ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023

# An Investigation and Evaluation of Cloud Computing Capabilities and Data Duplication Methods

**Dr. R. Murugadoss**

*Professor, Department of Computer Science and Engineering,*
*V.S.B College of Engineering Technical Campus,*
*Coimbatore, TamilNadu, India.*

## Abstract

Cloud computing has made it possible to modify, configure, and access large-scale distributed computer programmes over a network. More people are moving their data to cloud storage as a result of the rising popularity of cloud computing. The fast-growing data volume in the cloud, which is typically on one side, results in a considerable quantity of data duplication. On the other hand, if there is only one duplicate copy of the preserved symmetric information in the deduplicated cloud backup, tampering or the lack of a single copy could lead to an unexpected failure. Therefore, it is crucial to discuss how to conduct credibility audits and file deduplication in both academic and professional settings in a safe and efficient manner. We establish a new ownership system during file deduplication in order to uphold the consistency of the tagging and the symmetrical modelling and to confirm shared ownership. Researchers also offer a method for ownership policy maintenance. A user-helped key is used in in-user block deduplication to provide a stochastic key process and reduce key storage space. Not to mention, the security and effectiveness audit demonstrates that our system effectively manages data ownership while ensuring data correctness and integrity.

## Keyterms:

Data replication, cloud computing, consumer internet, ownership policies,

## Introduction

Computing resources became a public utility with the introduction of the cloud in the middle of the 2000s; this idea stretches back to the early 1960s. Infrastructure-as-a-Service (IaaS), one of the cloud computing paradigms, uses a pay-as-you-go model that enables anyone with a working credit card to rent numerous computational power from cloud - based data centers and only pay for what they actually use, avoiding an initial cost in hardware infrastructure [1]. Internet-scale Web services like Netflix, Vimeo, and Pinterest are frequently built on public IaaS clouds like Aws Elastic Cloud Compute (EC2). Public clouds therefore have a huge impact on consumer Internet. For instance, since about April 2012, sites created on the Amazon cloud alone draw one third of daily Internet users and generate and over 1% of all Consumer online traffic.

[2] In many corporate and scientific fields today, including e-commerce, e-government, engineering design and analysis, finance, healthcare, web hosting, and online social networks, cloud technologies are being successfully used. Because of the flexibility and flexibility in resource provisioning as well as the utilisation of cutting-edge virtualization and scheduling methods, these technologies in particular offer cost-effective scalable solutions. [3] Cloud workloads are made up of a variety of different apps and services, each of which has its very own resource and performance requirements as

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023

well as restrictions outlined in the form of SLAs (SLAs). [4] The variable resource and network circumstances, as well as the dynamical character of the loads, which intensity can abruptly increase or decrease as a result of a variety of causes, are just a few of the many variables that have an impact on cloud performance.

Performance reduction may result from the utilisation of virtualized time-shared resources, notably [5]. The co-location of diverse applications within the same physical infrastructure causes interference and resource contention, and the overheads brought on by the resource management strategies in use are the main causes of this degradation. [6] The combination of workloads running simultaneously on a particular virtual machine (VM) might also have unanticipated consequences on performance due to conflicting temporal patterns of resource utilisation. When the workload is spread across various cloud infrastructures in multi-cloud settings, these performance difficulties could get even worse.

[7] Mapping public clouds to workload characteristics in these intricate situations is quite difficult. However, it is crucial for the successful implementation of cloud technology and the achievement of the specified service levels. [8] Therefore, it is crucial to develop a thorough understanding of the characteristics and the evolution of public clouds in order to deal with resource management, provisioning, and online capacity planning, as well as, more broadly, to manage and anticipate quality of service and performance (QoS). As a result, it is necessary to think about structured and systematic approaches to workload categorization as being an essential part of each of these tactics.

[9] Despite their significance, the characterisation and prediction of public clouds have only been briefly discussed in the literature, and most of the time just at the base of the VMs, without taking into account the characteristics of the specific workload components operating on the VMs. The objective is to give a general overview of the key difficulties surrounding the deployment of workloads over their entire lifecycle in cloud environments. [10] More specifically, we define certain broad workload categories stated in terms of both qualitative and quantitative qualities after first identifying the most pertinent behavioural aspects of cloud workloads. A literature assessment of the use of rescheduling methods and fault diagnosis and prediction mechanisms in the context of cloud workloads completes this in-depth study of the state of the art.

[11] Multi-tenancy, in which users from many organisations use the same hardware infrastructure, is a distinctive characteristic of public clouds. As a result, cloud providers employ virtualization to grant consumers computational capabilities resources in the form of virtual servers (VMs), while still maintaining complete control over the underlying hardware infrastructure. This is done in place of granting direct hardware access. [12] Users should be given the impression that they have exclusive access to hardware, and there should be tight isolation between virtual machines which share physical servers, the data centre network, and other levels of the cloud architecture so that they do not interact with one another.

[13] Unfortunately, in public clouds, such isolation is sometimes broken due to competition for the multiplexed shared resources used by guest VMs, which may cause performance interference. For instance, the achievement of a workload with a space - time locality in its shared memory pattern heavily relies on the effectiveness of different tiers of CPU cache memory, but its neighbouring virtual machines on the same physical server may run workloads that frequently cause cache eviction, forcing repetitive memory locations access for the same, recently used content. [14] Virtualization establishes a semantic gap between guests, who handle application workloads, and hosts, who control the cloud infrastructure, making it difficult to mitigate make efforts among guest virtual servers in public clouds.The processes or even the goals of optimization at one layer are built without understanding those at another, and as a result, they frequently serve opposing ends. For instance,

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023

from the perspective of the cloud's guests, the resource schedulers, which are under the control of the host and are used by all guest VMs, determine the degree of resource contention, whereas from the perspective of the host, scheduling policies may be affected by the resource usage patterns of the applications, but only guest VMs are aware of these patterns.

[15]Deduplication can occur at the block level or at the file level, and at the file level, it removes duplicates of the same file. While guaranteeing data confidentiality, traditional encryption is inconsistent with data deduplication. In particular, conventional encryption calls for several clients to secure their data with separate keys. Deduplication will be difficult since different users' copies of the same data will result in distinct cypher texts. It has been suggested to use convergent encryption to compel data secrecy while enabling deduplication. It uses a Convergent key to encrypt or decode a data copy that is acquired by calculating the cryptographically value of the data copy's content. Users keep the keys after data encryption and key generation and submit the encrypted text to the cloud. Since the cypher text and convergent key are formed from of the data content and the encryption method is deterministic, compatible data copies will produce the same cypher text and convergent key.

## Proposed Method

In this section, the suggested deduplication system is presented and constructed. The amount of deduplication should be determined proportionally by the storage nodes as well as other number of nodes. The same duplication node must receive identical data in order to achieve a high duplicate removal ratio. We develop an integrated deduplication architecture, which is described in this part, to provide great duplication efficiency and a large deduplication ratio while utilising minimum computational power.

The following parts show how our proposed model is laid out. To enhance high reduction reliability and scalability performance, we present a suggested data routing solution.For quick duplication efficiency over the elimination of duplicate endpoints, this approach has been supported by an overview of implementation-oriented datatypes.

## Frame Work Overview

Figure 2 presents the proposed strategy of this paper. This article's architecture is symmetrical in that it includes both the end user and the end service provider. Users nodes, storage nodes, and management nodes are the major three modules in this symmetrical decentralised deduplication paradigm.

## User Nodes

The identity verification chunk, data redirection, and data categorization are the three main feature components of a user node. The user feature protects and recuperates data files, uses SHA and MD5 and other collision-resistant hashing algorithms in the data tend to cluster module to determine the small piece sizes of each stream of data, and uses the two-tier routing check to reroute traffic from each section to a depreciation backup base station that is highly similar.

## Storage Nodes

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023

The chunk management, clients similitudes database search, and chunk index cache storage are the three primary parts of the storage server section. It also includes dynamic routing, inserting hot new chunk signature into the chunks search buffer to accelerate up the searching process, replicating chunk while keeping the be had in larger, parallel storage units called as warehouses.

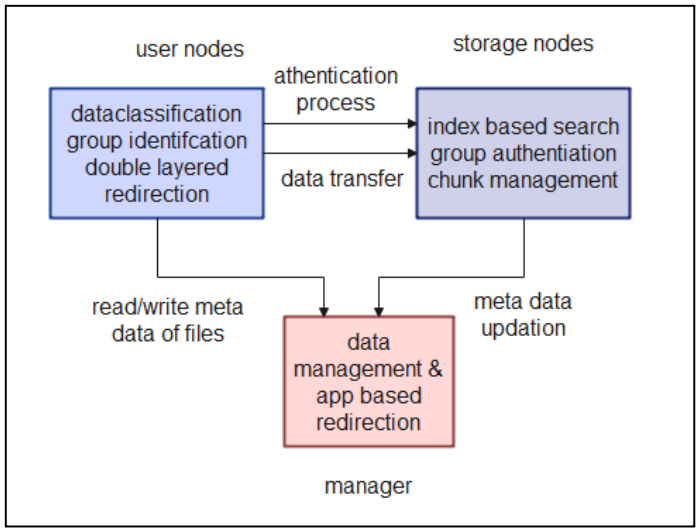


**Figure. 1. Model of the proposed system**

## Manager

Just on deduplicated storage node, this was in charge of managing and monitoring the file platform used for storing and retrieving data. It consists of transmitting decision-making and managing file reads. The file mapping for extraction impressions and the file reconstruction data are both stored in the document receipt management program. Metadata at the file level is kept by the director. The programme chooses a group of appropriate storage nodes for each file based on its comprehension of the forward recommendation framework and asks users for feedback on straight super chunk routing. To avoid the single node failure for high access in an efficient node failures, the director helps up to two nodes.

By downloading the documents from the data holders, our technology often reduces user data levels. Each file has a file-based key and an information module that are chosen by the file's owner. The label is then given to the supplier by a data user. Deduplication is finished by a service provider, who also produces reports for the data owners. Keep in mind that the user and provider must demonstrate the task method in order to guarantee tag correctness.

## Workload Profiling And Monitoring

The foundation for evaluating the quantitative and qualitative characteristics of the tasks is monitoring and profiling. Monitoring, in general, keeps tabs on the actions taken by the loads being processed as well as the condition of the allocated and available resources. The main goal of profiling is to explain how a workload makes use of cloud resources. Because of the complexity and dynamic nature of these settings, monitoring and profile in the clouds is extremely challenging. However, these actions are crucial when dealing with problems like:

- Resource management and capacity planning.
- Performance optimization.
- Billing.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023

- Troubleshooting and security.
- SLA confirmation

Different strategies have been developed to address particular monitoring difficulties. The load characteristics that can be seen at realtime to describe resource usages are the main emphasis of the sections that follow. The monitoring approach chosennamely, clouds providers and cloud usersdetermines the level of detail of the measurements taken in the clouds.

A different method used to gauge how much of a given resource is being used by each demanding activity is profiling. Particularly, profiling can be used by users of cloud services to optimise resource provisioning provision and by cloud service providers to fine-tune scheduling and placement rules for virtual machines. Profiling faces increased difficulties in cloud systems because of interference from other co-located VMs. In fact, the pooling of hardware resources may cause hardware elements like cache memories, CPU pipelines, and physical I/O devices to behave in ways that are unpredictable. Dynamic monitoring and sample hardware performance counters are common approaches for gathering pro ling measurements. An alternate strategy is focused on monitoring the general performance of the virtual machines (VMs) hosting the target apps at the hypervisor level.

## Process Of Deduplication

The user uploads the data if the service provider includes TF on its list of document tags but there is no TagF within the cloud. If not, the admin employs the Pows protocol used by the service provider, as shown in Algorithm 1, since the provider owns a copy. This user is allowed access if the procedure is successful with out file having uploaded to the saved file. By doing this, client-side file deduplication costs for CPU and connectivity are decreased.

**Algorithm 1**: Proposed Deduplication Model

**Step 1:** Client file management with file tag list.

**Step 2:** Create authentication tag collection as (a1, a2,..., am).

**Step 3**: Estimate Service provider or user behaviour modelling as Q = 1-(1-ß).

**Step 4**: Service provider authentication setup phase: <G,D,F(GD), Public Key>

**Step5:** Compute Authentication key value as: LF(E)and M'LOD.

**Step 6**: Encrypt the file tag information and file chunks.

**Step 7:** Encryption of keys as E(PK)

**Step 8**: Authenticator thread generation α," =D; F(GD).

**Step 9:** Execute query for deduplication Q = (f. G(i)).

**Step 10:** Hash index generation with object indexing process.

This is how data integrity verification is often done. Unlike the methods described above, the provider will add an users consider that cannot be kept for a group of authentication tags as well as a tag for the same block. In addition to preserving a provider's faith in the data, the concept of convergent encryption enables one to implement "existing facts" and c – c bond decryption, which would be frequently deducted encrypted data. While providing evidence, the supplier will mask the weighted D from of the samples block to the auditor's randomized masks S and public key in order to further

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023

avoid data leaks against the auditor. Despite having both S's key pair and that of the data holder, an auditor is mathematically unable to acquire the data owner's private information underneath the assumptions of hardness.

**Cloud Backup Server**

The client backup data is maintained by the online cloud server module. The file will be divided up into blocks using the Chunking Process Module. The block level deduplication module performs block signature creation and deduplication operations. The file level reduction module is made to carry out file level deduplication. The data backup values are protected by the data security module. Mobile phones' deduplication functionality is found in the Smart phones module's Deduplication section.

**Chunking Process**

The little files are separated using a file size filter. The huge size files are divided into chunks using an intelligent chunker. File format, static uncompressed files, and dynamically uncompressed files are the three types of backup files. Dynamic files can be edited, whereas static files cannot. A Whole File Like some (WFC) mechanism chunks compressed files. Static Chunking divides static uncompressed files into fixed-sized chunks (SC). By using Content Defined Chunking, dynamic uncompressed data are divided into chunks of different sizes (CDC). The deduplication mechanism has been modified for clients on computers and smartphones. The technology offers protection for the values of the backup data. The deduplication procedure also includes small file sizes. There are six main modules that make up the system.These include Deduplication in Smart Phones, Stacking Process, File - level Defragmentation, Multiple File Deduplication, Security Process, and Cloud Backup Server.

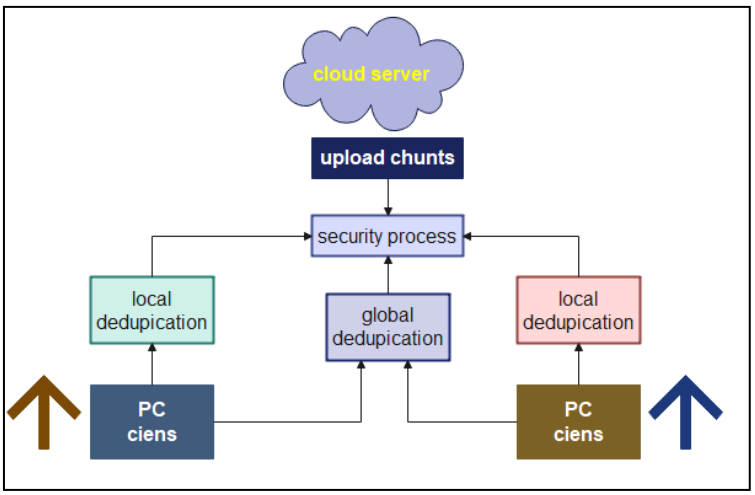


**Figure. 2 cloud server**

The cloud server is depicted in Figure 2. The client backup data is maintained by the online cloud server module. The file will be divided up into blocks using the Chunking Process Module. The block level deduplication module performs block identity generation and deduplication operations. The file level reduction module is made to carry out file level deduplication. The data backup values are protected by the data security module. In the Deduplication section of the Smartphones module, the deduplication process is carried out in mobile devices.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023

## Block Level Duplication

The hash engine generates chunk fingerprints. As a chunks fingerprint for localized duplicate data identification in compressed files, Rabin hash algorithms of 12 bytes are used. In compressed files, the global deduplication procedure is carried out using the Message Digest MD5 method. In uncompressed static files, deduplication is accomplished using the Secure Hash Algorithm (SHA1). The hash engine generates chunk fingerprints. As a chunk signature for localized duplicate data identification in compressed files, Rabin hash algorithms of 12 bytes are used. In compressed files, the global deduplication procedure is carried out using the Message Digest MD5 method. In uncompressed static files, deduplication is accomplished using the Secure Hash Algorithm (SHA1). The Message Digest (MD5) technique is used to hash dynamic uncompressed data. The native client and distant cloud both perform duplicate detection. Both a local and a global index of fingerprints exists. Trying to verify the fingerprint index values allows for deduplication to take place.

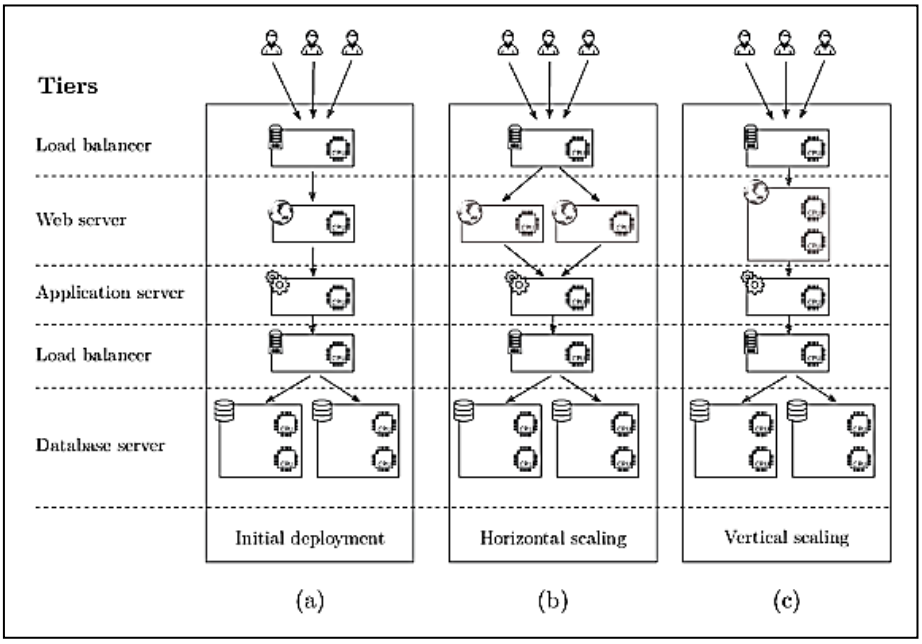


**Figure. 3 A five-tier architectural example that is common of big web applications Initial installation (a), as well as the web server's horizontal and vertical scaling (b, c).tier, correspondingly**

## File Level Duplication

Under a segment store environment, little files are kept. Deduplication at the file level is applied to files smaller than 10 KB. With the help of the Rabin hash function, file level signatures are produced. With the aid of a file-level fingerprint index verification technique, deduplication is carried out. Cloud computing is mostly used for the deployment of large-scale applications in domains like e-commerce, financial services, healthcare, gaming, and media servers when it comes to interactive workloads, which typically need to handle the changing behaviour of users. Multi-tier architectures are a popular approach to addressing these highly fluctuating load conditions, with each tier addressing a specific functionality and being deployed on one or more VMs. Figure 3 shows the architecture of a ve-tier web app as an illustration. The benefit of this approach is the ability to independently dynamically

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023

scale each tier both vertically and horizontally. In horizontal scaling, the number of Virtual machine is changed (see Fig. 3(b)). Vertical scaling, on the other hand, entails changing the quantity of funds allotted to individual VMs. The VM running the web service scales up it doubles the amount of cores it has, as seen in Figure 3(c).

## Results

All RPC apps are packaged to lower round trip overheads. We run an event-driven simulation of distributed deduplication strategies for duplication, load distribution, memory management, and overhead communications on one of our four servers. Our solution allows the client to simultaneously execute data splitting & fingerprinting before making a data routing decision. Program files can be chunked in two different ways. There are two methods: CDC whacking with variable-size chunks and SC stacking with fixed-size pieces for each type of programme. The grouping of a large number of successive small bits into the an amazingly for route optimization is the next phase. The OpenSSL library's implementation of hash fingerprints is its core. Results of deduplication for several Application models are shown in Table 1.

**Table 1. Results of deduplication for various application models**

| application | RAM usage(MB) | Bin Model (MB) | Deduplication(MB) |
|---|---|---|---|
| VM | 9.69 | 35.67 | 20.76 |
| Mail | 66 | 26.49 | 22.65 |
| Audio | 37.47 | 10.46 | 23.85 |
| OS | 59.31 | 25.109 | 22.46 |
| Photo | 22.67 | 23.66 | 21.36 |

The throughput comparison of the suggested approach and the existing one is shown in Figure 4. We compare the average likelihood directory of our apps to the common likelihood index for simultaneous deduplication in a single deduction storage node with several data sources. The results show how RAMFS data input and VM data output can be deduplicated simultaneously to reduce storage I/O blocks performance interference. We use a cold cache and an application to evaluate output, and we are conscious of the similarity index. When we first simultaneously deduplicate numerous words on the VM dataset, the segmentation fingerprint cache is empty, which is referred to as being in the "cold cache" in this situation. If we perform parallel deduplication with many streams, we go back to a dataset; "warm cache" refers to duplicated chunk fingerprint.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023

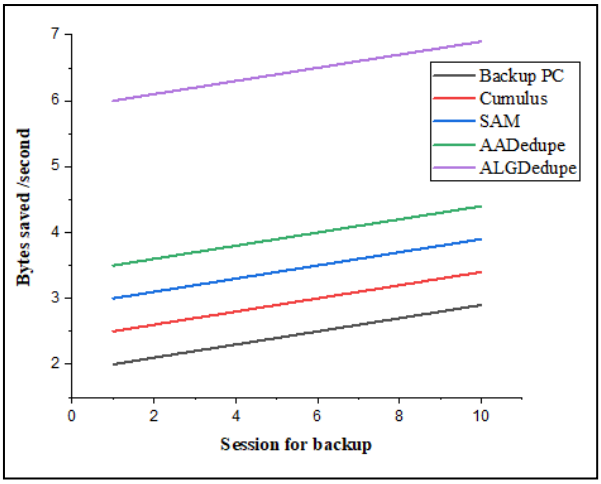


**Figure. 4. Throughput Comparison analysis.**

Figure 5 shows the spatial complexity analysis. We note that concurrent duplicating methods with software correlation will probably outperform naive similar duplication methods, and simultaneous warm-cache duplication techniques should perform much better than cold-cache regimes. With the increasing growth of mobile data traffic, comparable deduplication performance increases to 5.9 Gigabyte for both a software similarity measure as well as a warm supply.Due to extra rivalry for Index Position and Disk I/O, the percentage is 5.5GB/s lower with thirteen synchronised channels. According to our scientific findings, the providers would require each customer's total cloud storage capacity for the eight online cloud schemes for each backup session.

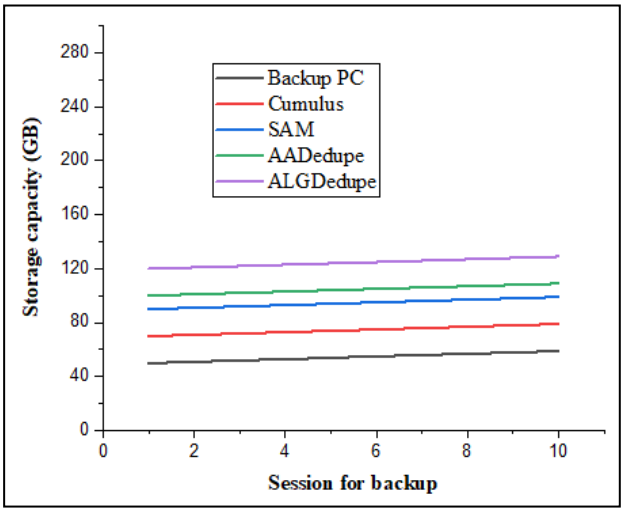


**Figure. 5.Space Complexity**

In source deduplication techniques, the Backup PC gross graining technology does not consistently outperform other fine graining mechanisms. The fine-grained Cumulus saves less storage than the local deductible-only AA Dedupe because it only performs a local replication scan and limits its search for undamaged data to blocks in the older models of the file. The threshold ratio of AA-Dedupe grows as a result of ALG-use Dedupe's of additional global deduplication levers in the cloud.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023

Additionally, it improves SAM, which combines global file origin deduplication with local chunk-level deduplication, by enhancing framework knowledge and improving overall deductive performance, as depicted in Figure 6.

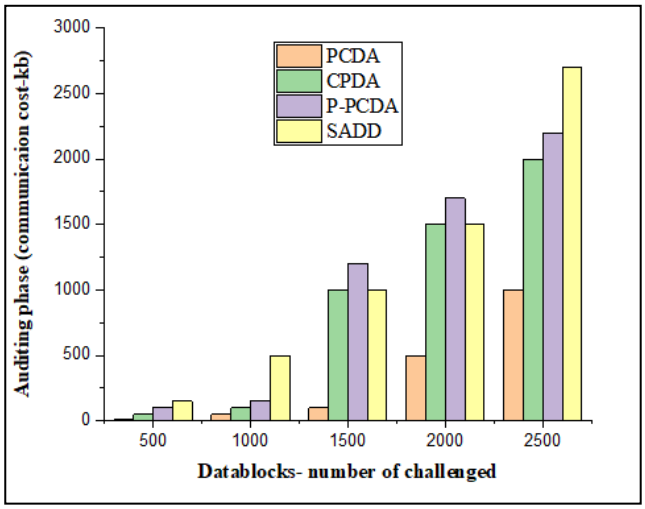


**Figure. 6.Data transfer cost.**

We utilise high-speed global replication checks. Figure 7 illustrates how we significantly improve the system's overall deduplication efficiency despite the high WAN latency. We examine the four cloud backup systems in the same cloud-related storage network.

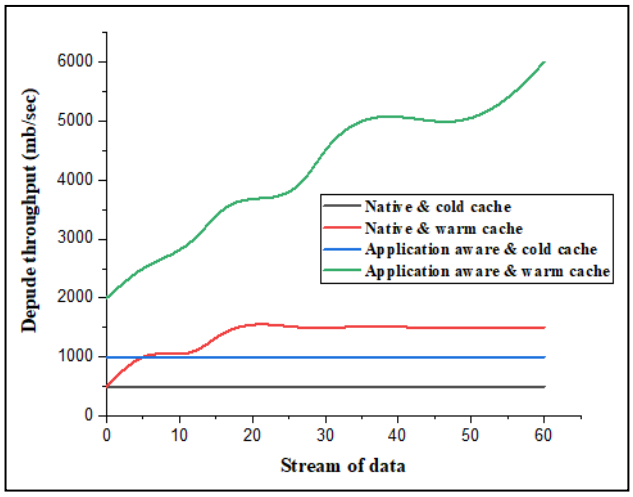


**Figure. 7.Deduplication efficiency analysis**

The performance test using deduplication in Figure 8 compares favourably to similar backup techniques with no overhead. The main benefits are the understanding of its use and the ubiquitous detection of duplicates throughout the deduplication process. It is 14% more effective than the present local machine, AA-Dedupe, and is roughly approximately twice the software SAM and 1.9 times the region dumping Cumulus quality because to its benefits in global design.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023

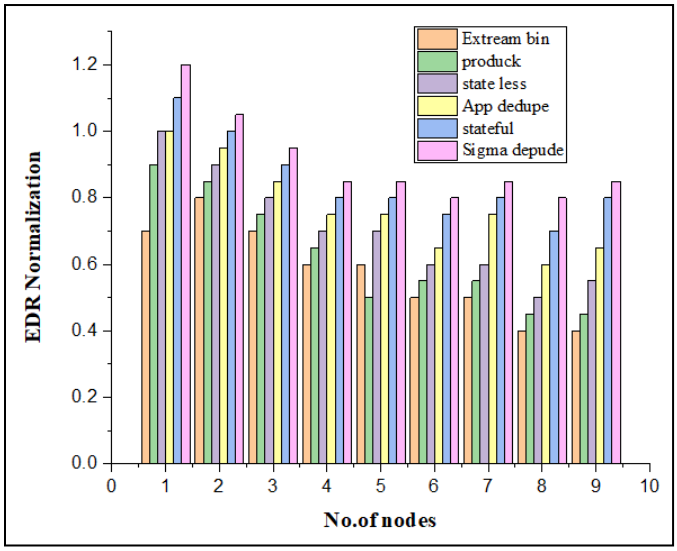


**Figure. 8. Distributed deduplication effectiveness**.

## Conclusions

In this work, we used app sensitivity, information uniqueness, and localization to construct a software adaptive inline decentralised big management deduplication architecture. This method strikes a balance between scaling efficiency and deduplication efficiency. To reduce pass data repetition with regulated overhead and trustworthy load balancing, a two data redirection structure is used. With somewhat more overhead structure in comparison to the strongly attachable idle groupings, it outperforms the domain specific close coupling architecture in terms of the cluster's enormous compression ratio. Second, it considerably improves the flexible stateless connectiveness techniques in the strong deductibility ratios throughout the cluster while maintaining the latter's high overhead scalability. Finally, we were able to demonstrate the effectiveness of the proposed method by contrasting the costs of communications, computation, and storage with those of modern techniques. This claim is further supported by numerical simulation and experimental findings.

## References


1. Fu, Yinjin, et al. (2017) Application-aware big data deduplication in cloud environment. IEEE transactions on cloud computing 7.4: 921-934.
2. Wu, Suzhen, et al. (2019) Pandasync: Network and workload aware hybrid cloud sync optimization. 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). IEEE.
3. Viji, D., and S. Revathy. (2019), Various data deduplication techniques of primary storage. 2019 International conference on communication and electronics systems (ICCES). IEEE.
4. Barik, Rabindra K., et al. (2021), GeoBD2: Geospatial big data deduplication scheme in fog assisted cloud computing environment. 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE.
5. Shakarami, Ali, et al. (2021), Data replication schemes in cloud computing: a survey. Cluster Computing 24.3: 2545-2579.
6. Ren, Ju, et al. (2019), A survey on end-edge-cloud orchestrated network computing paradigms: Transparent computing, mobile edge computing, fog computing, and cloudlet. ACM Computing Surveys (CSUR) 52.6 : 1-36.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 12, Iss 4, 2023


7. Devarajan, A. Augustus, and T. SudalaiMuthu. (2019), Cloud storage monitoring system analyzing through file access pattern. 2019 International Conference on Computational Intelligence in Data Science (ICCIDS). IEEE.
8. Fu, Yinjin, XiaofengQiu, and Jian Wang. (2019), F2MC: Enhancing data storage services with fog-toMultiCloud hybrid computing. 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC). IEEE.
9. Naseri, Afshin, and NimaJafariNavimipour. (2019), A new agent-based method for QoS-aware cloud service composition using particle swarm optimization algorithm. Journal of Ambient Intelligence and Humanized Computing 10.5: 1851-1864.
10. Zhou, Xijia, et al. (2019), An experience-based scheme for energy-SLA balance in cloud data centers. IEEE Access 7: 23500-23513.
11. Jia, Gangyong, et al. (2015), Coordinate memory deduplication and partition for improving performance in cloud computing. IEEE Transactions on Cloud Computing 7.2: 357-368.
12. Nashaat, Heba, NesmaAshry, and RawyaRizk. (2019), Smart elastic scheduling algorithm for virtual machine migration in cloud computing. The Journal of Supercomputing 75.7: 3842-3865.
13. Yan, Zheng, et al. (2017), Heterogeneous data storage management with deduplication in cloud computing. IEEE Transactions on Big Data 5.3: 393-407.
14. Kaur, Sandeep, and KiranbirKaur. (2019), Enhancing Reliability of Cloud Services Using Mechanism of Dynamic Replication and Migration of Data. International Journal of Applied Engineering Research 14.8: 1976-1983.
15. Gai, Keke, et al. (2020), Blockchain meets cloud computing: a survey. IEEE Communications Surveys & Tutorials 22.3: 2009-2030.