SURVEY ON SECURE PRIVILEGED BASED DATA DEDUPLICATION IN CLOUD USING TWIN CLOUD

Mr. Yendhe A.[1], Ms. Dumbre T.[2], Ms. Mahadik S.[3], Ms. Gholap A.[4], Prof. Gunjal A.[5]

[1] *Student, Department of Computer, SVCET, Maharastra, India*
[2] *Student, Department of Computer, SVCET, Maharastra, India*
[3] *Student, Department of Computer, SVCET, Maharastra, India*
[4] *Student, Department of Computer, SVCET, Maharastra, India*
[5] *Assistant Prof., Department of Computer, SVCET, Maharastra, India*

ABSTRACT

Deduplication of data is one of important data compression techniques for reducing duplicate copies of repeating data and has been more popular used in cloud storage to reduce the amount of cloud storage space and save bandwidth. Security in data deduplication can be provided with the use of convergent encryption technique which encrypts the data before uploading it to public system. The the limitations of convergent encryption drives researchers towards building more sophisticated data deduplication techniques which can fulfil current organizational needs. This paper makes an attempt to survey data deduplication system techniques available in cloud technology..

Keyword: - *Deduplication, authorized duplicate check, confidentiality, hybrid cloud, symmetric encryption, asymmetric encryption, convergent encryption, POR, POW.*

1. INTRODUCTION

Cloud computing provides many virtualized resources to users as services across the entire Internet, while hiding platform and implementation details. GMAIL is one of the best examples of cloud storage which is used by most of us regularly [2][24]. Cloud computing uses virtualization technique and thus hiding platform and implementation details. This provides unlimited resources to users on their devices with just internet connection. Cloud service providers provide highly available cloud database storage and slightly parallel computing resources as compare low costs. As cloud computing becomes obtain, an increasing amount of data is being collected in many resources and stored the cloud and share data by users with specified protocols, which shows the access authorized of the stored data. One difficult challenge of cloud storage services is the management of the ever-increasing amount of data. To control data management scalable in cloud computing, deduplication has been a well-known technique and has attracted more and mostly used in today's world [1][4][23][25].

Data de-duplication is a important data compression technique for removing replication copies of repeating data in cloud storage. This technique is used to increase storage utilization and can also be used to network data sharing to reduce the number of data bytes that must be sent. Therefore can be. keeping multiple data a copy with the same content, deduplication remove redundant data by keeping only one original copy and use other redundant data to that copy. Deduplication can be applied at the file level or the block level. For file level de-duplication, it removes duplicate copies of the same file. De-duplication can also take place at the block level, which removes duplicate blocks of data that occur in duplicate files.

The sequence of rest of paper is as follows: Section 2 covers literature review and section 3 contains concluding remarks.

2. LITURATURE SURVEY

1.1 A Hybrid Cloud Approach for Secure Authorized De-duplication

De-duplication of data has many forms. Typically, there is no one best way to implement data de-duplication across an whole an organization. Instead, to maximize the benefits, organizations may deploy more than one de-duplication strategy. Cloud data storage services mostly refer de-duplication, which removing redundant data by storing only single copy of every file or block [1]. It is very essential to know the backup and backup challenges, when selecting de-duplication as a solution.

Advantages:

This De-duplication technique reduces the space and bandwidth requirements of data storage services, and is most effective when applied with multiple users, a common practice by cloud storage offerings.

Limitations:

Data deduplication does not work with traditional encryption techniques. While using data deduplication technique it should not reduce fault tolerance mechanism.

Types of data de-duplication are described below:

File-level de-duplication:

This de-duplication technique is commonly called as single-instance storage, file-level data de-duplication compares a file that has to be archived or backup that has already been stored by checking all its attributes against the index. The index is updated and stored only if the file is unique, if not than only a pointer to the existing file that is stored references. Only the single instance of file is saved in the result and relevant copies are replaced by”stub” which points to the original file [1].

Block-level de-duplication:

Block-level data de-duplication operates on the basis of sub-file level. As the name implies, that the file is being broken into segments blocks or chunks that will be examined for previously stored information vs redundancy. The popular approach to determine redundant data is by assigning identifier to chunk of data, by using hash algorithm for example it generates a unique ID to that particular block. The particular unique Id will be compared with the central index. In case the ID is already present, then it represents that before only the data is processed and stored before. Therefore only a pointer reference is saved to the previously stored data. If the ID is new and does not exist, then that block is unique. The unique chunk is stored and the unique ID is updated in the Index. The size of the chunk which needs to be checked varies from vendor to vendor [1].

1.2 Content Addressable Storage

Eliminating multiple copies of any file is a form of the de-duplication. Single instance storage (SIS) environments can detect and eliminate redundant copies of identical files. After a file is stored in a single-instance storage system than, all the other references to same file, will refer to the original, single copy. Single instance storage systems compare the content of files to detect if the incoming file is identical to an existing file in the storage system. Content-addressed storage is typically combined with single-instance storage functionality [5]. While file-level de-duplication avoids storing files that are a duplicate of another file, many files that are considered unique by single-instance storage measurement may have a huge amount of redundancy within the files or between files. For example, it would take only one small element (e.g., a new date inserted into the title slide of a presentation) for single-instance storage to through two large files as being different and requiring them to be stored without further de-duplication [7].

Advantages:

CAS system provides higher searching speed for documents.

Limitations:

This system only provides performance benefits when there are more read operations than update operations.

1.3 Convergent Encryption

Convergent encryption provides data confidentiality in de-duplication. A user (or data owner) derives a convergent key from each original data copy and encrypts the data copy with the convergent key. The basic idea of convergent encryption (CE) is to derive the encryption key from the hash of the plaintext. The intelligible implementation of converging encryption can be defined as follows: Alice obtain the encrypt key from her message M such that K = H(M), where H is a cryptographic hash function; he or she can encrypt the message with the help of key, hence: C = E(K;M) = E(H(M);M), where E is a block cipher texts. By using this technique, two users with two identical plaintexts will obtain two identical constant size cipher texts since the encrypt key is the same; hence the cloud storage provider will be authorized to perform de-duplication on such cipher texts. Therefore, encryption keys are generated, keeping and protected by users. As the encryption key is deterministically generated from the plaintext, users do not have to communicate with each other for establishing an agreement on the key to encrypt a given plaintext. Therefore, converging encryption seems to be a good candidate for the adoption of encryption and deduplication in the cloud storage domain. In addition, the user derives a tag for the data copy, such that the tag will be used to detect duplicates [9][11].

A typical convergent encryption is insecure because brute-force attack launched by cloud server can recover files falling into known set. To understand this, consider that public cloud server knows the given cipher texts file is drawn from message space S = {F1, F2, …, Fn} of size n, then it it can recover files file F using at most n on-line encryptions. This can be done by encrypting each file Fi where i = {1, 2,…,n} to get encrypted text Ci. If C = Ci this means underlying file is Fi. This means convergent encryption is insecure for predictable files.

M. Bellare design a system, Duplication less than combines a CE-type scheme with the ability to derived message-obtain keys with the help of a key server (KS) shared by a group of users. The users interact with the KS by a protocol for oblivious PRFs, ensuring that the KS can cryptographically mix in secret data to the per-message keys while do not learning anything about files stored by users. These mechanisms obtain that Duplication less provides strong security against external attacks and that the security of Duplication less inviting degrades in the face of comprised systems. Require a user can be compromised, learning the plaintext underlying another users cipher text requires mounting an online brute force attacks [11][18]. M. Bellare is to formalize a new cryptographic primitive, Message-Locked Encryption (MLE), where the key encryption and decryption are performed is it derived from the message. MLE provides a way to accomplish secure de-duplication, a aim formerly targeted by numerous cloud database storage providers. They supply definitions of privacy and a form of integrity that they call tag consistency. They provide ROM security analyses of a natural family of MLE schemes that includes deployed techniques. They built connections with deterministic encryption, hash functions secure on correlated inputs [10].

Another type criteria is the location at which deduplication is applied if data are de-duplicated at the user, then it is called source-based de-duplication, otherwise target-based. In source-based de-duplication, the user first hashes every data segment he wishes to upload and sends these result performances to the storage provider to check whether such data are already stored: thus only not de-duplicated data segments will be actually uploaded by the user. While de-duplication at the user side can obtain bandwidth savings, it unfortunately can built the system vulnerable to side channel attacks whereby attackers can instantly discover whether a secure data is stored or not. On the other side, by de-duplicating data at the storage provider, the system is protected against side-channel attacks but such solution does not decrease the communication performance.

Wang et al. proposed a new system which provides secure and efficient access to outsourced data [5]. Here the end user sends a request for data access to the data owner, after that the data owner will send back an encryption key and access certificate to an end user, and then the end user will send that access certificate to the data storage provider and the data storage provider will send the encrypted data blocks to the end user. The advantage of their approach is that, it has a low storage overhead, but it requires cloud server support to enforce policies.

Roxana et al. proposed a new system which overcomes the drawback of wang's approach. System supports the use of multiple policies. Here we focus on new approach which is named as FADE [13]. Authors in their study have proposed a new protocol called vanish which provides data privacy and self deleting data. The study is mainly focused on the data and it could be able to access for a limited period of time. After the time expiry, the data is not accessible to the users nor to the data owner. Vanish protocol is applicable for only sensitive data. To has self-deleting property, the activities takes place are, Vanish first encrypts user's data locally by taking the help of encryption key and the encryption key will not be known to the user also, then it destroy local copies of key and after that it sprinkles bits in DHT randomly. The drawback of this system is it provides the assured deletion based on time. Even the legitimate users may not be able to access the data after time expiration [20].

Sven B. et al. [10] proposed twin cloud architecture for secure deduplication in cloud storage. As the name suggest their approach uses one public cloud and one private cloud, User communicates with a private cloud (organization maintained cloud) which encrypts data before outsourcing to public cloud. This private cloud is also responsible for verification of stored data in public cloud. Their architecture uses private cloud for operations requiring security whereas other kind of queries is processed by public cloud. Their technique allows maximum utilization of resources of private cloud, and only high load queries are processed on-demand by the public cloud.

Trusted Cloud requires constant amount of storage and is used constantly in the Setup Phase for pre-computing encryption. The public cloud provides large amount of storage and is used for time-critical Query operations.

Zhang et al. also proposed a hybrid cloud [1] [7] system named Sedic [7]. The system supports the privacy aware data computing. The system is based on MapReduce fuction. They address the problem of authorized deduplication of public cloud data. Here the private cloud is assumed as honest but curious.

Advantages:

Convergent encryption provides security while deduplication process. Security in deduplication process can be increased using twin cloud approach or hybrid cloud approach and using random encryption keys.

Limitations:

Increase in complexity in deduplication process is main limitation in above approaches.

1.2 Proof of Ownership (POW)

The POW protocol allows user to efficiently prove to a cloud server about his ownership, rather than short information about the file such as a hash value. This is somewhat similar to proofs of retrievability (POR) and proofs of data possession (PDPs) with a role reversal here client is the proover is cloud server. Pietro et.al [11] proposed three correlative protocols to achieve an efficient POW for deduplication. The main idea of their protocols is to challenge random K bits of file F. The probability that a malicious user is able to output the correct value of K bits of the file where each bit is selected at a random position is negligible in security parameter k, but their scheme cannot be adopted for encrypted files.

To overcome such attacks, Halevi et. al [9] introduced the notion of POW for client-side de-duplication. In addition, they presented Merkle-tree based schemes to allow a user to efficiently prove his ownership to the server, rather than some short information. However, their scheme cannot be adopted for encrypted file scenario, because encryption of the same file by different users with random keys results in different ciphertexts. The server cannot store the same hash root value for the ownership verification.

1.2 Proof of Retrieveability (POR)

A proof of retrievability (POR) is a compact proof by a file system (proover) to a client (verifier) that a target files F is intact, in the sense that the client can fully recover it. As PORs incur lower communication complexity as compare to transmission of F itself, they are an attractive building block for high-performance remote cloud storage systems. A POR is a protocol in which a server/archive proves to a client that target file integrity is valid, and thus client can recover their files whenever needed. In traditional POR, client needs to download file F and check the digital signature of that file to guarantee integrity [11]. The client can pre-process the file before uploading and insert some secret in that file, such that it can be used for checking consistency of file in PORs / PDPs technique.

Advantages:

Provides efficient mechanism for clients to verify the integrity of uploaded files on server.

Limitations:

The POR systems still have limitations including non-trivial (linear or quadratic) communication and computational cost, no support of public verifiability and malicious cloud server.

2. CONCLUSION

Data de-duplication is important technique used in cloud computing. But data deduplication technique can’t be used alone in cloud, because there is often need of data security. So data de--duplication and convergent encryption work in collaboration such that, data deduplication is possible with security of data. But convergent encryption does not provide much security, as it can be susceptible to guessing and brute force attacks. Also current data deduplication technique does not provide support for differential privilege level deduplication. This system is useful in currently changing industry where it is necessary to consider privilege levels of employees in data deduplication, so that, it will enhance data deduplication process and security. The POW is very useful in data deduplication as it makes system secure against several attacks.

REFERENCES

[1]. Jin Li, Yan Kit Li, Xiaofeng Chen,Patrick P. C. Lee and Wenjing Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication", IEEE Transaction On Parallel And Distributed System,Vol.PP,No.99, 2014.

[2]. Maneesha Sharma, Himani Bansal and Amit Kumar Sharma, " Cloud Computing: Different Approach & Security Challenge", IJSCE, Volume-2, Issue-1,March 2012.

[3]. Kangchan Lee, "Security Threats in Cloud Computing Environments", International Journal of Security and Its Applications, Vol. 6, No. 4, October, 2012.

[4]. Sashank Dara, "Cryptography Challenges for Computational Privacy in Public Clouds", International Journal of Security and Its Applications, Volume 4, 2002.

[5]. David Pointcheval, "Asymmetric Cryptography and Practical Security", International Journal of Security and Its Applications, Volume 4,2002.

[6]. Yogesh Kumar, Rajiv Munjal and Harsh Sharma, "Comparison of Symmetric and Asymmetric Cryptography with Existing Vulnerabilities and Counter-measures", International Journal of Computer Science and Management Studies, Vol. 11, Issue 03, Oct 2011.

[7]. Jan Stanek, Alessandro Sorniottiy, Elli Androulakiy, and Lukas Kencl, "A Secure Data De-duplication Scheme for Cloud Storage", IBM Research, Zurich, May 1996.

[8]. Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou, "Secure Auditing and De-duplicating Data in Cloud", IEEE Transactions on Computers,unpublished,2015.

[9]. Deepak Mishra and Sanjeev Sharma, "Comprehensive study of data de-duplication", International Conference on Cloud, Big Data and Trust,Vol.13.No.15,NOV 2013

[10]. Paul Anderson and Le Zhang, "Fast and Secure Laptop Backups with Encrypted De-duplication", Proceedings of Eurocrypt, Vol. 6,March 2013.

[11]. Mihir Bellare,Sriram Keelveedhi and Thomas Ristenpart, "Message-Locked Encryption and Secure De-duplication", Proceedings of Eurocrypt, Vol. 6,March 2013.

[12]. Pietro, R.D., Sorniotti, "Boosting Eciency and Security in Proof of Ownership for Deduplication", ACM Symposium on Information, 2012.

[13]. David Pointcheval,"Asymmetric Cryptography and Practical Security",International Journal of Security and Its Applications,Volume 4,2002.

[14]. Sashank Dara,"Cryptography Challenges for Computational Privacy in Public Clouds",International Journal of Security and Its Applications,Volume 4, 2002.

[15]. Sean Quinlan and Sean Dorward,"Venti: a new approach to archival storage",Bell Labs, Lucent Technologies,Vol. 6,1998.

[16]. Paul Anderson and Le Zhang,"Fast and Secure Laptop Backups with Encrypted Deduplication",Proceedings of Eurocrypt,Vol. 6,March 2013.

[17]. J. R. Waykole and S. M. Shinde,"A Survey Paper on Deduplication by Using Genetic Algorithm Along with Hash-Based Algorithm",Journal of Engineering Research and Applications,Vol.4,Issue.1, Jan 2014.

[18]. Zhe Sun,Jun Shen and Jianming Young,"A novel approach to data deduplication over the engineering-oriented cloud systems", Integrated Computer Aided Engineering, 2013.

[19]. C.-K Huang, L.-F Chien, and Y.-J Oyang, “Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs,” J. Am. Soc. for Information science and Technology, vol. 54, no. 7, pp. 638-649, 2003

[20]. C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013.

[21]. W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446. ACM, 2012.

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

[23]. Srivatsa maddodi, Girija V. Attigeri, Dr karunakar A.k, Data Deduplication Techniques and Analysis. Third International Conference on

[24]. Emerging Trends in Engineering and Technology IEEE, 2010

[25]. Borja Sotomayor, Rubén S. Montero and Ignacio M. Llorente, Ian Foster, Virtual Infrastructure Management in Private and Hybrid Clouds, Published by the IEEE Computer Society, 2009