Privacy Preserving Redundant Data Removal in Cloud Storage: RDRC

Srinivas Mudepalli, V. Srinivasa Rao, R. Kiran Kumar

Abstract*: With rapid increase and enlargement of the cloud storage technology, the users transmitting their data in the cloud storage. By uploading these data there are some issues in cloud storage, for example duplicate data, security problems, availability and vendor lock problem. To overcome these issues, we propose a redundant data removal in cloud storage: RDRC in this paper. RDRC is used to reduce the redundant data in cloud computing using the method data de-duplication and distributing the data among the cloud storage provider by data reference characteristics. To better utilize and overcome the security issue the data is encrypt with the blowfish algorithm. The performance and implementation is showing that the proposed methodology RDRC increases the performance and cost efficiency with existing schemes.

***Index Terms*: blow fish algorithm, cloud computing, data duplication, data consistency, Elliptic curve key**

I. INTRODUCTION

An advance cloud computing system platform is developed recently for the application of deploying, managing and presenting large scale data [1] [2]. At several larger scale many number of clients utilizes same framework, the services via internet on the basis of cloud computing framework is different client server architecture model. This is attracted because of the data is secured and accessed only in the form of encryption [3] [4]. By the rabid increase of cloud technology, the users uploading their data in to the cloud storage provider [5]. In recent many vendor lock problems are have been reported due to the cause of customers want a response with different service providers. By overloading the data in to the cloud storage the lack of fault tolerance in cloud service and monetary losses, heavy penalty may be caused. In 2013 this may cause failure in Microsoft Windows Azure servers and loss of millions of deterioration and dollars.

Usually, the private information of the users is not stored in cloud as the form of plain text even if law data privacy based is enforced because of the public nature [6]. Based on the improvement of cloud computing, cloud storage is described. For the stage of processing cloud large data, the cloud storage provides a desirable solution as a data storage and cloud computing management. Generally, cloud storage is classified into four layers which is user access layer, application interface layer, basic management layer and storage layer [7]. Basically, the cloud storage is a large data centre and giving users along with on-demand pay storage services with great flexibility, scalability and tolerance features. The cloud storage server may abuse the data management rights and cause exposed data to the threats for the security [8]. The Drop box faced a disconnection twice in 2014, this cause many problems in users so they maintained copy of their data in cloud storage.

In 2013 similar things happen in Nirvanix, so fault tolerance is the important feature for the efficient cloud storage provider. In cloud system the issues of security and maintenance of cost use the deduplication process [9]. Software bugs and operator errors are the example of application data from specific cloud storage provider. To secure the data the user encrypts the data with their own keys [10]. This may be result the identical data will be cipher text. By de-duplication the duplicated data cannot be eliminated therefore the encryption technique is utilized for data confidentiality. In a distributed cloud servers the data will be stored and files should be multiple files using encryption technique [11]. Data residing at the server is another important concern of the storage necessity. To overcome this concern one of the approach is deduplication process. These processes eliminate the redundant copies of data and stored only one unique data in the cloud storage [12]. Deduplication is used to identifying the similarity of data and removing redundant data in cloud storage [13].

II. RELATED WORK

Priyanka Singh *et al.* [16] had explained secure for the cloud based on deduplication concept. This paper mainly focuses on the secret sharing of data in a cloud. It provides secure data duplication scheme for the secret sharing of data in the cloud, it may be possible for data confidentiality, reliable key management and fault tolerance. Additionally, the secret sharing is based on CRT (Chinese Remainder Theorem), it should be minimized the key overhead. Yukun Zhou *et al.* [17] had examined Edeup (Encrypted deduplication) scheme to secure the cloud on the basis of flexible access control in the cloud. Target based duplicate-chunk checking and source-based similar-segment detection are combined by EDedup to guarantee deduplication efficiency and resist the attacks. Additionally, it manages the metadata by providing message-derived file keys for duplicate files and it reduces the metadata storage overheads.

MeixiaMiao *et al.* [18] had developed secure cloud deduplication. This paper is mainly focused on secure based on deduplication in cloud storage system. It allows the client to participate the deduplication and it secure the data of clients. Either than, this system shows the total cost of the cloud service provider, when the cost of client is decreased the cloud service provider profit will be increase.

Revised Manuscript Received on December 22, 2018.

Srinivas Mudepalli, Research Scholar, Department of CSE, Krishna University, Machilipatnam, Andhra Pradesh, INDIA

Dr. V. Srinivasa Rao[2], Professor, Department of CSE, V. R. Siddhartha Engineering College, Vijayawada, Andhra Pradesh, INDIA

Dr. R. Kiran Kumar[3], Assistant Professor, Department of CSE, Krishna University, Machilipatnam, Andhra Pradesh, INDIA.

Jinfeng Liu *et al.* [19] had introduced data deduplication on the basis of similarity in cloud data. This paper mainly focuses on to reduce the computation overhead in the deduplication files. This proposed scheme provides the efficiency and security for the deduplication files on the cloud.

Chao Yang *et al.* [20] had examined client side deduplication method. This paper is mainly focused on client side deduplication method. Additionally this paper introduced a key distribution scheme on the basis of proxy re-encryption. In client side knows nothing about the encryption key still this scheme helps for distributing the encryption file. The client can share the files with his encrypted key without establishing the secure channel. The server cannot can recover the private key of the client when the distribution phase

III. PROPOSED METHODLOGY OF RDRC

A novel proposed methodology has been introduced in this cloud storage for redundant data removal of cloud storage. The RDRC deleted the redundant data in cloud storage and administer the respective data over the several cloud storage modules.

A. The Architecture of RDRC

RDRC consists of four important function modules. They are data distribution, data deduplication, evaluation, performance & cost evaluation. In data deduplication technique is used for dividing the data block into multiple data blocks by Multi-level byte index checking and for securing the data, we encrypt the data after deduplication by blowfish algorithm. In data distribution allocates the data blocks corresponding to the cloud storage providers. From the enactment and cost of the cloud storage services are evaluated in the performance and cost evaluation.

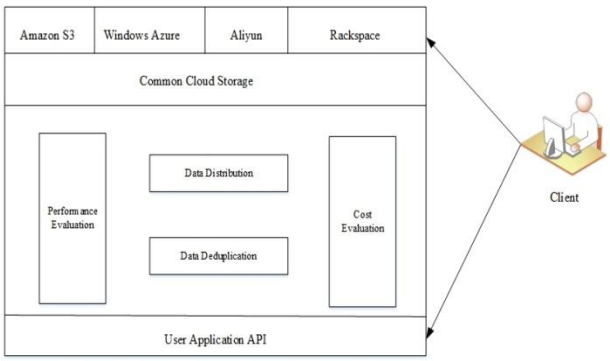

Figure 1: Architecture of RDRC

B. Data Deduplication

Data deduplication is used to eliminate the duplicated data or repeated data of files. First we data deduplication are done by multi-level byte index chunking then encrypt the data by using blowfish algorithm.

Multi-level byte index checking

According to the size of file the multi-level byte index checking is divided the file into multiple chunks. The deduplication process is done by 32kb size indexed table and 4mb chunk sized index table. The multi-level byte index chunking approaches consist of double index table for a file. 32kb and 4mb are the size of each chunk. In first level detects the large size identical data blocks by using 4mb chunk sized index-table. In the second level, the first level files similarity data block detection is find by using 32kb index table. This method gives accuracy data deduplication process. The working process is shown in Figure 2.

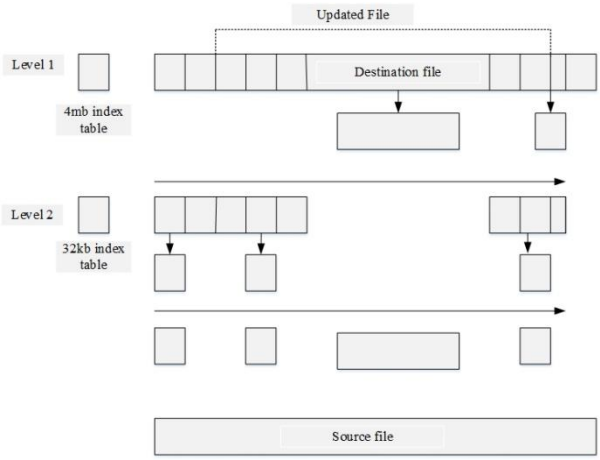

Figure 2: Multi-Level Byte Index Checking Blowfish algorithm

Key generation

The prime step for ECC is the base point which is given below,

$$E : y^2 = x^3 + ax + r \quad (1)$$

Here g and r are integers, satisfy the condition is $4g^3 + 27h \neq 0 \bmod p$; and the prime number is p_r and include a point called point at infinity. Below equation (2) is used to create a public key.

$$p = r^* c \qquad (2)$$

Where, the random number is r that to choose within the range between $1 to n-1$. c is the point on the curve, p is the public key and t is the private key.

Table 1: Blowfish Algorithm

Algorithm of Blowfish
Split a in to two 32-bit division : a_L, a_R
For $k = 1$ to 16;
$a_L = a_L XOR\ p_k$
$a_R = a_R XOR\ a_R$
swap a_L and a_R
Next k
Swap a_L and a_R (Undo the last swap)
$a_R = a_R XOR\ p_{17}$
$a_L = a_L XOR\ p_{18}$
Remerge a_L and a_R

B. Data Distribution

Replication method and eraser code technique are used for the data distribution process. Moreover, these techniques are accomplishing the reference characteristics in data deduplication. The eraser code based scheme is used to store the data blocks. The data blocks are written with eraser code scheme. Eraser code and replication code based plan is utilized in cloud to use the decent variety attributes of distributed storage suppliers. Eraser code gives better execution in storage productivity. In data distribution, appropriation circulates the data blocks relating to the cloud storage suppliers.

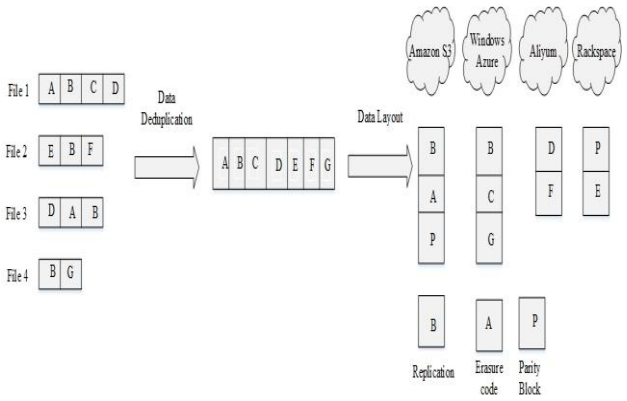

Figure 3: Data Distribution Using Reference Characteristics

C. Data Consistency

In RDRC design the data consistency mainly in three ways. In the cloud storage provider the write data must be accurately stored, in persisted storage the index data and metadata must be accurately stored and user read request must be fetch in the integrated data.

(i) In the cloud storage the write data stored, thus the RDRC utilize replica scheme used to distribute the data blocks. While utilizing the write request for includes many compose tasks and includes different distributed storage suppliers. When performing a write request for the RDRC ensure the data or information blocks and relating parity blocks are precisely written in the cloud storage suppliers. The write request is finished until all the operations are satisfied or else uncompleted write operation will be reperformed. (ii) RDRC stores the record information and metadata in the persisted storage to keep the loss of file information and metadata in case of intensity supply or framework crash. (iii) Multiple cloud storage provider the read request for information will stored by individual cloud storage provider. The requested data will be reproduced after every one of data blocks is fetched.

IV. RESULTS & DISCUSSION

This experiment is implemented on MATLAB tool installed in the system having the configuration of windows 7 with Matlab R2014a. The proposed work is implemented to redundant data removal in cloud storage. The experiment is carried out by existing cloud-of- cloud schemes called RACS, Dura Cloud and HyRD. The performance analysis with respect to average response time and cost evaluation. Our evaluations for three traces characteristics are mail, file server and web server.

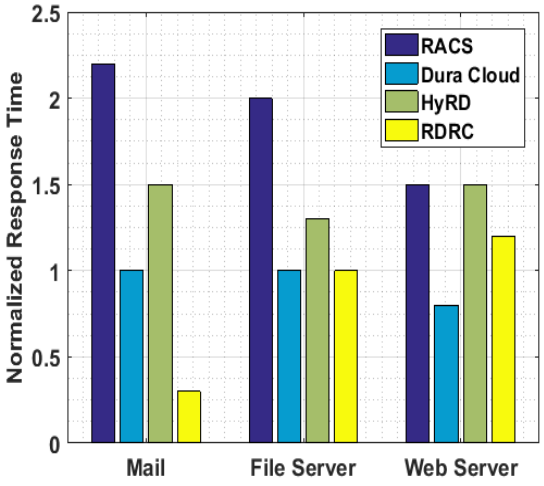

Figure 4: Normalized Response Time

Figure 4 shows the normalized response time for different cloud schemes. We can see that in graph RDRC is minimum average response time than other existing systems. The reason is that the RDRC is use replica scheme used to allocate the data blocks. While using the write request involves many write operations and involves several cloud storage providers. When performing a write request the RDRC make sure the parity blocks and data blocks are accurately written in the cloud storage providers. Thus the proposed methodology gives lowest response time comparing with other existing schemes.

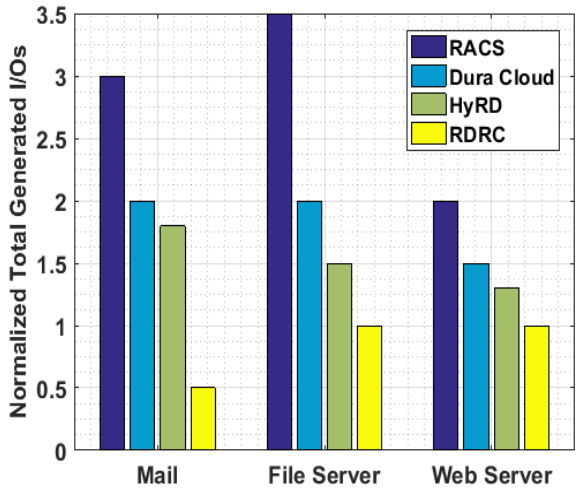

Figure.5. Normalized Total Generated I/O request

Figure 5 represent the normalized total generated I/O request. From figure we can see that RDRC is slightly increase the I/O request in the mail trace. The tree traces have great data redundancy, particularly the mail trace have high I/O request over the network. By using the eraser code and replication code scheme it will overcome the write amplification problem. Therefore the I/O request will be increase and reduce the average response time.

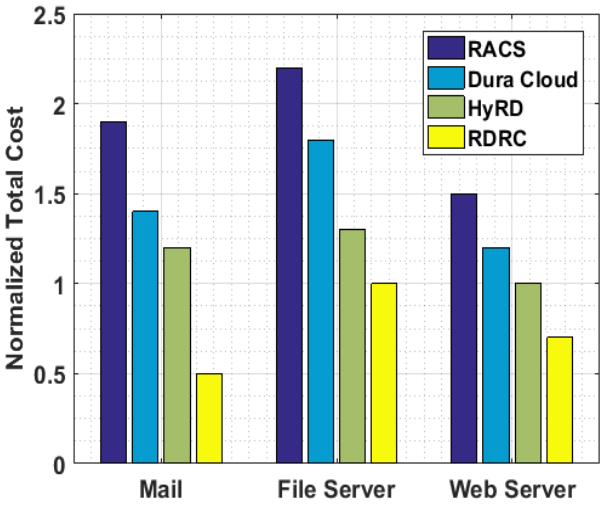

Figure 6: Normalized Total Cost

Figure 6 denoted the normalized total cost of proposed methodology and existing schemes. From figure we can see that total cost for our methodology is lowest comparing with other schemes. The reason of is that RDRC reduce the storage capacity and I/O requests. In figure we noted that RACS has the maximum total cost due to the increment of I/O request and data redundancy.

V. CONCLUSION

High enlargement of the cloud storage technology, the users uploading their data's in to the cloud computing. This will occur some issues in cloud storage. The issues like vendor lock problem, accessibility and security. To overcome these issues we propose a methodology RDRC for redundant data in cloud storage. This scheme used data deduplication process using multi-level byte index checking and to secure the data utilizing blowfish algorithm. Additionally for data consisting by using replication code scheme and eraser code scheme. This gives the better performance for our proposed methodology than other existing schemes.

REFERENCES

1. Sookhak, Mehdi, Abdullah Gani, Muhammad Khurram Khan, and RajkumarBuyya, "Dynamic remote data auditing for securing big data storage in cloud computing", Information Sciences,vol.380, pp. 101-116, 2017.
2. Wu, Suzhen, Kuan-Ching Li, Bo Mao, and Minghong Liao, "DAC: improving storage availability with deduplication-assisted cloud-of-clouds", Future Generation Computer Systems, vol.74, pp.190-198, 2017.
3. Rehman, Muhammad Habib, Victor Chang, Aisha Batool, and Teh Ying Wah. "Big data reduction framework for value creation in sustainable enterprises." International Journal of Information Management, vol.36, no. 6, pp.917-928, 2016.
4. Akhila, K., Amal Ganesh, and C. Sunitha, "A study on deduplication techniques over encrypted data",Procedia Computer Science,vol.87, pp.38-43, 2016.
5. Widodo, Ryan NS, Hyotaek Lim, and Mohammed Atiquzzaman, "SDM: Smart deduplication for mobile cloud storage" Future Generation Computer Systems, vol.70, pp.64-73, 2017.
6. Xu, Jiwei, Wenbo Zhang, Zhenyu Zhang, Tao Wang, and Tao Huang, "Clustering-based acceleration for virtual machine image deduplication in the cloud environment" Journal of Systems and Software,vol.121, pp.144-156, 2016.
7. Li, X., Li, J. and Huang, F, "A secure cloud storage system supporting privacy-preserving fuzzy deduplication", Soft Computing, vol.20, no.4, pp.1437-1448, 2016.
8. Shin, Youngjoo, Dongyoung Koo, JunbeomHur, and Joobeom Yun, "Secure proof of storage with deduplication for cloud storage systems", Multimedia Tools and Applications, vol.76, no. 19, pp.19363-19378, 2017.
9. Venish, A., and K. Siva Sankar. "Study of chunking algorithm in data deduplication," In Proceedings of the International Conference on Soft Computing Systems, pp. 13-20. Springer, New Delhi, 2016.
10. Kwon, Hyunsoo, Changhee Hahn, Daeyoung Kim, and JunbeomHur, "Secure deduplication for multimedia data with user revocation in cloud storage." Multimedia Tools and Applications, vol.76, no. 4, 5889-5903, 2017.
11. Yan, Zheng, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng. "Deduplication on encrypted big data in cloud." IEEE transactions on big data, vol.2, no. 2, pp.138-150, 2016.
12. Hur, Junbeom, Dongyoung Koo, Youngjoo Shin, and Kyungtae Kang. "Secure data deduplication with dynamic ownership management in cloud storage." IEEE Transactions on knowledge and data engineering, vol.28, no. 11, pp.3113-3125, 2016.
13. Li, Jingwei, Jin Li, DongqingXie, and Zhang Cai. "Secure auditing and deduplicating data in cloud." IEEE Transactions on Computers, vol.65, no. 8, pp.2386-2396, 2016.
14. Yan, Zheng, Mingjun Wang, Yuxiang Li, and Athanasios V. Vasilakos. "Encrypted data management with deduplication in cloud computing." IEEE Cloud Computing, vol.3, no. 2, pp. 28-35, 2016.
15. Xia, Wen, Hong Jiang, Dan Feng, Fred Douglis, Philip Shilane, Yu Hua, Min Fu,Yucheng Zhang, and Yukun Zhou, "A comprehensive study of the past, present, and future of data deduplication." Proceedings of the IEEE, vol.104, no. 9, pp.1681-1710, 2016.
16. Singh, Priyanka, Nishant Agarwal, and Balasubramanian Raman, "Secure data deduplication using secret sharing schemes over cloud", Future Generation Computer Systems, vol.88, pp.156-167, 2018.
17. Zhou, Yukun, Dan Feng, Yu Hua, Wen Xia, Min Fu, Fangting Huang, and Yucheng Zhang, "A similarity-aware encrypted deduplication scheme with flexible access control in the cloud", Future Generation Computer Systems, vol.84, pp.177-189, 2018.
18. Miao, Meixia, Tao Jiang, and Ilsun You, "Payment-based incentive mechanism for secure cloud deduplication." International Journal of Information Management, vol.35, no. 3, pp.379-386, 2015.
19. JinfengLiu, JianfengWang and XiaolingTao, "Secure similarity-based cloud data deduplication in Ubiquitous city", Pervasive and Mobile Computing, vol.41, pp. 231-242, 2017.
20. ChaoYang, MingyueZhang and QiJiang, "Zero knowledge based client side deduplication for encrypted files of secure cloud storage in smart cities", Pervasive and Mobile Computing, vol. 41, pp.243-258, 2017.