Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

# DATA DEDUPLICATION SCHEME WITH INTEGRITY WITH ENHANCED SECURITY USING IDENTITY BASED TECHNIQUE

LAVANYA MAYALURU[1], B.CHARISHMA[2]

[1]PG SCHOLAR (M.Tech., (CSE)), Dept of CSE, Srinivasa Institute of Technology And Science, Kadapa

[2]Assistant Professor, HOD, Dept of CSE, Srinivasa Institute of Technology And Science, Kadapa

**ABSTRACT:**

Cloud computing enables new business models and cost effective resource usage. In Cloud Computing Technology Data Storing and Data Sharing plays a major role. In Data Storing we face a main problem of Data deduplication. Various traditional deduplication systems are introduced for elimination of replicate check besides the data itself, but existing techniques are not able to decode compressed files. The proposed architecture provides duplicate check procedures to reduce minimal overhead compared to normal operations. The data stored in cloud will be in compressed format the paper introduces decoding data compression techniques for eliminating duplicate copies of repeating data, through this cloud storage space and upload and download bandwidths can be reduced. The work also presents various new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture. Security analysis exhibit that our scheme is protected in terms of the description particular in the projected security model. The work realize a prototype of proposed approved duplicate check scheme and carry out tested experiments by means of the prototype. We show that our planned authorized replacement check scheme incurs negligible transparency evaluate to normal operations for elimination of duplicate data from clouds.

Index Terms- Cloud Computing, Deduplication, Duplicate Removal, Hybrid Cloud and Secure Authorization.

## 1. INTRODUCTION

In Emerging Technologies like Cloud Computing make available various resource usages using central architecture. Cloud service supplier in today's technology offering together extremely obtainable storage and particularly similar computing reserve at comparatively low costs. As low cost and effective technology there is tremendous increase of data storage and Usage with various specified privileges. Main critical challenge in this cloud storage services is the ever-increasing volume of data and controlling duplication of data storage. Data deduplication is a specialized data firmness technique for duplicate copies of go over data in storage. Fig 1 shows the architecture of Cloud Resources.


**Figure 1.1: Architecture of Cloud Computing**

Duplicate Data uploading may occur at various levels it may take place at moreover the file level or the block level for file level deduplication, it eliminates duplicate copies of the same file. Various old encryption Techniques, while providing data discretion is mismatched with data deduplication. Specially, established encryption necessitate dissimilar users to encrypt their data with their own keys. Thus, identical data copies of unusual users will lead to dissimilar cipher texts making deduplication

unfeasible. Convergent encryption that has planned to enforce data discretion whereas creating deduplication probable. It encrypts/decrypts a data copy with a Convergent key, which is attained by computing the cryptographic hash value that is satisfied of the data copy. After key production and data encryption, users maintain the keys and send the cipher text to the cloud. Since the encryption operation is deterministic and is derivative from the data content, identical data copies will produce the same convergent key and hence the same cipher text. A Hybrid Cloud is a mutual form of private clouds and public clouds wherein some critical data inhabit in the enterprise's private cloud even as further data is stored in and Accessible from a public cloud. Hybrid clouds seek to distributed the benefit of scalability, dependability, speedy consumption and possible cost savings of public clouds with the security and enlarged control and supervision of private clouds.

The critical challenge of cloud storage or cloud computing is the management of the continuously growing volume of data. Data deduplication or Single Instancing fundamentally submits to the exclusion of redundant data. However, indexing of all data is still retained should that data ever be required. In general the data deduplication eradicates the duplicate copies of repeating data.

## II. RELATED WORK

In archival storage systems, there is a enormous quantity of duplicate data or unneeded data, which inhabit important extra equipments and power utilizations, largely lowering down resources utilization (such as the network bandwidth and storage) and imposing extra burden on management as the scale increases. So data de-duplication, the goal of which is to decrease the duplicate data in the inter level has been getting broad concentration both in intellectual and industry in recent years. In this paper, semantic data de-duplication (SDD) is designed, which is having the semantic information in the I/O path (such as file type, file format, application hints and system metadata) of the archival files to straight the separating a file into semantic chunks (SC). The following papers are studied to know assorted techniques and procedures which were used for duplicate exclusion.

"**A secure cloud backup system with assured deletion and version control**. **A. Rahumed**", **H. C. H. Chen, Y. Tang, P.P. C. Lee, and J. C. S.Lui** [1], Has presented Cloud storage is an emerging service model that facilitate exclusive and enterprises to outsource the storage of data backups to remote cloud provider at a low cost. Hence results shows that FadeVersion only adds negligible presentation overhead with a traditional cloud backup service that does not carry out secured removal. "**A reverse deduplication storage system** optimized for reads to latest backups", C. Ng and P. Lee. **Revdedup**

[2] Had present RevDedup, a de-duplication system designed for VM disk image backup in virtualization environments. RevDedup has various design goals: high storage competence, low memory usage, high backup presentation, and high return presentation for newest backups. They lengthily appraise our RevDedup prototype by means of dissimilar workloads and validate our design goals. **"**Role-based access controls**"**, D. Ferraiolo and R. Kuhn **[3],has described the Mandatory Access Controls (MAC) are** suitable for multilevel secure military applications, Discretionary Access Controls (DAC) are frequently apparent as meeting the security processing needs of industry and civilian government. "**Secure deduplication with efficient** and reliable convergent key management", J. Li, X. Chen, **M. Li, J. Li, P. Lee, andW. Lou** [4], had planned Dekey, an capable and dependable convergent key management scheme for secure de-duplication. They execute Dekey using the Ramp secret sharing scheme and make obvious that it incurs small encoding/decoding overhead evaluated to the network transmission below in the usual upload/download operations." **Reclaiming space from duplicate files in a** server less distributed file system", J. R. Douceur, A. **Adya, W. J. Bolosky, D. Simon, and M. Theimer.**[5], Has obtainable the Farsite distributed file system make available accessibility by replicating every file onto several desktop computers. Measurement of over 500 desktop file systems shows that almost half of all extreme space is engaged by replica files. The mechanism includes 1) convergent encryption, which make easy replacement files to come together into the space of a single file, even if the files are encrypted with dissimilar users' keys, and 2) SALAD, a Self Arranging, Lossy, Associative Database for aggregating file satisfied and location

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

in sequence in a decentralized, scalable, fault-tolerant manner. "**A secure data deduplication scheme for cloud storage**", **J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl** [6],has make available the private users outsource their data to cloud storage suppliers, recent data breach incidence that make end-to-end encryption an increasingly prominent requirement data deduplication can be efficient for accepted data, whilst semantically protected encryption protects detested satisfied. "**Weak leakage-resilient client-side** deduplication of encrypted data in cloud storage", J. **Xu, E.-C. Chang, and J. Zhou** [7], by describing the protected client-side deduplication scheme with the following benefits: our scheme benefit data improvement (and *some* partial information) alongside together exterior adversaries and honest-but-interested cloud storage server, even as Halevi *et al.* trusts cloud storage server in data secrecy. "**Secureand constant cost public cloud storage auditing with deduplication**", **J. Yuan and S. Yu**[8] Data integrity and storage capability are two significant compulsion for cloud storage. The author planned scheme is also distinguish by stable realtime statement and computational cost on the user side. "**Privacy aware data intensive computing on hybrid clouds**", **K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan** [9] has planned, the exterior of cost-effective cloud services offers organizations great instance to reduce their cost and enlarge efficiency. The system, called Sedic, leverages the special features of Map Reduce to frequently partition a computing job with the security levels of the data it works. "Gq and schnorr classification schemes Proofs of **security against impersonation under active and concurrent attacks**", **M. Bellare and A. Palacio** [10] has make available, the proof for GQ sustained on the not mentioned security of RSA under one more inversion, an calculation of the common one way statement that was introduced. Both results make bigger to conclude security against thought beneath immediate attack**.**

## III.OVERVIEW OF THE HYBRID CLOUD CONCEPTS HYBRID CLOUD

A hybrid cloud is a cloud computing environment in which an organization supply and supervise some resources in-house and has others make available superficially. For example, an organization may use a public cloud service, such as Amazon Simple Storage Service(Amazon S3) for archived data but continue to preserve in house storage for prepared customer data


The concept of a hybrid cloud is meant to bridge the gap among high domination, high cost "private cloud" and extremely callable, adaptable, low cost "public cloud". "Private Cloud" is normally used to describe a VMware deployment in which the hardware and software of the environment is used and managed by a single entity.

The concept of a "Public cloud" typically occupy some form of elastic/subscription based reserve pools in a hosting supplier datacenter that make use of multi-tenancy. The term public cloud doesn't mean less security, but instead refers to multi-tenancy. The thought turn around very much concerning connectivity and data portability. The use cases are numerous: resource burst-ability for regular demand, expansion and testing on a uniform platform lacking consuming local resources, disaster recovery, and of course excess capacity to make better use of or free up local consumption.

VM ware has a key tool for "hybrid cloud" use called "vCloud connector". It is afree plugin that allows the management of public and private clouds within the vSphere client. The tool offers users the capability to supervise the comfort view, power status and more from a "workloads" tab and offers the ability to copy virtual machine templates to and from a remote public cloud offering.

## VI.METHODOLOGY

The notion of authorized data deduplication was proposed to defend the data security by together with disparity privileges of users in the duplicate check. We also presented various new deduplication constructions behind authorized replacement check in hybrid cloud architecture, in which the duplicate

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

check tokens of files are generated by the private cloud serve with private keys. Security analysis exhibit that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of notion, we implemented a prototype of our projected authorized duplicate check scheme and carry out experiments on our prototype. We demonstrate that our authorized duplicate check scheme incurs smallest below evaluate to convergent encryption and network move.

## 1. Methodology

The basic purpose of this work is the problem of privacy preserving deduplication in cloud computing and a proposed System focus on these aspects:

1) **Differential Authorization:** every authorized consumer is able to get his/her unit token of his file to execute duplicate check based on his freedom.
2) **Authorized Duplicate Check:** Authorized user is able to use his/her individual private keys to generate query for certain file and the privileges he/she owned with the help of private cloud, while the public cloud performs duplicate check directly and tells the user if there is any duplicate.

**Proposed System**

In Proposed system, Convergent encryption has been used to enforce data confidentiality. Data copy is encrypted below a key beneath by confusion the data itself. This convergent key is used for encrypt and decrypt a data copy. Moreover, such not permitted users cannot decrypt the cipher text even conspire with the S-CSP (storage cloud service provider). Security analysis make obvious that that system is secure in terms of the description particular in the planned security model.

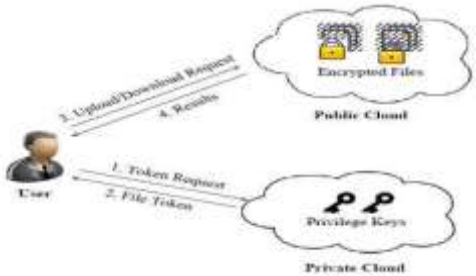



**Figure 1:** Architecture for Authorized **Deduplication**

This work known a company by where the employee data such as name, password, email id, contact number and designation is registered by admin or owner of the company based on his userid and password employees of the company able to perform operations such as file upload download and duplicate checks on the files based on his privileges.

There are three entities define in hybrid cloud architecture of authorized deduplication.

- **Data Users:** A user is an being that wants to outsource data storage to the S-CSP(storage cloud service provider) and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users. Each file is confined with the convergent encryption key and privilege keys to understand the authorized deduplication with discrepancy privileges.
- **Private Cloud:** This is new entity for facilitating users secure use of cloud services. The private keys for privileges are managed by private cloud, which provides the file token to users. Specifically, since the computing resources at data user/owner side are controlled and the public cloud is not fully trusted in carry out, private cloud is able to provide data user/owner with an finishing situation and infrastructure working as an interface among user and the public cloud.
- **S-CSP(storage cloud service provider):**This is an entity that provides a data storage service in public cloud. The S- CSP make available the data outsourcing service and stores data in support of the users. To decrease the storage cost, the S-CSP reducing the storage of redundant data via deduplication and keeps only unique data. In this paper, we assume that S-CSP is always online and has abundant storage capacity and computation power.

Our completion of the **Client** provides the following function calls to support token generation and deduplication along the file upload process.

- FileTag(File) - It computes SHA-1 hash of the File as File Tag;
- TokenReq(Tag, UserID) - It requests the Private Server for File Token generation with the File Tag and User ID;
- DupCheckReq(Token) - It requests the Storage

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

Server for Duplicate Check of the File by sending the file token received from private server;

- ShareTokenReq(Tag, {Priv.}) - It requests the Private Server to generate the Share File Token with the File Tag and Target Sharing Privilege Set;
- FileEncrypt(File) - It encrypts the File with Convergent Encryption using 256-bit AES algorithm in cipher block chaining (CBC) mode, where the convergent key is from SHA-256 Hashing of the file;
- FileUploadReq(FileID, File, Token) – It uploads the File Data to the Storage Server if the file is Unique and updates the
- File Token stored. Our completion of the Private Server includes matching request handlers for the token production and retain a key storage with Hash Map.
- TokenGen(Tag, UserID) - It loads the connected privilege keys of the user and produce the token with HMAC-SHA-1 algorithm

## V.RESULTS

Our evaluation focuses on comparing the overhead induced by authorization steps, including file token generation and share token generation, beside the convergent encryption and file upload steps. We appraise the overhead by unreliable various factors, together with 1) File Size 2) Number of Stored Files 3) Deduplication Ratio 4) Privilege Set Size. We break down the upload process into 6 steps, 1) Tagging 2) Token Generation 3) Duplicate Check 4) Share Token Generation 5) Encryption 6) Transfer . For each step, we record the start and end time of it and therefore obtain the breakdown of the total time spent. We present the regular time taken in every data set in the figures

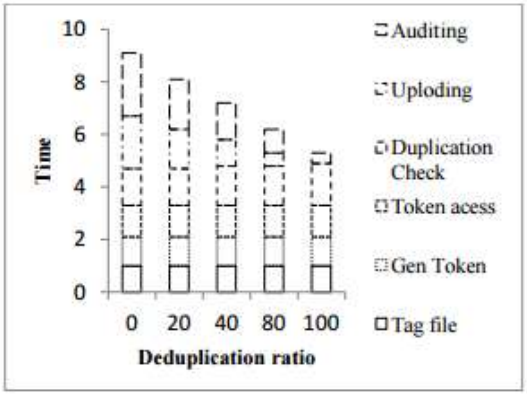


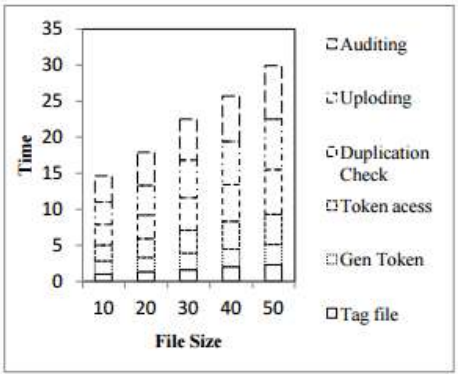

**Fig 2. Crack UP Time for Different Duplication Ration**

### File Size

To appraise the consequence of file size to the time spent on various steps, we upload 100 unique files (i.e., without any deduplication opportunity) of particular file size and record the time break down. Using the unique files enables us to evaluate the worst-case scenario where we have to upload all file data. The average time of the steps from test sets of different file size are plotted in Figure 2. The time spent on tagging, encryption, upload enlarge linearly with the file size, since these operations involve the actual file data and incur file I/O with the whole file.

**Number of Stored Files**

To evaluate the effect of number of stored files in the system, we upload 10000 10MB unique files to the system and record the breakdown for every file upload. From Figure 3, every step remains constant along the time. Token checking is done with a hash table and a linear search would be carried out in case of collision.




### Deduplication Ratio

To appraise the consequence of the deduplication ratio, we prepare two unique data sets, each of which consists of 50 100MB files. We first upload the first set as an initial upload. For the second upload, we pick a portion of 50 files, through given deduplication ratio from the initial set as duplicate files and remaining files from the second set as unique files. The average time of uploading the second set is presented in Figure 4.

Privilege Set Size

**Fig3.Crack UP Time for Different file size**

To evaluate the effect of privilege set size, we upload 100 10MB unique files with different size of the data owner and target share privilege set size. In Figure 5, it shows by taking token generation increases linearly as more keys are associated with the file and also the duplicate check time. While the number of keys increases 100 times from 1000 to 100000, the total time spent only increases to 3.81 times and it is noted that the file size of the experiment is set at a small level (10MB), the effect would become less significant in case of larger files.

**Conclusion and Future**

In this proposed architecture we have designed a new notion for removing data deduplication and to protect the data security through privileges of users and duplicate check. We had perform various new deduplication constructions behind authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are produced by the private cloud server with private keys. As the notion in this project we realize a prototype of our considered authorized duplicate check scheme and conduct test bed experiments on our prototype. From this project we show that our sanctioned duplicate check scheme acquire negligible overhead balance to convergent encryption and network relocate.

## References

[1] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In *Proc. of APSYS*,Apr 2013.

[2] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with efficient and reliable convergent key management. In *IEEE Transactions on Parallel and Distributed Systems*, 2013.

[3] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure datadeduplication scheme for cloud storage. In *Technical Report*, 2013.

[4] J. Xu, E.-C. Chang, and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In *ASIACCS*,pages 195–206, 2013.

[5] J. Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication. *IACR Cryptology ePrint Archive*, 2013:149, 2013.

[6] K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan. Sedic: privacy aware data intensive computing on hybrid clouds. In *Proceedings of the 18th ACM conference on Computer and communications security*,CCS'11, pages 515–526, New York, NY, USA, 2011. ACM.

[7] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In *USENIX Security Symposium*, 2013.

[8] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message- locked encryption and secure deduplication. In *EUROCRYPT*, pages 296–312, 2013.

[9] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. Secure data deduplication. In *Proc. of StorageSS*, 2008.

[10] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In *Proc. of USENIX LISA*, 2010.

[11] R. D. Pietro and A. Sorniotti. Boosting efficiency and security in proof of ownership for deduplication. In H. Y. Youm and Y. Won, editors, *ACM Symposium on Information, Computer and Communications Security*, pages 81–82. ACM, 2012.

[12] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In *Workshop on Cryptography and Security in Clouds (WCSC2011)*, 2011.

[13] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, *ACM Conference on Computer and Communications Security*, pages 491–500. ACM, 2011.

[14] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 441–446. ACM, 2012.

[15] Z. Wilcox-O'Hearn and B. Warner. Tahoe: the least- authority filesystem. In *Proc. of ACM StorageSS*, 2008.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023