Deduplication Based Storage and Retrieval of Data from Cloud Environment

[1]Gopinath kowdeed [2]Mr. S. Pradeep
[1]Mtech, Computer Science, SRM University, Chennai
[2]Assistant Professor, SRM University, Chennai

***Abstract* - Data deduplication is a method used for reducing the amount of storage space a company needs to save its data. Most of the organizations, contains the storage systems that contain duplicate copies of many blocks of data. For example, the same file may be saved in different places by the different users, or two or more files that aren't same may still include much of the same content data. Deduplication obviates these extra copies by saving just one copy of the data and replacing the other copies with pointers that guide back to the original data copy. Organizationas frequently use deduplication in backup and disaster recovery applications, but it can be used to clear up the space in primary storage as well. To abviate this duplication of data and to maintain the confidentiality in the cloud, the concept of twin cloud is introduced. To protect the confidentiality of sensitive data while supporting deduplication, the convergent encryption method has been proposed to encrypt the data before outsourcing.**

***Keywords* - Deduplication, authorized duplicate check, confidentiality, hybrid cloud**

I. INTRODUCTION

Cloud computing is a rising technology that recently has drawn vital attention from each trade and academia. It provides services over the web, cloud computing user will utilize the net services of various package rather than buying or placing them on their own computers. It has a service-oriented architecture in which services are divided into three categories mainly, they are: Infrastructure-as-a- Service (IaaS), which includes equipment such as Storage and networking components which are made accessible over the Internet. Platform-as-a-Service (PaaS), which includes hardware and software computing platforms such as virtualized servers, And Software-as-a-Service (SaaS), which includes software applications and other hosted services. A cloud service is different from traditional hosting in three principal aspects. It is provided on demand. It is elastic since users can use the service as much as or as little as they want at any given time and the service is fully managed by the provider. The increasing amount of data is being stored in the cloud and shared among the users with specific privileges, which define the access rights of the stored data. . One of the critical challenge of cloud storage services is the management of the ever-increasing volume of data. To make data management ascendable in cloud computing, deduplication has been introduced and it has attracted more and more attention recently. Data deduplication is a data compression technique for eliminating duplicate copies of repeating data in storage. Intelligent (data) compression and single-instance(data) storage are the techniques used to improve data storage utilization.

A Hybrid Cloud is a combination of private clouds and public clouds in which some of the critical data occupies in the organization's private cloud while the other data is stored in and accessible from a public cloud. Hybrid clouds try to deliver the advantages of scalability, dependability, fast deployment and potential cost savings of public clouds with the security and increased control and management of private clouds. As cloud computing becomes notable, an increasing amount of data is being stored in the cloud and used among users with specific privileges, which define the access rights of the stored data.

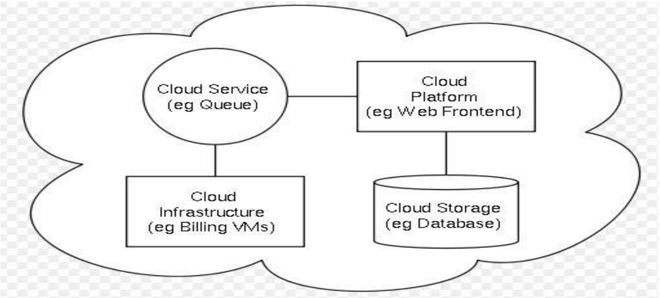

Figure: 1. Architecture of cloud computing

The critical challenge of cloud storage is the management of the increasing volume of data. Data deduplication essentially refers to the elimination of pleonastic data. In the process of deduplication, the duplicate copies of data are deleted, allowing only one copy to be stored. However, indexing of all data is still maintained 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. 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. Using the data deduplication within a hybrid cloud, the encryption becomes simpler.

The network consists of abundant amount of data, which is shared among the users and nodes in the network. Many organizations use the data cloud to store the data and share their data on the network. The node or the user, which is present in the network have full rights to upload or download data. But most of the times different users upload same data on the network, which creates the duplication in the cloud. When the user wants to retrieve the data or download the data from cloud, he has to use the two encrypted files of same data every time. The cloud performs the 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.

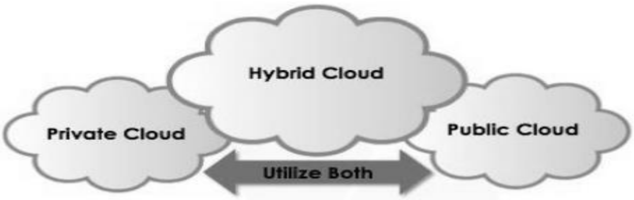

Figure: 2. Architecture of Hybrid cloud.

To avoid this duplication of data and to maintain the confidentiality in the cloud the concept of Hybrid cloud is introduced. The hybrid cloud is combination of public and private cloud.

Another way to think about data deduplication is by where it occurs. When the deduplication occurs where the data is stored then it is often cited as "source deduplication". The other one is the "target deduplication", in which the deduplication occurs where the data is stored. Source deduplication assures that data on the data source is deduplicated. This generally takes place within a file system. The file system will scan new files creating hashes and compare them to hashes of existing files at a regular time intervals.

When files with the same hashes are found then the file copy is removed and the new file points to the old file. Duplicated files are considered to be separate entities and if one of the duplicated files is later altered, then using a system called Copy- on-write a copy of that file or changed block is created. The deduplication process is pellucid to the users and the backup applications. choking a deduplicated file system will often cause duplication to occur resulting in the backups being bigger than the root data. Target deduplication is the process of removing duplicates of data in the thirdhand store. Generally this will be a backup stock such as a data repository.

One of the most common forms of data deduplication implementations works by comparing chunks of data to find the duplicates. For that, each chunk of data is assigned an identification, measured by the software, generally using cryptographic hash functions. In many implementations, the premise is made that if the identification is identical, then 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 identifiers are similar , 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, depending on the implementation, then it substitutes that duplicate data block with a link. Once the data has been deduplicated, upon read back of the file, wherever a link is found, the system simply substitutes that link with the already existing data block. The deduplication process is intended to be transparent to end users and applications

II. PROPOSED SYSTEM

In the proposed system, the data deduplication is achieved by providing the proof of data by the data owner. This proof is used at the time of uploading of the each 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 approach the files. Before submitting his duplicate check request for some data 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.

Encryption of Files

Here the common secret key *k* is used to encrypt as well as decrypt the data. This is used to convert the plain text to cipher text and again cipher text to plain text. Here, the basic three functions are used.

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*.

Confidential Encryption

The confidential encryption provides data confidentiality in deduplication. A convergent key is derived by the user from each original data copy and encrypted with that convergent key. In addition, a *tag* for the data copy is derived, such that the tag will be used to find the duplicates.

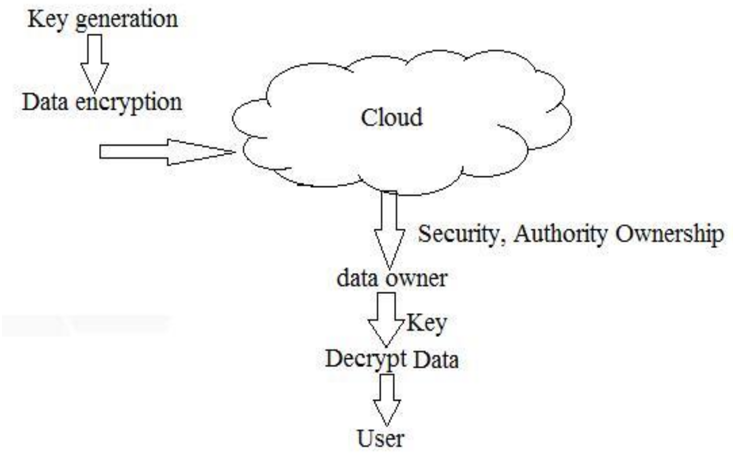

Figure: 3 confidential data encryption

Proof of Data

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

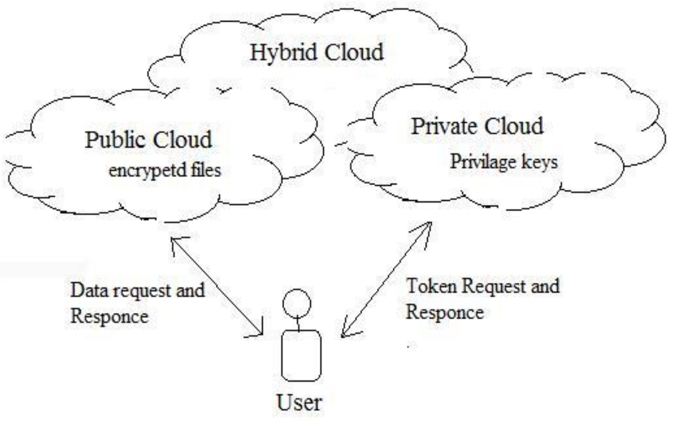

Figure: 4 System Architecture

III. CONCLUSION

Cloud computing has reached a maturity that leads it into a generative phase. The most of the main issues with cloud computing have been addressed to a degree that clouds have become interesting for full commercial development. Cloud computing is therefore still as much a research topic, as it is a market offering. For better confidentiality and security in cloud computing the proposed method has 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.

[11] S. Quinlan and S. Dorward. Venti: a new approach to archival storage. In *Proc. USENIX FAST*, Jan 2002.

[12] 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 *3rd International Workshop on Security in Cloud Computing*, 2011.

[13] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman. Role-based access control models. *IEEE Computer*, 29:38–47, Feb 1996.

[14] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure data deduplication scheme for cloud storage. In *Technical Report*, 2013.