Concept of minimizing the response time for reducing dynamic data redundancy in cloud computing

Wan Nurazieelin Wan Abd Manan, Mohamad Aizi Salamat
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia

Article Info	ABSTRACT
	Reduction of dynamic data redundancy in cloud computing is one of the best ways to maintain the storage capacity from being fully utilized. Cloud storage is a part of cloud computing technology which holds a high demand in any organization for reducing the cost of purchasing and maintaining storage infrastructures. Increase in the number of users will require a larger storage capacity for storing their data. Reduction of dynamic data redundancy allows service providers to be energy savvy and minimize maintenance cost. Recent researches focus more on static data nature despite its limited capability as compared to dynamic data characteristic in cloud storage. Therefore, this paper theoretically compares various techniques for reduction of redundant dynamic data in cloud computing and suggests the best technique for completing the task in terms of response time.

Corresponding Author:

Wan Nurazieelin Wan Abd Manan, Mohamad Aizi Salamat
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johor, Malaysia.
Email: gi180019@siswa.uthm.edu.my, wannurazieelin93@gmail.com , aizi@uthm.edu.my

1. INTRODUCTION

The Fourth Industrial Revolution (4th IR), also known as Industrial Revolution 4.0 (IR 4.0) has been a hot issue in Malaysia recently. Most aspect in our daily life is affected by the digital lifestyle and the Internet of Things (IoT) which is brought by the 4th IR [1]. Malaysian agencies such as the Malaysian Digital Economy Corporation (MDEC) are welcoming the change by providing numbers of initiatives to accelerate the process towards a digital lifestyle. In the meanwhile, cloud computing has expanded drastically, especially in terms of pooling and sharing of processing resources, network infrastructure, server, storage, services, and application. Since cloud computing technology has advanced, a number of organizations have switched their computing environment either partially or completely into the cloud-based system [2]. Cloud computing services has been at high demands due to a cheaper cost of services, scalability, high performance, reliability, as well as availability [2].

Generally, a lot of technique of reducing dynamic data redundancy either in aspect of security, accuracy, response time, availability and scalability by previous researcher. Through this paper, more focusing on review the technique for reducing redundancy dynamic data based on response time. Some of the studies focus on either reducing redundancy for static data-like archive or backup data only; or reducing redundancy for dynamic data on a more flexible operation such as insertion, modification, and deletion operation. Over the years, many approaches and techniques have been suggested with a variant of solutions, yet there are very few efforts to identify the best way to reduce data redundancy for dynamic data in cloud computing by comparing same technique such a deduplication technique for dynamic data in cloud computing with different data set.

A layout-free deterministic Verifiable Data Redundancy Protocol (VDRP) has been proposed where users can verify the deployed level of data redundancy remotely without having the file layout information in a cloud storage system [3]. VDRP was designed for static data, which makes it unsuitable to be used for dynamic data. In order to determine replication possession in the server, VDRP does not rely on the system response time [3]. Hence, improvements such as extending the verifiable data redundancy for dynamic data and implementing response time is required as alternative ways to measure the performance of verifying data redundancy. The improvement ideas relied on [4] which as a cloud service provider, nine (9) service need to provide to consumer which are application, data, runtime, middleware, operating system, virtualization, Server, Storage and Networking. From this research, considering runtime is the service need to enhance.

Several researchs works come with a different technique to solve the data redundancy for cloud computing for dynamic data nature. Most of the research works [3, 5-9], focuses on static data, while few focuses on dynamic data in cloud computing [2, 10-], some of researcher implement on mobile cloud computing [13]. In a meantime, dynamic data nature in cloud computing may give some impact on cloud storage when it stored redundant data without having a proper way for reducing it. Hence, through this research, we suggest an excellent technique to reduce redundancy data for dynamic data nature in cloud computing for maintaining storage and energy consumption for some providers or any consumer need do maintenance on their cloud storage based on response time.

Dynamic data redundancy in cloud computing really exists based on research works [14], to prove redundancy dynamic data in cloud computing through map-based provable multicopy dynamic data possession (MB-PMDDP) scheme. Besides, to explain about minimizing response time for redundancy data in cloud computing based on previous research works [15] by preventing redundant data being stored, it took of hours to tackle the challenge for data deduplication technology takes effect in storage device. Based on research need by [1] and deficiency in storage devices stated in [15], this research will measure performance based on response time to overcome the problem of response time challenge on minimizing redundancy dynamic data redundancy in cloud computing. In a business perspective, time is a valuable resource and constraint.

Objectives of this research are to (i) analyse current technique of reducing data redundancy for dynamic data in cloud computing, (ii) compare the response time of reducing data redundancy in cloud computing, and (iii) propose a better technique of reducing data redundancy in cloud computing based on current technique being used by researchers. Different researchers are suggesting a variety techniques for reducing data redundancy in cloud computing either for static data or dynamic data. Therefore, this paper offers a review and comparative analysis on current techniques and proposed a new technique based on best result from current technique being used by previous researchers which one can reduce data redundancy for dynamic data in cloud computing based on response time effectively.

The paper consists as follows: Section 1 covers the introduction and general brief about the work. Section 2 covers the related works, basic concepts of data redundancy for dynamic data in cloud computing, and previous works on the ways to reduce dynamic data redundancy reduction in cloud computing. Section 3 presents the comparative analysis of various data redundancy reduction techniques. Section 4 presents the conclusion and future works.

2.   REVIEW OF RELATED WORK

Selvi and Anbuselvi [16] states cloud computing is a practice that offers more adaptability in infrastructure and reduces cost than our traditional computing models [16]. Haoran et al. [5] state cloud computing provide a flexible, virtual and scalable resource of managing mode for internet enterprise. Cloud computing is fundamental for the storage architecture of cloud storage. Cloud storage is projected separately to attain the highly available and scalable storage. Cloud storage has supplied the enterprise a data warehouse solution. Nevertheless, when cloud storage becomes a centralized data store, huge storage space is needed to store the data. Xu and Tu [9] states that an increase of data in cloud storage with redundant data is a reason for a waste of plenty of storage. Islam and Hassan [3] states data redundancy is the common approaches for fault tolerance. Redundant data in distinct physical systems ensures fault-tolerant for data storage. However, the bigger size of data in cloud computing, a reduction in data volumes could help cloud computing providers to reduce the costs of running large storage system and save the consumption of energy stated by Leesakul et al [11]. Hence, data deduplication technique has been implemented in cloud storage to improve storage efficiency. By identifying and analyzing of reducing redundant dynamic data technique from previous researcher either by using combination of replication and erasure code by Suresh Patil, et al., [2] or DelayDedupe mechanism by Xu and Tu [12], or Convergent encryption, bilinear map and Merkle hash tree structure by Wu et al. [10] or deduplicators, cloud storage and redundancy manager method by Leesakul et al. [11].

3.   COMPARATIVE ANALYSIS

Table 1 summarizes the problems to be solved and the corresponding techniques used to handle the problem.

Table 1. The Comparative Analysis of Different Dynamic Data Redundancy Reduction Techniques

Authors, Year	Discussed Issue	Method	Results	Disadvantage/ Future Works
SureshPatil, et al., 2017 [2]	Managing data redundancy by using data distribution replication technique or erasure code scheme	Combination of replication and erasure code (Data owner, File Verification, Versioning, Hybrid Redundancy, HyRD)	Redundant data can be reduced and data can be maintained by implementing MD5 algorithm which is used for verification of the hash values of file and file version to maintain the availability and durability of data.	Need to integrate security features to the system and provide access control policies.
Wu et al., 2017 [10]	To suggest a technique that binds data operations dynamically with data deduplication in the privacy-preserving public auditing for protected cloud storage	Convergent encryption, Bilinear map and Merkle hash tree structure	Combination of dynamic data operations with data deduplication mechanism in privacy-preserving public auditing for secure cloud storage	No reference of performance response time in the work.
Xu and Tu, 2015 [12]	To improve the deduplication ratio and extended the DelayDedupe strategy, Chunk level deduplication are based on delayed target-deduplication for cloud storage system	Delayed deduplication method in snode	Storage load made more stabilize and response time of the system decreased by DelayDedupe	Need to optimize the response time to accurately predict hot-duplicated chunk.
Leesakul et al., 2014 [11]	To achieve a balance between fault tolerance demands and storage efficiency. To enhance efficiency in cloud storage systems that experience changes in data scenarios and user patterns.	Load Balancer, deduplicators, cloud storage and redundancy manager	A dynamic data deduplication scheme for cloud storage and improved performance in cloud storage systems.	Need to evaluate the availability and performance of the system (a dynamic data deduplication scheme).

The first comparison was done by using a combination of replication and erasure code by SureshPatil, et al., [2]. There are 4 modules have been proposed which are owner of data module, verification of the file module, versioning module, and Hybrid redundancy [HyRD]. For data owner module, in order to determine the owner of data is by identifying the username and password of the owner. Users will be able to download and upload any type of data once the login has been successful. The second module is the file verification module which utilizes the MD5 algorithm to verify the existence of a file in the system with the help of hash value comparison. In the case that the file exists on the cloud server, the path to the file is fetched from database and uploaded to the user's record. This process require 5 steps to generate the hash value of a file, for an example, adding the padding bits and increase the length, initializing MD buffer, process in 16-word block and their corresponding results. The third module is the versioning module which is used to reduce the file overwrite problem. In normal file upload processes, in the case that an existing file has the same name as the file that is about to be saved in the system, the updated file has been overwrited the existing file when we proceed to save the file. it overwrites the existing file when we proceed to save the updated file. But, in versioning module this problem overcome by generating 2 file as a substitute overwriting the old file. The new file updated in the system with the same name of file, the version number and same extension of the old file remain as it is in the server. Besides, the path with its version number of the new file updated to user's record.

The forth and the last module is the hybrid redundancy module. Through this module, after successful operations on saving a file, proceed to the file distribution module, where a hybrid redundancy technique had been use. This particular technique uses different method like replication and erasure code for data distribution in the cloud storage system. During this stage of the process, prior to storing the data, the size of the file is evaluated and categorized into two categories, either small or large file. File verification and file versioning method are then applied to the categorized file to evaluate whether it consist of small files or large files. Cloud storage selecting small size of files which have technique of replication to store data in cloud storage server makes multiple copies of data will store in cloud storage system. This technique helps to increase the durability and availability of the files in order to use the second cloud storage for large file which has erasure coding technique for distributing large files in cloud storage servers. The erasure code divides the object

into 'm' chunks and recodes these particular chunks into larger 'n' chunks. In addition, the original file can be reconstructed from any 'm' chunk which provides high efficiency and high availability.

The second comparison was done by using the DelayDedupe mechanism by Xu and Tu [12], a new approach used by researchers by combining client-side deduplication and target side deduplication. There are 4 modules in Client side. The first module is File Preprocess Module, it classifies the file into different group according to their types and calculate the fingerprints of file by MD5 Algorithm. Secondly, for local deduplication, it plays a role for discarding redundant data successively via the file-level and chunk-level deduplication. Thirdly, "Metadata manager" is responsible to preserve the fingerprints of files and chunks that have been uploaded to Snode in order to prevent duplicated data being uploaded redundantly. Lastly, "File transfer module" is aimed to transfer the metadata of these data processed by local deduplication to MS and it will not upload the new data until the message states "not found" received by MS. In addition, "MS" contains two modules which are Filter module and Update module. For Filter module, it is responsible for filtering redundant data from different clients while the Update module will update the metadata index in MS according to the metadata information that have been modified from Snode.

Furthermore, there are 4 modules in Snode, which are Metadata manager, Store Module, self-check & Report Module and DelayDedupe Module. Firstly, store module will keep the actual data block on the disk, whereas Metadata manager will retain the metadata information, including the reference for the data chunk stored in Snode and fingerprint. In the side note, Self-Check&Report Module will detect the redundant data that have been modified by other users and reports the modified metadata information to MS. Last but not least, DelayDedupe Module is responsible to determine whether the duplicated chunk is hot or not. Normally they won't removed for hot duplicated chunks until they are not hot. To ensure the functionality of each of these module towards this system, Client, MS, and Snode need to maintain the related data structure and tables. The architecture of delaydedupe system as shown in Figure 1.
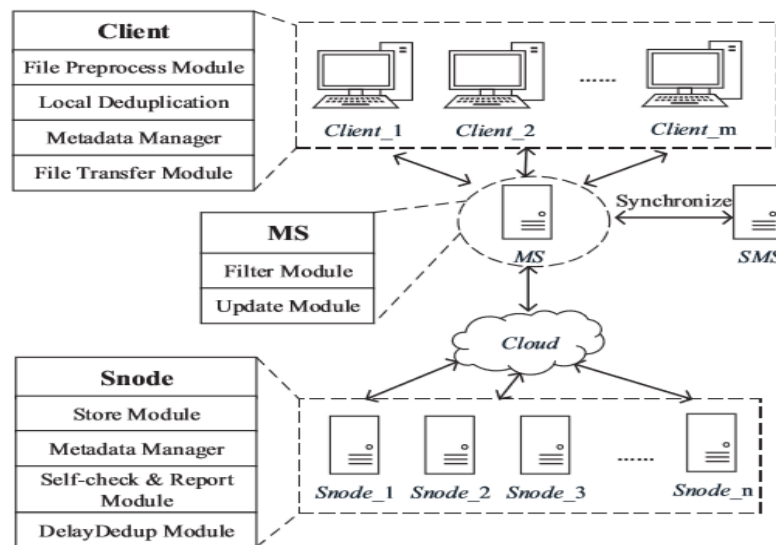

Figure 1. The architecture of DelayDedupe system [10]

To begin the detection of file level and chunk level duplication and elimination locally by client and globally by Metadata Server (MS) to improve the deduplication ratio. Furthermore, by using the DelayDedupe strategy, a delayed target-deduplication scheme followed the access frequency and the chunk-level deduplication of chunks in the Snodes. This method considers the values of duplicated data itself and determines whether new duplication for data modification is hot by the access frequency when combined with replica management. In this research, Client and Snode represent the location of the initial data to be uploaded and the location to keep the new data after deduplication.

The third comparison is done by using Convergent encryption, bilinear map and Merkle hash tree structure by Wu et al. [10]. By using convergent encryption, it provides data confidentiality in deduplication process with the important section of the encryption is convergent key. It is used to encrypt the duplicate of data which is generated from initial data copy. Moreover, check tag is used to identify redundancy that derived from initial copy that is which means, when two data copies are identical, with similar convergent keys are

same, then their check tags are also the same. When the user uploads the check tag with upload data, cloud server will analyse the tags to identify the whether or not there are redundancy. Besides, using bilinear map as an efficient algorithm for computing map, and Merkle hash tree structure which maintains the set of elements rather unmodified, unless with adequate authorization. Leaves are the hashes values of authentic data block in a binary tree which will be changed when authentic data block has been changed.

The forth comparison is done by using deduplicators, cloud storage and redundancy manager method by Leesakul et al. [11]. After identify the duplication, a deduplication system has been proposed as a mitigation medium, redundancy manager will calculate an optimum number for files have been copied based on number of references and level of Quality of Services (QOS) where necessary. Changes of reference number, QOS level and files demand will dynamically change the number of copies. Either a file being deleted by a user, or the QOS level file has been updated, the changes can be monitored, or redundancy manager will capture and recalculate an optimum number of copies.

By identifying and analyzing the concept of reducing dynamic data redundancy in cloud computing, the best technique to reduce redundant dynamic data will be determined based on the technique used by previous researcher such as using Combination of Replication and erasure code [2], DelayDedupe mechanism [12], Convergent encryption, bilinear map and Merkle hash tree structure [10] and deduplicators, cloud storage and redundancy manager method [11].

4. CONCLUSION

A comparison between various data redundancy reduction techniques has been presented in the paper as a proof of concept for the future work where the actual validation based on real dynamic data in correspondance to response time will be presented in the future work. From different technique of reducing redundancy dynamic data has been explained from previous researcher to enhance the availability and performance of cloud computing. This work as an review that can provide a direction to researchers in improving further research about reducing dynamic data redundancy in cloud computing based on response time.

REFERENCES

[1] D. B. Abdullah, et al., "A Review on the Concept of Fourth Industrial Revolution and the Government's Initiatives to Promote it among Youths in Malaysia," e-Bangi, 2017.
[2] D. SureshPatil, et al., "Improving the Availability and Reducing Redundancy using Deduplication of Cloud Storage System," *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), IEEE*, 2017.
[3] M. K. Islam and R. Hasan, "Verifiable Data Redundancy in the Cloud," *Proceedings of the Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), 2016 IEEE International Conferences on. IEEE,* pp. 29-36, 2016.
[4] Kumar, P. R., Raj, P. H., & Jelciana, P. (2018). *Exploring Data Security Issues and Solutions in Cloud Computing. Procedia Computer Science, 125, 691–697*
[5] W. Haoran, et al., "A data deduplication method in the cloud storage based on FP-tree," *Computer Science and Network Technology (ICCSNT), 2015 4th International Conference on, IEEE*, 2015.
[6] J. Hur, et al., "Secure data deduplication with dynamic ownership management in cloud storage," *IEEE Transactions on knowledge and data engineering*, vol. 28, pp. 3113-3125, 2016.
[7] L. Yu, et al., "CoRE: cooperative end-to-end traffic redundancy elimination for reducing cloud bandwidth cost," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, pp. 446-461, 2017.
[8] M. Kaur, et al., "Optimized cloud storage capacity using data hashes with genetically modified SHA3 algorithm," *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), IEEE,* 2017.
[9] Mao, B., et al. (2015). *Improving storage availability in cloud-of-clouds with hybrid redundant data distribution.* 2015 IEEE International Parallel and Distributed Processing Symposium (IPDPS), IEEE.
[10] Y. Wu, et al., "Dynamic Data Operations with Deduplication in Privacy-Preserving Public Auditing for Secure Cloud Storage," *Proceedings of the Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference on. IEEE*, pp. 562-567, 2017.
[11] W. Leesakul, et al., "Dynamic data deduplication in cloud storage," *Proceedings of the Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on. IEEE*, pp. 320-325, 2014.

[12] X. Xu and Q. Tu, "Data deduplication mechanism for cloud storage systems," *Proceedings of the Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2015 International Conference on. IEEE*,

[13] Chen, C., Bao, W., Zhu, X., Ji, H., Xiao, W., & Wu, J. (2015). AGILE: A terminal energy efficient scheduling method in mobile cloud computing. *Transactions on Emerging Telecommunications Technologies*, 26 (12), pp. 1323-1336.

[14] Ayad F. Barsoum and M. Anwar Hasan. (2015) "Provable Multicopy Dynamic Data Possession in Cloud Computing Systems", *IEEE Transactions On Information Forensics And Security*, Pages: 485-497.

[15] Kumar, N., et al. (2017). *Genetic optimized data deduplication for distributed big data storage systems.* Signal Processing, Computing and Control (ISPCC), 2017 4th International Conference on, IEEE.

[16] M. S. A. E. Selvi and R. Anbuselvi, "An Analysis of Data Replication Issues and Strategies on Cloud Storage System," *Proceedings of the International Journal of Engineering Research & Technology (IJERT), NCICN-2015 Conference Proceedings*, pp. 18-21, 2015.