International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

A Cloud Storage Monitoring System using Deduplication and File Access Pattern

Augustus Devarajan A, SudalaiMuthu T

Abstract*: Cloud computing is important on current demanding business requirements and it has been emerged as unavoidable technology. The usage of memory storage for Cloud Computing IaaS Service is expanding exponentially every year. The cloud storages are used by the cloud user due to less storage cost compare with other storage methods. The replication of files provides high availability, reliability. It helps in attractive the data availability which reduces the overall access time of the files, but at the same time it occupies more storage space and yielded high storage cost. The cloud user holds storage space twice than which is needed. It is a dire need of a system to find unwanted files in the cloud and the frequency of file access in order to optimize the storage space in cloud. This paper proposed a system, Cloud Storage Monitoring (CSM) system, which is monitoring the IaaS storage usage and analyzes 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 proposed system generates a ranking for each file which is also predicts future access pattern. The CSM system generates a dashboard to the user to reorganize or delete or archive the files to eliminate duplicate files in the cloud storage which can increase the availability of the files. The CSM system is experimented in the Cloud Sim environment, the results showed that the availability have been increased and the storage space is reduced as 10.91% lower than the normal system.

Keywords*: Cloud Computing, Cloud Storage, File Access Pattern.

I. INTRODUCTION

The data replication service of cloud storage duplicates the records in real time to increase the availability of the records which in turn increases the hardware cost. The data replication service consists of data replication, file replication, cloning infrastructure and remote storage replication. The cloud storage replication services 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 improves through the replication procedure. The replications also increase delay in request and response transaction in cloud environments. The predictive auto-scaling systems forecasting future storage workload of the cloud service and adjusting by compute storage capacity in order to meet the future needs. The system also generates a dashboard depends upon the future requirements usage which can be calculated through the performance indicator's value.

Manuscript published on 30 September 2019
* Correspondence Author
Augustus Devarajan A*, Research Scholar, Department of Computer Science and Engineering, Hindustan Institute of Technology and Science, Chennai, India. Email: aug_aa@yahoo.com
SudalaiMuthu T, Associate Professor, Department of Computer Science and Engineering, Hindustan Institute of Technology and Science, Chennai, India. Email: sudalaimuthut@gmail.com

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, 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 recognizes the frequent access patterns on history of usage. 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 can be extended to improve the consistency and replicating process with popular blocks of data files having similar properties.

III. CLOUD STORAGE ARCHITECTURE

The cloud architecture provides a cloud solution structure which is comprise of both "on-premises" and "cloud storage" resources as shown in Figure 1.

The software having different component services, middleware, and geo-location services. It is externally having visible parameters on the relationships between interfaces [5].

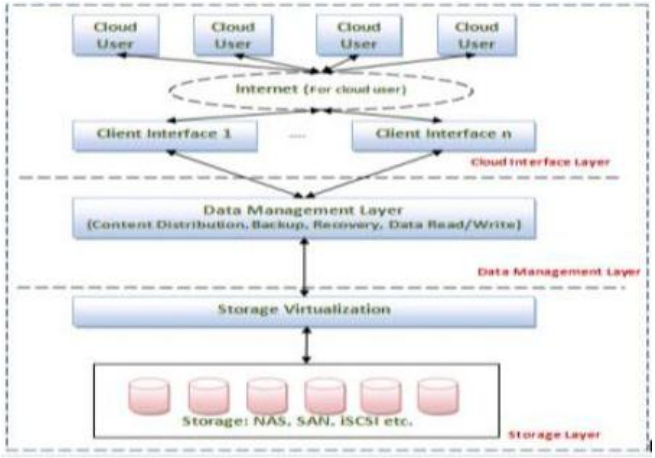

Fig.1 Cloud Storage Architecture

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 also has the authorization and authentication techniques in order to authenticate and validate the users with their credentials by using the single sign-on options.

B. Cloud Data Management Layer

The Data Management Layer also represents as a software layer which is used to manage and validate data of particular cloud users with respect to architecture and activities like data storage, data partitioning, synchronization, and content distribution across storage location, controlling data movement over the network, maintaining consistency backup and data recovery replication [6]. It provides capable data access with distribution of parameters values for the data layer. It is also stores repeatedly used SQL statements in memory areas by using the metadata in terms of performance reduction which can avoids the need of additional time-consuming recompilation at run-time. This layer data encrypted by using SHA keys while update in the database which can have backed up and automatically incorporate the security features.

C. Cloud Storage Layer

This cloud storage layer majorly consists of two parts: Storage virtualization and Basic storage.

Storage Virtualization: It gives maps towards distributed on heterogeneous storage devices and hardware having single allocated 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 the hardware devices which can encompass of different database servers and storage devices which having heterogeneous nature such as DAS, SAN, NAS etc. This storage also compiles on architecture layers on storage classification.

D. Design Principles of Cloud Storage:

The main design principles of cloud storage with requirements such as availability, scalability, cost reliability, simplicity, multi-tenancy, 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 from 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 specific boundaries with the actual requirement on physical storage allocation at the moment. When 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 drives constraints in energy consumption.

These cloud storage requirement services will be accessible at any point of time when user requires IaaS storage. The 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 can be residing on the different servers in different locations it can prevent a single point of failure. Suppose if primary system fails, backup system needs to take over. It increases the data availability features for the 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. In this case the cloud storage is commonly used without other parts of IaaS Storage. This productive features such as de-duplication increase storage utilization, thin provisioning and reduces the amount of storage.

E. Challenges on Cloud Storage:

Nowadays most of organizations are 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 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 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 challenge on the principles. Most of organizations uses 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 challenges in many organizations [14]. Most of enterprise having concerns on-premises critical applications. Vendor lock-in begins use on cloud IaaS storage vendor, the data transmission to a different vendors having in-house 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.

- Identifying the frequency on access pattern
- Provide prediction on file access
- Identifying the duplication of files on cloud storage
- Making storage efficient system.
- Increasing efficiency of the system.
- Improving search experience
- Blocking duplication of files in future

The proposed research work, CSM system is rank the files based on their popularity and the frequency of access [15]. The system generates ranking report of the files and helps to optimize the storage space and availability. The CMS system reduce the storage space by de-duplication and increase the availability by keep 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 in the 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

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(Extn)	File Types	File Size(MB)	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

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

- Comparison based on Files attributes
- Comparison based on delta version and hashing
- Data De-duplication

A. Comparison based on Files attributes

This technique is used to reduce the duplication of data at the file level. It is comparing parameters of the file system 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 identify the files which are having same attributes [17]. This process running 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 de-duplication compares individual files on inside data and variances within the files also compared 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.

C. Data Deduplication

The duplicate data elements have been removed through the deduplication algorithms or it can increase bandwidth through single instancing. The de-duplication process used in the cloud server can reduce the space requirement of the server. Data de-duplication can implement in many forms. There is multiple de-duplication strategies has been followed in different organizations. Before implementing the de-duplication techniques it is very essential 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 of redundant files from storage archives.

Data comparison has detected the duplicate copy by comparing files. If any match has been identified, then the file is discarded.

Data Compression: Data compression technique mainly used to compress the given file by reducing the size of files but not to recognize 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.

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 checked 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 be large amount date of file having redundancy.

Data Comparison: The de-duplication methodology has detected the duplicate copy of the file by comparing the first 50 bytes of the new and existing files and last 50 bytes of new and existing files. This comparison is done one by one byte with existing files in the IaaS storage.

V. RESULTS AND REPORTS

The proposed framework which has been experimented in Cloud Sim with five files and 3 storage servers with each 5GB storage. The test results are captured in Table 3.

Table.3 Simulation results with recommended CSM system

S.No	File Name(Extn)	File Types	File Size in MB	Servers			No. of Duplication
				Server1 5GB	Server2 5GB	Server3 5GB	
1	Main.java	docx	0.08	0	5	0	5
2	Check_ds	pdf	0.12	0	0	2	2
3	Image_01	jpeg	0.42	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 represented in Figure 2. The recommended CSM system yielded better performance in utilizing the storage space using de-duplication s. There are five different files with the size of 0.08 MB, 0.12 MB, 0.42 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 Name(Extn)	File Types	File Size in MB	Servers		
				Server 15GB	Server25 GB	Server35 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
Used Space in GB				0.8	1.9	1.4
Available Space in GB				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 de-duplication is carried out to reduce the usage of storage space. The average de-duplication is 3.8.

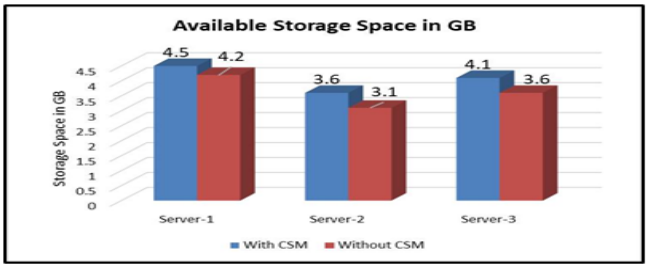

Fig.2 Available Storage Space – Comparative

VI. CONCLUSION

The Cloud Storage Monitoring (CSM) system is proposed to reduce the storage space in IaaS-Cloud Environment. The CSM system has improved the performance of IaaS storage through detecting mechanism. The identical files are removed by using the de-duplication techniques which is supporting authorized duplicate check in hybrid cloud architecture. A prediction framework is design to evaluate the ranking of files which is quantified and ranked based on the frequency of files. It also recommend the user to take easy decision on file related to moved or archived based on the ranking dashboard. 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 yielded the average de-duplication is 3.8. Thus proposed CSM system provided an efficient data storage. In future this system can be enhanced further for other series such as PaaS, SaaS in cloud computing.

REFERENCES

1. 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 (2017).
2. S.Annal Ezhil Selvi, Dr. R. Anbuselvi, S.Annal Ezhil Selvi, R.Anbuselvi, International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 4, Issue 9, September 2017.
3. 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.
4. 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.
5. Prabavathy Balasundaram∗, Chitra Babu and Subha Devi M "Improving Read Throughput of Deduplicated Cloud Storage using Frequent Pattern-Based Prefetching Technique" Advance Access publication on 18 March 2016.
6. Srinivasan, K., Bisson, T., Goodson, G. R. and Voruganti, K. (2012) iDedup: Latency-aware, Inline Data De-duplication for Primary Storage.Proc. FAST'12, San Jose, CA, February 15–17, pp. 1–14.
7. 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, 2017.
8. 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, 2017.
9. 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.
10. 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, 2016.
11. 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.
12. S.Annal Ezhil Selvi and Dr. R. Anbuselvi, An Analysis of Data Replication Issues and Strategies on Cloud Storage System, International Journal of Engineering Research & Technology (IJERT), NCICN-2015 Conference Proceedings, pp18-21, March 2015.
13. 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, Month 2015.
14. Navneet Kaur Gill and Sarbjeet Singh, Dynamic Cost-Aware Re-replication 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. Comput. (FICTA) 2014 – Vol. 2, Advances in Intelligent Systems and Computing 328, DOI: 10.1007/978-3-319-12012-6_5, 2015.
15. 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.
16. John D. Cook , Robert Primmer and Ab de Kwant, Compare Cost and Performance of Replication and Erasure Coding, WHITE PAPER , Hitachi Review Vol. 63, July 2014.
17. 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), pp367-381, October 2014.