International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Authorized Data Deduplication Using Hybrid Cloud Technique

Snehal Baravkar[1], Vaishali Mali[2]

[1]Shree Ramchandra College of Engineering, Pune

[2] Professor, Shree Ramchandra College of Engineering, Pune,

Abstract-Now a days use of cloud computing is increasing rapidly. Cloud computing is very important in the data sharing application. Daily use of cloud is increasing. But the problem in cloud computing is every day data uploaded on the cloud, so increasing similar data in cloud. Therefore we can reduce the size of similar data in cloud using the data DE duplication method. These method main aim is that remove duplicate data from cloud. It can also help to save storage space and bandwidth. Our proposed method is to remove the duplicate data but in which user have assigned some privilege according to that duplication check. Cloud DE duplication is achieve using the hybrid cloud architecture. We proposed method is more secure and consumes less resources of cloud. Also we have shown that proposed scheme has minimal overhead in duplicate removal as compared to the normal DE duplication technique.

Keywords: Authorization; data security; privilege; DE duplication; credentials; cloud.

1. Introduction

Current era is cloud computing era. Now a days cloud computing has wide range of scope in data sharing. Cloud computing is provide large amount of virtual environment hiding the platform and operating systems of the user. User use the resources for sharing data. But user have to pay as per the use of resources of cloud. Now cloud service providers are offering cloud services with very low cost and also with high reliability. User can upload the large amount data on cloud and shared data to millions of users. Cloud providers is 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. DE duplication is best method to make well data management in the cloud computing. This method is becoming more attraction for data DE duplication. This method is send the data over the network required small amount of data. This method has application in data management and networking. Data duplication is the technique of reducing the size of data Also it is the best compression method for the data DE duplication. This method is send the data over the network required small amount of data. This method have application in data management and networking. Instead of keeping redundant copies of the same data DE duplication 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 content level duplication check. In the file level duplication check is remove the same name file from the storage and block level DE duplication are removed the duplicate blocks. As the data DE duplication is considering the user data there must be need of the some security mechanism. It arises security and privacy concern of the user's sensitive data. In the traditional method user need to encrypt his own data by himself so there are different cipher files for each new user.

To avoid the unauthorized data DE duplication convergent data DE duplication is proposed in [8] to enforce the data confidentiality while checking the data duplication.The cloud providing many 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. We are using user credentials to check the authentication of the user. In the hybrid cloud is present two type of cloud such private cloud and public cloud. In private cloud store the user credential and user

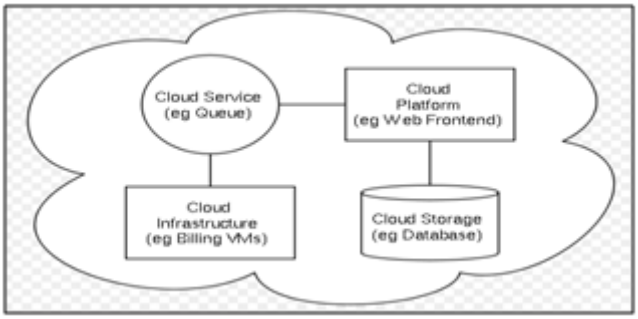

Figure 1: Cloud architecture and services

data present in public cloud. The hybrid cloud take advantages of both public cloud and private cloud as shown in the figure 2. public cloud and private cloud are present in the hybrid cloud architecture When any user forward request to the public cloud to 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. We have used a hybrid cloud architecture in proposed. The file name is check on primary level in file data duplication and data DE duplication is checked at the block level. If user wants to retrieve his data or download the data file he need to download both of the file from the cloud server this will leads to perform the operation on the same file this violates the security of the cloud storage.

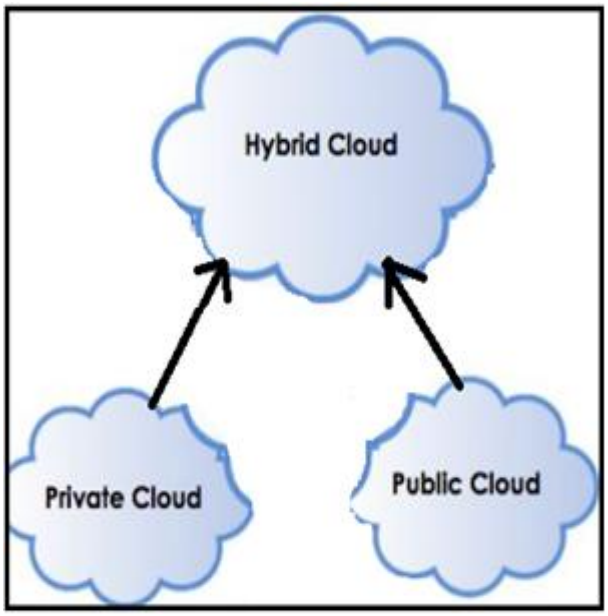

Figure 2: Hybrid Cloud Architecture

2. Literature Survey

There are so many researches have been done to secure duplication check of data on cloud. The cloud storage and data DE duplication are two methods present in existing system. First method of the data DE duplication is perform as post processing method [3] In this 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. Some problem of this the post processing method is not useful at all because it checks the file after storing it on the cloud server. Second method of the duplication check is the inline duplication check. It is check when new entries are to be added to the database the duplication of the file. It will checks for the block level duplication of the file before adding the new entry or new data to the database. This method have some 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 data duplication check have same output in the inline and post processing method. 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 DE duplication is source data DE duplication 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.

Chunk level duplication checker is another method of the duplication calculation. In this for each chunk identification is get assigned generated by the software. For the pre-processing 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.

3. Proposed System

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 is decide who can perform duplication check of the file. User is need to submit his file and proof of ownership of the file before sending the request to for the duplicate check Request to the cloud. When there is file on the cloud and also privileges of the user only that time to approved the duplicate check request.

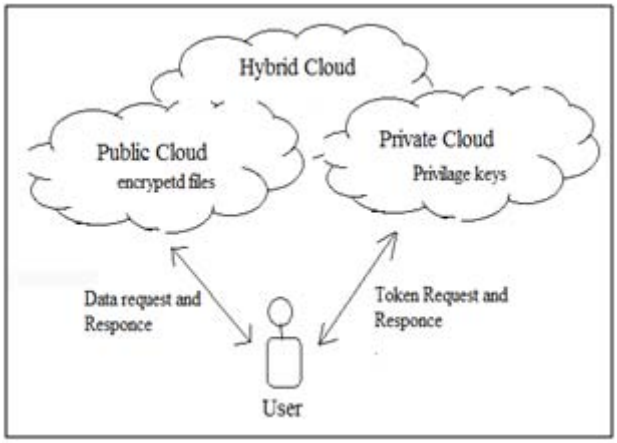

Figure 3: Overview of the system

Above fig. shows the proposed system architecture which comprises of public cloud, private cloud and user.

Proposed system architecture includes only one public cloud and one is private cloud. All data of user is contains in public cloud such as files. And private cloud consists of user credentials. User for each transaction with the public cloud need to take token from 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.

A. Encryption of File:

We are using secrete key resides at the private cloud to encrypt the user data. 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 follows:

1) Key GenSE: It is generate the secrete file by using security parameter. In this k is the key generation algorithm.

2) EncSE (k, M): In this we have generated a cipher text using formulae M is the text message and k is the secrete key
3) DecSE (k, C): In this we have to generate plain text using C is the cipher text and k is the encryption key.

B. 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 converged key generation algorithm to encrypt the user data. This will ensures the security, ownership and authority of the data.

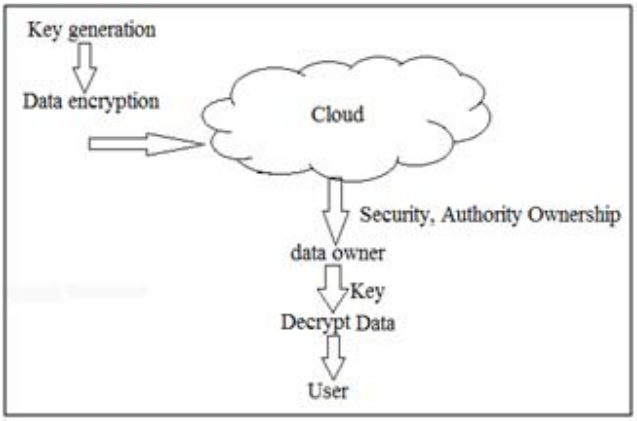

Figure 4: Confidential data encryption.

C. Proof of Data

When file upload and download user need to provide proof of the data. User need to submit his convergent key which was generated at the time of file upload. To generate the hash value of the data we have used MD5 massage digest version 5 algorithm to generate the hash value of the user data. If there is any change in data occur the hash value of that data get changed.

4. Mathematical Module

A. Set theory

User Authentication:-
Set (C) = {c0, c1, c2, c3, c4}
C0= Get User Id
C1=Get Cloud Id.
C2=Get Data Owner Info
C3=get the User Privilege Information
C4= Get key from hash table.
Data DE duplication:-
Set (T) = {c1, c2, d0, d1, d2}
d0= Get Data File Name.
d1=Data accessing user id.
d2=Get Cloud id

B. Union and Intersection of project:-

Set (C) = {c0, c1, c2, c3, c4}
Set (T) = {c1, c2, d0, d1, d2}
P union T – {c0, c1, c2, c3, c4, d0, d1, d2}
P Intersection T = {c1, c2}

C. Venn Diagram

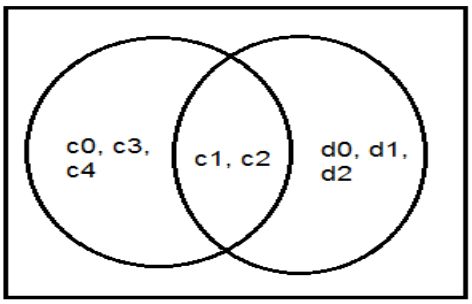

Figure 5: Vein Diagram

5. Experimental Analysis

We have modelled three systems one client program which is used as a user which can upload or download the file. Second one is Private server program which acts as a private cloud and perform operation such as users key management and token generation. Third is the server program which acts as S-CSP which contains the user files on it. We have used MD5 algorithm to generate the hash value of the data which is low complex as compare to other digesting algorithm. The proposed method is independent of the file size which is get upload. In the proposed system we are doing block level duplication check.

6. Comparison between Existing and Proposed System

Table 1: Comparison Between Existing and Proposed System

Existing System	Proposed System
Insecure and unauthorized approach.	Security is improved by using tokens.
No user verification was done.	User must verify here to upload or download data.
It causes data congestion over cloud.	This helps to avoid the data congestion and data maintains problem of the cloud server
Less security provided	Security is achieve by including differential privileges of users in the duplicate check.
Fail to prevent insider and outsider attacks	This system is secure in terms of insider and outsider attacks

7. Results

This system should prevent user from uploading duplicate data on cloud. Data stored on cloud must be in secure encrypted format. Malicious user not able to upload or download data on cloud. The user who has proof of ownership only that user can modify data.

Figure 6: Home

Figure 7: User Registration

Figure 8: User Login

Figure 9: Private Cloud Login

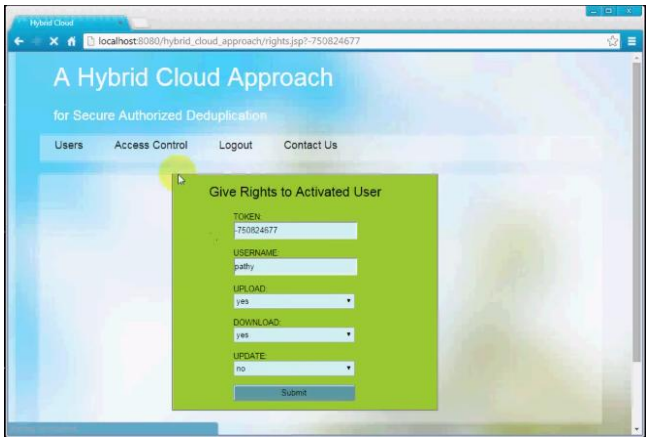

Figure 10: User activation by Admin

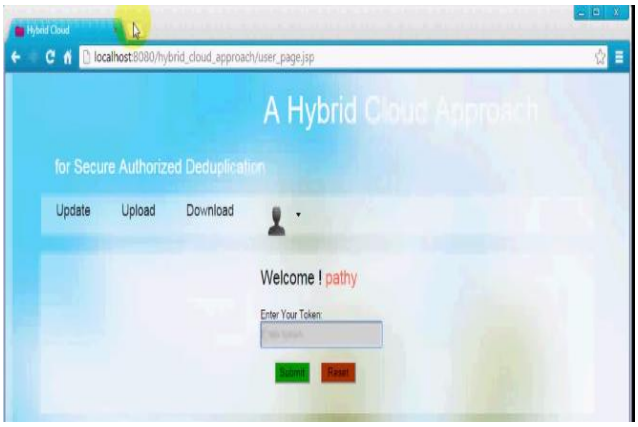

Figure 11: User Window

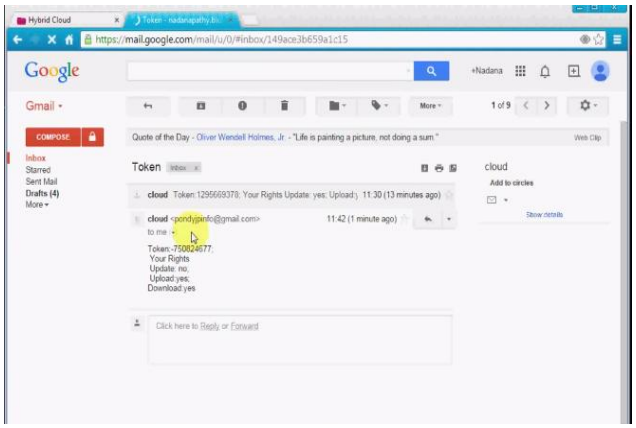

Figure 12: Activation token mail

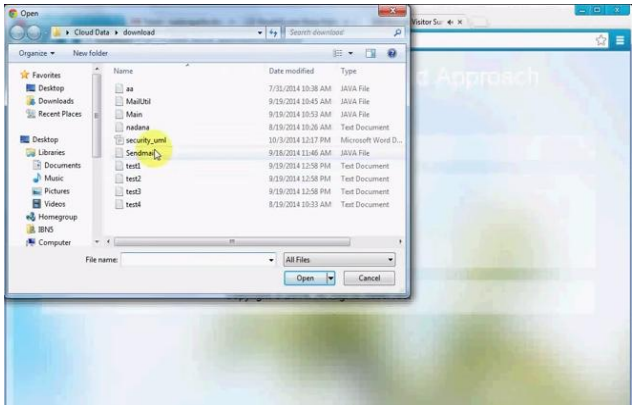

Figure 13: File Upload

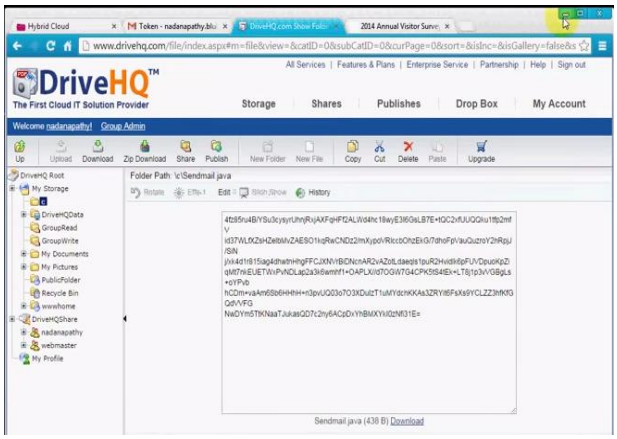

Figure 14: File on drive HQ

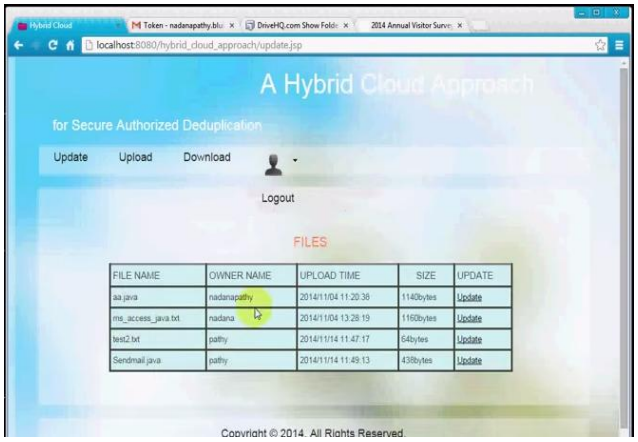

Figure 15: Admin Window

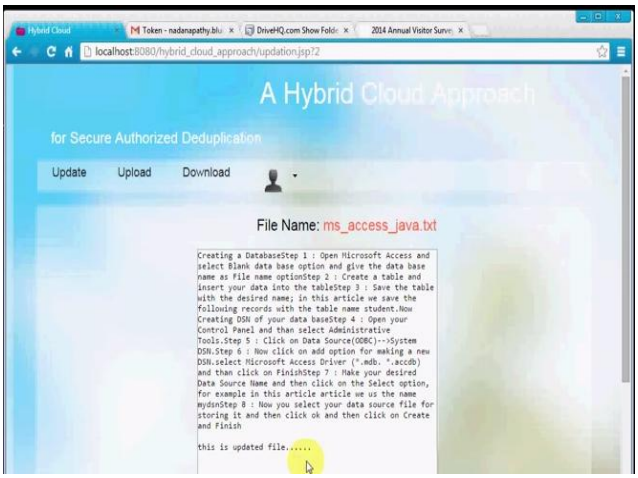

Figure 16: File Update

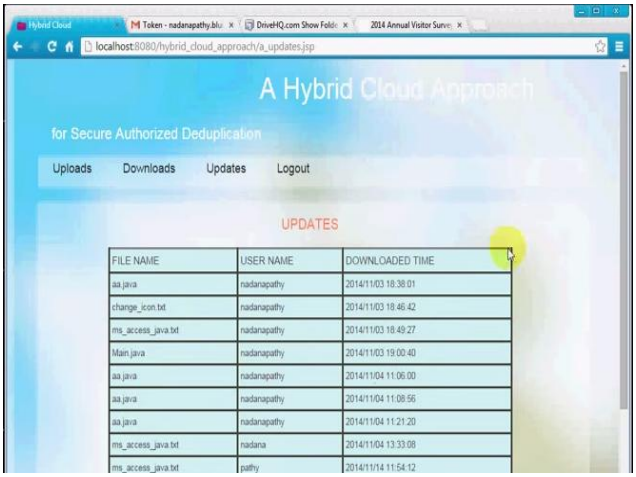

Figure 17: File Download

8. Conclusion

Here we can conclude that our proposed system data DE duplication of file is done authorizes way and securely. . In this we have also proposed new duplication check method which generate the token for the private file. The data user need to submit the privilege along with the convergent key as a proof of ownership. 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

9. Acknowledgment

Author would like to thank The Principal of Shree Ramchandra College of Engineering for giving her the opportunity to work on this project. She would like to show her sincere gratitude to her guide Prof. Vaishali Mali for her guidance and knowledge without which this paper would not be possible. She provided me with valuable advice which helped me to accomplish writing this paper. She is also thankful to her H.O.D Prof. Deepti Varshney (Department of Computer Engineering) for her constant encouragement and moral support. Also she would like to thank M.E Co-ordinator Prof. Baban Thombre for his support and encouragement, which helped her in correcting mistakes and proceed to complete the paper with the required standards.

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.
[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] 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.
[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 DE duplication scheme for cloud storage. In Technical Report, 2013.