ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Secure Data Deduplication with Efficient Key Management in Cloud Databases

Pasupuleti Pavani
M.Tech Student,
Department of Computer Science and Engineering,
Amalapuram Institute of Management Sciences and College of Engineering.

Mohammed Alisha
Associate Professor & HOD,
Department of Computer Science and Engineering,
Amalapuram Institute of Management Sciences and College of Engineering.

ABSTRACT:

In Cloud computing involves deploying groups of remote servers and software networks that allow centralized data storage and online access to computer services or resources. Clouds can be classified as public, private or hybrid. Cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. Data deduplication is an important technique for eliminating redundant data. Instead of taking no. of same files, it store only single copy of file. In most organizations, storage system contain many pieces of duplicate data. . For example, the same file may be saved in several different places by different users. Deduplication eliminates these extra copies by saving just one copy of the data and replacing the other copies with pointers that lead back to the original copy. It is data compression technique for improve the bandwidth efficiency and storage utilization. Data deduplication most widely used in cloud computing. It make data management scalable and storage problem in cloud computing. We show that our proposed authorized duplicate check scheme has minimal overhead compared to normal operations. As a proof of concept, we implement a prototype of our proposed authorized duplicate check scheme and conduct tested experiments using our prototype. This paper tries to minimize the data duplication that occurs in hybrid cloud storage by using various techniques.

Index Terms:

Deduplication, Authorized Duplicate Check, Confidentiality, Hybrid Cloud, C.

I.INTRODCUTON :

In computing, data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single-instance (data) storage.

This 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. In the deduplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. Given that the same byte pattern may occur dozens, hundreds, or even thousands of times (the match frequency is dependent on the chunk size), the amount of data that must be stored or transferred can be greatly reduced. A Hybrid Cloud is a combined form of private clouds and public clouds in which some critical data resides in the enterprise's private cloud while other data is stored in and accessible from a public cloud. Hybrid clouds seek to deliver the advantages of scalability, reliability, rapid deployment and potential cost savings of public clouds with the security and increased control and management of private clouds. As cloud computing becomes famous, an increasing amount of data is being stored in the cloud and used by users with specified privileges, which define the access rights of the stored data.

The data is encrypted before outsourcing it on the cloud or network. This encryption requires more time and space requirements to encode data. In case of large data storage the encryption becomes even more complex and critical. By using the data deduplication inside a hybrid cloud, the encryption will become simpler. As we all know that the network is consist of abundant amount of data, which is being shared by users and nodes in the network. Many large scale network uses the data cloud to store and share their data on the network. The node or user, which is present in the network have full rights to upload or download data over the network. But many times different user uploads the same data on the network. This will create duplication inside the cloud. If the user wants to retrieve the data or download the data from cloud, every time he has to use the two encrypted files of same data.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The cloud will do same operation on the two copies of data files. Due to this the data confidentiality and the security of the cloud get violated. It creates the burden on the operation of cloud.

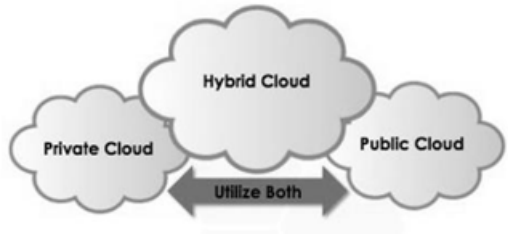

Figure: 1. Architecture of Hybrid cloud.

To avoid this duplication of data and to maintain the confidentiality in the cloud we using the concept of Hybrid cloud. It is a combination of public and private cloud. Hybrid cloud storage combines the advantages of scalability, reliability, rapid deployment and potential cost savings of public cloud storage with the security and full control of private cloud storage.

II.RELATED WORK

Single-Instance Storage:

Single-instance storage environments are able to detect and remove 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 determine if the incoming file is identical to an existing file in the storage system. 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 tremendous amount of redundancy within the files or between files.

Sub-file De-Duplication:

Sub-file de-duplication detects redundant data within and across files as opposed to finding identical files as in SIS implementations. Using sub-file de-duplication, redundant copies of data are detected and are eliminated-even after the duplicated data exist, within separate files. Sub-file data de-duplication has tremendous benefits even where files are not identical, but have data elements that are already recognized somewhere in the organization. Fixed length sub-file de-duplication uses an arbitrary fixed length of data to search for the duplicate data within the files. So most of the organizations widely use data depulication technology, which is also called as, single instance storage, intelligent compression, and capacity optimized storage and data reduction.

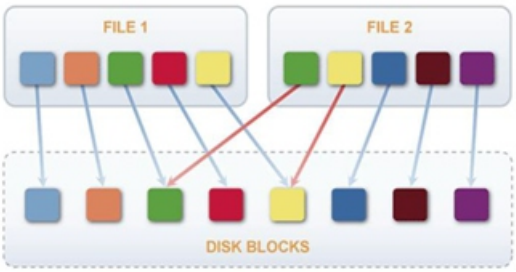

Fig 2: Example of De-duplication

Deduplication methods:

One of the most common forms of data deduplication implementations works by comparing chunks of data to detect duplicates. For that to happen, each chunk of data is assigned an identification, calculated by the software, typically using cryptographic hash functions. In many implementations, the assumption is made that if the identification is identical, the data is identical, even though this cannot be true in all cases due to the pigeonhole principle; other implementations do not assume that two blocks of data with the same identifier are identical, but actually verify that data with the same identification is identical. [6] If the software either assumes that a give identification already exists in the deduplication namespace or actually verifies the identity of the two blocks of data, depending on the implementation, then it will replace that duplicate chunk with a link. Once the data has been deduplicated, upon read back of the file, wherever a link is found, the system simply replaces that link with the referenced data chunk. The deduplication process is intended to be transparent to end users and applications.

» Chunking. Between commercial deduplication implementations, technology varies primarily in chunking method and in architecture. In some systems, chunks are defined by physical layer constraints (e.g. 4KB block size in WAFL). In some systems only complete files are compared, which is called Single Instance Storage or SIS. The most intelligent (but CPU intensive) method to chunking is generally considered to be sliding-block. In sliding block, a window is passed along the file stream to seek out more naturally occurring internal file boundaries.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

» Client backup deduplication. This is the process where the deduplication hash calculations are initially created on the source (client) machines. Files that have identical hashes to files already in the target device are not sent, the target device just creates appropriate internal links to reference the duplicated data. The benefit of this is that it avoids data being unnecessarily sent across the network thereby reducing traffic load.
» Primary storage and secondary storage. By definition, primary storage systems are designed for optimal performance, rather than lowest possible cost. The design criteria for these systems is to increase performance, at the expense of other considerations. Moreover, primary storage systems are much less tolerant of any operation that can negatively impact performance. Also by definition, secondary storage systems contain primarily duplicate, or secondary copies of data. These copies of data are typically not used for actual production operations and as a result are more tolerant of some performance degradation, in exchange for increased efficiency.

To date, data deduplication has predominantly been used with secondary storage systems. The reasons for this are two-fold. First, data deduplication requires overhead to discover and remove the duplicate data. In primary storage systems, this overhead may impact performance. The second reason whydeduplication is applied to secondary data, is that secondary data tends to have more duplicate data. Backup application in particular commonly generate significant portions of duplicate data over time. Data deduplication has been deployed successfully with primary storage in some cases where the system design does not require significant overhead, or impact performance.

POST-PROCESS DEDUPLICATION :

With post-process deduplication, new data is first stored on the storage device and then a process at a later time will analyse the data looking for duplication. The benefit is that there is no need to wait for the hash calculations and lookup to be completed before storing the data thereby ensuring that store performance is not degraded. Implementations offering policy-based operation can give users the ability to defer optimization on “active” files, or to process files based on type and location. One potential drawback is that you may unnecessarily store duplicate data for a short time which is an issue if the storage system is near full capacity.

IN-LINE DEDUPLICATION:

This is the process where the deduplication hash calculations are created on the target device as the data enters the device in real time. If the device spots a block that it already stored on the system it does not store the new block, just references to the existing block. The benefit of in-line deduplication over post-process deduplication is that it requires less storage as data is not duplicated. On the negative side, it is frequently argued that because hash calculations and lookups takes so long, it can mean that the data ingestion can be slower thereby reducing the backup throughput of the device. However, certain vendors with in-line deduplication have demonstrated equipment with similar performance to their post-process deduplication counterparts. Post-process and in-line deduplication methods are often heavily debated.

SOURCE VERSUS TARGET DEDUPLICATION:

Another way to think about data deduplication is by where it occurs. When the deduplication occurs close to where data is created, it is often referred to as “source deduplication.” When it occurs near where the data is stored, it is commonly called “target deduplication.” Source deduplication ensures that data on the data source is deduplicated. This generally takes place directly within a file system. The file system will periodically scan new files creating hashes and compare them to hashes of existing files.

Benefits:

Storage-based data deduplication reduces the amount of storage needed for a given set of files. It is mosteffective in applications where many copies of very similar or even identical data are stored on a single disk— a surprisingly common scenario. In the case of data backups, which routinely are performed to protect against data loss, most data in a given backup remain unchanged from the previous backup. Common backup systems try to exploit this by omitting (or hard linking) files that haven’t changed or storing differences between files. Neither approach captures all redundancies, however. Hard-linking does not help with large files that have only changed in small ways, such as an email database; differences only find redundancies in adjacent versions of a single file (consider a section that was deleted and later added in again, or a logo image included in many documents).

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

» Network data deduplication is used to reduce the number of bytes that must be transferred between endpoints, which can reduce the amount of bandwidth required. See WAN optimization for more information.

Virtual servers benefit from deduplication because it allows nominally separate system files for each virtual server to be coalesced into a single storage space. At the same time, if a given server customizes a file, deduplication will not change the files on the other servers—something that alternatives like hard links or shared disks do not offer. Backing up or making duplicate copies of virtual environments is similarly improved simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material.

III.HYBRID CLOUD FOR SECURE DEDUPLICATION:

At a high level, our setting of interest is an enterprise network, consisting of a group of affiliated clients (for example, employees of a company) who will use the S-CSP and store data with deduplication technique. In this setting, deduplication can be frequently used in these settings for data backup and disaster recovery applications while greatly reducing storage space. Such systems are widespread and are often more suitable to user file backup and synchronization applications than richer storage abstractions. There are three entities defined in our system, that is, users, private cloud and S-CSP in public cloud . The S-CSP performs deduplication by checking if the contents of two files are the same and stores only one of them. The access right to a file is defined based on a set of privileges.

The exact definition of a privilege varies across applications. For example, we may define a rolebased privilege according to job positions (e.g., Director, Project Lead, and Engineer), or we may define a time-based privilege that specifies a valid time period (e.g., 2014-01-01 to 2014-01-31) within which a file can be accessed. A user, say Alice, may be assigned two privileges “Director” and “access right valid on 2014- 01-01”, so that she can access any file whose access role is “Director” and accessible time period covers 2014-01- 01. Each privilege is represented in the form of a short message called token. Each file is associated with some file tokens, which denote the tag with specified.

A user computes and sends duplicate-check tokens to the public cloud for authorized duplicate check. Users have access to the private cloud server, a semitrusted third party which will aid in performing deduplicable encryption by generating file tokens for the requesting users. We will explain further the role of the private cloud server below. Users are also provisioned with per-user encryption keys and credentials

In this paper, we will only consider the file level deduplication for simplicity. In another word, we refer a data copy to be a whole file and file-level deduplication which eliminates the storage of any redundant files. Actually, block-level deduplication can be easily deduced from file-level deduplication, Specifically, to upload a file, a user first performs the file-level duplicate check. If the file is a duplicate, then all its blocks must be duplicates as well; otherwise, the user further performs the block-level duplicate check and identifies the unique blocks to be uploaded. Each data copy (i.e., a file or a block) is associated with a token for the duplicate check.

A.Architecture For Authorized Deduplication:

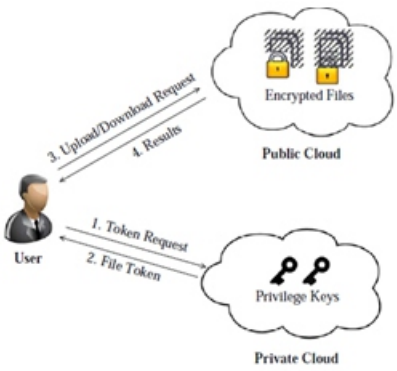

Fig3. Architecture for Authorized deduplication

• S-CSP. This is an entity that provides a data storage service in public cloud. The S-CSP provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data. In this paper, we assume that S-CSP is always online and has abundant storage capacity and computation power.

1. Start
2. Get unencrypted file tag
3. Accept the privilege based token from user

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

4. Validate the token and assert privilege level
5. Run deduplication check only on the privileged files
6. If the same tag is found along the privileged files, mark deduplication search as successful and grant access to the encrypted file
7. Stop

• Data Users. A user is an entity that wants to outsource data storage to the S-CSP and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users. In the authorized deduplication system, each user is issued a set of privileges in the setup of the system. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.

• Private Cloud. Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for facilitating user's secure usage of cloud service. Specifically, since the computing resources at data user/owner side are restricted and the public cloud is not fully trusted in practice, private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user and the public cloud. The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users. The interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively.

Notice that this is a novel architecture for data deduplication in cloud computing, which consists of a twin clouds (i.e., the public cloud and the private cloud). Actually, this hybrid cloud setting has attracted more and more attention recently. For example, an enterprise might use a public cloud service, such as Amazon S3, for archived data, but continue to maintain in-house storage for operational customer data. Alternatively, the trusted private cloud could be a cluster of virtualized cryptographic co-processors, which are offered as a service by a third party and provide the necessary hardware based security features to implement a remote execution environment trusted by the users.

1. Start
2. Authenticate the user based on their credentials
3. Generate the unique privilege based token
4. Return the token to the user
5. Stop

The public cloud does not have access to the user credentials. If compromised token generation keys can be changed regardless of the user credentials. Public cloud does not have access to decrypted files. Deduplication only happens on privileged files.

IV.IMPLEMENTATION:

We implement system with data deduplication,in which we model three entities as separate programs. A Client program is used to model the data users to carry out the file upload/download process. A Private Server program is used to model the private cloud which manages the private keys and handles the file token computation. A Storage Server program is used to model the S-CSP which stores and deduplicates files. Followings are function calls used in system:

•FileTag(File) - It computes SHA-1 hash of the File as File Tag;
•DupCheckReq(Token) - It requests the Storage Server for Duplicate Check of the file.
•FileEncrypt(File) - It encrypts the File with Convergent Encryption using 256-bit AES algorithm in cipher block chaining (CBC) mode, where the convergent key is from SHA-256 Hashing of the file;
•FileUploadReq(FileID, File, Token) – It uploads the File Data to the Storage Server if the file is Unique and updates the File Token stored.
•FileStore(FileID, File, Token) - It stores the File on Disk and updates the Mapping.

V.CONCLUSION:

In this paper, the idea of authorized data deduplication was proposed to protect the data security by including differential authority of users in the duplicate check. In public cloud our data are securely store in encrypted format,and also in private cloud our key is store with respective file. There is no need to user remember the key.So without key anyone can not access our file or data from public cloud. We also presented several 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. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

As a proof of concept, we implemented a prototype of our proposed authorized duplicate check scheme and conduct tested experiments on our prototype. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

REFERENCES:

[1] Open SSL Project. http://www.openssl.org/.

[2] P. Anderson and L.Zhang. Fast and secure laptop backups with encrypted de-duplication.In Proc.of USENIX LISA, 2010.

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

[4] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312, 2013.

[5] M. Bellare, C. Namprempre, and G. Neven. Security proofs for identity-based identification and signature schemes. Cryptology,22(1):1–61, 2009.

[6] M. Bellare and A. Palacio. Gq and schnorr identification schemes:Proofs of security against impersonation under active and concurrentattacks. In CRYPTO, pages 162–177, 2002.

[7] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twinclouds: An architecture for secure cloud computing. In Workshopon Cryptography and Security in Clouds (WCSC 2011), 2011.

[8] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer.Reclaiming space from duplicate files in a serverless distributedfile system. In ICDCS, pages 617–624, 2002.

[9] D. Ferraiolo and R. Kuhn. Role-based access controls. In 15thNIST-NCSC National Computer Security Conf., 1992.

[10]GNULibmicrohttpd.http://www.gnu.org/software/libmicrohttpd/.

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

[12] S. Kamara and K. Lauter, "Cryptographic Cloud Storage," in Proc. Financial Cryptography: Workshop Real-Life Cryptograph. Protocols Standardization, 2010, pp. 136-149.

[13] R. Geambasu, T. Kohno, A. Levy, and H.M. Levy, "Vanish: Increasing Data Privacy with Self-Destructing Data," in Proc. USENIX Security Symp., Aug. 2009, pp. 316-299..

[14] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. Secure data deduplication. In Proc. of StorageSS, 2008.

[15] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. Reclaiming space from duplicate files in a serverless distributed file system. In ICDCS, pages 617–624, 2.

Author Details:

Mohammed Alisha
is currently working as Associate Professor & Heading the department of Computer Science and Engineering, Amalapuram Institute of Management Sciences &College Of Engineering. He is a postgraduate in Computer Science and Engineering and had 9 years of teaching and research experience. His research interests include Spatial Data Mining, Compuer Networks, Web Mining and Data Warehousing.