(IJUSE) 2018, Vol. No. 4, Jan-Dec **e-ISSN: 2454-759X, p-ISSN: 2454-7581**

THE SURVEY OF DATA STORAGE OF DE DUPLICATION PROCESS, STRATEGIES AND ENCRYPTION TECHNIQUES

Akshata Rahul Pathak

Dempo Higher Secondary School of Science, Pace

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 distributed storage administrations, de duplication innovation is regularly used to diminish the space and transmission capacity prerequisites of administrations by wiping out the repetitive information and putting away a solitary duplicate of them. Deduplication procedure is best when numerous clients redistribute similar information to the distributed storage. Deduplication system utilizes various stages to characterize the duplication process. Information deduplication innovation is likewise used to improve the capacity framework that thusly diminishes the measure of information, and subsequently along these lines lessening vitality utilization and diminishing the warmth discharge. Information pressure can diminish the number of circles utilized in the activity to lessen plate vitality utilization costs. This paper examines the Deduplication process, its procedures and encryption strategies, for example, symmetric and lopsided methodologies

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 centres. 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 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.

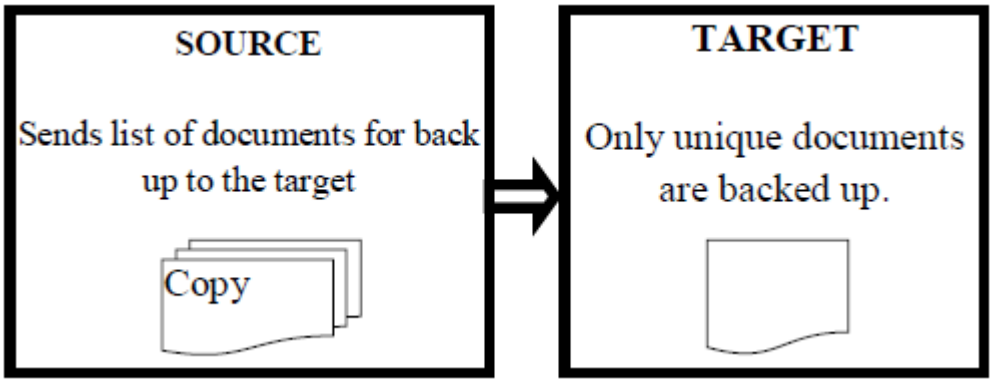

Fig.1: Schematic diagram of Data De-duplication

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. [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 has 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 allots 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]

(IJUSE) 2018, Vol. No. 4, Jan-Dec **e-ISSN: 2454-759X, p-ISSN: 2454-7581**

Step 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 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. TECHNIQUES USED IN DE DUPLICATION

Following are some basic methods used in de duplication:

Symmetric Encryption

It 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.

Each client encrypts the information with their very own encryption calculation. In these indistinguishable information duplicates that produce the unique figure message, this makes the de duplication process inconceivable [14].

Convergent Encryption

It encodes a data copy through a unified key, which is a resultant gotten by delivering a cryptographic hash estimation of the substance of the data [13]. Additionally, the customer surmises a tag for the data copy, which is used to see duplication After key age and data encryption, customers hold the keys and send the tag to the server-side to check for the similar copy. It is acknowledged that in case two copies are indistinct, by then their solid mark regards are moreover undefined. Since encryption is deterministic indistinct data copies will make the relative centered keys and same figure content. This empowers the cloud to execute de duplication on figure works which must be decoded through contrasting data owners and their combined keys [15].

Proof of ownership

Proof of ownership allows proprietorship of data copies 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 technique defines 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.