A brief survey on dynamic strategies of data replication in cloud environment: last five year study

[1]Faraidoon Habibi, [2]Nagesh Kumar
[1]M.tech (CSE), [2]Assosiate Professor
[1]Department of Computer Science & Engineering,
[1]A P Goyal Shimla University, Shimla, India

***Abstract*— Data replication provides availability of data in multiple sites to ensure efficient utilization of data. Because of the extreme growth of data usage, the cloud computing is getting more preferences nowadays. As most companies are using cloud computing to store and access data, it is mandatory to backup and replicates data offsite to ensure easy recovery of data in the event of downtime and in case of disaster. And the best practice for this purpose is using data replication which allows organizations to scale their offsite storage quickly for faster backup and recovery. Data replication is the best solution where there is a growing need for faster recovery as it offers high performance, availability, and reliability. Since now, a large variety of strategies have been followed to replicate data. Generally, we can categorize data replication techniques into two group named –static and dynamic replication. In static replication, once we decide strategy, it can't be modified while the operation is going on. To overcome this limitation, dynamic replication is employed which can create and delete replicas dynamically according to the need. Several numbers of researches have been done regarding this two categories of data replication in this review we will analyze some of the dynamic data replication techniques that are used in the cloud environment. Our focus will be on finding a better solution of the previous results in order to increase the efficiency, performance, and reliability of the dynamic data replication in the cloud.**

***Index Terms*— Data replication, dynamic, cloud computing, efficiency, performance.**

I. INTRODUCTION

At the present time, in different scientific regulation, a huge amount of data is an important and crucial part of shared resources. Such enormous mass of data is typically kept in the cloud data centers [1]. Cloud computing comprises of virtualized resources for computing and interconnected through a private network or global network (internet) [8].To increases availability, consistency, and reliability of resources in cloud environment we used data replication and it creates multiple copies of the same data on different sites which is required in cloud[17].Data replication is an appropriate technique used to manage a great deal of data that creates multiple copies of data in multiple sources in order to reduce access time and bandwidth consumption. It also guarantees data reliability and load balancing for the system. Data replication techniques can be divided into two categories, Static Replication and Dynamic Replication [9]. Dynamic data replication is more suitable for cloud environment than static data replication techniques because the static techniques have some limitations such as they do not adapt according to the changes in the environment [2]. They are not suitable for a large amount of data and a large number of users [3]. Dynamic replication strategy is the best solution for a service-oriented environment where the number and location of the users who are going to access data often have to be determined in a highly dynamic way [4] Therefore, it is more preferable for the cloud environment. Several approaches of dynamic data replication have been purposed by many researchers in this recent years. But there wasn't sufficient work done on dynamic data replication specifically for the cloud environment. This paper's main perspective is to do a brief survey of those papers in order to analyze the approaches they used to come up with a better solution that will emphasize the efficiency and performance of dynamic data replication in the cloud.

II. DYNAMIC DATA REPLICATION IN CLOUD

Dynamic data replication is intelligent in making choices about the location of data rely on the information of the current environment [1].Dynamic replication strategy is primarily more appropriate for a service-oriented environment where the number and location of the users who are going to access data should be determined in a highly dynamic manner [4]. It can optimize the use of resources as well as the efficiency by providing a dynamic number of replicas in data cloud system [5]. By combining both I/O and the communication cost it Optimizes total cost [6].Because of the process of intelligent higher cognitive and choosing the location of replica according to the need by considering the surrounding condition it makes dynamic replication ways higher and more efficient than the static replication ways. [7].

III. RELATED WORKS

D2RS (Dynamic Data Replication Strategy): This is a mathematical model which is formulated to show the system availability and the number of replicas relationship which was not available in previous researches [11]. The multi-tier hierarchical cloud system architecture was used to replicate data in cloud storage. In D2RS the replication of data file takes place automatically based on popularity [6].This strategy emphasizes on three parameters which are Bandwidth Expenditure, Availability& Number of Replicas.

It uses the temporal locality theory to make the decision about selection of data file for replication. This technique analyzes the data relevant to access information of the users. Two things are calculated Based on this analysis, the popularity degree and replica factor. With the placing popular data file pursuant to access history the D2RS increases the data availability, task execution of cloud system, response time and decreases the bandwidth consumption [10]. Replicas are placed within data nodes in a balanced way. Experimental results demonstrate the effectiveness of the system developed by the proposed D2RS strategy. After reviewing the whole process we found that there is still some improvement needed for reducing user waiting time, increasing data access and data availability. And this replication needs to be applied in a real cloud platform.

CDRM (Cost Effective Dynamic Replication Management): Qingsong Wei et.al [12] proposed cost-effective dynamic replication management scheme which is known as CDRM, in order to obtain the relationship between the availability and replica number. For a designated availability requirement of a file, it computes and maintains the minimal number of the replica to fulfill the requirement. The measurement of the availability of a file is based on block locations, the current number of blocks, number of replicas, network bandwidth, etc. The replica replacement in CDRM takes place based on data nodes capacity and their blocking probability in an efficient manner. This strategy in the heterogeneous cloud can dynamically redistribute workload across data nodes [12]. It creates new replicas dynamically if the current number of replicas in the data node is less than the minimum replica number and does not satisfy the availability requirement [10]. This technique is implemented on HDFS (Hadoop Distributed File System) the results demonstrate that this strategy is cost-effective and perform better the HDFS default replication management in terms of load balancing for huge volume cloud storage and performance [12]. CDRM is better than default replication technique of HDFS if popularity is small [10]. Although CDRM is an efficient algorithm the consistency of replicas was not ensured by this algorithm properly which needs to be considered in future.

CIR (Cost-effective Incremental Replication): Wenhao Li et al. proposed a cost-effective dynamic data replication strategy that implies an incremental replication technique [10]. The aim of this technique is to reduce the number of replicas as well as to satisfy requirement reliability in order to ensure the cost-effectiveness [13]. CIR generates replicas incrementally while existing replicas cannot ensure the reliability [10]. In this approach cost is the top priority that can be considered as an alternative solution to cloud storage [13]. In this approach, a data storage reliability model was proposed to deal with a large amount of data storage. The result of the practical experiment of this approach showed that CIR can preserve about 2/3rd of the storage cost that is enough conventional. It also saves memory space while storage duration is short. [10]. It does not provide any appropriate solution for data loss issue.

RTRM (Response Time-Based Replica Management): Bai et al (2013) proposed a response time-based replica management strategy which generates a replica for automatically increasing the number of replicas depending on the average of the response time. It predicts the bandwidth of the replica servers while receiving the new request that makes the replica selection accordingly and combines the number of replicas and the network transfer time [4]. RTRM technique focuses on the three issues which are the creation of the replica, selection of replica and placement of replica. This strategy defines a time interval for response time and it will enhance the number of replicas and create new replicas if the response time is higher than time interval [14]. RTRM is a NP-hard problem. This paper proposes a reduction algorithm to solve this problem based on graph theory. To measure the performance of RTRM, strategies were run in OptorSim. The simulation results show that replica management strategy performs better than the five built-in replica management strategies in OptorSim simulator in terms of service response time and network utilization [14].

LRM (locality replication manager): This algorithm suggested to reduce the using resources cost, energy cost and system delay and also enhance the availability of the system. The important task of LRM is to obtain the user's queries, gather condition of nodes in the cluster, and finally select the best host for block placement. LRM carryout this task with the cooperation of its other components and the final decision is made by LRM. HDFS architecture is used by this strategy for replication management. As LRM is compared with other algorithms it is found out that this algorithm performs better than the others in terms of using recourses and energy, providing availability and reducing system delay. LRM consider the needs of quality-of-service (QoS) function, as well as the physical locality of data blocks to obtain more optimum replication parameters [15].

QADR (QoS-Aware Data Replication): Jenn-Wei Lin has proposed two QoS-aware data replication (.QADR) algorithms in order to continuously support an application QoS requirements after data corruption in cloud computing system. The first algorithm for doing the data replication pursue the idea of high-QoS first-replication (HQFR). The next algorithm converts the QADR problem into the well-known minimum-cost maximum-flow (MCMF) problem. As there exist many polynomial-time MCMF algorithms, once the QADR problem converted to the MCMF problem, one of them can be used to achieve the optimal solution in polynomial time.At this point, the optimal algorithm is also known as the MCMF Replication (MCMFR) algorithm. MCMFR algorithm can increase the average recovery time of Hadoop algorithm by about 71% because it has the lowest average recovery time. They proffer node combination technique in order to make the proposed replication algorithm work on large-scale cloud computing system. Results of the simulation demonstrate that the recommended replication algorithms can efficiently execute the QoS-aware data replication in cloud computing systems. In the future, they plan to implement the recommended QADR algorithms in a real cloud computing platform in future. Furthermore, the replication algorithms will be also enhanced to focus on energy consumption [16].

IV. COMPARATIVE ANALYSIS

All In this paper we summarized the model used and the parameters considered in the above methods as well as the outcomes achieved to make a general comparison. The main goal is to think about the future works that can be done to improve the above

techniques and suggest some solution to existing problems. A summarization of the above algorithms are shown in the following table 1.

TABLE 1. COMPARATIVE ANALYSIS OF DYNAMIC DATA REPLICATION TECHNIQUES IN CLOUD ENVIRONMENTS

	Used model	Parameters	Outcomes	Future works
D2RS (Dynamic Data Replication Strategy)	• multi-tier hierarchical cloud system architecture • temporal locality theory	• Bandwidth consumption • Availability • Number of Replicas	• Increases data availability, task execution, response time • Decreases the bandwidth consumption • Replicas are placed in a balanced way.	• Improvement needed for reducing user waiting time • Should increase data access and data availability. • Needs to be applied in the real cloud platform.
CDRM (Cost Effective Dynamic Replication Management)	• HDFS(Hadoop Distributed File System)	• **B**lock locations • Current number of blocks • Number of replicas • Network bandwidth	• Cost effective. • Better performance than HDFS • Efficient	• Consistency of replicas was not ensured
CIR (Cost-effective Incremental Replication)	• Data storage reliability model	• Number of replicas • Requirement reliability • Cost-effectiveness	• Deal with large amount of data storage • Preserves about 2/3rd of the storage cost • Saves memory space for short storage	• Does not provide any appropriate solution for data loss issue.
RTRM (Response Time-Based Replica Management)	• Reduction algorithm • Graph theory	• Replica creation • Replica selection • Replica placement	• Better performance • Enhance number of replicas • Better response time and network utilization.	
LRM (locality replication manager)	• HDFS architecture	• Quality-of-service • Physical locality of data blocks	• Better than others in terms of using recourses and energy • Provides availability • Reducing system delay	• Needs to be implemented in the real cloud system.
QADR (QoS-Aware Data Replication)	• high-QoS first-replication (HQFR) • Minimum-cost maximum-flow (MCMF)	• Recovery time • Quality-of-service •	• Increase the average recovery time. • Works on large-scale cloud computing system	• Needs to be implemented in the real cloud system. • Should focus more on energy consumption. •

.

V. CHALLENGES & FUTURE SUGESSIONS

Still, there are many improvements need to be done in order to overcome the existing drawbacks of the dynamic data replication techniques. One of the major challenges we need to focus on is data loss prevention. Most of the previous works didn't give enough importance to data loss during replicating data in the cloud which should be a major consideration. In future, some modification can be applied in the replication techniques to deal with data loss issue in order to ensure more reliable replication. To reduce the probability of accidental data loss we should try to apply synchronize data replication in the cloud. Maintaining the quality-of-service (QoS) can be considered as another challenge for dynamic data replication. When data corruption occurs in the cloud, the QoS requirements of the application cannot be ensured [16]. The QoS awareness needs to be focused more during data replication in the cloud. The rate of change (ROC) is the amount of data that changes over time and needs to be replicated on the other sites. This is another challenge of data replication where data changes dynamically in the cloud environment. The ROC is used to determine the speed of data replication therefore, further improvisation should be done to analyze the ROC in order to speed up the replication. The time consumption of creating replicas was not given much priority in any of the previously proposed methods. In dynamic data replication, it is another challenge for future consideration. The replica creation should be fast enough to improve the availability and disaster recovery to remove the latency which is a major problem in the cloud. A more significant dynamic data replication model

needs to be developed for reducing the time required to create replicas. It has to be made sure that All the existing dynamic data replication techniques where experimented are implemented in the real cloud.

VI. CONCLUSION

In this paper, we did a brief review on some popular dynamic data replication techniques which are more suitable for the cloud environment. We analyzed several researches and found out six major dynamic data replication strategies for cloud computing each of this method focused some specific parameters such as Number of replicas, bandwidth consumption, availability, recovery time, quality-of-service etc. we summarized the features of all the techniques then we found out that some future improvisation and modification can be done to increase the efficiency and performance of dynamic data replication in the cloud. Most of all each of the techniques should be implemented in the real cloud environment. On the other hand, we should concentrate more on the cost-effectiveness and quality-of-service of the replication techniques. None of the methods have given much priority on time consumption of creating replicas which can be a future consideration. The future works that are mentioned in this paper can be implemented practically as research works.

REFERENCES

[1] Bahareh Alami Milani, Nima Jafari Navimipour, A Systematic Literature Review of the Data Replication Techniques in the Cloud Environments, In Big Data Research, 2017, ISSN 2214-5796,

[2] Bahareh Rahmati, Amir Masoud Rahmani , Ali Rezaei, Data Replication-Based Scheduling in Cloud Computing Environment ,Journal of Advances in Computer Engineering and Technology, 2017

[3] Nazanin Saadat, Amir Masoud Rahmani, PDDRA: A new pre-fetching based dynamic data replication algorithm in data grids, In Future Generation Computer Systems, Volume 28, Issue 4, 2012, Pages 666-681

[4] Bahareh Alami Milani, Nima Jafari Navimipour, A comprehensive review of the data replication techniques in the cloud environments: Major trends and future directions, In Journal of Network and Computer Applications, Volume 64, 2016, Pages 229-238

[5] Reza SOOKHTSARAEI, Ahmad FARAAHI, Hadi ADINEH "A locality-based replication manager for data cloud, Frontiers of Information Technology & Electronic Engineer, 2015

[6] Y. Huang and O. Wolfson, "A competitive dynamic data replication algorithm," Proceedings of IEEE 9th International Conference on Data Engineering, Vienna, 1993, pp. 310-317.

[7] Sonali Warhade, Prashant Dahiwale, M.M. Raghuwanshi, A Dynamic Data Replication in Grid System, In Procedia Computer Science, Volume 78, 2016, Pages 537-543, ISSN 1877-0509

[8] Khosravi, A & Buyya, R. (2017). Energy and carbon footprint-aware management of geo-distributed cloud data centers: A taxonomy, state of the art, and future directions. DOI: 10.4018/978-1-5225-2013-9.ch002.

[9] MyunghoonJeon, Kwang Ho Lim, Hyun Ahn, Byong Dai Lee, " Dynamic Data Replication Scheme In Cloud Computing Environment", Second IEEE Symposium on Network Cloud Computing and Applications, London, UK, November 2012,pp. 40-47

[10] Rashmi R Karandikar and M B Gudadhe. Article: Comparative Analysis of Dynamic Replication Strategies in Cloud. IJCA Proceedings on Trends in Advanced Computing and Information Technology TACIT 2016(1):26-32, August 2016.

[11] Sun DW, Chang GR, Gao S et al. Modeling a dynamic data replication strategy to increase system availability in cloud computing environments. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 27(2): 256{272 Mar. 2012. DOI 10.1007/s11390-012-1221-4

[12] Qingsong-Wei, Bharadwaj Veravalli, Bozhao Gong, Lingfang Zeng, Dan Feng," CDRM: A cost effective dynamic replication management scheme for cloud storage cluster", Proc. IEEE International Conference on Cluster Computing, Herakilon, Crete, Greece, September 2010, pp.188-196.

[13] W. Li, Y. Yang and D. Yuan, "A Novel Cost-Effective Dynamic Data Replication Strategy for Reliability in Cloud Data Centres," 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure

[14] Bai X., Jin H., Liao X., Shi X., Shao Z. (2013) RTRM: A Response Time-Based Replica Management Strategy for Cloud Storage System. In: Park J.J..H., Arabnia H.R., Kim C., Shi W., Gil JM. (eds) Grid and Pervasive Computing. GPC 2013.

[15] Reza SOOKHTSARAEI, Ahmad FARAAHI, Hadi ADINEH "A locality-based replication manager for data cloud, Frontiers of Information Technology & Electronic Engineer, 2015.

[16] J. W. Lin, C. H. Chen and J. M. Chang, "QoS-Aware Data Replication for Data-Intensive Applications in Cloud Computing Systems," in IEEE Transactions on Cloud Computing, vol. 1, no. 1, pp. 101-115, Jan.-June 2013.

[17] Nagamani H Shahapure and P Jayarekha. Article: Replication: A Technique for Scalability in Cloud Computing. International Journal of Computer Applications 122(5):13-18, July 2015