ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 6, Issue 1, January 2017

An Efficient Data Deduplication in Cloud Environment with Improved Reliability: An Overview

Miss. Namrata P. Kawtikwar, Prof. M.R. Joshi

***Abstract*— Day by day the use of memory is increases rapidly. The process of eliminating the repeated or duplicates copies of data is called as Data deduplication. This data deduplication process is widely used in cloud storage to decrease storage space and upload bandwidth. By using, deduplication system progress of storage utilization and reliability is increases. In addition, the dare of privacy for sensitive data also take place when they are outsourced by users to cloud. Planning to address the above security test, this paper constructs the first effort to celebrate the idea of scattered reliable deduplication system. The paper recommends a new distributed deduplication systems with upper dependability in which the data chunks are distributed from corner to cornering multiple cloud servers. The safety needs of data privacy and tag stability are also accomplish by introducing a deterministic secret sharing scheme in distributed storage systems, instead of using convergent encryption as in previous deduplication systems. A deduplication technique, on the other hand, can reduce the storage cost at the server side and save the upload bandwidth at the user side.**

Keywords: **Deduplication, reliability, distributed storage system, secret sharing etc**.

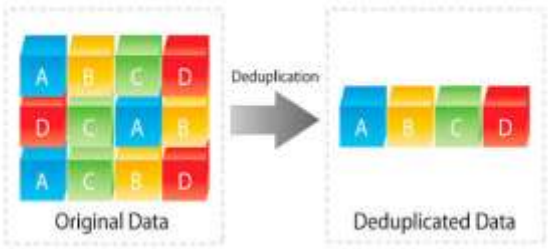
I. INTRODUCTION

Now days with the huge increasing of population and the using of technology, it leads to many problems. The growth in technology is increasing the amount of storage or communication and technique devices. By the unpredictable development of digital data, deduplication techniques are broadly engaged to backup data and decrease network and storage transparency by notice and eradicate redundancy among data. As an alternative of maintaining multiple data copies with the same content, deduplication reducing redundant data by maintaining only single copy and referring other redundant data to that copy. Deduplication has inward much concentration from both academic world and industry since it can really recover storage utilization and keep storage space, particularly for the applications with high deduplication ratio such as archival storage systems. A number of deduplication systems have been projected based on various deduplication scheme such as client-side or server-side deduplication, file-level or block-level deduplications. Specially, with the advent of cloud storage, data deduplication procedure grows to be more gorgeous and essential for the management of ever-increasing quantity of data in cloud storage services. Deduplication has received much attention from both academic and industry because it can more improves storage utilization and save storage space, especially for the applications with high deduplication ratio such as accession storage systems. For eliminating duplicate copies of data we use data deduplication technique. To reduce storage space and for uploading bandwidth mostly it has been used.

How deduplication works?

Data deduplication works by comparing objects (usually files or blocks) and removes objects (copies) that already exist in the data set. All the processes which are not unique are removed in this method.

In Data deduplication method we divide the input data into blocks and a hash value is calculated for each of these blocks. Then using these hash values we can determine whether another block of same data has already been stored. If a similar data file is found then replace the duplicate data with a reference to the object already present in the database.



II. EXISTING SYSTEM

Data de-duplication techniques are very interesting techniques that are widely employed for data backup in enterprise environments to minimize network and storage overhead by detecting and eliminating redundancy among data blocks. There are many de-duplication schemes proposed by the research community. The reliability in

Manuscript received Jan, 2017.
Miss. Namrata P. Kawtikwar
M.E. CSIT (2nd Year)
P.G. Department of Computer Science & Information Technology, HVPM, COET Amravati, Maharashtra,India.

Prof. M.R.Joshi
Asst.Professor P.G. Department of Computer Science & Information Technology, HVPM, COET Amravati,Maharashtra, India.

de-duplication has also been addressed. However, all of these works have not considered and achieved the tag consistency and integrity in the construction

Disadvantages of Existing System:

- Most of the previous deduplication systems have only been considered in a single-server setting.
- Data reliability is actually a very critical issue in a deduplication storage system because there is only one copy for each file stored in the server shared by all the owners.
- The traditional deduplication methods cannot be directly extended and applied in distributed and multi-server systems

III. PROPOSED SYSTEM

System Model

Two kinds entities will be involved in this deduplication system, including the user and the storage cloud service provider (S-CSP). Both client-side deduplication and server-side deduplication are supported in our system to save the bandwidth for data uploading and storage space for data storing.

• *User*. The user is an entity that wants to outsource data storage to the S-CSP 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. Furthermore, the fault tolerance is required by users in the system to provide higher reliability.

• *S-CSP*. The S-CSP is an entity that provides the outsourcing data storage service for the users. In the deduplication system, when users own and store the same content, the S-CSP will only store a single copy of these files and retain only unique data. A deduplication technique, on the other hand, can reduce the storage cost at the server side and save the upload bandwidth at the user side. For fault tolerance and confidentiality of data storage, we consider a quorum of S-CSPs, each being an independent entity. The user data is distributed across multiple S-CSPs.

To protect private data the secret sharing technique is used which is corresponding to distributed storage systems. Here the secret sharing technique is used for protection of private data. In detail a file is divides and encode into sections by using secret sharing technique. These sections will be distributed over many independent storage servers. A cryptanalysis hash value of the content will also be calculated and send to storage server as the mark of the fragment stored at each server. Only the data user who first upload the data is required to calculate and distribute such secret shares and following users own same data copy do not need to calculate and stores these shares. Retrieve data copies owner must access a minimum number of storage server by a validation and obtain the secret shares to alter the data. In different way, the authorized uses will access the secret shares data copy.
Here, the main motives of the proposed systems are-

- To authenticate a data and to make available integrity and validity assurances on the data.
- To Distributed Deduplication System.
- To implement Block-level Distributed Deduplication System.
- To maintain the consistency and integrity of data within file.

IV. PROPOSED TECHNIQUES

A. *File Splitting*

Data deduplication involves finding and removing duplication within data without compromising its fidelity or integrity. The goal is to store more data in less space by segmenting files into small variable-sized chunks (32–128 KB), identifying duplicate chunks, and maintaining a single copy of each chunk. Redundant copies of the chunk are replaced by a reference to the single copy. The chunks are compressed and then organized into special container files in the System Volume Information folder. After deduplication, files are no longer stored as independent streams of data, and they are replaced with stubs that point to data blocks that are stored within a common chunk store. Because these files share blocks, those blocks are only stored once, which reduces the disk space needed to store all files. During file access, the correct blocks are transparently assembled to serve the data without calling the application or the user having any knowledge of the on-disk transformation to the file. This enables administrators to apply deduplication to files without having to worry about any change in behavior to the applications or impact to users who are accessing those files.

B. *File-Level Distributed Deduplication System*

It support capable duplicate check, tags for each file will be calculated and send to storage cloud service provider. To prevent alignment invasion organized by the cloud based service provider, tag collected at different storage servers. System Setup: In our structure, the storage cloud service provider is considered to be n with identities denoted by id1, id2,…,idn respectively. To upload file F, the client communicate with cloud based service provider to perform the elimination of duplicate data .For downloading file F, the client downloads the secret shares of the file from k out of storage servers.

C. *Block-Level Deduplication System*

In this part, we appear how to derive the fine grained block level distributed deduplication. In this system, the client also demands to perform the file level deduplication before uploading file. The user partition this files into blocks, if noduplication is found and performs block-level deduplication system. The system set up is similar to file-level deduplication and also block size parameter will be defined.

- System Architecture design:

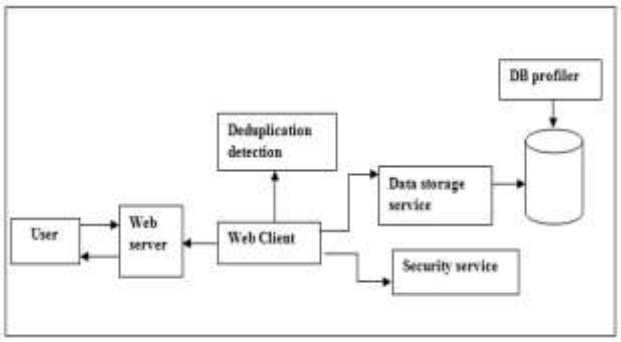


Fig. System Architecture

V. ADVANTAGES

1. This application helps in easy maintenance of data on the cloud platform so that no duplicate files are saved in the cloud which helps to save the storage space.

2. Proposed system provides authentication and integrity and validity assurances on the data.
3. The proposed constructions support both file-level and block-level deduplications.

VI. CONCLUSION

Design of an improved technique for storage in Cloud is deduplication technique. Deduplication aids in saving the storage space. This application helps in easy maintenance of data on the cloud platform so that no duplicate files are saved in the Cloud. With the evolution of Cloud computing, storage resources of commodity machines can be efficiently utilized. This allows every organization to build its own private cloud for a variety of purposes. In order to better utilize the limited storage available in a private cloud, a suitable approach for optimization has to be used.

ACKNOWLEDGEMENT

I express my sincere gratitude to Resp. Prof. P. L. Ramteke, Head of the Department of Computer Science & Information Technology & Resp. Prof. M. R. Joshi for providing their valuable guidance and necessary facilities needed for the successful completion of this paper throughout.

Lastly, but not least, I thank all my friends and well wishers who were a constant source of inspiration.

REFERENCES

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

[2] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, "Secure data deduplication," in Proc. Of StorageSS, 2008.

[3] P. Anderson and L. Zhang, "Fast and secure laptop backups with encrypted de- duplication," inProc. of USENIX LISA, 2010

[4] J. S. Plank, S. Simmerman, and C. D. Schuman, "Jerasure: Alibrary in C/C++ facilitating erasure coding for storage applications- Version 1.2," University of Tennessee, Tech. Rep. CS-08-627,August 2008.

[5] J. S. Plank and L. Xu, "Optimizing Cauchy Reed-solomon Codesfor fault-tolerant network storage applications," in NCA-06: 5th
IEEE International Symposium on Network Computing Applications,Cambridge, MA, July 2006.

[6] C. Liu, Y. Gu, L. Sun, B. Yan, and D. Wang, "R-admad: Highreliability provision for large-scale de-duplication archival storage
systems," in Proceedings of the 23rd international conference onSupercomputing, pp. 370–379.

[7] M. Li, C. Qin, P. P. C. Lee, and J. Li, "Convergent dispersal:Toward storage-efficient security in a cloud-of-clouds," in The 6thUSENIX Workshop on Hot Topics in Storage and File Systems, 2014.

[8] P. Anderson and L. Zhang, "Fast and secure laptop backups withencrypted de-duplication," in Proc. of USENIX LISA, 2010.

[9] Z. Wilcox-O'Hearn and B. Warner, "Tahoe: the least-authorityfilesystem," in Proc. of ACM StorageSS, 2008.

[10] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S.Lui, "A secure cloud backup system with assured deletion and
version control," in 3rd International Workshop on Security in CloudComputing, 2011.

[11] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, "Securedata deduplication," in Proc. of StorageSS, 2008.

[12] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A securedata deduplication scheme for cloud storage," in Technical Report,2013.

Authors Details:-

Miss. Namrata P. Kawtikwar
M.E. CSIT (2nd Year)
P.G. Department of Computer Science & Information Technology,
HVPM, COET Amravati, Maharashtra,India.

Prof. M.R.Joshi
Asst.Professor P.G. Department of Computer Science & Information Technology, HVPM, COET Amravati,Maharashtra, India.