iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50987

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: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Detecting Replicated Files in the Cloud

Mrs. G. Venkateswari[1], S. Sowjanya[2], S. Meenamrutha[3], K. S. Mounika[4], S. Hari Priyanka[5], M. Anvitha[6]

Abstract: ***Now-a-days all are using the cloud server to store their data and provides many features suitable for the users or customers. We have cloud servers like Google Cloud Platform,Microsoft Azure etc. For cloud also sometimes there will be storage problem to store the data of the users. We need the security to the data stored in the cloud. For hospitals and some private companies, the data should be secure and confidential. So we need both storage and thesecurity to our data stored in the cloud. Authorized Client-Side Deduplication Using CP-ABE here proposed which provides the security and provides Deduplication in the cloud.***

Keywords: Cloud storage, Deduplication, Encryption, Security, CP-ABE Algorithm

I. INTRODUCTION

With the ever-increasing popularity of cloud computing, the demand for cloud storage has also increased exponentially. Computing firms are no longer the only consumers of cloud storage and cloud computing, but rather average businesses, and even end-users, are taking advantage of the immense capabilities that cloud services can provide. While enjoying the flexibility and convenience brought by cloud storage, cloud users release control over their data, and particularly are often unable to locate the actual their data; this could be in-state, in-country, or even out-of-country. Lack of location control may cause privacy
breaches for cloud users (e.g., hospitals) who store sensitive data (e.g., medical records) that are governed by laws to remain within certain geographic boundaries and borders.

Another situation where this problem arises is with governmental entities that require all data to be stored in the same country that the government operates in; this challenge has seen difficulties with cloud service providers (CSPs) quietly moving data out-of-country or being bought out by foreign companies. For example, Canadian laws demand that personal identifiable data must be stored in Canada. However, large cloud infrastructure like the Amazon Cloud has more than 40 zones distributed all over the world [1], which makes it very challenging to provide guaranteed adherence to regulatory compliance. Even Hadoop, which historically has been managed as a geographically confined distributed file system, is now deployed in large scale across different regions (see Facebook Prism [2] or recent patent [3]).

To date, various tools have been proposed to help users verify the exact location of data stored in the cloud [4], [5], [6], with emphasis on post-allocation compliance. However, recent work has acknowledged the importance of a proactive location control for data placement consistent with adopters' location requirements [4], [7], [8], to allow users to have stronger control over their data and to guarantee the location where the data is stored.

A. Motivation

The motivation for detecting replicated files in the cloud is to ensure data integrity, save storage space, and reduce costs.

Replication is a common technique used in cloud storage to provide high availability and reliability of data. However , it can also lead to the creation of multiple copies of the same file, which can take up unnecessary storage space and increase storage costs. Moreover , it can also create inconsistencies in the data if changes are made to one copy of the file and not propagated to other copies. By eliminating duplicated files, organizations can reduce their storage footprint, maintain a single source of truth for their data, and reduce the risk of unauthorized access. Additionally, it can help organizations to improve performance by reducing latency and meet compliance requirements by ensuring the integrity of their data.

B. Objective

The objective of detecting replicated files in the cloud is to identify and eliminate duplicate copies of files or data within a cloud storage system. Replicated files can take up unnecessary storage space, lead to increased costs, and potentially cause synchronization issues or data inconsistencies.

C. Existing system.

The Problems with the existing systems are Integrity auditing and secure deduplication

D. Proposed System

In this paper, aiming at achieving data integrity and deduplication in cloud, we propose two secure systems namely SecCloud and SecCloud+.

1) SecCloud introduces an auditing entity with maintenance of a MapReduce cloud, which helps clients generate data tags before uploading as well as audit the integrityof data having been stored in cloud.

2) Besides supporting integrity auditing and secure deduplication, SecCloud+ enables the guarantee of file confidentiality.

3) We propose a method of directly auditing integrity on encrypted data.

E. Advantages

1) This design fixes the issue of previous work that the computational load at user or auditor is too huge for tag generation. For completeness of fine-grained, the functionality of auditing designed in SecCloud is supported on both block level and sector level. In addition, SecCloud also enables secure deduplication.

2) The challenge of deduplication on encrypted is the prevention of dictionary attack.

Our proposed SecCloudsystem has achieved both integrity auditing and file deduplication.

II. LITERATURE SURVEY

A. Proofs of Ownership in Remote Storage Systems

Proofs of ownership in remote storage systems are essential for ensuring the security and integrity of data stored remotely.
The main drawbacks of proofs of ownership in remote storage systems are managing cryptographic keys,scalability

B. Provable Data Possession at Untrusted Store.

Provable data possession (PDP) is a cryptographic technique that allows a client to store data on an untrusted server while ensuring the server possesses the correct data without actually retrieving it. This technique is useful in scenarios where the client wants to ensure the integrity of their data but does not fully trust the server or the communication channels between the client and server.
The drawbacks with the provable data possession are complexity , limited functionality and secure assumptions and dependence on the cloud provider.

III. SYSTEM MODEL

A. Modules

1) *Cloud Servers:* The Cloud Servers module is a component of cloud computing infrastructure that enables users to store and access data remotely over the internet. When it comes to detecting replicated files in the cloud, the Cloud Servers module can be used in several ways. One way is to implement a hashing algorithm that computes a unique hash value for each file stored in the cloud. This hash value can then be used to compare files and identify any duplicates.

2) *Data Users Module:* The Data Users module in cloud computing allows users to access and manage data stored in the cloud. To detect replicated files, the module provides tools like data search, retrieval, and manipulation. Users can use these tools to search for and compare files to identify duplicates, as well as delete or consolidate them.

3) *Auditor:* The Auditor module in cloud computing evaluates the security, reliability, and compliance of cloud services to ensure they meet industry standards. To detect replicated files, auditors perform compliance checks, analyze data for patterns that indicate duplicates, and perform risk assessments to identify and prevent data replication.

B. Algorithms

CP-ABE

CP-ABE (Ciphertext-Policy Attribute-Based Encryption) is a type of advanced encryption algorithm used to protect sensitive data in cloud computing environments.
In CP-ABE, a policy is associated with each encrypted file, and the decryption process depends on the attributes of the user requesting access to the file. The user's attributes must match the policy associated with the file in order to successfully decrypt it.

C. *Techniques*

1) *Hashing:* This technique involves generating a unique cryptographic hash value for each file in the cloud. Duplicate files will have the same hash value, making it easy to detect them.
2) *Data Deduplication:* This technique involves using algorithms to identify and eliminate duplicate data blocks within files, which can significantly reduce storage space and improve efficiency.
3) *Encryption:* Encryption is the process of transforming plain or readable data, often referred to as plaintext, into a coded or scrambled format, known as ciphertext. Encryption is used to protect the confidentiality and integrity of sensitive information by making it unreadable and unusable by unauthorized parties who may gain access to it.
4) *Decryption:* Decryption is the reverse process of encryption and involves transforming the ciphertext back into plaintext using the same encryption algorithm and key. The decryption process is only possible for authorized parties who have access to the secret key or password used to encrypt the data.

IV. RESULTS AND ANALYSIS

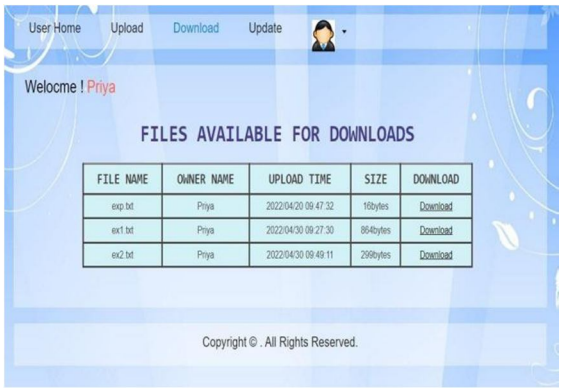
A. *Home Page*

B. *Register Page*

C. *Cloud User Login Page*

D. User Operations

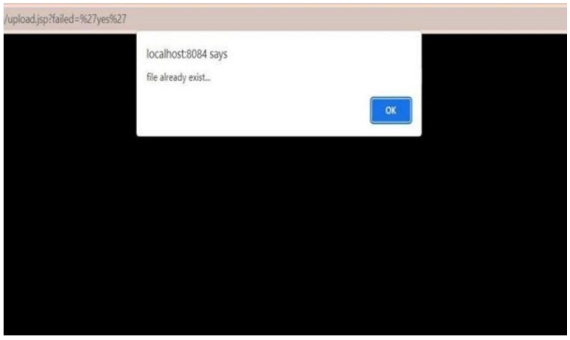
E. Uploading Files



F. Downloading Files

G. Encrypted Data

H. Deduplication



V. CONCLUSION

In conclusion, detecting replicated files in the cloud is an important task in managing cloud storage resources and ensuring data integrity. There are several techniques that can be used to detect replicated files in the cloud, including hashing, and data encoding. These techniques can be used alone or in combination with each other, depending on the specific needs of the project. It's important to consider factors such as storage capacity, data transfer costs, and computational requirements when choosing a detection technique. Regular monitoring and auditing of cloud storage can also help to identify and mitigate issues related to file replication. By implementing effective strategies for detecting replicated files, organizations can ensure the integrity and availability of their data in the cloud.

VI. FUTURE WORK

In this project we only done work with the text files in further development we can work on different files to store in the cloud server. So the system is developed to enhance the change by the requirements of the user, therefore these are opportunities and scope for future enhancement and upgrading are possible in this project. The project is flexible to adapt the changes efficiently without affecting the present system

REFERENCES

[1] Amazon, “Aws global infrastructure,” 2017. [Online]. Available: https://aws.amazon.com/about-aws/global-infrastructure/
[2] C. Metz, “Facebook tackles (really) big data with project prism,” 2012. [Online]. Available: https://www.wired.com/2012/08/facebook-prism/
[3] K. V. Shvachko, Y. Aahlad, J. Sundar, and P. Jeliazkov, “Geographically-distributed file system using coordinated namespace replication,” 2014. [Online]. Available:https://www.google.com/patents/WO2015153045A1?cl=zh
[4] C. Liao, A. Squicciarini, and L. Dan, “Last-hdfs: Location-aware storage technique for hadoop distributed file system,” in Proc. IEEE Int. Conf. Cloud Comput., 2016, pp. 662–669.
[5] N. Paladi and A. Michalas, ““One of our hosts in another country”: Challenges of data geolocation in cloud storage,” in Proc. Int. Conf. Wireless Commun. Veh. Technol. Inf. Theory Aerosp. Electron. Syst., 2014, pp. 1–6
[6] Z. N. Peterson, M. Gondree, and R. Beverly, “A position paper ondata sovereignty: The importance of geolocating data in the cloud,” in Proc. 3rd USENIX Conf. Hot Topics Cloud Comput., 2011, pp. 9–9.
[7] A. Squicciarini, D. Lin, S. Sundareswaran, and J. Li, “Policy driven node selection in MapReduce,” in Proc. 10th Int. Conf. Security Privacy Commun. Netw., 2015, pp. 55–72.
[8] J. Li, A. Squicciarini, D. Lin, S. Liang, and C. Jia, “Secloc: Securing location-sensitive storage in the cloud,” in Proc. ACM Symp. Access Control Models Technol., 2015, pp. 51–61.
[9] E. Order, “Presidential executive order on strengthening the cybersecurity of federal networks and critical infrastructure,” 2017. [Online]. Available: https://www.whitehouse.gov/the-pressoffifice/2017/05/11/presidential-executive-order strengthening cybersecurity-federal
[10] “Hdfs architecture,” (2018). [Online]. Available: http://hadoop. apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/ HdfsDesign.html
[11] R. Miller, “Inside amazon cloud computing infrastructure,” 2015. [Online]. Available: http://datacenterfrontier.com/inside-amazon_cloud-computing-infrastructure/
[12] T. Bujlow, K. Balachandran, S. L. Hald, M. T. Riaz, and J. M. Pedersen, “Volunteer-based system for research on the internet traffific,” Telfor J., vol. 4, no. 1, pp. 2–7, 2012
[13] M. Geist, “Location matters up in the cloud,” (2010). [Online]. Available: http://www.thestar.com/business/2010/12/04/geist_location_matters_up_in_the_cloud.html
[14] Z. N. Peterson, M. Gondree, and R. Beverly, “A position paper on data sovereignty: The importance of geolocating data in the cloud,” in Proc. 8th USENIX Conf. Netw. Syst. Design Implementation, 2011, pp. 9–9.

[15] K. Benson, R. Dowsley, and Shacham, "Do you know where your cloud files are?" in Proc. 3rd ACM Workshop Cloud Comput. Security Workshop, 2011, pp. 73–82.
[16] M. Gondree and Z. N. Peterson, "Geolocation of data in the cloud," in Proc. 3rd ACM Conf. Data Appl. Security Privacy, 2013, pp. 25–36.
[17] G. J. Watson, R. Safavi-Naini, M. Alimomeni, M. E. Locasto, and S. Narayan, "Lost: Location based storage," in Proc. ACM Workshop Cloud Comput. Security Workshop, 2012, pp. 59–70
[18] A. Albeshri, C. Boyd, and J. G. Nieto, "Geoproof: Proofs of geographic location for cloud computing environment," in Proc. 32nd Int. Conf. Distrib. Comput. Syst. Workshops, 2012, pp. 506–514.
[19] A. Albeshri, C. Boyd, and J. G. Nieto, "Enhanced geoproof: Improved geographic assurance for data in the cloud," Int. J. Inf. Security, vol. 13, no. 2, pp. 191–198, 2014.
[20] A. Michalas and K. Y. Yigzaw, "Locless: Do you really care where your cloud files are?" in Proc. IEEE Int. Conf. Cloud Comput. Technol. Sci., 2016, pp. 515–520.
[21] D. Lin, P. Rao, R. Ferrini, E. Bertino, and J. Lobo, "A similarity measure for comparing XACML policies," IEEE Trans. Knowl. Data Eng., vol. 25, no. 9, pp. 1946–1959, Sep. 2013.
[22] P. Rao, D. Lin, E. Bertino, N. Li, and J. Lobo, "Fine-grained integration of access control policies," Comput. Security, vol. 30, no. 2–3, pp. 91–107, 2011.

BIOGRAPHIES

Somu sowjanya [B.Tech], Student, Dept of Computer Science and Engineering, BWEC, Andhra Pradesh, India

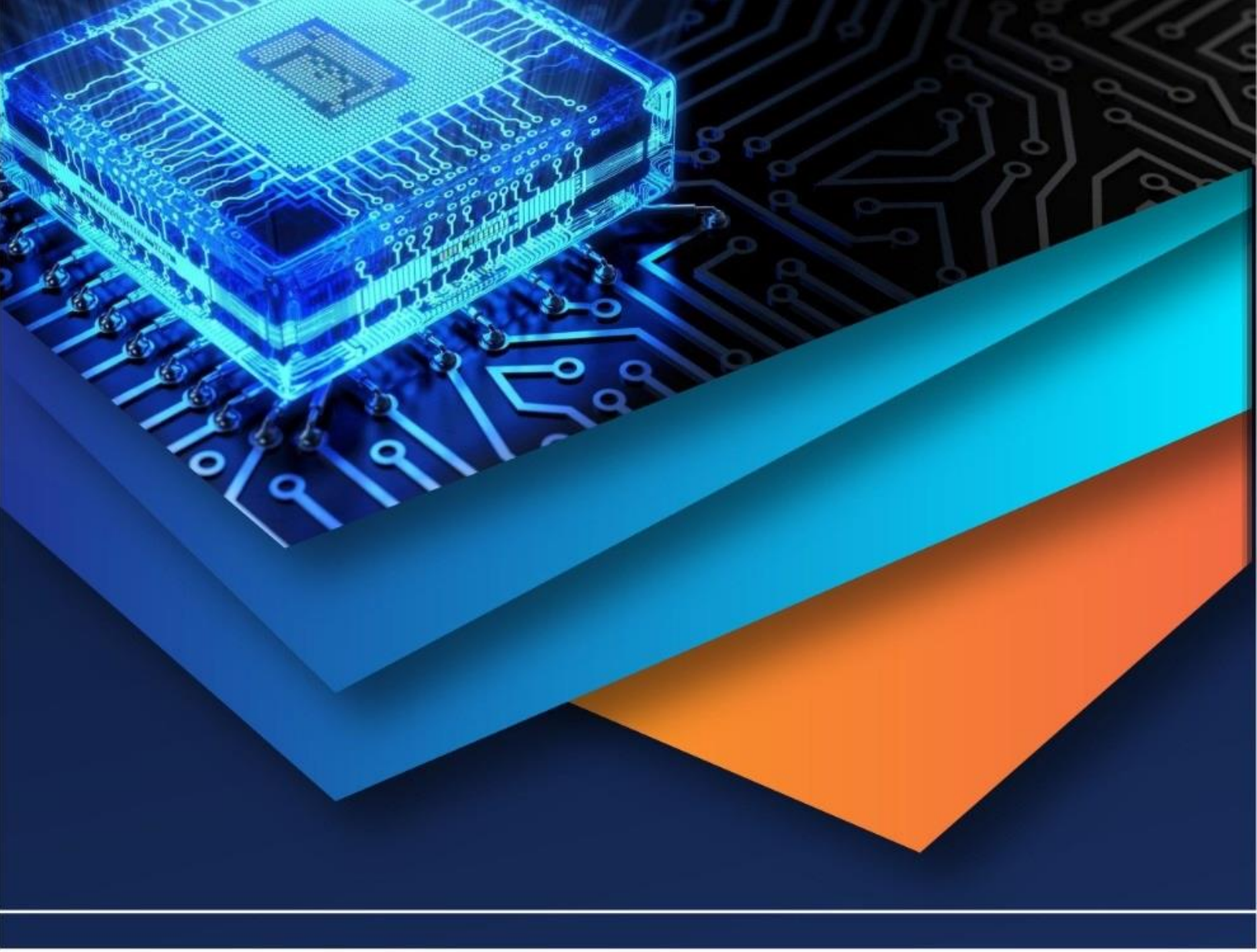
Sugguna Meenamrutha[B.Tech], Student, Dept of Computer Science and Engineering, BWEC, Andhra Pradesh, India

Kondi Sharanya Mounika [B.Tech], Student, Dept of Computer Science and Engineering, BWEC, Andhra Pradesh, India

Sugguna Hari Priyanka [B.Tech], Student, Dept of Computer Science and Engineering, BWEC, Andhra Pradesh, India

Marri Anvitha [B.Tech], Student, Dept of Computer Science and Engineering, BWEC, Andhra Pradesh, India

G. Venkateswari M. Tech, Head of the Department, Dept of Computer Science and Engineering, BWEC, Andhra Pradesh, India



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY