Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 87 (2016) 38 – 43

Procedia Computer Science

Fourth International Conference on Recent Trends in Computer Science & Engineering

Chennai, Tamil Nadu, India

A Study on Deduplication Techniques over Encrypted Data

Akhila K[a]*,Amal Ganesh[a],Sunitha C[a]

[a]*Department of CSE,Vidya Academy Of Science and Technology,Thrissur 680501,India*

Abstract

In the current digital world, data is of prime importance for individuals as well as for organizations. As the amount of data being generated increases exponentially with time, duplicate data contents being stored cannot be tolerated. Thus, employing storage optimization techniques is an essential requirement to large storage areas like cloud storage. Deduplication is a one such storage optimization technique that avoids storing duplicate copies of data. Currently, to ensure security, data stored in cloud as well as other large storage areas are in an encrypted format and one problem with that is, we cannot apply deduplication technique over such an encrypted data. Thus, performing deduplication securely over the encrypted data in cloud appears to be a challenging task. Various methods that address this challenge are studied in this paper.

Keywords : Deduplication,Cloud storage,Convergent Encryption;

1. Introduction

With numerous benefits of cloud storage such as cost savings, accessibility, scalability etc., users around the world tend to shift their invaluable data to cloud storage. As the data generation rates are increasing, it is a tedious task for cloud storage providers to provide efficient storage. Cloud storage providers uses different techniques to improve storage efficiency and one of leading technique employed by them is deduplication, which claims to be saving 90 to 95% of storage [1],[2]. Data Deduplication technique evolved as an simple storage optimization technique in secondary then widely adapted in primary storage as well as larger storage areas like cloud storage area. Now, data deduplication is widely used by various cloud storage providers like Dropbox [3],Amazon S3 [4], Google Drive [5],etc. Data once deployed to cloud servers, its beyond the security premises of the data owner, thus most of them prefer to outsource their in an encrypted format. Data encryption by data owners eliminates cloud service providers chance of deduplicating it since encryption and deduplication techniques have conflicting strategies, i.e., data encryption with a key converts data into an unidentifiable format called cipher text thus

* Corresponding author. Tel.: +91 953 949 4768;
E-mail address:akhilak777@gmail.com

Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016
doi:10.1016/j.procs.2016.05.123

encrypting, even the same data, with different keys may result in different cipher texts , making deduplication less feasible .However, performing encryption is essential to make data secure, at the same time, performing deduplication is essential for achieving optimized storage. Therefore, deduplication and encryption need to work in hand to hand to ensure secure and optimized storage. Various techniques and approaches used for deduplication over encrypted data are studied in this paper.

2. Background

2.1. Deduplication

Deduplication is basically a compression technique for removing redundant data. Fig 1 explains the deduplication process before storing data onto memory. Deduplication can be categorized as file level deduplication and block level deduplication based on granularity. File level deduplication takes into account the entire file, thus even small update or append makes the file different from previous version of it and thereby reducing deduplication ratio. Where as in case of block level deduplication data blocks are considered for deduplication. Deduplication can further categorized based on location of deduplication i.e., as client side deduplication and as source side deduplication. Performing deduplication at client side is ensures bandwidth saving since only hash value of file is sent to server, if duplicate is existing [6] , [7].Deduplication is widely is used various applications like backup, metadata management, primary storage, etc. for storage optimization [8].

2.2. Convergent Encryption

Convergent encryption [2], is an encryption approach that support deduplication. With convergent encryption, encryption key is generated out of hash of plain text. Thus applying these technique identical plaintexts would produce same cipher text, and this helps in performing deduplication further.

2.3. Proof Of Ownership

Deduplication works by computing cryptographic hash function on to data and using this hash value to determine similar data. Once a duplicate copy is found then new data is not uploaded but pointer to file ownership is updated thus saving storage and bandwidth. When it comes to client side deduplication, hash values of data are computed at client and send for duplicate check. An attacker, who gains access to hash value of a data which not authorized to him/her, may claim deduplication of file and thereby gaining access to the file. To defend such an attack, a Proof Of Ownership (PoW) has been proposed in[10], and various works like[10],[11], etc adapted this method. PoW works as an interactive algorithm between two parties - a prover and verifier to prove the ownership of the file. Verifier computes a short value of data M whereas, a prover need to compute short value of M and send it to verifier for claiming ownership of M [9],[10].

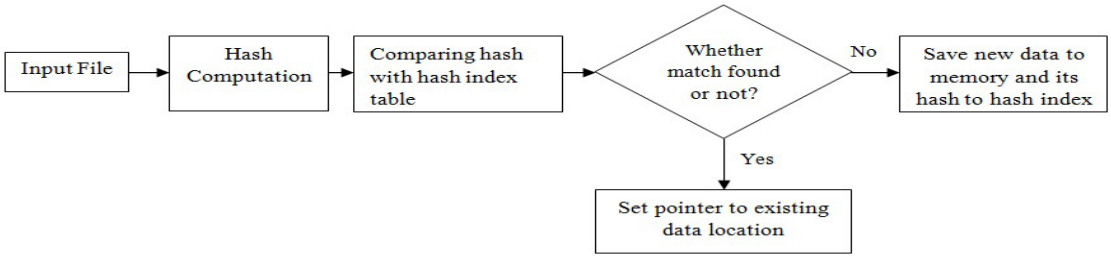

Fig. 1. Data Deduplication flow chart

3. Related Works

Bellare et.al [12] propose an encryption scheme wherein key for encryption and decryption are derived from message itself.MLE key generation algorithm maps the message M to a key K and further the encryption algorithm generates cipher text C of the message using key K. Ciphertext C is then mapped to a tag T, and this tag used for duplicate check by server. Keys used in MLE scheme are of fixed and shorter length thus does not result in much storage overhead.

Chen et.al [13] put forward a method to achieve dual level source based deduplication of large encrypted files with block key management and Proof of Ownership [10],[11]. Author claims that MLE scheme were proposed for target based file level deduplication and extending it to dual level deduplication requires much metadata management. In BL-MLE scheme with the given input file , a master key is generated and set of block keys for each message block in the file .With tag generation algorithms file tags and block tags are generated and further these tags are used checking equality of blocks and files ensuring security to it. Ownership of files or blocks proved and verified by using PowPrf and PowVrf algorithms in this approach.

In [14] encryption and decryption data is performed at client side and key for this is provided by key server located at cloud storage provider premises. Homomorphic encryption is used as the one of key management scheme in this approach.Data encryption key is first computed by the initial file uploader and further distributed consequent verified uploader by key server. Data encryption key used for encryption are further encrypted with the hash of file content.Data encrypted with data encryption keys are send to the storage server. HEDup ensures privacy while enabling deduplication. Key server discussed in this approach may become a bottleneck when number of clients increase in case of large scale deployment, and a decentralized deployment of key server is supposed as a solution.

In [15] Bellare et. al claim that Message locked encryption [13] are subject to Brute force attack and proposes a new architecture called DupLess where Brute force is resisted. Client receives message based keys, for encryption, from key server via a Oblivious Pseudorandom function (OPRF) protocol. With OPRF public key for encryption is shared among clients where as secret key resides with key server. With this method attackers cost of attack increased and chance is eliminated.

Puzio et.al in [16] propose ClouDedup,a secure and efficient storage service which assures block level deduplication[7] data confidentiality at the same time using convergent key encryption[2] added with block level key management .Architecture of ClouDedup proposes to prevent well known attacks against convergent encryption by embedding a user authentication mechanisms and access control mechanisms. Thus, a server encryption is applied on top of convergent encryption performed by user. For each data segment a signature is linked to it , and need to be verified for retrieving it. To deal with block level key management a metadata manager(MM) has been added to architecture.MM uses file table- to store meta data about file, pointer table-to manage storage and a signature table- to store meta data about signature for meta data management.

S. Bugiel et.al in [17] propose an approach that mainly involves two components - a trusted cloud and a commodity cloud. Trusted cloud is responsible for encrypting data and verifying operations performed on the commodity cloud. Security critical operations are performed by trusted cloud and queries to outsourced data are processed by commodity cloud. This approach claims protection against various security issues like leakage of data, computation manipulations, etc.

In [11] Li et.al propose a hybrid cloud approach to ensure security in deduplication which involves private cloud for providing tokens to access encrypted data in cloud. Data encryption technique employed here is convergent encryption [2] and PoW [9], [10] is used to ensure ownership eligibility to deduplicate the file. In [18] M. W. Storer et. al proposes to provide single server and distributed storage systems with data security and space savings. With this method key for encryption is generated out of data chunk. Even a full compromise of the system cannot reveal which data chunk is owned which user since the decryption information is encrypted with client's private keys. Two models for secure deduplication Authenticated model and Anonymous model are used in this method. An authenticated model is similar Convergent key construction [2]. Anonymous model hides identities of both authors and readers.

Li et.al, in [19] aims at addressing the problem of exposing and deduplicating sensitive data. With this approach, data chunks are distributed among multiple cloud servers. Furthermore to ensure tag consistency and data confidentiality, a deterministic secret sharing scheme is introduced in this distributed storage system. In contrast to the conventional deduplication-encryption method, here a secret sharing scheme is used instead of encryption method. Moreover, a Ramp secret sharing scheme [20],[21] is employed for key management.

The main objective of [22], is to address the problem of large key space overhead and to resist Brute force attack. For that, this method uses User Aware Convergent Encryption (UACE) and Multi Level Key Management (MLK).With UACE, cross-user file level and single user block level deduplication is achieved here. File level keys convergent encryption keys are generated by using a server –aided method whereas chunk level keys are generated via user-aided method. For reducing key space, chunk keys are encrypted using file level keys, thus increase in number of sharing users, key space is not increased. Furthermore, to eliminate the chance of single point of failure, this method uses multiple key servers, equipped with share-level keys that are generated out of file level keys and Shamir's secret sharing scheme [23] is used to communicate with these distributed servers.

In [24] data are differentiated based on popularity. A popular data implies it is shared among multiple user thus assumed to less sensitive and actively included in for deduplication with weaker security. Whereas unpopular data provided with security with semantically secure encryption. In [19] Xu et.al proposes a method that works with a weak leakage-resilient [9] than for cross user client side deduplication and providing security from outside adversaries and honest-but-curious cloud storage service providers.

Li et.al [26] address the problem of efficiently and reliably managing huge number of convergent keys for secure deduplication. With this Dekey approach both file level and block level deduplication is supported. Li et. al proposes a base line approach where user maintains a master key to encrypt convergent keys and develops Dekey approach out of it. With baseline approach user need to protect and manage large set of master keys , which a tedious task, thus a Dekey approach was proposed. In Dekey approach user need not manage any keys but distribute convergent keys among multiple servers. Ramp secret sharing scheme is used by Dekey approach for securely sharing convergent keys.

Table I does comparison between various methods, to make deduplication work with encrypted data.

Table 1 comparison between deduplication techniques carried over encrypted data

Approach	Encryption Scheme	Deduplication Strategy used
Message-locked encryption and secure deduplication	Message locked encryption	File level
BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication	Block Level Message locked encryption	Dual level: File level and Block level
HEDup: Secure Deduplication with Homomorphic Encryption	Homomorphic encryption	File level
DupLESS: Server-Aided Encryption for Deduplicated Storage	Enhanced Message level encryption to support security against Brute force attack	File level
ClouDedup:Secure Deduplication with Encrypted Data for Cloud Storage	Convergent encryption with added access control mechanisms	File level
Secure Deduplication with Efficient and Reliable Convergent Key Management	Convergent encryption	Block level
Twin clouds: An architecture for secure cloud computing	Convergent encryption	File level
A hybrid cloud approach for secure authorized deduplication	Convergent encryption	File level
Secure Data Deduplication	Convergent encryption	File level
A secure data deduplication scheme for cloud storage	Symmetric encryption on data categorized based on popularity	File level
Secure Distributed Deduplication Systems with Improved Reliability	Deterministic secret sharing scheme	File level and fine grained block level
SecDep: A User-Aware Efficient Fine-Grained Secure Deduplication Scheme with Multi-Level Key Management	User aware convergent encryption	File level and chunk level

4. Conclusion

Deduplication is a method available in cloud storage for saving bandwidth and storage capacity. But, deduplication is less feasible with encrypted data since, different key encryptions convert same data into different formats. In this paper various methods are discussed where deduplication methods are carried out on encrypted data in a large storage area. Most of the methods studied here work on the basis of convergent encryption, which is a simple approach that makes deduplication compatible with encrypted data. In this information dense world, we cannot compromise on both security and duplication of data across storage areas. A strategy needs to be formulated

which will enhance storage optimization without negotiating on encryption method; by providing deduplication technique in data storage servers where the available data is encrypted.

References

1. OpenDedup. OpenDedup, Global inline deduplication for Block Storage and Files. [online] 2010 Available from: http://opendedup.org/index.php.
2. J. Douceur, A. Adya, W. Bolosky, D. Simon, and M. Theimer. “Reclaiming space from duplicate files in a serverless distributed file system. In Distributed Computing Systems”, 2002. Proceedings. 22nd International Conference on, pages 617{624. IEEE, 2002.
3. Dropbox http://www.dropbox.com.
4. AmazonS3 http://aws/amazon.com/s3s.
5. GoogleDrive http://www.drive.google.com.
6. SNIA,"Advanced Deduplication Concepts,"[online] 2011.Available from http://www.snia.org/sites/default/education/tutorials/2011/fall/DataProtectionManagement/ThomasRiveria_Advanced_Dedupe_Concepts_FINAL.pdf
7. http://searchdatabackup.techtarget.com/tip/Where-and-how-to-use-data-deduplication-technology-in-disk-based-backup
8. Dutch T Meyer and William J Bolosky.”A study of practical Deduplication”. *ACM Transactions on Storage (TOS),* 7(4):14, 2012.
9. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, “Proofs of ownership in remote storage systems,” in Proc. ACM Conf. Comput. Commun. Security, 2011, pp. 491–500
10. Yang, Chao, Jianfeng Ma, and Jian Ren. "Provable Ownership of Encrypted Files in De-Duplication Cloud Storage." *Ad Hoc & Sensor Wireless Networks*26.1-4 (2015): 43-72.
11. Li, Jin, Yan Kit Li, Xiaofeng Chen, Patrick PC Lee, and Wenjing Lou. "A hybrid cloud approach for secure authorized deduplication." *Parallel and Distributed Systems, IEEE Transactions on* 26, no. 5 (2015): 1206-1216.
12. Bellare, Mihir, Sriram Keelveedhi, and Thomas Ristenpart. "Message-locked encryption and secure deduplication." *Advances in Cryptology–EUROCRYPT 2013*. Springer Berlin Heidelberg, 2013. 296-312.
13. Chen, Rongmao, Yi Mu, Guomin Yang, and Fuchun Guo. "BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication." (2015). *Information Forensics and Security, IEEE Transactions on* 26(2015), no. 12: 2643-2652.
14. Miguel, Rodel, and Khin Mi Mi Aung. "HEDup: Secure Deduplication with Homomorphic Encryption." In *Networking, Architecture and Storage (NAS), 2015 IEEE International Conference on*, pp. 215-223. IEEE, 2015.
15. Bellare, Mihir, Sriram Keelveedhi, and Thomas Ristenpart. "Dupless: Server-aided encryption for deduplicated storage." *Proceedings of the 22nd USENIX conference on security*. USENIX Association, 2013.
16. Puzio, Pasquale, Refik Molva, Melek Önen, and Sergio Loureiro."ClouDedup:Secure Deduplication with Encrypted Data for Cloud Storage." .In *Cloud Computing Technology and Science(CloudCom),2013 IEEE 5th International Conference on (Volume:1)p.363 – 370.*
17. S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider, “Twin clouds: An architecture for secure cloud computing,” in Proc.*Workshop Cryptography Security Clouds*, 2011, pp. 32–44.
18. M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, “Secure data deduplication,” in *Proc. 4th ACM Int. Workshop Storage Security Survivability*, 2008, pp. 1–10.
19. Li, Jie, Xia Chen, Xumin Huang, Song Tang, Yingmeng Xiang, Mehdi Hassan, and Abdul Hameed Alelaiwi. "Secure Distributed Deduplication Systems with Improved Reliability."*Computers, IEEE Transactions on* 64, no. 12(2015): 3569 – 3579.
20. G. R. Blakley and C. Meadows, “Security of ramp schemes,” in *Proc. Adv. Cryptol*., 1985, vol. 196, pp. 242–268.
21. A.D. Santis and B. Masucci, ‘‘Multiple Ramp Schemes,’’ *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1720-1728, July 1999.
22. Zhou, Yukun, Dan Feng, Wen Xia, Min Fu, Fangting Huang, Yucheng Zhang, and Chunguang Li. "SecDep: A User-Aware Efficient Fine- Grained Secure Deduplication Scheme with Multi-Level Key Management." *Mass Storage Systems and Technologies (MSST), 2015 31st Symposium on,* pp. 1-14.
23. A. Shamir,“How to share a secret,” Commun. ACM, vol. 22, no. 11,pp. 612–613, 1979.
24. J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, “A secure data deduplication scheme for cloud storage,” *Tech. Rep. IBM Research, Zurich*, ZUR 1308-022, 2013.
25. Xu, Jia, Ee-Chien Chang, and Jianying Zhou. "Weak leakage-resilient client-side deduplication of encrypted data in cloud storage." *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*. ACM, 2013.
26. Li, Jin, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick PC Lee, and Wenjing Lou. "Secure deduplication with efficient and reliable convergent key management." *Parallel and Distributed Systems, IEEE Transactions on* 25, no. 6 (2014): 1615-1625.