ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Indian Journal of Science and Technology, *Vol 12(20), DOI: 10.17485/ijst/2019/v12i20/145116, May 2019*

Data Fragmentation and Duplication in Cloud for Secure Performance

L. M. Nithya* and P. Anu Priya

Department of Information Technology, SNS College of Technology, Saravanampatti - 641035,
Coimbatore, Tamil Nadu, India

Abstract

The Fragmentation and duplication in data sets is used to overcome the increased data overloading issue in the cloud servers. The increase in data usage and processing in cloud has brought new challenges to data management in cloud computing. We propose the idea which helps in reducing the data load in the cloud and reduces the storage and management cost for the users. Duplication finding plays a very major role in data management. Data de-duplication method finds the restricted fingerprint for every data chunk by using hash algorithms such as MD5 and SHA. The recognized fingerprint is then compared touching other available chunks in a database that is dedicated for storing the chunks. Though, there is simply one copy for every file stored in cloud, it will not be immobile if such a file is owned by a massive number.

As a part the de-duplication system improves storage consumption whereas dropping reliability. Aiming to contradict with the above safety challenges, this proposed idea makes the first attempt to provide the idea of distributed dependable de-duplication system. This new distributed de-duplication system comes with increased reliability in which the data chunks are distributed diagonally to various cloud servers. This allows the redundancy of all data is eliminated. The security needs of data privacy and tag consistency are also achieve by introducing a deterministic furtive sharing system in distributed storage systems, as an option of using convergent encryption as in foregoing de-duplication system.

Keywords: Cloud, Cloud Storage, Data Mapping, File Data Security, Fragmentation, Graph Colouring Algorithm, Graphical Representation, Node Allocation, Performance

1. Introduction

One of the developing paradigms of distributed computing is cloud computing which become root for business, technical and social perspective. Cloud applications are more popular due to the availability, scalability and utility model a high demand on interactive application which attracts the user in great demand due to the availability. Data intensive and analysis model of the cloud. Cloud basically a physical environment which provides a virtualization environment in which the user can the usage by the internet services (Figure 1).

The most highly developed application such as Matlab, Mathematica which does not run by a single desktop system due to their rate and speed of memory performance; hence they use this cloud environment for the data representation. Thus the cloud environment is major platform for the numerous applications in our day to day life. The centralized data which should be confidential and reliable to the entire user are provided by various algorithm methods to provide data security and high performance.

2. Literature Survey

2.1 Secure and Constant Cost Public Cloud Storage Auditing with Deduplication

Proof of Ownership is the major idea that proposed in this paper, which is obtained by two concepts namely proof of reliability and proof of possession. The combination of data integrity and storage is analysed will achieve non-trivial duplicate of metadata which is

**Author for correspondence*

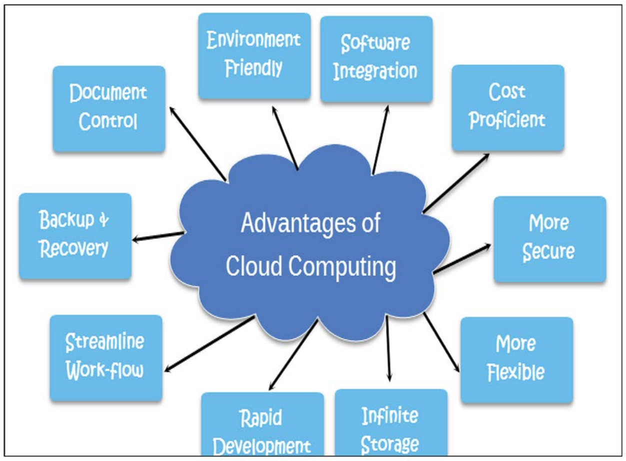

Figure 1. Cloud vision.

called as authentication tags that are provided to each set analysis that could determine in several ways. Lots of communication cost and computational errors are due the lot of replication data along with the replicated sets which are forbidden and find to review the original data. This problem is over by the method proposed in the paper where novel based on techniques including polynomial-based authentication tags to segment and homomorphic linear authenticators are the important sets. Duplication of the data and the authentication tags are the most familiar steps that are used to overcome by the ownership mechanism of design.

2.2 A Survey on Deduplication in Cloud Computing

The paper has detailed study of all the proposed algorithms and tabulated view of all variability outcomes in every stage. Virtualization is the method used to process the data in every aspect. Through the virtual machine runs on the host environment machine is the way that will help the user of internet and also the user can access the application for the usage for every critical analysis. However the current method will use the static method which restrict to certain extreme. Hence in this paper it deliver a dynamic de duplication scheme for cloud storage, which aiming to that improve storage efficiency with shrunk of segment analysis and maintaining redundancy for fault tolerance method ways. This method is carried out in step by step process in which the first level of access is by file level. Hence this paper clearly picturize that the duplication method should focus on block level to improve the space and security.

2.3 Dynamic Data De-Duplication in Cloud

From the customers point of view the main advantage of cloud storage is to reduce their expenditure and more redundant everyway by purchasing and then maintaining storage infrastructure of the cloud platform that will pay for requested amount for the scaled-up and down upon demand in all the basis of requirements. As compared to the previous days, a larger requirements in the data size of cloud computing. A reduction in data volumes could help the providers in reducing the large storage system as well as cost redundant saving system. The saving of energy consumption is also an important factor of this proposed system so that the data reduplications techniques which were introduced to improve the storage efficiency in cloud storages. And also the dynamic nature of data in cloud storage system, data usage in cloud changes from decades of over the time in which some data the chunks may be read frequently in period of time for all the requirements, but may not be used in another time period due to more expensiveness of the storage maintenance. There are lots of data segments in which the datasets may be frequently acquired and are accessed and also updated by the multiple users from which the same time analysis are measured in storage methods, while others may need the high level of redundancy in the data set that are more reliability requirement in all sets of needs. Hence this was crucial to support the dynamic feature that implemented in the cloud storage. Therefore the methodologies have certain drawbacks namely the redundant factor will be very low for the online access and modification of the large volume of data.

2.4 Data Protection and Deduplication in Cloud

CP-ABE uses file access tree structure (folder inside the folder) to encrypt data. In this paper, proposed the idea of Equality Checking Algorithm to check the files/data whether it's duplicate or not in the stored data methods. Any duplication files present in the storage system will intimate the data owner about the duplication. Here, the Symmetric Algorithm is used to encrypt the files/data for security purpose and it is implemented for AmazonS3 Cloud.

An Amazon cloud did not detect duplicate files, it will check file names and if you upload same name and format are different then the content in the Amazon cloud will replace in existing file. So the upload content has different name and same content then the file will be uploaded. It will not check the content of the file. In other clouds like Dropbox, cloud, etc. will never check the duplicate files, it change the name of the file (1), and file (2), etc. Hence the proposed system checks content also.

2.5 Energy Efficient and Replication in Cloud Computing

The paper put forward the view of data replication in cloud computing data centres. As of the other approaches available in the literature, it considers the both energy efficiency and bandwidth consumption of the system and proposed the following views. This is in addition to the improved quality of service QoS obtained as a result of the reduced communication delays in the systems. The evaluation results, obtained from both mathematical model l and extensive simulations that helps to unveil performance and energy efficiency trade-offs methods to guide the design of future data replication solutions in all data centres.

2.6 De-Duplication of Data in Cloud

In this paper author insist on the architecture that the intrusion detection and preventive methods are performed automatically in the way of defining rules for the major attacks hence it alert the system automatically in all aspects. The major attacks and events that include vulnerabilities cross site scripting (XSS), SQL injection, cookie poisoning, and wrapping of all systems. Data deduplication technique allows the cloud users which manages the cloud storage space effectively by avoiding storage of respective data's and save bandwidth. The data are that are stored in cloud server namely CloudMe in all aspects. To ensure data confidentiality the data are stored in an encrypted type using Advanced Encryption Standard (AES) algorithm which probably reduces the duplication and increases the efficiency.

2.7 Drawbacks

The first problem is integrity auditing. The cloud server is able to relieve clients from the heavy burden of storage management and maintenance.

The second problem is secure deduplication. The rapid adoption of cloud services is accompanied by increasing volumes of data stored at remote cloud servers (Table 1).

Table 1. Various algorithm efficiency analyses

Algorithm	Cost	Security	Storage	Time	Key Usage
AES	M	L	H	H	H
SHA-I	H	L	M	L	H
MD5	L	M	M	M	L

H-High M-Medium L-Low

3. Proposed System

In the proposed system it strongly envisage on collectively approaches of the issues of security and performance as a secure data replication is the problem constrain. The system presents judicially fragments user files into pieces and replicates them at strategic locations within the cloud. The division of a file into fragments is performed based on the given user criteria such that the individual fragments do not contain any meaningful information.

In order to find the duplication on the file, the very first step is that the data authorized owner is allowed to upload the file. Then the second step is, the admin accepts the file and upload it to the database. Before it uploads it compared with the existing file and the file uploaded. Hence the algorithm framed will check for the duplication. The detailed step is as follows in methodologies (Figure 2).

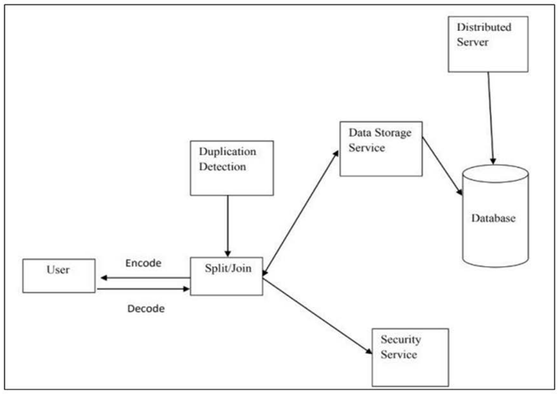

Figure 2. System architecture.

Each of the cloud nodes (we use the term node to represent computing, storage, physical, and virtual machines) contains a distinct fragment to increase the data security.

4. Methodologies

In methodologies, files are divided into various segments which are familiarly called as fragments. So these nodes will have the data information, which will ensure that, at every successful attack no information will be revealed that ensures the security. Each steps that detailed as below:

4.1 Fragmentation

In order to improve the process of the system duplication and replication, the data files which are divided into several segmentation? The segmentation can also be described in a way that the data fragmented into n number of blocks. Each block is coded with encryption algorithms. That blocks are represented as nodes. Each node is given by t-coloring algorithm, framed according to each sets.

4.2 Steps to Fragment

4.2.1 Requester 1

S1. The registration is to be done for providing the details.
S2. The credential details will be given to the registered cloud user.
S3. Next step is to upload the data.
S4. N number of files that can upload and based on the requirement, user can choose the file.
S5. Choose the file which has to be fragmented and then spilt the files.
S6. When the data files are fragmented, just verify the split details and view each fragments.
S7. Finally the data files that are split will be transferred to various nodes over the servers.
S8. If a data is needed, it can be retrieved from the server nodes.

4.2.2 Requester 2

This mainly concern about admin access and steps for an issue handling of the users.
S1. Admin have to login with the credentials.
S2. The option such, attack request, access of several details will provided if any request made.
S3. Server down details, path request is accepted.
S4. The authentication requested is accepted after the detail analysis made.

4.2.3 Module Details of Servers

S1. Request and response are given by the servers to accept several messages.
S2. When the server receives the file of the fragments, it will be listed to every sort of nodes.
S3. The user can view the details of the fragments once it gets uploaded to various server nodes.
S4. Hence, it is the responsibility of the server to duplicate and replication of all the server details.

Data Owner: The one who uses to login and upload the files and performance various functions.

Data Servers: Since the data that are updated functioned as database to the various functionality to accept and process the file request.

Attacker: The intruders who search for file and performance various attack in server nodes. Fragmenting nodes at each initialization algorithm as follows (Figure 3):

STEP 1:

INPUT: User file upload after the registration in cloud environment.

STEP 2: Consider the N is the number of nodes initialized denoted by n1, n2, n3…

STEP 3: The size of the node specified as S foe respective nodes n as s1, s2, s3...

STEP 4: Colouring are specified for the nodes for uniqueness of node to stop the intruder attack.

STEP5: Maximum optimality fragmented loops are provided.

STEP 6: Duplication are identified with file name and content are initialized.

OUTPUT: User will request for the data modification such as adding, deleting and copying for node data will help in replicated data set matching and the middleware will work on the replacement fragmented algorithm in which the duplication data are avoided. The original data stored will be easily accessed with high level secure end.

4.3 First node Selection

Always the first node selection plays the major roles in follow up of each data to continue the data set to complete. Hence the starting of the data set that numbered in way that, the continued set with all functionalities. Sometimes the data sets that are kept in static states this static state always numbered in a way from which the duplication that occurred in lot ways. This follows up with

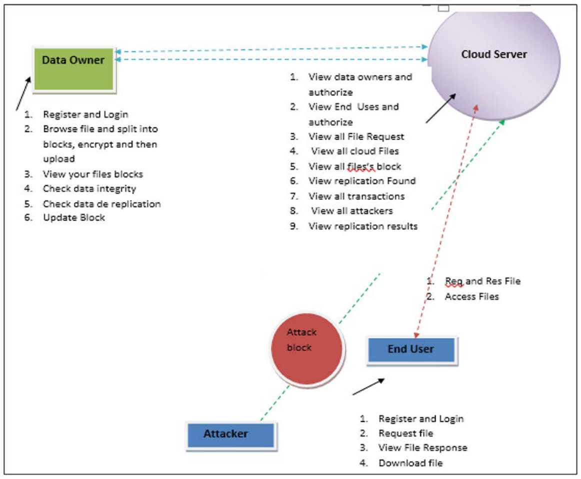

Figure 3. Step to access the nodes.

4.4 Deduplication

Data fragmentation into several compression techniques from the data to eliminate the duplicate set data from which the fragmented data into several duplicate data to which file level data duplication set. The block size parameter will be defined to store the data sets in order to analysis particular level pre-defined values.

4.5 Data Sharing

Data nodes on the cloud from which coloring concepts with the certain distance each node concept which is similar to distance node for the securing node data allocation methods the data input with high level security level to data allocation with randomly generated output of the selected nodes.

4.6 Mapping

Share divides secret S into (k-r) fragments of same size, which produces r for random fragments of the equal size. The translates into simple language the k fragments using a non-systematic k-of–n ration code into n shares of the similar size The nodes that are separated on each region will be addressed by unique identification to map and find the duplicated data sets.

5. Distributed Storage

This one of the method of storage efficient method to enhance the data method allocation here the data redundant chunk will point to the node from which the functionality can be pre-mapped. During the data mapping set the data match with the original data set for the defined method of adopting.

6. File Access

The security of the large scale system that depends on individual node access with all certain accessing point a successful intrusion each will replicate in each node hence securing each access is very important. Therefore the node encryption credential is very important at each level of authentication. However compromising a single file will enhance the data effort to penetrate to each node.

7. Result and Discussion

The data set input such as file upload and link sharing are interfaced with the GUI and back end support for query mapping with cloud data set backbone. The graphical representation will be shown how the prolonged data and its efficiency will be shown by up and downs of data duplication and redundancy. Hence the analysis of all the fragmented data will enhance the storage efficiency and cost effective. Hence it become user friendly and allows the fragmented data which decides upon itself.

8. Conclusion

Cloud computing generally faces the theft of data, hence this decision making data will allows the high level data performance at each level. The business at high level which interact with customer end data such as data mapping, data end at high level that provides the faster access as well the storing of the data. The performance at high level of data retrieval wills all efficient method. Hence the random generated data will enhance the secure level through which each data set is mapping in all ways.

9. Future Work

There are scopes in reducing the time latency with the new searching algorithm and to increase the security, additional encryption methods can also be increased. An improved searching and comparison algorithm can reduce the processing time and accuracy can be increased.

10. References

1. Bilal K, Khan SU, Zhang L, Li H, Hayat K, Madani SA, Min-Allah N, Wang L, Chen D, Iqbal M, Xu CZ, Zomaya AY. Quantitative comparisons of the state of the art data center architectures, Concurrency and Computation: Practice and Experience. 2013; 25(12):1771-83. https://doi.org/10.1002/cpe.2963.
2. Bilal K, Manzano M, Khan SU, Calle E, Li K, Zomaya A. On the characterization of the structural robustness of data center networks, IEEE Transactions on Cloud Computing. 2013; 1(1):64-77. https://doi.org/10.1109/TCC.2013.6.
3. Chen Y, Paxson V, Katz RH. Whats new about cloud computing security. University of California, Berkeley Report No. UCB/EECS-2010-5; Jan. 20, 2010.
4. Deswarte Y, Blain L, Fabre J-C. Intrusion tolerance in distributed computing systems, In: Proceedings of IEEE Computer Society Symposium on Research in Security and Privacy, Oakland CA; 1991. p. 110-21.
5. Grobauer B, Walloschek T. Stocker E. Understanding cloud computing vulnerabilities, IEEE Security and Privacy. 2011; 9(2):50-57. https://doi.org/10.1109/MSP.2010.115.
6. Wayne A. Jansen. Cloud Hooks: Security and Privacy Issues in Cloud Computing. Proceedings of the 44th Hawaii International Conference on System Sciences - 2011. https://doi.org/10.1109/HICSS.2011.103
7. Kappes G, Hatzieleftheriou A, Anastasiadis SV. Virtualization-aware Access Control for Multitenant File systems, IEEE. 2014; 978(1)4799-5671. [3]. Aiqiang Gao, Luhong Diao, Lazy Update Propagation for Data Replication in Cloud Computing, IEEE. 2010; 978(1); 4244-9142.
8. Yang Tang, Patrick PC. Lee, John CS. Lui, Radia Perlman. Secure overlay cloud storage with access control and assured deletion, IEEE Transactions on Dependable and Secure Computing. Nov-Dec 2012; 9(6). https://doi.org/10.1109/TDSC.2012.49.
9. Mazhar Ali, Samee U. Khan. DROPS: Division and replication of data in cloud for optimal performance and security, IEEE. 2015. [2]. Wylie JJ, Bakkaloglu M, Pandurangan V, Bigrigg MW, Oguz S, Tew K, Williams C, Ganger GR, Khosla PK. Selecting the right data distribution scheme for a survivable storage system. Carnegie Mellon University, Technical Report CMU-CS-01-120. May 2001.
10. Khan SU, Ahmad I. Comparison and analysis of ten static heuristics-based Internet data replication techniques, Journal of Parallel and Distributed Computing. 2008; 68(2):113-36. https://doi.org/10.1016/j.jpdc.2007.06.009.
11. Boru D, Kliazovich D, Granelli F, Bouvry P, Zomaya AY. Energy-efficient data replication in cloud computing datacenters, In: IEEE Globecom Workshops; 2013. p. 446-51. https://doi.org/10.1109/GLOCOMW.2013.6825028.
12. Loukopoulos, Ahmad I. Static and adaptive distributed data repli-cation using genetic algorithms, Journal of Parallel and Distributed Computing. 2004; 64(11):1270-85. [6] Bilal K, Khan SU, Zhang L, Li H, Hayat K, Madani SA, Min-Allah N, Wang L, Chen D, Iqbal M, Xu CZ, Zomaya AY. Quantitative comparisons of the state of the art data center architectures, Concurrency and Computation: Practice and Experience. 2013; 25(12):1771-83. https://doi.org/10.1002/cpe.2963.
13. Bertino E, Paci F, Ferrini R, Shang N. Privacy-preserving digital identity management for cloud computing, IEEE Data Eng. Bull. Mar 2009; 32(1):2127. [13] Skopik F, Schall D, Dustdar S. Start trusting strangers bootstrapping and prediction of trust. In: Proc. 10th Int. Conf. Web Inf. Syst. Eng.; 2009. p. 275-89.
14. Guo H, Huai J, Li Y, Deng T. KAF: Kalman filter based adaptive maintenance for dependability of composite services. In: Proc. 20th Int. Conf. Adv. Inf. Syst. Eng.; 2008. p. 328-42. https://doi.org/10.1007/978-3-540-69534-9_26.
15. Wei Y, Blake MB. Service-oriented computing and cloud computing: Challenges and opportunities, IEEE Internet Comput. Nov-Dec 2010; 14(6):72-75. https://doi.org/10.1109/MIC.2010.147.
16. Fung B, Wang K, Chen R, Yu P. Privacy-preserving data publishing: A survey of recent developments, ACM Comput. Surv. 2010; 42(4):1-53. https://doi.org/10.1145/1749603.1749605.