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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Survey on Identical Encoded Data Management in Cloud Storage

Pallavi Bangale, Prof. Rekha Kulkarni

M.E. Student, Dept. of Computer Engineering, Pune Institute of Computer Technology, Pune, India.

Dept. of Computer Engineering, Pune Institute of Computer Technology, Pune, India

ABSTRACT: Cloud Computing is the emerging distributed computing technology that provides the huge physical storage for the end user. The underlying technique of cloud virtualization helps to improve the availability of physical storage. As the data increases, to reduce the complexity of handling such big data deduplication is necessary. While handling such big data, security concerns are also considered. But deduplication doesn't support the encrypted security in that flexible way. The existing solutions compromise about security for encrypted big data. In this paper, we propose a solution that works on the encrypted data for deduplication using the proxy re-encryption to deduplicate the data on cloud storage without using the additional server. By using re-encryption, the redundancy due to encrypted data is avoided and efficiency of cloud storage can be increased.

KEYWORDS: Big data, Cloud Computing, Data Deduplication

I. INTRODUCTION

Cloud Computing is the distributed technology used for the ubiquitous computing from many machines located remotely. This technology enables to use the resources of the remote machines as their own and support the high performance computing by performing the application on those machines. The backbone of the Cloud Computing is the virtualization in which the virtual version of the resourcesis created for the computation. One type of the virtualization is the storage virtualization in whichmultiple physical storage system from a single logical storage. This virtual storage has memory to store the big data and user unable to tell the difference between the physical and logical storage system. As the cloud service provider(CSP) has to maintain the logical pool of such large memory, the efficiency and performance of storage should be superior. But as the data storage increases, the data management becomes complicated. The major reason behind the complex data management is the duplicate files reside on the cloud. These files consume the network, storage resources and also causes the energy wastage.

Deduplication of the data is one of the technique to reduce the cost, and storage demand for the cloud application. It avoids the uploading of the redundant data and provides links to that single copy irrespective of number of clients requesting that file. Deduplication divides the file in chunks and then each chunk is compared with the data chunks in the cloud storage to avoid the double uploading of the data. Deduplication can provide the 90-95 percent saving in the backup of the data [3]. Deduplication also reduces bandwidth requirement of the cloud service. More than 90% of the bandwidth is saved because of the deduplication [2]. The advantages of the deduplication reflects in the economic and performance benefits of the cloud service. But as the performance increases, more number of customers attract towards the technology, and security issues need to be concerned then. As the unauthorized access to the confidential information and data is one of the most important user concerns, cloud service provider should secure the cloud storage from malicious attacks performed by the intruder and third parties.

In practice, the security to the data is provided by the cryptosystem. In this, the data is encrypted using the keys by data sender and then it is decrypted by data receiver using the same or different keys depending on the type of encryption. Encryption of the data changes the original data into the ciphertext which is random unrecognizable stream of characters. Now as the data gets changed, then it is difficult to compare the data with another data. But deduplication

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

works only when the chunks of the data can be compared. Sothe encrypted data produces new challenges for data deduplication. Also, many attempts show that deduplication of encrypted data can be achieved, but they have to compromise with the security of the data. Also the additional servers are added to perform the deduplication of the data which further leads to the increased complexity of the structure. Also the cloud service provider takes time to serve the request to the client because of the additional server.

Our work proposes he system that works on the deduplication on the encrypted data without using any additional servers. The technique of the proxy re-encryption is going to be used in the system in case of the duplicate files. Also system would take the help of the ownership challenge to identify the original data ownership of the uploaded file. Using the proxy re-encryption and data ownership challenge, system can perform the data deduplication. This paper proposes the survey of the methods of the deduplication and the proposed system architecture and analysis of the system.

II. RELATED WORK

In Cloud storage service is provided by many organizations and institutes such as Google, Amazon, Dropbox which performs the data deduplication on the users data and then unnecessary upload is avoided. Google Drive uses the data ownership to identify the owner so that owner should have the additional authority to access the cloud storage. Whenever any modification is done it is only the file owner and all the other clients are then notified about the change. And also the old files are stored in the revision history of the files, with the default version is the updated one. Even Dropbox provides the data storage which provides the data deduplication. Experiments prove that uploading one. Experiments prove that uploading the same file again takes less time to upload, as at was never uploaded, just the ownership of the file is known to the clients so that the data can read, written by the permission of the owner. Reduction of time is the result of the less bandwidth requirements for the cloud access. Also the Amazon AWS provides the deduplication with the help of the StorReduce, which deals with petabytes of unstructured data for deduplication [5]. The removal of the unwanted files on Microsoft Azure is done by the Disk Deduplication in which Windows Server performs the data deduplication on virtual hard disks that are attached to virtual machines as backup[4].

Combining the deduplication and encryption is the hot topic in the research these days. Many researchers have developed methods to achieve the deduplication of encrypted data but they lack the security issue that needs to be considered. Such that in paper [1], author proposed a system to deduplicate which uses the additional server authorized party which is trusted and used by the cloud clients for re-encryption. But additional of extra server costs much and also complexity is increased. Another work [6] proposed the system that use two cloud servers, one for the storage and one for fingerprint which is the hash value of the data. Again another additional resources unnecessarily leads to the deduplication on block-level and they used extra component for the key management among the other clients and the data owner.

Also the system [8] developed by Zheng Yan used attribute based encryption to deduplicate the encrypted data. But attribute-based-encryption also suffers challenges like key coordination, key revocation. Also they lack non-efficiency and non-existence of attribute revocation method. Also the another system [9] propose the deduplication and proxy re-encryption and proof of ownership to deduplicate the file.

In order to reduce workloads, another system is proposed which uses Index Name Server(INS) [10] to manage not only the data deduplication but also the compression of file, IP information. Another work [11], TIN-Yu-Wu proposed same Index Name Server which integrates data deduplication with the facility of the automation in the reduction of the numbers. In the given INS system we cannot distinguish between the different file formats. So we have the same bandwidth for each type of the file including the text, audio, video files. This will cause unequal balancing to the same bandwidths of the different file types. Suppose any request to the server contains the access to the text but slow to the video files. Then the same bandwidth will be provided to these two files. This will cause the fast access to text files.

So current INS server cannot deduplicate the encrypted data. C. Fan and S. Y. Huang proposed a data deduplication [13] on hybrid cloud which supports deduplication on plain text and ciphertext. But this mechanism cannot support encrypted data deduplication very well. It assumes that CSP knows the encryption key, but that's not the case always. CSP cannot be fully trusted by the data holders and owners. Many developers use Elliptic Curve Cryptography (ECC),Proxy Re-encryption(PRE), or any asymmetric curve cryptography. But in this system we used the Advanced

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Encryption Standard (AES) which is also symmetric encryption algorithm. It has serval advantages over other encryption methods like it is more secure and it also supports larger key sizes. Also it is more faster.

III. PROPOSED SOLUTION

We propose a scheme to deduplicate encrypted data at CSP by applying the proxy re-encryption to issue keys to different authorized data holders based on the data ownership challenge. It is useful in situation where data holders are not available for deduplication control. The system consist of three entities:

1. CSP: Storage services are provided by CSP.
2. User: Data is uploaded by user. The system can have number of users which could save thw same encrypted data at CSP.
3. Data Owner: Data owner is the user that actually uploads the file and has ownership of that file.

The architecture of the system is shown in the Fig 1. This system has client and server side and connector is the bridge between them. Connector accepts the request from the client and then further sends to re-encryption to check the duplication. This system contains the main aspect of identifying the duplicate file. It will identify the file by using fingerprint which are the unique IDs assigned to each data chunk. Data chunks are blocks of the data also known as the token. So as the tokens are uploaded, they are compared with the already uploaded token. If token match is positive, it means that data deduplication occurred. When deduplication occurs, then Cloud Service Provider(CSP) challenges the data ownership, which checks the eligibility of the user, then it applies encryption again with the new key so that only the eligible user can have that key and only that data holder can decrypt that file. this technique of encrypting the data again with new key is called as re-encryption and all the authorized data holders have the re-encryption key.

When the data owner tries to delete the file, then CSP does not delete the file, it just makes the data owner unauthorized sp that he no longer access the file. If that file has no duplication record which means that if that file is not tried to uploaded by any other user, then that file will be deleted from cloud storage. In case, the data owner updates the original file such that the encryption key foe that file is changed, then all the other data holders are updates about the change by the cloud service provider. As we are using AES encryption technique for the system, it is more secure than the other encryption techniques.

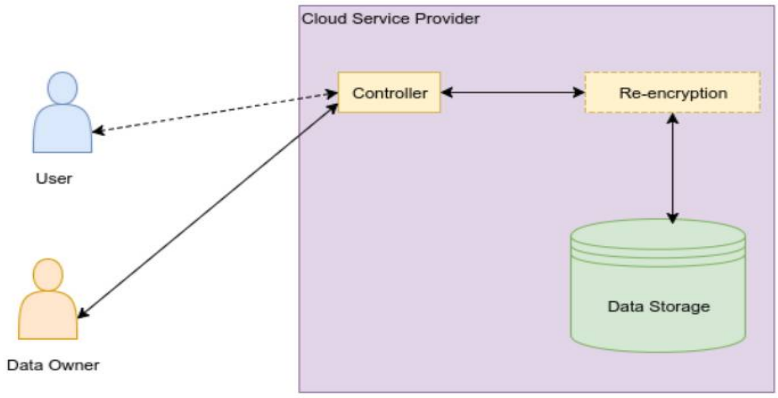

Fig. 1. Architecture of System

Our approach consist of following main aspects:

1. **Encrypted Data Upload**: If the data deduplication check is negative, its data holder encrypts its data using a randomly selected symmetric key DEK, in order to ensure the security and the privacy of the data, and stores the encrypted data at CSP together with the token used for deduplication check. The user encrypts DEK with pkCSP and passes the encrypted key to CSP.
2. **Data Deduplication**: Data deduplication occurs at the time when user u tries to store the same data that has been stored already at CSP. This is checked by CSP through token comparison. If the comparison is positive CSP

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

challenges data ownership, checks the eligibility of the user, and then issues a re-encryption key that can convert the encrypted DEK to a form that can only be decrypted by the eligible user.

3. **Data Deletion**: When the user deletes data from CSP, CSP firstly manages the records of duplicated users by removing the duplication record. If the rest records are not empty, the CSP will not delete the stored encrypted data, but block data access from the holder that requested data deletion. If the restof the records are empty, the encrypted data should be removed at CSP.
4. **Data Owner Management**: In case of a real data owner uploads the data later than the user, the CSP can manage to save the data encrypted by the real data owner at the cloud with the owner generated DEK and later on, CSP supports re-encryption of DEK at CSP for eligible users.
5. **Encrypted Data Update**: In case that DEK is updated by a data owner with DEK 0 and the new encrypted raw data is provided to CSP to replace old storage for the reason of achieving better security, CSP issues the new re-encrypted DEK 0 to all data users.

Our approach provides a secure approach to protect and deduplicate the data stored in the cloud by concealing plaintext from CSP. The security of the proposed scheme is ensured by AES symmetric encryption technique. Hardware environment is provided using the CPU having specification Intel Core 2 Quad Q9400 2.66 GHZ and having memory of 4GB SDRAM. Software environment includes the operating system Ubuntu 14.04, and Windows Home Ultimate editions 64 bits. Programming environment includes Java and Opennebula 5.02 version cloud platform.

In this approach, the operation time for the data encryption and decryption with different AES key sizes (128 bits, 192 bits, 256 bits) and different data size (from 10 megabytes to 600 megabytes). We observed that even when data is as big as 600 MB, the encryption/ decryption time is less than 12 seconds if applying 256-bit AES key. Applying symmetric encryption for data protection is a reasonable and practical choice. The time spent on AES encryption and decryption is increased with the size of data. This is inevitable in any encryption scheme. Since AES is very efficient on data encryption and decryption, thus it is practical to be applied for big data.

A. Proxy Re-encryption Evaluation: The efficiency of each operation of 1024-bit proxy re-encryption with different sizes of AES symmetric key(128 bits, 192 bits, 256 bits) is calculated. We observed that the operation time for PRE key pair generation (KeyGen), re-encryption key generation (ReKeyGen), encryption (Enc), re-encryption (ReEnc) and decryption (Dec) is not related to the length of an input key. For the tested three AES key sizes, the encryption time is less than 5 seconds. The approach does not introduce heavy processing load to data owners and holders. We also observe that the computation time of each operation does not vary too much with the different length of AES key size. In particular, the PRE related operation for deduplication are not influenced by the size of stored data.

IV. CONCLUSION AND FUTURE WORK

Huge cloud size and large number of end users cause the duplication of the files and so it is hard to manage the storage and it degrades the performance. Therefore deduplication of data needs to be done. Managing encrypted data is also important in practice for achieving a successful cloud storage service. In this paper, we propose a data ownership and proxy re-encryption method to deduplicate the encrypted file. Using only cloud service provider to achieve deduplication we propose a method to avoid the duplication of data encrypted using the AES encryption method.

REFERENCES

1. Zeng Yan, Wenxiu Ding, Xixun Yu, Haiqi Zhu, Robert H. Deng, "Deduplication on encrypted Big Data in Cloud", IEEE Transactions on Big Data, April 2016,pp.138-150.
2. Junbeom Hur, Dongyoung Koo, Youngioo Shin, Kyungate Kang, "Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage", IEEE Transaction on Knowledge and Data Engineering 2016.
3. Opendedup (2016).[Online]. Available: http://opendedup.org/
4. docs.microsoft.com, "Azure backup", 2016. [Online] Available: https://docs.microsoft.com/en-us/azure/backup/backup-introduction-toazure-backup.
5. aws.amazon.com, "Cloud Dedupliction On-Demand StorReduce an APN Technology Partner", 2016 [Online]. Available: http:// aws.amazon.com/blogs/apn/cloud-deduplication-on-demandstorereduce-an-apn-technology-partner/

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

6. Zhaocong Wen, JInman Luo, Huajun Chen, Jiaxiao Meng, XuanLi, Jin Li, "A Verifiable Data Deduplication Scheme in Cloud Computing," Intelligent Networking and Callaborative System(INCoS), 2014 International Conference, 2015.
7. Pasquale Puzio, Refik Molva, Melek Onen, Sergio Loureiro, "ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage," Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference, 2013.
8. Zheng Yan, Mingiun Wang, Yuxiang Li, Athanasios V. Vasilakos, "Encrypted Data Management with Deduplication in Cloud Computing," IEEE Cloud Computing 2016.
9. Xuexue Jin, Lingbo Wei, Mengke Yu, Nenghai Yu, Jinyuan Sun, "Anonymous deduplication of encrypted data proof of ownership in cloud storage," Communications in China(ICCC), 2013 IEEE/CIC International Conference, 2013.
10. Y. Wu, J. S. Pan, and C. F. Lin, "Improving accessing efficiency of cloud storage using deduplication and feedback schemes," IEEE Syst. J., vol. 8, no, 1, March 2014, pp.208-218.
11. Tin-Yu Wu, Wei-Tsang Lee, Chia Fan Lin, Cloud Storage Performance Enhancement by Realtime Feedback Control and De-duplication Wireless Telecommunication Symposium (WTS), 2012.
12. opennebula.org, "The Simplest Cloud Management Experience", 2016. [Online].
13. C. Fan, S. Y. Huang, and W. C. Hsu, "Hybrid data deduplication in cloud environment," in proc. Int. Conf. Inf. Secure. Intell. Control, 2012, pp.174 to 177, doi:101109/ISIC.2012.6449734.