iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** XII **Month of publication:** December 2018

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue XII, Dec 2018- Available at www.ijraset.com

Client-Side Deduplication and Auditing on Big Data-Cryptographic Algorithms

Prof. Pritam Ahire[1], Aditya Wadkar[2], Aniket Chikode[3], Akash Kakare[4], Gaurav Patil[5]
[1, 2, 3, 4, 5]*Computer Engineering, First-Third University*

Abstract: ***In current scenario, there is a considerable increase in the amount of data stored in storage services along with dramatic evolution of networking techniques. In cloud storage services with vast amount of data, the cloud servers may want to reduce the volume of stored data, and the clients may want to check the integrity of their data with a minimum cost, since the cost of the functions related to data storage increase with respect to the size of the data. To achieve these goals, secure deduplication and integrity auditing techniques have been studied. The system also computes the results using serial and parallel data processing and calculated the time, which show the performance of serial and parallel results. Final implementation results show the frequent outcome for a particular file to upload on cloud and using these results the user get the comparative research on serial and parallel operation. In this paper, the architecture describes the combined techniques which performs both secure deduplication of encrypted data and public integrity auditing of data. To support two functions, the proposed scheme performs challenge response protocols using the BLS signature based homomorphic linear authenticator. System utilizes a third party auditor for performing public audit, in order to help low-powered clients. The proposed system satisfies all the security requirements. It also proposes two variances that provide higher security and better performance.***

Keywords: Cloud storage, Data security, Information security, Public audit, Secure deduplication, Chunking techniques, Encryption, Decryption, Proof of Ownership, Homomorphic

I. INTRODUCTION

Cloud storage is a service that enables people to store their data on a remote server[5]. With a rapid growth in user base, cloud storage providers tend to save storage costs via cross-user deduplication: if two clients upload the same file, the storage server detects the duplication and stores only a single copy. Deduplication achieves high storage savings and is adopted by many storage providers. It is also adopted widely in backup systems for enterprise workstations. Cloud storage provides customers with benefits, ranging from cost saving and simplified convenience, to mobility opportunities and scalable service. These great features attract more and more customers utilize and storage their personal data to the cloud storage: according to the analysis report, the volume of data in cloud is expected to achieve 40 trillion gigabytes in 2020. Even though cloud storage system has been widely adopted, it fails to accommodate some important emerging needs such as the abilities of auditing integrity of cloud files by cloud clients and detecting duplicated files by cloud servers[4][6]. System illustrate both problems below.

Deduplication is the most efficient technique, a process of identifying and eliminating redundant data[6][7]. On the client side deduplication, data is duplicated on the client side where the client sends only new, unique data across the network, which results in reduced storage capacity and network bandwidth savings[7]. The benefits of deduplication include reduced infrastructure costs, reduced management costs, many cloud storage providers such as Dropbox, Memopal and Mozy use client side deduplication in order to save resources which results in avoiding storage of redundant data in cloud storage servers and network bandwidth savings by eliminating transmission of same contents several times[4][7].

Even though there are advantages in client side deduplication it has issues related to security, for example, attackers will mainly target the bandwidth and confidentiality which is related to privacy of legitimate cloud users[4][7].

In order to solve these concerns, Proof of Ownership (POW) schemes is introduced where they allow the storage server to check a client data ownership, based on a hash value[4][5][7]. Even though existing scheme deals with different properties of security, but there still is a need for careful consideration of potential attacks which includes data leakage and poison attacks, which mainly target on privacy preservation and data confidentiality. In the baseline approach which is a proof of Ownership (POW) scheme a new cryptographic method which uses the hashing technique and encryption, which results in efficient data deduplication while providing data security in cloud storage systems and providing dynamic sharing between users.

The scheme supports both secure deduplication and integrity auditing in a cloud environment. In particular, the proposed scheme provides secure deduplication of encrypted data. Our scheme performs POW for secure duplication and integrity auditing based on

the homomorphic linear authenticator (HLA), which is designed using BLS signature. The proposed scheme also supports public auditing using a TPA (Third Party Auditor) to help low-powered clients. The proposed scheme satisfies all fundamental security requirements, and is more efficient than the existing schemes that are designed to support deduplication and public auditing at the same time. The main improvement in this paper is that system propose two variations to provide higher security and better performance. In the first variant, which is designed for stronger security, architecture assume that stronger adversary and provide a counter measure against the adversary. In the second variant, system design a technique that supports a very low-powered client and entrusts more computation to the cloud storage server in the upload procedure[1][4][7].

II. MOTIVATION

Cloud computing offers a new way of Information Technology services by rearranging various resources(e.g., storage, computing) and providing them to users based on their demands. The most important and popular cloud service is data storage service. Cloud users upload personal or confidential data to the data center of a Cloud Service Provider (CSP) and allow it to maintain these data. Since intrusions and attacks towards sensitive data at CSP are not avoidable ,it is users may upload duplicated data in encrypted form to CSP, especially for scenarios where data are shared among many users. Although cloud storage space is huge, data duplication greatly wastes network resources, consumes a lot of energy, and complicates data management. Our proposed preventable to assume that CSP cannot be fully trusted by cloud users. Moreover, the loss of control over their own personal data leads to high data security risks, especially data privacy leakages. Due to the rapid development of data mining and other analysis technologies, the privacy issue becomes serious . Hence, a good practice is to only outsource encrypted data to the cloud in order to ensure data security and user privacy. But the same or different system secure deduplication and integrity auditing solves the problem of duplication of data and security of data.

III.RELATED WORK

Reconciling deduplication and client-side encryption is an active research topic. Douceur et al. describes the deduplication problem present in multi-tenant environment. It used Message-Locked Encryption (MLE) as a prominent manifestation for convergent key encryption. Letting M be a file's data, a client first computes a key K <- H(M)by applying a cryptographic hash function H to M, and then computes cipher-text C E(K,M); via a deterministic symmetric encryption scheme. A second client B encrypting the same file M will produce the same C, enabling deduplication. However, CE is subject to an inherent security limitation, namely, susceptibility to offline brute- force dictionary attacks The authors described about convergent encryption where keys are derived from the hash of data[2][4]. Then, in 2008 Storer et al. proposed two approaches for secured data deduplication. Where it has disadvantages in security and in deduplication open areas for exploration exist, multiple levels of permissions can be utilized for future designs.

The proof of Ownership (POW) is proposed by Halevi[1][5]. It is a challenge response protocol that enables a storage server to verify whether a request is from the data owner, which is based on hash value. Whenever a client or owner uploads a data file to the cloud server, he has to compute a hash value and sends this value to the cloud storage. The cloud storage server has a database of shared values (hash value) of all stored files. If the hash value is present in the cloud server then the file is already outsourced and it informs the data owner that the uploading of file is not required[2][5][6][7].

1) *Security Analysis:* Despite having the significant resource saving advantages, POW schemes comes along with a number of security challenges that may create a dangerous environment for sensitive data[4].
2) *Data Confidentiality Disclosure:* Data Confidentiality is an important concern[4].
3) *Privacy Violation:* Sensitive data leakage is a major critical challenge that was not addressed by Halevi et al. The cloud storage should not build user profiles and access the data stored by the user in the cloud[4].
4) *Poison Attack:* The data file is encrypted by using some random encrypted key. Now the cloud server cannot verify the uploaded file and the hash present in its database as values are different and attacker can easily replace encrypted original file with a malicious file [4].

In our scheme, each user has to generate the integrity tags, even for the file in the cloud. Moreover, the file is available in its plain form on the cloud side. Li et al. proposed an integrity auditing scheme for encrypted deduplication storage. This scheme is based on homomorphic verifiable tags and Merkle hash tree. A user encrypts his file by using a convergent encryption technique and uploads the file to a fully trusted TPA.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue XII, Dec 2018- Available at www.ijraset.com

IV.LITERATURE SURVEY

TABLE I LITERATURE SURVEY

Sr.No	Paper Title	Approaches	Limitations
1	"Secure Deduplication with efficient and reliable Convergent Key Management"[6].	Convergent Key Management, Encryption, Decryption, Tag generation, Secret Key Sharing.	It does not efficient data security focus on in the cloud.
2	"DupLESS: Server-Aided Encryption for Deduplicated Storage" [7].	Message-Locked Encryption, Decryption, Security, Key Generation.	Performance overhead, Able to perform only on small size data,It can resist to offline brute force attacks
3	"Secure Deduplication of Encrypted Data without Additional Independent Servers" [5].	PAKE(password authenticated key- exchange), Encryption, Decryption, Cross user deduplication. .	Additional attacks are possible when considering the long term operation of the system since it involves multiple rounds of protocols
4	"A hybrid cloud approach for secure authorized deduplicat ion"[9].	Server-Side Deduplication, Encryption, Decryption.	Server-side deduplication can improve storage utilization, but does not result in any bandwidth. It does not work on good on encrypted as well.

V. PROPOSED SYSTEM

All title and author details must be in single-column format and must be centered. The data outsourcing model is first formulated. There are four entities, namely user or client, Load servers, Cloud Storage servers, Third Party Auditor (TPA).

1) *Client (or user):* Outsources data to cloud storage. CE-Encrypted data are first generated, and then uploaded to the cloud storage to protect confidentiality. The client also needs to verify the integrity of the outsourced data. To this client delegates integrity auditing to the TPA[1][4].
2) *Cloud Storage Server(CSS):* Provides data storage services to the users. Deduplication technology is applied to save storage space and cost. System consider that the CSS may act maliciously due to the insider/outsider attacks , software/hardware malfunctions, intentional saving of computational resources, etc. During the duplication process, the CSS carries out PoW protocol to verify that the client owns the file. Moreover, in the integrity audit process, it is necessary to generate and respond to a proof corresponding to the request of the TPA[1][2][4][5].
3) *TPA (Third Party Auditor):* Performs integrity auditing on behalf of the client to reduce the client's processing cost. Instead of the client, the auditor sends a challenge to the storage server to periodically perform an integrity audit protocol. TPA is assumed to be a semi-trust model, that is, an honest but curious model. Under the assumption, it is assumed that the TPA does not collude with other entities[1][5][7].
4) *Load Servers:* Load Servers are used to handle the load of various incoming requests. The load is allocated to the server on count and priority basis.

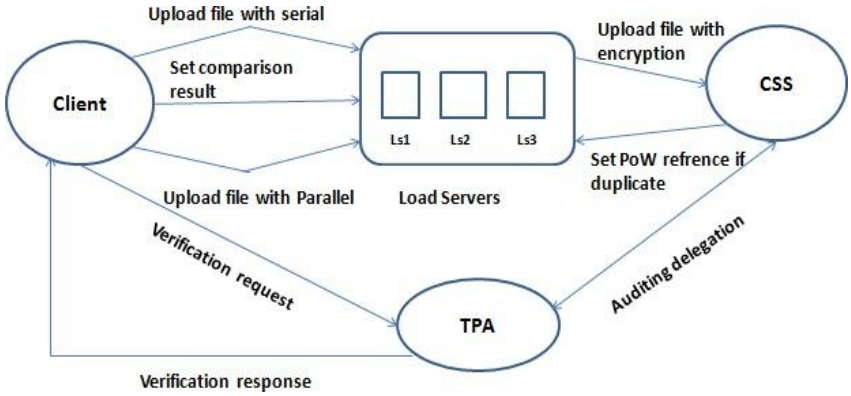


Fig. 1 System Model

In this work, design shows a data copy to be either a whole file or a small-sized blocks, and this leads to two types of deduplication: 1) file-level deduplication, which eliminates the storage of any redundant files, and 2) block-level deduplication, which divides a file into a smaller fixed size or variable-size blocks and the eliminates redundant blocks and store only the blocks which are unique. Using fixed-size blocks simplify the computations of block boundaries, while using variable-size blocks provide better deduplication efficiency. System used deduplication mechanism in both file and block levels. Specifically to upload a file, a client first performs the file level duplicate check. If the file is duplicated, then all its blocks must be duplicated as well; otherwise, the user further performs the block level duplicate check and identifies the unique blocks to be uploaded. Each file or a block is associated with a tag for the duplicate check. All the data copies and tags will be stored in the CSS[4][7].
System also does operation on big data processing, means the file of huge size will also be processed. It will use both the serial and parallel technique to upload the file. Thus, it shows the comparison result between the serial and parallel technique of the operation. Finally, it will show the timeline chart regarding the serial and parallel techniques implemented.

VI. ALGORITHMS

A. *Chunking Algorithm:* It is used to divide the file into variable size chunks or blocks. The number of chunks in the data depend on the content of the file.
B. *Advanced Encryption Standard (AES):* AES is a symmetric key algorithm for public security. It converts the plain text into cipher text.
C. *Message Digest 5 (MD5):* MD5 is based on hashing technique. It generates the hash value for each and every type of data. Duplicate data is checked based on the hash value.

VII. CONCLUSION

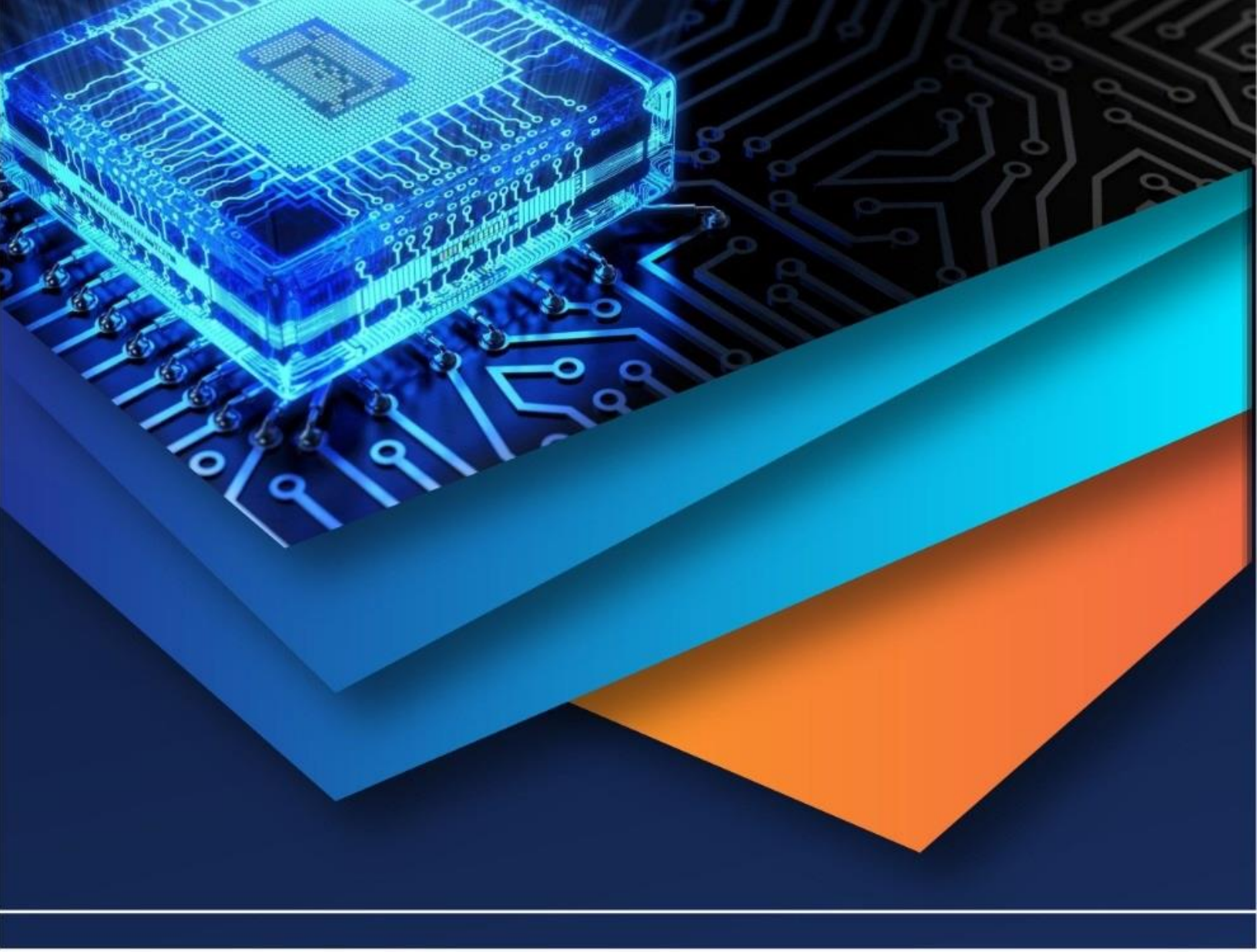
In this paper, the design implementation dealt with the dilemma that cloud storage providers want to use deduplication to save cost, while users want their data to be encrypted on client-side. In addition, cloud servers want to use their storage more efficiently. To satisfy both the requirements, system proposed a scheme to achieve both secure deduplication and integrity auditing in a cloud environment. To prevent leakage of important information about user data, the proposed scheme supports a client side deduplication of encrypted data, while simultaneously supporting public auditing of encrypted data. Also, it makes operation on big data and it shows the comparison result of serial and parallel technique.

VIII. ACKNOWLEDGEMENT

The authors would like to thank the publishers, researchers for making their resource available and teachers for their guidance. We thank the college authority for providing the required infrastructure and technical support. Finally, we extend our heartgelt gratitude to friends and family members.

REFERENCES

[1] Taek-Young Youn1, Ku-Young Chang1, Kyung Hyune Rhee2, AND Sang UK Shin2,".Efficient Client-Side Deduplication of Encrypted Data with Public Auditing in Cloud Storage", 10.1109/ACCESS.2018.2836328, IEEE Access.
[2] JZheng Yan, Senior Member, IEEE, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, Fellow, IEEE, "Deduplication on Encrypted Big Data in Cloud", IEEE Transactions on Big Data,
[3] J. Li, J. Li, D. Xie and Z. Cai, "Secure auditing and deduplicating data in cloud," IEEE Transactions on Computers, vol. 65, no. 8, pp. 2386–2396, Aug. 2016.
[4] Naveen AN,V Ravi,"Client Side Deduplication Scheme for Secured Data Storage in Cloud Envioronments",International Journal of Research & Technolgy(IJERT),VOL NO .4,Issue 05,May 2015
[5] Jian Liu,N.Asokan,Benny Pinkas,"Secure Deduplication of Encrypted data without Additional Independent Servers", IEEE Conferences on Cloud Computing,Aug. 2015
[6] Jin Li,Xiaofeng Chen,Mingqiang,Jingwei Patrick P.C. Lee, and Wenjing Lou, "Secure Deduplication with Efficient and Reliable Convergent Key Management",IEEE Transaction on parallel and Distributed Systems, VOL
[7] S. Keelveedhi and M. Bellare and T. Ristenpart, "DupLESS: server aided encryption for deduplicated storage," in Proc. of the 22nd USENIX Security Symposium (USENIX Security 13), Washington, D.C. USA, 2013, pp. 179–194.
[8] Jian Liu,N.Asokan,Benny Pinkas," Secure Deduplication of Encrypted Data without Additional Independent Servers", IEEE Conferences on Cloud Computing.
[9] J. Li, Y. K. Li, X. Chen," *A hybrid cloud approach for secure authorized deduplication*"IEEE Conferences on Cloud Computing.
[10] "Y. Dodis, S. Vadhan and D. Wichs, "Proofs of retrievability via hardness amplification," in Proc. of the 6th Theory of Cryptography Conference on Theory of Cryptography (TCC'09), San Francisco, CA, USA, 2009, pp. 109–127.
[11] A.T. Clements, I. Ahmad, M. Vilayannur, andJ. Li, "Decentralized Deduplication in San Cluster File Systems," in Proc. USENIX ATC,2009,p.8.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY