International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 17 (2018) pp. 13000-13004

A Survey on Secure Data De-Duplication in Cloud Storage

[1]M.SriRama Lakshmi Reddy, [2]Dr. K.Rajendra Prasad

[1]Research Scholar JJT University, Jhunjhunu, Churela, Rajasthan-333001, India.

[2]Professor & HOD of CSE Department, Institute of Aeronautical Engineering, Dundial, Hyderabad-501401, India.

Abstract

The basic notion of de-duplication is storage of duplicate data only for a single time. Thus, a user willing to upload a stored file will have to be first added by the cloud provider in the owner list for that particular file. This is the reason why de-duplication has been rapidly adopted by various providers of cloud storage. Today, it has become a popular approach for minimizing storage space and for uploading bandwidth and assists largely in increasing the scalability of data. De-duplication also eliminates the fear of surplus data by maintaining a single physical copy and refers any surplus data to this copy and is the best alternative of multiple data copies having the same data. This survey of literatures has critically evaluated multiple algorithms and techniques on secure de-duplication approaches through efficient and reliable methodologies. The finding show that a blend of secured de-duplications approaches with enhanced security features would offer extraordinary security options for a successful and reliable de-duplication.

Keywords: Encryption; De-Duplication; convergent encryption.

INTRODUCTION

The rising popularity of cloud computing can be attributed to lower costs, easily usable processing resources and increased storage. There has also been an unexpected rise in the use of online digital data which multiplies the significance of cloud storage for effective costing and better utilization of power. With increasing volume in data, the Total Cost of Ownership (TCO) is also increasing including human organization, management and storage setup cost. Thus, the most important thing with respect to cloud storage systems is to reduce the volume of transferable and stored data. This is quite beneficial for administrative and storage costs and application presentation [3]. This makes data de-duplication the most popular and significant feature with respect to saving on costs. It comprises of methods for storing a solo copy of the dismissed data and also provides links to reach that copy rather than storing this data's original copies. It also serves an important purpose for backup as it assists in transition of the services from a tape to the disk. Data de-duplication provides for savings on network bandwidth and disk space by transmitting and storing a single copy of the duplicate data [1].

Management of huge volumes of data is the key challenge of all the cloud storage services. The volume of data is expected to cross 50 trillion gigabytes by the year 2020. Constant rise in the total number of the users and the data size has made de-duplication indispensible for cloud storage. De-duplication is of two different types like file-level at the block-level and at the file-level. The former is associated with the entire file whereas the latter is associated with a variable or fixed sized data block. On the other hand, convergent encryption offers a good option for maintaining data privacy and realizes de-duplication at the same time [7]. It is a kind of cryptosystem which creates vague cipher text from similar plaintext files without considering the encryption keys. A convergent key is used for decrypting/encrypting data derived mainly by evaluating the value of cryptographic hash for the data copy. The keys are retained following data encryption and generation of key and the cipher text is then sent to the concerned cloud. Encryption is known to be deterministic which is why similar data copies result in similar convergent keys and cipher text [2]. Thus, de-duplication becomes possible on these cipher texts. Also, these cipher texts are decrypted only by their respective data owners and with their own convergent keys only.

The main objective of this paper is to explore various approaches and methodologies proposed by different researchers and scholars for a secure de-duplication approach. The paper critically evaluates these techniques and presents an extensive evaluation.

A. Existing Primitive Methods in Cloud De-Duplication

A number of security methods have to be applied for securing de-duplication; this section discusses old time approaches of secure de-duplication.

B. Encryption

A study by Storer et.al (2008) asserts that traditionally, encryption mandated users to use their own keys for data encryption. Thus, similar data copies belonging to different users would result in varying cipher text. This is the reason why de-duplication is not compatible with this form of encryption [4].

C. De-duplication with Reliable and Efficient Convergent Key Management

Li et.al (2014) discusses the secure De-duplication with Reliable and Efficient Convergent Key Management. As

described by these researchers the basic problem of reliable and efficient key management is addressed by de-duplication. The Ramp secret sharing system along with Dekey is implemented for the adoption of varying confidentiality and reliability levels. Data security is also maintained using convergent encryption and key management approach. De-duplication is carried out on both block level and file level [1]. One must note that the convergent keys are spread across a number of servers but key servers are always limited. Attention must be paid to key space overhead.

D. SecDep

He et.al [5] proposes a system namely SecDep which is a Fine-Grained and User-Aware Secure De-duplication System using Multi-Level Key Management. SecDep makes use of User Aware Convergent Encryption that assists in minimizing the computation costs also resisting brute force attack to a great extent. Usage of Multi-Level Key Management reduces the key space overheads as the file-level keys are split into share-level keys and are spread across various servers ensuring reliability and security of the file-level keys [5]. The main drawback is that multi-level key management also helps in reducing the time overhead.

E. Secure De-duplication and Message-Locked Encryption (MLE)

According to Puzio et.al (2013) secure De-duplication and Message-Locked Encryption (MLE) is a cryptographic primitive and the message helps in deriving the key for which decryption and encryption is carried out. It plays a significant role in achieving secure de-duplication which is the key target of many cloud-storage providers today. It also serves as a definition for both privacy and integrity and is referred to as tag consistency [7]. This is further supported by Yuon and Yu (2013) who describe that the different mechanisms for reclaiming the space from supplementary duplication which makes it open for measured file replication. The primary thing included here is convergent encryption which merges the duplicate files into the space meant for a solo file even when the files have been encrypted by varying user keys. The second important thing here is SALAD which is an abbreviation for Self-Arranging, Lossy, Associative Database meant for aggregating the location information and file content in a fault-tolerant, scalable and de-centralized manner [8].

MLE is an advanced cryptographic primitive where decryption and encryption are carried out under keys acquired from the messages. In a way, MLE is the best option for attaining secure de-duplication. It also serves as the definition for tag consistency as it provides for both integrity and privacy of data. Thus, theoretical and practical contributions are being made on this basis as well [12]. Practically, a ROM security analysis is also provided for MLE schemes comprising of the deployed methods. Theoretically, the key challenge lies in the standard model. The drawback of this method is that convergent encryption results in various convergent keys that cannot be managed easily with increase in number of users. It is significantly affected by a brute-force attack [9].

F. Server-Guided Encryption

A study by Min et.al (2011) discusses a Server-Guided Encryption for De-duplicated Storage (DupLESS) architecture. DupLESS can be defined as an architecture, which assists in secure de-duplicated storage and also resists brute-force attacks. The PRF protocol generates the key server which provides for the message-based keys which are then used by the clients to encrypt. It assists the clients in storing encrypted data using existing service which performs the de-duplication on part of the clients achieving good confidentiality at the same time [11].The drawback of this method is that the Get and Put operations are highly time-consuming and computational costs are quite large at the chunk level [7].

G. Proofs-of-ownership

A study by Ng et.al (2012) discusses the Proofs-of-ownership in Remote Storage Systems. The solutions associated with encoding and Merkle Trees have been discussed along with identification of attacks which exploit the client-side de-duplication efforts for identifying de-duplication. The concept of Proofs-of-ownership (PoWs) states that the Client holds the entire file data and not just the basic information about it and this is what is proved to the server. This approach helps in producing a proof that the user is capable of retrieving the target file F for a back-up service or archive [14]. This indicates that the archive transmits and retains file data reliably and allows the user to recuperate F entirely. POR is a form of cryptographic Proof of Knowledge (POK) designed specifically to deal with a file F which is large in size. Also, Proofs of Ownership (POW) are introduced to deal with attacks and permits the client to prove efficiently to the server that the file is retained by them. The notion of POW is formalized based on security definitions and competence requirements of the Petabyte scale systems of storage [12]. The drawback of this method is indicated by performance measurements that this system incurs quite a smaller overhead when compared with de-duplication taking place on the client's side.

LITERATURE SURVEY

A. Distributed De-duplication Systems with Better Reliability

Stanek et.al (2014) talks about secure Distributed De-duplication Systems with Better Reliability. A Distributed De-duplication system has been proposed by the author exhibiting greater reliability along with integrity and confidentiality of the data. It also supports de-duplication at the block and file level. Ramp Secret Sharing Scheme RSSS has been used for better reliability and for providing greater fault tolerance. Tag Generation Algorithm and RSSS are used for confidentiality. Message Authentication Code (MAC) is used for integrity which is a great support for the process of de-duplication. The main drawback is that it considers only two of the attack types like attack for Collusion and Attack for Dishonest System [11].

A detailed implicit or explicit definition on security attacks or proofs for various signature schemes and identity-based identification has been provided by Bellare et.al (2013). The framework given here explains the ways in which the schemes have been acquired along with modular security analyses assisting in unifying, simplifying and understanding previous works. The key focus is on construction of standard folklore which offers signature schemes and identity-based identification devoid of randomized oracles [4]. Authors stress on a converse de-duplication storage system which has been optimized for updated backups. It has been seen that de-duplication eradicates duplicates but at the same time reduces read performance as it brings about fragmentation [17].

B. Dekey approach

Li et.al (2015) stress on the concept of scheme and security in concentrated security framework to carry out Symmetric encryption. They highlight different security notions and study the concrete reduction complexity existing between them. The next step is providing concrete security investigation of the different encryption methods with the help of a block cipher which includes two common methods like Counter Mode and Cipher block chaining [6].

Also as suggested by Li et.al (2015) and Cochran (2012) the Dekey approach has been explained here with the help of the Ramp secret sharing scheme (RSSS) for storing the convergent keys. Talking specifically, RSSS includes n, k and r where n > k > r>= 0 and generates 'n' number of shares from a particular secret. The primary step is to recover the secret from 'k' number of shares but not less than that. The next step is to not deduce any information from the secret from 'r' number of shares. It is well-known that RSSS turns out to be the (n, k) when r =0 which is referred to as the Information Dispersal Algorithm (IDA). Also, when r =k-1, the (n, k, k-1), RSSS turns out to be the (n, k) referred to as the Secret Sharing Scheme (SSSS) [6] [10].

Two different models were developed by Li.et.al (2015) for both anonymous and authenticated secure de-duplicated storage. Both the designs indicate that it is possible to merge security with de-duplication such that it provides various security features. These models offer security by using convergent encryption. This method had initially been introduced with respect to the Farsite method providing a deterministic approach of creating an encryption key so that both the users are capable of encrypting data to the common cipher text [2] [1]. A map is developed in both anonymous and authenticated models for every file describing ways to rebuild a particular file from its basics. A unique key is used to encrypt this file. Scholars have shown ways to safeguard the confidentiality of data by changing the foreseeable messages into unpredictable messages. A key server is also implemented as a third party for generating file tags for carrying out duplicate checks. An advanced encryption approach was presented for providing differential security for unpopular and popular data [1].

C. Twin clouds Architecture

Bugiel et.al (2015) has proposed the architecture of secure outsource of data and computations to un-trusted cloud. The user communicates with trusted cloud, that does both encryption and verification of stored data and processes occurred in un-trusted cloud. This divides computations so that trusted cloud is used in security-critical operations in set up phase which are less time-critical, whereas queries to outsourced data are being processed parallel with fast cloud on the encrypted data [15].

D. Private Data Deduplication protocols used in Cloud Storage

One of most significant issues in cloud storage is storage capacity utilization. This paper considers two types of data deduplication strategies and extend fault tolerant scheme of digital signature ass proposed by Ng et.al (2017) by examining the redundancy of blocks for achieving data deduplication. Proposed scheme reduces cloud storage capacity and also enhances speed of the data deduplication [16]. Also, the signature is evaluated for each of the uploaded file to verify integrity of the files.

E. Advanced Secure deduplication using Convergent Key Management

The data deduplication is made use of in removing duplicate data and is widely used in cloud storage for reducing both storage and uploading bandwidth. Though it is promising, there is a challenge to achieve secure deduplication in the cloud storage, convergent encryption is being used extensively for secure deduplication, there is an uncertain issue of having convergent encryption to be practical and manage large number of convergent keys managed efficiently. The techniques are key management and convergent Encryption [17].

CRITICAL COMPARISON OF PROMINENT APPROACHES OF CLOUD DE-DUPLICATION

Research Authors	Technique	Key Features	Outcome
Bugial et.al (2015)	Twin Clouds-Architecture for secured cloud computing	• Secure computing • Low latency • Stores huge amounts of data • Environment for secure execution	Client uses trusted cloud as proxy, which provides clearly defined interface for managing outsourced programs, data and queries.
Li et.al (2015)	Dekey approach	• It is possible to merge security with de-duplication such that it provides various security features. • It offers security by using convergent encryption.	A unique key is used to encrypt this file. Scholars have shown ways to safeguard the confidentiality of data by changing the foreseeable messages into unpredictable message

Ng et.al (2017)	Private data de-duplication	• Improved speed of data duplication • Reduces cloud storage capacity • Fault tolerant	Enhanced efficiency of data
Bosman et.al (2016)	Advanced Secure deduplicatio n	• Efficient • reduces bandwidth and storage space • reliable key management • Offers confidentiality	Convergent key share over multiple server
Bellare et.al (2013)	DupLESS Server-aided encryption for the de-duplicated storage	• Saving of space • Resolving cross user de-duplication • Security: Against external attacks • High performance	Simple Interface
Forman et.al (2017)	Distributed De-duplication	• Saves time • High security • Identification of attacks • Savings in bandwidth	Performance measurements show that scheme has only small overhead compared to client side deduplication

The critical comparison of different prominent methodologies shown in the above table highlights key findings like:

In Twin Clouds architecture, a secure cloud computing has a Client using trusted Cloud as proxy providing clearly defined interface for managing outsourced data, queries and programs and it has low latency and provides secure execution environment. DupLESS Server Aided Encryption for Deduplicated Storage is deployed for simple storage interface which also offers a strong security against external attacks such as brute force attacks. It also has high performance and resolves cross user duplication. The distributed deduplication and advanced de-duplication provides Performance measurements indicating that scheme has only a small overload in comparison with naïve client-side deduplication. This also saves bandwidth and identifies attacks. The protocols for Private data deduplication in cloud storage boost the efficiency of data and also speed of data duplication. A dekey approach having reliable and efficient key management reduces storage space along with bandwidth. Convergent key can be shared across multiple servers.

CONCLUSION

The best choice for optimizing the storage space and upload bandwidth over cloud is Source Based De-duplication. The process of distributed de-duplication helps in achieving reliability, confidentiality and security of data. A combination of both the methods helps in achieving greater de-duplication ratio and data reliability. Also, good de-duplication ratio can be achieved by modifying de-duplication algorithm further.

Some of the authors stressed on encrypted de-duplication with secure and fast laptop backups. Today, people have started storing large volumes of corporate and personal data on their personal computers and laptops. The problem is discontinuous and poor connectivity which makes their data susceptible to theft and sometimes failure of the hardware is also an issue.

This paper has reviewed various approaches and algorithms which can be used for making the most of the data common to multiple users for increased backup speed and for reducing the storage needs. These algorithms provide great support in client-end per-user encryption and maintaining the confidentiality of personal data [4].

The basic foundation here is that secure de-duplication services are deployable using additional features on security for both outside and inside attacker with the help of detecting masquerade activities. A very important role is played by the deterrence effect, attacker's confusion and additional expenses in prevention of masquerade activities by the risk averting attackers. Thus, it is believed that a blend of such security features would offer extraordinary security options for de-duplication.

REFERENCES

[1]. Li, J., Chen, X., Li, M., Li, J., Lee, P. P., & Lou, W. (2014). Secure deduplication with efficient and reliable convergent key management. *IEEE transactions on parallel and distributed systems*, *25*(6), 1615-1625.

[2]. Li, J., Li, Y. K., Chen, X., Lee, P. P., & Lou, W. (2015). A hybrid cloud approach for secure authorized deduplication. *IEEE Transactions on Parallel and Distributed Systems*, *26*(5), 1206-1216.

[3]. Bellare, M., Keelveedhi, S., & Ristenpart, T. (2013, May). Message-locked encryption and secure deduplication. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques* (pp. 296-312). Springer, Berlin, Heidelberg.

[4]. Storer, M. W., Greenan, K., Long, D. D., & Miller, E. L. (2008, October). Secure data deduplication. In *Proceedings of the 4th ACM international workshop on Storage security and survivability* (pp. 1-10). ACM.

[5]. He, Q., Li, Z., & Zhang, X. (2010, October). Data deduplication techniques. In *Future Information Technology and Management Engineering (FITME), 2010 International Conference on* (Vol. 1, pp. 430-433). IEEE.

[6]. Li, J., Chen, X., Huang, X., Tang, S., Xiang, Y., Hassan, M. M., & Alelaiwi, A. (2015). Secure distributed deduplication systems with improved reliability. *IEEE Transactions on Computers*, *64*(12), 3569-3579.

[7]. Puzio, P., Molva, R., Onen, M., & Loureiro, S. (2013, December). ClouDedup: secure deduplication with encrypted data for cloud storage. In *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on* (Vol. 1, pp. 363-370). IEEE.

[8]. Yuan, J., & Yu, S. (2013, October). Secure and constant cost public cloud storage auditing with deduplication. In *Communications and Network Security (CNS), 2013 IEEE Conference on* (pp. 145-153). IEEE.

[9]. Li, A., Jiwu, S., & Mingqiang, L. (2010). Data deduplication techniques. *Journal of Software*, *21*(5), 916-929.

[10]. Cochran, W. T. (2012). *U.S. Patent No. 8,199,911*. Washington, DC: U.S. Patent and Trademark Office.

[11]. Min, J., Yoon, D., & Won, Y. (2011). Efficient deduplication techniques for modern backup operation. *IEEE Transactions on Computers*, *60*(6), 824-840.

[12]. Kogelnik, C. (2012). *U.S. Patent No. 8,117,464*. Washington, DC: U.S. Patent and Trademark Office.

[13]. Stanek, J., Sorniotti, A., Androulaki, E., & Kencl, L. (2014, March). A secure data deduplication scheme for cloud storage. In *International Conference on Financial Cryptography and Data Security* (pp. 99-118). Springer, Berlin, Heidelberg.

[14]. Ng, W. K., Wen, Y., & Zhu, H. (2012, March). Private data deduplication protocols in cloud storage. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing* (pp. 441-446). ACM.

[15]. Bugiel, S., Nurnberger, S., Sadeghi, A., & Schneider, T. (2011, March). Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011) (Vol. 1217889).

[16]. Ng, W. K., Wen, Y., & Zhu, H. (2012, March). Private data deduplication protocols in cloud storage. In Proceedings of the 27th Annual ACM Symposium on Applied Computing (pp. 441-446). ACM.

[17]. Bosman, E., Razavi, K., Bos, H., & Giuffrida, C. (2016, May). Dedup est machina: Memory deduplication as an advanced exploitation vector. In 2016 IEEE symposium on security and privacy (SP) (pp. 987-1004). IEEE.