Deduplication of Data Using MLE Algorithm in Cloud Computing

Stanly Jayaprakash[1], Arulselvan[2], Bregit Raj[3], Dhinesh[4], Jeeva[5]
[1]Assistant Professor, Dept of Computer Science
[2, 3, 4, 5]Dept of Computer Science
[1, 2, 3, 4, 5]Mahendra Institute Of Technology, Tamil Nadu, Namakkal DT – 637 503

Abstract- *The cloud storage auditing with deduplication can verify the integrity of data stored in the cloud while the cloud needs to keep only a single copy of duplicated le. To the best of our knowledge, all the existing cloud storage auditing schemes with deduplication are vulnerable to brute-force dictionary attacks, which incurs the leakage of user privacy. In this paper, we focus on a new aspect of being against brute-force dictionary attacks on cloud storage auditing. We propose a cloud storage auditing scheme with deduplication supporting strong privacy protection, in which the privacy of user's le would not be disclosed to the cloud and other parties when this user's le is predictable or from a small space. In the proposed scheme, we design a novel method to generate the le index for duplicate check, and use a new strategy to generate the key for le encryption. In addition, the user only needs to perform lightweight computation to generate data authenticators, verify cloud data integrity, and retrieve the le from the cloud. The security proof and the performance evaluation demonstrate that the proposed scheme achieves desirable security and efficiency.In chunk-based deduplication systems, logically consecutive chunks are physically scattered in different containers after deduplication, which results in the serious fragmentation problem. The fragmentation significantly reduces the restore performance due to reading the scattered chunks from different containers. Existing work aims to rewrite the fragmented duplicate chunks into new containers to improve the restore performance, which however produces the redundancy among containers, decreasing the deduplication ratio and resulting in redundant chunks in containers retrieved to restore the backup, which wastes limited disk bandwidth and decreases restore speed. To improve the restore performance while ensuring the high deduplication ratio, this paper proposes a cost-efficient submodular maximization rewriting scheme (SMR). SMR first formulates the defragmentation as an optimization problem of selecting suitable containers, and then builds a submodular maximization model to address this problem by selecting containers with more distinct referenced chunks. Move over, this paper further leverages the grouped form, i.e., GSMR, to reduce the fragmented chunks caused by the accumulated differences among backup versions. We implement SMR in the deduplication system, which is evaluated via three real-world datasets. Experimental results demonstrate that SMR is superior to the state-of-the-art work in terms of the restore performance as well as deduplication ratio, and GSMR further improves the restore performance. We have released the source code of SMR in GitHub for public use.*

I. INTRODUCTION

The rapid development of cloud computing, cloud storage has been widely accepted by individuals and enter-praises for its advantages of universal access, low costs and on-demand service. Users can outsource complex computations to the cloud to reduce their computational bur-den [40], [41]. In addition, users also can outsource their large-scale data to the cloud to release their local storage burden. Under such a trend, it becomes urgent to guarantee the quality of data storage services for the users and the cloud. On one hand, the outsourced data might be corrupted or lost due to the inevitable operation errors or soft- ware/hardware failures in the cloud. Thus, it is critical to develop cloud storage auditing, by which users can verify the integrity of cloud data without downloading the whole data from the cloud. On the other hand, lots of data stored in the cloud are duplicated.

The cloud cannot deduce or derive the convergent key from the content of file since a secret ``seed" is embedded in the convergent key. Unfortunately, the key server is able to guess or derive the content from the file's hash value sent from the user by launching the brute-force dictionary attacks. Therefore, this scheme cannot fully prevent the brute-force dictionary attacks. In addition, all users who want to upload file to the cloud need to generate a file index and send it to the cloud for duplicate check. With the file index, the cloud can verify whether the file uploaded by the user is duplicated or not. If the file index has been kept by the cloud, then the subsequent users do not need to upload data to the cloud any more. Most of deduplication schemes set the hash value of the file as the file index. It will result in the data privacy leakage because the malicious cloud or other parties might guess or derive the content of file by performing the brute-force dictionary attacks.

1.1 CLOUD STORAGE AUDITING

A cloud auditor is a party that can perform an independent examination of cloud service controls with the intent to express an opinion thereon. Audits are performed to verify conformance to standards through review of objective evidence. In Cloud computing, the term cloud is a metaphor for the Internet, so the phrase Cloud computing is defined as a type of Internet-based computing, where different services are delivered to an organization's computers and devices through the Internet. Cloud computing is very promising for the Information Technology (IT) applications; however, there are still some issues to be solved for personal users and enterprises to store data and deploy applications in the Cloud computing environment. Data security is one of the most significant barriers to its adoption and it is followed by issues including compliance, privacy, trust, and legal matters.

1.2 CLOUD STORAGE AUDITING

A cloud auditor is a party that can perform an independent examination of cloud service controls with the intent to express an opinion thereon. Audits are performed to verify conformance to standards through review of objective evidence. In Cloud computing, the term cloud is a metaphor for the Internet, so the phrase Cloud computing is defined as a type of Internet-based computing, where different services are delivered to an organization's computers and devices through the Internet. Cloud computing is very promising for the Information Technology (IT) applications; however, there are still some issues to be solved for personal users and enterprises to store data and deploy applications in the Cloud computing environment. Data security is one of the most significant barriers to its adoption and it is followed by issues including compliance, privacy, trust, and legal matters.

1.3STRONG PRIVACY PROTECTION

Cloud computing proposes the opportunity to organizations that would merely connect to the cloud and use the available resources on a Pay Per use basis that avoids the company's capital expenditure on supplementary of premises infrastructure resources. It promptly scales up and scales down rendering to business requirements. It consists of cloud client, services, application platform, storage, and infrastructure measured services. Thus, the cloud computing is highly automated utility-based paradigm shift comprises of efficient and optimized framework that includes virtual desktops, servers and allocates services for computer network over the internet suggesting software applications and platform for easy and agile deployment of the secure data management.

1.4 DATA SECURITY

Data security refers to the process of protecting data from unauthorized access and data corruption throughout its lifecycle. Data security includes data encryption, hashing, tokenization, and key management practices that protect data across all applications and platforms.

1.5 CLOUD STORAGE

Cloud storage is a model of computer data storage in which the digital data is stored in logical pools, said to be on "the cloud". The physical storage spans multiple servers (sometimes in multiple locations), and the physical environment is typically owned and managed by a hosting company. These cloud storage providers are responsible for keeping the data available and accessible, and the physical environment protected and running. People and organizations buy or lease storage capacity from the providers to store user, organization, or application data

II. EXISTING SYSTEM

In general, we can divide these approaches into four categories: simple ciphertext access control, hierarchical access control, access control based on fully homomorphic encryption and access control based on attribute-based encryption (ABE). All these proposals are designed for non-mobile cloud environment. Tysowski et al. considered a specific cloud computing environment where data are accessed by resource-constrained mobile devices, and proposed novel modifications to ABE, which assigned the higher computational.

III. PROPSED SYSTEM

In this project, we focus on a new aspect of being against brute-force dictionary attacks on cloud storage auditing. We propose a cloud storage auditing scheme with deduplication supporting strong privacy protection, in which the privacy of user's le would not be disclosed to the cloud and other parties when this user's le is predictable or from a small space. In the proposed scheme, we design a novel method to generate the le index for duplicate check, and use a new strategy to generate the key for le encryption. In addition, the user only needs to perform lightweight computation to generate data authenticators, verify cloud data integrity, and retrieve the le from the cloud. The security proof and the performance evaluation demonstrate that the proposed scheme achieves desirable security and efficiency.

IV. MODULE DESCRIPTION

- Admin Module
- Data User Module
- Private Cloud Module
- De-Duplication Module
- Privacy Preserving Module

Admin Module

In this module, admin is the main person who has a facility to monitor all the activities that are been processed by cloud user. The facilities are like upload, download and update.

Data User Module

In this module, the data user is the person who wishes to store the data securely in the cloud. He will store the data which is to be stored safely in the cloud server and in turn downloads the data whenever he wants to access that data. He has the following facilities like upload, download and update. For accessing these privileges he should get access rights from the private cloud service provider.

PRIVATE CLOUD MODULE

This is the module in which he will give allow or deny permission to the cloud user at the time of registration.He also has the facility like giving access privileges for the users who request for getting upload, download or update privileges for their uploaded files.He can also view the uploaded file information.

DEDUPLICATION

In this module the data will not be duplicated to store in the cloud data base. If the file which is already kept in the server by an user like _X‘ can‘t be stored once again by any other user including _X‘.This method is known as data de-duplication technique where the data once uploaded on cloud server can‘t be uploaded once again in the cloud server.

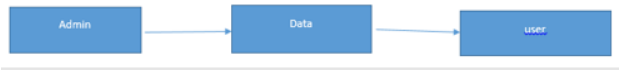
PRIVACY PRESERVING MODULE

In this module we can provide privacy preserving for the uploaded data in the cloud server by encrypting the data which is stored on the cloud server.

V. DATA FLOW DIAGRAM

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

LEVEL 0



VI. DEPLOYMENT

Component diagrams are used to describe the components and deployment diagrams shows how they are deployed in hardware. UML is mainly designed to focus on the software artifacts of a system. However, these two diagrams are special diagrams used to focus on software and hardware components.

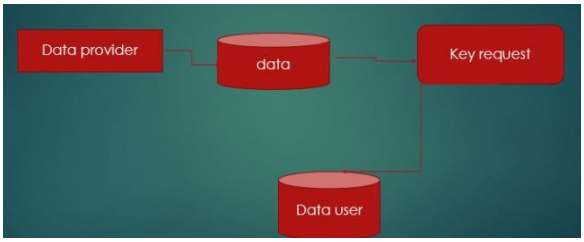


Fig.Deployment Diagram

VII. CONCLUSION AND FUTURE ENHANCEMENT

In this project, we study on how to solve the problem of user's privacy leakage in cloud storage auditing with deduplication when brute-force dictionary attacks are launched. We design a lightweight cloud storage auditing scheme with deduplication supporting strong privacy protection. In the proposed scheme, the privacy of user can be well preserved against the cloud and other parties. The user

relieves the heavy computation burden for generating data authenticators and verifying data integrity. The security proof shows that the proposed scheme is secure. We also provide detailed comparisons among our proposed scheme and other existing schemes by experiments. Experimental results show the proposed scheme achieves higher storage efficiency and is more efficient.The security proof shows that the proposed scheme is secure. We also provide detailed comparisons among our proposed scheme and other existing schemes by experiments. Experimental results show the proposed scheme achieves higher storage efficiency and is more efficient in authenticator generation phase and auditing phase.

VIII. DECLARATION

Conflicts of Interest
No conflict of interest in this manuscript
Authors Contributions
Stanly Jaya Prakash, Arulselvan was involved in data collection, data analysis & manuscript writing.
Author,Bregit Raj, Jeeva, Dhinesh was involved in conceptualization, data validation, and critical review of manuscripts.

Acknowledgment

The authors would like to express their gratitude towards Mahendra Institute of Technology (Formerly known as Mahendra University) for successfully carrying out this work.

REFERENCES

[1] The Gnu Multiple Precision Arithmetic Library (GMP). Accessed: Oct. 2019. [Online]. Available: http://gmplib.org/
[2] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, ``Provable data possession at untrusted stores," in Proc. 14th ACM Conf. Comput. Commun. Secur. (CCS), 2007, pp. 598609.
[3] M. Bellare, S. Keelveedhi, and T. Ristenpart, ``Message-locked encryption and secure deduplication," in Proc. Annu. Int. Conf. Theory Appl.Crypto- graph. Techn. Berlin Germany: Springer, 2013, pp. 296312.
[4] H. Cui, R. H. Deng, Y. Li, and G. Wu, ``Attribute-based storage supporting secure deduplication of encrypted data in cloud," IEEE Trans. Big Data, vol. 5, no. 3, pp. 330342, Sep. 2019.
[5] R. Ding, H. Zhong, J. Ma, X. Liu, and J. Ning, ``Lightweight privacy- preserving identity-based veritable IoT-based health storage system," IEEE Internet Things J., vol. 6, no. 5, pp. 83938405, Oct. 2019.
[6] J. R. Douceur, A. Adya, W. J. Bolosky, P. Simon, and M. Theimer, ``Reclaiming space from duplicate files in a server less distributed file system," in Proc. 22nd Int. Conf. Distrib. Comput. Syst., Jul. 2002, pp. 617624.
[7]] Y. Fan, X. Lin, G. Tan, Y. Zhang, W. Dong, and J. Lei, ``One secure data integrity verification scheme for cloud storage," Future Gener. Comput. Syst., vol. 96, pp. 376385, Jul. 2019.
[8] J. Gantz and D. Reinsel. (2012). The Digital Universe DecadeAre You Ready (2010). [Online]. Available: http://www. emc.com/collateral/analystreports/idcdigital-universe-are-you-ready.pdf
[9] X. Ge, J. Yu, H. Zhang, C. Hu, Z. Li, Z. Qin, and R. Hao, ``Towards achieving keyword search over dynamic encrypted cloud data with symmetric-key based verification," IEEE Trans. Dependable Secure Com- put., to be published.