International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 30th December 2017. Link :
http://www.ijiemr.org/downloads.php?vol=Volume-6&issue=ISSUE-13

Title: A New Cloud Design For Secure Verified Deduplication.

Volume 06, Issue 13, Page No: 251 - 257.

Paper Authors

*** ANAMANAMUDI SIREESHA, K. RAJA RAO.**

* Dept of CSE, St.Mary's Women's Engineering College.

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

A NEW CLOUD DESIGN FOR SECURE VERIFIED DEDUPLICATION

***ANAMANAMUDI SIREESHA, **K. RAJA RAO**
*PG Scholar, Dept of CSE, St.Mary's Women's Engineering College,Budampadu, Guntur.
**Assistant Professor, Dept of CSE, St.Mary's Women's Engineering College, Budampadu, Guntur.

ABSTRACT:

Data deduplication is a technique for reducing the amount of storage space an organization needs to save its data. In most organizations, the storage systems contain duplicate copies of many pieces of data. For example, the same file may be saved in several different places by different users, or two or more files that aren't identical may still include much of the same data. 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. Companies frequently use deduplication in backup and disaster recovery applications, but it can be used to free up space in primary storage as well. To avoid this duplication of data and to maintain the confidentiality in the cloud we using the concept of Hybrid cloud. To protect the confidentiality of sensitive data while supporting deduplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this paper makes the first attempt to formally address the problem of authorized data deduplication.

Keywords: Deduplication, Authorized duplicate check, Confidentiality, Hybrid cloud

I. INTRODUCTION

In computing, data deduplication is a specialized 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

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

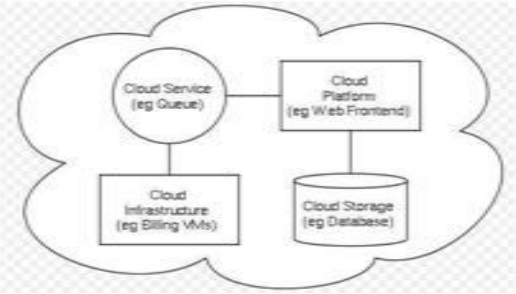

Figure: 1. Architecture of cloud computing.

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. In the deduplication process, duplicate data is deleted, leaving only one copy (single instance) of the data to be stored. 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.

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. Which will create a 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. 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.

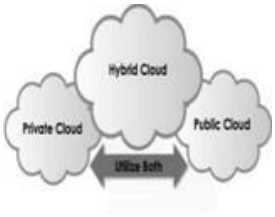

Figure: 2. Architecture of Hybrid cloud

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. LITERATURE SURVEY

In previous deduplication systems cannot support differential authorization duplicate check, which is important in many applications. In such an authorized deduplication system, each user is issued a set of privileges during system initialization. The overview of the cloud deduplication is as follow:

[2.1] 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 the 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.

[2.2] 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.

[2.3] 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.

When files with same hashes are found then the file copy is removed and the new file points to he old file. Unlike hard links however, duplicated files are considered to be separate entities and if one of the duplicated files is later modified, then using a system called Copy- on-write a copy of that file or changed block is created. The deduplication process is transparent to the users and backup applications. Backing up a deduplicated file system will often cause duplication to occur resulting in the backups being bigger than the source data. Target deduplication is the process of removing duplicates of data in the secondary store. Generally this will be a backup store such as a data repository or a virtual tape library.

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. If the software either assumes that a given identification already exists in the deduplication namespace or actually verifies the identity of the two blocks of data,

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

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.

III. PROPOSED SYSTEM

In the proposed system we are achieving the data deduplication by providing the proof of data by the data owner. This proof is used at the time of uploading of the file. Each file uploaded to the cloud is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files. Before submitting his duplicate check request for some file, the user needs to take this file and his own privileges as inputs. The user is able to find a duplicate for this file if and only if there is a copy of this file and a matched privilege stored in cloud.

[3.1] ENCRYPTION OF FILES

Here we are using the common secret key k to encrypt as well as decrypt data. This will use to convert the plain text to cipher text and again cipher text to plain text. Here we have used three basic functions, $KeyGen_{SE}$: k is the key generation algorithm that generates κ using security parameter 1. Enc_{SE} (k, M): C is the symmetric encryption algorithm that takes the secret κ and message M and then outputs the ciphertext C; Dec_{SE} (k, C): M is the symmetric decryption algorithm that takes the secret κ and ciphertext C and then outputs the original message M.

[3.2] CONFIDENTIAL ENCRYPTION

It provides data confidentiality in deduplication. A user 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.

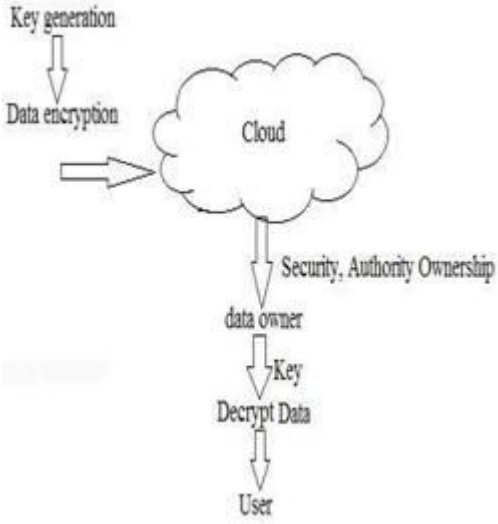

Figure: 3 confidential data encryption

[3.3] PROOF OF DATA

The user have to prove that the data which he want to upload or download is its own data. That means he have to provide the convergent key and verifying data to prove his ownership at server.

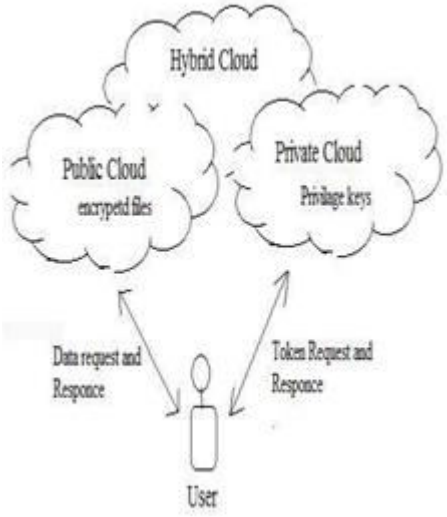

Figure: 4 System Architecture.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IV. EXISTING SYSTEM APPROACH

From the above literature survey we have concluded that existing data de-duplication systems, the private cloud are involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges. Such architecture is practical and has attracted much attention from researchers. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud. We present an advanced scheme to support stronger security by encrypting the file with differential privilege keys. In this way, the users without corresponding privileges cannot perform the duplicate check. Furthermore, such unauthorized users cannot decrypt the cipher text even collude with the S-CSP.

V.RELATED WORK

Secure Deduplication. With the advent of cloud computing, secure data deduplication has attracted much attention recently from research community proposed a deduplication system in the cloud storage to reduce the storage size of the tags for integrity check. To enhance the security of deduplication and protect the data confidentiality, Bellare showed how to protect the data confidentiality by transforming the predictable message into unpredictable message. In their system, another third party called key server is introduced to generate the file tag for duplicate check. Stanek presented a novel encryption scheme that provides differential security for popular data and unpopular data. For popular data that are not particularly sensitive, the traditional conventional encryption is performed. Another two-layered encryption scheme with stronger security while supporting deduplication is proposed for unpopular data. In this way, they achieved better tradeoff between the efficiency and security of the outsourced data. Li addressed the key-management issue in block-level deduplication by distributing these keys across multiple servers after encrypting the files

Convergent Encryption Convergent encryption ensures data privacy in deduplication. Bellare formalized this primitive as message-locked encryption, and explored its application in space-efficient secure outsourced storage. Xu et al. also addressed the problem and showed a secure convergent encryption for efficient encryption, without considering issues of the key-management and block-level deduplication. There are also several implementations of convergent implementations of different convergent encryption variants for secure deduplication It is known that some commercial cloud storage providers, such as Bitcasa, also deploy convergent encryption.

Proof of ownership. Halevi et proposed the notion of "proofs of ownership" (PoW) for deduplication systems, such that a client can efficiently prove to the cloud storage server that he/she owns a file without uploading the file itself. Several PoW constructions based on the Merkle-Hash Tree are proposed to enable client-side deduplication, which include the bounded leakage setting. Pietro and Sorniotti proposed another efficient PoW scheme by choosing the projection of a file onto some randomly selected bit-positions as the file proof. Note that all the above scheme111111111111s do not consider data privacy. Recently, Ng et al. extended PoW for encrypted files, but they do not address how to minimize the key management

overhead.

Twin Clouds Architecture: Recently, Bugiel provided an architecture consisting of twin clouds for secure outsourcing of data and arbitrary computations to an entrusted commodity cloud. Zhang et also presented the hybrid cloud techniques to support privacy-aware data-intensive computing. In our work, we consider to address the authorized deduplication problem over data in public cloud. The security model of our systems is similar to those related work, where the private cloud is assume to be honest but curious

VI. CONCLUSION

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. This however does not mean that all the problems listed above have actually been solved, only that the according risks can be tolerated to a certain degree. Cloud computing is therefore still as much a research topic, as it is a market offering. For better confidentiality and security in cloud computing we have proposed 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. Proposed system includes proof of data owner so it will help to implement better security issues in cloud computing.

REFERENCES

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

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

[3] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with efficient and reliable convergent key management. In *IEEE Transactions on Parallel and Distributed Systems*, 2013.

[4] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, *ACM Conference on Computer and Communications Security*, pages 491–500. ACM, 2011.

[5] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with efficient and reliable convergent key management. In *IEEE Transactions on Parallel and Distributed Systems*, 2013.

[6] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In

Proc. of APSYS, Apr 2013.

[7] 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.

[8] 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.

[9] 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.

[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*, pages 81– 82. ACM.

Mrs Anamanamudi Sireesha, Scholar, M.Tech, Department of Computer Science & Engineering, St.Mary's Women's Engineering College, Budampadu, Guntur. .

Mr K Raja Rao, Assistant Professor, Department of Computer Science & Engineering, St.Mary's Women's Engineering College, Budampadu, Guntur.