Data Deduplication using Hybrid Cloud Architecture

Gaurav kakariya
Department Of Information Technology
Siddhant College of Engineering Sudumbare
Pune,Maharashtra
E-mail:gauravkakariya007@gmail.com

Sonali Rangdale
Department Of Information Technology
Siddhant College of Engineering Sudumbare
Pune,Maharashtra
E-mail: sonali_rangdale@rediffmail.com

***Abstract*—Cloud computing is getting wide range of scope now a days. The infrastructure as a service facility of the cloud is most popular among those. The data deduplication in the cloud computing is the process of removing same files from the cloud and save the user space on the cloud. In large organizations same data is stored on the different places by different users. This will increase the storage size. In the duplicate removal process one can remove the file duplicate with the original file and make space empty for the further storage. To avoid the data duplication and also to maintain the user confidentiality we have proposed a novel method data deduplication by using hybrid cloud. Proposed system also maintain the confidentiality of the user data as all credentials are stored on the private cloud. We have proposed a novel method to improve the data storage. Proposed method insures data deduplication securely. This method ensures the data deduplication by secure way.**

Keywords- data deduplication, hybrid cloud, public cloud, credentials, cloud storage.

I. INTRODUCTION

Current era is cloud computing era. Use of cloud computing is increasing rapidly. The amount of data over the cloud is also increasing every day. The main problem with cloud computing is the large volume of data present on it. This may leads to the space problem. In the cloud computing user have to pay as per his use. If user uploaded the same file over the cloud it may leads to the loss of resource. So there is need of removing duplicate files of repeating data. This method useful to utilize the storage space and also reduces the amount of bandwidth required to send the data over the internet. In the deduplication method duplicate chunk or pattern or file name are replace by the small data.

The hybrid cloud used in the proposed system comprises of public an private cloud in the private cloud all user credentials are presents and in the public cloud data assessable publically is stored. Hybrid cloud provides all easy management and storage facilities an also well resources utilization. As the popularity of cloud is increasing day by day. The use therefor the data resides on the cloud is also increasing daily. In our proposed method we have add the some credentials to provide security to the user data. At the time of file upload the admin ask to user for enter the file password and at the time of file download user need to provide the password that was used by the user at the time of uploading the file on the cloud server. The figure 1. Shows the architecture of the basic cloud computing. In this architecture cloud platform cloud billing and cloud virtual machines are there. The cloud computing architecture is shown in the following figure.

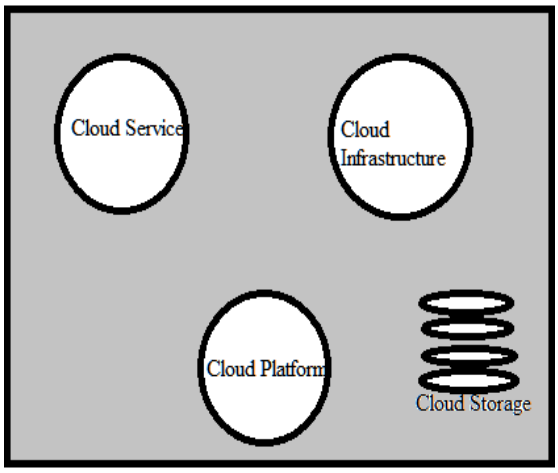

Figure. 1 Cloud architecture.

In the cloud computing user have to pay as per his use. If user uploaded the same file over the cloud it may leads to the loss of resource. So there is need of removing duplicate files of repeating data. This method useful to utilize the storage space and also reduces the amount of bandwidth required to send the data over the internet. In the deduplication method duplicate chunk or pattern or file name are replace by the small data.

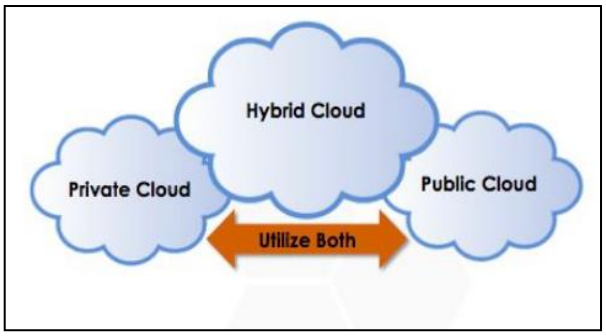

Figure.2 Architecture of hybride cloud

Continuously increasing data on the cloud server is main problem in the cloud computing. Data deduplication or single instancing refers to the removal of duplicate files. In the data deduplication method duplicate files are removed keeping only one instance or data copy on the server side but the indexing all the data is also very critical process. In simple way data deduplication removes the duplicate files from the server side keeping only one file on the server. Data is encrypted before uploading over the cloud to provide the security to the uploaded data. This will take more time and it is more complex to encrypt the data before store it on the cloud. The encryption of the file become more complex for the large size files. Because of this time consuming task of the data encryption we are using data deduplication to remove the duplicate files and to minimize the time required to encrypt the data every time. As the network consist of large amount of data , which is being shared by users and nodes in the network. The user having access to the cloud have full right of upload and download the file on the cloud server. The main reason behind increase in the data on the cloud is number of users are uploading the same data on the cloud and due to this same files present over the cloud no of duplicate file are increasing every day. . If user wants to download the data from the cloud server user need to download same files even if his data is present in only one file. 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. The following figure shows the hybrid cloud. It comprises of private and public cloud in it.

II. LITERATURE SURVEY

In this section we will see the existing methods studied about the cloud computing. In the existing system the data duplication was not checked. In previous system each user have assigned some privilege to access the particular file and only that file can user access [2],[4]. This not useful to reduce the duplicate files present on the cloud server.In the old method of the deduplication technique if device found the new file on the cloud server then it will simply refer the previous file stored on the server. The main benefit of the inline duplication is it not required to check the duplicate data over the cloud server. On the other side if data duplicate is present on the other side of the server this duplicate data cannot be removed by using the in line deduplication method. But some vendors have proposed a method to remove the inline deduplication data. Post processing and inline data deduplication are in debate. Another method of the data deduplication is remove the duplicate data from where it is created that is the source of the duplicate file is need to find out. This is nothing but the source deduplication [5]. This source side data deduplication insures that the data get deduplicate on the source side. This is generally happens in the file system. The file are periodically scan and hash function of the file is generated and if new created hash matches with the existing hash function then file get removed. The data deduplication technique is to remove the duplicate files from the secondary side that is from the user side file removal. The data duplication is carried out in such a way that a chunk is taken and applying the algorithm on that chunk unique id is generated that is nothing but the hash value of each file. As compare to the original file the size of the hash function is very less. But if you change the file name value of that hash is also get changed.

Each chunk of data get assigned by the software calculated hash values. In many case it get assume that data is identical, and chunk are also same but this is not true in al the cases due to pigeonhole phenomenon. Second assumption is that the data with same hash value are same but this is not true in all the cases the data may be identical or not. There are two issue with this if software assume that Once the data has been deduplicated, depending upon the readback if file found to be duplicate it get removed from the database. The existing methods not supports the secure deduplication system. In the existing all the system duplication check is done on the basis of name of the files. If the file with same name is found then it get replace by the link to the previous file present already in the database. The convergent cipher text allows user to check the duplication of the file on the name of that file. Proof of ownership is avoid the file access by unauthorized user on the cloud.

III. PROPOSED SYSTEM

In the proposed deduplication technique we are achieving the data deduplication by applying the proof of ownership applied by the owner of the data. The proof is applied at the time of uploading the file. The file uploaded to the cloud is bounded by some value that specifies which user can have the access to the particular file. Without this proof of ownership user is not able to perform the duplication check on the file. Figure 3 shows the system architecture of the proposed system.

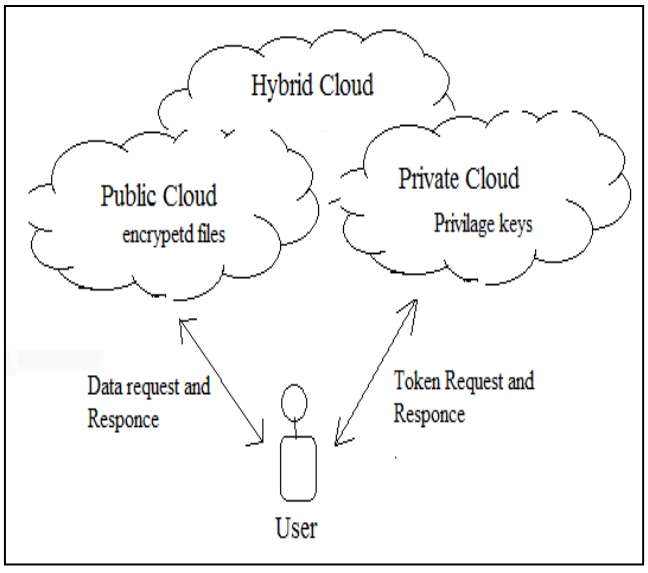

Figure 3. System Architecture

User need to submit his proof of ownership and access level privilege to access the file which can get the user from the private cloud. User is able to find the file

duplication of his own file if and only if he match the privileges of his file stored on the public cloud.

A. *Encryption of Files and data:*

In this we are using the common key to encrypt and decrypt the text. This secrete key k is used to convert the plain text to cipher text and get plain text back from the cipher text. To achieve this we have used three basic function $KeyGen_{SE:}$ it is the key generation algorithm which generate a key to generate the key value. Second one is Enc_{SE} (k, *M*): *C* is the second function which takes Message M as a input and apply key K on it and generate the ciphertext C.

B. *Confidential Encryption Method:*

This helps to generate the confidentiality check. Each original file get encrypted by using a convergent key. Also in the proposed system user can also generate the tag which are helpful to remove the duplicate files. The key generation algorithm is used to generate the key and this generated key is used to encrypt the data and then proof for the data ownership is need to provide . Then the data is get encrypted and stored on the cloud storage. At the time of downloading the data user need to provide his proof of ownership and also need a key to get original content back. This method ensures that the duplication is not in the file. The confidential encryption method is shown in the following figure.

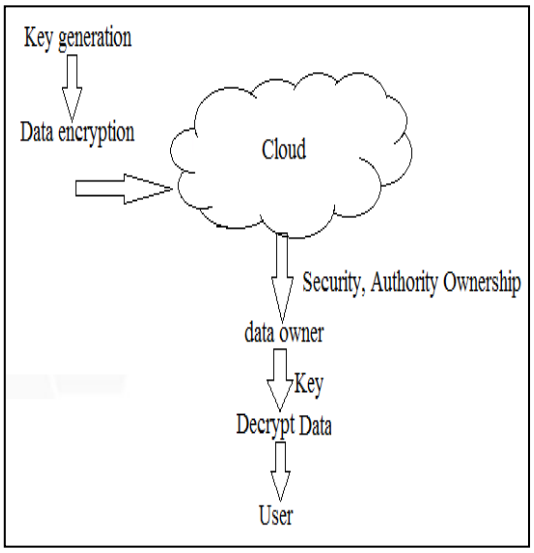

Figure 4. Confidential data encryption

C. *Proof of data:*

At the time of file download or file upload user need to provide his proof of ownership that means he need to submit his file details and convergent key at the time of data upload and data download.

In the proposed system we are using data duplication check by using the file content. The duplication is check at the content level of the file. User needs to provide his proof of ownership at the time of uploading or downloading the data from the cloud. To solve the problem of the existing file duplication check we have introduced novel approach which can check the file duplication at the content level of the file. The user credential is managed at the private clout side. To get the assess to the particular file user needs to send the token request to the private cloud and then the according to the privilege set at the private cloud user get the token for the file. The authentication of the user is done both the places at the private cloud and at the public cloud also.

D. *Procedure for Deduplication:*

1. Cloud User Requests to store file Fi on cloud Ci.
2. Before file contents are checked, File with same name availability is checked from private cloud token list.
3. If token already assigned to same filename, user is asked to check for block level duplication.
4. A new token is assigned to the file Fi . consider that to ken to be Ti. Ti is computed using MD5 Algorithm.
5. If the Previously available token and current computed token are same.
6. Drop the file to be uploaded as its duplicate file getting uploaded,
7. If tokens are different, i.e. file contents are different.
8. Upload the file to requested cloud.

IV. EXPEREMENTAL ANALYSIS

We will do this experiment with the text file. To generate the hash function of the uploaded file we are using MD5 algorithm. This algorithm generate the hash value of the file get uploaded on the cloud. Our proposed system will check the file content level duplication. Our method is independent of the file size. To complete the experiment we have take three system one of which is client Which acts as a user and performs the operation such as file upload, download etc. The second system acts as Private Cloud which manage the key and generate the token for the user file access. Third is Server system which acts as C-CSP which contains data of the user on it.

V. CONCLUSION

In this paper, the file deduplication is addressed. We have proposed a novel method to securely check the duplication. Cloud computing has reach to his maturity level. That means the cloud is used by commercial user in very large amount. It does not mean that all the problems of cloud computing are solved totally. This problems are only tolerated according to the situation. Due to this cloud computing is more research part. We have proposed novel scheme to remove the duplicate files from the cloud storage. This method insures that the proposed method successfully remove the duplicate files from the cloud storage.

ACKNOWLEDGMENT

Author would like to take this opportunity to express our profound gratitude and deep regard to my guide prof. Sonali Rangdale for his exemplary guidance, valuable

feedback and constant encouragement throughout the duration of the project. His valuable suggestions were of immense help throughout my project work. His perceptive criticism kept me working to make this project in a much better way. Working under him was an extremely knowledgeable experience for me.

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.-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.

[7] 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] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] [6]C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In *Proc. of APSYS*, Apr 2013.