International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 46-50
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

ISSUES & BENEFITS OF DATA DE-DUPLICATION IN CLOUD COMPUTING A REVIEW

Rohini Sharma
Computer Science & Engineering
SVIET, Banur, Punjab, India

Amritpal Kaur
Assistant Professor CSE
SVIET, Banur, Punjab, India

Abstract— **Cloud calculating environment is a scheming instance, wherever a huge pool of schemes are connected in private or public systems grid, to deliver dynamically climbable infrastructure for request, data and file storage. Data de-duplication is the method which wrappings the information by eliminating the duplicate copies of identical information and it is lengthily used in cloud storing to save bandwidth and minimalize the storage space. For defense of data security, this paper makes a challenge to primarily attend to the predicament of approved data de-duplication. We propose advance which works with Hashing algorithm, genetic algorithm and also reducing memory consumption. These hashing algorithms have their own assets like their output size, block size, rounds and performance. The genetic algorithm is the study of genes, what they are , what they do and how they work. The advantage of de-duplication unfortunately comes with high cost in terms of new security and privacy challenges.**

Keywords— ***Cloud Computing, De-duplication, Cloud Storage and application hosting***

I. INTRODUCTION

Cloud computing is the delivery of computing facilities above the Internet. Cloud facilities permit individuals and businesses to use software and hardware that are achieved by third parties at out-of-the-way locations. Samples of cloud services include online file storage, web-mail, social networking web-sites & online commercial requests. Cloud computing model allows access to information and computer properties from wherever that a system connection is obtainable. Cloud computing provides a common pool of resources, grids, computer processing power, including information storage space, and specialized corporate and user applications. [1] The cloud marks it likely for you to access your information since anywhere at any time. While a traditional computer setup needs you to be now the same place as your data storing device, the cloud takes away that step. The cloud removes the requirement for you to be in the similar physical place as the hardware that stores your data.[2]

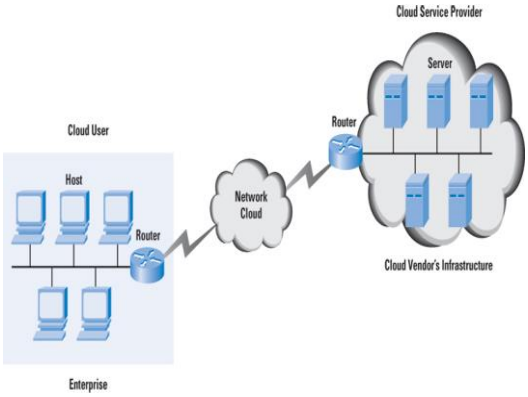

Fig. 1. Cloud Computing[11]

II. BENEFITS OF CLOUD COMPUTING

The following are some of the possible benefits for those who offer cloud computing-based facilities and requests: [3]

- Cost Investments
- Scalability/Flexibility
- Reliability
- Maintenance
- Mobile Available

III. DATA DE-DUPLICATION

Data de-duplication is progressive expertise that can melon dramatically decrease the quantity of backup information stored by eliminating redundant data. Data de-duplication exploits storage consumption while permitting IT to recall more near line backup data for a longer time. This tremendously recovers the competence of disestablished backup, altering the way data is protected. In general, data de-duplication compares novel information with current information from preceding backup or archiving jobs, and eliminates the redundancies. Advantages include better storage competence and budget savings, as glowing as bandwidth minimization for fewer expensive and faster offsite repetition of reserve information.[4][5] The de-duplication expertise is capable of recognizing and then eliminating the redundant data, which makes the storage space of backup dramatically

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 46-50
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

decrease, and then further enables the enterprise to possess a much longer storage of backup data to ensure the instant recovery (the better recovery of RTO), backup more frequently and create much more RPOs.[5]

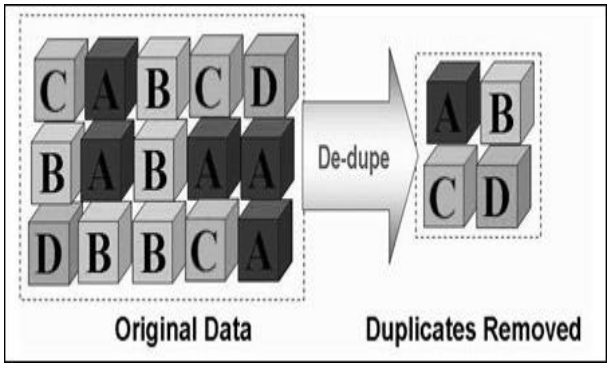

Fig. 2. Data De-duplication[12]

De-duplication Algorithm using SHA
Various types of SHA:-

- SHA 0
- SHA 1
- SHA 256
- SHA 512

The Secure Hash Algorithm is one of a numeral of cryptographic hash functions. There are currently three generations of Secure Hash Algorithm:

- SHA-1 is the original 160-bit hash function. The similar to the earlier MD5 algorithm.
- SHA-2 is a relation of two similar hash functions, with different block sizes, known as SHA-256 and SHA-512. They differ in the word size; SHA-256 uses 32-bit words where SHA-512 uses 64-bit words.
- SHA-3 is a future hash function standard still in development.

The SHA algorithm uses 5 state variables, each of which is a 32 bit integer (an unsigned long on most systems). These variables are sliced and dice and are (eventually) the message digest. The variables are initialized as follows:
h0 = 0x67452301
h1 = 0xEFCDAB89
h2 = 0x98BADCFE
h3 = 0x10325476
h4 = 0xC3D2E1F0
There are 80 rounds in SHA Algorithm.
The hash value generated by the sha hash function.

Genetic Algorithm

Genetic algorithm is computer programs that simulator the processes of natural evolution in order to solve difficulties and to model evolutionary systems.
Different types of three operators [18]:

- The selection operator selects those chromosomes in the populace that will be allowed to replicate, with better chromosomes producing on average more spring than less ones.
- Crossover exchanges subparts of two chromosomes, roughly replicating biological re-combination between two single gene organisms.
- Mutation casually changes the allele values of some positions in the chromosome; and transposal reverses the order of a connecting section of the chromosome, thus re-arranging the order in which genes are organised [13].

The Genetic Procedure is a model of machine knowledge which derives its performance from image of the processes of Evolution in environment. This is done by the creation within a machine of a Populace of Individuals represented by Chromosomes, in spirit a set of character strings that are similar to the base-4 chromosomes that we see in our own DNA. The individuals in the populace then go through a process of evolution.

IV. ISSUES IN DE-DUPLICATION

Backups can take longer and use a lot of CPU cycles in the process of de-duplicating data, possibly introducing performance issues on production machineries. Though, as we'll converse future in the white paper, a new technology called performance-optimized source de-duplication can exclude maximum of source de-duplication's presentation tradeoffs. All copies that existed prior to de-duplication essential be sent above the grid, possibly causing a bandwidth bottleneck. The choice of source versus target de-duplication will be contingent on which restraint client CPU handling above or bandwidth considerations is most important to your organization.

V. RELATED WORKS

Harsha Nagarajaiah *et.al.* [6] represented the embedded processors are capable of providing the need computational provision if they were to holder security functions in the field. When likened to the algorithmic presentation on extraordinary end scheme, viz. Intel Core 2 Duo CPU, the positive results obtained make a case for by the Atom CPU in networked requests employing mobile plans. The system may be applied to conservative de-duplication difficulties such as originate in address management as glowing as more progressive problems such as banned image recognition. The scheme usages the AURA design match approaches instigated within facility oriented structural design. The method shapes on the PMS & PMC expertise industrialized in the DAME science project.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 46-50
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

Jim Austin *et.al.* [7] introduced the data de-duplication capability to resolution the problematic of dismissed material in the course of backup by scheming and applying a backup scheme with bright data de-duplication named Backup Dedup which includes four de-duplication strategies, who is CDC,SIS, FSP& SW.

Guofeng Zhu *et.al.* [5] proposed Backup Dedup supports the online base sideways de-duplication & is proficient of selecting dissimilar de-duplication procedures according to the corresponding data types. Temporarily, it agreements the information dependability and safety in the backup process. The experimental test results show that Backup Dedup employs multi de-duplication approaches concurrently to substantially remove redundant data in the backup process so as to spread the goal of efficiently saving storing space and grid bandwidth.

Yueguang Zhu *et.al.*[8]**]**represented block-level data de-duplication joint with alike file recognition. At the interval of assuring the de-duplication elimination ratio, we narrow the variety of information to decrease the meta-data and eliminate presentation bottlenecks. We present a detailed evaluation of our technique and additional current information de-duplication techniques, and we appearance that our method meets its enterprise goals as it recovers the de-duplication relation while reducing overhead costs.

Tin-Yu Wu, *et.al.* [9] represented the index name servers (INS) to achieve not individual file storing, information de-duplication, enhanced node collection, and server capacity balance, but similarly file compression, chunk identical, real-time response control, IP info, and busy level index monitoring. To manage and enhance the storing nodes established on the client-side transmission station by our planned INS, all knobs must elicit optimum presentation and offer appropriate resources to clients. In this method, not only can the performance of the storage system be better, but the files could also be sensibly dispersed, lessening the workload of the storing nodes.

Vasilios *et.al.* [14] 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 & 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.*[15]proposed relocation methods taking into account (dynamic, receptive & 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 & the Clients/Server mode. They can likewise handle unpredicted the glimmer group activity with little cost. It likewise demonstrates that the cloud cost & server transmission capacity picked assume the most essential parts in sparing expense, while the distributed storage size & cloud substance upgrade system assume the key parts in the client experience change.

C. Ward *et.al.* [16]represented Acquainted the augmentations with a coordinated mechanization capacity called the Darwin structure that empowers on load movement for this situation & talk about the effect that computerized relocation has on the expense & dangers ordinarily connected with relocation to cloud.

Kang *et.al.*[17] proposed the migration algorithm .The VM to its best PM specifically, with the proviso that it has adequate capability. Then, if the relocation restriction is gratified, we transfer a different VM from this PM to oblige the new VM. In addition, they are learning a hybrid system where a lot is working to recognize forthcoming VMs for the on-line expansion. Assessment results establish the high competence of our method.

VI. BENEFITS OF DE-DUPLICATION

1. Reduced storage allocation: De-duplication approach eliminates the duplicate file from the storage and reduces the storage space during live production phase.
2. Fast processing speed: while the dataset reduced with de-duplication approach the processor need to search small dataset instead of a large amount of so data so that the response time will be reduced and users can get faster speed.
3. Efficiently growth network bandwidth: If the user's data upload will be discards with de-duplication approach. The bandwidth will automatically optimize and utilized.
4. A greener situation can be reached: de-duplication reduce the extra usage of the resources and reduce the volume of the structure. So if the things will be reduced then the production will automatically effect. So that in this process greener situation can be reached
5. Fast Recoveries ensure that line-of business process continue unimpeded.
6. This property in your storing appliance assistances in quicker recovery and ensures that data continuity and disaster recovery plans are very well set-up.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 46-50
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

7. Since you're purchasing and preserving less stowage, fast return on investment can be obtained and thus reduces overall storage costs.

VII. How to use Data De-duplication in Cloud?

Data de-duplication is a method to recognize that information which have the similar contents and only store one reproduction of them. Therefore, data de-duplication can spend less the cloud storing volume and utilize cloud stowage more properly. According to the original cloud storing structures, some of structures store the entire file hooked on the storage server without any de-duplication. Thus, if there are two like files, the cloud storing server would collect redundant blocks among these two like files. Therefore, the cloud storing volume cannot be used correctly. There are certain cloud storage vendors using the technique of data de-duplication while storing the uploaded records, Drop Box for example. Some data de-duplication schemes calculate hash implication for all file used to approve whether now terminated hash assessment amongst uploaded files in the cloud storage is. Others interpret a folder hooked on n blocks and then compute a hash value to signify every block; consequently, the cloud storing server can inspect the redundancy of each hash value of chunks.[10]

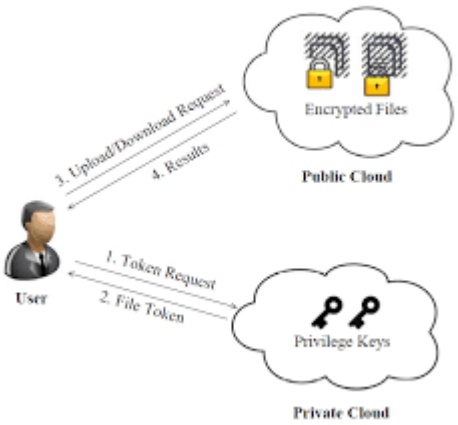

Fig. 3. De-duplication approach in cloud storage network[13]

VIII. Conclusion

Data de-duplication technology tremendously improves the efficiency of disk-based backup, decreases the quantity of deposited information, and changes the way information is threatened. Several key characteristics distinguish Falcon Storage data de-duplication results since other de-duplication explanations the design and application of a backup scheme be contingent on intelligent info de-duplication. De-duplication was deliberate to safeguard the information safety by counting differential benefits of customers in the identical copy checked. The performance of a little new de-duplication developments supporting authorized identical copy in crossbreed cloud structural design, in that the duplicate check symbols of leaflets are manufactured by the private cloud server taking private keys an intention to eliminate redundant data in backup process.

IX. References

[1] R. Barga, "Cloud computing architecture and application programming," *ACM SIGACT News*, vol. 40, no. June, p. 94, 2009.

[2] A. Huth and J. Cebula, "The basics of cloud computing," *United States Comput.*, pp. 1–4, 2011.

[3] M. Meenakshi, "An overview on cloud computing technology," *... Adv. Comput. Inf. Technol.*, no. 3, pp. 244–246, 2012.

[4] D. Geer, "Reducing the storage burden via data deduplication," *Computer (Long. Beach. Calif).*, vol. 41, no. 12, pp. 15–17, 2008.

[5] G. Zhu, X. Zhang, L. Wang, Y. Zhu, and X. Dong, "An intelligent data de-duplication based backup system," *Proc. 2012 15th Int. Conf. Network-Based Inf. Syst. NBIS 2012*, pp. 771–776, 2012.

[6] H. Nagarajaiah, S. Upadhyaya, and V. Gopal, "Data de-duplication and event processing for security applications on an embedded processor," *Proc. IEEE Symp. Reliab. Distrib. Syst.*, pp. 418–423, 2012.

[7] J. Austin, A. Turner, and S. Alwis, "Grid enabling data De-duplication," *e-Science 2006 - Second IEEE Int. Conf. e-Science Grid Comput.*, pp. 1–6, 2006.

[8] Y. Zhu, X. Zhang, R. Zhao, and X. Dong, "Data de-duplication on similar file detection," *Proc. - 2014 8th Int. Conf. Innov. Mob. Internet Serv. Ubiquitous Comput. IMIS 2014*, pp. 66–73, 2014.

[9] T. Y. Wu, J. S. Pan, and C. F. Lin, "Improving accessing efficiency of cloud storage using de-duplication and feedback schemes," *IEEE Syst. J.*, vol. 8, no. 1, pp. 208–218, 2014.

[10] Lin, Iuon-Chang, and Po-ChingChien,."Data Deduplication Scheme for Cloud Storage."*International Journal of Computer and Control(IJ3C)*,Vol1 2(2012).

[11] Cloud Computing -A Primer - The Internet Protocol,*http://www.cisco.com/c/en/us/about/press/internet-protocol-journal/back- issues /table-contents-45/123-cloud1.html*.

[12] Data De-duplication, *http://www.cd-datahouse.co.uk/solutions/de-duplication-of-data*.

[13] Choudary, Bhutan, and Amit Dravid. "A Study on Authorized Deduplication Techniques in Cloud Computing." *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* Volume 3 (2014).

[14] 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.

[15] Haitao Li, LiliZhong, Jiangchuan Li, , Bo Li, KeXu, *"* Cost-effective Partial Migration of VoD Services toContent Clouds", *2011 IEEE 4th International Conference on Cloud Computing,* pp. 203-110, 2011.

[16] 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.

[17] Kangkang Li, HuanyangZheng, & JieWu . "Migration-based Virtual Machine Placement in Cloud Systems", *2013 IEEE 2nd International Conference on Cloud Networking (CloudNet, IEEE,* pp. 83-90, 2013.

[18] Goldberg, David E., and John H. Holland. "Genetic algorithms and machine learning." Machine learning 3.2 (1988): 95-99.