Deduplication of Encrypted Big Data In Cloud

1. *Khurud Pallavi B., Information Technology, Sanjivani COE, Kopargaon, Maharashatra, India*
2. *Kurhade Jyoti B., Information Technology, Sanjivani COE, Kopargaon, Maharashatra, India*
3. *Late Rohini A., Information Technology, Sanjivani COE, Kopargaon, Maharashatra, India*
4. *Lawar Shreya A., Information Technology, Sanjivani COE, Kopargaon, Maharashatra, India*

ABSTRACT

With the continuous and exponential increase of the number of users and the size of their data, data deduplication becomes more and more a necessity for cloud storage providers. By storing a unique copy of duplicate data, cloud providers greatly reduce their storage and data transfer costs. The advantages of deduplication unfortunately come with a high cost in terms of new security and privacy challenges. We propose ClouDedup, a secure and efficient storage service which assures block-level deduplication and data confidentiality at the same time. Although based on convergent encryption, ClouDedup remains secure thanks to the definition of a component that implements an additional encryption operation and control mechanism.Furthermore, as the requirement for deduplication at block-level raises an issue with respect to key management, we suggest to include a new component in order to implement the key management for each block together with the actual deduplication operation.We show that the overhead introduced by these new components is minimal and does not impact the overall storage and computational costs.

Keyword: *Access control, big data, cloud computing, data deduplication, proxy re-encryption*

INTRODUCTION

With the potentially infinite storage space offered by cloud providers, users tend to use as much space as they can and vendors constantly look for techniques aimed to minimize redundant data and maximize space savings. A technique which has been widely adopted is cross-user deduplication. The simple idea behind deduplication is to store duplicate data (either files or blocks) only once. Therefore, if a user wants to upload a file (block) which is already stored, the cloud provider will add the user to the owner list of that file (block).

Deduplication has proved to achieve high space and cost savings and many cloud storage providers are currently adopting it. Deduplication can reduce storage needs by up to 90-95 percent for backup applications and up to 68 percent in standard file systems[5]. Along with low ownership costs and flexibility, users require the protection of their data and confidentiality guarantees through encryption.Unfortunately, deduplication and encryption are two conflicting technologies. While the aim of deduplication is to detect identical data segments and store them only once, the result of encryption is to make two identical data segments indistinguishable after being encrypted. This means that if data are encrypted by users in a standard way, the cloud storage provider cannot apply deduplication since two identical data segments will be different after encryption. On the other hand, if data are not encrypted by users, confidentiality cannot be guaranteed and data are not protected against curious cloud storage providers. A technique which has been proposed to meet these two conflicting requirements is convergent encryption whereby the encryption key is usually the result of the hash of the data segment.Although convergent encryption seems to be a good candidate to achieve confidentiality and deduplication at the same time, it unfortunately suffers from various well-known weaknesses including dictionary attacks: an attacker who is able to guess or predict a file can easily derive the potential encryption key and verify whether the file is already stored at the cloud storage provider or not.

Figure 1: System model

Cloud computing offers a new way of Information Technology services by rearranging various resources (e.g., storage, computing) and providing them to users based on their demands. Cloud computing provides a big resource pool by linking network resources together.It has desirable properties, such as scalability, elasticity, fault tolerance, and pay-peruse.Thus, it has become a promising service platform. The most important and popular cloud service is data storage service. Cloud users upload personal or confidential data to the data center of a Cloud Service Provider (CSP) and allow it to maintain these data. Since intrusions and attacks towards sensitive data at CSP are not avoidable[6] , it is prudent to assume that CSP cannot be fully trusted by cloud users. Moreover, the loss of control over their own personal data leads to high data security risks, especially data privacy leakages[2]. Due to the rapid development of data mining and other analysis technologies, the privacy issue becomes serious[2]. Hence, a good practice is to only outsource encrypted data to the cloud in order to ensure data security and user privacy.But the same or different users may upload duplicated data in encrypted form to CSP, especially for scenarios where data are shared among many users. Although cloud storage space is huge, data duplication greatly wastes network resources, consumes a lot of energy,and complicates data management. The development of numerous services further makes it urgent to deploy efficient resource management mechanisms. Consequently, deduplication becomes critical for big data storage and processing in the cloud. Deduplication has proved to achieve highcost savings, e.g., reducing up to 90-95 percent storage needs for backup applications and up to 68 percent in standard file systems. Obviously, the savings, which can be passed back directly or indirectly to cloud users, are significant to the economics of cloud business. How to manage encrypted data storage with deduplication in an efficient way is a practical issue. However, current industrial deduplication solutions cannot handle encrypted data. Existing solutions for deduplication suffer from brute-force attacks[3][1] . They cannot flexibly support data access control and revocation at the same time[4] . Most existing solutions cannot ensure reliability, security and privacy with sound performance.In practice, it is hard to allow data holders to manage deduplication due to a number of reasons. First, data holders may not be always online or available for such a management,which could cause storage delay. Second, deduplication could become too complicated in terms of communications and computations to involve data holders into deduplication process.Third, it may intrude the privacy of data holders in the process of discovering duplicated data. Forth, a data holder may have no idea how to issue data access rights or deduplication keys to a user in some situations when it does not know other data holders due to data super-distribution. Therefore, CSP cannot cooperate with data holders on data storage deduplication in many situations.

LITERATURE SURVEY

2.1 Zheng Yan[7], provided proxy re-encryption and ownership challenge to deduplicate encrypted data stored in cloud. The main advantage of these techniques is that users can share data even when they are offline. The main disadvantage of these techniques is that optimization of design is necessary so that CSP functions properly in deduplication management.

2.2 M. Bellare[1], provided DupLESS that provides secure deduplicated storage to resist brute-force attacks.In DupLESS, a group of affiliated clients (e.g., company employees) encrypt their data with the aid of a Key Server (KS) that is separate from a Storage Service (SS).The main advantage of these techniques is that brute force attacks are avoided and clients can encrypt their data with key server which is different from separate storage server. The main drawback of these technique is that flexibility to other data users can not be provided.

2.3 C. Y. Liu[3], a policy-based deduplication proxy scheme was proposed but it did not consider duplicated data management (e.g., deletion and owner management) and did not evaluate scheme performance. The main advantage of this technique is that it establishes trust relation among cloud storage components with policy-based deduplication. The main disadvantage of this technique is that data deletion and owner management is not considered by policy-based deduplication.

2.4 T. Y.Wu,[8] proposed Index Name Servers (INS) to manage not only file storage, data deduplication, optimized node selection, and server load balancing, but also file compression, chunk matching, real-time feedback control, IP information, and busy level index monitoring .The main advantage of this technique is that Index Name Servers algorithm help to reduce workloads of resources and improve the performance of system.INS also handles server load balancing.– The main disadvantage of this technique is that encrypted data cannot be deduplicated.

2.5 C. Fan[9], have provided the system based on the assumption that CSP knows the encryption key of data.Thus it cannot be used in the situation that the CSP cannot be fully trusted by the data holders or owners. The main advantage of these technique is that it support deduplication on plaintext and ciphertext. The main disadvantage of these technique is that it does not support encrypted
data deduplication.

EXISTING SYSTEM

A technique which has been proposed to meet conflicting requirements is convergent encryption whereby the encryption key is usually the result of the hash of the data segment.Although convergent encryption seems to be a good candidate to achieve confidentiality and deduplication at the same time, it unfortunately suffers from various well-known weaknesses including dictionary attacks: an attacker who is able to guess or predict a file can easily derive the potential encryption key and verify whether the file is already stored at the cloud storage provider or not. In our project, we cope with the inherent security exposures of convergent encryption and propose ClouDedup, which preserves the combined advantages of deduplication and convergent encryption. The security of ClouDedup relies on its new architecture whereby in addition to the basic storage provider, a metadata manager and an additional server are defined: the server adds an additional encryption layer to prevent well-known attacks against convergent encryption and thus protect the confidentiality of the data; on the other hand, the metadata manager is responsible of the key management task since block-level deduplication requires the memorization of a huge number of keys. Therefore, the underlying deduplication is performed at block-level and we define an efficient key management mechanism to avoid users to store one key per block.

SYSTEM ARCHITECTURE

In our project, we cope with the inherent security exposures of convergent encryption and propose ClouDedup, which preserves the combined advantages of deduplication and convergent encryption. The security of ClouDedup relies on its new architecture whereby in addition to the basic storage provider, a metadata manager and an

additional server are defined: the server adds an additional encryption layer to prevent well-known attacks against convergent encryption and thus protect the confidentiality of the data; on the other hand, the

metadata manager is responsible of the key management task since block-level deduplication requires the memorization of a huge number of keys. Therefore, the underlying deduplication is performed at block-level and we define an efficient key management mechanism to avoid users to store one key per block. In this above architecture, there are following working sides. One is the client side and another one is the server side. Using following methods we will acquire the cloud deduplication in a programmatic manner . Data holder is the one who uploads and save its data at CSP .It is possible to havenumber of eligible data holders that could save the same encrypted raw data at CSP.The data holder that produces or creates the file is regarded as data owner.It has higher priority than other normal data holders.CSP offers storage services.It permits information owner to keep any kind of information and can not be fully trusted by users.User fully trusts AP. AP verifies data ownership and handles data deduplication.

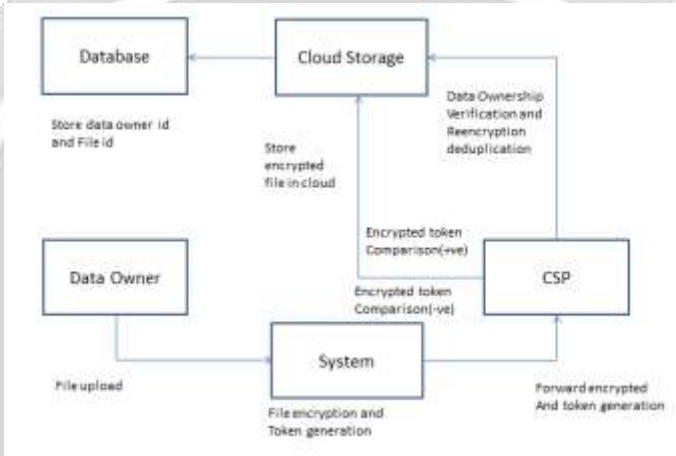

Fig2:System Architecture

Our scheme contains the following main aspects:

1. Encrypted Data Upload:

If data duplication check is negative, the data holder encrypts its data using a randomly selected symmetric key DEK in order to ensure the security and privacy of data, and stores the encrypted data at CSP together with the token used for data duplication check. The data holder encrypts DEK with pk AP and passes the encrypted key to CSP.

2. Data Deduplication.:

Data duplication occurs at the time when data holder u tries to store the same data that has been stored already at CSP. This is checked by CSP through token comparison. If the comparison is positive, CSP contacts AP for deduplication by providing the token and the data holder's PRE public key. The AP challenges data ownership, checks the eligibility of the data holder, and then issues a re-encryption key that can convert the encrypted DEK to a form that can only be decrypted by the eligible data holder.

3. Data Deletion:

When the data holder deletes data from CSP, CSP firstly manages the records of duplicated data holders by removing the duplication record of this user. If the rest records are not empty, the CSP will not delete the stored encrypted data, but block data access from the holder that requests data deletion. If the rest records are empty, the encrypted data should be removed at CSP.

4. Data Owner Management:

In case that a real data owner uploads the data later than the data holder, the CSP can manage to save the data encrypted by the real data owner at the cloud with the owner generated DEK and later on, AP supports re-encryption of DEK at CSP for eligible data holders.

5. Encrypted Data Update:

In case that DEK is updated by a data owner with DEK 0 and the new encrypted raw data is provided to CSP to replace old storage for the reason of achieving better security, CSP issues the new re-encrypted DEK 0 to all data holders with the support of AP.

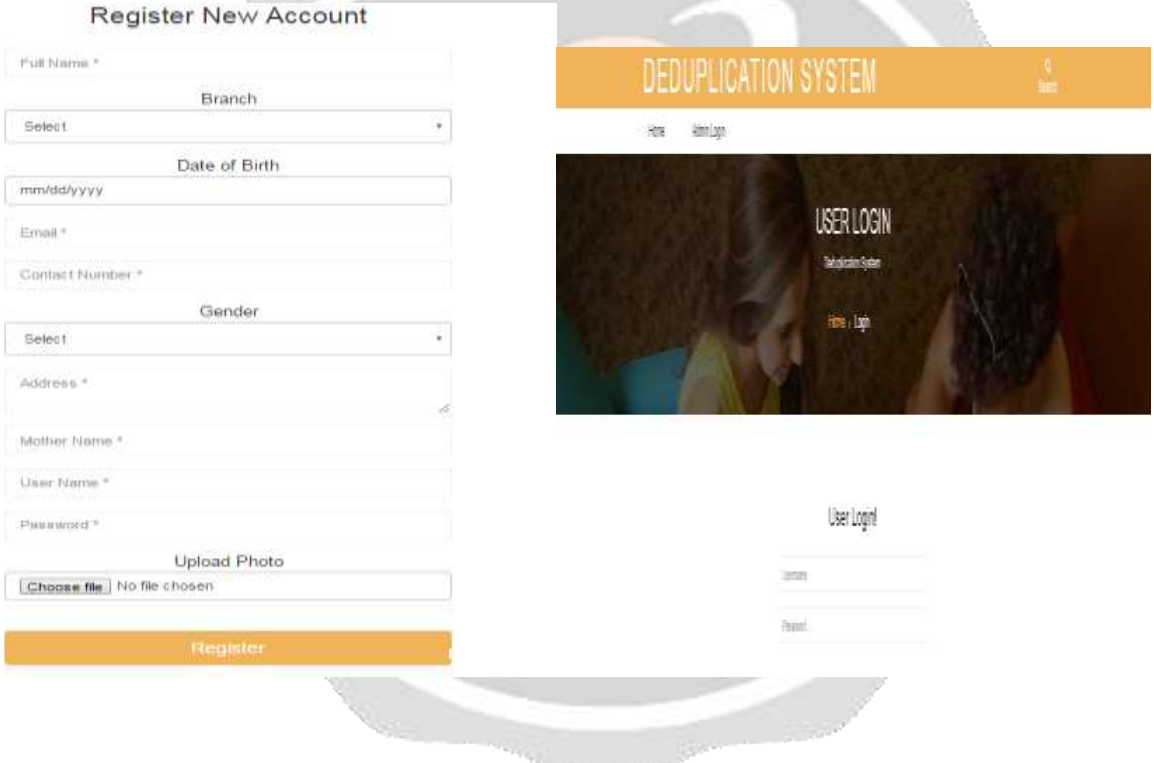

Fig.2. Registration and Login System

CONCLUSIONS

We have proposed a new system called deduplication which saves storage space of cloud enabling to reduce the data duplication by saving only single copy of data for multiple users and also provides security mechanism with the help of AES algorithm and SHA1 algorithm for hash key generation.

REFERENCES

[1]M. Bellare, S. Keelveedhi, and T. Ristenpart, "DupLESS: Server aid encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Secur., 2013, pp. 179–194.

[2] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: A survey," J. Big Data, vol. 2, no. 1, pp. 1–32, 2015, doi:10.1186/s40537-015-0030-3.

[3] C. Y. Liu ,X. J. Liu, and L. Wan, "Policy-based deduplication in secure cloud storage," in Proc. Trustworthy Comput. Serv., 2013, pp. 250–262, doi:10.1007/978-3-642-35795-4_32.

[4] P. Puzio, R. Molva, M. Onen, and S. Loureiro, "ClouDedup: Secure deduplication with encrypted data for cloud storage," in Proc. IEEE Int. Cof. Cloud Comput. Technol. Sci., 2013, pp. 363–370,doi:10.1109/CloudCom.2013.54.

[5] Opendedup. (2016). [Online]. Available: http://opendedup.org/

[6] T. T. Wu, W. C. Dou, C. H. Hu, and J. J. Chen, "Service mining for trusted service composition in cross-cloud environment," IEEE Systems Syst. J., vol. PP, no. 99, pp. 1–12, 2014, doi:10.1109/ JSYST.2014.2361841.

[7] Zheng Yan, Wenxiu Ding, Xixun Yu,Haiqi Zhu,and ROBRT H.Deng, Deduplication on Encrypted Big data in Cloud, IEEESyst.J.,vol.2,no.2,pp.138,Apr-Jun.2016.

[8] T. Y.Wu, J. S. Pan, and C. F. Lin, Improving accessing efficiency of cloud storage using de-duplication and feedbackschemes,IEEESyst.J.,vol.8.

[9] C. Fan, S. Y. Huang, andW. C. Hsu, Hybrid data deduplication in cloud environment, in Proc. Int. Conf. Inf. Secur. Intell. Control,2012, pp. 174177.