A POLICY BASED DEDUPLICATION OF ENCRYPTED BIG DATA IN CLOUD STORAGE

[1] Chethana C, [2] K. Thippeswamy
[1]Student, [2] Professor and Chairman
[1] Department of Computer Science and Engineering
[1]Visvesvaraya Technological University Department of PG Studies, Regional Office
Mysuru, Karnataka, India

Abstract : The number of users continuous and exponential increase of the size of their data, data deduplication becomes more and more a necessarily for cloud storage providers. Cloud computing offers a new way of service provision by re-arranging kinds of resources over the Internet. However, encrypted data introduce new challenges for cloud data deduplication, which becomes crucial for big data storage and processing in cloud. In order to store the privacy of data holders, data are often stored in cloud in an encrypted form. Traditional data deduplication schemes cannot work on encrypted data. Existing solutions of encrypted data deduplication suffer from security weakness. They cannot flexibly support data control and revocation. Therefore, few of them can be readily deployed in practice. In this paper, we propose a policy based deduplication encrypted data stored in cloud based on ownership challenge and proxy re-encryption. It integrates cloud data deduplication with access control. The results show the superior efficiency and performance effectiveness of the policy for potential practical deployment, especially for the big data deduplication in cloud storage and also we show that the overhead introduced by these new components is minimal and does not impact the overall storage and computational costs.

Index Terms- Big data; access control; cloud storage; cloud computing; data deduplication; convergent Encryption; proxy re-encryption.

I. INTRODUCTION

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). Cloud computing offers a new way of Information Technology services by rearranging various resources 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-per-use. Therefore 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. The rapid development of data mining and other analysis technologies, the privacy issue becomes serious. 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.

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 data holders may not be always online or available for such a management, which could cause storage delay. Deduplication could become too complicated in terms of communications and computations to involve data holders into deduplication process it may interrupt the privacy of data holders in the process of discovering duplicated data. 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. Deduplication has proved to achieve high cost 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 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. They cannot flexibly support data access control and revocation at the same time.

In this paper, we propose a scheme based on data owner-ship challenge and Proxy Re-Encryption (PRE) to manage encrypted data storage with deduplication. We aim to solve the issue of deduplication in the situation where the data holder is not available or difficult to get involved. Meantime, the performance of data deduplication in our policy is not influenced by the size of data, thus applicable for big data.

II. RELATED WORK

Zheng Yan[1], 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.

M. Bellare[2], 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.

C.Y. Liu[5], 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.

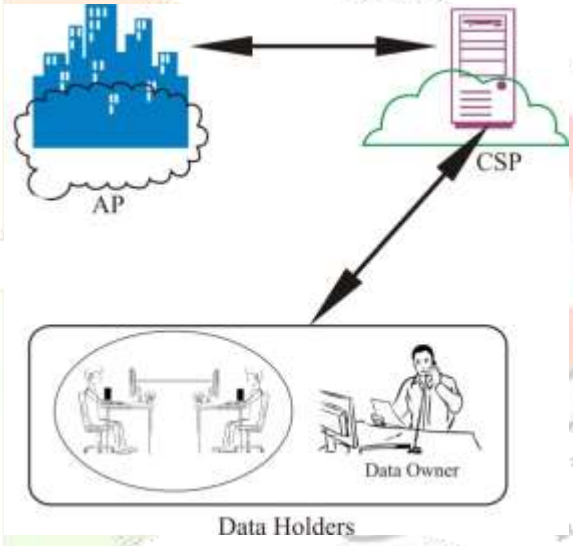

Figure 2. System Model

C. Fan[7], 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.

T.Y. Wu[9], 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.

Deyan Chen, Hong Zhao[10] discusses the paper that it is well-known that cloud computing has many potential advantages and many enterprise applications and data are migrating to public or hybrid cloud. But regarding some business-critical applications, the organizations, especially large enterprises, still wouldn't move them to cloud. The market size the cloud computing shared is still far behind the one expected. From the consumers' perspective, cloud computing security concerns, especially data security and privacy protection issues, remain the primary inhibitor for adoption of cloud computing services.

III. MOTIVATION

The contributions of this paper can be summarized as below :

- We motivate to save cloud storage and preserve the privacy of data holders by proposing a scheme to manage encrypted data storage with deduplication. Our scheme can flexibly support data sharing with deduplication even when the data holder is offline, and it does not intrude the privacy of data holders.
- We propose an effective approach to verify data ownership and check duplicate storage with data access control in a simple way, thus reconciling data deduplication and encryption.
- We prove the security and assess the performance of the proposed scheme through analysis and simulation. The results show its efficiency, effectiveness and applicability.

IV. EXISTING SYSTEM

Cloud users upload personal or confidential data to the data center of the cloud service provider (CSP). CSP cannot be fully trusted by cloud users. The loss of control over own personal data leads to high data security risks, especially data privacy leakages. Deduplication becomes critical for big data storage and processing in the cloud. Deduplication has proved to achieve high cost savings, reducing up to 91-96% storage needs for backup application and up to 68% in standard file system.

To make data management scalable in cloud computing, deduplication has been a well-known technique and has attracted more and more attention recently. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Convergent encryption has been proposed to enforce data confidentiality while making deduplication feasible. It encrypts/decrypts a data copy with a convergent key, which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the ciphertext to the cloud. Since the encryption operation is deterministic and is derived from the data content, identical data copies will generate the same convergent key and hence the same ciphertext.

V. SYSTEM ARCHITECTURE

System architecture is the conceptual model that defines the structure, behavior and more views of a system. An architecture description in figure 2 is a formal description and representation of a system, organized in a way that supports reasoning about the structures of the system which comprise system components, the externally visible properties of those components, the relationships between them and a plan from which products can be procured, and systems developed, that will work together the overall system. The system architecture of the project is shown below.

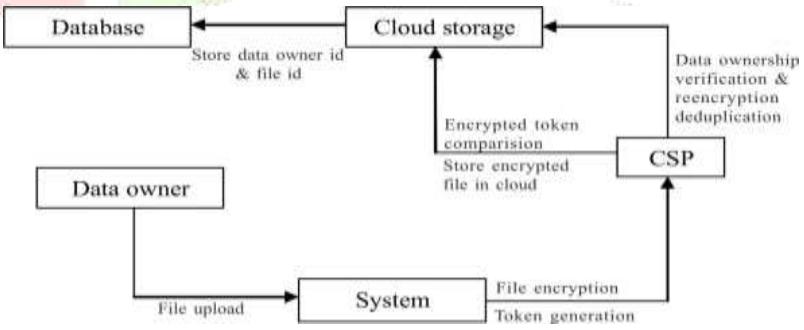

Figure 5. System Architecture

Convergent encryption has been used to enforce data confidentiality. Data copy is encrypted under a key derived by hashing the data itself. This convergent key is used for encrypt and decrypt a data copy. Furthermore, such unauthorized users cannot decrypt the cipher text even collude with the S-CSP. Security analysis demonstrates that system is secure in terms of the definitions specified in the proposed security model.

System architecture is primarily concerned with the internal interfaces among the system's components such as admin, owner, cloud server and user, and the interfaces such as registration, encryption of key, key modification and requesting data between the system and its external environment, especially the user.

VI. CONCLUSION

The notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. In this project we perform several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. As a proof of concept in this project we implement a prototype of our proposed authorized duplicate check scheme and conduct test bed experiments on our prototype. From this project we show that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

VII. Results and Outcome

Managing encrypted data with deduplication is important and significant in practice for achieving a successful cloud storage service, especially for big data storage. This scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only authorized data holders can obtain the symmetric keys used for data decryption. Extensive performance analysis and test showed that our scheme is secure and efficient under the described security Model and very suitable for big data.

VIII. ACKNOWLEDGMENT

The author would like to thank Dr. K. Thippeswamy, Professor and Chairman, Dept. of Studies in Computer Science and Engineering VTU Regional Office, Mysuru and anonymous reviewers encouragement and constructive piece of advice that of prompted us for new round of rethinking of our research, additional experiments and clearer presentation of technical content.

REFERENCES

[1 Zheng Yan, Wenziu Ding, Xixum Yu, Haiqi Zhu, Robert H. Deng, "Deduplication on Encrypted Big Data in Cloud", IEEE Transactions on Big Data.

[2] M. Bellare, C. Namprempre, and G. Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology.

[3]M. Bellare and A. Palacio. Gq and schnorr dentification schemes: Proofs of security against impersonation under active and concurrentattacks.

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

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

[6] Zheng Yan, Wenxiu Ding, Xixun Yu, Zhu, and ROBRT H. Deng, Deduplication on Encrypted Big data in Cloud, IEEE Syst. J., Vol. 2, No.2, pp.138, Apr-Jun. 2016.

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

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

[9] T.Y. Wu, J.S. Pan, and C.F. Lin, Improving accessing efficiency of cloud storage using de-duplication and feed back schemes, IEEE Syst. J., Vol. 8.

[10] Deyan Chen, Hong Zhao, Youngjoo Shin, "Secure data Deduplication With dynamic ownership management in cloud storage", IEEE transactions on knowledge and Data Engineering.