ISSN 2320 2629

Volume 2, No.4, July – August 2013

International Journal of Information Technology Infrastructure

Available Online at http://warse.org/pdfs/2013/ijiti01242013.pdf

A Novel approach for unique Data Backup in Cloud Storage

Ch.Anilkumar
Research Scholar, DRKIST,
Hyderabad, India
anilkumar08515@gmail.com

Dr.R.V.Krishnaiah
Principal, DRKIST,
Hyderabad, India
r.v.krishnaiah@gmail.com

ABSTRACT

Today's growth of data in IT environments, the IT managements efforts and investments are getting raises in proportion with Data Center expansion, providing data security, adhering to the compliance policies, cooling expanses and increase in staff to manage Data Center.

Deduplication is a data reduction techniques where only unique data will be stored as part of backup .Reduction in the size of the data stored in backup destination will address the problems related to Data Center expansion, data security, excessive power consumption and buying expensive data storage devices like file servers, NAS devices etc.

Cloud storage in an online storage service over the network where the data will be stored in the cloud. Cloud storage will help to offload the active data on premise to cloud with pay-per-usage model. It'll address the problem related to Data Center expansion, adhering to compliance policies, cooling expanses and increase in staff to manager Data Center.

The aim of this paper is to propose a complete backup solution to address the above problems by integrating "Deduplication" and "Backup to cloud" functionalities in Backup application.

Keywords: Deduplication, Backup application, Cloud storage, Green computing,Hashing

1.INTRODUCTION

As the IT management expenses (especially Data center management) are exponentially increasing, now the IT focus is on reducing the data center operating expenses. Also now the IT trend is towards the green computing. Green computing [1] refers to environmentally sustainable IT by implementing energy efficient applications and reducing the power consumption. Utilizing the storage efficiently and effectively with minimal or no impact on the environment and with less IT management expenses should be focus while designing backup application. with the growth of data and its usage, nowadays more data is available online. This increase in online storage has increased power consumption and cooling expenses. Reducing the amount of data stored on storage media, while still the same amount of data is online will help to reduce the power consumption and helps to Green IT. Data centers, which have been criticized for its extraordinary high-energy demand, are a primary focus for proponents of green computing. The federal government has set a minimum 10% reduction target for data center energy usage by 2011[2].

Carefully designing a backup solution with Deduplication and cloud storage tightly integrated will help to reduce the IT expenses and environment impact.

2.RELATED WORK

Backup applications are meant for protecting the data by making copies of data in secondary storage devices like hard disk, tape, cloud etc. Backup will ensure that mission critical data can be always recovered in the case of data loss\corruption or disasters.

a. Data Deduplication

In backup technology, Data Deduplication [3] is a process of eliminating the redundant data. Data will be examined for duplicate chunks and only unique chunks will be backed up. All the subsequent redundant chunks will be replaced with a reference that points to the chunk already stored.

An advantage of deduplication includes:

1) Reduce the network bandwidth
2) Reduce the backup window, as the less data will be transferred to storage media
3) Reduce the disk space requirements and hence IT expenses

As Deduplication will be referring to previous data blocks, it'll be performing more random access operations. So Deduplication will be ideally implemented on disk-based storage, rather than on tapes, which are sequential. The implementation approaches are as follows:

Block level: Entire data will be divided into blocks (irrespective of files) and comparison will be done across these blocks will be replaced with a reference to the block that is already stored. This is more efficient in finding duplicates and need more processing power[4].

File level: Data in a file will be divided into blocks and comparison will be done. This is almost similar the regular data compression. It is less efficient, but need less processing power[4].

Application level: It uses the same approach as the block level Deduplication. But instead of entire data, only the application specific data will be divided into blocks and compared. For example, if there are multiple exchange servers in the environment, then data from each exchange server will be divided into blocks and compared with the blocks of other exchange servers [4].

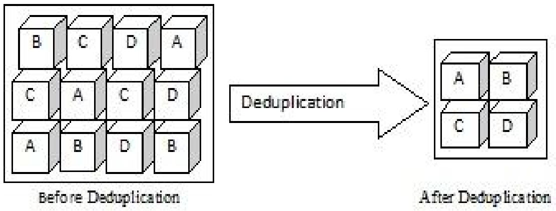

Figure 1: Deduplication example

Deduplication can be implemented phases are

1)***In-line***: When the data is read from the disk for backup, Deduplication will be performed as real time. Hence only unique data will be stored on storage media any time and needs lesser disk storage [5].

2) ***Post-backup****:* Entire data will be written to storage media first and then user can schedule deduplication to happen when the processing resources are less burdened [5].

Deduplication can be implemented at various places (considering the backup application to use client-server model):

1) ***Agent side***: Agent will perform the deduplication and only unique data will be traveled to the server over the network saving the network bandwidth

2) ***Server side:*** Entire data will be transformed from agent to server and server will perform the deduplication. This will help to decrease burden on the agent machine and only one high-end server machine need to be dedicated for reduplication process.

Deduplication savings depends on pattern of data, application type, metadata added to the files, frequency of data change and block size. If it is assumed that data change rate is 5% every day, then Deduplication compression ratio can be expected up to 90% for daily backups.

Deduplication % and ratio:
A= Size of source data
B= Size of data written to storage media
Deduplication Ratio=A / B
Deduplication %=A-B / A

For example, if the source data is 50 GB and 20 GB is written to the storage media after performing Deduplication,

Then deduplication ratio would be 2.5 (1 block of data written to storage media for every 2.5 blocks) and % would be 60 (60 % of disk space is saved)

B. Cloud Storage

Cloud Storage is a service delivered over a network (Internet / Intranet) in which data is maintained, managed and backed up remotely and made available to users. It provides on-demand data storage as a service.

Advantages of Cloud Storage

Reduced Cost: Cloud storage [6] is paid incrementally, reducing the initial investment. Also the storage cost is cheap, compared to hosting the storage in the LAN. Offloading the on-premise data will also helps in reducing the expenses with Data Center, Power consumption, IT staff etc.

More Mobility: Data can accessed from anywhere over the Internet.

Simplified Cost and Consumption Model: With the pay-per-usage model, only required storage space can be utilized for the required amount of time and storage space can be returned when done.

Right size to Address Business Changes: Storage can be expanded on demand to scale up the changing business needs.

Ease of Integration: It's cheap, doesn't require installation, doesn't need replacing / adding hardware, has backup and recovery systems, has no physical presence, requires no environmental conditions, requires no personnel and doesn't require energy for power or cooling.

Highly Secure Infrastructure: Data stored in cloud is secured and users no need to worry about the compliance as well.

3. BACKUP TO CLOUD STORAGE WITH DEDUPLICATION

Backup application will read the source data and the source data will be divided into blocks. Block size plays key role in the dedupliction process. Having smaller block size will increase the chances of finding duplicates, but more processing power &time will be required for hash calculation. Also more metadata has to be stored. Having bigger block size will be fast and less hash calculation need to be performed, but the chances of finding duplicates will be less. So carefully choosing the block size will provide the right balance between the deduplication ratio and computing power required. Data block size between 8 to 24 KB will provide optimum compression ratio with optimal computing power. While dividing the source data into blocks, blocksize can be fixed or variable. Fixed block size means, the block size will be pre-defined and entire data will be divided into blocks based on this size. Variable block size means, block size will be decided in the run time based on the amount of changed data between the backups.

Fixed block size Deduplication is inefficient, if there is any change in the data, but it is less complex and not much intelligence is required. Variable block size approach provides very good Deduplication ratios with built-in intelligent algorithm.

Source data divided into blocks:

Dedupl	ication an	d cloud	are lat	test tr	ends in
storage	technolo	gy			

Source data after small changes in 5th block with fixed block size of 8 characters:

Dedupl	ication an	d cloud	are lat	test tr	ends in
storage	technolo	gy			

Source data after small change in 5th block with variable block size:

Dedupl	ication an	d cloud	are lat	test tr	ends in
storage	technolo	gy			

■ changed data blocks

In the above example, fixed block size approach changed more blocks and hence Deduplication ratio will be very less. With variable block size, only one block is changed and other blocks remaind same resulting better Deduplication ratio.

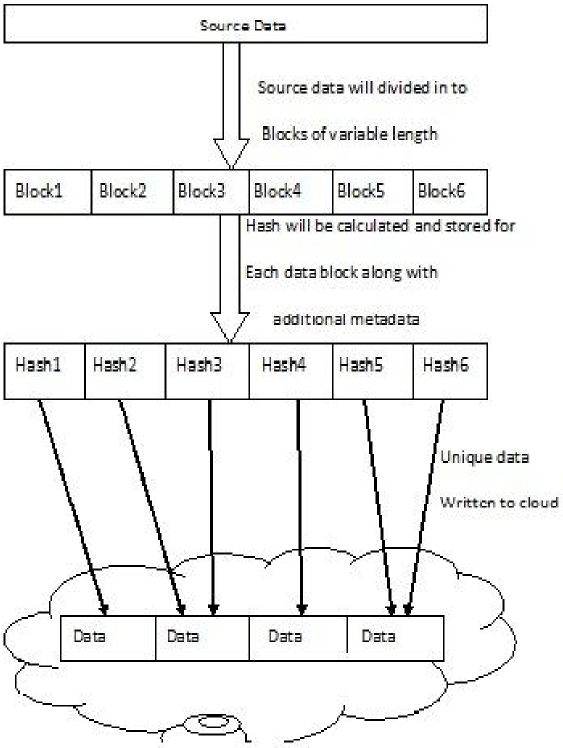

Figure 2: Dedpulication process with backup to cloud storage

Once the data is divided into blocks, hash algorithm will be applied on data blocks. Hash value will be calculated for each data block and compared with the existing hash values. If the hash values of current data block matches with any previous hash values, then it means that the data is identical. Hence no need to write the block to cloud. Instead, a reference needs to be added for current block to the block matching the hash value. This will ensures that only unique data blocks are written to cloud and still the complete data is protected.

Along with the hash value for each block, other metadata such as the block number, previous block number matching the hash value, Cloud storage location will be stored.

Hashing [7] is the transformation of a string of characters into a usually shorter fixed-length value or key that represents the original string. Hashing is used to index and retrieve items because it is faster to find it using the original value. For calculating the hash values, there are lots of widely accepted algorithms available such as MD5, SHA1, SHA 256 and Fletcher's checksum etc. Fletcher-32[8] checksum algorithm can be used for calculating the hash value. Fletcher checksum will divide the source data to be protected into "blocks" and calculates the check sum for those.

Block number	Hash value	previous block Matching hash	cloud block info

Figure 3: Metadata format

1	ABCDEF	0	Aws.amazon.com/ Account/myaccount/ <object1>
2	PQRSTU	0	Aws.amazon.com/ Account/myaccount/ <object2>
3	PQRSTU	2	Aws.amazon.com/ Account/myaccount/ <object3>

Figure 4: Metadata format example

In the Figure 2:, it is assumed that data in Block 2 and Block 3 are identical and also the data in block 5 and block 6 are identical.

In the Metadata format example in the figure 5, it is assumed that backup is performed to Amazon S3 cloud and cloud block info contain the Amazon S3 account name, bucket name and object information.

In the figure 4, block 1 with hash value 'ABCDEF' is not having any previous matching hash values and hence 'previous block matching hash' will be marked as '0'. As the data corresponding to block1 is not already stored, it'll be stored in the cloud and corresponding cloud account information will be updated in 'Cloud account info'. Data in block 2 also will be written to cloud in the same manner. But in the case of block 3, its hash value matches with the block 2 hash values. In this case, 'previous block matching hash' will updated with '2' and 'cloud account info' also will be updated with the value of block2 and no actual data will be written to the storage media i.e. cloud.

Meta data created, as part of deduplication will be stored locally instead of transferring to the cloud. This will ensure that even though intruder gains access to the data, it cannot be intercepted without metadata will be accessed frequently, it is better to keep it in the local premises to provide faster access. This will ensures that security concerns will be taken care automatically.

Deduplication ratio may not be promising with the first backup of source data, as the hash comparison will be done only between the data blocks within the backup. From the second backup onwards, hash comparison will be performed with Hash sets of the latest previous backup. Hash comparison can be performed with hash sets of the latest previous backup. Hash comparison can be done with hash sets from all previous backups, instead of latest previous backup. But this approach will take more time and processing power. Also there is a high probability that duplicates will be found when compared against the previous backup, rather than the entire backup sets. As the savings are not promising compared to the processing power utilized when hash values are compared with all previous backup sets, it is recommended to compare the hash values against the latest previous hash set. Assuming 5% of change in the source data every day, Deduplication savings can be more than 90% from the second backup onwards.

While calculating Deduplication benefits, data retention time on storage media also need to be considered. For example, if source data is 100 GB and retention period is 30 days, then Deduplication ratio from second backup onwards would be 1:64 and Deduplication percentage would be 95%

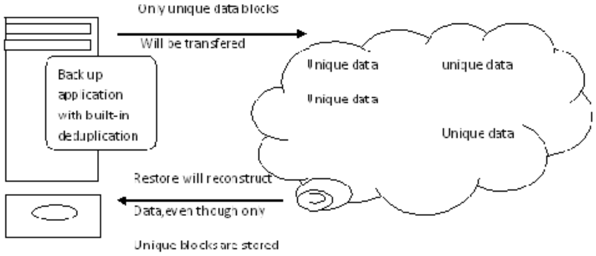

Figure 5: High level implementation

When the restore of data is attempted from the cloud storage, the metadata will be read first. From this metadata, data corresponding to each block will be restored from cloud using the cloud account information. The data blocks that are not actually stored in cloud, but referenced to other unique data blocks also can be restored using the same cloud account information of unique blocks. In this way entire source data can be reconstructed as part of restore.

With the help of Deduplication approach discussed here, significant disk space(of approximately 90%) will be saved and results in reducing the IT expenses related to data center expansion, data security, excessive power consumption and buying expensive data storage devices. With the help of involving the cloud storage in the backup process will address the challenges related to data center expansion, adhering to compliance polices, cooling expenses and increase in staff to manager data center.

Integrating Deduplication and cloud storage with the proposed approach will provide significant savings in the IT expenses and also helps to the green computing.

4. CONCLUSION

Management of continuously expanding data center and increasing IT expenses are the major concerns that every organization is facing with rapid growth of data and need for its online availability. This paper provides a design approach to solve these problems by integrating Dedupliation with backup to cloud storage along with moving towards green computing. This implementation recommends Inline block level Deduplication with variable block size for better results. Further work to this paper includes handling hash collision with deduplication and increasing data transfer rates over WAN to and from cloud storage.

REFERENCES

[1] Green computing San Murugesan, "Harnessing Green IT: Principles and Practices," IEEE IT Professionals, January-February 2008, pp 24-33.

[2] S. Murugesan, "Going Green with IT: Your Responsibility toward Environmental Sustainability," *Cutter Business—IT Strategies Executive Report*, vol. 10, no. 8, 2007.

[3] S. Pritchard, "IT Going Green: Forces Pulling in Different Directions," *Financial Times*, 30 May 2007.

[4] Y. Tan, H. Jiang, D. Feng, L. Tian, and Z. Yan. CABdedupe: A Causality-Based Deduplication Performance Booster for Cloud Backup Services. In *2011 IEEE International Parallel & Distributed Processing Symposium*, pages 1266-1277,2011.

[5] L.Xu, J. Hu, S. Mkandawire, and H. Jiang. SHHC: A Scalable Hybrid Hash Cluster for Cloud Backup Services in Data Centers. In *31st International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pages 61-65, June 2011.

[6] D. Meister and A. Brinkmann. dedupv1: Improving deduplication throughput using solid state drives (SSD). In *IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, pages 1-6, 2010.

[7] BHAGWAT, D., ESHGHI, K., LONG, D., AND LILLIBRIDGE, and M.Extreme Binning: Scalable, Parallel Deduplication for Chunk based File Backup. In *Proceedings of the* 17*th IEEE International Symposium on Modeling, Analysis, and Simulation (MASCOTS)* (2009), pp. 1–9.

[8] HONG, B., AND LONG, D. D. E. Duplicate Data Elimination in a SAN File System. In *Proceedings of the* 21*st IEEE /* 12*th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST)* (2004), pp. 301–314.