The Survey of data storage of De-duplication Process, Strategies and Encryption Techniques

Manpreet Kaur[1], Anurag Jain[2]
[1]*M.Tech Student,* [2]*Associate Professor*
Computer Science Department, Chandigarh Engineering College, Landran, Mohali, India

Abstract: Data de-duplication could minimize the cost and enhance accuracy in backup systems. Currently, it becomes highly popular to apply this approach in the primary storage system, in which information is intensively used by e-Commerce applications. In cloud storage services, de-duplication technology is normally used to decrease the space and bandwidth requirements of services by eliminating the redundant data and storing a single copy of them. De-duplication process is most effective when multiple users outsource the same data to the cloud storage. De-duplication technique use different stages to define duplication process. Data de-duplication technology is also used to optimize the storage system that in turn reduces the amount of data, and hence thereby reducing energy consumption and decreasing the heat emission. Data compression can decrease the number of disks used in the operation to reduce disk energy consumption costs. This paper discusses about the De-duplication process, its strategies and encryption techniques i.e. symmetric and asymmetric approaches.

Keyword *- Data De-duplication, De-duplication Process Sta ges, De-duplication Strategy, encryption techniques.*

I. INTRODUCTION

Recently, cloud computing is becoming increasingly more imperative and being more used. The amount of data over the network or stored in a computer is continually increasing. Thus, the dispensation of this increasing mass of data requires more computer equipment to meet the several needs of organizations [1]. Cloud computing is an inescapable trend in the future expansion of computing technology. Its critical importance lies in its proficiency to provide all the users with high presentation and consistent calculation. Cloud computing is the progression of dispersed computing, grid figuring, and many other systems. In cloud computing, data is growing from desktop system for data centers. By means of virtualization technology, one corporeal host can be virtualized into numerous virtual hosts and use these clouds as a basic computing unit. Data de-duplication enables data storage systems to find and remove duplication within data without compromising its availability. [1] The goal of data de-duplication is to store more data in less space by storing and maintaining files (blocks in fine grained de-duplication manner) into a single copy, where the terminal copies of data are replaced by a reference to this copy. Data De-duplication in the cloud is a technique to identify those data which have the same contents and only store one copy of them[2]. Therefore, data de-duplication can economize the cloud storage capacity and utilize cloud storage more effectively. According to the original cloud storage schemes, many of them store the complete file into the storage server without any de-duplication. Thus, if there are two similar files, the cloud storage server would store redundant blocks of two similar files. Therefore, the cloud storage capability cannot be used efficiently. There are several clouds, storage vendors using this technique of data de-duplication for storing the uploaded files, the Drop Box for example. Some data de-duplication scheme calculates a hash value for each file and use that hash value to check whether there already exists a redundant hash value among uploaded files in the cloud storage. While other schemes transform a file into n blocks & then calculate a hash value to represent e-very block; therefore, the cloud storage server can examine the redundancy of every hash value of new uploaded blocks .De-duplication is defined with different stages of process. Duplication is a data reduction technique, commonly used in disk-Based backup systems, storage systems designed to reduce the use of storage capacity [3].

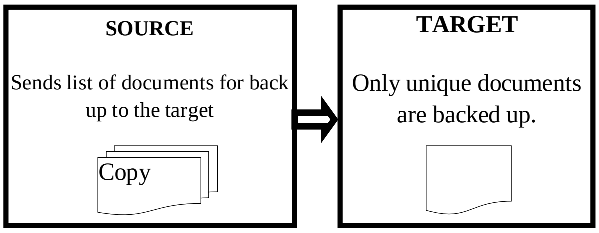
SOURCE
Sends list of documents for back up to the target
Copy
TARGET
Only unique documents are backed up.

Fig.1: Schematic diagram of Data De-duplication

Data De-duplication is widely used in back-up and archiving system to minimize storage space usage and energy consumption. Currently, both academic and ecommerce communities are defining to apply this approach in primary storages. Though, data de-duplication is mature in backup and attain systems, various challenges appear in the primary storage system environment. Nearly impressive data de-duplication ratio, a practical in line de-duplication system should deliver satisfactory execution with minimal operational over-head and sufficiently high throughput/accuracy. Data reliability is a must for the principal storage. To pursue increase accuracy, most of the state of the art de-duplication systems use fingerprint comparison in its place of byte to byte comparisons [4].

II. RELATED WORK

Anurag Jain et al. [5] discussed the calculating taxonomy and their association with cloud computing. Then they have discussed the important characteristics, layered service model construction and deployment model of cloud environment. In the end they have acknowledged the several exploration challenges, cloud adoption experiments along with the applications of cloud computing. This paper is for those who heard the term "cloud computing" for the first time and require knowing about its taxonomy. Also this paper delivers an idea of enterprise contests of cloud computing and helps in classifying significant research directions in this area.

Anurag Jain et al. [6] discussed adaptable nature of cloud computing and random behavior of users, and also discussed load balancing as the main issue in cloud computing paradigm. An effectual load balancing technique can advance the performance in terms of efficient resource operation and higher customer satisfaction. Load balancing can be applied through task scheduling, resource allocation and task migration. Numerous parameters to analyze the presentation of load balancing method are response time, cost, data dispensation time and throughput. This paper validates a two level load balancer Method by combining join idle queue and join direct queue method.

Yi Lu et al. [8] have discussed the Join Idle Queue (JIQ) development method for load balancing. Authors have realized a two level preparation. To understand the concept of two levels scheduling, writer-s have used the dispersed scheduler. Number of schedulers is very less in assessment to quantity of virtual machines. Every scheduler will reserve a queue of idle virtual machines. On getting a task, scheduler first refers its idle file. If it discovers any virtual machine which is idle then it directly assigns the task to that virtual machine and eliminates that virtual machine from its idle queue. If it does not discover any idle virtual machine then it aimlessly allot that task to any virtual machine. Virtual machine afterward job conclusion, update about its position to any of the arbitrarily chosen idle queue connected with a scheduler.

Aggeliki Sgora et al, [9] main difficulties in wireless multichip networks is the development of programs in a fair and efficient manner. Time Division Multiple Access appears to be one of the central solutions to realize this area, since itis a modest arrangement and can protract the devices' generation, by allowing them to communicate only helping of the time during chat. For that reasons numerous TDMA scheduling procedures may be found in his works. The scope of this paper is to categorize the current TDMA preparation procedures based on several factors, i.e. the object that is planned, the network topology material that is needed in command to produce or uphold the schedule and the entity/entities that achieve the computing for creating and preserving the lists, and to converse the advantages and drawbacks of each category.

Table no: 1 Description of the Related Work

Year	Techniques used	Performance Parameters
2014	Distributed Computing	No
2016	Join Shortest Queue	Response Time and Cost
2011	Join-Idle-Queue algorithm and randomized	Mean response Time
2013	Time Division Multiple Access	Throughput, Delay and Complexity

III. PROCESS OF DE-DUPLICATION

De-duplication process involves:

Identifying file types[7]dividing file data into chunks Calculating fingerprints of chunks, and Identifying and storing non-identical data. This de-duplication process is defined with different stages. All stages are categorizes are defined below:

Step 1: File Level De-duplication: For each incoming file, compute its fingerprint or hash value. Compare the fingerprint of incoming file with those already stored in the metadata using hash value as the key. If hash value matches with the existing one; then this file is not considered for the backup, because it is already being stored and is not modified later. If it is found that file is not identical with any of the previously stored file(s), continue with step 2. [10]

Step 2:- Chunk Formation: Divide the entire file into chunks using different chunking methods. Depending on the chunk granularity, compute its hash value using various hash algorithms MD5 (Message Digest), SHA-1 (Secure Hash Algorithm), tiger hash. Individual chunks are recognized by their unique chunk numbers.

Step 3:- De-duplication process :is applied on these chunks. Duplicate chunks are identified by matching them with the existing ones. Unique chunks are stored. If a duplicate chunk is found, then metadata (chunk index table) is updated with the duplicate reference. Performance of lookup process on chunk index table can be improved by caching a part of table entries, which can also avoid I/O lookup and disk bottleneck.

IV. STRATEGIES OF DE-DUPLICATION

Data de-duplication technology is used to identify duplicate data, eliminate redundancy and reduce the need to transfer or store the data in the main storage capacity. Data de-duplication technology use mathematics for every data element, "hash" processes to deal with, and get a single code called a hash authentication number. Each number is compiled into a list; this list is often referred to as hash index[11].

At present mainly the file level, block-level and byte-level deletion strategy, they can be optimized for storage capacity.

File-level data de-duplication Strategy :File-level de-duplication is often denoted as Single Instance Storage (SIS)[9], it checks the index back up or collection files stored in the storage for comparison. If the same file does not exist, it will store & update the index; else, shares the already present pointer to an existing file. Therefore, the same file saved only one instance, and then copy all the "stub" alternative, while the "stub" pointing to the unique file.

Block-level data de-duplication Strategy: Block-level data de-duplication Strategy divides the data stream into blocks, check the data block, and determine whether it met the similar data(usually on the application of the hash algorithm for each data block to form a digital signature or unique identifier) [12] [13]. If the block is unique and was written to disk, its identifier is also added in the index; otherwise, only the pointer to store the same data block's original location is added. This method pointer through a small-capacity alternative to the duplication of data blocks, slightly than storing duplicate data blocks over, thus saving disk storage space. Hash algorithm used to judge duplicate information, may lead to conflict between the hash error.

Byte-level data de-duplication: Analysis of data using byte stream level data de-duplication is another way. It stores bytes of data stream one by one, to achieve advanced accuracy. With byte-level technology products are generally able to "identify the content," In other words, the supplier of the backup procedure, examines the data flown to learn how to retrieve the file name, file type, date/time stamp and other information. In determining duplicate data, this method can reduce the computational load.

V. TECHNIQUES USED IN DE-DUPLICATION

Following are some basic methods used in de-duplication:

Symmetric Encryption

Symmetric encryption uses a communal secret key to encrypt & decrypt information. A symmetric encryption scheme is made up of three primary functions.

1) KeyGen SE (1λ) →: k is the key generation algorithm that generates k using security parameter 1λ;

2) Enc SE (k, M) → C: is the symmetric encryption algorithm that takes the secret k, and message M & then outputs the cipher text C, and

3) Dec SE (k, C) → M: is the symmetric decryption algorithm that receipts the secret k and cipher text C and then outputs the original message M.

Every user encrypts the data with their own encryption algorithm. In these identical data copies that produce the dissimilar cipher text, this makes the de-duplication process impossible [14].

Convergent Encryption

Convergent encryption encrypts a data copy through a convergent key, which is a resultant obtained by generating cryptographic hash value of the content of the data [13].In addition user derives a tag for the data copy, which is used to notice duplication After key generation and data encryption, users retain the keys and send the tag to the server side to check for the similar copy. It is assumed that if two copies are identical then their consistent tag values are also identical. Since encryption is deterministic identical data copies will make the similar convergent keys and same cipher text. This allows the cloud to execute de-duplication on cipher texts which can only be decrypted through corresponding data owners with theirs convergent keys [15].

Proof of ownership

Proof of ownership allows proprietorship of data co-pies on the server side. When tag value is similar in storage then it should be proven that which user/owner owns that file.

Table no. 2 Different between Symmetric Encryption and Convergent Encryption

Symmetric Encryption	Use Common secret Key
	Use three primary Function
	Encryption Algorithm
Convergent Encryption	Use convergent key
	Use hash Value
	Use cipher Text for execution de-duplication.

VI. CONCLUSION

This paper described that Data de-duplication process enables the data storage systems to find and remove duplication within the data. Data de-duplication is a process that is used in storage systems. The different techniques used in de-duplication process like Symmetric Encryption, Convergent Encryption and Proof of ownership techniques .Every techniques define a different process to solve de-duplication problem in storage systems. In Symmetric Encryption, techniques show Common secret key, primary function and Encryption Algorithm. In Convergent Encryption, it uses Convergent key, Hash function, and Chipper text. So every technique uses different way to solve the problem.

VII. REFERENCES

[1]. Lin, IuonChang, and PoChingChien. "Data de-duplication scheme for cloud storage." International Journal of Computer and Control (IJ3C), Vol1 2 (2012).

[2]. Jiang, Tao, Xiaofeng Chen, Qianhong Wu, Jianeng Ma, Willy Susilo, and Wenjing Lou. "Secure and Efficient Cloud Data De-duplication with Randomized Tag." IEEE Transactions on Information Forensics and Security(2016).

[3]. Fu, Min, Patrick PC Lee, Dan Feng, Zuoning Chen, and Yu Xiao. "A simulation analysis of reliability in primary storage

deduplication."In Workload Characterization (IISWC), 2016 IEEE International Symposium on, pp. 1-10.IEEE, 2016.

[4]. Jiang, Tao, Xiaofeng Chen, Qianhong Wu, Jianfeng Ma, Willy Susilo, and Wenjing Lou. "Secure and Efficient Cloud Data De-duplication with Randomized Tag." IEEE Transactions on Information Forensics and Security(2016).

[5]. Jain, Anurag, and Rajneesh Kumar. "A Taxonomy of Cloud Computing." International Journal of Scientific and Research Publications 4, no. 7 (2014): 1-5.

[6]. Jain, Anurag, and Rajneesh Kumar. "A multi stage load balancing technique for cloud environment." In Information Communication and Embedded Systems (ICICES), 2016 International Conference on, pp. 1-7. IEEE, 2016.

[7]. Meyer, Dutch T., and William J. Bolosky. "A study of practical deduplication." ACM Transactions on Storage (TOS) 7, no. 4 (2012): 14.

[8]. Lu, Yi, QiaominXie, Gabriel Kliot, Alan Geller, James R. Larus, and Albert Greenberg. "Join-IdleQueue: A novel load balancing algorithm for dynamically scalable web services." Performance Evaluation 68, no. 11 (2011): 1056-1071.

[9]. Sgora, Aggeliki, Dimitrios J. Vergados, and Dimitrios D. Vergados. "A survey of TDMA scheduling schemes in wireless multi hop networks." ACM Computing Surveys (CSUR) 47, no. 3 (2015): 53.

[10]. Jain, Anurag, and Rajneesh Kumar. "A multi stage load balancing technique for cloud environment." In Information Communication and Embedded Systems (ICICES), 2016 International Conference on, pp. 1-7. IEEE, 2016.

[11]. He, Qinlu, Zhanhuai Li, and Xiao Zhang. "Data de-duplication techniques." In Future Information Technology and Management Engineering (FITME), 2010 International Conference on, vol. 1, pp. 430-433. IEEE, 2010.

[12]. Min, Jaehong, Daeyoung Yoon, and Youjip Won. "Efficient de-duplication techniques for modern backup operation." IEEE Transactions on Computers60, no. 6 (2011): 824-840.

[13]. Bellare, Mihir, SriramKeelveedhi, and Thomas Ristenpart. "Message-locked encryption and secure de-duplication."In Annual International Conference on the Theory and Applications of Cryptographic Techniques, pp. 296-312.Springer Berlin Heidelberg, 2013.

[14]. Li, Jin, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick PC Lee, and Wenjing Lou. "Secure deduplication with efficient and reliable convergent key management." IEEE transactions on parallel and distributed systems 25, no. 6 (2014): 1615-1625.

[15]. Cochran, William T. "Secure encryption algorithm for data de-duplication on untrusted storage." U.S. Patent 8,199,911, issued June 12, 2012.