ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 1 , January 2017

Survey on Secure De-duplication with Encrypted Data for Cloud Storage

Vishalakshi N S, S.Sridevi

P.G. Student, Department of Computer Science, New Horizon College Of Engineering, Bangalore, Karnataka, India
Assistant Professor, Department of Computer Science, New Horizon College Of Engineering, Bangalore, Karnataka ,India

ABSTRACT: Cloud Computing provides many resources to users as services such as highly available storage space. To manage the ever-increasing volume of data in cloud is a critical task. To make data management scalable in cloud computing, data deduplication is a technique. It is the technique of data compression for eliminating duplicate copies of repeating data in cloud storage to reduce the amount of storage space.It has been widely used in the cloud storage to reduce the amount of storage space and save bandwidth.

The advantage of deduplication unfortunately come with high cost in terms of new security and privacy challenges . Data deduplication is the new data compaction technology which removes duplicates in data. “Data Deduplication is the process of examining a data-set or I/O stream at the sub-file level and storing and /or sending only unique data”. It differs from the compression techniques by working on the data at sub-file level where as compression encodes the data in the file to reduce its storage requirement. We propose Clouded up, a secure and efficient storage service which assures block-level deduplication and data confidentiality at the same time. Although based on convergent encryption, Clouded up remains secure thanks to the definition of a component that implements an additional encryption operation and an access control mechanism.

KEYWORDS: Cloud computing, Deduplication, Convergent Encryption.

I. INTRODUCTION

Cloud computing provides a low-cost, scalable, location- independent infrastructure for data management and storage. Owing to the population of cloud service and the increasing of data volume, more and more people pay attention to economize the capacity of cloud storage than before .Therefore how to utilize the cloud storage capacity well becomes an important issue nowadays. Cloud service providers offer highly available storage space and massively parallel computing resources at relatively low costs. The advent of Cloud Storage motivates enterprises and organizations to outsource data storage to third party cloud providers. 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. Gmail is an example of cloud storage which is used by most of us regularly.

One of the major issues of cloud storage services is the management of the ever increasing volume of data. To make data management scalable in cloud computing, deduplication is a technique and has attracted more and more attention recently. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeated data in storage. Data deduplication is also known as single instancing or intelligent compression technique. This technique is used to improve storage utilization. Instead of keeping multiple data copies with the same content on cloud, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy.
Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. strategies as follow, differentiated by the type of basic data units.

A) File-level deduplication: A file is a data unit when examining the data of duplication, and it typically uses the hash value of the file as its identifier. If two or more files have the same hash value, they are assumed to have the same contents and only one of these files will be stored.

B) **Block-level deduplication**: This strategy segments a file into several fixed-sized blocks or variable-sized blocks, and computes hash value for each block for examining the duplication blocks.

A technique which has been proposed to meet these two conflicting requirements is convergent encryption whereby the encryption key is usually the result of the hash of the data segment. Although convergent encryption seems to be a good candidate to achieve confidentiality and deduplication at the same time, it unfortunately suffers from various well-known weaknesses including dictionary attacks: an attacker who is able to guess or predict a file can easily derive the potential encryption key and verify whether the file is already stored at the cloud storage provider or not. Deduplication lowers storage costs as fewer disks are needed. It improves disaster recovery since there's far less data to transfer. Backup/archive data usually includes a lot of duplicate data.

The similar data is stored over and over again, consuming unwanted storage space on disk or tape, electricity to power and cool the disk/tape drives and bandwidth for replication. This will create a chain of cost and resource inefficiencies within the organization. While providing data confidentiality, traditional encryption is incompatible with data deduplication. Specifically, it requires different users to encrypt their data with their own keys. Thus, indistinguishable data copies of different users will lead to different cipher texts, making deduplication unfeasible.

Convergent encryption has been proposed to impose data confidentiality while making deduplication feasible. It encrypts and decrypts a data copy with a convergent key, which is obtain by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users preserve the keys and send the cipher text to the cloud. Because the encryption operation is deterministic and is derived from the data content, indistinguishable data copies will generate the same convergent key and hence the same cipher text. To avoid unauthorized access, a secure PoW (proof of ownership protocol) is also needed to provide the confirmation that the user indeed owns the same file when a duplicate is found. After the confirmation, consequent users with the same file will be provided a pointer from the server without needing to upload the similar file. A user can download the encrypted file with the pointer from the server, which can only be decrypted by the equivalent data owners with their convergent keys. Thus, convergent encryption will allow the cloud to do deduplication on the cipher texts and the proof of ownership (PoW) prevents the unauthorized user to access the file.

II. LITERATURE SURVEY

Data deduplication technology is method for maximizing the usage of available data storage. Deduplication helps to identify similarities among different files to save disk space. Data deduplication inspects data down to block-level and bit level and, after the initial occurrence, only the changed data they find are saved. The rest are discarded and replaced with a pointer to the previously saved information. Block-level and bit-level deduplication methods are able to achieve compression ratios of 20x to 60x, or sometime even higher, under the right conditions.
There is file-level deduplication, called single instance storage in file-level deduplication, in this if two files are identical, one copy of the file is kept while subsequent iterations are not. File-level deduplication is not as efficient as block-level and bit-level storage because even a single changed bit results in a new copy of the whole file being stored. For the purposes data deduplication is defined as operating at block level and bit level.

Data deduplication reduces the amount of data that is needed to be stored. This means that less media has to be bought and it takes long to fill up disks and tapes. Data can be backed up more quickly, which means shorter backup windows and quicker restores time. A reduction in the amount of space taken up in disk systems and VTLs [13] for example, longer retention periods are possible, bringing quicker restores to users direct from disk and reducing dependence on tape storage and its management. Less data also means less bandwidth taken up, which means data deduplication can speed up remote backup, replication and also disaster recovery processes.

Bellare et.al 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. Cipher text 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. put forward a method to achieve dual level source based deduplication of large encrypted files with block key management and Proof of Ownership. Author claims that MLE scheme were proposed for target based file level deduplication and extending it to dual level deduplication requires much meta data 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 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 up loader and further distributed consequent verified unloader 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 clientsincrease in case of large scale deployment, and a decentralized deployment of key server is supposed as a solution.

In Bellare et. al claim that Message locked encryption 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 Pseudo random 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 propose Clouded up,a secure and efficient storage service which assures block level deduplication data confidentiality at the same time using convergent key encryption[2] added with block level key management. Architecture of Clouded up 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 meta data 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.

III. PROPOSEDARCHITECTURE

Data deduplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and deduplication is used in cloud storage to reduce the need for amount of storage space and it's also used to save bandwidth. To protect the confidentiality of sensitive data while supporting deduplication the techniques like convergent encryption technique has been proposed and used to encrypt the data before outsourcing. This paper makes the attempt to formally address the problem of authorized data deduplication. Here we aim at efficiently solving the problem of handling huge volume of data by chunk level deduplication with differential privileges in cloud computing, we consider a cloud architecture.

For the data duplication check in the proposed system we are doing duplication check in authenticated way. For the file duplication check proof of ownership is also set at the time of file upload the proof is added with the file this proof will decide the access privilege to the file. It will define who can perform duplication check of the file. For the send duplicate check request user need to submit his file and proof of ownership of the file. The duplicate check request get only approved when there is file on the cloud and also privileges of the user are there.

A. System Architecture:

The proposed system architecture is shown in the figure 1. shows the proposed system architecture which comprises of public cloud, private cloud and user. In the proposed system architecture shown in Figure 1. There are one public cloud and one is private cloud. Public cloud contains all data of the user such as files and private cloud consist of user credentials. For each transaction with the public cloud user need to take token for the private cloud. If the user credentials stored at the public cloud and private cloud are get matched then user can have assess for the duplicate check. Following operations are need to be done in the authenticate duplicate check.

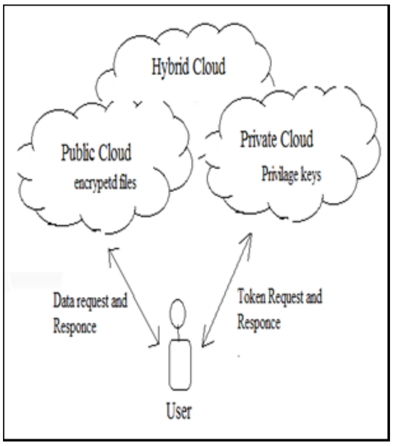

Fig 1. System architecture

B. Encryption of File:
Aim To encrypt the user data we are using secrete key resides at the private cloud. This key is used to convert plain text to cipher text and again for the decryption of the user data. To encrypt and decrypt we have used three basic functions as follow:
KeyGenSE: In this k is the key generation algorithm which can generate the secrete file by using security parameter.
EncSE (k, M): in this formulae M is the text message and key is the secrete key by using this both we have generated a cipher text C.
DecSE (k, C): Here C is the cipher text and k is the encryption key by using cipher text and secrete key we have to generate plain text.

C. Confidential Encryption of data:
This ensures a data confidentiality in the duplication. User derives a convergent key from each original data and encrypt the data copy with the generated convergent key. User also add the tag for the data so that the tag will helps to detect the duplicate data. By using convergent key generation algorithm key is get generated this key is used to encrypt the user data. This will ensures the security, ownership and authority of the data.

IV. CLOUDED UP

The scheme proposed in this paper aims at deduplication at the level of blocks of encrypted files while coping with the inherent security exposures of convergent encryption. The scheme consists of two basic components: a server that is in charge of access control and that achieves the main protection against COF and LRI attacks; another component, named as meta data manager (MM), is in charge of the actual deduplication and key management operations.

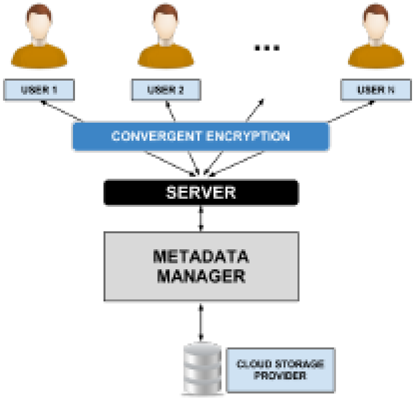

Fig. 2. High-level view of ClouDedup

A. The Server

A simple solution to prevent the attacks against convergent encryption (CE) consists of encrypting the cipher texts resulting from CE with another encryption algorithm using the same keying material for all input. This solution is compatible with the deduplication requirement since identical cipher texts resulting from CE would yield identical outputs even after the additional encryption operation. Yet, this solution will not suffer anymore from the attacks targeting CE such as COF and LRI.

We combine the access control function with the mechanism that achieves the protection against CE through an additional encryption operation. Indeed, access control is an inherent function of any storage system with reasonable security assurance. Enhancing the trusted component of the storage system, that implements access control, with the new mechanism against COF and LRI attacks, seems to be the most straightforward approach. The core component of Clouded up is thus a server that implements the additional encryption operation to cope with the weaknesses of CE, together with a user authentication and an access control mechanism embedded in the data protection mechanism. Each data segment is thus encrypted by the server in addition to the convergent encryption operation performed by the user. As to the data access control, each encrypted data segment is linked with a signature generated by its owner and verified upon data retrieval requests. The server relies on the signature of each segment to properly identify the recipient.

B. Block-level Deduplication and Key Management

Even though the mechanisms of the server cope with the security weaknesses of CE, the requirement for deduplication at block-level further raises an issue with respect to key management. As an inherent feature of CE, the fact that encryption keys are derived from the data itself does not eliminate the need for the user to memorize the value of the key for each encrypted data segment. Unlike file-level deduplication, in case of block-level deduplication, the requirement to memorize and retrieve CE keys for each block in a secure way, calls for a fully-fledged key management solution. We thus suggest to include a new component, the meta data manager (MM), in the new ClouDedup system in order to implement the key management for each block together with the actual deduplication operation.

C. Threat Model

The goal of the system is to guarantee data confidentiality without losing the advantage of deduplication. Confidentiality must be guaranteed for all files, including the predictable ones. The security of the whole system should not rely on the security of a single component (single point of failure), and the security level should not collapse when a single component is compromised. We consider the server as a trusted component with respect to user authentication, access control and additional encryption. The server is not trusted with respect to the confidentiality of data stored at the cloud storage provider. Therefore, the server is not able to perform off line dictionary attacks. Anyone who has access to the storage is considered as a potential attacker, including employees at the cloud storage provider and the cloud storage provider itself. In our threat model, the cloud storage provider is honest but curious, meaning that it carries out its tasks but might attempt to decrypt data stored by users.

We do not take into account cloud storage providers that can choose to delete or modify files. Our scheme might be extended with additional features such as data integrity and proofs of retrievability . Among the potential threats, we identify also external attackers. An external attacker does not have access to the storage and operates outside the system. This type of attacker attempts to compromise the system by intercepting user's account. External attackers have a limited access to the system and can be effectively neutralized by putting in place strong authentication mechanisms and secure communication channels.

D. Security

In the proposed scheme, only one component, that is the server, is trusted with respect to a limited set of operations, therefore we call it semi-trusted. Once the server has applied the additional encryption, data are no longer vulnerable to CE weaknesses. Indeed, without possessing the keying material used for the additional encryption, no component can perform dictionary attacks on data stored at the cloud storage provider. The server is a simple semi-trusted component that is deployed on the user's premises and is in charge of performing user authentication, access control and additional symmetric encryption. The primary role of the server is to securely retain the secret key used for the additional encryption. In a real scenario, this goal can be effectively accomplished by using a hardware security module (HSM) . When data are retrieved by a user, the server plays another important role. Before sending data to a given recipient, the server must verify if block signatures correspond to the public key of that recipient. The meta data manager (MM) and the cloud storage provider are not trusted with respect to data confidentiality, indeed, they are not able to decrypt data stored at the cloud storage provider. We do not take into account components that can spontaneously misbehave and do not accomplish the tasks they have been assigned.

V. CONCLUSION

The idea of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. Several new deduplication constructions supporting authorized duplicate check in cloud .Cloud computing has reached a maturity that leads it into a productive phase. This means that most of the main issues with cloud computing have been addressed to a degree that clouds have become interesting for full commercial Exploitation. We designed a system which achieves confidentiality and enables block-level deduplication at the same time. Our system is built on top of convergent encryption. We showed that it is worth performing block-level deduplication instead of file level deduplication since the gains in terms of storage space are not affected by the overhead of meta data management, which is minimal. To protect confidentiality of data in case of deduplication environment is important. Security is provided using convergent encryption technique to encrypt the data.

REFERENCES

[1] Atul Adya, William J Bolosky, Miguel Castro, Gerald Cermak, Ronnie Chaiken, John R Douceur, Jon Howell, Jacob R Lorch, Marvinv Theimer, and Roger P Wattenhofer. Farsite: Federated, available, and reliable storage for an incompletely trusted environment. ACM SIGOPS Operating Systems Review, 36(SI):1–14, 2002.

[2] Mihir Bellare, Alexandra Boldyreva, and Adam ONeill. Deterministic and efficiently searchable encryption. In Advances in Cryptology-CRYPTO 2007. Springer, 2007.

[3] Mihir Bellare, Sriram Keelveedhi, and Thomas Ristenpart. Dupless: Server-aided encryption for deduplicated storage. 2013.

[4] Mihir Bellare, Sriram Keelveedhi, and Thomas Ristenpart. Message-locked encryption and secure deduplication. In Advances in Cryptology–EUROCRYPT 2013. Springer, 2013.

[5] Kevin D. Bowers, Ari Juels, and Alina Oprea. Hail: a high-availability and integrity layer for cloud storage. In Proceedings of the 16th ACM conference on Computer and communications security, CCS '09, New York, NY, USA, 2009. ACM.

[6]M. Bellare and A. Palacio. Gq and schnorr identification schemes: Proofs of security against impersonation under active and concur- rent attacks. In CRYPTO, 2002.

[7] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloudcomputing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.

[8]Pooja S Dodamani, Pradeep Nazareth, "A Survey on Hybrid Cloud with De-Duplication", International Journal of Innovative Research in Computer and Communication Engineering, December 2014.

[9] Boga Venkatesh, Anamika Sharma, Gaurav Desai, Dadaram Jadhav, "Secure Authorised Deduplication by Using Hybrid Cloud Approach", November 2014.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. ACM, 2012.

[10] R. D. Pietro and A. Sorniotti. Boosting efficiency and security in proof of ownership for deduplication. In H. Y. Youm and Y. Won, editors, ACM Symposium on Information, Computer and Communications Security. ACM, 2012.

[11]Pasquale Puzio Refik Molva Melek Sergio Onen L. ClouDedup: Secure Deduplication with Encrypted Data for cloud storage. In Proceeding in 2013 IEEE International Conference on Cloud Computing Technology and Science

[12] I. Clarke, O. Sandberg, B.Wiley, and Hong T.W. Freenet: A distributed anonymous information storage and retrieval system.