Various Data De-Duplication Methods to Utilize Memory Structure in Cloud Storage System

Manpreet Kaur[1], Daljit Kaur[2], Jaskiran Kaur[3]
[1]*M.tech Scholar, Chandigarh Engineering College Landran*
[2]*Assistant Professor, Chandigarh Engineering College Landran*

Abstract - Cloud system is becoming very popular with the constant and exponential increase in the number of users and the size of data. Data Deduplication becomes increase for cloud storage sources. Data de-duplication is one if the significant data compression approaches for reducing duplicate copies of repeating data. It has been widely used in the cloud storage to eliminate the amount of storage space and save memory. The benefits of de-duplication unfortunately come with high cost in terms of novel security and privacy challenges. In this paper not only the overcomes the cloud storage capacity but also advances the speed of data de-duplication. To safe confidentiality of data while supporting de-duplication the encryption technique has been planned to encrypt the data before outsourcing. This paper makes the first attempt to statement the problems of approved data.

Keywords - Cloud Storing system, Data de-duplication, compression approaches and advantages of de-duplication.

I. INTRODUCTION

Cloud computing, in adding to other services offers various organizations as service. Storage-as-a-service is one of the most significant [1] and widely used organizations provided by cloud computing knowledge. With the increasing request of computers and other computer based services, the demand for data storage is also collective day by day. In this scenario cloud computing proposals best solutions for rapid, elastic, reliable, and measured storage.

The accumulative demand of cloud storage condition has led to the procedure of de-duplication. The period data de-duplication mentions to devices that store only a single copy of terminated data, and deliver links to that copy in its place of storing other definite copies of this data. [1]The de-duplication process is used to defend the cloud server from storing terminated data. If two users poverty to upload the similar file, only a single file will be uploaded on the cloud server and the users will be providing with a link that will fetch the whole file for them when they want to retrieve it.

Example: Suppose user1 on cloud stores a file A. He will demand to upload the file and the file will be effectively uploaded now ,when a user2 will upload the same file, the cloud will de-duplicate the folder if user2 link the file A, which is already present on the cloud. Thus 'n' number of users can be permitted to admission same file with only copy stored on cloud.

The de-duplication can be achieved on the cloud server. If the [2] whole file is first transported to the cloud server before any de-duplication, this is server-based method for de-duplication. This development saves the storage creation in above mentioned way but the network bandwidth for sending the terminated data is wasted. Thus client side de-duplication is used to save system bandwidth as well as storage space. Though the de-duplication development has the capability to save both storage space and network bandwidth but this procedure give rise to a security problematic in cloud computing, the side channel attack. The cross virtual machine users can use several attacks to find the private data related to each other as well as the administrator. Thus cross user de-duplication leads to the susceptibility of side channel attacks in cloud computing. De-duplication approach can be characterized into two main strategies as follow, differentiated by the type of basic data units.

1).File-level de-duplication: A folder is a data unit when grouping the data of duplication, and it classically uses the hash value of the folder as its identifier. If two or more files have the same confusion worth, they are expected to have the same insides and only one of these files will be stored.

2).Block-level de-duplication: This strategy sections a file into several fixed-sized tablets or variable-sized blocks, and calculates hash value for each block for examining the duplication blocks.

II. VARIOUS CHALLENGES DE-DUPLICATION

It is not prepared to oversee by the cloud executive arrange [3] that prompts the system programme and multifaceted nature of apparatus. To beat this there are plentiful document lumping and information compression strategies are utilized. One thought is that the remarkable issues connected with dispersed computing security have not been apparent. Another thought is that the precise security necessities for distributed computing have not been all around considered inside of the group. One worry is that the clients would prefer not to uncover their information to the cloud management supplier. Unease is that the clients are undefined about the uprightness of the information they get from the cloud. Here, more than customary security mechanisms will be needed for information security. One of the fundamental difficulties that keep end clients from embracing cloud storage administrations is the concern of losing information or information destruction.

III. INFORMATION DE-DUPLICATION RESTRICTIONS

a) Interpretation support Server-side information

De-duplication is just relying on specific renditions of capacity rulers or most recent servers. Henceforward for ideal aptitude when utilizing server side information [4] de-duplication, move up to the upheld check adaptation. Moreover, Client-side information De-duplication is likewise should be reshaped which is critical.

b) Competent capacity pools

Information on unbalanced access storage can't be de-duplicated. Just material away pools that are associated with consecutive get to i.e. record can be de-duplicated. You must authorize document storing pools for information de-duplication. Customer records must be sure to an administration class that regulates a de-duplication empowered capacity pool.

c) Encrypted documents

A safety measure, you can take one or a greater quantity of the accompanying steps:

1. Qualify capacity gadget encryption organized with customer side information de-duplication.
2. Use client side information de-duplication just for hubs that are secure.
3. If you are unverifiable about system security, empower [5] Secure Sockets Layer.

d) Document size

Just documents that are more than 2 KB are de-duplicated. Records that are 2 KB or less are not de-duplicated.

IV. RELATED WORK

Vasilios et.al. [6]Presents a migration support network, in which fundamental elements are cost effective system. They planned a three level outline that contents all 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.

Kang et.al. [8] 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 hybrid scheme where a batch is employed to accept upcoming VMs for the on-line development. Evaluation results prove the high efficiency of our algorithms.

Xian Xin et.al. proposed a dynamic prototype system termed Cyber Live App to support application sharing and migration on demand among various [9] users. 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. proposed the Saas infrastructure for the improvement of administrations. Distributed computing can be a solitary capacity application, a framework [10] 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.

V. DE-DUPLICATION PROBLEMS OF CLOUDS

Storage competence functions such as de-duplication [11] afford packing providers better operation of their storage back ends and the capability to serve more customers with the same infrastructure. It is the procedure by which a storage worker only stores a single copy of a file owned by numerous of its users and there are four different de-duplication strategies, depending on whether de-duplication happens at the client side or at the server side, and whether de-duplication ensues at a file level or at a block level. Deduplication is most pleasing when it is triggered at the client side, as it also saves upload bandwidth but for these reasons, de-duplication is a serious enabler for a number of popular and positive storage services which offers a cheap, remote storing to the wide-ranging public by performing client-side de-duplication, thus it will saving both the network bandwidth and stowing costs. Indeed, data de-duplication is perhaps one of the main reasons why the prices for cloud storage and cloud backup facilities have dropped so suddenly. As the world transfers to digital storage for archival resolutions, there is a growing demand for systems that can deliver a secure data storage in a cost effective manner. By recognizing the common amounts of data both in and between files and storing them only once, by this de-duplication can yield cost savings by increasing the utility of a given measure of storage but Unfortunately, de-duplication exploits identical content, while encryption challenges to make all content appear random, when the same satisfied encrypted with two different keys consequences in very different ciphertext. Thus, in encryption joining the space efficiency of de-duplication with the secrect aspects is difficult. Though data de-duplication brings a lot of benefits to cloud user, security and privacy distresses arise as user's sensitive data are susceptible to both insider and stranger attacks. Specifically, traditional encryption necessitates different users to encode their data with their own keys. Thus, identical data copies of different users will lead to a unlike ciphertext, which makes de-duplication incredible. Thus Convergent encryption has been proposed to enforce data discretion while making de-duplication feasible.

VI. GAP IN STUDY

IT budgets are also growing. One estimate put increases in IT spending at 3% for 2011 and improvement over the past couple of years, but modest nevertheless. However, the impression this 'new money' has on lessening the data growth problem is a little more difficult. Exertions to address data growth will certainly [12] receive a portion of IT budgets, but not all of it. And some trainings show that perchance more spending should go to optimizing standing storage and not just buying raw capacity. So where does this leave IT executives who are tackled with finding half again [13] as much storage capacity each year just to keep up? While funds are growing and storage costs are dropping, the real question is "Will IT is able to cover the meal between projected budgets and projected costs of needed storage?" For many the answer is "no". Trying to cover an assessed 50% growth in data with a 25% decrease in storage costs and a small increase in IT spending leaves a significant gap.

This break between possible storage capacity needs and the predictable ability of businesses to afford that capacity is very real. Basically, the typical company will be faced with an overtaxed substructure and an overreached staff as they challenge to find ways to make ends meet. Just keeping up will mean shifting budget dollars to storage and away from funds in expansion, innovation and even people. With these restrictions the data affordability gap will be a drain on short term productivity and longer term competitiveness.

VII. CONCLUSION

Cloud computing has stretched a maturity that leads it into a dynamic phase. This means that most of the main problems with cloud computing have been lectured to a degree that clouds have become interesting for full commercial use. This however does not mean that all the problems enumerated above have actually been solved, only that the according dangers can be tolerated to a certain degree. Cloud computing is therefore still as much a exploration topic, as it is a market offering. For better privacy and refuge in cloud computing. Several new de-duplication developments supporting endorsed copy sign up a cross breed cloud plan. Security examination shows that this topic is secure as far as the definitions laid out in the organized security model.

VIII. REFERENCES

[1] Hashizume, Keiko, et al. "An analysis of security issues for cloud computing." Journal of Internet Services and Applications 4.1 (2013): 1-13.

[2] Forman, George Henry, Fereydoon Safai, and Bin Zhang. "Data de-duplication." U.S. Patent No. 7,200,604. 3 Apr. 2007.

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

[4] Meyer, Dutch T., and William J. Bolosky. "A study of practical deduplication." ACM Transactions on Storage (TOS) 7.4 (2012): 14.

[5] Ju, Jiehui, et al. "A survey on cloud storage." Journal of Computers 6.8 (2011): 1764-1771.

[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.

[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] 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.

[9] 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.

[10] 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.

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

[12] Sengar, Seetendra Singh, and Manoj Mishra. "A Parallel Architecture for In-Line Data De-duplication." Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on. IEEE, 2012.

[13] Wu, Tin-Yu, Jeng-Shyang Pan, and Chia-Fan Lin. "Improving accessing efficiency of cloud storage using de-duplication and feedback schemes."Systems Journal, IEEE 8.1 (2014): 208-218.