Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

# Data Deduplication: A Technique for Efficient Storage in Cloud

Aparna Shetty[1], S.Girish[2]
[1]4th sem M.Tech, Dept of CS&E, Sahyadri College of Engineering and Management, Adyar, Mangalore, India
[2]Assistant Professor, Dept of CS&E, Sahyadri College of Engineering and Management, Adyar, Mangalore, India

**ABSTRACT:** Cloud Computing allows users to outsource storage and computation to servers using internet. In this paper we propose a data Deduplication technique to reduce the amount of storage space and to save bandwidth. The convergent encryption technique has been proposed to protect the confidentiality of sensitive data. Our scheme has the feature of access which allows only valid users to share the data over the cloud storage. Data Deduplication is not only an efficient means of storage over the cloud but also enforces security over the stored data. We ensure that the authorized duplicate check scheme does not create any overhead. The hybrid cloud approach is used for data storage which is very different from the traditional systems.

**KEYWORDS**: cloud computing; Deduplication; duplicate check; hybrid cloud; convergent encryption

## I. INTRODUCTION

Problem Statement: This paper addresses the following problems: How to improve the ways of storing data in the cloud, which focuses on using the least resource to store the data as well as incurs less overhead. Here hybrid cloud architecture is used, where the user's details and duplicate check is done in the private cloud while the actual data is encrypted and stored in public cloud.

Cloud computing is receiving a lot of attention from both academic and industrial worlds. Users using the internet can outsource both computation and storage to the cloud. Cloud provides unlimited resources to the users, while hiding the details of platform and implementation. Today a huge amount of data is stored in the cloud and shared among users. Management of data stored in the cloud is thus a very important task. Along with data management security also plays a very important task. Most of the data stored in the cloud is sensitive data, for example, social networks and personal records. Security and privacy are the most important issues in cloud computing. On one end user authentication has to be done before initiating any transaction and on the other end the data must be protected from the cloud.

Data Management becomes scalable in cloud computing through a well known technique called Deduplication. Deduplication eliminates redundancy by eliminating multiple copies of the same data and just storing one physical copy. This technique improves storage utilization and can be applied to data transfer over the network also, which reduces the number of bytes sent over a network. With Deduplication there is no need to maintain multiple copies of the same data by keeping only one physical copy and referring to other redundant data to that single copy.

The possibility of failure of the storage servers in arbitrary ways is very high. The cloud servers are also prone to modification of data and also server colluding attacks. The adversary can compromise storage servers so that they can tamper the data files as long as the files are internally consistent.

 

Hence data modification should be taken into account when designing efficient storage techniques. Homomorphic encryption technique ensures that the cloud is not able to read the data while performing computations on them. The cloud doesn't know the data it is operated on while the users can decode the result.

## II. RELATED WORK

Even though there are a lot of benefits from data Deduplication, security and privacy are the two important concerns as user's sensitive data are stored in the cloud. With traditional encryption techniques each user requires to encrypt the data with his/her own keys. Here similar data copies of different users when encrypted with their own keys leads to different ciphertext, making Deduplication incompatible. In order to make data Deduplication feasible a convergent encryption technique [1] has been proposed. Here the convergent key is used; this is obtained by calculating cryptographic hash value of the data copy. Since the encryption key is calculated from the data content itself similar data copies will produce the same convergent key. When a duplicate is found in order to prove that a file will be owned by another user also a proof of ownership protocol [2] is used. According to this protocol (POW) a user with a file which is already uploaded, will be provided a pointer from the server. So there is no need to upload the same file again. The traditional systems even though provided data confidentiality did differential authorization duplicate check. Here a user is provided with certain privileges, which defines his/her access to the stored data.

*Convergent encryption*: Data confidentiality is provided through Convergent Encryption [1] [3]. A user encrypts the data with the convergent key which is derived from the data itself. In addition each data copy is also assigned a tag, which is used to detect duplicates. Data Deduplication can take place at either block level or file level. [4]. Here we consider only file level Deduplication which eliminates duplicate files. The following four primitives:

- The key generation algorithm which derives a convergent key k from the data copy M.
- The symmetric encryption algorithm that outputs the cipher text C taking the data copy M and convergent key K as inputs.
- The decryption algorithm that outputs the data copy M taking inputs as cipher text C and key K.
- Tag generation algorithm that outputs a tag for the original data copy.

*Proof of ownership*: The POW [2] algorithm here the verifier generates the token for the data copy. The proover needs to submit the same token which the verifier had initially calculated..

*Identification Protocol*: The identification protocol is used to ensure that only valid users are used to access the data. Here the user's needs to be registered initially. Then he needs to be approved by the admin. Then each time he wants to access the data he need to submit his credentials that is username and password. He can login successfully only on providing the right username along with the password [5][6].

Current Works which are based on access control in the cloud are centralized [7] [8] [9] [10] [11] [12] [13].The attribute based encryption (ABE) is used in [9] [10] [11] [12] [13]. Authentication is not supported in [7] [9] [10] [13]. Work by Zhao et al. [12] proposed a access control scheme that provides privacy along with user authentication .In centralized approach there is only one key distribution centre (KDC) where all the user credentials are stored. A decentralized approach was proposed by yang et al. [14] but doesn't ensure the anonymity property while authenticating users. We extend the previous works with additional features that ensures authentication of users without revealing their identity that is anonymous authentication. A distributed access scheme proposed by ruj et al. [15] but this scheme failed to authenticate users. The other disadvantage was that it provided only read access to other users.

 

A. *Our Contributions*

A Hybrid cloud architecture comprising of a public cloud and private cloud is used, targeting the problem of Deduplication which considers the different privileges for accessing the cloud.

A private cloud is used as a proxy that allows users with different privileges to perform the duplicate check. In order to provide a strong security the file is encrypted with the different privileges of users.

Decentralized access of data stored in the cloud so that only valid users can access the data. In the decentralized approach key management is done at multiple KDC's. Anonymous authentication doesn't disclose the user's identity while accessing the cloud.

## III. SYSTEM MODEL

A. *Architecture for secure authorized Deduplication:*

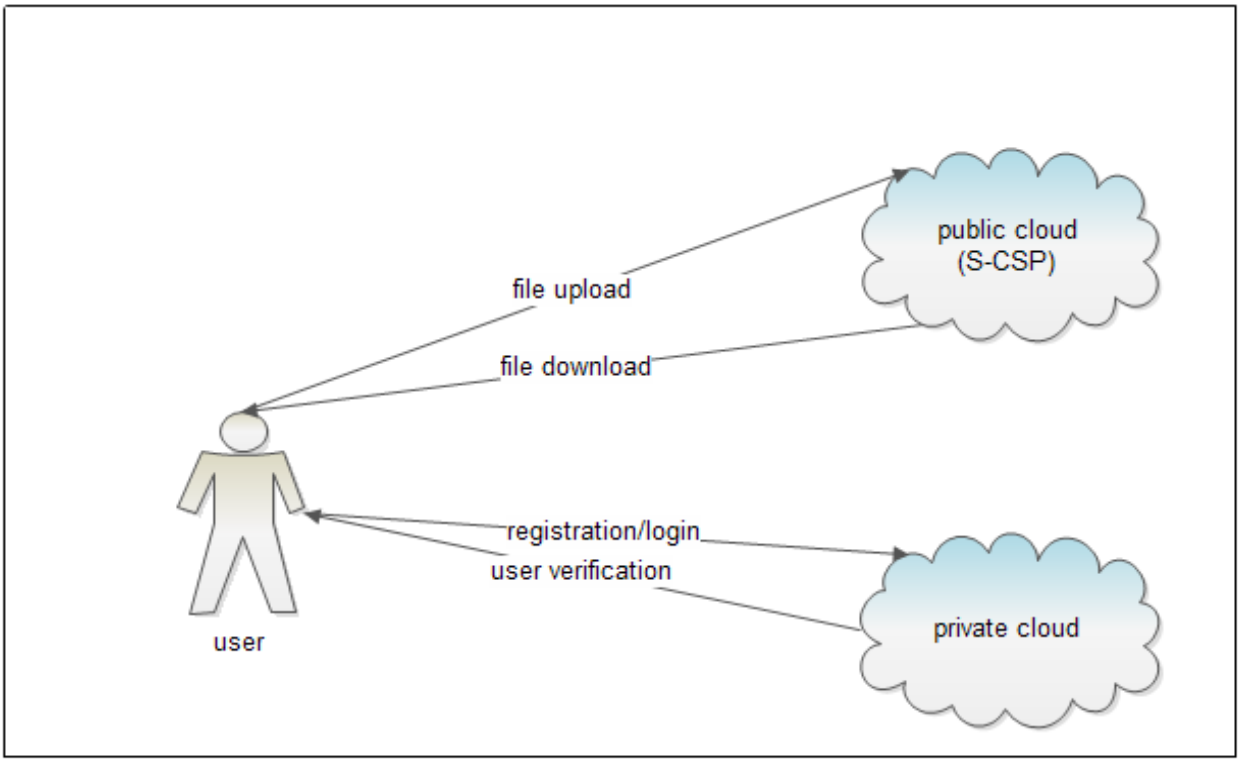


Fig 1. Architecture for secure authorized deduplication

The high level of interest is a enterprise network (for example company employees) who uses the Deduplication technique in order to store their data in the cloud. Data Deduplication greatly reducing the storage space also provides data backup services along with disaster recovery .The architecture shown in Fig. 1 includes the users, the private cloud and the public cloud. Deduplication is performed in the public cloud by comparing if two files have the same content, if true it stores only one copy of the file. Each file is attached a file token, each user needs to compute and send the token to perform the duplicate check. The private cloud is intermediate of users and public cloud. The private cloud is used to verify each user before accessing the public cloud for data storage. All the user credentials are stored in the private cloud.

- *S-CSP*: Storage-cloud service provider entity which provides the public cloud service that is storage as a service. Using this service only unique data copies are stored in the cloud eliminating redundant data. S-CSP always has a huge amount of storage capacity and a high computation power.

- *Users of Data*: Any entity who wishes to outsource their data storage to the cloud. A user also can share their data with the particular group of users. In order to save the storage bandwidth the user

only uploads unique copy of the files. Each user has a username and password for secure authorization.

- *Private Cloud*: Private cloud is introduced for users to securely access the cloud storage. User verification during login is performed at the private cloud. All the user credentials are stored in the private cloud.

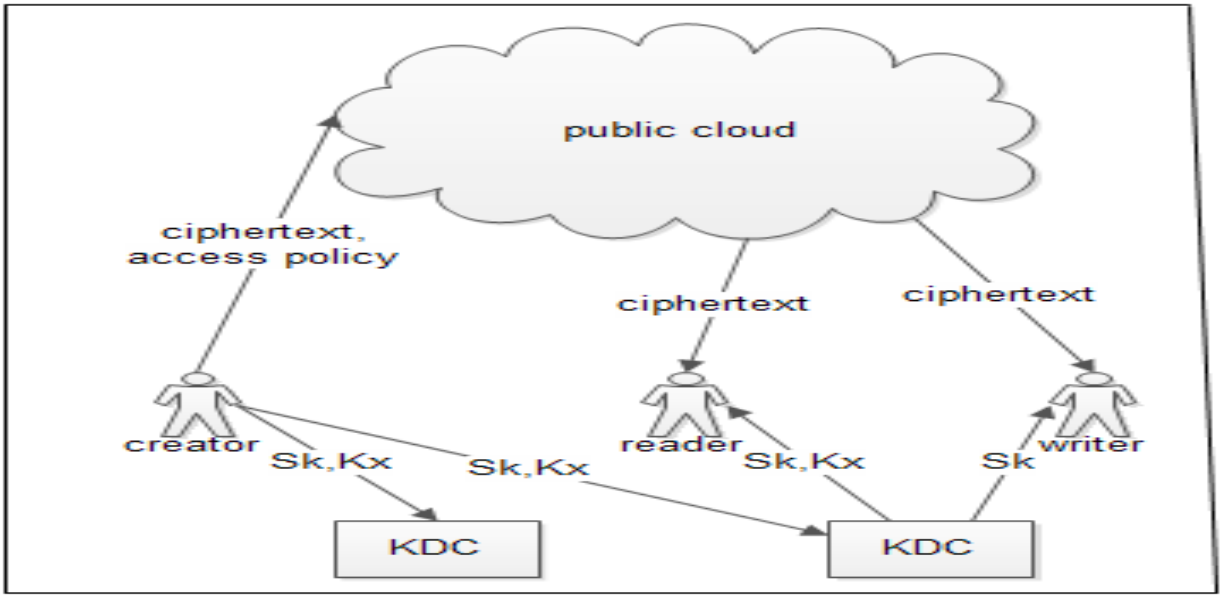


Fig 2. Architecture for decentralized access control

*B: Architecture for Decentralized access control:*

The architecture shown in Fig. 2 represents three users a creator, reader and writer. Initially a creator submits his/her id to the trustee. On submitting the id, the trustee returns a token. A trustee is assumed to be honest. There are multiple key distribution centers (KDC). The creator on submitting the token to KDC's receives encryption/decryption keys and keys for signing. Here the file to be uploaded is encrypted using access policy. Access policy indicates who has the right to access the data. SK is the secret key used to decrypt the data. The key for signing is represented by Kx. The ciphertext is sent to the cloud, along with the signature. When a reader wishes to read the data he gets the ciphertext from the cloud. He is able to decrypt the ciphertext only if his attributes matches according to the access policy. Write needs to perform the same steps as the creator. Here the user is freed from the time consuming verifications, which is performed in the cloud itself.

## IV. PROPOSED METHODOLOGY

We use hash functions for generating encryption key and tokens. The hash functions generate the convergent key used for encryption based on the contents of the file. So verified users can decrypt the file only if they have the convergent key which is computed based on the contents of the file. The duplicate check token is also computed based on the contents of the file. So users who own the file only can compute the duplicate check token. Then submit the token to the S-CSP to perform the duplicate check.

*System Initialization*: The user who wants to use the system needs to be registered. Only after a user is activated by the admin he can access the cloud storage. Only users who have registered and verified can access system. User needs to perform a secure login each time he/she enters the system. All the user details are stored in the private cloud. This performs the verification when the user tries to enter into the system.

*File Uploading*: If a user wants to upload a file and share it with a group of users. He needs to perform the duplicate check initially, by submitting the file token. If the token matches with the file

 

tag he need not upload the file again. Here he will be assigned a pointer from the server. Hence he will be referring to the file which is already uploaded. Otherwise if no duplicate is found. He needs to encrypt the file using convergent encryption technique. After encryption he needs to upload the file to the cloud.

*File Downloading*: The user needs to send a request to S-CSP. The S-CSP verifies if the user. After user verification, if he/she is a valid user, it sends the encrypted file to the cloud. The user can decrypt the file only on submitting the correct key

## V. CONCLUSION

Data Deduplication makes storage over the cloud very efficient. It not only reduces a huge amount of storage space but also provides a means of secure data storage over the cloud. The duplicate check by submitting the file token allows only authorized users to check for duplicates. Decentralized access control prevents the problem of single point failure, since the secret keys are stored in multiple KDC's. The cloud performs user verification but doesn't identify users through anonymous authentication.

## Acknowledgment

We sincerely thank all those who have guided and assisted us in coming up with the project. Our sincere gratitude to the organization, Sahyadri College of Engineering and Management for all the facilities and support in completing the project as a part of the curriculum of Visvesvaraya Technological University, Belagavi.

## REFERENCES

[1] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer.
Reclaiming space from duplicate files in a serverless distributed file system. In ICDCS pages 617–624, 2002.
[2] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs ofownership 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.
[3] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312, 2013.
[4] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. Secure deduplication
with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.
[5] M. Bellare, C. Namprempre, and G. Neven. Security proofs for identity-based identification and signature schemes. J . Cryptology ,22(1):1–61, 2009.
[6] M. Bellare and A. Palacio. Gq and schnorr identification schemes:Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162–177, 2002.
[7] W. Wang, Z. Li, R. Owens, and B. Bhargava, "Secure and efficient access to outsourced data," in ACM Cloud Computing Security Workshop (CCSW) 2009.
[8] http://securesoftwaredev.com/2012/08/20/xacml-in-the-cloud.
[9] M. Li, S. Yu, K. Ren, and W. Lou, "Securing personal health records in cloud computing: Patient-centric and fine-grained data access control in multi-owner settings," in SecureComm, pp. 89–106, 2010.
[10] S. Yu, C. Wang, K. Ren, and W. Lou, "Attribute based data sharing with attribute revocation," in ACM ASIACCS, pp. 261–270, 2010.
[11] G. Wang, Q. Liu, and J. Wu, "Hierarchical attribute-based encryption for fine-grained access control in cloud storage services," in ACM CCS,pp. 735–737, 2010.
[12] F. Zhao, T. Nishide, and K. Sakurai, "Realizing fine-grained and flexible
access control to outsourced data with attribute-based cryptosystems," in ISPEC, ser. Lecture Notes in Computer Science, vol. 6672. Springer, pp.83–97, 2011.
[13] S. Ruj, A. Nayak, and I. Stojmenovic, "DACC: Distributed access control in clouds," in IEEE TrustCom, 2011.
[14] Kan Yang, Xiaohua Jia and Kui Ren, " DAC-MACS: Effective Data Access Control for Multi-Authority Cloud Storage Systems", IACR Cryptology ePrint Archive, 419, 2012.
[15] Jin Li, Xiaofeng Chen, Patric P. C. Lee, Wenjing Lou ,"A Hybrid Cloud Approach for Secure Authorized Deduplication"IEEE transactions on parallel and distributed systems ,VOL:PP NO:99 YEAR 2014.
[16] Sushmith Ruj, Milos Stojmenovic , Amiya Nayak, "Decentralized Access Control with Anonymous Authentication of Data Stored in Clouds" "IEEE transactions on parallel and distributed systems, VOL:25 NO:2 YEAR 2014.
[17] R.Ranjith, D.Kayathri Devi,"Secure Cloud Storage using Decentralized Access Control with Anonymous Authentication", International Journal of Avanced Research in Computer and Communication Engineering, VOL:2,Issue 11, November 2013.