ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

Ranking Algorithm Based on File's Accessing Frequency for Cloud Storage System

Mrs. S.Annal Ezhil Selvi[1], Dr. R. Anbuselvi[2]
Assistant Professor in Computer Science, Bishop Heber College, Trichy, TamilNadu, India. [1]
Associate Professor in Computer Science, Bishop Heber College, Trichy, TamilNadu, India. [2]

Abstract*:* The number of cloud storage users has improved abundantly at recent times. The reason behind is, the Cloud Storage system minimizes the burden of maintenance and it has less storage cost compare with other storage methods. It provides high availability, reliability and it is most suitable for high volume of data storage. In order to provide high availability and reliability, the systems introduce redundancy. In replicated system, the cloud storage services are hosted by multiple geographically distributed data centres. But the file Replication is rendering little bit threat about the Cloud Storage System for the users and for the providers it is a big challenge to offer efficient Data Storage. Since the increasingly expanded utility of Cloud storage, the improvement of resources management in the shortest time to respond to the user's requests and the geographical constraints are of prime importance to both the Cloud service providers and the users. The data replication 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 storage cost. In order to rectify the above mentioned problems, need to identify the popularity of the file. So this paper proposed new ranking algorithm which lists the most often accessed files and less frequently accessed files. In future the least accessed a file's replications going to be reduced likewise most accessed file's replications going to be increased based on their SLA.

Keywords: Cloud Computing, Data Replication, Popularity Degree, Distribution Networks Cloud Storage, Storage Cost.

I. INTRODUCTION

Cloud computing is a system which works "on demand" or "Pay per Use" concept. In cloud computing all the computational resources (like storage, data) are shared among the users [1, 2, 3 and 4].Service Level Agreement (SLA) is connecting the user and the service provider. This agreement defines QoS parameters (like availability, Reliability, Scalability and cost etc.).

Cloud storage is a representation of data storage in which the digital records are stored in logical collection. The physical storage data stored on multiple servers (and frequent locations), manage by a hosting company [5]. These cloud storage sources are responsible for assuring the records available and accessible. Peoples and organizations buy or let storage capacity beginning the providers to store user. Today, popular Internet companies, such as Google, Yahoo, and Microsoft offers more services for millions of users every day. These services are hosted in datacentres that contain thousands of servers, as well as power delivery (and backup) and networking infrastructures. Because users demand high availability and low response times, each service is mirrored by multiple datacentres that are geographically distributed [6]. Each datacentre is supposed to serve the requests of the users that are closest (in terms of network latency) to them. If this datacentre becomes unavailable or unreachable, these requests are forwarded to a mirror datacentre.

A Cloud storage data replication service is a managed service in which stored or archived records is duplicated in real time over a storage area network. Further terms for this type of service consist of file replication, data replication, and remote storage replication. The appearance can also refer to a program or grouping that facilitates such duplication. Cloud Storage replication services provide an extra determine of redundancy that can be invaluable if the main storage backup system fails. The instant of that the cloud user can access to the replicated data to minimize downtime and its associated costs [7]. The services, if accurately implement, can clustering based make more efficient disaster recovery process by generating a replica copy of the entire backed-up files on a continuous basis [8].

In cloud inconvenience are normal slightly incomparable, so high availability; high performance and high fault tolerance are important factors to be considered. Concept of replication is used in order to get high availability, high performance and high fault tolerance [10]. Replication is the method to store multiple copies of a data files at data centers for performance and availability reasons. Seeing that cloud is on demand model, therefore the user will pay for using the cloud storage. User will prefer that service provider who will guarantee their maximum demand about the data storage. As the result, replication is used to reach highest availability [1 to 4, 9 and 10]. But at the same time it is not needed that the benefits accrued from the replication

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

will be greater than the cost incurred. Hence cost of replication is essential concept to be considered [2].

This research work proposed Ranking Algorithm for cloud storage system which is used to rank the files based on their popularity. This ranking algorithm lists the most often accessed files and less frequently accessed files. In future the least accessed a file's replications going to be reduced likewise most accessed file's replications going to be increased based on their SLA .

The rest of paper is organized as follows. The related works are presented in section 2. Existing system is described in section 3. The proposed files Ranking Algorithm is described in section 4. Results and discussions are described in section 5. Finally, section 6 concludes the work with future scope.

II. RELATED WORKS

As mentioned in previous work of Annal et al. (2015) [2], there are two major strategies used to obtain a replication system in cloud storage. They are Static mechanism and Dynamic mechanism. In static method of replication the availability and reliability is predetermined. But unwanted use of storage, no flexibility, no scalability and high amount of cost received from the user for storage. In dynamic methods the availability is high and the percentage of problem is less compare with static methods. There is number of dynamic heuristic methods used to reduce the percentage the problems.

Navneet et al. (2014) [11] developed an algorithm named as Dynamic Cost-aware Re-replication and Re-balancing Strategy (DCR2S). The algorithm optimizes the cost of replication using the knapsack problem concept. The algorithm is simulated using CloudSim tool. Rajalakshmi et al. (2014) [12] proposed an algorithm for dynamic data replication in cloud. The replication management system allows users to create, register, manage and update the replication if the original datasets are modified. The proposed algorithm is suitable for optimal replication selection and placement to increase availability of data in the cloud. Replication method used to increase availability of resources, low access cost and shared bandwidth usage. John D. Cook et al. (2014) [13] examines the trade-offs in cost and performance between replicated and erasure-encoded storage systems. Data replication placement mechanism analysed by Zhang et al. (2013) in [14] developed a heuristic algorithm for data replica placement. The simulation shows that the algorithm has a better performance in a storage sensitive environment.

Masoud et al. (2014) in [15] proposed self-configurable geo-replicated cloud storage. Tuba is their replicated key value store which allows application to specify desire consistency and dynamically selects replicas in order to maximize the utility delivered to the read operation. It reconfigures itself automatically while respecting user constraints so that it adapts to changes in users locations or request rates. The work confirms that automatic reconfiguration can yield substantial benefits which are resizable in practice. Zhen et al. (2015) in [16] designed to minimizing data redundancy for high reliable cloud storage system. The evaluations through analysis and experiments validate of their claims are the optimum storage allocation scheme can significantly reduce the search space, the calculation process can be easily simplified, accelerated and the redundancy can be efficiently save by the scheme.

Dhananjaya et al. (2013) [17] implemented an automatic replication of data from local host to cloud System. Data replication is implemented by using HADOOP which stores the data at different data centres. If one node goes down then data can be getting from other places seamlessly. Wenhao et al (2011) [18] designed a novel cost-effective dynamic data replication strategy which facilitates an incremental replication method to reduce the storage cost and meet the data reliability requirement at the same. This strategy works very well especially for data which are only used temporarily and/or have a relatively low reliability requirement. The simulation result shows that replication strategy for reliability can reduce the data storage cost in data centres significantly. Yaser Mansouri et al. (2013) [19] suggested an algorithm that determines the minimum replication cost of objects such that the expected availability for users is assured. And they also developed an algorithm to optimally select data centres for striped objects in such a way that the expected availability under a given budget is maximized.

III. EXISTING SYSTEM MODEL

Navneet et al. (2014) [7] used heterogeneous architecture rather than homogeneous. The data centres in one tier will have different configuration, then the data centres in other tiers. Super data centres stores the original copy of each data file. In order to meet availability, replicas are spread from super to main data centres and from main to ordinary data centres. User sends a file access request to the broker, which furthermore interacts internally with the Replica catalog. Replica catalog contains information where the replicas of requested data file are placed. Replica catalog will send the list of data centres having that file to the broker. Broker will analyse the received list of data centres, and will schedule the request to the nearest data centre. Basic unit of storage is block. Since a file can be replicated at more than one data centre, therefore the blocks of a file will have different block available probabilities at different types of data centres. Super data centres have higher cost, reliable hardware; hence it will have highest block.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

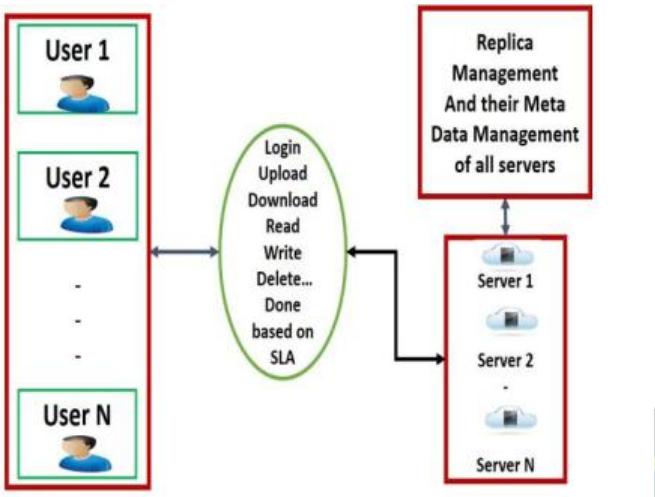


Figure 1 Existing Cloud Storage Architecture

In order to optimize the cost, storage without affecting reliability and availability in future this work was designed file Ranking Algorithm which is used to predict the popularity of the file.

IV. PROPOSED SYSTEM

The important key features of the cloud storage system are high availability, high reliability and high performance. The replication system is used to attain these important key features in CSS. Due to the replication system, the cost of the replication storage and maintenance may be little bit high. So, the users may hesitate to switch over to the cloud storage. Even though recently the number of cloud storage users increases. So the proposed systems plans to optimize the cloud storage to get efficient storage. For that, this research proposed a new file ranking algorithm which is used to rank the file based on their accessing frequency.

A. System Model

In distributed data storage the data files are stored in different data Centre. The data Centre are located in different geographical areas. Each data Centre has additional data Centre such as, secondary or ordinary data Centre. And it all has different managers and catalogs to manage the data warehouse. One important manager is replication manager and replication catalog. Every data centre connected with all other data centre. Figure 2 shows the process when the user initiate to store, access or delete their file on cloud.

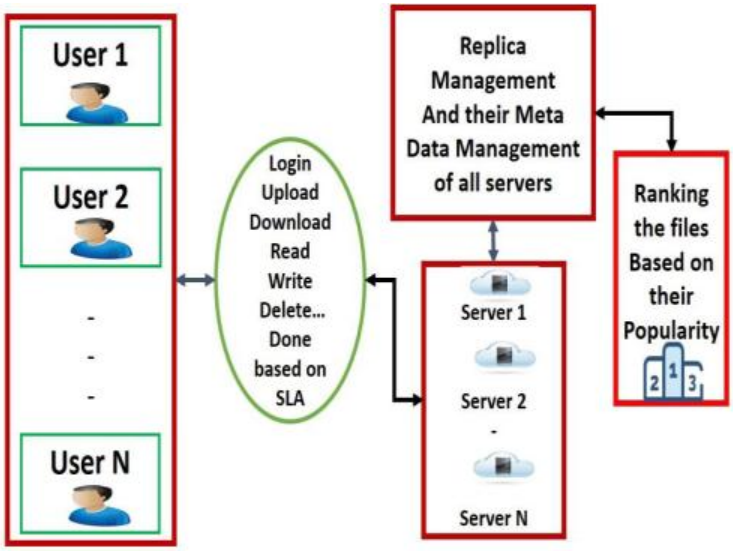


Figure 2 Enhanced Cloud Storage Architecture

Ranking Algorithm:

File ranks (k-Rank)

Input: N (S), N(F), i=0, q, k=0, ∅, t=currentTime

Output: Rank { q's result set}

1. While t >=0
2. If file upload
3. i=1
4. End if
5. While N(S)= no. of iteration i ∈ M(F)
6. If file access
7. ∅ = i+1;
8. End if
9. End while
10. for each N(S) do
11. for each N(S) do
12. for each N(F) do
13. K=k+ ∅
14. End for
15. End for
16. Rank. insert (k)
17. End for
18. If Rank =q
19. Return Rank
20. End if
21. End while

Where S is server in cloud storage system N(S) is number of servers. M (F) is file's Meta data (log file). 'i' is the number access frequency of a particular file, q query result, k sum of access frequency of all servers. '∅' is access frequency of a particular file in all servers individually.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

This algorithm works 24/7 in replication system and deals with Meta data. Initially the ∅ value is 0 then its value changed based on the operation. If the file initiated to store on cloud storage the ∅ value is 1. After that the ∅ value may incremented based on the number of access frequency. Finally all ∅ values are summed together from all servers for each file which is k value.

S.No	File Name	File Type	Frequencies of access per week						$\sum_{i=1}^{n} Si$
			Server 1	Server 2	Server 3	Server 4	-	Server N	Frequencies /week
1	Array_Java	docx	0	4	5	2	-	-	11
2	Tree_ds	mp4	0	0	0	0	-	-	0
3	HelloEnglish	mp3	4	8	2	10	-	-	24
4	CS_C	pdf	2	1	1	1	-	-	5
5	Img_001	jpeg	1	3	0	2	-	-	6

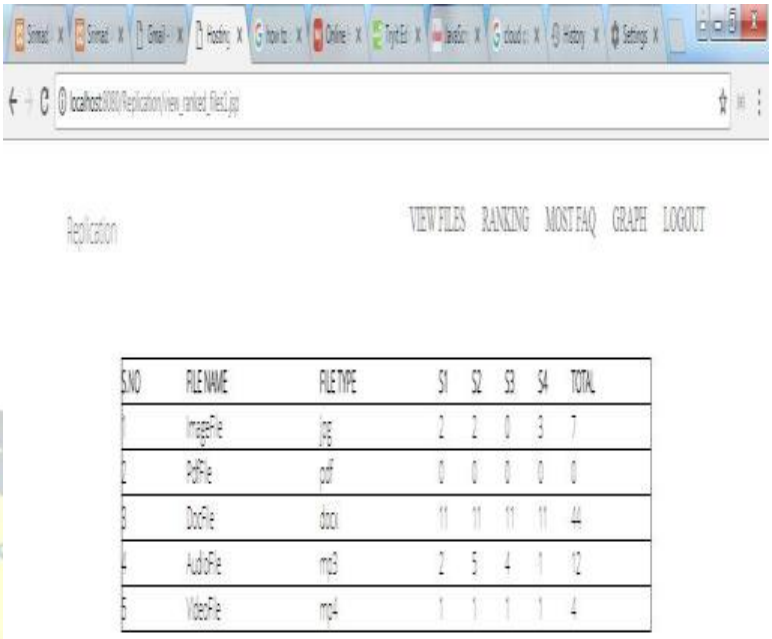
Table 1 File Ranking based on their Popularity

The above table (table 1) denotes the N number of user can avail the Cloud storage services. S_1, $S_2 \ldots S_n$ are the N number of data centre (Servers). The users can store their files (any type) F_n on any data centre (Servers) of cloud storage based on their SLA. The ranking done based on the files popularities from the log file (Meta data) which means how many number of times that files are accessed. That frequencies are calculated individually. Finally, in order to rank that files all individual frequencies are needed to sum by using the following equation.

$$\sum_{i=1}^{n} S\,i$$

V. RESULTS AND DISCUSSIONS

The proposed ranking algorithm implement in CloudMe cloud storage through JSP technology. Finally this researched obtained the expected outcome.



By the above screen shot the research shows the propose work was done successfully. That is, the ranking algorithm ranked the most frequently accessed files and less frequently accessed files. Based on the above result the research can move on the next stage.

VI. CONCLUSION AND FUTURE SCOPE

As mentioned in abstract the Content Distribution Networks have attracted and emerged abundantly in recent year. But it has some issues due to the replication system. So this research plans to increase the efficiency of the cloud storage as well as reduce the threads without affecting the key features of the cloud Storage system. So, this paper proposed ranking algorithm which was ranked the files based on their accessing frequencies. Accessing frequency is nothing but the popularity of the particular file. From the result and discussion section this paper clearly shows the primary objective of this work was obtained. Accessing frequency was calculated in two steps. First, it was calculated for each server for each file. Then it was summed together for each file from all servers. In future the low ranked files which are less accessed files replications going to be reduced. Likewise, the high ranked files which are most frequently accessed a file's replications going to be increased.

REFERENCES

[1]. 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.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

[2]. 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.

[3]. Anuradha.R and Dr. Y. Vijayalatha, A Distributed Storage Integrity Auditing for Secure Cloud Storage Services, International Journal of Advanced Research in Computer Science and Software Engineering, August 2013.

[4]. Ayed F. Barsoum and M. Anwar Hasan, On Verifying Dynamic Multiple Data Copies over Cloud Servers, August 15, 2011.

[5]. 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

[6]. Wenhao Li, Yun Yang "Ensuring Cloud data reliability with minimum replication by proactive replica checking" IEEE Transactions On Computers Manuscript Id

[7]. 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

[8]. 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.

[9]. Nicolas Bonvin, Thanasis G. Papaioannou and Karl Aberer , A Self-Organized, Fault-Tolerant and Scalable Replication Scheme for Cloud Storage, SoCC'10, Indianapolis, Indiana, USA, June 10–11, 2010.

[10]. Priya Deshpande, Aniket Bhaise, Prasanna Joeg , A Comparative analysis of Data Replication Strategies and Consistency Maintenance in Distributed File Systems, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-1, March 2013.

[11]. 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.

[12]. 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.

[13]. 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.

[14]. ZHANG Tao, Data Replica Placement in Cloud Storage System, International Workshop on Cloud Computing and Information Security (CCIS), 2013.

[15]. 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.

[16]. Zhen Huang, Jinhang Chen, Yisong Lin, Pengfei You and Yuxing Peng, Minimizing Data Redundancy for Hign Reliable Cloud Storage Systems, Published by ELSEVIER(COMPNW 5513), No. of Pages 14, February 2015.

[17]. Dhananjaya Gupt, Mrs.Anju Bala, Autonomic Data Replication in Cloud Environment, International Journal of Electronics and Computer Science Engineering, ISSN 2277-1956/V2N2-459-464, Volume2, Number 2, 2013.

[18]. Wenhao LI, Yun Yang and Dong Yuan, A Novel Cost-effective Dynamic Data Replication Strategy for Reliability in Cloud Data Centers, IEEE International Conference, 2011.

[19]. Yaser Mansouri, Adel Nadjaran Toosi and Rajkumar Buyya, Brokering Algorithms for Optimizing the Availability and Cost of Cloud Storage Services, Cloud Computing Technology and science (CloudCom), IEEE 5th International Conference, Volume: 1, pp 581 – 589, , Dec. 2013.

[20]. Zheng Yan, Lifang Zhang, Wenxiu Ding, and QinghuaZheng, "Heterogeneous Data Storage Management with Deduplication in Cloud Computing" IEEE Transactions On Big Data, Vol. pp, No.99, May 2017

Mrs. S. Annal Ezhilselvi received her M.C.A and M.Phil degree in Computer Science from Bharathidasan University, Trichy, and Tamilnadu, India in 2006 and 2011 respectively. And she cleared SET and NET exams in 2016. She was published 1 book and now she is working as an Assistant Professor in Bishop Heber College (Autonomous), Trichy, and Tamilnadu, India. She has 9 years of teaching experience. Her research interest is Cloud Storage. She is now pursuing her PhD under the guidance of Dr. R. Anbuselvi.

Dr. R. Anbuselvi received her PhD degree in Computer Science from Mother Teresa University, Kodaikanal, and Tamilnadu, India in 2013. She is now working as a Assistant Professor in Bishop Heber College (Autonomous), Trichy, Tamilnadu, India. She has 19 years of teaching experience. Her research interests are Artificial Intelligence, Cloud Computing and Data Mining. She is now guiding more than 8 PhD research scholars.