Mathematical Statistician and Engineering Applications
ISSN:2094-0343
2326-9865

Securing Data Deduplication Using Convergent Encryption Algorithm

K. P. Saurabh[1]
[1]Research Scholar, Department of Computer Science and Engineering, Dr. A. P. J. Abdul Kalam University, Indore, Madhya Pradesh
Dr. Kailash Patidar[2]
[2]Supervisor, Department of Computer Science and Engineering, Dr. A. P. J. Abdul Kalam University, Indore, Madhya Pradesh

Article Info
Page Number: 13287 – 13292
Publication Issue:
Vol 71 No. 4 (2022)

Article History
Article Received: *25 October 2022*
Revised: *30 November 2022*
Accepted: *15 December 2022*

Abstract
Deduplication is a technique for minimizing data storage needs by getting rid of duplicate copies of files. The exponential growth in both the number of users and the amount of data stored in the cloud has made data deduplication an absolute must for cloud service providers. Their cloud providers will save money on storage and data transport by keeping only one copy of duplicate data. The goal is to explicitly address the issue of permitted data deduplication in order to better preserve data security. In addition to the data itself, we also take into account the varying levels of access each person has in the duplication check process. We have numerous novel deduplication architectures to provide approved duplicate check in a hybrid cloud environment. To prove the scheme is safe according to the suggested security model's definitions, a security analysis must be carried out. We develop a prototype of the approved duplication check mechanism we propose and test it on a testbed.
Keywords: - Deduplication, Cloud Storage,ConvergentKeys, Encryption, Algorithm

I. Introduction

Increased storage space, more readily available processing power, and lower prices all contribute to cloud computing's meteoric rise in popularity. Unpredictable growth in digital data usage has also increased the value of cloud storage for efficient resource management and lower operational expenses. The cost of human resources needed for data organization, administration, and storage infrastructure rises in tandem with data volume. As a result, the primary concern with regards to cloud storage systems should be the minimization of data transit and storage volumes. This is great news for application presentation and administrative expenditures. Given its importance to reducing expenses, data de-duplication has become a popular and widely adopted function. Methods for keeping a single copy of the discarded material and providing links to that copy instead of the originals are part of this strategy. With its help, services may be moved from a tape to a disk, which is a crucial step in the backup process. Since only one copy of the duplicate data needs to be transmitted and stored, data de-duplication helps reduce data transfer and storage costs.

The goal of the data-deduplication approach is to reduce unnecessary data storage. Traditional deduplication systems seek to identify and save just one copy of duplicate data chunks in storage in an effort to save on storage space. Instead of keeping duplicate data, we construct logical references to the other copies. By eliminating redundant data, deduplication can save valuable disk space and traffic. However, these methods often have unintended consequences for the fault tolerance of a system. In the

event of a failure, the availability of multiple files that depend on the same data piece might be compromised. This issue has prompted the development of several strategies and methods for boosting storage efficiency and fault tolerance. The deduplication process is followed by these methods, which offer data redundancy.

The biggest obstacle for all cloud storage systems is managing massive amounts of data. Data storage capacity is projected to exceed 50 trillion gigabytes by 2020. With an ever-increasing user base and larger data sets, de-duplication has become a necessity for cloud storage. File-level de-duplication and block-level de-duplication are the two main categories of de-duplication. In contrast to the former, the latter is linked to a chunk of data whose size might change over time or remain constant. Convergent encryption, on the other hand, is a viable alternative for protecting user anonymity while still enabling deduplication. Using the same or comparable plaintext files, this type of cryptosystem generates cryptic cipher text without taking into account the encryption keys. The value of the cryptographic hash for the data copy is the primary factor in determining the convergent key used for decryption/encryption. After data is encrypted and a key is generated, the keys are stored and the cipher text is uploaded to the relevant cloud. Convergent keys and cipher text generated from identical data copies are a well-established property of deterministic encryption. That's why these cipher messages can be de-duplicated now. In addition, only the data owners themselves, using their unique convergent keys, may decode the cipher messages.

II. Data Deduplication

In a nutshell, it's as easy as reiterating the same thing over and over again until the desired result is achieved. As a byproduct of data compression, duplication has been around for quite some time. Compressing data in-file to remove redundant information and add the original value to the index. By extending this idea by repetition, we get:

- Within a single file (complete agreement with data compression)
- Cross-document
- Cross-Application
- Cross-client
- Across time

Disk-based backup systems, which are storage systems created to minimize the use of storage space, frequently employ the data reduction approach of duplication. It's effective at fmding duplicate files in various storage locations and of varying data block sizes, but it operates in a distinct time frame. There is an indication replacing duplicate data blocks. It's a well-known fact that those who have a lot of experience with computers and know how to use them effectively have a significant advantage over others who are still learning the ropes. Further, data deduplication tools help users share information quickly and easily across several locations, which is crucial for a cost-effective data replication back-up strategy.

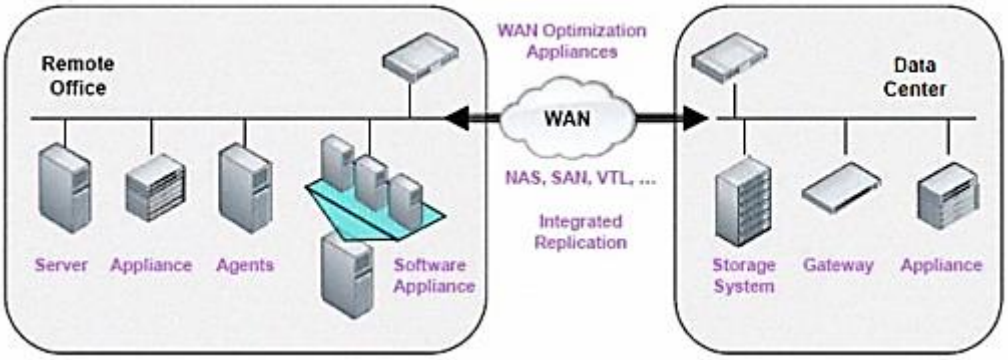

Figure 1: Where Deduplication Can Happened

To accomplish this, compression methods look for and remove duplicate bytes inside a document before saving it, but duplicate data can also be disseminated via the technique to be removed from several copies of the same file or data block.

Deduplication is a new kind of data compression that operates differently than traditional methods. To avoid saving duplicate information, data compression will only include the first data point in the index if two files are found to be similar. This is because the index will use the first data point to determine whether or not the two documents are different. In addition, it has data compression, as the first file does not include redundant information and so is smaller.

To further distinguish itself from incremental backups, duplication stands on its own. Incremental backups' primary purpose is to save only newly created data, while data de-duplication's primary goal is to store only a single copy of each piece of data, making the latter a more efficient tool for reducing data storage needs. The standard capacity of data storage systems is 1120, although most vendors say their data deduplication devices can lower it by a factor of 11. The foundation of data de-duplication technology is the identification and replacement of the single occurrence of a pointer inside a data block.

III. Methodology

To achieve this goal, we take into account a hybrid cloud architecture, which combines a public cloud and a private cloud, to store both public and private cloud data. Private cloud involvement as a proxy allows data owner/users to safely execute duplicate check with differential rights, unlike existing data deduplication solutions. Because of its usefulness, this kind of design has received a lot of attention from academics.

In the end, it all comes down to the same thing: the data owner knows best, therefore he or she uses that knowledge to their advantage. on this hybrid cloud architecture, the SCSP is hosted on the public cloud, and a novel deduplication technology that allows for differential duplicate check is presented.

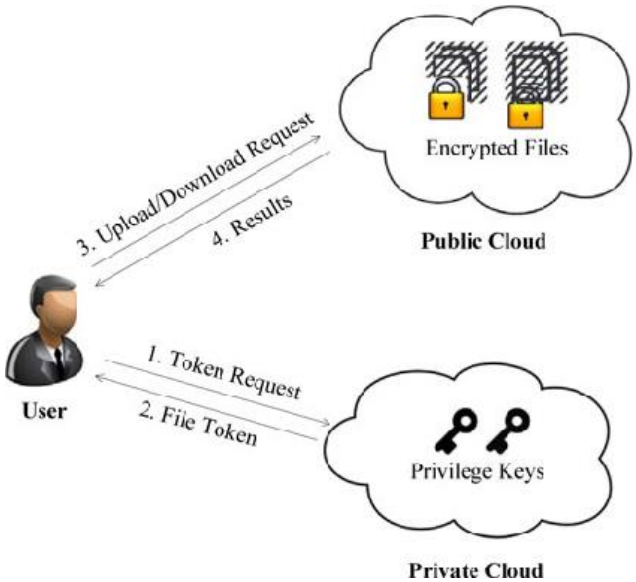

Figure 2: Architecture for secure data Deduplication

The user can only check for duplicates in files that have been given the appropriate permissions. Additionally, we tighten up the security of our system. By encrypting the file using differential privilege keys, we provide a sophisticated approach to support increased security. Users without the proper permissions will be prevented from doing the duplication check. Furthermore, even in concert with the SCSP, such unauthorized users are unable to decipher the encrypted text. Our system has been shown to be secure, according to the definitions provided by the suggested security model. We conclude by

developing a working model of the allowed duplication check proposal and conducting testbed tests to measure the prototype's overhead. Compared to standard convergent encryption and file upload methods, we demonstrate that the overhead is small.

Symmetric encryption

Symmetric encryption encrypts and decrypts data with the same secret key κ. There are three fundamental operations that make up a symmetric encryption method.

Algorithm

- Step 1: $KeyGen_{SE}(1^{\lambda})$ → The key k is generated via the k key generation technique with the 1^{λ} security parameter.
- Step 2: Enc_{SE} (k, M) → The cipher text C is the result of a symmetric encryption algorithm (C) that inputs the secret key k and the plaintext message M.
- Step 3: Dec_{SE} (k, C) → Using the key k and the cipher text C, the symmetric decryption method M can recover the original message M.

Convergent encryption

Deduplication data privacy may be ensured by using convergent encryption. The user or data owner takes the original data copies and uses the derived convergent key to encrypt the copies. The user also creates a tag for the duplicate data that may be used to spot further instances of the same file. It is assumed here that if two data copies are identical, then their tags are identical as well (this is known as the "tag correctness property"). Users can check for duplication by sending a tag to a central server, which will then determine whether or not the same item has previously been saved. Note that the tag cannot be used to determine the convergent key and so cannot be used to undermine data secrecy. There will be a server-side copy of the encrypted data and an accompanying tag. It is possible to formally describe a convergent encryption system using only four elementary operations.

Algorithm

- Step 1: $KeyGen_{CE}$(M) -- A convergent key K is generated by the key generation technique that transfers a data copy M to K.
- Step 2: Enc_{CE}(K,M) -- The convergent key K and the data copy M are the inputs of a symmetric encryption method, and the resulting cipher text C is the output.
- Step 3:Dec_{CE} (K,C) -- M is the decryption technique that, given the cipher text C and the convergent key K, returns the unaltered copyM; and
- Step 4:TagGen(M) -- The algorithm for creating tags, T(M), maps the source data copy M and returns a tag, T(M).

IV. Results And Discussion

We determined how much more data could be sent once deduplication took place in storage. We utilized TCP dump to achieve this, filtering out everything that wasn't connected to Drop Box and deduplication. As you can see, we're not the only ones that think so. It costs roughly 7 KB to upload a file via the Drop Box API, even if it's extremely little. Deduplication's bandwidth usage as a percentage of basic Dropbox usage grows as file sizes grow. Overhead significantly decreases as file sizes increase due to deduplication's modest constant size of the additional file delivered.

Using distinct information, we quantify how well deduplication often works. At the outset, the server is empty. The client processes transfer 128GB of data to the server, all of which is distinct from any other data in the world. The obtained Linux command is then used by a client process to get the information. We also take a look at the raw disk performance by using our testbed's native file system

to read and write data. Deduplication reduces write throughput by 13–19% compared to raw write throughput. Since there are fewer segments to process, unique write performance improves with increasing segment size in deduplication. Deduplication's read throughput, on the other hand, is virtually identical to raw read throughput.

Confidentiality of data is crucial in situations when it must be protected from prying eyes. To achieve this goal, the protected authorization system can be made available to each user upon registration for data ownership.

V. Conclusion

In order to run a trustworthy cloud storage service, especially one that handles data processing, it is crucial to manage encrypted data with deduplication. When it comes to making the most of your cloud storage and data transfer speeds, Source Based De-duplication is your best bet. Distributed de-duplication is an effective method for improving data security, privacy, and integrity. The de-duplication ratio and data integrity can both be improved by using a hybrid approach that combines the two techniques. Moreover, by tweaking the de-duplication method, a decent de-duplication ratio may be accomplished. We have conducted extensive testing of the program to show that the proposed way is the most effective strategy for eliminating redundant data in the cloud. We develop our suggested approved duplicate check system and run experiments on a testbed to see how well it performs.

References: -

[1] Burramukku, Tirapathi&Padmasree, Ms&KrishnaKanth, Mr&KumarReddy, Mr. (2018). Secure Data Deduplication by Using Convergent Key Technique. International Journal of Engineering & Technology. 7. 459. 10.14419/ijet.v7i2.32.15741.

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

[3] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C.Lee, and Wenjing Lou, A Hybrid cloud approach for secure authorized deduplication, in: IEEE Transactions on parallel and distributed systems, May 2015.

[4] 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.

[5] T. Y. Wu, J. S. Pan, and C. F. Lin, "Improving Accessing Efficiency of cloud storage using de-duplication and Feedback schemes," IEEE Syst.J., vol. 8, no. 1, pp. 208– 218, Mar. 2014.

[6] 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.

[7] 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.

[8] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message- Locked encryption and secure deduplication," in Proc. CryptoloEUROCRYPT 2013, pp. 296–312.

[9] 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.

[10] 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.

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

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