e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 5, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.54

International Journal Of Multidisciplinary Research In Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506057 |

Data Deduplication for Efficient Storage Based on File Scanning

Kavitha M[1], Midhunkumar K[2], Shankar R R[3], Surya S[4]

Assistant Professor, Department of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, Salem, TamilNadu, India[1]

Student, Department of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, Salem, TamilNadu, India[2,3,4]

ABSTRACT*:* Data deduplication is a technique for eliminating duplicate copies of data, and has been widely used in cloud storage to reduce storage space and upload bandwidth. However, there is only one copy for each file stored in cloud even if such a file is owned by a huge number of users. As a result, deduplication system improves storage utilization while reducing reliability. Furthermore, the challenge of privacy for sensitive data also arises when they are outsourced by users to cloud. Aiming to address the above security challenges, this paper makes the first attempt to formalize the notion of distributed reliable deduplication system. We propose new distributed deduplication systems with higher reliability in which the data chunks are distributed across multiple cloud servers. The security requirements of data confidentiality and tag consistency are also achieved by introducing a deterministic secret sharing scheme in distributed storage systems, instead of using convergent encryption as in previous deduplication systems. Security analysis demonstrates that our deduplication systems are secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement the proposed systems and demonstrate that the incurred overhead is very limited in realistic environments.

KEYWORDS:

VPN Virtual Private Network
NaaS Network as a Service
CaaS Communication as a Service
VNO Virtual Network Operator
CSP Cloud Service Provider

I. INTRODUCTION

Cloud computing is a type of computing that relies on *sharing computing resources* rather than having local servers or personal devices to handle applications. In cloud computing, the word cloud (also phrased as "the cloud") is used as a metaphor for "*the Internet,"* so the phrase *cloud computing* means "a type of Internet-based computing," where different services -- such as servers, storage and applications -- are delivered to an organization's computers and devices through the Internet.

Cloud computing is comparable to grid computing, a type of computing where unused processing cycles of all computers in a network are harnesses to solve problems too intensive for any stand-alone machine. Cloud computing is an on-demand service that is obtaining mass appeal in corporate data centers. The cloud enables the data center to operate like the Internet and computing resources to be accessed and shared as virtual resources in a secure and scalable manner. Like most technologies, trends start in the enterprise and shift to adoption by small business owners.

In its most simple description, cloud computing is taking services ("cloud services") and moving them outside an organizations firewall on shared systems. Applications and services are accessed via the Web, instead of your hard drive. In cloud computing, the services are delivered and used over the Internet and are paid for by cloud customer (your business) -- typically on an "as-needed, pay-per-use" business model. The cloud infrastructure is maintained by the cloud provider, not the individual cloud customer.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506057 |

Currently, the standards for connecting the computer systems and the software needed to make cloud computing work are not fully defined at present time, leaving many companies to define their own cloud computing technologies.

II. LITERATURE

Similarity and Locality Based Indexing for High Performance Data Deduplication- IEEE Transactions on Computers (Volume: 64, Issue: 4, April 2015).

Try Managing Your Deduplication Fine-Grained-ly: A Multi-tiered and Dynamic SLA-Driven Deduplication Framework for Primary Storage -2016 IEEE 9th International Conference on Cloud Computing (CLOUD).

HPDV:A Highly Parallel Deduplication Cluster for Virtual Machine Images - 2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID).

Secure Textual Data Deduplication Scheme Based on Data Encoding and Compression - 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).

Achieving Efficient and Privacy-Preserving Multi-Domain Big Data Deduplication in Cloud - IEEE Transactions on Services Computing (Volume: 14, Issue: 5, Sept.-Oct. 1 2021).

.

III. METHODOLOGY

- The goal of cloud computing is to apply traditional supercomputing, or high-performance computing power, normally used by military and research facilities, to perform tens of trillions of computations per second, in consumer-oriented applications such as financial portfolios, to deliver personalized information, to provide data storage or to power large, immersive computer games.
- The standards for connecting the computer systems and the software needed to make cloud computing work are not fully defined at present time, leaving many companies to define their own cloud computing technologies. Cloud computing systems offered by companies, like IBM's "Blue Cloud" technologies for example, are based on open standards and open source software which link together computers that are used to to deliver Web 2.0 capabilities like mash-ups or mobile commerce.
- Cloud computing has started to obtain mass appeal in corporate data centers as it enables the data center to operate like the Internet through the process of enabling computing resources to be accessed and shared as virtual resources in a secure and scalable manner. For a small and medium size business (SMB), the benefits of cloud computing is currently driving adoption. In the SMB sector there is often a lack of time and financial resources to purchase, deploy and maintain an infrastructure (e.g. the software, server and storage).
- In cloud computing, small businesses can access these resources and expand or shrink services as business needs change. The common pay-as-you-go subscription model is designed to let SMBs easily add or remove services and you typically will only pay for what you do use.

A. Software as a Service(SaaS)

In the business model using software as a service (SaaS), users are provided access to application software and databases. Cloud providers manage the infrastructure and platforms that run the applications. SaaS is sometimes referred to as "on-demand software" and is usually priced on a pay-per-use basis. SaaS providers generally price applications using a subscription fee.

B. Networkas a service(Naas)

A category of cloud services where the capability provided to the cloud service user is to use network/transport connectivity services and/or inter-cloud network connectivity services. Naas involves the optimization of resource allocations by considering network and computing resources as a unified whole. Traditional Naas services include flexible and extended VPN, and bandwidth on demand. Naas concept materialization also includes the provision of a virtual network service by the owners of the network infrastructure to a third party (VNP – VNO).

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506057 |

C. Private Cloud

Private cloud is cloud infrastructure operated solely for a single organization, whether managed internally or by a third-party and hosted internally or externally. Undertaking a private cloud project requires a significant level and degree of engagement to virtualize the business environment, and requires the organization to re-evaluate decisions about existing resources. When done right, it can improve business, but every step in the project raises security issues that must be addressed to prevent serious vulnerabilities. Self-run data centres are generally capital intensive. They have a significant physical footprint, requiring allocations of space, hardware, and environmental controls. These assets have to be refreshed periodically, resulting in additional capital expenditures.

D. Public Cloud

A cloud is called a "public cloud" when the services are rendered over a network that is open for public use. Technically there may be little or no difference between public and private cloud architecture, however, security consideration may be substantially different for services (applications, storage, and other resources) that are made available by a service provider for a public audience and when communication is effected over a non-trusted network. Generally, public cloud service providers like Amazon AWS, Microsoft and Google own and operate the infrastructure and offer access only via Internet (direct connectivity is not offered).

E. Deduplication Data Flow

Deduplicate data and where the process will take place, there are also different approaches taken to when the operation is carried out. When the first deduplication appliances were developed, most hardware platforms could not bring data into the system and deduplicate it inline fast enough to keep up with the requirements for midrange and Enterprise data protection. The early pure inline systems were limited to relatively small systems, and higher performance was achieved by devices that allowed users to move deduplication out of the backup window. Some systems were designed to use a fully deferred post-processing data flow that brings all of a user's data into the appliance during the ingest window, and then carries out deduplication at a later time.
These systems were designed to provide faster ingest capability, with the tradeoff of requiring extra disk space for a landing area and delaying replication operations. Some post process operations could also extend the total data protection time—in other words, the time to ingest, replicate, and dedupe might take longer than a more concurrent process.

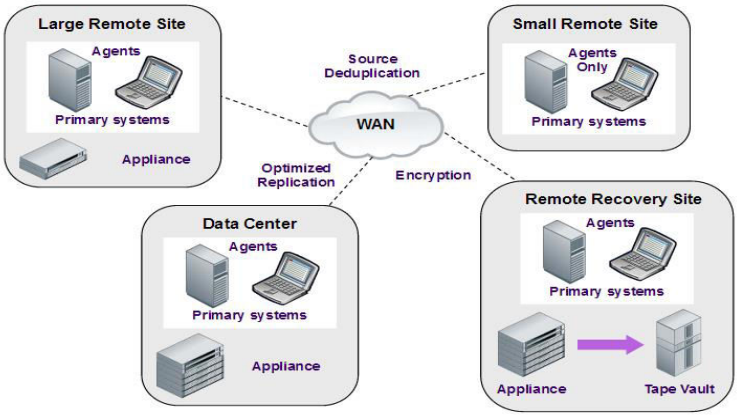

Fig 1 Backup Remote Office Source Deduplication

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506057 |

IV. RESULT

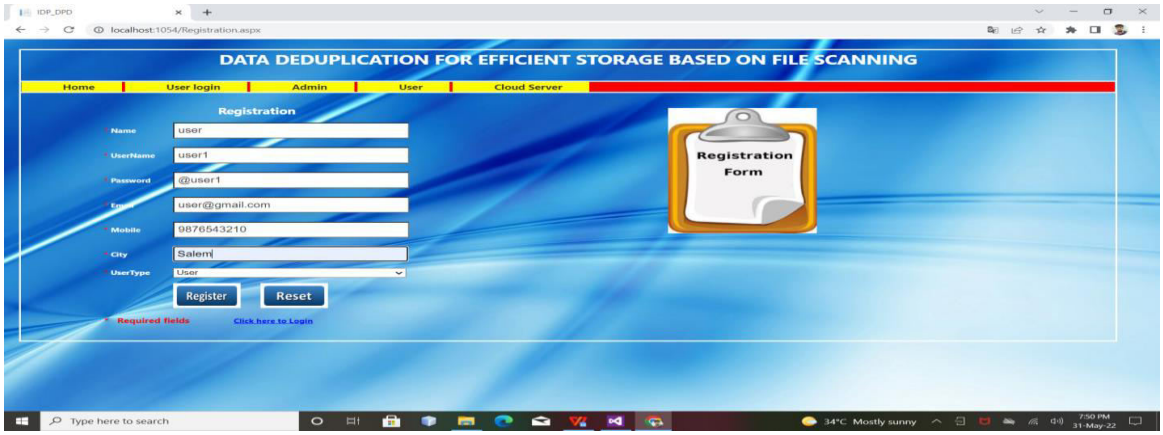

Fig 2 New Registration

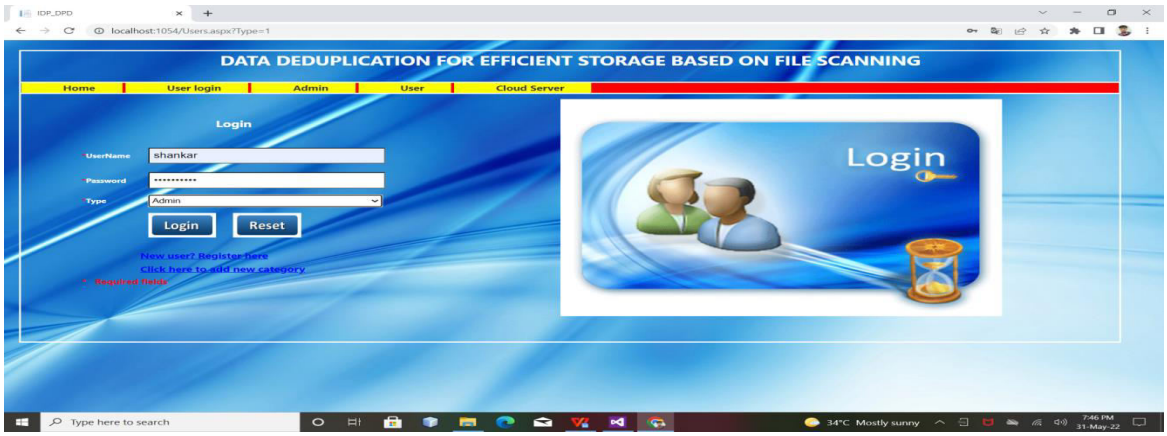

Fig 3 Login Page

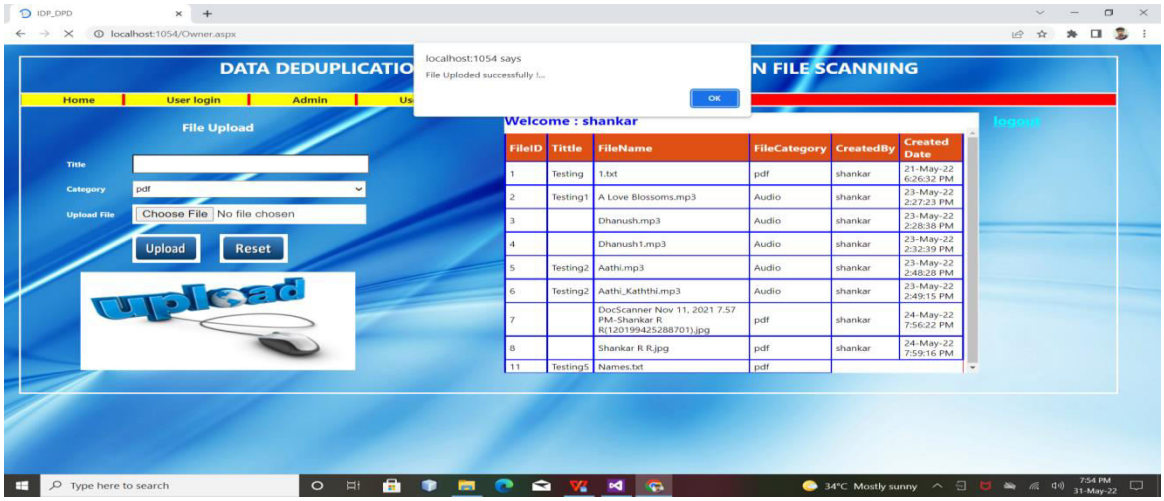

Fig 4 File uploading page

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506057 |

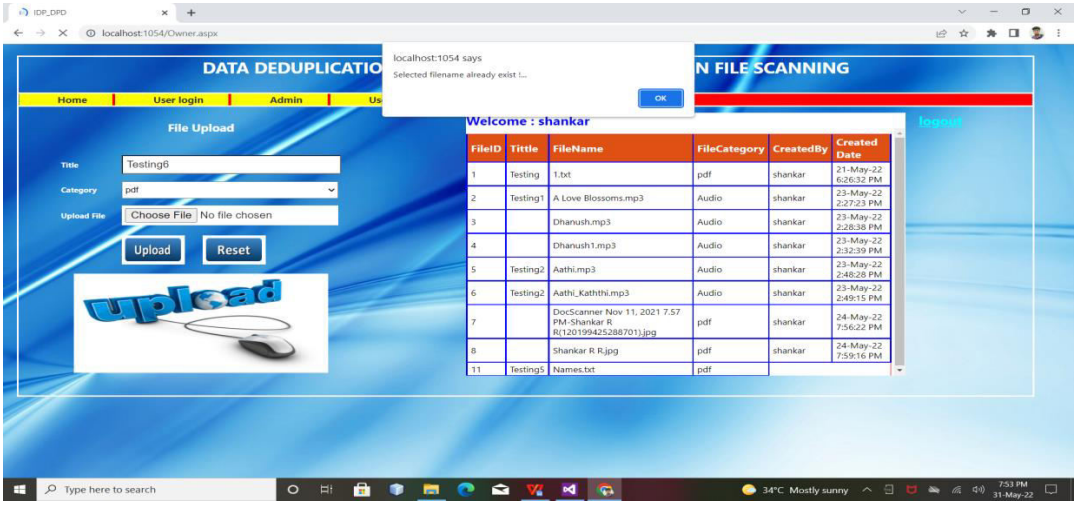

Fig 5 File Name Scanning

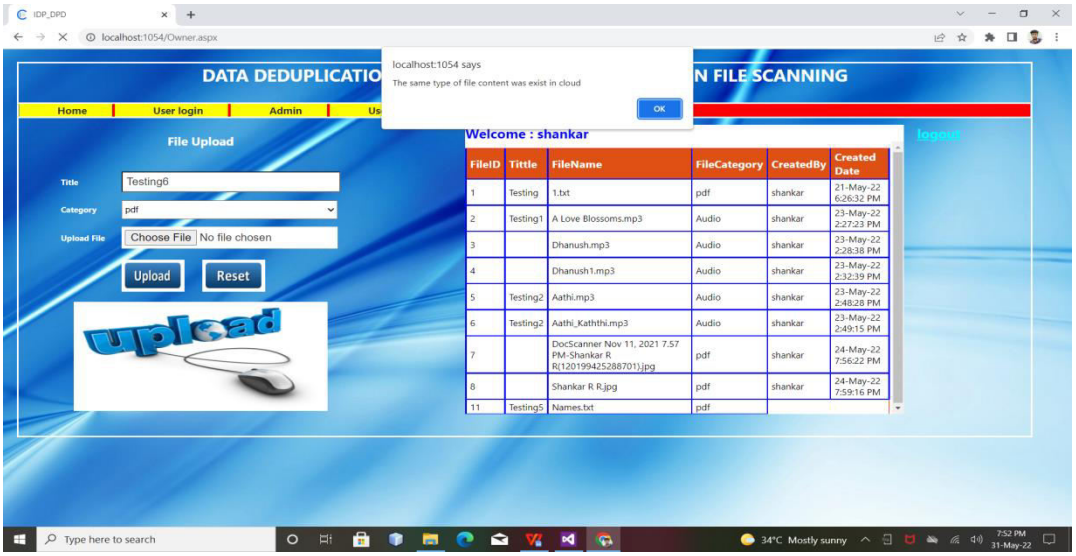

Fig 6 File Content Scanning

V. CONCLUSIONS

The distributed reduplication systems to improve the reliability of data while achieving the confidentiality of the users' outsourced data without an encryption mechanism. Four constructions were proposed to support file-level and fine-grained block-level data deduplication. The security of tag consistency and integrity were achieved. We implemented our deduplication systems using the Ramp secret sharing scheme and demonstrated that it incurs small encoding/decoding overhead compared to the network transmission overhead in regular upload/ download operations.

REFERENCES

[1] M. O. Rabin, "Fingerprinting by random polynomials,"Center forRes. Comput. Technol., Harvard Univ., Tech. Rep. TR-CSE-03-01, 1981.

[2] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, "Reclaiming space from duplicate files in a serverless distributed file system," in Proc. 22nd Int. Conf. Distrib. Comput. Syst., 2002, pp. 617–624.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506057 |

[3] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Serveraided encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Secur. Symp., 2005, pp. 179–194.
[4] M. Bellare, S. Keelveedhi, and T. Ristenpart," Message-locked encryption and secure deduplication," in Proc. EUROCRYPT, 2008, pp. 296–312.
[5] G. R. Blakley and C. Meadows, "Security of ramp schemes," in Proc. Adv. Cryptol., 2008, vol. 196, pp. 242–268.
[6] A. D. Santis and B. Masucci, "Multiple ramp schemes," IEEE Trans. Inf. Theory, vol. 45, no. 5, pp. 1720–1728, Jul. 2009.
[7] M. O. Rabin, "Efficient dispersal of information for security, load balancing, and fault tolerance," J. ACM, vol. 36, no. 2, pp. 335–348, Apr. 2009.
[8] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 6, pp. 1615–1625, Jun. 2010.
[9] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," in Proc. ACM Conf. Comput. Commun. Secur., 2011, pp. 491–500.
[10] J. S. Plank, S. Simmerman, and C. D. Schuman, "Jerasure: A library in C/C++ facilitating erasure coding for storage applications—Version 1.2," Univ. of Tennessee, TN, USA: Tech. Rep. CS-08-627, Aug. 2011.
[11] J. S. Plank and L. Xu, "Optimizing Cauchy Reed-Solomon Codes for fault-tolerant network storage applications," in Proc. 5th IEEE Int. Symp. Netw. Comput. Appl., Jul. 2012, pp. 173–180.
[12] C. Liu, Y. Gu, L. Sun, B. Yan, and D. Wang, "R-admad: High reliability provision for large-scale de-duplication archival storage systems," in Proc. 23rd Int. Conf. Supercomput., 2013, pp. 370–379.

Impact Factor
7.54

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com