e-ISSN:2582 - 7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 4, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 5.928

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407019|

Secure Data Deduplication in Public Cloud Computing Environment

Miss. Dumbre Anita S[1], Prof. Monika D. Rokade. [2]

PG Student, Department of Computer, SPCOE, Dumbarwadi (Otur) Pune, India[1]

Assistant Professor (ME Co-coordinator), Department of Computer, SPCOE, Dumbarwadi (Otur) Pune, India[2]

ABSTRACT: an overview of cloud computing, cloud file services, their usability, and storage is given by this project. The de-duplication review of existing data de-duplication methods, procedures, and implementations for the benefit of cloud service providers and cloud users also considers storage optimization. The project also proposes a time-saving method for detecting and removing duplicates by calculating the digest of files using file checksum algorithms. This method suggests deleting duplicate data, but according to the duplication quest, the user has assigned some privileges, and each user has a unique token. Using the hybrid cloud model, cloud deduplication is accomplished. This proposed technique is more reliable and uses less cloud resources. It has also been shown that, in comparison to traditional deduplication techniques, the proposed scheme has a low overhead in duplicate removal. This paper examines both content and files level deduplication of file data in the cloud.

KEYWORDS: Data deduplication, Delta compression, Storage system, Index structure, Performance evaluation

I. INTRODUCTION

As shown by the production of approximately 1 terabyte and 2 terabytes of data in 2019 and 2020, the volume of digital data is increasing exponentially. As a result of these "data delegations," maintaining storage and lowering costs in a mass storage system may be one of the most challenging and critical tasks. Data replication is a data reduction technique that not only saves storage space by eliminating duplicate data, but also eliminates redundant data transmission in low-bandwidth networks. In recent years, data replication has grown in popularity as a highly effective data reduction method. Cloud computing is an evolving trend in information and communication technology for the modern century. The current study uses file data checksum extraction to minimize the time it takes to rule out false positives. The target file, on the other hand, stores a number of attributes, including the user id, filename, height, extension, checksum, and date-time table. Whenever a user uploads a file, the device first calculates the checksum, which is then compared to the checksum data stored in the database. If the file already exists, the record will be updated; otherwise, a new entry will be created in the database. Masters of data (owners), cloud servers (servers), and data customers (consumers) (users). Virtualization, distributed computing, networking, applications, and web services are all part of cloud computing. Clients, datacenters, and distributed servers are all components of a cloud. It includes features such as fault tolerance, high availability, scalability, versatility, lower user overhead, lower total cost of ownership, and on-demand services, among others. Data de-duplication is a method of detecting data duplication in storage space. Identifies data de-duplication strategies and eliminates non-unique data.

II. LITERATURE REVIEW

Kaiping Xue [1] propose a new heterogeneous architecture to solve the single-point performance bottleneck problem and provide a more robust access control scheme with an auditing mechanism Multiple attribute authorities are used in our system to distribute the burden of user legitimacy verification. Meanwhile, a CA (Central Authority) is implemented in our scheme to create hidden keys for users whose legitimacy has been tested. Unlike other multiauthority access control systems, ours handles the entire attribute collection individually for each authority. We also suggest an auditing mechanism to detect the AA (Attribute Authority) has conducted the validity verification procedure improperly or maliciously to improve protection.

Kan Yang and et. Al.[2], proposed a revocable multi-authority CP-ABE scheme, and use it to design the data access control scheme's underlying techniques. Both forward and backward protection can be achieved with ease using our attribute revocation tool. In multi-authority cloud storage systems, where multiple authorities coexist and each authority

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407019|

may issue attributes separately, the system often design an expressive, reliable, and revocable data access control scheme.

Zhongma Zhu and et. Al The system [3] proposed a secure method for anti-collusion key distribution that does not depend on third-party networks, and users can get their private keys from the group owner in a secure manner. Second, this approach can have fine-grained access control; any user in the community can access the cloud source, and revoked users cannot re-access the cloud after being revoked. Third, the mechanism will protect the scheme from collusion attacks, which ensures that even if revoked users merge with an untrusted cloud, they will not be able to access the actual data file. In this method, the system can complete a secure client negation conspire by using polynomial capability; finally, this plan can achieve fine performance, implying that past clients do not need to refresh their revoked from the community.

N. Attarpadung and et. Al [4] proposes the most important feature of the key-approach feature is that it is based on KP-ABE with non-monotonic access structures and standard cipher text size. The system also proposes the first Key-Policy Attribute-based Encryption (KPABE) approach that supports non-granted access structures (i.e., those with negated attributes) and has a constant cipher text size. To accomplish this, the framework first demonstrates that in the selective set model, a certain class of identity-based broadcast encryption schemes yields monotonic KPABE systems. The system then describes a new identity-based revocation mechanism that, when combined with a specific instance of our general monotonic construction, yields the first genuinely expressive KP-ABE realization with constant-size cipher text.

F. Zhang and K. Kim [5] proposed an Both methods are focused on bilinear pairings and the Java pairing library, and both are based on ID-based ring signatures. In addition, the system evaluates their security and performance in comparison to various existing strategies. For data encryption and decryption, the Java Pairing library (JPBC) was used. Some user access management policies are designed for end users while also protecting the data owner's privacy and confidentiality.

J. Han and et.al [6], propose The first Identity-based threshold ring signature method without java pairings. It proposes the first threshold verifiable ring signature technique based on identity. The method also examines whether the individual signers' privacy is preserved even though the Identity-based system's PK generator (PKG) is used. Finally, the device demonstrates how to incorporate identity collusion and other existing base schemes. The framework proposed in this paper actually form a suite of Identity based thresh-old ring signature methods, which are analogous to many real-world systems with varying degrees of signer inscrutability they support.
J. Yu and et. al [7], system first validates the security requirements of whole architecture, and after that adds to in the security architecture. System proposed AES 128 16 bit encryption approach for end to end user verification and data encryption/ decryption purpose.

Kan Yan [8], System proposed CP-ABE (Cipher text-Policy Attribute-based Encryption) is a promising method for controlling access to encrypted data. It necessitates the management of all attributes and the distribution of keys in the device by a trusted authority. Multiple authorities coexist in cloud storage environments, and each authority has the ability to issue attributes independently. Due to the inefficiency of decryption and revocation, current CP-ABE schemes cannot be explicitly extended to data access control for multi-authority cloud storage systems. In this paper, framework proposes DAC-MACS (Data Access Management for Multi-Authority Cloud Storage), an efficient decryption and revocation data access control scheme. In particular, the system develops a new multi-authority CP-ABE scheme with efficient decryption as well as an efficient attribute revocation method that provides both forward and backward protection.

Guangyan Zhang[9] the proposed CaCo is an effective Cauchy coding technique for cloud data storage. To begin, CaCo generates a matrix collection using Cauchy matrix heuristics. Second, CaCo generates a sequence of schedules for each matrix in this collection using XOR schedule heuristics. CaCo selects the shortest schedule from all the produced schedules in the second step. In this way, CaCo can find an ideal coding scheme for any given redundancy configuration that is within the capabilities of the current state of the art. CaCo is also implemented in the Cloud distributed file system, and its performance is compared to that of "Cloud 2.5." Finally, the author suggested that this method improve the security of distributed file systems by employing an efficient data storage scheme.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407019|

Ibrahim Adel [10] defines HDFS now has a new replica placement strategy. The problem of load balancing is addressed in this paper by distributing replicas equally among cluster nodes. As a result, there is no need for any load balancing software. The simulation results show that IDPM can produce replica distributions that are perfectly even and adhere to all HDFS replica placement laws. IDPM is intended for use in clusters where all cluster nodes have the same computing capabilities. The new proposal has a lot of potential for future work. HDFS replica placement policy Since data block replicas cannot be uniformly distributed across cluster nodes, HDFS currently relies on a load balancing utility to balance replica distributions, which takes more time and resources. These difficulties necessitate the creation of intelligent methods for resolving the data placement problem and achieving high efficiency without the use of a load balancing utility.

Monika Rokade and Yogesh Patil [11] proposed a system deep learning classification using nomaly detection from network dataset. The Recurrent Neural Network (RNN) has classification algorithm has used for detection and classifying the abnormal activities. The major benefit of system it can works on structured as well as unstructured imbalance dataset.

The MLIDS A Machine Learning Approach for Intrusion Detection for Real Time Network Dataset has proposed by Monika Rokade and Dr. Yogesh Patil in [12]. The numerous soft computing and machine learning classification algorithms have been used for detection the malicious activity from network dataset. The system depicts around 95% accuracy ok KDDCUP and NSLKDD dataset.
Monika D. Rokade and Yogesh Kumar Sharma [13] proposed a system to identification of Malicious Activity for Network Packet using Deep Learning. 6 standard dataset has sued for detection of malicious attacks with minimum three machine learning algorithms.

Sunil S. Khatal and Yogesh kumar Sharma [14] proposed a system Health Care Patient Monitoring using IoT and Machine Learning for detection of heart and chronic diseases of human body. The IoT environment has used for collection of real data while machine learning technique has used for classification those data, as it normal or abnormal.
Data Hiding In Audio-Video Using Anti Forensics Technique For Authentication has proposed by Sunil S.Khatal and Yogesh kumar Sharma [15]. This is a secure data hiding approach for hide the text data into video as well as image. Once sender hide data into specific objects while receivers does same operation for authentication. The major benefit of this system can eliminate zero day attacks in untrusted environments.
Sunil S.Khatal and Yogesh Kumar Sharma [16] proposed a system to analyzing the role of Heart Disease Prediction System using IoT and Machine Learning. This is the analytical based system to detection and prediction of heart disease from IoT dataset. This system can able to detect the disease and predict accordingly.

IV. PROPOSED SYSTEM DESIGN

The framework is proposed to include efficient de-duplication with system stability for file-level both block-level de-duplication, respectively, for safe de-duplication. When a user attempts to upload a file, our system performs a first-level replication scan. If file is duplicate it will be rejected by storage server and this check saves the space equal to file length. If file is not duplicate then file is divided into blocks of fixed size. Data is divided into fragments and stored at various nodes using secure secret sharing (RBAC) schemes. Before uploading this blocks block level duplication is performed. If the blocks are duplicate then these blocks are not uploaded to the server. This saves the amount of space equivalent to avoided duplicate blocks. The security of the system will be assessed in terms of two aspects: duplicate check authorization and data confidentiality. Convergent encryption, symmetric encryption, and the POW scheme are used to create the stable de-duplication scheme. Data protection is achieved by encrypting data before sending it to the storage server.

Data Deduplication
When a user wants to store the same data that has already been stored in the cloud, data replication occurs. The data owner verifies this by comparing hashes (tokens). It detects the duplicate file using hash comparison if the files are identical. When analyzing the data of each file for de-duplication, a file is a data content file, and it simply uses the hash value of the file as an identifier. If more than one file has the same hash value, they are expected to contain the same data information, which means a duplicate file has been created. When a new file is submitted by a user, the hash value must be checked first to ensure that only one specific data file is being stored.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407019|

System Architecture

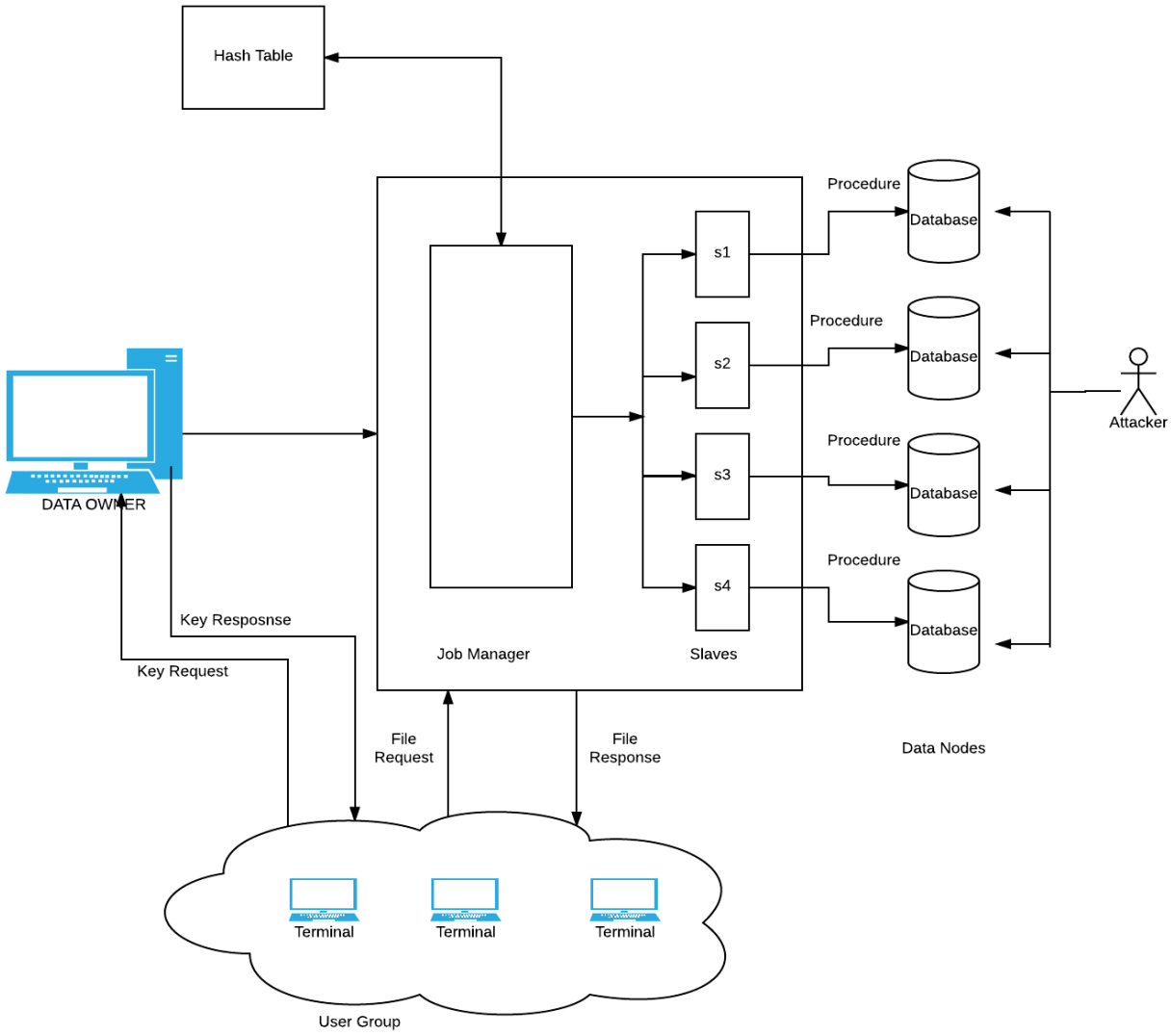

Fig.1. System Architecture

Secret Sharing Scheme:

Secret sharing scheme performs two operations namely Share and Recover. The secret is divided and shared by using Share. With enough shares, the secret can be extracted and recovered with the algorithm of Recover. The input to this module is file. It performs dividing of file into fixed size blocks or shares. These blocks are then encoded and allocated on cloud server at different nodes. When user request for file these blocks are decrypted and by combining these blocks file is given to user.

Tag Generation:

In this tag similarity is considered a kind of semantic relationship between tags, measured by means of relative co-occurrence between tags, known as J. coefficient. The input to this block is file blocks. This module assigns tags to each block for duplication check. The output of this module is blocks with tag assigned.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407019|

Mapreducemodule :
In this module system will process all the execution in parallel, we used one hash table at the time of data insertion once data has insert into database it will make a history into the hash table. For the efficient retrieval we can use the hash table.

Convergent Encryption Module
Traditional encryption, while providing data confidentiality, is incompatible with data de-duplication. Specifically, traditional encryption requires different users to encrypt their data with their own keys. Thus, identical data copies of different users will lead to different cipher texts, making de-duplication impossible.Convergent encryption has been proposed to enforce data confidentiality while making de-duplication feasible. It encrypts/ decrypts a data copy with a convergent key, which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since the encryption operation is deterministic and is derived from the data content, identical data copies will generate the same convergent key and hence the same cipher text.

V. CONCLUSION

To protect data confidentiality along with secure de-duplication, notion of authorized de-duplication is proposed. To carry duplicate check firstly privileges assigned to user are checked Instead of data itself duplicate check is based on differential privileges of users. In this paper, the issue of privacy preservation in cloud de-duplication is discussed, and an advanced structure supporting differentiated authorization and allowed duplicate check is proposed. This project addresses the issue in authorized de-duplication to achieve better security. We demonstrated that when comparison to convergent encryption and network transfer, our allowed duplicate check structure incurs minimal overhead

REFERENCES

[1] Xue K, Xue Y, Hong J, Li W, Yue H, Wei DS, Hong P. RAAC: Robust and auditable access control with multiple attribute authorities for public cloud storage. IEEE Transactions on Information Forensics and Security. 2017 Apr;12(4):953-67.
[2] Kan Yang and Xiaohua Jia, Expressive, E_cient, and Revocable Data Access Control for Multi-Authority Cloud Storage, IEEE Transactions on parallel and distributed systems, VOL. 25, NO. 07, July 2014.
[3] Zhongma Zhu and Rui Jiang proposed A Secure Anti-Collusion Data Sharing Scheme for Dynamic Groups in the Cloud in IEEE TRANSACTIONS ON PAR- ALLEL AND DISTRIBUTED SYSTEMS, VOL. 27, NO. 1, JANUARY 2016.
[4] N. Attarpadung, B. Libert, and E. Pana_eu, Expressive keypolicy attribute based encryption with constant-size ciphertexts, in 2011.
[5] F. Zhang and K. Kim. ID-Based Blind Signature and Ring Signature from Pairings. In ASIACRYPT 2002, volume 2501 of Lecture Notes in Computer Science, pages 533547. Springer, 2002.
[6] J. Han, Q. Xu, and G. Chen. E_cient id-based threshold ring signature scheme. In EUC (2), pages 437442. IEEE Computer Society, 2008.
[7] J. Yu, R. Hao, F. Kong, X. Cheng, J. Fan, and Y. Chen. Forward secure identity based signature: Security notions and construction. Inf. Sci., 181(3):648660, 2011
[8] Yang K, Jia X. DAC-MACS: E_ective data access control for multi-authority cloud storage systems. InSecurity for Cloud Storage Systems 2014 (pp. 59-83). Springer, New York, NY.
[9] Guangyan Zhang at. al. proposed CaCo: An Efficient Cauchy Coding Approach for Cloud Storage Systems in IEEE Feb 2016.
[10] Ibrahim Adel Ibrahim at. al. proposed Intelligent Data Placement Mechanism for Replicas Distribution in Cloud Storage Systems in 2016 IEEE International Conference on Smart Cloud.
[11] Monika D.Rokade ,Dr.Yogesh kumar Sharma,"Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic."IOSR Journal of Engineering (IOSR JEN),ISSN (e): 2250-3021, ISSN (p): 2278-8719
[12] Monika D.Rokade ,Dr.Yogesh kumar Sharma"MLIDS: A Machine Learning Approach for Intrusion Detection for Real Time Network Dataset", 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407019|

[13]Monika D.Rokade, Dr. Yogesh Kumar Sharma. (2020). Identification of Malicious Activity for Network Packet using Deep Learning. *International Journal of Advanced Science and Technology*, *29*(9s), 2324 - 2331.
[14] Sunil S.Khatal ,Dr.Yogesh kumar Sharma, "Health Care Patient Monitoring using IoT and Machine Learning.", **IOSR Journal of Engineering (IOSR JEN),** ISSN (e): 2250-3021, ISSN (p): 2278-8719
[15]Sunil S.Khatal ,Dr.Yogesh kumar Sharma, "Data Hiding In Audio-Video Using Anti Forensics Technique ForAuthentication ", IJSRDV4I50349, Volume : 4, Issue : 5
[16]Sunil S.Khatal Dr. Yogesh Kumar Sharma. (2020). Analyzing the role of Heart Disease Prediction System using IoT and Machine Learning. *International Journal of Advanced Science and Technology*, *29*(9s), 2340 - 2346.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 5.928

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

9710 583 466 9710 583 466 ijmrset@gmail.com

www.ijmrset.com