International Journal of Future Generation Communication and Networking
Vol. 13, No. 1, 2020 pp.1594-1600

Secure Deduplication of Encrypted Data in Cloud

Dhore M L [#1], Varpe K M[*2], Dhore R M[#3]

[#]*Computer Engineering, Professor, Vishwakarma Institute of Technology, Pune, SPPU, India*

[*]*Computer Engineering, Asst. Prof., Vishwakarma Institute of Technology, Pune, SPPU, India*

[#]*Software Developer, Siemens Technology and Services Private Limited, Pune, India*

Abstract

Cloud computing platform supports high amount of storage space to store data and provides access to data from remote machines. For data storage and access for data on cloud, the user have to pay the amount on rental basis. As data is stored on remote machines, it must have a facility to avoid the duplicate data; if it is not done the user have to pay additional amount for extra storage occupied by duplicate files. Data deduplication is a mechanism used to avoid data duplication. This is one of the best data compression techniques being used for removing the duplicate copies of repeated data in recent cloud storage. As this data deduplication technique is used to avoid the duplicate files not to be stored on cloud, it also needs to protect the data confidentiality. The private cloud server generates the privileges for the user as read, write, update, modify and delete. If the user gets the access privileges then only he can operate on data through a token id received from the cloud service provider which is unique. The proposed system provides a solution for preserving the data in cloud with the deduplication and also uses a data deduplication load balancing technique for distributing workload across multiple computing resources which includes cluster of computers and network links. In order to protect the confidentiality of sensitive data, AES encryption technique is used. The objective of work is to save space as well as bandwidth by using deduplication.

Keywords— *Cloud Computing, Deduplication, Secure Storage, Encryption, Decryption, Load Balancing.*

I. INTRODUCTION

Nowadays there is tremendous growth in information. Cloud storage providers provides infinite amount of storage space for their users and also provides different methods to save the space using compression methods. Users can access the information according to their needs and most of the users access the same information again and again. In a current scenario, the cost of data computation, application hosting, data storage and delivery is reduced significantly. The cloud storage makes it possible for users to access their data from anywhere and anytime. The data that we store on cloud may be personal, private or secret data. Cloud computing provides a data compression technique called data deduplication that splits data into chunks for removing duplicate copies of data. As this technique improves the storage space in turn it reduces the access time during download and saves the bandwidth too. Proposed system uses AES and RSA algorithms for uploading and sharing files respectively. AES to encrypt the data in the files before storing onto the cloud and RSA algorithm is used at file sharing side. Cloud storage is always large and store the files but storing duplicate copies of files just having different file name but data content are same occupies the extra storage space which is not required. In this paper authors have also proposed a method called load balancing which is used to distribute workload across multiple computing resources from security point of view when user downloads the data. Load balancing system splits the data in files into the chunks and stores into the different locations such as cluster of computers, or network links [1]. In order to balance the load across several nodes, the data need to be migrated across the multiple nodes. The key purpose of this work is to propose a dynamic load balancing algorithm based on data deduplication to balance the load across the storage nodes. Cloud computing have mainly four characteristics. The first one is on-demand self-service. In this case, the user avail services from the cloud service provider by making request and pays for the usage made. Second characteristic is network access in which user gets burly connectivity to all said resources regularly. A third characteristic is about pooling the resources from cloud infrastructure. In this case they provide multiuser services to multiple clients along with

ISSN: 2233-7857 IJFGCN

scalable services. Last characteristic is measuring usages of services from accounting point of view as well as many other monitoring purposes. Similar to characteristics, cloud computing has mainly five security issues. Following are the security issues in cloud computing. First is user authentication. Authentication must match the user's credentials in a database of authorized users. The unique authentication key is provided by service provider to the customer. Customer interacts with cloud using the authentication key provided only. Second is confidentiality. It means to protect the personal information during the transaction between service provider and customer.. Third is data integrity. It indicates that the information should be altered due to transmission errors during transaction. Fourth is availability. It assures to provide services to all legal users from time to time as well as whenever required. Fifth one is regarding security. During any transaction it identifies malicious customers to avoid the obstruction.

II. OVERVIEW OF DATA DEDUPLICATION

Data deduplication is a lossless technique used to minimize network bandwidth by saving storage space by detecting the data redundancy. In data compression, file is compressed by using either lossless or lossy compression technique by looking consecutive series of either zeros or ones while in data-duplication it looks on the similarity of content within a file at file level, block level or byte level.

A. File Level Deduplication

This method treats file as a single entity. Whole file is considered as a single instance and the hash value for whole file is generated in File Level Deduplication. It is used to find similarities between two files and hence the file level de-duplication.

B. Block Level Deduplication

Block Level Deduplication divides the single file into multiple data blocks and hence exploits data redundancy at block level. File can be divided either into smaller fixed-size blocks or variable-size blocks. It identifies and removes similar contents among the data blocks.

C. File Level Deduplication

This technique finds the similarity in the data chunks at the byte level and hence called byte level data deduplication. Byte Level Deduplication compares data chunks byte by byte and finds similar contents. This method requires many more input-output operations compared to earlier two methods.

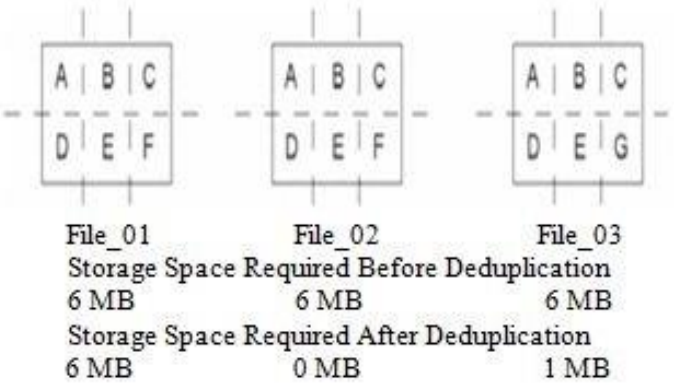

Fig. 1 Example of data deduplication process

The process of deduplication is depicted in figure 1. There are three files where two files have same contents while third file is bit different. Without deduplication, it occupies 18 MB while after deduplication it requires only 7 MB storage space.

III. LITERATURE REVIEW

In 2013, Chun-Ho Ng et.al proposed RevDedup algorithm to find and remove duplicates from virtual machine images. Whenever new VM image comes, the RevDedup find the similarity with old data and removes it from old data [2]. In the same year, Mihir Bellare et.al proposed a cryptographic approach called Message-Locked Encryption (MLE). In MLE the keys used for encryption and

decryption are derived from message itself. It was the secured way to carry out deduplication [3]. In 2014, Zhou Lei et.al proposed a mechanism using fixed size block method to store images. This method calculates compact digest called fingerprint for each image file and hence make the directory of fingerprints. For new image input it calculates fingerprint and compares with available fingerprint library [4]. In the same year, Waraporn Leesakul et.al proposed a new scheme to improve efficiency of cloud storage space using dynamic data deduplication. This scheme improved storage space along with maintaining the redundancy [5]. In the same year, Issa M. Khalil et.al identified 28 cloud security issues through his survey on security issues in clouds and security solutions [6]. In 2015, N. Jayapandian et.al proposed the scheme based on authorization. This system has the feature to protect user data confidentiality using differential privileges based on duplicate check [7]. In the same year, Mi Wen et.al developed a scheme using convergent encryption technique for secure deduplication scheme [8]. In the same year, Lakshmi Pritha et.al developed a system using RSS key to provide secure access to cloud resources and demonstrated ALG technique for data deduplication [9]. In the same year, Chun-I Fan et.al proposed check block mechanism for encrypted data deduplication [10]. In the same year, Mr. Dame Tirumala Babu et.al presented a method for data deduplication based on authorization to secure data [11]. In 2016, Shuai Wang et.al proposed a RRMFS file system to support data deduplication [12]. In the same year, Zheng Yan et.al presented a scheme for ownership and re-encryption to deduplicate encrypted data stored in cloud [13]. In the same year, Naresh Kumar et.al performed a comparative analysis of various deduplication techniques is done using destor tool. Data deduplication technique uses several chunking algorithms fixed length and variable length chunking [14]. In the same year, Jun Ren et.al proposed a method based on differential privacy for secure data deduplication [15]. In the same year, Saurabh Singh et.al provided cloud security survey with discussion about security issues and challenges [16]. In the same year, Feilong Tang et.al introduced an approach called Load Balanced Flow Scheduling approach for dynamic load balancing and to maximize network throughput [17]. In 2017, Danoing Li et.al proposed a method called CSPD using modified DCT-based Perceptual Image Hash (D-phash) to improve the accuracy of the duplicate check [18]. In the same year, Hui Cui et.al implemented an ABE encryption system for cloud storage based on attributes [19]. In the same year, Rayan Dasoriya et.al presented a dynamic load balancing algorithm to distribute the load across multiple connected network links [20]. In the same year, Shunrong Jiang et.al proposed data confidentiality and ownership management system for data deduplication based on Proof of Ownership (PoW) technique [21]. In the same year, Himshai Kambo et.al implemented a secure deduplication mechanism based on CDC and MD5 algorithm. CDC used to break the data streams using randomization and MD5 algorithm creates the hash values for the segments or chunks created by CDC. It was used to improve network bandwidth [22].

IV. ANALYSIS OF EXISTING SYSTEMS

There are systems developed to provide efficient storage space and security for deduplication process. Most of the previous deduplication system only provides storage space on the server, but not avoid duplicate data. The deduplication system should provide reliability comparable to other available systems. Existing system have the major challenge about privacy of users sensitive data outsourced to cloud. Most of existing systems have provided confidentiality of user's data before outsourcing data on data center. Commercial service providers are reluctant to use encryption technique as it makes the deduplication process more complex.

Traditional system only provides data deduplication but there is less security for file upload, download and delete. Mainly three disadvantages are observed. First one is deduplication checked only with file name but not file content. Second is to maintain high data reliability in deduplication system is a critical one. Last is system only provides data deduplication but does not provide load balancing technique.

V. SYSTEM ARCHITECTURE AND IMPLEMENTATION

Proposed system mainly focuses on data deduplication and load balancing concepts. Data deduplication is the data compression technique that avoids duplicate copies of data from storage server and store only one unique copy of data.

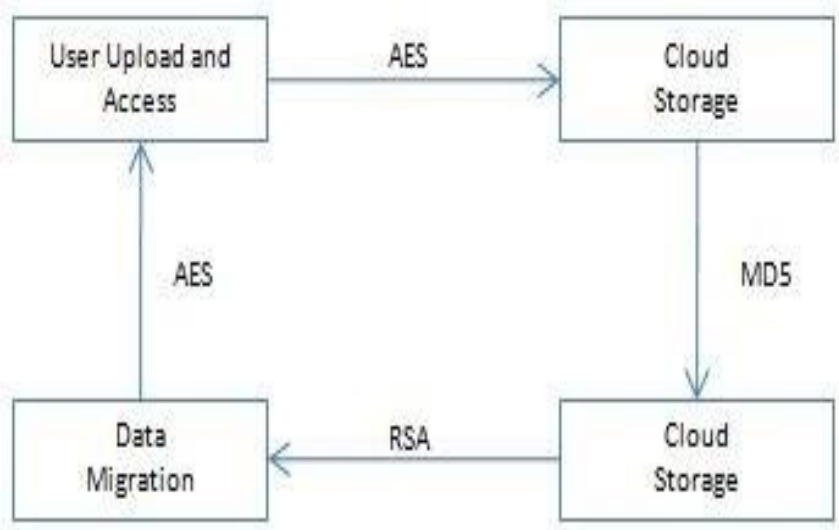

Fig. 2 System architecture

Load balancing is a method of distributing workload across multiple resources such as network link, cluster of computers etc. The goal of this system is to save storage space and preserve privacy of data holders by proposing a scheme to manage encrypted data with deduplication.

The overall architecture of system and its functional modules are depicted in figure 2. The four functional modules are user upload and access, cloud storage, data deduplication and data migration respectively. First module is used to upload as well as to access data to and from cloud. For upload user can select any file such as .docx, .txt, .pdf, .mp3, .mp4, .jpeg etc. and, encrypts it using AES encryption before upload on cloud storage.

After receiving file from user, cloud storage server calculates hash value of that file using MD5 algorithm by considering the contents in file and file name. The same module calculates the two hash values for each file as file name plus contents in file and excluding file name. It is done because most of the cases file names and contents are same but sometimes contents may be same but file name is different.

The third module is data deduplication. This module does the function of matching the file hash values of newly uploaded files with the file already exists on cloud server for the respective user only. Data deduplication process removes the repeating copies of same file and saves only one copy of that file on the basis of hash value. The major contribution of our work is to provide principle security to sensitive data from various kinds of attacks by providing data migration that is by changing the location of data with the user defined intervals. It is implemented with the help of thread time. When user wants to access data, he needs authentication key which is sent to his via mail. After authentication data is accessed using AES encryption.

VI. EXPERIMENTAL RESULTS

Storage space used before and after deduplication as well as time required for file uploading and downloading is shown in Table I. Experimental results shows that our duplicate check and load balancing scheme maximizes throughput, minimize response time, saves network bandwidth and storage space.

Table I Results

No. of Files	Before deduplication storage space used (MB)	After deduplication storage space used (MB)	File Upload Time (sec)	File Download Time (sec)
46	305	300	80	25
24	109	104.16	17	10
3	490	465.15	112	45

51	271	266.44	73	24
15	215	213.22	37	21
1346	135	109.58	35	12
109	397	390.42	98	37

Storage space before and after data deduplication is depicted in figure 4.

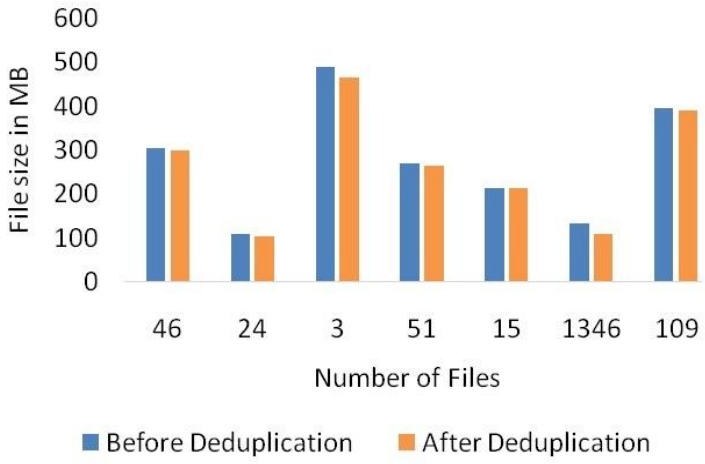

Fig. 3 Before and after data deduplication

Upload time before data deduplication and download time after data deduplication is depicted in figure 4.

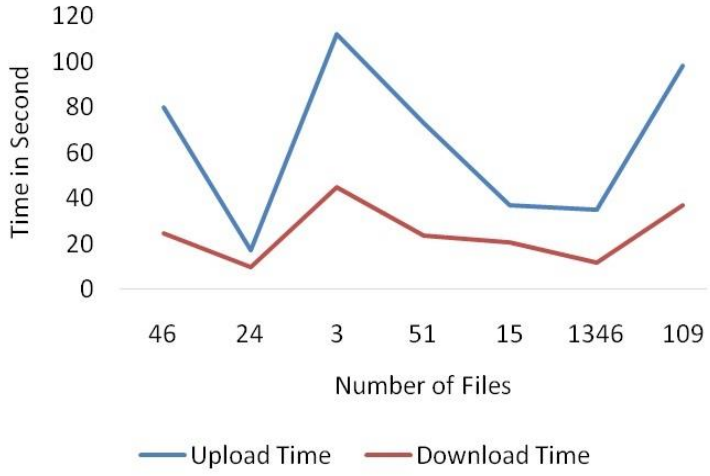

Fig. 4 Download time saving after deduplication

Experimental results shows that our duplicate check and load balancing scheme maximizes throughput, minimize response time, saves network bandwidth and storage space.

VII. CONCLUSION AND FUTURE SCOPE

Authors proposed the secured data deduplication system to maintain the confidentiality as well as load balancing is achieved by providing thread time to manage the load on server. We implemented our data deduplication system by matching the content of file using MD5 algorithm and security is provided for users outsourced data by encrypting the file using AES algorithm. The major

contribution of this work is principle security to sensitive data from various kinds of attacks by providing data migration that is by changing the location of data with the user defined intervals. Overall this system saves network bandwidth, storage space and provides the efficient data deduplication.

In future, we can extend our framework to implement various encryption algorithm and better load balancing technique to improve the security and also to implement in real time audio and video deduplication storage systems.

ACKNOWLEDGMENT

The authors wish thanks to all the referees involved in the above mentioned work of review.

REFERENCES

[1] Shyam Patidar, Dheeraj Rane, Pradesh Jain, A Survey Paper on Cloud Computing, Proceeding ACCT '12 Proceedings of the 2012 Second International Conference on Advanced Computing & Communication Technologies, pp 394-398, January 07 - 08, 2012.

[2] Chun-Ho Ng, Patrick P. C. Lee, RevDedup: A Reverse Deduplication Storage System Optimized for Reedstop Latest Backups, Proceeding APSys '13 Proceedings of the 4th Asia-Pacific Workshop on Systems, Article No. 15, Singapore, July 29 - 30, 2013.

[3] Mihir Bellare, Sriram Keelveedhi, and Thomas Ristenpart, Message-Locked Encryption and Secure Deduplication, Annual International Conference on the Theory and Applications of Cryptographic Techniques, EUROCRYPT 2013: Advances in Cryptology – EUROCRYPT, Lecture Notes in Computer Science, vol 7881, Springer, Berlin, Heidelberg, pp 296-312, 2013.

[4] Zhou Lei, Zhao Xin Li, Yu Lei, Yan Ling Bi, Luokai Hu, Wenfeng Shen, An Improved Image File Storage Method Using Data Deduplication, TrustCom 2014, The 13th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, Beijing, China, pp 638-643, 24-26 September 2014.

[5] Waraporn Leesakul, Paul Townend and Jie Xu, Dynamic Data Deduplication in Cloud Storage, SOSE 2014, IEEE Eighth International Symposium On Service-Oriented System Engineering Oxford, United Kingdom, pp. 7-11 April 2014.

[6] Issa M. Khalil, Abdallah Khreishah and Muhammad Azeem, Cloud Computing Security: A Survey, Article in 'Computers', Open Access Journal, Vol and Issue 3(1), pp. 1-35, 3 February 2014.

[7] N.Jayapandian, Dr A M J Md Zubair Rahman and I.Nandhini, A Novel Approach for Handling Sensitive Data with Deduplication Method in Hybrid Cloud, Online International Conference on Green Engineering and Technologies, November 2015.

[8] Mi Wen, Kejie Lu, Jingsheng Lei, Fengyong Li, Jing Li, BDO-SD: An Efficient Scheme for Big Data Outsourcing with Secure Deduplication, the Third International Workshop on Security and Privacy in Big Data, IEEE 2015.

[9] N. Lakshmi Pritha and N.Velmurugan, Deduplication Based Storage and Retrieval of Data from Cloud Environment in International Conference on Innovation Information in Computing Technologies, Chennai, pp. 1-6, IEEE 2015.

[10] Chun-I Fan and Shi-Yuan Huang, Encrypted Data Deduplication in Cloud Storage, Article in 'ASIAJCIS' 15 Proceedings of the 2015 10th Asia Joint Conference on Information Security, pp.18-25, May 24-26, 2015, IEEE Computer Society, Washington, ISBN: 978-1-4799-1989-5.

[11] Dama Tirumala Babu and Yaddala Srinivasulu, A Survey on Secure Authorized Deduplication Systems, International Research Journal of Engineering and Technology. Volume: 02 Issue: 05. Aug-2015.

[12] Shuai Wang and Jianhai Du A Storage Solution for Multimedia Files to Support Data Deduplication, 2016 2nd International Conference on Cloud Computing and Internet of Things, Dalian, China, pp-78-8, 2016.

[13] Zheng Yan and Wenxiu Ding, Deduplication on Encrypted Big Data in Cloud, IEEE Transactions On Big Data, Vol. 2, No. 2, April-June, 2016.

[14] Naresh Kumar, Preeti Malik, Sonam Bhardwaj, Sushil Chandra Jain, Comparative Analysis of Deduplication Techniques for Enhancing Storage Space, 4th International Conference on Parallel, Distributed and Grid Computing. IEEE, 2016.
[15] Jun Ren and Zhiqiang Yao, A Secure data deduplication scheme based on differential privacy, IEEE 22nd International Conference on Parallel and Distributed System, pp-1241-1246, 2016.
[16] Saurabh Singh and Young-Sik Jeong, A Survey on Cloud Computing Security: Issues, Threats, and Solutions, in Journal of Network and Computer Applications, pp-1-30, 2016.
[17] Feilong Tang and Laurence T. Yang, A Dynamical and Load-Balanced Flow Scheduling Approach for Big Data Centers in Clouds, IEEE Transactions On Cloud Computing 2016.
[18] Danping Li, Chao Yang, Chengzhou Li, Qi Jiang, Xiaofeng Chen, Jianfeng Ma, and Jian Ren, A Client-based Secure Deduplication of Multimedia Data, Communication and Information Systems Security Symposium. IEEE, 2017.
[19] Hui Cui, Robert H. Deng, Yingjiu Li, and Guowei Wu, Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud, IEEE Transactions on Cloud computing, year: 2017.
[20] Mr. Rayan Dasoriya, Ms. Purvi Kotadiya, Ms. Garima Arya, Mr. Priyanshu Nayak, Dynamic Load Balancing in Cloud: A Data-Centric Approach, International Conference on Networks & Advances in Computational Technologies. IEEE, 2017.
[21] Shunrong Jiang, Tao Jiang and Liangmin Wang, Secure and Efficient Cloud Data Deduplication with Ownership Management, IEEE Transactions on Services Computing. IEEE, 2017.
[22] Himshai Kambo, Bharati Sinha, Secure Data Deduplication Mechanism based on Rabin CDC and MD5 in Cloud Computing Environment, 2nd IEEE International Conference on Recent Trends in Electronics Information & Communication Technology (RTEICT).Bangalore, pp 400-404, May 19-20, 2017, India.