A Survey On De-duplication Technique Over Distributed Cloud Environment With Improved Security

Mahesh B.Gunjal[1], Prof . R.L.Paikrao[2]

[1]ME Student, Department of Computer Engineering, Amrutvahini COE, Sangamner, Maharashtra, India
[2]Associate Professor, Department of Computer Engineering, Amrutvahini COE, Sangamner, Maharashtra, India

ABSTRACT

De-duplication of data is the process of deleting redundant copies of stored data. It is single instance storage. It helps to achieve security requirements of data confidentiality in the cloud. As de-duplication has ability to improve storage utilization, it is most popular in academic as well as industry area. With the benefits and popularity of de-duplication it suffers from the problem such as, data reliability. Previously, existed systems have single server setting, which is not capable to preserve only single copy of the data in it due to privacy issue. Data privacy is a very challenging issue; it arises when more and more sensitive information is outsourced in cloud system. There is problem with encryption technique which is used by existing system, which required different cipher-texts for different users to share identical data. Today, there is need of achieving data confidentiality and reliability in distributed system by preserving data security requirements. Distributed de-duplication systems are efficient in which data blocks are spread across many cloud servers.

Keyword: *- De-duplication; distributed storage system; reliability; secret sharing.*

1. Introduction

De-duplication of data is a compressed form of the data. It is mostly used to keep back up of data as well as to reduce overheads of data storages. In network area data deduplication is applied to minimize the size of byte which is transferred. In this, single block of data and also byte patterns are identified during the process of storage analysis. Many deduplictionsystems such as, client-server side or file/block level deduplication systems are available with different de-duplication mechanisms [3]. These systems cannot protect the security of predictable files. Previously existed de-duplication systems have many limitations corresponding to data reliability and data confidentiality as only single copy of the data are preserved on server side. There is possibility of losing the confidentiality of data due to unavailability of small chunk or block of data.

Users are preferring cloud to host their important data and to manage their systems. Cloud is always a semi-trusted entity. Hence data on cloud is not safe sometimes. Private data de-duplication protocol (i.e. de-duplication checking of files at user level) in cloud storage is studied. Systems are having one more scenario in which data is not private. Data is shared with different users. In this case, data duplication is possible which low down the public cloud performance and is wastage of storage space. “Pay-As-You-Use” is basic principle followed by cloud. Hence there is need of de-duplication technique for encrypted data. For de-duplication check, comparing encrypted data of multiple users creates a challenge. Solution for this problem is the Ramp Secret sharing technique [1]-[6]. The de-duplication check need not to be the user dedicated it needs to check the de-duplication at cloud storage level. Data is shared among multiple users so there need to be some rules that restrict the de-duplication check for unregisterd user. Basic Problem is to provide de-duplication check for file level as well as at the block level for shared encrypted data. To store data, user needs to create encryption key and share with other user. To store keys on storage CSP violates the security. Hence a third party server is needed for key generation and management and for user authentication. Also if encrypted blocks are saved on the server then there are chances for data retrieval from the cloud hence distributed server need arises [4][14].

In cloud computing, there is need of designing secure de-duplication systems having higher reliability. The major focus is to develop the smart system that works well for cloud data de-duplication. System should use efficient network bandwidth. It must provide smart solution for encrypted file duplication along with this it must able to maintain secrecy as far as access to file is concern[4][6]. Ramp secret sharing scheme provides higher reliability and confidentiality levels. Storage cloud service provider (S-CSP) and user both are included in de-duplication system in cloud environment . In this user outsourced the data to S-CSP and then access it later [1]. Whereas, S-CSP reduces storage cost by storing unique copy of data and bandwidth of upload data at user side.

2. Related Work

Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang and Yang Xiang [1], proposes distributed de-duplication system which tends to achieve reliability as well as confidentiality of user's outsourced data. This paper implement, deduplication system using Ramp secret sharing as it incurs small overheads of data encoding and decoding. This system provides better fault tolerance. Secret sharing technique used spilt and encode the file into fragments.

J. Gantz and D. Reinsel[2] and J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer[3],This paper researched about the IDC analysis report that in 2020 data volume will be reaching up to 40 trillion gigabytes as the world expected. Author Farsite's, identified the problem of file identification in distributed system. In this paper proposed cryptosystem identifies the identical files located on cloud. This system also stores the replicas of each file in free space.

D M. Bellare, S. Keelveedhi, and T. Ristenpart[4],In this paper, proposed system Dup-Less works to obtain key-server using PRF protocol to encrypt the client message-based keys. By doing this task authors were willing to show that better performance can be achieved using encryption for de-duplicated storage. DupLess works against external security attacks. In this, clients are interacting with Key server with the help of PRF protocol. It uses less number of SS interactions.

G. R. Blakley and C. Meadows [5], A. D. Santis and B. Masucci[6] and A. Shamir [7] in this paper authors were formally introducing multiple ramp schemes. Multiple ramp schemes are required for sharing secret among multiple participants. In this entropy approach is used. Author, A. Shamir, provides the robust key management scheme in cryptographic system to share secret. Their techniques of sharing secrets are good for storing and also distributing encryption keys.

S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg[8], aims to overcome the problem of hash signature of files by introducing PoW i.e. Proofs-of-ownership's. In this scheme client keep a copy of file rather than short information about the file. Solution provided in this paper is based on Merkle-tree queries and it is applicable for relevant file-level deduplication only.

J. S. Plank, S. Simmerman, and C. D. Schuman [9], this paper describes interface as well as techniques and algorithms for code. Therefore, it is referred as a quasi-tutorial and a programmer's guide. Bit matrix is used for encoding and decoding.

M. Li, C. Qin, P. P. C. Lee, and J. Li [10], In this paper, convergent dispersal is proposed to provide efficient security for cloud storage system. In this system original data is used to derive deterministic cryptographic hash information. Furthermore, random information is replaced with cryptographic hash. Two convergent dispersal algorithms are proposed in this paper, namely CRSSS and CAONT-RS.

P. Anderson and L. Zhang [11], this paper defined evaluation of the proposed system using a local server to avoid the problem in sharing the data i.e. Time and cost of typical transfers as well as multi-user authentication. This problem can be reduced in this system by fetching and pre-processing the transmitting the required data over the cloud.

A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui[12],In this paper, authors extends the FADE system (File Assured Detection) to avoid weaker areas of older FADE to protect the data stored in the cloud and also to overcome the overhead in FADE. This scenario provides the guarantee of limited access control for data in cloud storage.

M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller [13], In this paper, authors Mark W. Storer Kevin Greenan Darrell D. E. Long Ethan L. Miller provide the solution for space management and data security in single server system as well as distributed file system. In this scenario, encryption keys are generated from chunk data in consistence manner and whole file utilizes the hash value as its identifier.

J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl[14], This paper firstly examines the popularity of the data; with respective to the popularity different level of protection is provided for the data in cloud. Proposed encryption system provides the guarantee of semantic security for unpolar data as well as less security and better security with proper bandwidth benefits for popular data.

D. Harnik, B. Pinkas, and A. Shulman-Peleg [15], In this paper, authors were exploring that how deduplication can be used as a side channel. They studied about cross-user deduplication which provides guarantees of higher privacy with slightly reducing bandwidth savings in cloud storage.

J. Xu, E.-C. Chang, and J. Zhou [16], This paper represents the deduplication cross different users in which identical duplicated files from multiple users are detected and removed safely.
In this two aspects are described as:

1. An efficient hash function is constructed.
2. Raised and simplified the convergent encryption method.

W. K. Ng, Y. Wen, and H. Zhu [17], Private data deduplication protocols is described in this paper and also formalized the context of two-party computations. Private data deduplication protocol is secure in simulation-based framework. In this system the result of private data deduplication is evaluated.

J. S. Plank and L. Xu[18] and C. Liu, Y. Gu, L. Sun, B. Yan, and D. Wang[19], In this paper, authors were studied about Cauchy Reed-Solomon coding for the construction of distribution matrix. Distribution matrix is constructed for encryption and decryption of data. High reliability provision mechanism i.e. R-ADMAD is proposed by D. Wang. R-ADMAD is dynamic and distributed recovery process in the cloud storage. It works same as the RAID system. But it is more reliable than RAID as it helps to reduce redundancy.

3. CONCLUSIONS

In this review paper, we studied about some previously existed de-duplication system. In the study of literature survey, we examined that existed systems have many problems such as, data reliability, data security, overheads of data storage etc. we also noticed that data de-duplication reduces the data storage overheads by preserving single copy of data in server side. In the study of base paper, we analyze that S-CSP reduces storage cost by storing unique copy of data and bandwidth of upload data at user side. Whereas, Ramp Secret sharing technique is efficient for sharing data securely in distributed system as it support to higher reliability and confidentiality levels of data. According to our evaluation in this paper, Ramp secret sharing and S-CSP both are powerful to achieve improved reliability in decentralized de-duplication.

5. ACKNOWLEDGEMENT

I am very much thankful of Prof. R.L. Paikrao for their guidance and consistent encouragement in this paper work.

6. REFERENCES

[1]. Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang and Yang Xiang Senior Member, IEEE and Mohammad Mehedi Hassan Member, IEEE and Abdulhameed Alelaiwi Member, IEEE, “Secure Distributed Deduplication Systems with Improved Reliability”.
[2]. J. Gantz and D. Reinsel, “The digital universe in 2020: Big data, bigger digi tal shadows, and biggest growth in the far east,” http://www.emc.com/collateral/analyst-reports/idcthe- digital-universe-in-2020.pdf, Dec 2012.
[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].Santosh S. Varpe and Prabhudev Irabashetti ,’Visual Cryptography for Providing Privacy to Biometric data’, International Journal of Current Engineering and Technology(IJCET), Vol.5, No.4 ,pp.912804-2807, 15 Aug 2015.

[6].G. R. Blakley and C. Meadows, "Security of ramp schemes," in Advances in Cryptology: Proceedings of CRYPTO '84, ser. Lecture Notes in Computer Science, G. R. Blakley and D. Chaum, Eds. Springer-Verlag Berlin/Heidelberg, 1985, vol. 196, pp. 242–268.
[7].A. D. Santis and B. Masucci, "Multiple ramp schemes," IEEE Transactions on Information Theory, vol. 45, no. 5, pp. 1720–1728, Jul. 1999.
[8]. A. Shamir, "How to share a secret," Commun. ACM, vol. 22, no. 11, pp. 612–613, 1979.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.
[9].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.
[10].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.
[11].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.
[12].P. Anderson and L. Zhang, "Fast and secure laptop backups with encrypted de-duplication," in Proc. of USENIX LISA, 2010.
[13]. 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 Cloud Computing, 2011.
[14].M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, "Secure data deduplication," in Proc. of StorageSS, 2008.
[15].J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A secure data deduplication scheme for cloud storage," in Technical Report, 2013.
[16].D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels in cloud services: Deduplication in cloud storage." IEEE Security & Privacy, vol. 8, no. 6, pp. 40–47, 2010.
[17].J. Xu, E.-C. Chang, and J. Zhou, "Weak leakage-resilient client-side deduplication of encrypted data in cloud storage," in ASIACCS, 2013, pp. 195–206.
[18].W. K. Ng, Y. Wen, and H. Zhu, "Private data deduplication protocols in cloud storage." in Proceedings of the 27th Annual ACM Symposium on Applied Computing, S. Ossowski and P. Lecca, Eds. ACM, 2012, pp. 441–446.
[19].J. S. Plank and L. Xu, "Optimizing Cauchy Reed-solomon Codes for fault-tolerant network storage applications," in *NCA-06: 5th IEEE International Symposium on Network Computing Applications*, Cambridge, MA, July 2006.
[20].C. Liu, Y. Gu, L. Sun, B. Yan, and D. Wang, "R-admad: High reliability provision for large-scale de-duplication archival storage systems," in *Proceedings of the 23rd international conference on Supercomputing*, pp. 370–379.

BIOGRAPHIES

	**Mr. Mahesh Bhaskar Gunjal** is Pursuing Master in Engineering from Amrutwahini College of Engineering Sangamner. Received BE degree from University of Pune. He is member of ACM.
	Prof. R. L. Paikrao is Associate Professor in Amrutvahini College of Engineering, Sangamner. Pursuing PhD degree . His Research interests includes Cloud Computing.