted files.

Access control mechanism is provided to the secure authorities to conduct the data Deduplication On the data that was outsourced. [6] Controlling Access Mechanism is employed using encryption techniques. It uses randomised convergent encoding and stable team's owner digital signatures to enable the cloud service provider to handle outsourced data access even when possessing excitations. This prevents data from being leaked not only to those who have had their rights removed who formerly owned the data, and also it is to a trustworthy yet suspect Server for cloud computing. Furthermore, the scheme that is proposed protects label inconsistencies and attacks compromise integrity of data.

The rest of the paper is organised the following: works on Data Deduplication on cloud computing paradigm. Section 3 defines and designs the proposed Secure Security Aware High Scalable paradigm for Data Deduplication in Big Data cloud on evolving data volumes. In Section 4, experimental results and performance of the proposed model has been detailed against state of art approaches on large volumes of data. Finally, Section 5 brings the paper to a close.

2. Related work

In this section, various existing model enumerating data deduplication with secure authority on evolving volume of data has analysed on basis of access control mechanism and encryption on outsourced data towards intra node deduplication and internode deduplication in detail as follows

2.1 In the Cloud, Encrypted Big Data Deduplication

In this method, outsourced information has been preserved from the data holders. This data is often encrypted and gathered in the cloud. The proprietorship problem and re-encryption of the proxy are used to rehab cloud-stored encrypted data. It combines deduplication in some cloud with access control [7]. When inter-file similarity is low, this method limits deduplication; it also has a higher rate of context switching and data bias. For large-scale distributed deduplication, however, it is unable to achieve high capacity savings while maintaining load balance.

2.2 Secure Approved Deduplication Using a Hybrid Cloud Approach

In this method, the convergent encryption technique was investigated to encrypt the information until redistribution, preserving the confidentiality and security of data while facilitating deduplication. In addition to data itself, the differential rights of users are taken into account in the duplicate search. Finally, in a hybrid cloud architecture, deduplication structures supporting approved duplicate checks have been presented. These complex changes in ownership can occasionally occur in
a functional cloud service system, and as a result, they must be properly controlled to prevent the cloud service's security from deteriorating[8].

3. Proposed model

In this section, a Security Aware High Scalable paradigm for Data Deduplication in Big Data cloud has been proposed with design steps on terms of security and granularity of the deduplication as follows

3.1 Big Data Cloud architecture

The following entities must be used to create a stable scheme for deduplication for encryption in data with dynamic management in the ownership capabilities.

- Owner of the data

It is a customer who has data that he or she wishes to hold in the database. The data is encrypted and sent to cloud storage along with the index information, or tag, by the data owner. If a cloud data transcoded information to the cloud storage that would not actually reside, it is referred to as an early uploader; if the data already occurs, he is referred to as a prior user because other owners might have already submitted a certain data. A sharing group[9] is a list of data owners who share the same data in cloud storage.

Device for Storage in Cloud

It is a company which offers solutions for cloud computing. This is made up of two parts: storage and a server for cloud. If required, the public cloud deduplicates and stores the data that users have outsourced. The server of cloud keeps track of stored data ownership lists, which are made up of a unique identifier for the data and the names of its person who owns. The cloud server manages (e.g., conflicts, revokes, and updates) group keys for each ownership group and controls depending on the ownership lists, access to the stored data as a group key authority Figure 1.

![Figure 1](image)

**Figure 1.** Architecture of the proposed Secure Deduplication paradigm

The computer in the cloud is believed to become trustworthy and yet a little suspicious. So, this will carry out some tasks assigned by the system honestly, but it wishes to understand as much as possible about the contents which were encrypted. Thus a result, even though it is truthful, should be prevented through gaining access to the encrypted data's plaintext. The proposed architecture paradigm is depicted in Figure 1 for secure deduplication in the big data cloud.

- Attack Model
Attack Model is established to verify the security strength of the proposed model through the following attack processing scheme. Users which don’t have authorized access to the data accessing Collude to expose the data. Attack algorithm generates ciphertext attacks and plaintext attack[10].

1.1. Attribute based Encryption – Access Control Mechanism

Generation

Key Generation is carried out for Data Authority to measure data deduplication and Data owner through RSA algorithm. In this RSA is considered as Ring generation algorithm

- A public/private key pair is generated by the user by doing these following steps:
  - p, q are two big primes chosen at random.
  - calculating their n=p.q device modulus
  - Observe the (n)=(p-1) (q-1)

Consider K as a protected parameter, and the public signing key is returned.

Set Up

It's used to get the number of data classes d and the protection parameter p for the public parameters param.

Key generation

A file owner who is able to exchange her data registers with her own public and pairs of private keys. The owner of data is in charge of categorising each data file into distinct groups. It produces the data owner's PK stands for public key, and MSK stands for master secret key.

Encryption Method

These file classes are encrypted using AES technique and stored in the cloud. It obtains the Public Key parameter PK, data class d and data as plain text T. outputs the ciphertext C.

\[ C = (C_0, C_1, C_2) = (Q_{(PK) + Q_i}, m.e(P_n, Q_n)) \]

Sign

It outputs the signature σ on the Message and the key of the Ring (V_{sign1}, V_{sign2}, V_{sign3}, V_{sign4}) in the group

- A group consists of n members G1,G2,G3,G4,G5….Gn
- Users which are outside the group G is U1,U2,U3,U4,……Un
- Each group member have their own (private, public) key pairs
- Any group member can upload encrypted file in cloud which is signed by ring signature
- Ring signature of a member = his/her private key and all group members public key

Output = Signature Message Pair (Sig, M)

1.2. Secure aware high Scalable Data deduplication

Hashing

Hash (Signature Message Pair)

Hash Value –H – Key for Ring

Verify
It verifies the signature $\sigma$ regarding the message $m$ in relation to the ring of the public key $R$. It takes hash value and produces the signature message pair, on processing of signature message pair, it output message.

### Decrypt

This aggregate key can be used by an approved data user to decrypt any message from any class. To produce the decrypted code, it obtains the cipherText $C$, data class $d$, and aggregate key $AK$. 

$$M = \sum_{i=0}^{n} d(P, Q) \ast C$$

### 4. Experimental Results

We will look at the output of in this segment, the secure deduplication mechanism of the proposed model against the exploring large volume of data sizes and its performance is evaluated on the encryption time, decryption time in terms of graphs and tables of results. The system's reliability is calculated by how well it uses proxy server technology to compute data deduplication while maintaining system protection. It will make delegating access privileges to privileged users much easier for data owners.

![Performance Comparison](image)

**Figure 2.** Performance Evaluation of the Decryption Time on Models proposed and current in different file sizes

The performance of the proposed model and current models in a big data cloud environment composed of outsourced data is described and summarised in Figure 2 and Table 1. Decryption time is calculated to measure the redundant computation by deduplication mechanism.

| Technique                       | No. of File or File Size | Time Take in ms | No of Decryption |
|--------------------------------|--------------------------|-----------------|------------------|
| Secure High Scalable Deduplication – Proposed | 5                        | 36000           | 4                |
| Deduplication                  | 5                        | 24000           | 2                |

Data protection for outsourced data can be easily assured against unauthorised users who have never been the owners of the information. It's a toxic assault on consistency in tag and is quickly found in some proposed system. As a result, proposed scheme protects data integrity from a tag accuracy poisoning attack. Whenever hardly a user attempts to load data into cloud storage which has been uploaded at some stage in the past, the main for the subsequent expansion team is modified individually.
also automatically and safely delivered to the legitimate owners who own the data instantly. As a result, the proposed scheme guarantees the forward confidentiality of the outsourced results.

5. Conclusion

We designed and implemented the secure high scalable data deduplication architecture in the big cloud environment using authorized accessing control mechanisms in the outsourced data. This encryption is carried using RSA. Authorization is set on attribute based encryption. The employed method saves the storage space and network bandwidth. The proposed system involves a data encryption method again which allows for regular up-dates in database storage in the event of leadership changes. As a result, the proposed system improves privacy in data and security in storage of cloud against unauthorised utilizers as well as a trustworthy yet suspicious cloud server.

References

[1] Y. Yang, H. Zhu, H. Lu, J. Weng, Y. Zhang, and K. R. Choo, Cloud based data sharing with fine-grained proxy re-encryption, Pervasive and Mobile Computing, vol. 28, pp. 122–134, 2016.

[2] M. Bellare, S. Keelveedhi, and T. Ristenpart, Message-locked encryption and secure deduplication, in Advances in Cryptology - EUROCRYPT 2013, 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Athens, Greece, May 26-30, 2013. Proceedings, ser. Lecture Notes in Computer Science, vol. 7881. Springer, 2013, pp. 296–312.

[3] M. Bellare and S. Keelveedhi, Interactive message-locked encryption and secure deduplication, in Public-Key Cryptography – PKC 2015 - 18th IACR International Conference on Practice and Theory in Public-Key Cryptography, Gaithersburg, MD, USA, March 30 – April 1, 2015, Proceedings, ser. Lecture Notes in Computer Science, vol. 9020. Springer, 2015, pp. 516–538.

[4] H. Anandakumar and K. Umamaheswari, A bio-inspired swarm intelligence technique for social aware cognitive radio handovers, Computers & Electrical Engineering, vol. 71, pp. 925–937, Oct. 2018. doi:10.1016/j.compeleceng.2017.09.016

[5] R. Arulmurugan and H. Anandakumar, Early Detection of Lung Cancer Using Wavelet Feature Descriptor and Feed Forward Back Propagation Neural Networks Classifier, Lecture Notes in Computational Vision and Biomechanics, pp. 103–110, 2018. doi:10.1007/978-3-319-71767-8_9

[6] Y. Fu, H. Jiang, N. Xiao, L. Tian, F. Liu, L. Xu. ApplicationAware Local-Global Source Deduplication based Cloud Backup Services for Personal Storage. In Proc. of IEEE Transactions on Parallel and Distributed Systems, 25(5): 1155-1165, 2014.

[7] T. Yang, H. Jiang, D. Feng, Z. Niu, K. Zhou, Y. Wan. DEBAR: a Scalable High-Performance Deduplication Storage System for Backup and Archiving, Proc. of the 24th IEEE International Parallel and Distributed Processing Symposium (IPDPS’10), pp. 1-12, Apr. 2010

[8] S. Keelveedhi, M. Bellare, and T. Ristenpart, Dupless: Serveraided encryption for deduplicated storage, in Proceedings of the 22th USENIX Security Symposium, Washington, DC, USA, August 14-16, 2013. USENIX Association, 2013, pp. 179–194.

[9] B. Zhu, K. Li, and R. H. Patterson, Avoiding the disk bottleneck in the data domain deduplication file system, in 6th USENIX Conference on File and Storage Technologies, FAST 2008, February 26-29, 2008, San Jose, CA, USA. USENIX, 2008, pp. 269–282.

[10] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, J. C. S. Lui, A secure cloud backup system with assured deletion and version control, Proc. International Workshop on Security in Cloud Computing, 2011