ISSN 2395-1621

Detecting Deduplication With Secure File Storage

[#1]Malti Lamkhade, [#2]Sneha Meshram, [#3]Pramila Rajmane, [#4]Jyoti Panpatte, [#5]Prof. Bombale G. R.

[1]maltilamkhade98@gmail.com,
[2]snehameshram1352@gmail.com,
[3]pramilarajmane16@gmail.com,
[4]jyotipanpatte56@gmail.com

[#12345]Department of Computer Engineering

Parvatibai Genba Moze College of Engineering, Wagholi, Pune.

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 (e.g., to determine whether plaintexts of two encrypted messages are identical). In this paper, we investigate proposed architecture, and propose an efficient and privacy-preserving big data deduplication in server side storage. Proposed structure achieves both privacy-preserving (encryption technique) and data availability. In addition, we take accountability into consideration to offer better privacy assurances than existing schemes.

Keywords:- Encryption Algorithm, Data Privacy, Data Redundancy verification , Notification System for Alert generation.

ARTICLE INFO

Article History

Received: 4th December 2019

Received in revised form :

5th December 2019

Accepted: 9th December 2019

Published online :

9th December 2019

I. INTRODUCTION

The Cloud is network for storage, access resources, where data is stored in pools of storage which are generally hosted by third parties. Cloud storage provides users with benefits, ranging from cost saving and simplified convenience, to mobility opportunities and scalable service. These properties used for customers to use and storage their personal data to the cloud storage: according to the analysis result in worlds, the volume of data in cloud is expected to achieve 40 trillion gigabytes in 2020. Even though cloud storage system has been widely used, it fails to accommodate some main emerging needs such as the abilities of auditing integrity of uploaded data cloud files by cloud clients and detecting duplicated files by cloud servers. We analysis both problems below. The first problem is integrity auditing in the cloud computing. The cloud server is able to remove clients from the heavy burden of storage management and maintenance.

The cloud storage used for storage is that the data is transferred via Internet and stored in an uncertain domain, not under control of the clients at all, which raises clients great concerns on the integrity of their data. These concerns originate from the fact that the cloud storage is susceptible to security threats from both outside and inside of the cloud [1], and some data loss from the clients may be hidden by the uncontrolled cloud servers to maintain the reputation. The most important thing is that for an ordinary clients the data which is rarely accessed is deliberately deleted by the servers to maintain the cost and space Considering the large size of the outsourced data files and the clients' constrained resource capabilities, the first problem is as how can the client efficiently

perform periodical in verifications even without the local copy of data files. The second problem is secure deduplication. The increased volumes of data stored at remote cloud servers accompanies the rapid adoption of cloud services is.

II. REVIEW OF LITERATURE

GDup: De-duplication of Scholarly Communication Big Graphs 2018 Claudio Atzori, Paolo Manghi, Alessia Bardi, In this paper, author propose the GDup system, an integrated, scalable, general-purpose system for entity deduplication over big information graphs.[1]

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.[2]

An Online Data Deduplication Approach for Virtual Machine Clusters 2018 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.[3]

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.[4]

RCDSD: RSA based Cross Domain Secure Deduplication on Cloud Storage2018,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.[5]

III. SYSTEM ARCHITECTURE / SYSTEM OVERVIEW

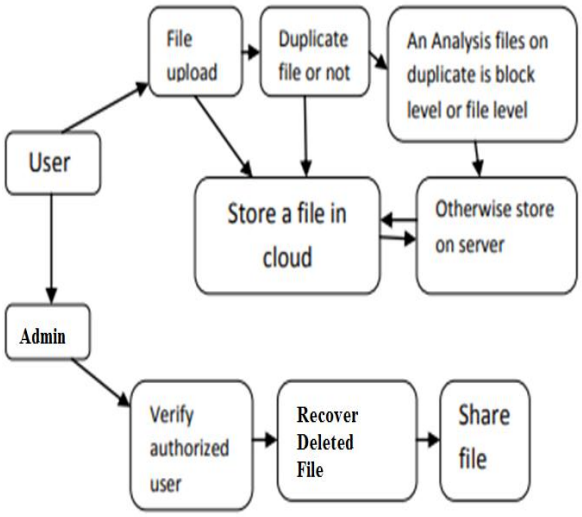

Fig 1. Block diagram

User Module:

Registration:

In this module, each user registers his user details for using files. Only registered user can able to login in cloud server.

File Upload:

In this module, user upload a block of files in the cloud with encryption by using his secret key. This ensures the files to be protected from unauthorized user. Here we analysis the duplicate data using the SHA algorithm.

Download:

This module allows the user to download the file using his secret key to decrypt the downloaded data verify the data and re-upload the block of file into cloud server with encryption. This ensures the files to be protected from unauthorized user.

Admin Module:

View Files:

In this module, public auditor view the all details of upload, download, blocked user, upload.

File Upload:

In this module, admin can also upload a block of files in the cloud with encryption by using his secret key. This ensures the files to be protected from unauthorized user.

IV. MATHEMATICAL MODEL

System Description:

Input:
Upload file ()
U : Upload file on cloud.
E : Encryption File.
S : Splitting file for security.
H : Hash value for each file.
Output:
Check Duplicate file on cloud storage

Input
Function Recovery (id, request, file)
ID : unique id for each file.
Request : User request for recovery of file.
File : Check file on cloud.
Output:
File will recover to data owner.

Encryption Process:

KeyGenCE(M) → K is the key generation algorithm that maps a data copy M to a convergent key K;

EncCE(K,M) → C is the encryption algorithm that takes both the convergent key K and the data copy M as inputs and then outputs cipher text C;

DecCE(K,C) → M is the decryption algorithm that takes both the cipher text C and the convergent key K as inputs and then outputs the original data copy M;

TagGen (M) → T(M) is the tag generation algorithm that maps the original data copy M and outputs a tag T(M)

V. 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.

ACKNOWLEDGMENT

I wish to express my profound thanks to all who helped us directly or indirectly in making this paper. Finally I wish to thank to all our friends and well-wishers who supported us in completing this paper successfully I am especially grateful to our guide Prof. Bombale G. R. Sir for him time to time, very much needed, valuable guidance. Without the full support and cheerful encouragement of my guide, the paper would not have been completed on time.

REFERENCES

[1] Claudio Atzori, Paolo Manghi, Alessia Bardi, “GDup: De-duplication of Scholarly Communication Big Graphs”IEEE 2018.

[2] Charles R. Haruna, MengShu Hou, Moses J. Eghan, “Cost-Based and Effective Human-Machine Based Data Deduplication Model in Entity Reconciliation”, IEEE 2018.

[3] Zhongwen Qian, Xudong Zhang, Xiaoming Ju, Bo Li, “An Online Data Deduplication Approach for Virtual Machine Clusters”, IEEE 2018.

[4] Ya Fan,Yong Wang, Miao Ye,” An improved small file storage strategy in Ceph File System”, IEEE 2018.

[5] Shivansh Mishra, Surjit Singh, Syed Taqi Al,” RCDSD: RSA based Cross Domain Secure Deduplication on Cloud Storage, IEEE 2018.

