ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

# VARIOUS FORMATS OF DATA STORAGE MANAGEMENT USING DEDUPLICATION IN CLOUD COMPUTING

**Mrs. K. SWATHI**

Assistant Professor, Department of CSE, Sri Mittapalli College of Engineering, Tummalapalem, NH-16, Guntur, Andhra Pradesh, India.

**1) A.VARDHANI, 2) D.M PRAVALLIKA , 3) D.TEJASWI , 4) G.CHANDINI**

B. Tech Students, Department of CSE, Sri Mittapalli College of Engineering, Tummalapalem, NH-16, Guntur, Andhra Pradesh, India.

***Abstract*:** Cloud storage as one of the most important services of cloud computing helps cloud users break the bottleneck of restricted resources and expand their storage without upgrading their devices. In order to guarantee the security and privacy of cloud users, data are always outsourced in an encrypted form. However, encrypted data could incur much waste of cloud storage and complicate data sharing among authorized users. We are still facing challenges on encrypted data storage and management with deduplication. Traditional deduplication schemes always focus on specific application scenarios, in which the deduplication is completely controlled by either data owners or cloud servers. They cannot flexibly satisfy various demands of data owners according to the level of data sensitivity. In this paper, we propose a heterogeneous data storage management scheme, which flexibly offers both deduplication management and access control at the same time across multiple Cloud Service Providers (CSPs). We evaluate its performance with security analysis, comparison and implementation. The results show its security, effectiveness and efficiency towards potential practical usage.

## I. INTRODUCTION

Cloud computing allows centralized data storage and online access to computer services or resources. It offers a new way of Information Technology (IT) services by re-arranging various resources and providing them to users based on their demands. Cloud computing has greatly enriched pervasive services and become a promising service platform due to a number of desirable properties, such as scalability, elasticity, fault tolerance, and pay-per-use.

Data storage service is one of the most widely consumed cloud services. Cloud users have greatly benefited from cloud storage since they can store huge volume of data without upgrading their devices and access them at any time and in any place.

However, cloud data storage offered by Cloud Service Providers (CSPs) still incurs some problems.

First of all, various data stored at the cloud may request different ways of protection due to different data sensitivity. The data stored at the cloud include sensitive personal information, publicly shared data, data shared within a group, and so on. Obviously, crucial data should be protected at the cloud to prevent from any access of unauthorized parties. Some unimportant data, however, have no such a requirement. As outsourced data could disclose personal or even sensitive information, data owners sometimes would like to control their data by themselves, while on some occasion, they prefer to delegate their control to a third party since they cannot be always online or have no idea how to perform such a control. How to make cloud data access control adapt to various scenarios and satisfy different user demands becomes a practically important issue. Access control on encrypted data has been widely studied in the literature. However, few of them can flexibly support various requirements on cloud data protection in a uniform way, especially with economic deduplication management. Second, flexible cloud data deduplication with data access control is still an open issue. Duplicated data could be stored at the cloud [39] in an encrypted form by the same or different users, in the same or different CSPs. From the standpoint of compatibility, it is highly expected that data deduplication can cooperate well with data access control. That is the same data (either encrypted or not) are only stored once at the cloud, but can be accessed by different users based on the policies of data owners or data holders (i.e., the eligible data users who hold original data). Although cloud storage space is huge, duplicated data storage could

ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

greatly waste networking resources, consume plenty of power energy, increase operation costs, and make data management complicated.

ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

Economic storage will greatly benefit CSPs by decreasing their operation costs and reversely benefit cloud users with reduced service fees. Obviously, cloud data deduplication is particularly significant for big data storage and management. However, the literature still lacks studies on flexible cloud data deduplication across multiple CSPs.

In this paper, we propose a holistic and heterogeneous data storage management scheme in order to solve the above problems. The proposed scheme is compatible with the access control scheme proposed in [33]. It further realizes flexible cloud storage management with both data deduplication and access control that can be operated by either the data owner or a trusted third party or both or none of them. Moreover, the proposed scheme can satisfy miscellaneous data security demands and at the same time save storage spaces with deduplication across multiple CSPs. Thus it can fit into various data storage scenarios. Our scheme is original and different from the existing work. It is a generic scheme to realize encrypted cloud data deduplication with access control, which supports the cooperation between multiple CSPs. Specifically, the contributions of this paper are:

We motivate to save cloud storage across multiple CSPs and preserve data security and privacy by managing encrypted data storage with deduplication in various situations.

We propose a heterogeneous data management scheme to support both deduplication and access control according to the demands of data owners, which can adapt to different application scenarios. Our scheme can support data sharing among eligible users in a flexible way, which can be controlled by either the data owners or other trusted parties or both of them.

We justify the performance of the proposed scheme through security analysis, comparison with existing work and implementation based performance evaluation. The results show its security, advantages, efficiency and potential applicability.

# II. RELATED WORK

## A. Access Control on Encrypted Data

Existing researches proposed to encrypt data before outsourcing it to the cloud in order to prevent data privacy from being invaded at CSP.

Access control on encrypted data requests that only authorized entities can decrypt the encrypted data.

An ideal approach is to encrypt each data once and issue relevant keys to authorized entities only once. However, due to the changeability of trust relationships, key management becomes complicated due to frequent key update.

Access Control Lists (ACLs) were applied to ensure data security in a distrusted or semi-trusted party (e.g., CSP). Before uploading data to CSP, the data owner first classifies the data into different groups, and then encrypts each group with a symmetric key, which is only distributed to the users in the ACL of the group. In this way, this group of data is only accessible by the users in the ACL. The shortcoming of this scheme mainly comes from the fact that the number of symmetric keys increases linearly with the number of groups. Moreover, the trust relationship change between one individual user and the data owner could cause essential update of relevant symmetric keys, which impacts other users in the same ACL. Thereby, this approach is impractical to be applied in many real applications where the trust relationship between different users changes frequently. Combining a traditional symmetric cryptosystem and an asymmetric cryptographic system was proposed for cloud data access control. However, the computation cost of key encryption increases linearly with the number of users in the ACL.

Attribute-Based Encryption (ABE) was proposed to achieve access control on encrypted cloud data. It specifies a set of attributes to identify users and encrypts data based on an access structure specified by attributes. Thus, encrypted data can only be decrypted by the users that hold such attributes that can satisfy the access structure. ABE is classified into two divisions: key-policy ABE (KP- ABE) and cipher text-policy ABE (CP-ABE) according to how the attributes link to cipher texts and decryption keys. ABE has such advantages as scalability and high flexibility in terms of attributes based access policies and fine-grained access control. It has been widely applied to secure cloud data storage in recent years.

However, all above existing solutions about access control on encrypted data did not consider how to solve the issue of duplicated data storage in cloud computing in a holistic and comprehensive manner, especially for encrypted data in various data storage scenarios. This issue is practically significant for big data secure storage over the cloud.

ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

### B. Encrypted Data Deduplication

It is a hot research topic to reconcile deduplication and client-side encryption. Existing industrial solutions fail to perform deduplication on encrypted data, e.g., Dropbox, Google Drive and Mozy. Message-Locked Encryption (MLE) was proposed to resolve this tension . Convergent Encryption (CE), the most prominent manifestation of MLE, was introduced. In CE, a user computes the key of data based on its hash code and encrypts with . Another user holding the same data can produce the same encrypted data, thus realizing deduplication. The CE suffers from offline brute-force dictionary attacks. As a result, CE can ensure high security only when the underlying data is drawn from a large space that is too big to exhaust. In addition, CE cannot support data access controlled by data owners, as well as other authorized parties. It is hard to support data revocation because generating a same new encryption key is hard to achieve for both the data owners and the data holders to re-encrypt the data.

A number of schemes were proposed to overcome the weakness of CE. Bellare et al. proposed DupLESS to resist the above-mentioned brute-force attacks. In DupLESS, users encrypt their data using the keys obtained from a Key Server (KS). They are generated based on the data with an oblivious Pseudo Random Function (PRF) protocol.

### C. Other Related Work

Yang et al. proposed a scheme called Provable Ownership of the File (POF), which allows a user to prove to a server that it really possesses a file without the need to upload the entire file. Data ownership proof is an essential process of data deduplication, especially for encrypted data. But this scheme does not consider flexible deduplication control across multiple CSPs. Yuan and Yu proposed a scheme to achieve data deduplication and secure data integrity auditing at the same time. It supports both public and batch auditing. This work applied different technologies (i.e., polynomial-based authentication tags and homomorphic linear authentication) from ours and focused on solving a different research issues.

Wu et al. developed Index Name Servers (INS) to reduce the workload caused by duplicated data. But this work cannot support the deduplication on encrypted data.

A hybrid data deduplication mechanism was proposed by Fan et al. It can deduplicate both plaintext and ciphertext. However, this mechanism has such a drawback that CSP knows the key that is used for data encryption. Therefore, it cannot be applied into such a situation that the CSP cannot be fully trusted by data owners. Li et al. formally addressed the problem of authorized data deduplication. Different from traditional deduplication systems, the differential privileges of users are further considered in duplicate check besides the data itself in a hybrid cloud architecture. All above work focused on solving different research issues from ours.

## III. SYSTEM AND SECURITY MODEL

It contains four types of entities:

1. Key Generation Center (KGC) that is fully trusted and responsible for system parameter generation and certification issuance.
2. The Cloud Service Provider (CSP) that offers a data storage service. Multiple CSPs could exist in the system. Thus, a cloud user can choose one of them to manage its uploaded data and seek advanced usage experiences. In addition, CSPs can cooperate with each other under a business agreement to save storage spaces through deduplication
3. The Data Owner or the Data Holder that uploads and stores data at CSPs. Different CSPs may serve the data holders. Multiple eligible data holders or a single cloud user could store the same encrypted or plain data at one CSP or across CSPs;
4. The Authorized Party (AP) that is responsible for controlling data access as a delegate of data owners as they expect to support deduplication.

In this system, AP is trusted by all entities. All CSPs cannot be fully trusted. That is, they are curious about the raw data of cloud users but follow system design and protocols strictly. We hold such an assumption that the AP would never collude with the CSPs due to different business incentive and interests. Any collusion would worsen the reputation of the CSPs, which lead to final loss of their business.

We additionally hold following assumptions. The data holder provides the correct hash code set of its data for data ownership verification. The first eligible data holder that uploads the data is regarded as the data owner. Multiple APs could exist in the system and can be supported by the underlying scheme. For simplification, we assume one AP in the system for easy presentation.

ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

## IV. SYSTEM DESIGN

We propose a scheme for heterogeneous data storage management with deduplication. It can be flexibly applied into such scenarios that cloud data deduplication is handled.

1) Only by the data owner;
2) By any trusted third party;
3) By both the data owner and the trusted third party; (i.e., plain data is stored at the cloud);
4) By either the data owner or the trusted third party.

Concretely, we use the hash code of data $M$ to check data duplication during data storage at the cloud. The data holder signs the hash code of the data for passing the originality verification of CSP. Meanwhile, a number of hash codes of randomly selected specific parts of the data are calculated with their indexes

### Fundamental Algorithms

1. System Setup
2. ABE Key Generation
3. Data Encryption and Decryption
4. Symmetric Key Management

## V. SECURITY ANALYSIS

The security of our scheme relies on ABE theory, PRE theory, symmetric key encryption and PKC. The security of PRE and ABE was proved in our previous work. Symmetric key encryption and PKC theory play as a security foundation in many security schemes. We assume that the applied key sizes of these two cryptosystems are long enough to satisfy the security requirements of our system. In what follows, we analyze the security of our scheme regarding data ownership verification and data deduplication.

Proposition 1) To pass data ownership verification, a cloud user must really hold data $M$.

Proposition 2) Data $M$ can be deduplicated in a secure way and only eligible users can access it if data owner $u$, CSP and AP cooperate without collusion.

Based on the above analysis, the CSP that stores $CT$ cannot obtain $M$ through $DEK$. The scheme can guarantee that $M$ is securely stored at CSP during deduplication, which can be only accessed by eligible data holders.

## A. Comparison with Existing Work

We compare our scheme with the previous work. One of them realizes deduplication managed by AP and the other manages deduplication by the online data owner. the proposed scheme can flexibly support various scenarios with similar computation complexity to existing work. Thereby, we compare their main properties in Table 3. We can see that the proposed scheme is a heterogeneous solution. It can realize both fine-grained and offline access control, thus it has better flexibility than previous work. In addition, the random hash code challenge is applied to verify data ownership, which can guarantee that the data holders really have the original data rather than its hash code Though possession proof has been achieved in by applying Elliptic Curve Cryptography (ECC) (with ownership verification time about 1.2 millisecond), hash code set employed in this paper is also very efficient if we make challenged part of data is small. If the challenged part of data is very small, e.g., within 1 kilobyte, we can achieve much better performance than considering the fast operation time of the hash function. Moreover, our scheme can cope with the situations of deduplication across multiple CSPs, which was not considered at all in previous work. In general, our scheme has distinct advantages compared with existing work in terms of high flexibility and advanced properties.

In our implementation, we applied AES for symmetric encryption, RSA for PKC and SHA-1 as a hash function to generate hash codes and hash sets. Our implementation was in C++ and adopted MySQL 5.5.46 to build a database. The experiments were conducted in a virtual machine running a 64-bits Ubuntu operating system on Amazon EC2 cloud service with Intel Xeon CPU E5-2670, 2.50GHz processor and 1-GB RAM We tested the correctness of our implementation in terms of each procedure described in Section 4.3. Herein, we only take the case of data deduplication with heterogeneous control as an example to illustrate our implementation due to paper size limitation.

under different control policies. The data owner deletion process involves generating a new $DEK$ and encrypting it. Therefore, there is no much difference under different access control policies. For a file without any access control, the deletion just needs to update related CSP file records and thus very efficient.

So, the proposed scheme achieves similar performance to the existing work. Considering its advanced properties and high flexibility, we conclude that our scheme outperforms the existing work.

ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

## VI. CONCLUSION AND FUTURE WORK

### A. CONCLUSION

Data deduplication is important and significant in the practice of cloud data storage, especially for big data storage management. In this paper, we proposed a heterogeneous data storage management scheme, which offers flexible cloud data deduplication and access control. Our scheme can adapt to various application scenarios and demands and offer economic big data storage management across multiple CSPs. It can achieve data deduplication and access control with different security requirements. Security analysis, comparison with existing work and implementation based performance evaluation showed that our scheme is secure, advanced and efficient.

Our scheme supports data privacy of cloud users since the data stored at the cloud is in an encrypted form. One way to support identity privacy is to apply pseudonyms in Key Generation Center (KGC), where a real identity is linked to a pseudonym, which is verified and certified by the KGC.

### B. FUTURE WORK

In our future work, we will further enhance user privacy and improve the performance of our scheme towards practical deployment. In addition, we will conduct game theoretical analysis to further prove the rationality and security of the proposed scheme.

## VII. 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) J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attribute-based encryption," in Proc. of *IEEE* Symp. Secure. *Privacy (SP'07)*, pp. 321-334, 2007

4) 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.

5) Sahai, and B. Waters, "Fuzzy identity-based encryption," in Proc. of 24th Int. Conf. Theory App. Cryptographic Tech., pp. 457– 473, 2005.

6) 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.

7) Z. Yan, Trust Management in Mobile Environments – Usable and Autonomic Models, IGI Global, Hershey, Pennsylvania, 2013.

8) Z. Yan, W. X. Ding, X. X. Yu, H. Q. Zhu, and R. H. Deng, "Deduplication on encrypted big data in cloud," *IEEE* Trans. on Big Data, vol. 2, no. 2, pp. 138-150, April-June 2016.

9) Z. Yan, X. Y. Li, M. J. Wang, A.V. Vasilakos, "Flexible data access control based on trust and reputation in cloud computing," *IEEE Trans. Cloud Comput.*, 2015. Doi: 10.1109/TCC.2015.2469662.

10) [34] J. Hur; D. Koo; Y. Shin; and K. Kang, "Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 11, pp. 3113-3125, 2016.

11) [35] J. Li, X. F. Chen, M. Q. Li, J. W. Li, P. P. C. Lee; and W. J. Lou, "Secure Deduplication with Efficient and Reliable Convergent Key Management," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 6, pp. 1615-1625, 2014.

.