A Review on Deduplication-Cost Efficient Method to Store Data Over Cloud Using Convergent Encryption

Asst. Prof. Ashwin Selokar
Department of Information Technology
DMIETR, RTMNU
Wardha, India

Asst. Prof. Jayant Rohankar
Department of Information Technology
TGPCET RTMNU
Nagpur, India

Abstract—This paper represents that, many techniques are using for the elimination of duplicate copies of repeating data, out of those techniques, the most important data compression technique is data deduplication. Convergent technique has been used to encrypt data before outsourcing for privacy and security point of view. In the proposed system, we apply the technique of cryptographic tuning to make the encryption more secure and flexible. In previous systems, there was a limitation of convergent encryption. Data deduplication does not allow the storage of repetitive blocks. It also puts the pointer to the existing blocks so that the data owner have the freedom of selecting users, to have access to the published file. Access control is provided into the application. The integrity of data outsourced to the cloud is managed by the hash calculation of any content following the proof-of-ownership module. Proposed system calculates the hash value of the data content on both sides i.e.; destination as well as source side. Request hash for the cloud side to predict the tampering of data. The expected analysis shows the improvement in execution time and development cost.

Keywords-*Deduplication, Convergent encryption, Cryptography, Hash value.*

*****

I. Introduction

Cloud computing enables new business models and cost effective resource usage. Instead of maintaining their own data center, companies can concentrate on their core business and purchase resources when it will be needed. Especially when combining publicly accessible clouds with a privately maintained virtual infrastructure in a hybrid cloud, the hybrid cloud technology can open up new opportunities for businesses. Today's cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified privileges, which define the access rights of the stored data. One critical challenge of cloud storage services is the management of the ever-increasing volume of data. To make data management scalable in cloud computing, deduplication has been a well-known technique and has attracted more and more attention recently.

Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Deduplication can take place at either the file level or the block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files [12].Although data deduplication brings a lot of benefits, security and privacy concerns arise as users' sensitive data are susceptible to both inside and outside attacks. Traditional encryption, while providing data confidentiality, is incompatible with data deduplication. Specifically, traditional encryption requires different users to encrypt their data with their own keys. Thus, identical data copies of different users will lead to different cipher text, making deduplication impossible.

Convergent encryption has been proposed to enforce data confidentiality while making deduplication feasible. It encrypts/decrypts a data copy with a convergent key, which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since the encryption operation is deterministic and is derived from the data content, identical data copies will generate the same convergent key and hence the same cipher text.

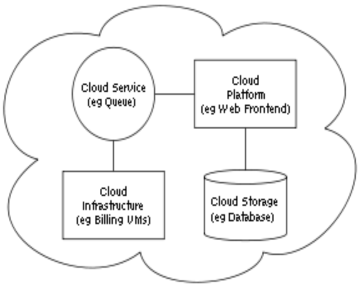
Cloud Service (eg Queue)
Cloud Platform (eg Web Frontend)
Cloud Infrastructure (eg Billing VMs)
Cloud Storage (eg Database)

Figure.1 Architecture of Cloud Computing

The critical challenge of cloud storage or cloud computing is the management of the continuously increasing volume of data. Data deduplication or Single Instancing essentially refers to the elimination of redundant data. However, indexing of all data is still retained should that data ever be required. In general the data deduplication eliminates the duplicate copies of repeating data [11].

Data deduplication is one of the hottest technologies in storage right now because it enables companies to save a lot of money on storage costs to store thedata and on the bandwidth costs to move the data when replicating it offsite. This is great news for cloud providers, because if you store less, you need less hardware. If you can deduplicate what you store, you can better utilize your existing storage space, which can save money by using what you have more efficiently. If you store less, you also back up less, which again means less hardware and backup media. If you store less, you also send less data over the network in case of a disaster, which means you save money in hardware and network costs over time. The business benefits of data deduplication include:

- Reduced hardware costs;
- Reduced backup costs;
- Reduced costs for business continuity / disaster recovery;
- Increased storage efficiency; and
- Increased network efficiency.

II. Literature Survey

A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S.Lui [1],has presented Cloud storage isan emerging service model that enables individuals and enterprises to outsource the storage of data backups to remote cloud providers at a low cost. Hence results shows that FadeVersion only adds minimal performance overhead over a traditional cloud backup service that does not support assured deletion.

C. Ng and P. Lee [2], had present RevDedup, a de-duplication system designed for VM disk image backup in virtualization environments. RevDedup has several design goals: high storage efficiency, low memory usage, high backup performance, and high restore performance for latest backups. They extensively evaluate our RevDedup prototype using different workloads and validate our design goals.

D. Ferraiolo and R. Kuhn [3],has described the Mandatory Access Controls (MAC) are appropriate for multilevel secure military applications, Discretionary Access Controls (DAC) are often perceived as meeting the security processing needs of industry and civilian government.

J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou [4], had proposed Dekey, an efficient and reliable convergent key management scheme for secure deduplication. They implement Dekey using the Ramp secret sharing scheme and demonstrate that it incurs small encoding/decoding overhead compared to the network transmission overhead in the regular upload/download operations.

J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer [5],has presented the Farsite distributed file system provides availability by replicating each file onto multiple desktop computers. Measurement of over 500 desktop file systems shows that nearly half of all consumed space is occupied by duplicate files. The mechanism includes 1) convergent encryption, which enables duplicate files to coalesce into the space of a single file, even if the files are encrypted with different users' keys, and 2) SALAD, a Self Arranging, Lossy, Associative Database for aggregating file content and location information in a decentralized, scalable, fault- tolerant manner.

J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl [6],has provided the private users outsource their data to cloud storage providers, recent data breach incidents make end-toend encryption an increasingly prominent requirement data deduplication can be effective for popular data, whilst semantically secure encryption protects unpopular content.

J. Xu, E.-C. Chang, and J. Zhou [7],has described the secure client-side deduplication scheme, with the following advantages: our scheme protects data confidentiality (and some partial information) against both outside adversaries and honest-but-curious cloud storage server, while Halevi et al. trusts cloud storage server in data confidentiality.

J. Yuan and S. Yu[8], has proposed Data integrity and storage efficiency are two important requirements for cloud storage. The author proposed scheme is also characterized by constant realtime communication andcomputational cost on the user side.

K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan [9] has proposed, the emergence of cost-effective cloud services offers organizations great opportunity to reduce their cost and increase productivity.The system, called Sedic, leverages the special features of Map Reduce to automatically partition a computing job according to the security levels of the data it works.

M. Bellare and A. Palacio[10] has provided the proof for GQ based on the assumed security of RSA under one more inversion, an extension of the usual onewayness assumption that was introduced. Both results extend to establish security against impersonation under concurrent attack.

Neal Leavitt [13] has described the hybrid cloud is the architecture that provides the organization to efficiently work on both the private and public cloud architecture in combination by providing the scalability to adopt. Some of the basic concepts and idea proposed by authors and how best and easy to adopt this environment are explained.

III. METHODOLOGY

A. Motivation

Cloud storage services are becoming very popular now a day. Cloud provides a better way of storage with efficient cost. One major problem with cloud is to manage huge amount of data. In order to manage data de-duplication technique is used. Although, de-duplication has many advantages but it has some security issues. This motivates us to propose a model which manage the security issues of de-duplication and provide authorized de-duplication in cloud.

B. Existing System

To make data management scalable in cloud computing, deduplication has been a well-known technique and has attracted more and more attraction recently. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Deduplication can take place at either the file level or the block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files.

Drawbacks of existing system are User's sensitive data are susceptible to both insider and outsider attacks. Security is not provided in existing system. To manage increasing volume of data it requires extra bandwidth. Sometimes deduplication is impossible.

C. System Architecture

If user wants to upload the files on the public cloud then user first encrypt that file with the key and then send it to the public cloud at the same time user also generates the key for that file and sends that key to the private cloud for the purpose of security. In the public cloud we use one algorithm for deduplication. Which is used to avoid the duplicate copies of files which is entered in the public cloud. Hence it also minimizes the bandwidth. That means we require the less storage space for storing the files on the public cloud. In the public cloud any person that means the unauthorized person can also access or store the data so we can conclude that in the public cloud the security is not provided.

In general for providing more security user can use the private cloud instead of using the public cloud. User generates the key at the time of uploading file and stores it to the private cloud. When user wants to downloads the file that he/she upload, he/she sends the request to the public cloud. Public cloud provides the list of files that are uploads the many user of the public cloud because there is no security is provided in the public cloud. When user selects one of the file from the list of files then private cloud sends a message like enter the key!. User has to enter the key that he generated for that file. When user enter the key the private cloud checks the key for that file and if the key is correct that means user is valid then private cloud give access to that user to download that file successfully. then user downloads the file from the public cloud and decrypt that file by using the same convergent key which is used at the time of encrypt that file in this way user can make a use of the architecture.

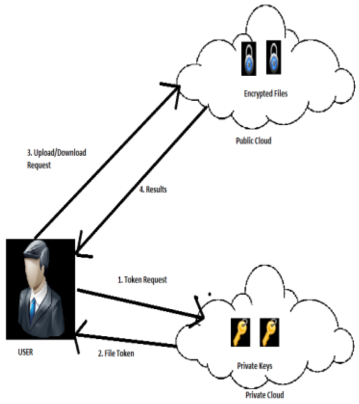

Figure.2 System Architecture

D. Convergent Encryption

Convergent encryption provides data confidentiality in deduplication. A user or data owner derives a convergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user also derives a tag for the data copy, such that the tag will be used to detect duplicates. Here, we assume that the tag correctness properties, if two data copies are same, then their tags are same. To detect duplicates, the user first sends the tag to the server side to check if the identical copy has been already stored. Both the encrypted data copy and its corresponding tag will be stored on the server side. Convergent encryption scheme can be defined with the following primitive functions:

- $KeyGen_{CE}(M)$ - K is the key generation algorithm that maps a data copy M to a convergent key K.
- $Enc_{CE}(K,M)$ - C is the symmetric encryption algorithm that takes both the convergent key K and the data copy M as inputs& then outputs a cipher text C.
- $Dec_{CE}(K,C)$ - M is the decryption algorithm that takes both the cipher text C & the convergent key K as inputs and then outputs the original data copy M.

In our next paper we will explain the working of the system using algorithm called as Advanced Encryption Standard used for data deduplication.

E. Applications

Hybrid clouds are mainly built to suit any of the IT environment or architecture, whether it might be any enterprise wide IT network or any department. Public data which is stored can be analyzed from statistical analyses which are done by social media, government entities can be used to enhance and analyze their own corporate data.

IV. Future Scope

It excludes the security problems that may arise in the practical deployment of the present model. Also, it increases the national security. It saves the memory by deduplicating the data and thus provides us with sufficient memory. It provides authorization to the private firms and protects the confidentiality of the important data.

V. Conclusion

The notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. We also presented new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. We use convergent encryption with modification version to deal with brute-force attack using domain separation and cryptographic tuning to make better authorized deduplication technique.

References

[1] 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 Proc. 3rd Int. Workshop Secutiry Cloud Comput., 2011, pp. 160–167.

[2] C. Ng and P. Lee, “Revdedup: A reverse deduplication storage system optimized for reads to latest backups,” in Proc. 4th Asia- Pacific Workshop Syst., Apr. 2013.

[3] J. Li, X. Chen, et al., Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.

[4] 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.

[5] J. Xu, E.-C. Chang, and J. Zhou, “Weak leakage-resilient client- side deduplication of encrypted data in cloud storage,” in Proc. 8th ACM SIGSAC Symp. Inform., Comput. Commun. Security, 2013, pp. 195–206.

[6] J. Yuan and S. Yu, “Secure and constant cost public cloud storage auditing with deduplication,” IACR Cryptology ePrint Archive, 2013:149, 2013

[7] K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan, “Sedic: Privacy-aware data intensive computing on hybrid clouds,” in Proc. 18th ACM Conf. Comput. Commun. Security, 2011, pp. 515–526.

[8] M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Server- aided encryption for deduplicated storage,” in Proc. 22nd USENIX Conf. Sec. Symp., 2013, pp. 179–194.

[9] M. Bellare, S. Keelveedhi, and T. Ristenpart, “Message-locked encryption and secure deduplication,” in Proc. 32nd Annu. Int. Conf. Theory Appl. Cryptographic Techn., 2013, pp. 296–312.

[10] 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.