An Efficient Framework for Deduplication in Cloud

Kunal R. Mahajan, Akshay D. Nagawade, Ravina Patil, Deepak Choudhari

***Abstract*—Data deduplication is important method of data compression which removes duplicate data copies over cloud so as to increase the storage space in cloud. To protect the confidentiality of compassionate data while supporting deduplication, encryption technique is used .**
To make data management scalable, deduplication has been a well known technique to condense storage space and upload bandwidth in cloud. Unfortunately, many systems have security problems, wastage of hardware resources and increase the complexity of data centre. To overcome above flaws, proposed system put forward an idea of efficient framework for deduplication by using many concepts in project.

***Index Terms*—Correlation, File Tagging, Reverse Circle Cipher, Subset vector creation.**

I. INTRODUCTION

Cloud computing is type of internet based computing. Cloud Computing provides unlimited virtual resources for user as services across the internet. The cloud computing provides unlimited storage and parallel computing resources at low cost. Now a day's most of the users are using the services of cloud computing, the increase in amount of data is stored over the cloud and shared by user with privileges such as accessing right of the stored data. The management of increasing volume of stored data over cloud is critical challenge in front of the service providers.

Data management is done efficiently by using technique known as Deduplication. Data deduplication is important method of data compression which removes duplicate data copies over cloud so as to increase the storage space in cloud and improve the network bandwidth. In this paper, the goal is at efficiently solving problems of deduplication along with distinctive privileges in cloud computing. Cloud architecture consists of Private cloud and Public cloud. Apart from existing data deduplication systems, the private cloud is included as a proxy to allow data user to perform duplication check over cloud. The user can only check for the duplicate files with privileges. The security efficiency is improved in the proposed system.

The basic idea of this project comes from the fact that cloud is big platform to store and to retrieve the data in huge capacity. Where there is greater possibility of duplication of the data can be happen due to this there will be huge storage space is used unnecessarily.

Instead of storing the multiple copies of same content, deduplication avoids that, by storing only one physical copy and referring other to that copy. Deduplication takes place at either block level or file level. For block level, duplicate blocks of data is eliminated that found in non-identical files while file leveleliminates the duplicate files. Convergent encryption has been proposed to keep user's security and privacy from insider and outsider attacks. Secure proof of ownership protocol is required to prevent from unauthorized access. After getting the proof of ownership with the same file will be provided with reference from server without uploading the duplicate file again. User can download the encrypted file with the reference from server and it can be decrypted only when the convergent key is matched with the owners.

Before uploading a file, the hash key generated using md5 hashing algorithm which is the most secure hashing algorithm. The generated hash key of every single file which is being uploaded on cloud is stored in that table along with the server information. The user can find duplicate file if and only if there exist a copy of same file and matched the key in the cloud.

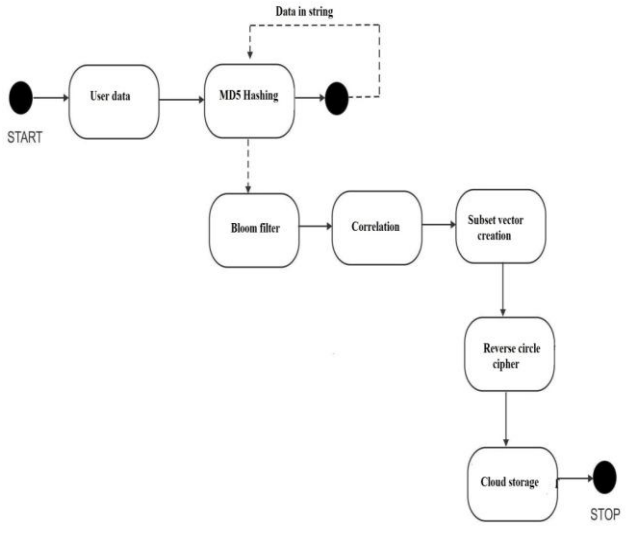
The rest of this paper is organized as follows. Section I illustrates the introduction related to the works and some background information. Section II introduces the proposed methodology. Section III presents literature review which defines the background study of topic. Conclusion offered in Section IV.

.

II. PROPOSED METHODOLOGY

To handle the concept of deduplication there are many methodologies are supporting like Reverse Circle cipher[1], Correlation[2], Inverted index[3], Bloom Filter[4].

The diagrams shows the flow of the proposed method to understand better.



ISSN: 2278 – 1323

The user data is referenced to the file user wants to upload over cloud and the processes are explained below:
[1] Introduces concept of Reverse Circle cipher, which is special encryption algorithm which uses "Circular substitution" and "Reversal Transposition" to utilize the advantages of confusion and diffusion. This scheme uses an random variable key length couple with an arbitrarily reversal cipher. Simple block cipher schemes to effectively reduce time and space complexity and still provide adequate security for both domains both domains of security. For example, algorithm divides the content in blocks and operation takes place in every single block of content. Every word in block is circulated in clockwise and that character is substituted by its ASCII value.
[2] Narrates concept of Correlation, is a statistical technique that can show whether and how strongly pairs of strings are related with each other. To avoid the duplication of names and sentences correlation is used.
[3] Describes concept of Inverted Index, An inverted index is storing a mapping from content, like numbers and words, to its locations in a database file.
[4] Proposed the Subset vector creation creates the set of uploading files which includes filenames along with user id. For example, User A is uploading a file named abc.txt and User B is uploading a file xyz.txt of same content then subset vector creation creates the set as {(abc.txt,A),(xyz.txt,B)} for duplicate files.

III. Literature Review

Deduplication losing it importance in association with end-to-end encryption. End-to-end encryption in a storage system is process by which is encrypted prior to ingress into storage system. It is becoming an increasingly remarkable requirement due to tightening of sector-specific laws and regulations and number of security incidents to leakage of unencrypted data[1]. If semantically secure encryption is used, file deduplication is impossible, as no one apart from the owner of the decryption key can resolve whether two cipher texts correspond to the same plaintext.

Several deduplication strategies have been proposed by the research community [2–4] showing how deduplication allows very appealing reductions in the management of storage resources [5, 6]. Most works do not consider security as a concern for deduplicating systems; recently, Harnik et al. [7] have presented a number of attacks that can lead to data leakage in storage systems of client-side deduplication is in place. To prevent such attacks, proof of ownership concept has been introduced [8, 9]. Convergent encryption is a cryptographic primitive introduced by Douceur et al. [10, 11], attempting to combine data confidentiality with the possibility of data deduplication. Convergent encryption of a message consists of encrypting the plaintext using a symmetric encryption scheme with a key which is deterministically derived solely from the plaintext. When two users individually attempt to encrypt the same file, they will generate the same ciphertext which can be deduplicated easily. Unfortunately, convergent encryption does not provide syntactic security as it is vulnerable to content guessing attack. [12] formalized convergent encryption under the name message locked encryption. The security analysis presented in [12] highlights that message-locked encryption offers confidentiality for unpredictable messages only, clearly failing to achieve syntactic security. Xu et al. [13] present a PoW strategy, for random oracle model they provide a security proof for their solution, presented DupLESS [14], a server-aided encryption for deduplicated storage. Similarly to ours, their solution uses a modified convergent encryption scheme with the aid of a secure component for key generation. are using *Word*, use either the Microsoft Equation Editor or the *MathType* add-on (http://www.mathtype.com) for equations in your paper (Insert | Object | Create New | Microsoft Equation *or* MathType Equation). "Float over text" should *not* be selected.

IV. Conclusion

To overcome the some major disadvantages in deduplication, this paper tries to proposed a method using some best idea which includes correlation,inverted index, subset vector creation and reverse circle cipher. Here in this paper many of the ideas have been analyzed.

REFERENCES

1. Open Security Foundation: DataLossDB (http://datalossdb.org/).
2. Meister, D., Brinkmann, A.: Multi-level comparison of data deduplication in ī backup scenario. In: SYSTOR '09, New York, NY, USA, ACM (2009) 8:1{8:12
3. Mandagere, N., Zhou, P., Smith, M.A., Uttamchandani, S.: Demystifying data deduplication. In: Middleware '08, New York, NY, USA, ACM (2008) 12{17
4. Aronovich, L., Asher, R., Bachmat, E., Bitner, H., Hirsch, M., Klein, S.T.: The design of a similarity based deduplication system. In: SYSTOR '09. (2009) 6:1{6:14
5. Dutch, M., Freeman, L.: Understanding data de-duplication ratios. SNIA forum(2008)
http://www.snia.org/sites/default/files/Understanding_Data_Deduplication_Ratios- 0080718.pdf.
6. Harnik, D., Margalit, O., Naor, D., Sotnikov, D., Vernik, G.: Estimation of deduplication ratios in large data sets. In: IEEE MSST '12. (april 2012) 1 {11
7. Harnik, D., Pinkas, B., Shulman-Peleg, A.: Side channels in cloud services: Deduplication in cloud storage. Security Privacy, IEEE 8(6) (nov.-dec. 2010) 40 {47
8. Halevi, S., Harnik, D., Pinkas, B., Shulman-Peleg, A.: Proofs of ownership in remote storage systems. In: CCS '11, New York, NY, USA, ACM (2011) 491{500
9. Di Pietro, R., Sorniotti, A.: Boosting e_ciency and security in proof of ownership for deduplication. In: ASIACCS '12, New York, NY, USA, ACM (2012) 81{82
10. Douceur, J.R., Adya, A., Bolosky, W.J., Simon, D., Theimer, M.: Reclaiming space from duplicate _les in a serverless distributed _le system. In: ICDCS '02, Washington, DC, USA, IEEE Computer Society (2002) 617{632
11. Storer, M.W., Greenan, K., Long, D.D., Miller, E.L.: Secure data deduplication. In: StorageSS '08, New York, NY, USA, ACM (2008) 1{10
12. Bellare, M., Keelveedhi, S., Ristenpart, T.: Message-locked encryption and secure deduplication. In: Advances in Cryptology{EUROCRYPT 2013. Springer 296{312
13. Xu, J., Chang, E.C., Zhou, J.: Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In: 8th ACM SIGSAC symposium. 195{206
14. Bellare, M., Keelveedhi, S., Ristenpart, T.: DupLESS: server-aided encryption for deduplicated storage. In: 22nd USENIX conference on Security. (2013) 179{194
15. Improving Accessing Efficiency of Cloud Storage Using De-Duplication and Feedback Schemes, Tin-Yu Wu, Jeng Shyang Panand Chia-Fan Lin, IEEE Journal volume 8, March 2014

ISSN: 2278 – 1323

 **Kunal R. Mahajan** is pursuing his Bachelors of Engineering in Computer Sciences from Savitribai Phule Pune University.
.

 **Akshay D. Nagawade** is pursuing his Bachelors of Engineering in Computer Sciences from Savitribai Phule Pune University.

Ravina Patil is pursuing his Bachelors of Engineering in Computer Sciences from Savitribai Phule Pune University.

Deepak Choudhari is pursuing his Bachelors of Engineering in Computer Sciences from Savitribai Phule Pune University.

ISSN: 2278 – 1323