A Review on Empirical Study of De-Duplication In Distributed System Using File & Block Level De-Duplication Technique

T. P. Adhau[1], Dr. V. M. Deshmukh[2]
[1]Dept of CSE
[2]Professor, Dept of CSE
[1, 2]PRMIT & R Badnera , Amravati, India

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 de-duplication. This data de-duplication process is widely used in in cloud storage to decrease storage space and upload bandwidth. By using, de-duplication 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 de-duplication system. This paper recommends a new distributed de-duplication 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 de-duplication systems.A de-duplication technique, on the other hand, can reduce the storage cost at the server side and save the upload bandwidth at the user side.*

Keywords- Block-level, de-duplication, Distributed, Encryption file level..

I. INTRODUCTION

By the unpredictable development of digital data, de-duplication 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, de-duplication reducing redundant data by maintaining only single copy and referring other redundant data to that copy. De-duplication 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 de-duplication ratio such as archival storage systems. A number of de-duplication systems have been projected based on various de-duplication scheme such as client-side or server-side de-duplication, file-level or block-level de-duplications.Specially, with the advent of cloud storage, data de-duplication procedure grow to be more gorgeous and essential for the management of ever-increasing quantity of data in cloud storage services which inspires Endeavour and club to outsource data storage to third-party cloud providers. Today’s commercial cloud storage services, such as Dropbox, Google Drive and Mozy, have been applying de-duplication to save the network bandwidth and the storage cost with client-side de-duplication. generally comprises up to 5 to 7 pages. These are:multi-label learning, more than one class can be assigned to an instance. With the increase in the number of data

II. RELATED WORK

M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Serveraided encryption for deduplicated storage,” in USENIX Security Symposium, 2013.It introduced the idea of security and scheme for symmetric encryption in concentrate security framework. They give different idea of security and analyze the good involution of reduction among them. They provide method of encryption using a block cipher, cipher block chaining and counter mode. Its have two goals .First is to study the idea of security for symmetrical encryption and second is to provide concrete security analysis of fixed symmetric encryption device.M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, “Secure data de-duplication,” in Proc. Of StorageSS, 2008.They developed a solution that provides both data security and space efficiency in single-server storage and distributed storage systems to solve the problem such that de-duplication exploits identical content, while encryption tries to make all content appear random ,the same content encrypted with two different keys results in very different cipher text. De-duplication and encryption are opposed to one another. De-duplication takes benefit of data similarity to achieve a reduction in storage space & the goal of cryptography is to make cipher text indistinguishable from theoretically random dataP. Andersonand L. Zhang, “Fast and secure laptop backups with encrypted de- duplication,” in

Proc. of USENIX LISA, 2010.They presents an algorithm which takes benefits of the data which is common between users to reduce the storage requirements, and increase the speed of backups. This algorithm supports clientend per-user encryption which is important for confidential personal data, also supports a unique feature that allows immediate detection of common sub trees, avoiding the necessity to query the backup system for every file. This system has shown that a community of laptop users shares a considerable amount of data in between. This gives the potential to significantly decrease backup times and storage requirements. However, they have shown that manual selection of the relevant data -eg, backing up only home directories is a poor strategy; this become fails to take backup of important files, at the same time as unnecessarily duplicating other files.

III. PROPOSED SYSTEM

To protect private data the secret sharing technique is used which is corresponding to distributed storage systems. In this paper 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. Another distinguishable feature of our proposal is that data completeness encloses tag consistency, can be derived. To explain further if the same value is stored in various cloud storage then de-duplication check by methods. It cannot oppose the collision attack established by many servers. To our knowledge no related work on secure de-duplication can rightly address, the reliability and tag consistency problem. The file level and block level de-duplication is used for higher reliability. The secret splitting technique is used for protect data. Our proposed structure supports both traditional de-duplication methods. Privacy, credibility and integrity can be achieved in our proposed system. In solution to kind of secret agreement attacks are considered. These are the attack on the data and the attack against servers. The data is secure when the opponent control limited number of storage servers.

Proposed Techniques

File Splitting

Data de-duplication 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 de-duplication, 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 de-duplication to files without having to worry about any change in behavior to the applications or impact to users who are accessing those files.

File-Level Distributed De-duplication 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.

Block-Level De-duplication System

In this part, we appear how to derive the fine grained block level distributed de-duplication.In this system, the client also demands to perform the file level de-duplication before uploading file. The user partition this files into blocks, if no duplication is found and performs block-level de-duplication system. The system set up is similar to file-level de-duplication and also block size parameter will be defined.

IV. CONCLUSION

We proposed the distributed de-duplication systems to improve the reliability of data while achieving the confidentiality of the users' outsourced data without an encryption mechanism. We proposed the file level de-duplication, block level de-duplication, tag generation and message authentication code technique. The security of tag consistency and integrity will be achieved.

REFERENCES

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

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

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

[4] J. S. Plank, S. Simmerman, and C. D. Schuman, "Jerasure: A library 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 Codes for fault-tolerant network storage applications," in NCA-06: 5thIEEE International Symposium on Network Computing Applications, Cambridge, MA, July 2006.

[6] C. Liu, Y. Gu, L. Sun, B. Yan, and D. Wang, "R-admad: High reliability provision for large-scale de-duplication archival storagesystems," in Proceedings of the 23rd international conference on Supercomputing, 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 6th USENIX Workshop on Hot Topics in Storage and File Systems, 2014.

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

[9] Z. Wilcox-O'Hearn and B. Warner, "Tahoe: the least-authority filesystem," 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 andversion control," in 3rd International Workshop on Security in Cloud Computing, 2011.

[11] M. W. Storer, K. Greenan, D. D. E. Long, and E. L.Miller,
"Secure data de-duplication," in Proc. of StorageSS, 2008.

[12] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A secure data de-duplication scheme for cloud storage," in Technical Report, 2013