Elimination of Redundant information in Cloud Storage using Hash technique

[1]M .SriRama Laksmi Reddy [2] Dr.M.Janga Reddy

1. RESEARCH SCHOLAR JJT UNIVERSITY, CHURELA, JHUNJHUNU, RAJASTHAN
2. PROFESSOR, DEPT OF CSE, CMRIT KANDLAKOYA(V), HYDERABAD, INDIA

Abstract: **With a rise within the usage of cloud storage, effective ways got to be used to cut back hardware prices, meet the information measure needs and to extend storage potency. this could beachieved victimization information Deduplication. informati on Deduplication could be a methodology to cut back the storage would like by eliminating redundant information. therefore by storing less information you'd would like less hardware and would be able to higher utilize the present cupboard space.**

I.INTRODUCTION

The utilization of cloud for putting away information by organizations for reinforcement and ordinary citizens for sharing data among companions has expanded radically in the course of recent years. This has made a test to the cloud specialist organizations to keep up this huge information and to offer these administrations at a lower cost to the clients. In all actuality, the majority of the information put away in the servers is frequently rehashed. For instance, an administration may contain a few examples of same information document; putting away every one of these occasions would require a lot of storage room. This issue can be comprehended by utilizing Data Deduplication procedure.

Information deduplication stores just a single one of a kind occurrence of the information sort on the plate or tape. In this technique excess information is supplanted with a pointer to the extraordinary information duplicate. This decreases the equipment used to store information and the data transmission costs required for transmitting and getting purposes. Square and bit level deduplication strategies can accomplish pressure proportions of 20x to 60x, or much higher, under the appropriate conditions.

II. DEDUPLICATION vs. COMPRESSION

Deduplication is sometimes confused with compression, another method for reducing storage requirements [5]. While deduplication eliminates redundant data, compression uses algorithms to save data more concisely. Compression may be lossless compression or lossy compression but when you consider the case of deduplication, no data is lost as it only eliminates extra copies of data. Deduplication often has a larger impact on bacakup file size than compression. In a typical enterprise backup situation, compression may reduce backup size by a ratio of 2:1 or 3:1, while deduplication can reduce backup size by up to 25:1, depending on how much duplicate data is in the systems

III. HOW DEDUPLICATION WORKS?

Information deduplication works by contrasting articles (typically documents or pieces) and expels objects (duplicates) that as of now exist in the informational index. Every one of the procedures which are not one of a kind are evacuated in this strategy.

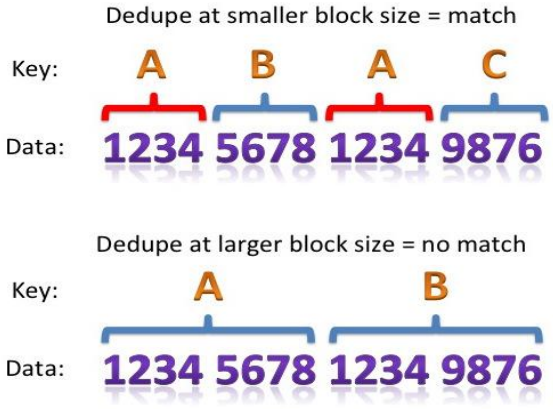
In Data deduplication strategy we partition the information into pieces and a hash esteem is computed for each of these squares. At that point utilizing these hash values we can decide if another piece of same information has as of now been put away. On the off chance that a comparable information record is discovered then supplant the copy information with a reference to the question effectively introduce in the database.

Fig: DEDUPLICATION

Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single-instance (data) storage

Hash-based algorithms:

Hash based deduplicatoin methods use algorithms to identify chunks of data. If the hash is already created, the data is identified as a duplicate and is not stored. Commonly used algorithms are Secure Hash Algorithm[8] 1(SHA1) and Message- Digest Algorithm 5(MD5). consists of four similar stages,termed rounds. Each round consists of 16 similar operations based on a non- linear operation F. There are four possible functions of this function and a different one is used for each round. Constants The main algorithm then uses each of these messages in turn to modify the state. The processing of a message block



SHA-1:

This was devised to create cryptographic signatures for security application. The 160-bit value created by SHA-1 is unique for each piece of data, it breaks data into "chunks" which are either fixed or variable in length. This processes the "chunk" with hashing algorithm to create a hash, If the hash already exists, the data is deemed to a duplicate and is not stored. If the hash does not exist, then the data is stored and the hash index is updated with the hash.[8]

IV.TECHNICAL CLASSIFICATION

MD5:

This 128-bit has was also designed for cryptographic uses. In this method the 128-bit state is divided into four 32-bit words, denoted A, B, C and D. These are initialized to certain fixed

Post-process deduplication (PPD):

It is otherwise called offbeat deduplication or disconnected deduplication. It includes the expulsion of excess information after reinforcement is finished and information has as of now been composed to capacity. The advantage of this strategy is that reinforcement information is direct and takes less time in light of the fact that the figurings of hash qualities and query happens simply after the majority of the information is put away.

In-line deduplication:

This process involves the calculation of hash values as the data enters the system. The benefit of this process over post-process is that it will take very less space because the calculation of hash values and the lookup process is completed before the data enters the database. So only one instance of a particular data is stores and the duplicate data is reference to the data present in the server.

Source deduplication:

This type of deduplication is the best suited to use at remote offices for backup to the cloud. The deduplication takes place typically within[3] the system by regularly scanning new files creating hashes and compares them to the hashes of existing files. It offers a number of benefits, including the reduction of bandwidth and the amount of data that has to be sent to the cloud server.

Target deduplication:

This is best suited for the use in the data center for the reduction of massive data sets. In this case, the client is unmodified and is not aware of any deduplication. Target deduplication requires that the target backup server or dedicated Hardware target appliance handles all of the deduplication. This process requires more network resources compared to source deduplication because the original data, with all its redundancy, must go over the network.

File and Sub-file Level Deduplication*:*

The full file level duplicates easily can be eliminated by calculating single checksum of complete file data and comparing it against existing checksums of the already backed up files. This method of deduplication is simple and fast, but the extent of deduplication is less, as this process does not address the problem of duplicate files or data-sets.

The sub-file level [2]. deduplication breaks the file into smaller fixed or variable size blocks, and uses hash based algorithm to compare these blocks and find similar blocks.

Fixed-Length Blocks:

A fixed-length block approach[8] divides the files into fixed size length blocks and uses a simple checksum-based approach (MD5/SHA etc.) to find the duplicates. This process has a limited effectiveness. The reason for this is that the primary opportunity for data reduction is in finding

duplicate blocks in two transmitted datasets that are mostly-but not completely of same data segment.

Variable Length Data Segment technology[1].: This technique divides the data stream into variable-length data segments using a methodology that can find the same block boundaries in different locations and contexts. This allows the boundaries to float within the data stream so that changes in one part of the dataset have little or no impact on the boundaries in other location of the dataset.

Algorithm:

Step 1: start Deduplication
Step2: identify point of application(source or destination)
Step3: if it is source backup and upload file
Step3: else identify time of application(post process or inline process)
Step4: if post process do sub file in fixed or variable length

V.PRACTICAL APPLICATIONS

Data deduplication helps to achieve data optimization and capacity scaling goals. It offers practical ways for the cloud service providers to achieve these goals. These ways include the following.

Capacity optimization: Data deduplication reduces the physical space used for storing data. This achieves greater storage efficiency than was possible by using features such as Single Instance Storage (SIS) or NTFS compression. This method uses subtle variable-size chunking and compression, which deliver optimization ratios of 2:1 for general file servers and up to 20:1 for virtualization data.

Scale and Performance:

Data deduplication can process up to 50 MB per second in typical windows derver 2012 R2, about20 MB of data per second in Windows Server 2012. It can work
on multiple volumes simultaneously without effecting other workloads on the server.

Reliability and data integrity: Data deduplication process maintains the integrity of data. This method uses checksum, consistency and identity validation to ensure data integrity. For all metadata and most frequently referenced data, data deduplication maintains redundancy to ensure that the data is recoverable in the event of data corruption.

Bandwidth efficiency with BranchCache: Through integration with BrachCache, the same optimizator techniques are applied to data transferred over the WAN to a branch office. The result is faster file download times and reduced bandwidth consumption.

Optimization management with familiar tools: Data deduplication functionality built into Server Manager and Windows PowerShell. Default settings can provide savings immediately, or administrators can fine-tune the settings to see more gains. One can easily use Windows PowerShell cmdlets to start an optimization job or schedule one to run in the future. Installing the Data Deduplication feature[6] and enabling deduplication on selected volumes can also be accomplished by using an Unattend.xml file that calls a Windows PowerShell script and can be used with Sysprep to deploy deduplication when a system first boots.

VI. SECURITY

The main real downside with it is a security gap in one of its essential properties. Consider a record being transferred, then the question emerges "Has anybody put away an earlier duplicate?" That implies is this specific document as of now put away or not? This question is replied by the aggressor, asking for to transfer a duplicate of the document and checking whether de-duplication happens. This being a limited inquiry, the appropriate response is either valid or false which does not give any data about who played out the assignment. Likewise, in the fundamental type of assault the assailant can just demand this question once. Once the inquiry is asked for by transferring the document, it is spared at the transfer benefit and consequently the response to the question is constantly positive. Assist, the data can be eradicated by the accompanying strategy; the aggressor begins transferring a document and

checks if de-duplication happens. In the event that de-duplication does not happen, a full transfer begins and the assailant close down the correspondence channel and the transfer ends. Thus, the duplicate of the record kept by the aggressor is not spared at the server, this thusly empowers the assailant to rehash a similar test at a later time and check if the document was transferred. Consequently, by utilizing this procedure at standard interims, the time window of the transferred record can be gotten.

VII.CONCLUSION

This paper examines the data about information deduplication for the cloud based frameworks. It incorporates the strategies that are utilized to accomplish savvy stockpiling and compelling data transfer capacity utilization by deduplication. The center idea includes dispensing with the copy duplicates

of the rehashed information by utilizing hashing calculations. Be that as it may, unwavering quality and speed are in question. The future test subsequently lies in recognizing more viable hashing calculations for enhancing the speed of putting away information and security. Be that as it may, information deduplication is the most urgent component for enhancing effectiveness of the cloud framework. This strategy will assume a noteworthy part in the cloud based administrations for putting away reinforcement information by both medium and huge endeavors.

VIII.REFERENCES

[1]. https://technet.microsoft.com/en-us/library/hh831602.aspx
[2]. http://www.computerworld.com/article/2474479/data-center/data-deduplication-in-the-cloud-explained--part-one.html
[3]. http://searchdatabackup.techtarget.com/definition/post-processing-deduplication
[4]. http://searchstorage.techtarget.com/definition/data-deduplication
[5]. https://en.wikipedia.org/wiki/Data_deduplication
[6]. http://www.webopedia.com/TERM/D/data_deduplication.html
[7]. http://www.druva.com/blog/understanding-data-deduplication/
[8.] https://pibytes.wordpress.com/2013/02/09/deduplication-internals-hash-based-part-2/

M.SriRama Laksmi Reddy received B.Tech from Annauniversity. He completed Master of Technology (Software engg from JNTU-H). And Pursuing Phd from JJTU Rajastan. My Research interest are Cloud computing, Network Security.

Dr.M. Janga Reddy Principal CMRIT Hyderabad, His Research Interest are Network Security, Cloud computing .