ISSN 2395-1621

Data Duplication Avoidance and Cloud Computing Security Using AES and SHA Algorithm

[#1]Swati O. Rathod, [#2]Dr.V.S.Gulhane, [#3]Dr.A.P.Jadhao

DEPARTMENT OF COMPUTER ENGINEERING

Dr. Rajendra Gode Institute of Technology & Research Amravati.

ABSTRACT

Secure data deduplication can significantly reduce the communication and storage overheads on server side services, and has potential applications in our big data-driven society. Existing data deduplication schemes are generally designed to the mobile flash storage application ensure the efficiency and data availability, but not both conditions. We are also not aware of any existing scheme that achieves accountability, in the sense of reducing duplicate information disclosure. In this system, we investigate proposed architecture, and propose an efficient and privacy-preserving big data deduplication in server side storage. Proposed structure achieves both privacy-preserving (AES encryption algorithm) and data availability. In addition, we take backup and recovery with accountability into consideration to offer better privacy assurances than existing schemes.

ARTICLE INFO

Article History

Received: 2nd June 2023

Received in revised form :

2nd June 2023

Accepted: 5th June 2023

Published online :

6th June 2023

Keywords: AES, SHA, Data Privacy, Data Encryption, Deduplication Analysis.

I. INTRODUCTION

Our propose technique provides data security using data encryption in cloud environment. For effective usage of storage space, we provide de-duplication check at file level. We also provide new deduplication constructions supporting authorized duplicate check in cloud architecture, in which the duplicate-check is done at local cloud server. This avoids multiple transaction of file tag over network while checking de duplication. We introduce a relative addressing method in which data will check at entry level when user uploading phases. Nowadays, cloud computing is very important in the Information Technology. Cloud computing enables access to a shared pool of configurable computing resources like servers, storage and applications, etc. The storage services provided to users are though internet. There is chances of cloud disaster like problem in connection, performance, privacy & security, data management. To solve connection problem we can implement offline storage & sync mechanism. To improve performance, load balancing is being an important task for doing operations in cloud and so as de-duplication also. As cloud computing has been growing and many clients all over the world are demanding more services and better results, so load balancing is necessary. Load balancing assure efficient resource utilization to customers on their demand and build up the overall performance of cloud. Every increasing volume of back up data in cloud storage may be a vital challenge so we can use de-duplication mechanism for eliminating the duplicate data. Many algorithms have been developed for allocating client's requests to available remote nodes. The key idea behind this project is to develop a offline store & sync mechanism, dynamic load balancing algorithm based on de-duplication to balance the load across the storage nodes during the expansion of private cloud storage. Data is the key factor in the modern era. Starting from Food Items, Groceries and till the high end satellite and rocket mechanisms, datas play an important role. A separate study is being implemented in Grade Schools and Institutions on data analysis so as to cater to the ever growing era of data science and its applications. More than 2.5*10^18 bytes of information are produced consistently. More than 90% of the above information has been made just in the recent years alone. This number will arrive at 35 ZB in 2025, which has demonstrated to be too bigger than imagination and control. Data storage mediums ranged during the earlier days were selector n tube, Punched card , Punched tapes , Drum memory and then the IBM HDD. Depending on the type of data storage medium, the data storage type also changed and this digital era, we store data in the server with the specialized keys for security concerns. Use of cloud computing is increasing. Cloud optimization is increasing. Effective use of cloud resources is the want of this time, as redundant datas are stored in the cloud again and again. This causes inefficient data storage in cloud and also affects the upload bandwidth. Data security is the major criteria while accessing data from and to the server. We need to reduce the load on the server or cloud storage so as to make it free and perform hassle free data transactions. Removal of duplicate data from the cloud and provide and access to the files will be the prime address issue of this paper implementation. The paper aims to free space, bandwidth and storage in cloud. The suggested approach is to remove the redundant data, where every user has been assigned some access according to the duplication check & each user have their priority token. Hybrid cloud organization is deployed to accomplish the deduplication in cloud. Cloud processing gives versatile, minimal effort and area free administrations over the web. The administrations gave ranges from basic reinforcement administrations to distributed storage foundations. The quick development of information volumes has enormously expanded the interest for systems for space and transmission capacity [1]. Distributed storage administrations like Drop box, Google drive pick a deduplication procedure where the cloud server stores just a solitary duplicate of repetitive information and makes connects to the duplicate as opposed to putting away genuine duplicates. The security of clients' information turns into another test. Consequently the clients encode the information before redistributing to the cloud. During that process we will be facing a problem of duplication. To solve that we have to perform data deduplication. There is a huge increment in the measure of information produced every day and in 2020 it is normal 44 zettabytes of information will be delivered. But storing and managing these large amounts of data is really a difficult task. Cloud computing offers a new way of service provisions by rearranging the resources over the internet. Cloud storage is the most popular among all the storage providers as cloud storage is the most efficient one. Data duplication occurs when the same data is being shared to the cloud storage by multiple users [2]. Data de-duplication keeps only one physical copy and eliminates multiple data copies. Through this consumption of resources will be reduced and saves the disk space and network bandwidth. Cloud user's upload their information. Security and privacy are the major issues though data deduplication promises lots of benefits. Data needs to be encrypted and store in the cloud which ensures security and user privacy[3].Let us consider three users user1, user2, user3 they are uploading some amount of data through this we will get an idea how deduplication results in.

User1 uploads -----→ a, b, c

User2 uploads -----→d, a, b

User3 uploads -----→d, c, a

The similar kind of data has been uploaded which reduces storage and efficiency of resources [4]. By de-duplication only the files "a, b, c, d" will only be stored into the cloud. This eliminates the repeated data and only stores first unique instance of any data. Whenever the user tries to store the data which is already present in the stored in the cloud it only creates a pointer to the existing one rather than creating redundant data. In block level for each file or chunk a unique hash value will be generated using hashing algorithms like SHA. Whenever user needs to upload a file a hash value will be generated foe that and it will be compared with the existing hash values. If the hash value is not present then the hash value and the file will be stored else it will not store. But in this we will be facing a problem whenever the hash algorithm produces same hash values for different chunks of data then collision occurs. This hash collision leads to data loss [5].

II. LITERATURE SURVEY

Bhavesh R. Birhade, Swati S. Shinde, [1]2021, have introduced SRRS system which comprise of convergent algorithm to maintain data confidentiality and used role reencryption algorithm to accomplish authorized data deduplication effectively. Management center is introduced to manage keys and user's roles. With the introduction of management center in the system, computational cost and overhead gets reduced on the client side. The SRRS system performs data deduplication and reduces storage space requirement and bandwidth consumption.

Bhavesh R. Birhade, Swati S. Shinde, [2]2021, have proposed novel Attribute-Based Storage system which supports secure and efficient de duplication. It also explained drawback of standard Attribute-based encryption technique which does not support secure de duplication. The system works on hybrid cloud environment where private cloud is in charge of identical copies detection and public cloud opts for managing storage.

Bhavesh R. Birhade, Swati S. Shinde, [3]2021, the author has introduced scheme which uses Symmetric Encryption algorithm, Hashing technique, Convergent encryption algorithm and token generation scheme to provide authorized duplication of data. Here the user data confidentiality and security is maintained. The data is protected both form passive and active attacks.

Bhavesh R. Birhade, Swati S. Shinde, [4]2021, has surveyed various methodologies and technologies for implementing data deduplication. They have also shown comparison of various technologies. The data confidentiality is compromised at different extent while performing data deduplication is depicted in the paper.

GDup: De-duplication of Scholarly Communication Big Graphs, 2018, Claudio Atzori, Paolo Manghi, Alessia Bardi, In this paper, author propose the GDup system, an Data Duplication Avoidance and Cloud Computing Security Using AES and SHA Algorithm Dr. Rajendra Gode Instiude of Technology & Research, Amravati Page 6 integrated, scalable, general-purpose system for entity deduplication over big information graphs.[5]

Cost-Based and Effective Human-Machine Based Data Deduplication Model in Entity Reconciliatio 2018 Charles R. Haruna, MengShu Hou, Moses J. Eghan, In this paper, a hybrid human-machine system was proposed where machines were firstly used on the data set before the humans were further used to identify potential duplicates.[6]

An Online Data Deduplication Approach for Virtual Machine Clusters2018 Zhongwen Qian, Xudong Zhang, Xiaoming Ju, Bo Li However, due to the heavyweight nature of virtual machine technology, a large amount of space is consumed when taking snapshot of VMC. To address the above issues, we propose an online deduplication mechanism which aims at improving storage efficiency without sacrificing the performance of VMC.[7]

An improved small file storage strategy in Ceph File System 2018, Ya Fan,Yong Wang, Miao Ye, the proposed scheme aims to achieve a better trade-off among the utilization of space of hard-disk and bandwidth resources, file access time, hard-disk I/Oas well as the cluster performance in Ceph FS by eliminating duplicate copies of repeating data, merging similar small files, and introducing the cache module.[8] . RCDSD: RSA based Cross Domain Secure Deduplication on Cloud Storage 2018, Shivansh Mishra, Surjit Singh, Syed Taqi Al, In this paper author propose a scheme RSA based Cross Domain Secure Deduplication (RCDSD), of coordination between distributed storage managers without revealing too much information about the actual data stored by the clients[9].

Junbeom Hur, Dongyoung Koo, Youngjoo Shin, and Kyungtae Kang [10] 2016, presented reduction of replicas withdifferent quality constraints by using deduplication scheme to get the good performance and efficiency. The performance has been done with dynamic ownership management. They used to minimize the cost and bandwidth. The experimental results are based on the ownership management.

Data Duplication Avoidance and Cloud Computing Security Using AES and SHA Algorithm Dr. Rajendra Gode Institude of Technology & Research, Amravati Page 7 Pasquale Puzio, RefikMolva and MelekOnen [11] 2016, publish additional encryption operation and access control mechanism. Their goal is to get security and privacy challenges. They propose Cloudedup to handle different constraints. They used to reduce the storage space and save the storage space. The result of this experiment had being partially succeeded.

Dimitrios vasilopoulos and Melek Onen [12] 2016, they presented proof of retrievability with MLE. Hence the data used is identical. It is performed on the setup phase with uploaded material. It introduces new encoding algorithm ML encode. But it fails because of current POR.

Pyla. Naresh, K. Ravindra, Dr. A. Chandra Sekhar [13] 2016, they deal with the danger of data stockpiling the data security as well as data integrity and data deduplication on Cloud. They proposed framework of D-Cloud.It create hash estimate before transferring, auditing, integrity of data put into Cloud.

Dipti Bansode and Amar Buchade [14] 2015, the study of deduplication technique describes how to secure the data on Cloud. This system has two components front end and back end. It proposed uses application aware index structure. The result of these experiments achieves reliability in deduplication. In future need to focus on data acess and deletion.

Xinyi Huang, Shaohua Tang and Yang Xiang [15] 2015, their first attempt to formalize the notion of distributed reliable deduplication system. They proposed new distributed deduplication in which data distributed across multiple Cloud servers. It shows that the incurred overhead is very limited in realistic environments.

Wen Xia and Min Fu [16] 2015,explained that cross- user redundant data are arising from duplicate files. They encrypted using method of convergent encryption. Its main aim is to backup the Cloud storage, performs deduplication to save space and network bandwidth. The solution is to achieve minimum storage space compared with existing. So user aware of convergent key encryption and multilevel key management has been done under this technique. The experimental result is to provide better performance. Data Duplication Avoidance and Cloud Computing Security Using AES and SHA Algorithm

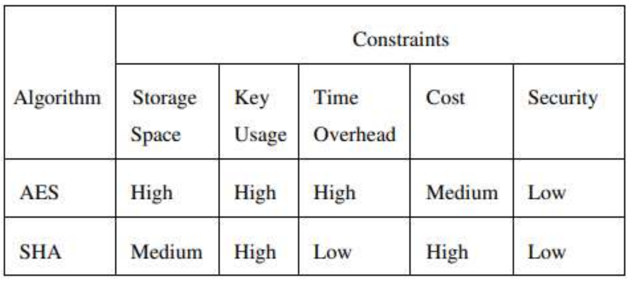
Dr. Rajendra Gode Institute of Technology & Research, Amravati Page 8 Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee and Wenjing Lou [17] 2014, introduced the base line approach where user just keeps the master keys. The proposed scheme is used for efficient and convergent key management. They use different constraints to achieve its target using proposed Dekey. Due to Dekey user need not manage the key by own. The overall result of this experiment is the convergent keys are distributed across multiple servers. Therefore it has been partially succeeded in key management.

Mihir Bellare, sriram Keelveedhi and Thomas Ristenpart [18] 2013, stated about message locked encryption for to resolve the duplication of files. This encryption is used to give increase the strong confidentiality of outsourced file and guarantee. They handled it with storage plain text by knowing its structure and size. In Cloud it is used to give an optimal solution for the proposed work.

MihirBellare, Sriramkeelveedhi and Thomas Ristenpart [19] 2013, introduced encryption and decryption performed from message. They presented it for the purpose of achieve secure deduplication. They extract paradigm to deliver schemes under different assumptions and for various classes of message sources. But the result does not provide deduplication to expected level. It suffered in brute force attack.

Arthur Rahumed, Henry C.H chen, Yang Tang, Patrick P.C.Lee and John C.Lui [20] 2011, their goal is to take backup Table 2.1 Research Issues for outsourced data with low cost. They used the fade version which eliminates redundancy among the data. Fade version had minimal performance overhead than other traditional Backup Service.

Table 2.1 Research Issues



Algorithm	Constraints				
	Storage Space	Key Usage	Time Overhead	Cost	Security
AES	High	High	High	Medium	Low
SHA	Medium	High	Low	High	Low

III. PROPOSED METHODOLOGY

The secure cloud data storage used for further access and the data is transferred via Internet connection and stored in a selected domain, but network and domain not under control of the clients at all. These problems not originate which user take the unauthorized access from the cloud storage is susceptible to security threats from both outside and inside of the cloud, during these process some data will loss from the clients may be hidden by the uncontrolled cloud servers to maintain the reputation. The most important parameter is that for clients are the data security which is less accessed is deliberately deleted by the servers to maintain the cost and space. This system considering the large data size of the outsourced data files uploaded by registered user and the clients' constrained resource capabilities, the first problem is as how can the client efficiently analysis the verifications based on the proposed algorithm even without the local copy of data files. Then we solve this above all problem using the detecting is secure de-duplication file on cloud storage. The remove increased data on cloud server by the cloud server provider stored at remote cloud servers accompany the rapid adoption of cloud services.

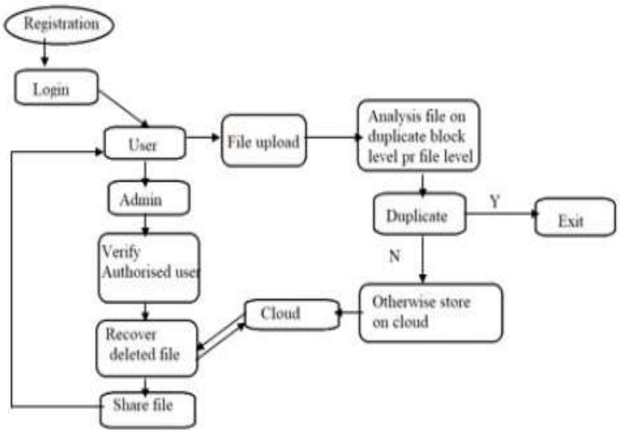


Fig. 2 Flow Chart of Methodology

IV. CONCLUSION

We implemented our deduplication systems using the Encryption and Hashing algorithm scheme and demonstrated that it overhead compared to the network transmission over-head in regular upload/download operations. Data de-duplication merchants are providing their products with enhanced capabilities which happen to increase its rate of adoption in and overthrow the challenges from substitutes such as cloud, tape and sometimes regular disks in a server. The systems are becoming more mature from disk storage systems involving de-duplication to genuinely complete data protection devices that can be integrated into applications and backup software's for upgraded management and efficiency. 3.Data de-duplication is the emerging trend and the secured deduplication is the important concerns of the cloud users. This paper focuses on the basics of de- duplication and how deduplication is done, what are the various papers based in this with different algorithms, and how this is different from others. The proposed plan will be applied for block level deduplication and the outcomes represent there diction in storage size and take less effort to store efficient information. At present, an improved strategy for capacity has been tried just for text files. In future work, it may extend for different sorts of files, for example, video and audio files.

V. REFERENCES

[1] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable data possession at untrusted stores," in Proceedings of the 14th ACM Conference on Computer and Communications Security, ser. CCS '07. New York, NY, USA: ACM, 2007, pp. 598– 609.

[2] G. Ateniese, R. Di Pietro, L. V. Mancini, and G. Tsudik, "Scalable and efficient provable data possession," in Proceedings of the 4th International Conference on Security and Privacy in Communication Netowrks, ser. SecureComm '08. New York, NY, USA: ACM, 2008, pp. 9:1– 9:10.

[3] C. Erway, A. K¨upc¸ ¨u, C. Papamanthou, and R. Tamassia, "Dynamic provable data possession," in Proceedings of the 16th ACM Conference on Computer and Communications Security, ser. CCS '09. New York, NY, USA: ACM, 2009, pp. 213– 222.

[4] F. Seb´e, J. Domingo-Ferrer, A. Martinez-Balleste, Y. Deswarte, and J.-J. Quisquater, "Efficient remote data possession checking in critical information infrastructures," IEEE Trans. on Knowl. and Data Eng., vol. 20, no. 8, pp. 1034–1038, 2008.

[5] H. Wang, "Proxy provable data possession in public clouds," IEEE Transactions on Services Computing, vol. 6, no. 4, pp. 551–559, 2013.

[6] Y. Zhu, H. Hu, G.-J. Ahn, and M. Yu, "Cooperative provable data possession for integrity verification in multicloud storage," IEEE Transactions on Parallel and Distributed Systems, vol. 23, no. 12, pp. 2231– 2244, 2012.

[7] H. Shacham and B. Waters, "Compact proofs of retrievability," in Proceedings of the 14th International Conference on the Theory and Application of Cryptology and Information Security: Advances in Cryptology, ser. ASIACRYPT '08. Springer Berlin Heidelberg, 2008, pp. 90– 107.

[8] Q. Wang, C. Wang, J. Li, K. Ren, and W. Lou, "Enabling public verifiability and data dynamics for storage security in cloud computing," in Computer Security – ESORICS 2009, M. Backes and P. Ning, Eds., vol. 5789. Springer Berlin Heidelberg, 2009, pp. 355–370.

[9] J. Xu and E.-C. Chang, "Towards efficient proofs of retrievability," in Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, ser. ASIACCS '12. New York, NY, USA: ACM, 2012, pp. 79– 80. Data Duplication Avoidance and Cloud Computing Security Using AES and SHA Algorithm Dr. Rajendra Gode Institude of Technology & Research, Amravati Page 75.

[10] E. Stefanov, M. van Dijk, A. Juels, and A. Oprea, "Iris: A scalable cloud file system with efficient integrity checks," in Proceedings of the 28th Annual Computer Security Applications Conference, ser. ACSAC '12. New York, NY, USA: ACM, 2012, pp. 229– 238.

[11] M. Azraoui, K. Elkhiyaoui, R. Molva, and M. O¨ nen, "Stealthguard: Proofs of retrievability with hidden watchdogs," in Computer Security - ESORICS 2014, ser. Lecture Notes in Computer Science, M. Kutyłowski and J. Vaidya, Eds., vol. 8712. Springer International Publishing, 2014, pp. 239–256.