International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 78-81
e-ISSN:2278-621X

USE OF POINT IN TIME, SNAPSHOTS, DATA DEDUPLICATION AND HASH ALGORITHM

Riyaz Mohammed[1], Jackson Sequeira[2] and Vishaal Dharwadkar[3]

Abstract- Today, data protection remains a big task for almost many organizations.A big part of that can be attributed to the fact that organizations are creating more and more data.And, many businesses have very little tolerance for doing it, so fast restores are more important for protecting the data.There are many technologies for the data backup such as SAN,NAS,CAS and so on.But these tehcnologies used by the legacy organizations and very old ones.For the replacements for those technologies the data scientist came up with the new technologies like PIT,Snapshots,Datadeduplication and the Hashing Algorithms.Having said that these technologies are very effective while having the data backups.
Keywords –SAN,NAS,CAS,PIT,Snapshots,Datadeduplication,Hashing Algorithms

I. INTRODUCTION

Data backup is one of the most important areas of Information Technology and yet is also one of the most ignored. Data is essential to the smooth running of any business and is the essential start to any business plan. 'Backing up' means making a copy of your most important files; this can then be used if the original copy is lost. Preferably the second copy should be held at a different location to the original and be kept in a secure environment.Growing numbers of computer viruses are also a risk to business information, as once they have infected your machine they often delete or corrupt your data. This is another reason why backing up your data is such an important thing to do.

Data loss can happen in many ways, the most common causes are physical failure of your PC, accidental error, theft or disasters like fire, flood and dropped coffee mugs!.Thats why we are introducing the technologies like Point In Time,Snapshots,Datadeduplication and the Hashing Algorithms.With the help of these new technologiesthe organizations can handle the data backups very effectively.Let us discuss these technologies in details.

II.DATA BACKUP TECHNOLOGIES

A . POINT-IN-TIME BACKUP-

A point-in-time snapshots is a copy of a storage volume, file or database as they appeared at a given point in time and are used as method of data protection. in the event of a failure, users can restore their data from the most recent snapshot before the failure. many point-in-time snapshots are read-only. there are two main methods of keeping point-in-time snapshots up to date with changes:

[1] *AIMIT,St Aloysius College Mangalore, Karnataka, India*
[2] *St Aloysius College, AIMIT,Mangaluru, Karnataka, India*
[3] *St Aloysius College, AIMIT,Mangaluru, Karnataka, India*

- Pointer remapping -- When new copies of a point-in-time snapshot are made, the more recent copy will maintain a mapping to the original copy.

- Copy-on-write -- When changes are made to data, only the data that is modified will be copied again, rather than make another full copy of the data set.

Point-in-time snapshots are beneficial because a user can select a specific time they want to restore data from.Also user can have create multiple replica at different point in time.Figure1.shows an example in which a copy is created every six hours from the same source.

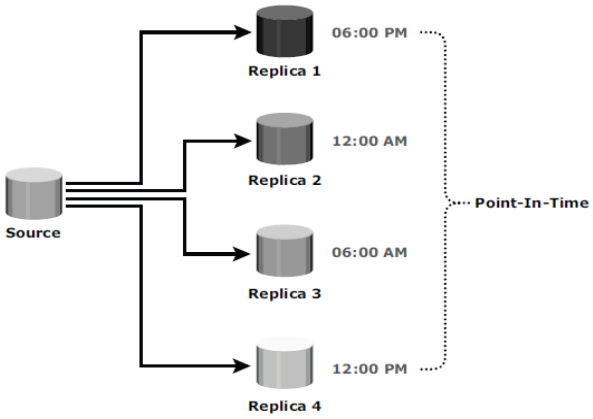

Figure 1.Multiple replicas created at different points in time

B. Snapshots Technology –

Snapshot technologies have long been a part of enterprise-class storage arrays. The ability to quickly create disk-based, point-in-time copies of data enables the data to be more easily accessed and used for business processes and administrative operations across both primary and secondary application environments. Snapshots are logical copies of data at a specific point in time so they can be created very rapidly. They can also be very space efficient because none data is copied up front just metadata representing the logical volume. Snapshots can be read only or read/write. The creation of a read-only snapshot effectively requires little or no space. During the creation of a read/write snapshot, some storage capacity is often reserved to accommodate the predicable changes to the writeable snapshot volume. Snapshots are a backup alternative that have been available for over a decade on most storage systems. In fact, most storage systems can now support hundreds of snapshot copies without any significant performances.The problem is these snapshots have limited ability to assist in the finding and recovering of isolated data. Essentially, most snapshots are only realistically usable to recover the last known good copy of an entire volume, not a subset of data that may be a old one. The potentially larger worry is that snapshots are exposed if the primary storage system that created them fails. If this happens, all snapshot data is lost. These technologies can be graded across four metrics:

- How efficiently copies dominate raw storage capacity for both metadata and data
- How long it takes to create the copy as well as the impact of copy operations on provision services
- Operational performance of the copy and what impact the copy's presence has on the source volume performance
- Management limitations the copy enforce on administrators

C .DataDeduplication-

Data deduplication is an advanced technology that can dramatically reduce the amount of backup data stored by eliminating redundant data. Data deduplication maximizes storage utilization while allowing IT to retain more nearline backup data for a longer time. This tremendously improves the efficiency of disk-based backup, changing the way data is protected. In general, data deduplication compares new data with existing data from previous backup or archiving jobs, and eliminates the redundancies. Advantages include improved storage efficiency and cost savings, as well as bandwidth minimization for less expensive and faster offsite replica of backup data.

Data deduplication works by comparing blocks of data or objects (files) in order to detect duplicates. Deduplication can take place at two levels that is file and sub-file level. In some systems, only complete files are compared, which is called Single Instance Storage (SIS). This is not as fluent as sub-file deduplication, as everyfiles have to be stored again as a result of any minor modification to that file.

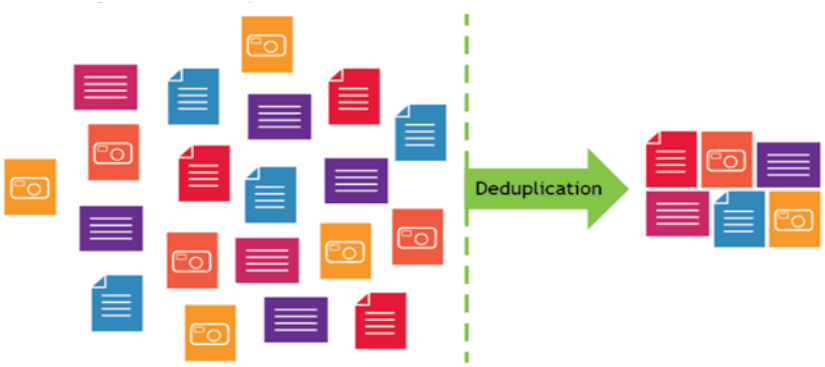

Figure 1.Duplication reduces the amount of stored data

D) Hash Algorithms-

Hash based data de-duplication methods use a hashing algorithm to identify “chunks” of data. Commonly used algorithms are Secure Hash Algorithm 1 (SHA-1) and Message-DigestAlgorithm 5 (MD5). When data is prepared by a hashing algorithm, a hash is created that represents the data. A hash is a bit string (128 bits for MD5 and 160 bits for SHA-1) thatrepresents the data handled. If you processed the same data through the hashing algorithm multiple times, the same hash is created each time.Hash based de-duplication breaks data into “chunks”, either fixed or variable length, and processes the “chunk” with the hashing algorithm to create a hash. If the hash already exists, the data is deemed to be a duplicate and is not stored. If the hash does not exist, then the data is stored and the hash index is modernize with the new hash.

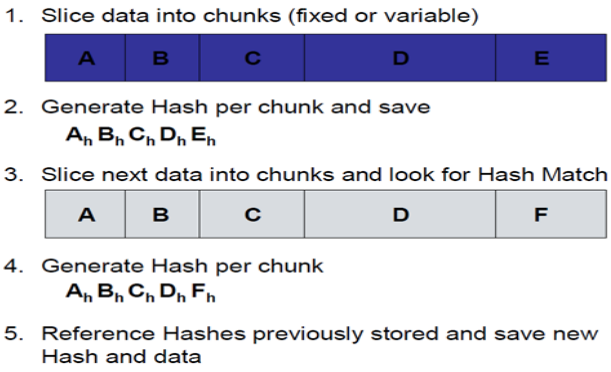

Figure 1. Hash based de duplication

In Figure 1, data "chunks" A, B, C, D, and E are processed by the hash algorithm and creates hashes Ah, Bh, Ch, Dh, and Eh; for purposes of this example, we assume this is all new data.Later, "chunks" A, B, C, D, and F are processed. F generates a new hash Fh. Since A, B, C, and D generated the same hash, the data is presumed to be the same data, so it is not stored again. Since F generates a new hash, the new hash and new data are stored.

III. DIFFRENCES BETWEEN DATA DEDUPLICATION AND HASHING ALGORITHM

Data deduplication works by measure blocks of data or objects or files in order to catch duplicates. Deduplication can take place at two levels — file and sub-file level. In some systems, only complete files are matched, which is called Single Instance Storage (SIS).

Hashing Algorithm works just same as data deduplication.But the main difference is in data deduplication the repeated data files and also called as redudantfiles.But in Hashing techniue the data scientists have the previlages which redundant files has to be removed from the backup so that any of the important files to miss or damanged from the storage medias.

IV.CONCLUSION

Here we discussed the methods or techniques for the efficient data backup that can be very usefull for the modern organizations to protect their important files and the informations.All these technologies are well suited for the backup usage.All these technologies can be very well implemented in big data.We will just look into this things.Big data can provide data more accurately and detailed. Big data can help in business to get decisions based on data facts that are more detailed, accurate and are processed and executed within minimum amount of time. Several companies spend quality amount of time in analyzing data pattern and then do strategies based on data patterns. Using Big Data we can get this data in minimum possible time which is helping any business to take decisions faster and grow.So these technologies are well suited for the implementation in big data analytics.

REFERENCES

[1] A. Venish and K. Siva Sankar"Study of Chunking Algorithm in DataDeduplication"
[2] Shikha Malik, Rajiv K Nath"point in time analysis-gathering bigdata over a small duration"
[3] AFA Snapshot Implementations Will Broaden Use of Snapshot Technology 'Sponsored by EMC'
[4] How Data DeduplicationWorks byFalconStor Software
[5] Andrew Burton and Brien M. Posey"Modern Data Protection Is More Than Backup "