ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A Survey on Secure Authorized Deduplication using Private and Public Cloud

S. B. Chaudhari , Supriya G. Jadhav

Asst. Professor, Dept. of Computer, Trinity College of Engineering, Pune, Maharastra, India

M.E Student, Dept. of Computer, Trinity College Of Engineering, Pune, Maharastra, India

ABSTRACT: The data deduplication is a technique to reduce the data space required on cloud to store the files. In this paper we have introduce the new deduplication technique which can check the duplication of data on file content level. 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: Hybrid cloud, public cloud, credentials, cloud storage, proof of ownership.

I. INTRODUCTION

Cloud computing having a wide range of scope now a days. Cloud provides most of the virtual environment hiding the platform and operating systems of the user. User get use of resources. User have to pay as per the use of the resources of the cloud. Now cloud service providers are offering cloud services with very low cost and also with high reliability. Large amount of data is get uploaded on the cloud and shared by millions of the users. Cloud providers offer different services such as infrastructure as a service, platform as a service, etc. User not need to purchase the resources. As the data is get uploaded by the user every day it is critical task to manage this ever increasing data on the cloud. In the cloud computing deduplication of the data [7] is best method to make well data management. This method for data deduplication check is becoming more attraction now a days. Data duplication is the technique of reducing the size of data or it is the best compression method for the data deduplication. The deduplication method have application in the data management and in the networking also to send the data over the network required small amount of data. Instead of keeping redundant copies of the same data deduplication only keep original copy and provide only references of the original copy to the redundant data. There are two methods of the duplication check, one is file level duplication check and other is block or content level duplication check. In the file level duplication check the file with same name are removed from the storage and in the block level deduplication the duplicate blocks are removed. There must be need of the some security mechanism as the data deduplication is considering the user data. It will be generate privacy concern of user’s data as it is sensitive data. User need to encrypt his own data by himself in the traditional method so there are different cipher files for each new user. To avoid the unauthorized data deduplication convergent data deduplication is proposed in [8] to enforce the data confidentiality while checking the data duplication).

The cloud provide the services as shown in the above figure such as platform, services, infrastructure as a service, and database as a service. In this we are using in cloud storage as a service. To check the authorized duplicate check we are using user credentials to check the authentication of the user. In the hybrid cloud the user credentials are present at the private cloud and data of the user is at public cloud. The hybrid cloud take advantages of both public cloud and private cloud as shown in the figure 1. In the hybrid cloud architecture there are public cloud and private cloud is there.

The user credentials are present at the private cloud in the hybrid cloud and data of the user is at public cloud. The hybrid cloud take advantages of both public cloud and private cloud as shown in the figure 2. In the hybrid cloud architecture there are public cloud and private cloud is there. When any user forward request to the public cloud to

DOI: 10.15680/IJIRCCE.2015. 0312008

access the data he need to submit his information to the private cloud then private cloud will provide a file token and user can get the access to the file resides on the public cloud. In the proposed system we have used a hybrid cloud architecture.

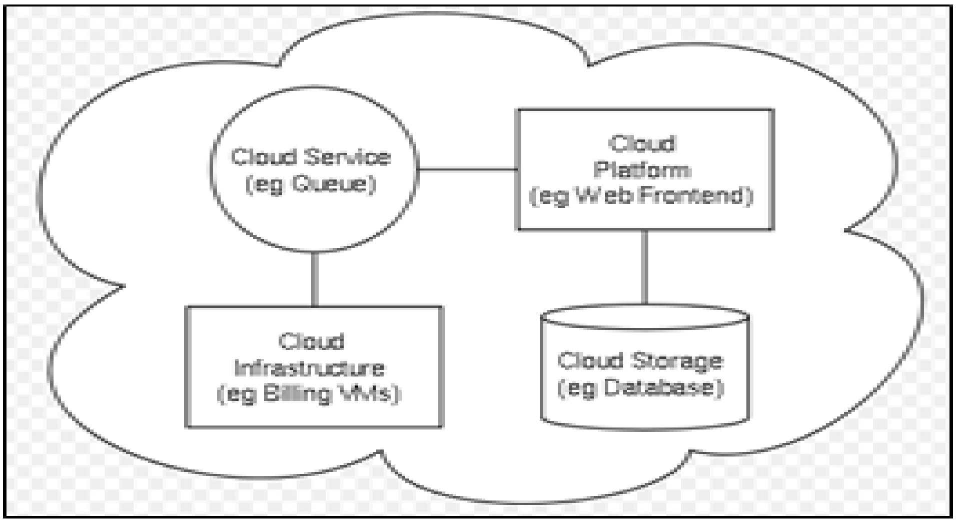

Fig. 1. Cloud architecture and services

The file data duplication is check on the primary level on the file name and then deduplication is checked at the block level of the data. When user want to retrieve data, access data file he need to be download both file from the cloud server .This will leads to perform the operation on the same file this violates the security of the cloud storage.

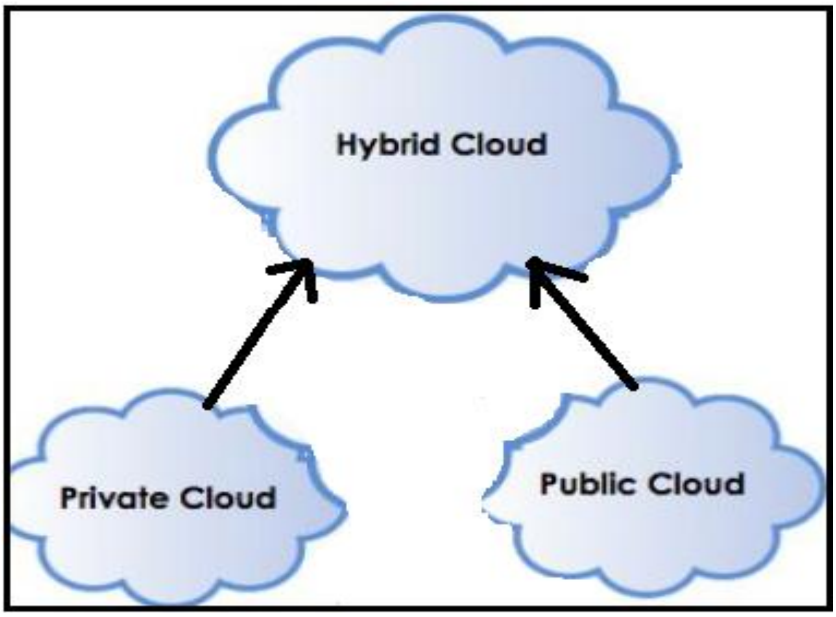

Fig. 2. Hybrid Cloud architecture

II. RELATED WORK

There are number of systems available as discussed in different articles. We have studied different methods for cloud data integrity this methods are as follow:

The paper number [9] this paper defines methods to achieving privacy and security in the Cloud and also briefly discuss the secure data sharing methods. This paper provided a survey on privacy and security in the Cloud focusing on how privacy laws should also take into consideration Cloud computing and what work can be done to prevent privacy and security attacks of one's personal data present on the Cloud. This elaborate the factors that affect the security of information of present on the cloud. It explains the needs of security for enterprises to understand the dynamics of information security in the Cloud.

J. Benaloh, M. Chase, E. Horvitz, and K. Lauter, "Patient Controlled Encryption: Ensuring Privacy of Electronic Medical Records," [10]. This paper define the broadcaster can transmit encrypted data or information to a set of users by broadcaster encryption so that only a targeted subset of users can decrypt the data. Other than above feature it is also allow Group monitor to include new member by storing previous data and user decryption secret keys need not be computed again and again, the Aggregation logic and size of cipher texts are remain same and the group encryption required different key but to decrypt the data only one key is required.

Cheng-Kang Chu, Sherman S.M. Chow, Wen-Guey Tzeng, Jianying Zhou, and Robert H. Deng," Key-Aggregate Cryptosystem for Scalable Data Sharing in Cloud Storage". This system utilize the data of cloud to encrypt and to decrypt the cloud data. The original data get divided into number of parts and some parts are used for encryption and decryption purpose. When revocation is needed owner of data required some slice to encrypt or decrypt the data. The owner of data can retrieve this signature by using intermediator and then he can allow user to upload or download the data over the clou In the existing methods of cloud storage and data deduplication. First method of the data deduplication is post processing method [3] in which data is first store on the storage device and then duplication check is applied on the data. The use of this method is there is no need to wait for calculating the hash function and the speed of storage not get downgrade. The main drawback with this system is that if storage capacity of the device is low then the file storage may get full. The post processing method is not useful at all because it checks the file after storing it on the cloud server. The another method of the duplication check is the inline duplication check [5] as it check the duplication of the file when new entries are to be added to the database. Before adding the new entry or new data to the database it will checks for the block level duplication of the file. Till this method have drawback such as each time need to calculate the hash function which may lead to slower throughput of the storage device. But the some of the vendors have proof that the inline and post processing data duplication check have same output. Another method of duplication check is source duplication check in which the file duplicate contents are checks for duplication before storing it on the cloud server. Third method of deduplication is source data deduplication in which data duplication is done at the side of the source. The file duplication is check before it get uploaded on the cloud server. The duplication is checked at the target level in which file get scanned periodically and hash get generated for the software can check for the hash value if both value get new matched with the existing hash value then the new file not get uploaded on the cloud server only link to that data is to be provide to the file user. If new file is to be added to the cloud server and it get match the hash function of the old file then it only remove the new file and just provide hard link to the old file resides on the cloud server.

Another method of the duplication calculation is chunk level duplication checker. In this for each chunk identification is get assigned generated by the software. For the preprocessing file checking we have to make some assumption that identification is same then data is also same but this is not true in all the cases due to the pigeonhole principal. It will produce wrong result that if for two blocks of the data same identification number is get generated it simply remove the one block of the data.

III. Proposed Algorithm

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 3.

Figure 3 shows the proposed system architecture which comprises of public cloud, private cloud and user. In the proposed system architecture shown in Figure 3. 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.

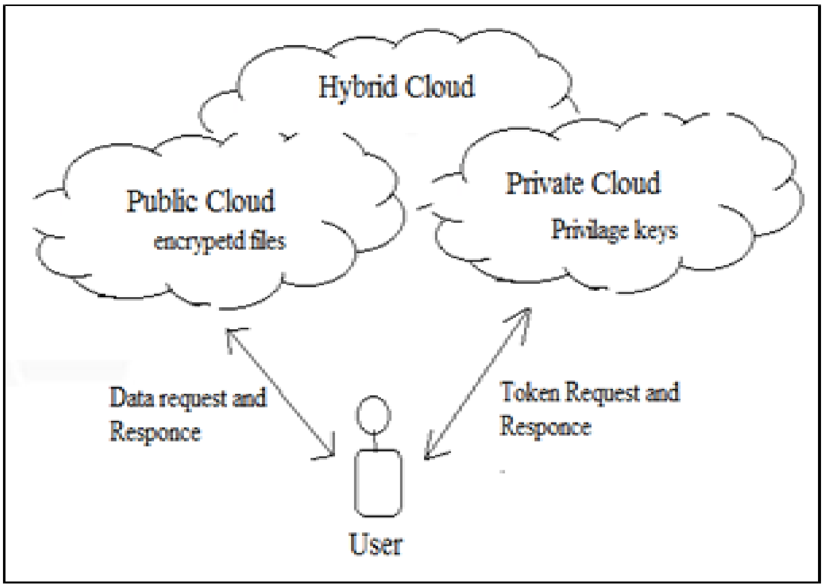

Fig. 3.System architecture

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 converget 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.

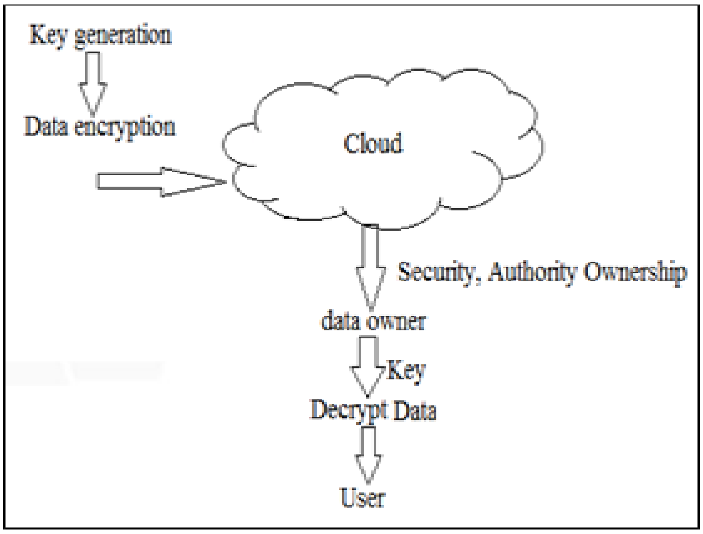

Fig. 4.Confidential data encryption

IV. Conclusion and Future Work

This paper shows that the proposed method for data deduplication is authorized and securely duplication of the file is done. In this we have also proposed new duplication check method which generate the token for the private file. As a proof of ownership of the data user need to submit the privilege along with the convergent key. We have solved more critical part of the cloud data storage which is only tolerated by different methods. Proposed methods ensures the data duplication securely. In future we are designing this project to check duplication of image or pdf file.

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 TermSuggestion in Interactive Web Search Based on ContextualInformation in Query Session Logs," J. Am. Soc. for Information science and Technology, vol. 54, no. 7, pp. 638-649, 2003.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.
8. 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.
9. S. Quinlan and S. Dorward. Venti: a new approach to archival storage. In Proc. USENIX FAST, Jan 2002.
10. 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 3rd International Workshop on Security in Cloud Computing, 2011.

BIOGRAPHY

S. B. Chaudhari is Assistant Professor at, Department of Computer Engineering, Trinity college of Engineering Pune, Maharastra, India.

Supriya Gorakh Jadhav is ME, Student at Computer Dept. Trinity College Of Engineering,Pune, Maharastra, India.

DOI: 10.15680/IJIRCCE.2015. 0312008