DOI: 10.13140/RG.2.1.4635.6085

A review: Various approaches for reduction of data elements with sharing of file on different file systems

Harpreet Kaur[1], Maninder Kaur[2]

[1]*Student (M.Tech), Doaba Institute of Engineering and Technology, Kharar*

[2]*Head of Department, Doaba Institute of Engineering and Technology, Kharar*

Abstract- Cloud Storage System is becoming increasingly popular with the continuous and exponential Increase of the quantity of operators and the opportunity of data. Information duplication changes more and more a necessity for cloud storage providers. Data de-duplication is one of the significant data looseness techniques for eliminating duplicate copies of repeating data. It has been widely used in the cloud storage to reduce the amount of storage space and save bandwidth. The advantage of de-duplication unfortunately come with high cost in terms of new security and privacy challenges .The proposed scheme in this paper not simply the decreases the cloud storage size but also improves the speed of data de-duplication. Data de-duplication titles a class of methods that reduce the storage capacity needed to store data or the quantity of data that has to be transmitted over a network. These approaches detect coarse-grained redundancies within a data set, e.g. a file system; Data de duplication not only reduces the storage space requirements by eliminating dismissed data but also reduces the network communication of duplicate data in the network storage systems.

Keywords- Cloud Computing, Data de-duplication recovers the speediness and minimalize network.

I. INTRODUCTION

Cloud computing provides a low-cost, scalable, position independent arrangement for data management and storage. Owing to the populace of cloud service and the growing of data size, more and more people pay devotion to economize the capacity of cloud storage than before .Therefore how to apply the cloud storage size well becomes important issue now a days. Cloud computing enables new commercial models and cost operative resource usage. Instead of keep up their own data centre, companies can concentrate on their fundamental commercial and purchase resources when it will desired. Especially when [1] combining publicly accessible clouds with a secretly maintained virtual arrangement in a hybrid cloud, the hybrid cloud technology can open up new chances for big business. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and the data common by dissimilar users with specified privileges, which define the access rights of the kept data. One critical task of cloud storage services is the management of the ever-increasing volume of data on cloud.

Cloud computing, in addition to additional facilities provides various arrangements as service. Storage-as-a-service is one of the most significant and widely used arrangements provided by cloud computing technology. With the increasing request of computers and other computer based facilities, the demand for data storage is also increasing day by day. In this situation cloud computing proposals best solutions for rapid, elastic, reliable, and measured storage 7. The [2] growing demand of cloud storage condition has led to the process of de-duplication. The term data de-duplication refers to methods that store only a single duplicate of redundant data, and provide links to that copy instead of storing other actual copies of this data 1. The de-duplication procedure is used to defend the cloud server from storing redundant data. If two operators want to upload the identical file, only a single file will be uploaded on the cloud server and the users will be providing with a link that will get the whole file for them every time they want to retrieve it. Suppose user1 on cloud supplies a file A. He will appeal to upload the file and the file will be successfully uploaded now when a operator, user2 will upload the identical file, the cloud will de-duplicate the file by providing user2 the link of file A, which is previously existent on the cloud. Thus ‘n’ number of users can be allowed to entrée identical file with a particular copy kept on cloud. Information duplication is a particular data compression [3] technique for eliminating identical copies of recapping data.

The earlier de-duplication systems cannot support to differential permission and identical check, which is very significant in many applications. In an authorized de-duplication scheme each user is issued a set of honours during scheme initialization. Each file uploaded to the cloud is restricted by a set of privileges which stipulate which kind of users is allowed to execute the duplicate checked and access right of the files. Before submitting his matching check demand for some file, the user needs to take his file and his own privileges as inputs. The operator is able to find a matching for this file if and only if there is a copy of this file and a matched privilege stored in cloud [4].

Deduplication strategy can be categorized into two main strategies as follow, differentiated by the type of basic data units:

Sr no.	Data unit	Description
1	File-level data de-duplication	The information is a facts element when examining the data of duplication, and it typically uses the hashing value of the information as its identifier. If two or more files have the same hash value, they are assumed to have the identical contents and only one of such files will be put in storage.
2	Block-level de-duplication	This approach sections a file into some fixed-sized blocks or variable-sized [5] blocks, and computes hash value for each block for examining the duplication blocks.

II. LITERATURE SURVEY

Vasilios et.al. [6] Presents a migration support network, in which fundamental elements are cost effective system. They proposed a three level framework that satisfies al the necessity in view of cost assumption. They utilized the windows azure policy as a part of creating prototyping model. Besides, the ability to consolidate necessities for numerous administration sorts, e.g., information stockpiling and systems administration, is imagined to be given, encouraging the choice making in relocation sorts past the off-stacking of the application stack on a VM. Haitao et.al. [7] proposed relocation methods taking into account (dynamic, receptive and shrewd procedures), albeit basically in light of the present data, can make the mixture cloud-helped VoD organization set aside to 30% transmission capacity cost contrasted and the Clients/Server mode. They can likewise handle unpredicted the glimmer group activity with little cost. It likewise demonstrates that the cloud cost and server transmission capacity picked assume the most essential parts in sparing expense, while the distributed storage size and cloud substance upgrade system assume the key parts in the client experience change. C. Ward et.al. [8]Acquainted the augmentations with a coordinated mechanization capacity called the Darwin structure that empowers workload movement for this situation and talk about the effect that computerized relocation has on the expense and dangers ordinarily connected with relocation to cloud. Kang et.al.[9]Proposed the migration algorithm .The VM to its best PM specifically, with the proviso that it has adequate capacity. Then, if the migration constraint is gratified, we transfer another VM from this PM to oblige the new VM. In addition, we study a crossbreed scheme where a batch is in employment to accept future VMs for the on-line development. Evaluation results prove the high efficiency of our algorithms. Xian Xin et.al. [10] proposed a dynamic prototype system termed Cyber Live App to support request allocation and migration on response among various operators. CyberLiveApp gives two key administrations: a safe multi-client sharing administration for the virtual desktop of a VM and multi-VM application sharing and movement.R Maggiani et.al. [11] proposed the Saas infrastructure for the improvement of administrations. Distributed computing can be a solitary capacity application, a framework on which these applications (and numerous others) can run an arrangement of administrations that offer the benefits of enormous measures of processing assets, and the capacity to store a lot of information remotely. Numerous organizations and instructive infrastructures are simply starting to understand the advantages of cloud-based applications that have generally obliged site permitting, establishment, and support.

III. SCOPE OF STUDY DE-DUPLICATION

The other key to understanding the full benefits of de-duplication is to spread its scope by combining record and/or backup data stores or other requests so that these collective data stores are globally [12]de-duplicated. Even though no necessities exist for companies to store backup and collection data on dispersed systems with different constructions, this common approach added segregates de-duplicated data by application or scope and, similar to the scalability restriction, reduces the effectiveness and value of data de-duplication. De-duplication incompetence due to incomplete request scope with utilization based approaches is particularly true for archive and backup information, since information that is archived has practically always already been earlier backed up multiple times. Most corporations possess a great deal of redundant data between their collection and backup data stores that cannot be worldwide de-duplicated across the disparate instances and architectures of de-duplicating appliances.

IV. IMPROVED TECHNIQUE DE-DUPLICATION

Data De-duplication is an actual method for optimization of requirements of data kept in cloud storage [13]. Deduplication can be classified addicted to chunk and file level data de-duplication. Chunk level data de-duplication method enforces the storage of unique chunks by comparing every received chunk for matching identification. This method achieves better de-duplication efficiency because it does exact de-duplication [14]. However, the throughput is low as it checks every incoming chunk for repetition. File level de-duplication technique applies the chunks of similar files to be compared against duplicates. Files with incremental modifications are referred to as a like files. This method achieves better output as it associates every received chunk only with chunks of similar files. Though, the data de-duplication efficiency is comparatively low as some duplicate chunks may be originate across dissimilar groups. Therefore, this technique performs only approximate de-

DOI: 10.13140/RG.2.1.4635.6085

duplication. In order to better make use of the storage space, copies among the files requirement to be identified. Each inward file is separated into chunks [3]. Depends on how the received chunk is checked upon contradiction of duplicates, de-duplication can be categorized into two types, namely, chunk and file level.

a) Chunk level Deduplication:

Whenever a information stream has to be printed, every chunk in the stream is verified for copies before writing. This is termed as chunk level de-duplication. Since every incoming chunk is tested for possible repetition, only unique chunks occupy the cloud memory. Therefore, chunk level de-duplication has improved data de-duplication efficacy. Though, as each received chunk is tested against a large list of chunk catalogues, the amount of disk I/O processes is large. This has a significant impact on de-duplication amount. Storage of out-dated backup consignment demands worthy de-duplication productivity as it includes large data redundancy amongst different assignments. Hence, this de-duplication approach is best suited for such workloads [3].

b) File Level Deduplication:

Whenever a data stream has to be printed, each chunk in the stream is tested in contradiction of the chunks of similar files. This is labelled as file level data de-duplication. This technique provides a climbable solution with the division of chunk index into two tiers namely Main and Minor index [3]. In this method, all the Chunk IDs that establish a file and the smallest Chunk ID between them are found.

V. ATTACK MODELS

Few attack models have been exposed, which can principal to the manipulation of de-duplication towards an insecure storage technique. However looking at the outstanding possibilities of enhancing storage efficiency, several solutions have been future. This section defines various attack representations. The first attack can be used to predict an already known file controlled by the user. The additional attack is related to generating a secret channel for removing information however the third attack is related to distribution of any file among various users of cloud storage.

A. Attack Model I:

Forecasting files this attack can be used to forecast whether a particular file is controlled by a specific user 1. Furthermore this attack can be additional professionally used to forecast a file if the file contains data with limited possibilities for sample yes or no in situation of a medicinal test report. Supposing the attacker wants to find out whether user1 retains information, File A. He will submitted a copy of file A if the information gets uploaded this will indicate that the file is not possessed through the user1. While in the other situation the attacker will be able to discovery out if the file is possessed by user1. Also, in instruction to hide his uniqueness as an attacker he will dismiss the connection as soon as the file uploading starts.

B. Attack Exemplary II:

Generating a secret channel If the attacker achieves to install any malicious software on the machine of user1, this program can be used to start a secret location among the user and the other user as attacker 1.There are some ways of producing this type of station one of them is to avoid the firewall then communicate with its control server. Reflect this example; supposing user1 is using the scheme with harmful software installed, the software will generate two files in two different conditions. When user1 will back up his files on control server this file [14] will be stored on the server. Currently, attacker can simply use the attack labelled in previous unit to find which file was stored by the program. This attack can use any amount of restricted information, and thus can be very harmful as well as untraceable.

C. Attack Exemplary III:

The information Sharing Attack the content [13] delivery attack can be used to distribute a exact file to various users devoid of providing the uniqueness of the supplier. The type of file can be a plagiarized video or a file containing a virus etc. The users in de-duplication are enabled to use a file if they are included in the access control list of the file. The access to a file is gained concluded a hashing, h(F) where F is the file conferred above and the h(F) is the hashing value consistent to the file F, designed by the control server. The attacker can get h(F) by submitting the essential file and distribute it to the few users in order to let them access the file.

VI. REFERENCE

[1] He, Qinlu, Zhanhuai Li, and Xiao Zhang. "Data deduplication techniques."Future Information Technology and Management Engineering (FITME), 2010 International Conference on. Vol. 1. IEEE, 2010.

[2] Lee, Ju-pyung. "DATA DEDUPLICATION METHOD." U.S. Patent No. 20,150,339,316. 26 Nov. 2015.

[3] Deepu, S. R. "Performance Comparison of Deduplication techniques for storage in Cloud computing Environment." Asian Journal of Computer Science & Information Technology 4.5 (2014).

[4] Raut, Bhavanashri Shivaji, and H. A. Hingoliwala. "A Review of Secure Authorized Deduplication with Encrypted Data for Hybrid Cloud Storage."

[5] Shinde, Priyanka K., and Avinash P. Wadhe. "Review paper on Authorized Duplication Checker in Hybrid C cloud."

[6] Vasilios Andrikopoulos, Zhe Song, Frank Leymann , “Supporting the Migration of Applications to the Cloud through a Decision Support System”, Institute of Architecture of Application Systems, IEEE, pp. 565-672, 2013.

DOI: 10.13140/RG.2.1.4635.6085

[7] Haitao Li, Lili Zhong, Jiangchuan Li, , Bo Li, Ke Xu, " Cost-effective Partial Migration of VoD Services toContent Clouds", 2011 IEEE 4th International Conference on Cloud Computing, pp. 203-110, 2011.

[8] C. Ward, N. Aravamudan, K. Bhattacharya, K. Cheng, R. Filepp, R. Kearney, B. Peterson, L. Shwartz, C. C. Young, "Workload Migration into Clouds – Challenges, Experiences, Opportunities", 2010 IEEE 3rd International Conference on Cloud Computing, pp. 164-171, 2010.

[9] Kangkang Li, Huanyang Zheng, and Jie Wu . "Migration-based Virtual Machine Placement in Cloud Systems", 2013 IEEE 2nd International Conference on Cloud Networking (CloudNet, IEEE, pp. 83-90, 2013.

[10] Jianxin Li, Yu Jia a, Lu Liub, Tianyu Woa, " CyberLiveApp: A secure sharing and migration approach for live virtual desktop applications in a cloud environment, Elsevier, Vol. 29, pp.334-340, 2013.

[11] R. Maggiani, "Cloud computing is changing how we communicate," 2009 IEEE International Professional Communication Conference, IPCC 2009,Waikiki, HI, United states ,pp 1, July 2009.

[12] Singh, Deepika, and Preetika Singh. "New Challenges for Security against Deduplication in Cloud Computing." International Journal 2.1 (2014).

[13] D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels in cloud services: Deduplication in cloud storage," Security Privacy, IEEE, vol. 8, no. 6, pp. 40 – 47, nov.-dec. 2010

[14] National Institute of Science and Technology. "The NIST Definition of cloud computing, Luis M. Vaquero1, Luis Rodero-Merino1, Juan Caceres1, Maik Cloud Computing".p.7. Retrieved July 24 2011