HYBRID CLOUD APPROACH FOR AUTHORIZED DATA DEDUPLICATION

Arpitha.D[*],Dhanya.S.K[*],Krithika.P.A[*],Kruthika.G.S[*],Swarnalatha.K[#]
Student, Associate Professor[#]*
Department of Computer Science and Engineering,
GSSS Institute of Engineering and Technology for Women, Mysuru, India

Abstract – Data deduplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. To protect the confidentiality of sensitive data while supporting deduplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this paper makes the first attempt to formally address the problem of authorized data deduplication. The differential privileges of users are further considered in duplicate check besides the data itself. We also present several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture[2].

Keywords – Data deduplication, authorized duplicate check, confidentiality, hybrid cloud.

I. INTRODUCTION

Cloud computing provides shared computer processing resources and data to computers and other devices on demand. Cloud computing and storage solution provide users and enterprises with various capabilities to store and process the data either privately owned or third party data centers that may be located far from the data users. Security can be improved due to centralization of data, increased security focused resources etc. Security is often as good as or better than the other traditional system. Cloud computing provides seemingly unlimited "virtualized" resources to users as services across the whole Internet, while hiding platform and implementation details. Today's cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified privileges, which define the access rights of the stored data. One critical challenge of cloud storage services is the management of the ever-increasing volume of data. 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.

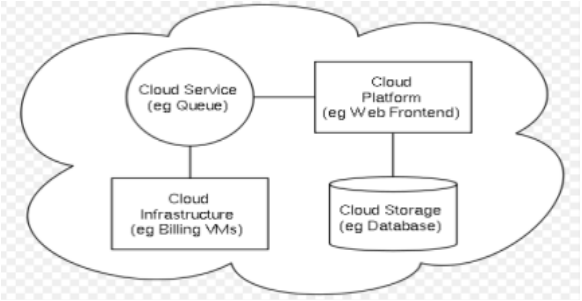

Figure 1.1: Architecture of Cloud Computing[3].

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. Deduplication can take place at either the file level or the block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files[4][6].

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 cipher text 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 cipher text. To prevent unauthorized access, a secure proof of ownership protocol is also needed to provide the proof that the user indeed owns the same file when a duplicate is found. After the proof, subsequent users with the same file will be provided a pointer from the server without needing to upload the same file. A user can download the encrypted file with the pointer from the server, which can only be decrypted by the corresponding data owners with their convergent keys. Thus, convergent encryption allows the cloud to perform deduplication on the cipher texts and the proof of ownership prevents the unauthorized user to access the file.

Previous deduplication systems cannot support differential authorization duplicate check, which is important in many applications. In such an authorized deduplication system, each user is issued a set of privileges during system initialization. Each file uploaded to the cloud is also bounded by a set of privileges to specify which kind of

users is allowed to perform the duplicate check and access the files. Before submitting his duplicate check request for some file, the user needs to take this file and his own privileges as inputs. The user is able to find a duplicate for this file if and only if there is a copy of this file and a matched privilege stored in cloud. For example, in a company, many different privileges will be assigned to employees. In order to save cost and efficiently management, the data will be moved to the storage server provider (SCSP) in the public cloud with specified privileges and the deduplication technique will be applied to store only one copy of the same file. Because of privacy consideration, some files will be encrypted and allowed the duplicate check by employees with specified privileges to realize the access control[5][9].

Traditional deduplication systems based on convergent encryption, although providing confidentiality to some extent; do not support the duplicate check with differential privileges. In other words, no differential privileges have been considered in the deduplication based on convergent encryption technique. It seems to be contradicted if we want to realize both deduplication and differential authorization duplicate check at the same time[2].

II. RELATED WORK

"A Hybrid Cloud Approach for Secure Authorized De-duplication" Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou [2] have proposed how Data de-duplication is achieved by providing the proof of data by the data owner. Address the problem of authorized data deduplication.

"Role-based access controls", D. Ferraiolo and R. Kuhn [6], has described the Mandatory Access Controls (MAC) are appropriate for multilevel secure military applications, Discretionary Access Controls (DAC) are often perceived as meeting the security processing needs of industry and civilian government.

"Secure deduplication with efficient and reliable convergent key management", J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou [3], had proposed Dekey, an efficient and reliable convergent key management scheme for secure de-duplication. They implement Dekey using the Ramp secret sharing scheme and demonstrate that it incurs small encoding/decoding overhead compared to the network transmission overhead in the regular upload/download operations.

"Reclaiming space from duplicate files in a server less distributed file system", J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. [4], has presented the Farsite distributed file system provides availability by replicating each file onto multiple desktop computers. Measurement of over 500 desktop file systems shows that nearly half of all consumed space is occupied by duplicate files.

"A secure data deduplication scheme for cloud storage", J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl [5], has provided the private users outsource their data to cloud storage providers, recent data breach incidents make end-to-end encryption an increasingly prominent requirement data deduplication can be effective for popular data, whilst semantically secure encryption protects unpopular content.

III. METHODOLOGY

The basic objective of this work is the problem of privacy preserving deduplication in cloud computing and a proposed System focus on these aspects:

1. Differential Authorization**:** Each authorized user is able to get his/her individual token of his file to perform duplicate check based on his privileges.
2. Authorized Duplicate Check: Authorized user is able to use his/her individual private keys to generate query for certain file and the privileges he/she owned with the help of private cloud, while the public cloud performs duplicate check directly and tells the user if there is any duplicate.

i. Proposed Work

In Proposed system, 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 (storage cloud service provider). Security analysis demonstrates that that system is secure in terms of the definitions specified in the proposed security model.

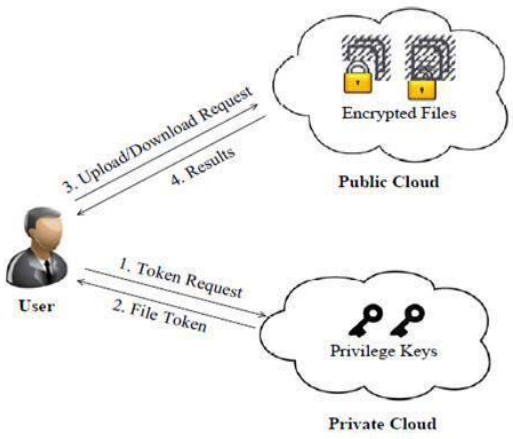

Figure 1.2: Architecture for Authorized Deduplication[2].

There are three entities define in hybrid cloud architecture of authorized deduplication.

☐**Data Users:** A user is an entity that wants to outsource data storage to the S-CSP (storage cloud service provider) and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.

☐ **Private Cloud:** This is new entity for facilitating users secure use of cloud services. The private keys for privileges are managed by private cloud, which provides the file token to users. Specifically, since the computing resources at data user/owner side are restricted and the public cloud is not fully trusted in practice, private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user and the public cloud.

☐ **S-CSP (storage cloud service provider):** This is an entity that provides a data storage service in public cloud. The S-CSP provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data. In this paper, we assume that S-CSP is always online and has abundant storage capacity and computation power[7][2].

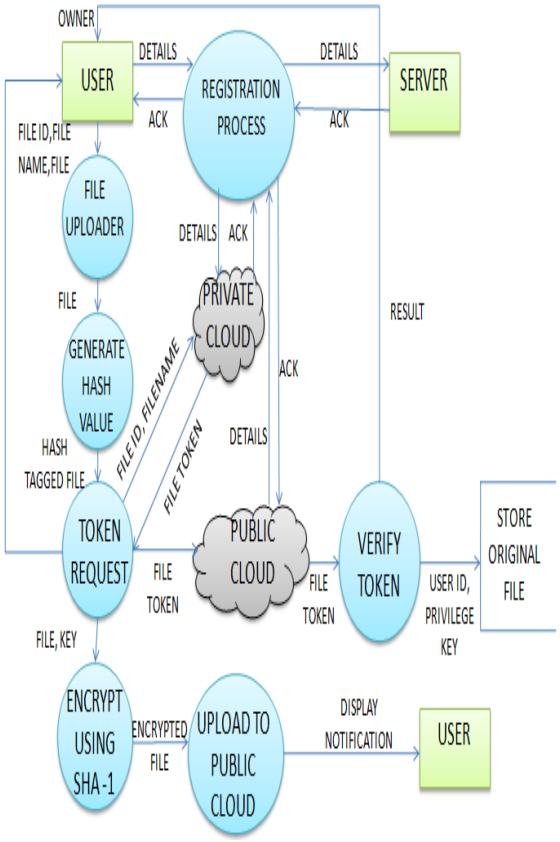

Figure 1.3: Data flow diagram of the proposed work.

ii. SHA1 Algorithm Description

In the proposed system convergent key for each file is generated by using secure hashing algorithm-1 the steps of this algorithm is given below:

Step 1: Padding.

- Pad the message with a single one followed by zeroes until the final block has 448 bits.
- Append the size of the original message as an unsigned 64 bit integer.

Step 2: Initialize the 5 hash blocks (h0, h1, h2, h3, h4) to the specific constants defined in the SHA1 standard.

Step 3: Hash (for each 512bit Block)

- Allocate an 80 word array for the message schedule
- Set the first 16 words to be the 512bit block split into 16 words.
- The rest of the words are generated using the following algorithm

Step 4: word [i3] XOR word [i8] XOR word [i14] XOR word [i16] then rotated 1 bit to the left.

- Loop 80 times doing the following.
- Calculate SHA function() (it calculates the hash value of the string that is given as input) and the constant K (these are based on the current round number).
- e=d
- d=c
- c=b (rotated left 30)
- b=a
- a = a (rotated left 5) + SHAfunction() + e + k + word[i]
- Add a,b,c,d and e to the hash output.

Step5: Output the concatenation (h0,h1,h2,h3,h4) which is the message digest[11].

iii. Requirements

Functional Requirements:

Software Requirements

- Operating system: Windows
- Technology :Java
- Eclipse neon 2
- JDK 8

Hardware Requirements

- Processor: Intel core i3
- RAM: 2GB minimum

Non Functional Requirements:

- Security
- Adaptability
- Performance
- Scalability
- Manageability
- Capacity
- Availability

iv. Installation Details

For implementing this project eclipse neon 2 and JDK 8 software is required.

Design tools: Back end is designed using the eclipse and frond end is designed using HTML, CSS, Java script and JQuery. The database that is used to store all the uploaded documents/files is PostgreSQL.

v. Flow Chart

The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users and this interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively. The server takes the details from private and public could and acknowledges. The details are verified with the database[12][14].

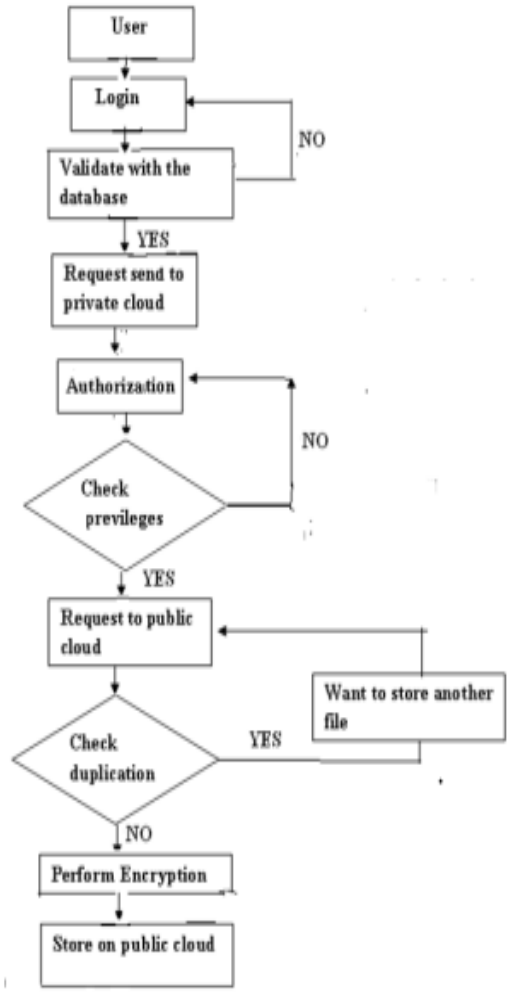

Figure 1.3: Flow Diagram of the proposed work.

In deduplication system, hybrid cloud architecture is introduced to solve the problem of unauthorized deduplication of file. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server. The user needs to send a request to the private cloud server to get a file token. The user needs to get the file token from the private cloud server to perform the duplicate check for some file. The user either uploads this file or proves their ownership based on the results of duplicate check. If it is passed, the private cloud server will find the corresponding privileges of the user from its stored table list and send to the user then user can upload his files. The same way user can download his file from storage cloud[10][5].

IV. CONCLUSION AND FUTURE ENHANCEMENT

In this Project, 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 testbed 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.

Futures work: It increases the national security. It saves the memory by deduplicating the data and thus provides us with sufficient memory. It provides authorization to the private firms and protects the confidentiality of the important data

REFERENCES

[1] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS,Apr 2013.

[2] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.

[3] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure datadeduplication scheme for cloud storage. In Technical Report, 2013.

[4] J. Xu, E.-C. Chang, and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In ASIACCS,pages 195–206, 2013.

[5] J. Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication. IACR Cryptology ePrint Archive, 2013:149, 2013.

[6] K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan. Sedic: privacy aware data intensive computing on hybrid clouds. In Proceedings of the 18th ACM conference on Computer and communications security,CCS'11, pages 515–526, New York, NY, USA, 2011. ACM.

[7] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013.

[8] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312, 2013.

[9] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. Secure data deduplication. In Proc. of StorageSS, 2008.

[10] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.

[11] R. D. Pietro and A. Sorniotti. Boosting efficiency and security in proof of ownership for deduplication. In H. Y. Youm and Y. Won, editors, ACM Symposium on Information, Computer and Communications Security, pages 81–82. ACM, 2012.

[12] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC2011), 2011.

[13] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.

[14] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441– 446. ACM, 2012.