A Survey on Distributed Deduplication System for Improved Reliability and Security

B. Sateesh Kumar[1], V. Uma Rani[2], T. Akshay Raj[3]

[1]*Asst. Prof of CSE, JNTUH College of Engineeering, Jagitial, Nachupally, Kodimial, Karimnagar District, Telangana, India*
[2]*Asst. Prof of CSE, School Of Information Technology-JNTUH, KPHB Village, Kukatpally Mandal, Ranga Reddy District, Telangana, India*
[3]*M.Tech Student, Department of CSE, School Of Information Technology-JNTUH, KPHB Village, Kukatpally Mandal, Ranga Reddy District, Telangana, India*

---***---

Abstract - *Data Deduplication is a specialized strategy for eliminating repeated copies of data, and has been extensively used in distributed storage to shorten storage space and save bandwidth. In order to achieve the confidentiality and tag consistency the concept of deterministic secret sharing scheme has been proposed as a replacement to convergent encryption in distributed storage systems. Data deduplication checks for back-up of sequences of bytes over a comparison window of definite size. Certain Sequence of data(about 8Kb) are compared to the history of added such sequences and it is perfect for redundant operations like backup where iterative calls are made for copying and storing the similar data set a couple of times for data recovery purpose. In order to handle such attacks, the notion of proofs-of-ownership (PoWs) has been found, which lets a user effectively prove to a server that the applicant holds a file.*

To reduce the number of bytes dispatched through network, the idea of data Deduplication can be applied to network data. In spite of the fact that Data Deduplication gives a considerable measure of privacy and security concerns emerges as client's sensitive information is vulnerable to both insider and outsider attacks.

***Key Words*: Data Deduplication, Distributed Deduplication System, Reliability, Secret Sharing scheme, Security, Reliability**

1. INTRODUCTION

Today, With the Exponential growth in volume of data over the past few years many organizations are struggling to manage the data leading to an expensive problem. For this purpose, Data Deduplication is considered as the next evolutionary step in the field of Backup Technology due to its ability to reduce cost of the storing data. Data Dedeplication(DeDupe) is considered as "Integral" for all every organizations to remain competitive by operating efficiently.

The basic idea behind deduplication is to store repeated copies of data (either blocks or files) only once, which ultimately helps in improving the search results much more efficiently and quickly. For Instance, Consider a typical email system which may contain over 100 instances of same 1 megabyte (MB) file attachment. When the email platform is archived or backup, all those 100 instances are saved which requires 100 MB of storage. By applying data deduplication, only single instance of the attachment is stored actually and the successive instances are referred back to the one saved copy. In this case, demand for 100 MB is reduced to only 1 MB [10].

Deduplication can shorten storage utilization by up to 90 to 95% for backup applications[8] and up to 68% in standard file systems[9]. Data Deduplication plays a strategic role of saving on storage costs. It plays a crucial role in disaster recovery since there is considerably less data to transfer. In case of backup or archive data includes lot of redundant data. Similar data is saved multiple times, consumes unnecessary storage space, power and bandwidth. Such process creates a chain of consumes unnecessary storage space, power and bandwidth. Such process creates a chain of cost and inefficiency of resources within an organization.

Various deduplication frameworks have been proposed based on different deduplication methodologies for example, Server-side deduplication or Client-side deduplication or block level or file level deduplication. Deduplication can be applied either at block level or at file level. In file level deduplication, it takes out repeated copies of identical file. Deduplication can also be applied at block level that removes duplicate data blocks present in non identical files.

Table 1 Comparison of various Deduplication Approaches

Deduplication Approach	Storage Utilization	Bandwidth Utilization	Throughput	Cost	Efficiency	Deduplication Ratio
Source based	Medium	Low	Medium	Low	Medium	Medium
Target Based	High	High	Medium	High	Medium	Medium
Block Level	High	Medium	Low	Medium	High	High
File Level	Medium	Low	High	Low	Less	Low
Inline	Low	Low	Low	Low	Medium	Low
Post process	Low	High	Medium	High	High	High

2. RELATED WORK

[1] M. Bellare, C. Namprempre , Gregory Neven et al implement either security proofs or attacks for a huge number of identity based identification and signature recognition scheme characterized either directly or indirectly in the existing. This scheme works that one hand it clarifies how the plans are inferred and empowers the modular security analysis. It includes interaction with IBI(an authority which has master public key and a master private key).

IBI provides a secret key to user based on his identity. The client, playing the character of a actual user, could At that point, present himself to a auditor in a conventional way in which the auditor starts by knowing the asserted identity of the client and the master key of the authority. IBS plan is comparable with the exception of that the client signs messages, instead of distinguishing itself and validation of a signature need details about identity of signer and the public master key.

[2] Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou et al, proposed a system that aims to achieve proficient and trustworthy key management using convergent key deduplication and handle the issue of managing the huge number of Convergent keys generated with increase in users with efficiency and reliability. The author proposed two methodologies for this system.

They are

1. Convergent Encryption: This ensures security against unauthorized users. The data block or file is encoded using hash key generated from hash function further this hash key is encrypted by owner's key.
2. Dekey Methodology: Creates secret divisions on original convergent key and transmit them to multiple key management authorities. Different users who shares same block may access the same convergent keys respectively.

[3] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer et al demonstrate a mechanism to recover space from accidental duplication for controlled file replication to make it accessible.

This mechanism includes

1) Convergent encryption, which encodes the file using hash function later hash value, is encoded using the user's public key.

2)Self Arranging, Lossy, Associative Database(SALAD) is utilized for combining file content and data location in a fault tolerant, decentralized and scalable features. Huge scale reenactment examinations show that the duplicate file merging system is scalable, very effective and fault tolerant.

[4] M. Bellare, S. Keelveedhi and T. Ristenpart et al proposed DupLess Server aided encryption for deduplicated storage performs deduplication to safeguard space by saving single replica of each file uploaded. Message lock encryption may be used to clear the dilemma of clients encrypting their file however the saving are lock. Dupless is proposed to fulfil secure deduplicated storage and in addition to storage resisting brute force attacks. Customers encode under message-predicated keys acquired from a key-server by means of inattentive PRF protocol in

dupless server. It authorizes clients to store encrypted data with an available service. It demonstrates that encryption for deduplicated storage could successfully achieve desired performance and space saved close to that of using the storage accommodation with plaintext data.

Characteristics**:**

- Resolve the issue of message lock Encryption.
- Increased Security.
- Effortlessly conveyed answer for encryption that supports deduplication.

[5] In 2008 Mark W. Storer et al. proposed two models for secure deduplicated storage are authentication and anonymous. These two designs exhibit that security can be merged with deduplication in such way that it gives a distinct scope of security attributes. In the models they show that the security is given using convergent encryption. This procedure, initially presented with regards to the Farsite framework, gives a deterministic method for producing an encryption key, such that two distinct clients can encode information to the same cipher text. Both authentic and anonymous models, a map is made for each record on how to construct the entire file from blocks of data. This record itself encrypted using a distinct key.

To improve the security of deduplication and ensure the Confidentiality, Bellare et al. demonstrated to ensure the information secrecy by changing the anticipated message into erratic message. In their framework, another third party called key server is acquainted to produce the file tag for copy check. Q. Wang et al. introduced a novel encryption plan that gives differential security to well known information and disliked information.

[6] P.Anderson, L.Zhang et al to depict traditional backup algorithm which takes favorable circumstances of the data which is regular between clients to boost the speed of backups and lower the storage required. This algorithm backs client- end user encryption which is essential for personal information. It additionally backs an interesting feature which permits quick identification of common sub trees, keeping away from the need to query the backup source.

To descries a model execution of this algorithm for Apple OS X, and present an examination of the potential adequacy, utilizing genuine information acquired from a set of typical users. Deduplication of a hashing function can be utilized to generate a unique key for a data block, with respect to contents of data; if two or more users have same data, then hashing function must return the same key of keys user index for saving data, any attempt to save numerous duplicates of same data block will be identified promptly. Encrypting data nullifies Deduplication, two same blocks encoded by different keys. Two same data blocks yield same encrypted blocks which can be copied in the ordinary way. Every block has a different encryption key.

[7] S.Halevi, D.Harnik, B.pinkas, A.Shulman-Peleg et al, client side deduplication tries to determine the opportunities of deduplication at client and save the bandwidth by obstructing the transfer of duplicate files to the server. In order to determine the attacks at client side that allows an attacker to gain access to a particular file of other users based on signature of the file. An attacker having signature of file can depict to server that he owns the file and download it.

To overcome such attacks, author introduced the concept of Proof-of-Ownerships (POW's) an interactive algorithm run by client using which a client efficiently prove a server that he owns the file without uploading the file to sever. To formalise the idea of Proof-of-Ownership, under thorough security definitions and thorough effectiveness prerequisites of data peta scale storage framework. To tackle the issue of utilizing a little hash values has an intermediary for the whole record, outline an answer where a customer evidences to the server that it in fact has the document. Few PoW concepts based on Merkle Hash Tree were proposed to enable client side deduplication along with identification of bounded leakage setting.

3. CONCLUSION:

In this paper, we have surveyed various methodologies available for performing deduplication and their effectiveness in terms of performance and security. The techniques from the above mentioned papers could be used to enhance the security and protect the system from unauthorized users. In this survey paper we have examined the existing system that has many problems in terms of data reliability, security and storage overheads. Aiming to achieve both reliability and security, we propose Distributed Deduplication System for Improved Reliability and Security that ensure there are no duplicate copies of data and achieve reliability and data confidentiality without even using an encryption mechanism. The use of concepts like block level and file level deduplication and ramp secret sharing scheme which uses tag generation algorithm. The combination of such features gives unmatched levels of security for the deduplication.

REFERENCES:

1. Bellare, C. Namprempre, and G. Neven. Security proofs for identity-based identification and signature schemes. *J. Cryptology*, 22(1):1–61, 2009.
2. 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*, 2014, pp. vol.25(6), pp. 1615–1625.
3. 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*, 2002, pp. 617–624.
4. M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Serveraided encryption for deduplicated storage,” in *USENIX Security Symposium*, 2013.
5. M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, “Secure data deduplication,” in *Proc. of StorageSS*, 2008.
6. P. Anderson and L. Zhang, “Fast and secure laptop backups with encrypted de-duplication,” in *Proc. of USENIX LISA*, 2010.
7. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, “Proofs of ownership in remote storage systems.” in *ACM Conference on Computer and Communications Security*, Y.Chen, G. Danezis, and V. Shmatikov, Eds. ACM, 2011, pp.491–500.
8. http://opendedup.org/
9. Dutch T Meyer and William J Bolosky. A study of practical deduplication. ACM Transactions on Storage (TOS), 7(4):14, 2012.
10. https://en.wikipedia.org/wiki/Data_deduplication .