A Review on Data De-Duplication Techniques for Managing Data into Cloud

Barkha Sharma
Central College of Engineering and Management
Dept. of Computer Science and Engineering
Raipur, Chhattisgarh, India
Barkha.sharma.jobs@gmail.com
Mob 9630095363

Dr.P.Uday Kumar
Central College of Engineering and Management
Dept. of Computer Science and Engineering
Raipur, Chhattisgarh, India
Uday.uday08@gmail.com
Mob 9893263785

Abstract

The pattern of cloud has expanded with the expanding information in internet. Cloud depends on virtualization which is raising the virtual machine images in data centers, and keeping up the first and reinforcements make the necessity of space in a few TB. This outcomes in making of reproductions of existing information or new information at time of entry, coming about duplication. For this, de-duplication of information is required. Here, we have made a review on various strategies of de-duplication and which approach suits the best. Thus here in this paper we would be discussing data de-duplication techniques along with securing techniques thus forming secure de-duplication.

Keywords— *Data Deduplication, Cloud, Big data, secure deduplication.*

I. INTRODUCTION

The utilization of cloud computing to store, share their information for the diverse purposes has expanded now a days. Diverse clients are transferring and putting away their information for different circumstances [6]. It needs a lot of putting away space to store that information. It might happen that distinctive clients transfer same information and ordinarily a similar client transfers the information more than once purposely or unconsciously. On the off chance that the information is put away over and over, it needs a lot of storage room. To spare the storage room checks can be connected when the information is transferred by the client. On the off chance that the information as of now exists then it will tell client that there is a copy information else it will store in the cloud server. De-duplication is the strategy of expelling the excess information.

With the fast reception of cloud administrations, more volume of information is put away at remote servers, so strategies to spare plate space and system transfer speed are required. A key idea in this setting is de-duplication, in which the server stores just a solitary duplicate of every record, paying little respect to what number of customers need to store that document. All user having that document just utilize the connection to the single duplicate of the record put away at the server.

II. DATA DE-DUPLICATION

Information De-duplication is a particular information is basically a compression strategy for dispensing with copy duplicates of repeating information [7].

A. Benefits

Fundamentally, it can decrease the storage room involved by the information [8]. This will bring the following advantages:

a. IT reserve funds stores (need not bother with the additional space expected to expand speculation)

b. Reduce the reinforcement information, information depictions of the extent of the cost-sparing, sparing time, and so on.

c. Less power pressure on account of less hard, less tape, and so on.

d. Save network bandwidth on the grounds that exclusive less information.

e. Because of the need space of less storage, disk backup possible.

B. Data De-Duplication Process

The fundamental steps to erase duplicate information comprises of five phases:

1. The principal period of information accumulation stage, by contrasting the old and the new reinforcement information

reinforcement, reducing the extent of the information.

2. The second period of the way toward recognizing information, in bytes, of the information gathering stage denote a similar information objects.

3. The information is re-collected, new information is spared, and the past stage was checked duplicate information is spared information pointer substitution. The final product of this procedure is to deliver a duplicate of the erased after the reinforcement assemble see.

4. Actually expel all the copy information before playing out an information trustworthiness check efficiency.

5. Finally evacuate the repetitive storage of information, the arrival of already involved disk space for different uses.

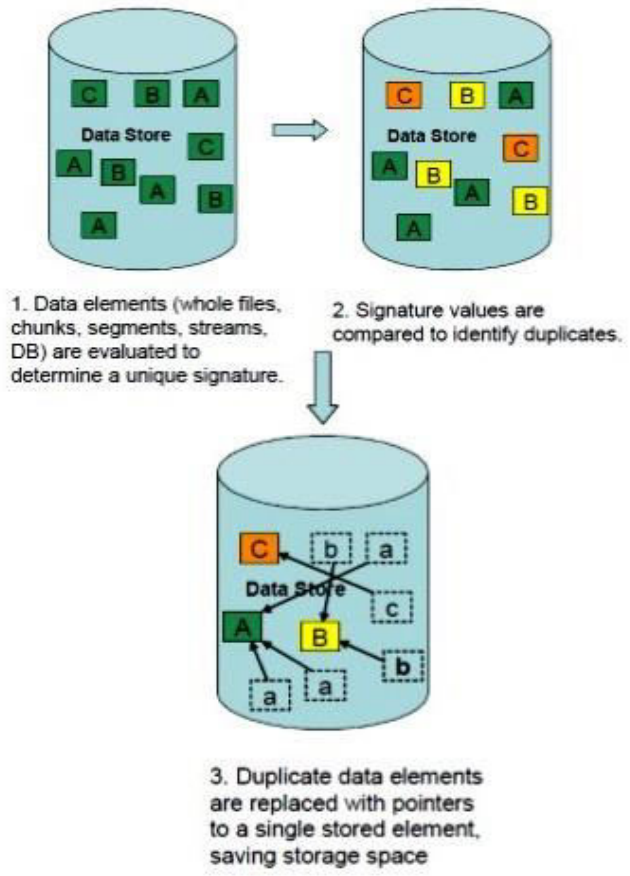

Fig.1. shows the workflow of data de-duplication implementation

III. LITERATURE SURVEY

Li, X. Chen et al. [1], introduces a baseline approach in which each user holds an independent master key for encrypting the convergent keys and outsourcing them to the cloud. However, such a baseline key management scheme generates an enormous number of keys with the increasing number of users and requires users to dedicatedly protect the master keys.

Yukun Zhou et al. [2], presents an observation that cross-user redundant data. They are mainly from the duplicate files, motivates us to propose an efficient secure de-duplication scheme SecDep. SecDep employs UserAware Convergent Encryption (UACE) and Multi-Level Key management (MLK) approaches. (1) UACE combines cross-user file-level and inside-user chunk-level deduplication, and exploits different secure policies among and inside users to minimize the computation overheads. Specifically, both of file-level and chunk level de-duplication use variants of Convergent Encryption (CE) to resist brute-force attacks.

M. Bellare et al. [3], propose an architecture that provides secure de-duplicated storage resisting brute-force attacks, and realize it in a system called DupLESS. In DupLESS, clients encrypt under message-based keys obtained from a key-server via an oblivious PRF protocol. It enables clients to store encrypted data with an existing service, have the service perform de-duplication on their behalf, and yet achieves strong confidentiality guarantees. We show that encryption for de-duplicated storage can achieve performance and space savings close to that of using the storage service with plaintext data.

S. Halevi et al. [4], introduces the notion of proofs-of-ownership (PoWs), which lets a client efficiently prove to a server that that the client holds a file, rather than just some short information about it. We formalize the concept of proof-of-ownership, under rigorous security definitions, and rigorous efficiency requirements of Petabyte scale storage systems. We then present solutions based on Merkle trees and specific encodings, and analyze their security. We implemented one variant of the scheme. Our performance measurements indicate that the scheme incurs only a small overhead compared to naive client-side de-duplication.

Jin Li et al. [5], proposes new distributed de-duplication systems with higher reliability in which the data chunks are distributed across multiple cloud servers. The security requirements of data confidentiality and tag consistency are also achieved by introducing a deterministic secret sharing scheme in distributed storage systems, instead of using convergent encryption as in previous de-duplication systems.

Table I. *shows various existing methods and its limitations*

Ref. No.	Method Used	Data Source	Approach	Strength	Limitation
1	De-duplication method	Outsourced Data at S-CSP	Introduce a baseline approach in which each user holds an independent master key for encrypting the convergent keys and outsourcing them to the cloud.	demonstrate that Dekey incurs limited overhead in realistic environments	Convergent keys are distributed across Multiple servers but the key servers are limited. Key space overhead needs to be taken care.
2	server-aided method	real-world datasets	Proposes redundant data distribution on cross-user file-level and inside-user chunk-level to perform different security policies	SecDep ensures data confidentiality and key security	Time overhead comes with multi-level key management can be reduced.
3	Dupless architecture	encrypted data	Author shows that encryption for deduplicated storage can accomplish performance and space savings close to that of using the storage service with plaintext data.	secure de-duplicated storage resisting brute-force attacks	Get and Put operations are time consuming. o Large computational overheads for chunk level
4	PoWs method	real-world datasets	author discusses solution based on Merkle Trees and specific encoding which identify attacks that exploit client side de-duplication attempts to identify de-duplication.	Proofs of-ownership (PoWs) concept in which Client proves to the server that it in fact holds the data of the file and not just some short information about it.	Performance measurements indicate that this scheme incurs small overhead compared to naïve client side de-duplication
5	De-duplication method	real-world datasets	Author has proposed a distributed De-duplication system with higher reliability (in storage over cloud) in addition to achieving confidentiality and integrity over data.	Message Authentication Code (MAC – use short cryptographic hash function) are used which also support process of secure de-duplication system.	Only two types of attacks are considered. Type 1 Attack for Dishonest system and Type 2 attack for Collusion

Hash Code is used to search the similar file. Normally files are stored and hash is applied on that particular file. If new file is uploaded, hash function checks whether it's available or not via Hash code.

The hash function is generally expressed by formula:

$$h(m) = h^{-1} \oplus (m \ll p) \otimes (m \gg q)$$

IV. TOOLS USED

There are many tools available for searching and mapping existing files. The tools used in subsequent literature reviews are discussed below:

a. hashLib: This module implements a common interface to many different secure hash and message digest algorithms.
b. Pure-Python RSA: For RSA encryption algorithm implementation.
c. Dropbox Package: The Core API is based on HTTP and OAuth and provides low-level calls to access and manipulate a user's Dropbox account.

V. CONCLUSION

Cloud storage services offers on request virtualized capacity resources and clients pay for the space they really expended. As the expanding interest and information store in the cloud, information de-duplications one of the systems used to enhance storage effectiveness. Information de-duplication is a specific information compression strategy for deleting copy duplicates of information away. With a specific end goal to upgrade transfer data transmission and storage room over cloud, Source Based De-duplication is one of the best choices.

Distributed de-duplication frameworks accomplishes security, classification and unwavering quality if information. In this manner if both the methodologies are joined then one can accomplish better de-duplication ratio along with reliability of information. Further de-duplication algorithm can be altered to accomplish better de-duplication proportion.

REFERENCES

[1] Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management", IEEE Transactions on Parallel and Distributed Systems vol. 25(6), Year – 2014, pp. 1615-1625

[2] Yukun Zhou, Dan Feng, Wen Xia, Min Fu, Fangting Huang, Yucheng Zhang, Chunguang Li, "SecDep: A User-Aware Efficient Fine-Grained Secure Deduplication Scheme with Multi-Level Key Management", IEEE Mass Storage Systems and Technologies (MSST) 2015 31st Symposium, Year – 2013, pp. 1-14

[3] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Server-aided encryption for deduplicated storage", ACM SEC'13Proceedings of the 22nd USENIX conference on Security, Year – 2013, pp. 179-194.

[4] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems", ACM Conference on Computer and Communications Security, Year – 2011, pp. 491-500

[5] Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang, Yang Xiang, Mohammad Hassan and Abdul hameed Alelaiwi, "Secure Distributed Deduplication Systems with Improved Reliability", IEEE Transactions on Computers Volume: PP, Year– 2015, pp. 3569-3579.

[6] P. Anderson and L. Zhang, "Fast and secure laptop backups with encrypted de-duplication," in Proc. 24th Int. Conf. Large Installation Syst. Admin., 2010,pp. 29–40.

[7] W. K. Ng, Y. Wen, and H. Zhu, "Private data deduplication protocols in cloud storage," in Proc. 27th Annu. ACM Symp. Appl. Comput., 2012, pp. 441–446.

[8] J. Xu, E.-C. Chang, and J. Zhou, "Weak leakage-resilient client side deduplication of encrypted data in cloud storage," inProc.8th ACM SIGSAC Symp.Inform. Comput. Commun. Security, 2013, pp. 195–206.