IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.047 ⁜ Vol. 10, Issue 6, June 2022

DOI: 10.17148/IJIREEICE.2022.10605

Cloud Storage Monitoring System analyzing through File Access Pattern

Ms. Manya S[1], Ms. Sahana M[2], Mr. Srinivas B V[3]

Student, Department of Information Science and Engineering, Atria Institute of Technology[1,2]

Assistant Professor, Department of Information Science and Engineering, Atria Institute of Technology[3]

Abstract: Cloud computing is an important technology on current demanding business requirements and it has been emerged as unavoidable technology. The usage of IaaS Service storage for Cloud Computing is being expanding exponential every year. The cloud storages are used by the cloud user due to its economy compared with other storage methods. The replications of files helps user for easy access with high availability which reduces the overall access time of the files, but at the same time it occupies more storage space and result in high storage cost. The cloud user holds multiple times of the storage than what he is actually needed. It is a dire need of system to find unwanted files in the cloud and also optimize the storage space by evaluating through file access frequency.

This paper propose Cloud Storage Monitoring (CSM) system, which monitor the IaaS storage usage and analyze the file access patterns by various parameters to identify the frequency of access, size, future access prediction, replication of files in the cloud storage. This allocates a ranking for each file which also predicts future access pattern. This generates a recommendation dashboard to the user who can decide on the operations such as reorganize, delete or archive the files and eliminate duplicate files in the cloud storage which can increase the space for future use. This system is experimented in the CloudSim environment and validate through multiple simulations testing, by using comparison techniques related to file attributes, delta versionhashing, Data de-duplication. The ranking algorithm technique applied on frequency distribution shows that increase in the storage space up to 10.91% higher than the normal system. It also helps to forecast towards future files usability prediction and prevents the duplicate entries.

I. INTRODUCTION

The data replication services of cloud storage duplicate the files in real time to increase the availability of the files which in turn increase the hardware cost. The data replication service consists of data replication, file replication, cloning infrastructure and remote storage replication. The cloud storage replication service determine of redundancy which is invaluable on main storage when backup system fails.

As the result, replication is used to reach highest availability at high cost. It is degrading the performance of the service when the cost benefits accrued from the replication. This also increase delay in request and response transaction in cloud environment.

The predictive auto-scaling technique forecast future storage workload of the cloud service and adjusts the cloud storage capacity in order to meet the future needs. The system generates the recommendation dashboard to forecast future files usability and it also eliminate the duplicate entries.

II. RELATED WORK

The monitoring of cloud storage is one among the emerging research filed in cloud research. There are many active research works have contributed to the field in last ten years.

Ali et al., Samuel A. Ajila and Chung-Horng Lung [1] have developed a system "An autonomic prediction suite for cloud resource provisioning". Their proposed system predicted the workload pattern. The work load patterns have assumed the database layer, which has no negative impact on prediction with respect to auto-scaling accuracy. Their system is required to be enhanced for handling the other pattern of workloads also. Annal et al., [2] have reported a research plan to increase the efficiency of the cloud storage as well as reduce the threads without affecting the key features of the cloud storage system. They also presented a ranking algorithm which was ranked the files based on their accessing frequencies. The results showed that there is an improvement on optimizing used space, server performance, and response time delay. However, the algorithm required a high cost of operation. Prabavathy et al., [3] have presented "Improving Read Throughput of Deduplicate Cloud Storage using Frequent Pattern-Based Prefetching Technique" to recognize the frequent access patterns on history of usage.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.047 ⁜ Vol. 10, Issue 6, June 2022

DOI: 10.17148/IJIREEICE.2022.10605

It is also predicting the "combine of fingerprints" which is most probably be accessed in the immediate future on the cache values. Sarbjeet Singh et al., [4] have designed dynamic rebalancing strategy. The strategy was using the reduplication technique to rebalance the dynamic requests. It was simulated in CloudSim simulator. The result showed the significant improvement on effectiveness of the dynamic rebalancing. The presented algorithm middleware, and geo-location services. Externally having visible parameter with relationship between interfaces.

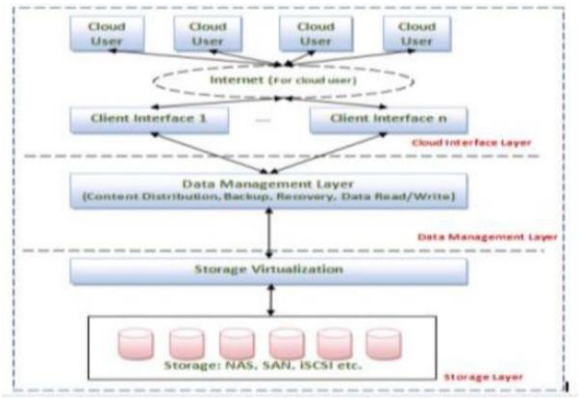

Figure 1: Cloud Storage Architecture

A. Cloud Client Interface Layer

This layer represents as a software layer which is provided by Cloud storage providers to connect different Cloud users to avail the Cloud IaaS storage service by using Internet connection. This layer has authorization and authentication techniques in order to authenticate and validate the users with their credentials by using the single sign-on.

B. Cloud Data Management Layer

The Data Management Layer which is used to manage and validate data of particular cloud users with respect to data architecture and activities like data storage, data partitioning, synchronization, and content distribution across storage location, control data movement over the network, maintain consistency backup and data recovery replication [6]. It provides necessary data access with distribution of parameters for the data layer. It also stores repeatedly used SQL statements in memory areas by using the metadata. This layer data encrypted by SHA keys while update in the database which can have back up and automatically incorporate the security features.

C. Cloud Storage Layer

This cloud storage layer majorly consists of two parts:

Storage Virtualization: It maps the heterogeneous storage devices and hardware having single allocation storage space which can create a shared platform through dynamic storage layers [7]. The storage virtualization technology provides built-in availability, scalability and security to applications.

Basic storage: It contains hardware device which can encompass different database servers and storage devices having heterogeneous nature such as DAS, SAN, and NAS etc.This layer includes architecture layer on storage classification.

D. Design Principles of Cloud Storage

The main design principle of cloud storage has requirements such as availability, scalability, cost reliability, simplicity, multitenancy, speed and bandwidth limitation. However, while many of these design principles and patterns are not particular to the cloud, and could be applied locally, they become necessary when building reliable cloud services [8]. The Cloud storage should able to meet requirements of unlimited and concurrent users and experiment without affecting the system performance and usage speed. It uses virtualization and prevents over limit provisioning and enhances the efficiency which has specific boundaries with the actual requirement on physical storage allocation at the moment. When

DOI: 10.17148/IJIREEICE.2022.10605

application grows, the storage blocks will automatically increase with system requirements. It also decreases the amount of storage requirement on service layer of applications which minimize the disk drive constraints on energy consumption.

These cloud storage requirement services will be accessible at any point of time when user requires IaaS storage. These techniques related to decentralization such as replication, erasure codes which are used for better availability and fault–tolerance process related to cloud services in architecture layer. There is chance of data replication which resides on different servers in different locations which can prevent a single point of failure. Suppose if primary system fails, backup system needs to take over. This increases the data availability feature for users [9]. At any point the data can be retrieved from any combination of fragments which can decompose the original structure. The other techniques such as snapshot, replicating framework and cloning services can be used for duplication of data for better availability and reliability. This snapshot can simplify access to stored data which can speed up the process of data recovery. It can also increase the redundancy of data.

IaaS creates virtual hardware devices such as virtual networks, virtualized storage and virtual machines [10]. This IaaS layer is tightly coupled on concepts of virtualization and the higher level requirements of PaaS and SaaS. The productive features such as de-duplication increase storage utilization, thin provisioning reduces the amount of process time.

E. Challenges on Cloud Storage

Nowadays most organizations understand the benefits of migrating data to a cloud storage service but at the same time cloud services also having its own risks and drawbacks. In future cloud storage services will replace the storage network in the data center, mostly due to high sensitive transactional applications, data-intensive, low-response time, and deals with critical data. Most of use cases are related to organizations and companies having substantial on-premise storage requirements related to cloud storage from various vendors in a Public/Private/Hybrid model deployment. The organization is making difficult on enforcing cloud storage data management policies and best practices on storage optimization features.

Security on public cloud is not more secure than in-house storage; Most of IT managers aren't comfortable when dealing with sensitive data on public environment. The sensitive data has been shared to cloud provider which is having multi-tenancy infrastructure which is accessed by public. The Cloudian expressed concerns 62 percent of survey related to organizations security issues which is most common challenge in cloud storage management [11].

Cost related to cloud storage services based on the amount of storage capacity which was consumed by users and the number of IOPs accomplished with respect of the amount of consumed bandwidth. It may reduce cloud storage costs through optimization on de-duplication and also taking advantage of pay-as-you-grow options on choosing the respective cloud storage service provider [12,13]. It meets all other challenges on the principle. Most of organizations use the public cloud which leads to optimize the cloud storage costs.

Interoperability of many organizations incorporating hybrid cloud IaaS storage principles, it related with on-premise layer of infrastructure on a key challenge in many organizations [14]. Most of enterprise having concerns on-premise critical applications. Vendor lock-in begins use on cloud IaaS storage vendor, the data transmission to a different vendors having inhouse becomes a costly and complex operation. Almost 20 percent of enterprises had vendor lock-in issues related to public/hybrid cloud storage.

IV. CLOUD STORAGE MONITORING(CSM) SYSTEM

A prediction and ranking based system is proposed to handle the de-duplication in cloud storage with the following design objectives.

- Identify the frequency on access pattern Provide prediction on file access
- Identify the duplication of files on cloud storage
- Build storage efficient system.
- Increase efficiency of the system.
- Improve search experience
- Block duplication of files in future

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.047 ⁜ Vol. 10, Issue 6, June 2022

DOI: 10.17148/IJIREEICE.2022.10605

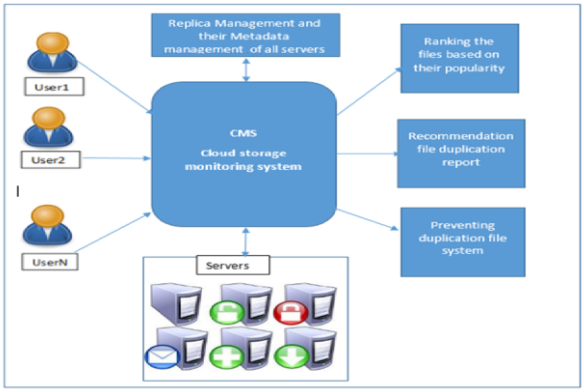

Figure 2: Cloud Storage Monitoring Architecture

A prediction and ranking based system is proposed to handle the de-duplication in cloud storage with the following design objectives.

- Identify the frequency on access pattern
- Provide prediction on file access
- Identify the duplication of files on cloud storage
- Build storage efficient system.
- Increase efficiency of the system.
- Improve search experience
- Block duplication of files in future

The proposed research work, CSM system rank the files based on their popularity and the frequency of access [15]. The system generates ranking dashboard which helps to optimize the storage space and availability. The CMS system reduces the storage space by de-duplication and increase the availability by having the files ready for access. The ranking is determined by using the frequency of access and future access prediction of files with the weight value of 0.6 and 0.4 respectively. The CSM system is simulated at CloudSim with the simple cloud storage environment [16]. A sample file accessed environment is generated as Table 1.

Table 1: File Accessed per week Vs Transaction Id

Transaction id	Files accessed
T1	File1, File2
T2	File2, File3, File4
T3	File1, File3, File4, File5
T4	File1, File4, File5
T5	File1, File2, File3
T6	File1, File2, File3, File4
T7	File1
T8	File1, File2, File3
T9	File1, File2, File4
T10	File2, File3, File5

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.047 ⁜ Vol. 10, Issue 6, June 2022

DOI: 10.17148/IJIREEICE.2022.10605

The detailed file access history is given in Table 2. The ranking base is shown in Table 2.

Table 2: Ranking Base on Popularity

File Name(.Ex	*File Types*	*File SizeinM*	*Rank*	*T1*	*T2*	*T3*	*T4*	*T5*	*frequency*
Main.java	docx	0.08	2	1	1	1	1		4
Check_ds	pdf	0.11	4	1	1				2
Image_01	jpeg	0.41	3	1	1	1			3
Help_VS	mp3	0.55	1	1	1	1	1	1	5
Eng_TST	mp4	1	4	1	1				2

K-means ranking Algorithm: **Input**: k (the number of clusters),
D (a set of lift ratios)
Output: a set of k Clusters

Method:
Arbitrarily choose k objects from D as the initial cluster centers:

Repeat:

1. Reassign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
2. Updated the cluster means, i.e.; calculate the mean value of the objects for each cluster **Until** no change;

Description:

1. Identify the clusters data into k groups or points where k is assigned with initial group centroids.
2. Categorize or Select k groups or points with random cluster centers in data.
3. Assign the data points to their nearest cluster center towards the Euclidean distance function.
4. Evaluate the positions of K centroids with all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same centroids or points are assigned to each cluster in consecutive rounds.

The de-duplication process is carried out by using the following methods:

- Comparison based on File attributes
- Comparison based on delta version and hashing
- Data de-duplication

A. Comparison based on File attributes

This technique is used to reduce the duplication of data at the file level. It is comparing parameters of the file system which includes the file size, file type, file name, and date-modified information of two files with the same attributes including file name being stored in a system. It scans specific folders and identifies the files which are having same

DOI: 10.17148/IJIREEICE.2022.10605

attributes [17-19]. This process runs as usual applications which could minimize the storage without demanding overheads to the cloud performance.

B. Comparison based on delta version and hashing

The file-level duplication compares individual files on inside data and variances within the files which compare to updates into a file and then store those variances of "delta" details to the original file. This file version techniques associates on file updates and it stores the deltas value in other versions. The comparison process through hashing techniques creates unique values on mathematical "hash" representation of files. The hash values compared for new file storage. This match on hashing provides assurance that the files are same, and it can be removed. The delta-encoding technique is used to identify the files having similar attributes [20,21].

C. Data de-duplication

The redundant data have been eliminated through the data deduplication or single instancing. The de-duplication process used in the cloud server can reduce the space requirement of the server. There are multiple de-duplication strategies which have been followed in different organizations. Before implementing the de-duplication techniques, it is very effective to take backup. Data de-duplication majorly having three types. **Compression** is technique frequently used for long time. **Single-instance storage (SIS)** which has used to remove redundant files from storage archives. **Data comparison** detects the duplicate copy by comparing files. If any match has been identified, then the file is discarded [22].

D. Data Compression

Data compression technique mainly used to compress the given file by reducing the size of files but not to reorganize or eliminate duplicate file. By using data compression empty space that appears inside file can be removed. But still duplicate data is available in local file and remains independent and data segments within those files. The advantages of data compression just compress the space which becomes isolated to each particular file.

E. Single-Instance Storage

This technique also removes multiple copies of any file. It can also detect and eliminate redundant copies of similar files. The main usage is to keep only the single Instance where the pointers are created for all other users having ownership of same file. The SIS also checks the content of files and determines whether the files are identical towards the existing file while uploading in cloud storage. In some instance the user insert or change only header level and may be most of file having redundant data in the same file. For example: Most of the time user insert or change only the title slide of a presentation and may have large amount of data redundancy.

F. Data Comparison

This methodology detects duplicate copy of the file by comparing first 50 bytes of the new and existing files also comparing the last 50 bytes of new and existing files. This comparison is done on byte by byte with existing files in the IaaS storage.

V. RESULTS AND DISCUSSION

The proposed K-means algorithm allocate ranking for files. It can validate by simulation testing through CloudSim with having samples of five files and 3 storage servers with each 5GB storage. The results are tabulated as shown in Table 3.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.047 ⁜ Vol. 10, Issue 6, June 2022

DOI: 10.17148/IJIREEICE.2022.10605

Table 3. Simulation results with Proposed CSM system

S.No	File	File Types	File	Servers			No. of
				Server15(	Server25 B	Server35(	
1	Main.java	docx	0.08	0	5	0	5
2	Check_ds	pdf	0.11	0	0	2	2
3	Image_01	jpeg	0.41	0	0	3	3
4	Help_VS	mp3	0.55	3	4	0	7
5	Eng_TST	mp4	1	0	0	2	2
Used Space in GB				0.5	1.4	0.9	
Available Space in GB				4.5	3.6	4.1	

The similar environment is also simulated without using the proposed CSM system as shown in Table 4 and the results are compared as shown in Figure 2. The proposed CSM system yield better performance in utilizing the storage space using deduplication technique. There are five different files with the size of 0.08 MB, 0.11 MB, 0.41 MB, 0.55 MB and 1 MB are used for the experiments. The average is taken from 100 independent trails.

Table 4. Simulation results without CSM system

S.No	File	File Types	File	Servers		
				Serve GB 15	Server2 GB	Server3 GB
1	Main.java	docx	0.08	0	5	0
2	Check_ds	pdf	0.11	0	0	2
3	Image_01	jpeg	0.41	0	0	3
4	Help_VS	mp3	0.55	3	4	0
5	Eng_TST	mp4	1	0	0	2

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.047 ⁜ Vol. 10, Issue 6, June 2022

DOI: 10.17148/IJIREEICE.2022.10605

Used Space in GB	0.8	1.9	1.4
Available Space in G B	4.2	3.1	3.6

The CSM system has reduced the usage space as 6.66%, 13.88%, 12.19% for sever-1, server-2 and server-3 respectively than the system without using CSM system. The deduplication is carried out to reduce the usage of storage space. The average de-duplication is 3.8 GB.

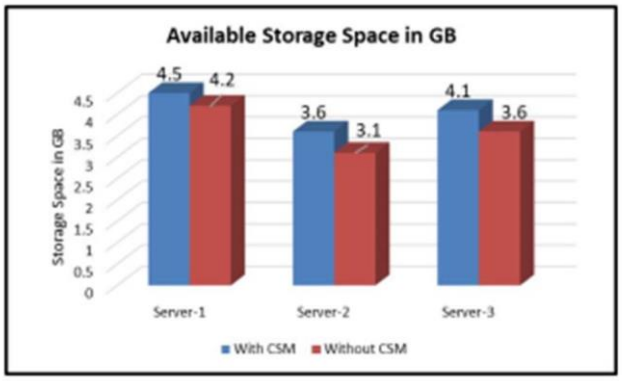

Figure 3: Available Storage Space – Comparative

VI. CONCLUSION AND FUTURE ENHANCEMENT

The Cloud Storage Monitoring (CSM) system is proposed to increase the storage space availability in IaaS-Cloud environment. The frequency of files is quantified and ranked. The frequency and popularity of the files are used for ranking. A prediction algorithm evaluates the ranking of files. The files are moved or archived based on the ranking of the files. The identical files are removed by using the de-duplication technique. The simulated experiments have been carried out with five files with the range from 0.11 MB to 1.00 MB. The CSM system has given better performance as average of 10.91% reduction more than “without using CSM” system and yield the average de-duplication as 3.8 GB. Thus proposed CSM system provides an efficient data storage mechanism. In future this system can be enhanced further for other series such as PaaS, SaaS in cloud computing.

REFERENCES

[1] A.Rajalakshmi, D.Vijayakumar, Dr. K .G. Srinivasagan, An Improved Dynamic Data Replica Selection and Placement in Hybrid Cloud, International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 3, March 2014.

[2] Ali Yadavar Nikravesh, Samuel A. Ajila* and Chung-Horng Lung "An autonomic prediction suite for cloud resource provisioning” Nikravesh et al. Journal of Cloud Computing Advances, systems and Applications, December 2017.

[3] A. Augustus Devarajan and Dr.T. SudalaiMuthu,"Cloud Storage Monitoring System analyzing through File Access Pattern", International Journal of Engineering & Technology (IJET), Vol 7 No 3.32 (2018): Special Issue 32, October 2018.

[4] Jonathan L. Krein, Lutz Prechelt “Multi-Site Joint Replication of a Design Patterns Experiment using Moderator Variables to Generalize across Contexts” IEEE Transactions On Software Engineering, Vol. X, No. X, April 2016.

[5] M. Du and F. Li, "ATOM: Efficient Tracking, Monitoring, and Orchestration of Cloud Resources", IEEE Transactions on Parallel & Distributed Systems, Vol. 28, No.8, pp. 2172-2189, April 2018.

[6] Masoud Saeida, Ardekani, Douglas B. Terry, A Self-Configrable Geo-Replicated Cloud Storage Systems, 11th USENIX Symposium on Operating System Design and Implementation (OSDI’ 14), pp367381, October 2014.

[7] Navneet Kaur Gill and Sarbjeet Singh, Dynamic Cost-Aware Rereplication and Rebalancing Strategy in Cloud System, © Springer International Publishing Switzerland 2015 S.C. Satapathy et al. (eds.), Proc. of the 3rd Int. Conf. on Front. of Intell. Computer. (FICTA) 2014 – Vol. 2, Advances in Intelligent Systems and Computing 328, DOI: 10.1007/978-3-319-12012-6_5, January 2015.

DOI: 10.17148/IJIREEICE.2022.10605

[8] Prabavathy Balasundaram, Chitra Babu and Subha Devi M "Improving Read Throughput of Deduplicated Cloud Storage using Frequent Pattern-Based Prefetching Technique", The Computer Journal Vol. 60, No.3, pp. 444-456, 2017

[9] Ranjana P, George, "A Improving network capacity for effective traffic management using graphs", International Journal of Applied Engineering Research, vol. 10, no.7, pp. 16853-16863, 2015.

[10] Runhui Li, Yuchong Hu, and Patrick P. C. Lee "Enabling Efficient and Reliable Transition from Replication to Erasure Coding for Clustered File Systems" IEEE Transactions On Parallel And Distributed Systems, Vol. pp, No. 99, March 2017.

[11] S.Souravlas, and A. Sifaleras, "Binary-Tree Based Estimation of File Requests for Efficient Data Replication", IEEE Transactions on Parallel & Distributed Systems, Vol. 28, No. 7, pp. 1839-1852, February 2017.

[12] S.Annal Ezhil Selvi and Dr. R. Anbuselvi, A Detailed Analysis of Cloud Storage Issues, International Conference on Mathematical Methods and Computation (ICOMAC 2015), January 2015.

[13] Sathish N, Ranjana P, "Secure remote access fleet entry management system using UHF band RFID", Advances in Intelligent Systems and Computing, vol. 216, pp. 141-149, 2014.

[14] Sathish N, Ranjana P, "Secure remote access fleet entry management system using UHF band RFID", Advances in Intelligent Systems and Computing, vol. 216, pp. 141-149, 2014.

[15] Srinivasan, K., Bisson, T., Goodson, G. R. and Voruganti, K. iDedup: Latency-aware, Inline Data De-duplication for Primary Storage.Proc. FAST'12, San Jose, CA, February 15–17, pp. 1–14, October 2012.

[16] T. S. Muthu and K. R. Kumar, "A Log-based predictive approach for replica replacement in data grid," 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, 2017, pp 1-6.

[17] T. S. Muthu, R. Vadivel, A. Ramesh and G. Vasanth, "A novel protocol for secure data storage in Data Grid environment," Trendz in Information Sciences & Computing (TISC2010), Chennai, 2010, pp. 125-130.

[18] T.SudalaiMuthu and K.RameshKumar, "Hybrid predictive approach for replica replacement in data grid," 2017 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2017.

[19] T. SudalaiMuthu and K. RameshKumar, "A Value based dynamic replica replacement strategy in datagrid," International Journal of Control Theory and Applications, vol. 10, no. 26, pp. 448-462, 2017.

[20] W. Li, Y. Yang, and D. Yuan, "Ensuring Cloud Data Reliability with Minimum Replication by Proactive Replica Checking", IEEE Transactions on Computers, Vol. 65, No. 5, pp. 1494-1506, February 2017.

[21] YaserMansouri, Adel NadjaranToosi, and Rajkumar Buyya "Cost Optimization for Dynamic Replication and Migration of Data in Cloud Data Centers" IEEE Transactions On Cloud Computing, Vol. pp, No. 99, January 2017.

[22] Zheng Yan, Lifang Zhang, Wenxiu Ding, and QinghuaZheng, "Heterogeneous Data Storage Management with De-duplication in Cloud Computing" IEEE Transactions on Big Data, Vol. pp, No.99, May 2017.