ISSN 2395-1621

Cloud Based Data Integrity Analysis Using Data Analytics

Navnath Gutte, Anjali Randhire, Darshan Rathod, Kalpana Ubale

nagutte0@gmail.com
anjalirandhire123@gmail.com
darshanr189@gmail.com
kalpanaubaleak@gmail.com

DEPARTMENT OF COMPUTER ENGINEERING

JSPM'S IMPERIAL COLLEGE OF ENGINEERING AND RESEARCH
WAGHOLI, PUNE 412207 SAVITRIBAI PHULE PUNE UNIVERSITY

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: 3rd May 2021

Received in revised form : 3rd May 2021

Accepted: 5th May 2021

Published online : 5th May 2021

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

I. INTRODUCTION

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 1. Simple Structure

Problem Statement:

Cloud computing provides a low-cost, scalable, location independent infrastructure for data management and storage. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. To implement cloud based data integrity using data analytics.

II. LITERATURE SURVEY

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 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.[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 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.[5]

III. PROPOSED SYSTEM

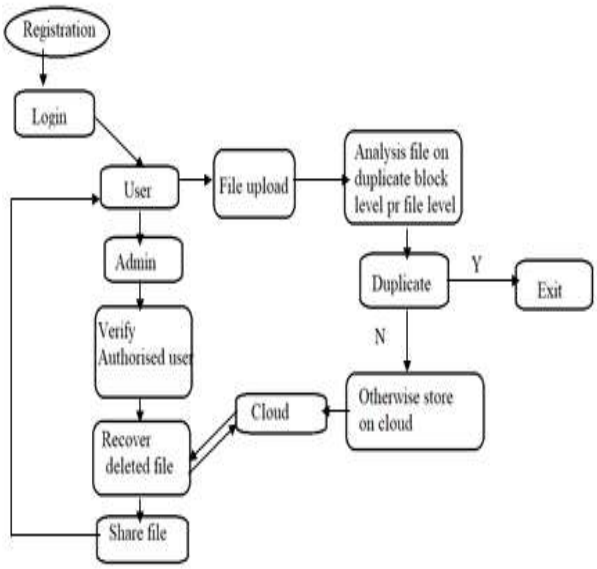

Fig 2. System Architecture

A. Description:

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 the servers to maintain the cost and space deliberately delete which is less accessed. 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.

B. Mathematical Model:

The 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.

Si: Success Condition
When duplicate files detected and data security also applies.

Fi: Failure Condition
When server fails then overall system will fails.

IV. ACKNOWLEDGEMENT

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 heartily thankful to my project guide for his valuable guidance and inspiration. In spite of their busy schedules they devoted their self and took keen and personal interest in giving us constant encouragement and timely suggestion. Without the full support and cheerful encouragement of my guide, the paper would not have been completed on time.

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.

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.

