2010 International Conference on Future Information Technology and Management Engineering

Data Deduplication Techniques

Qinlu He, Zhanhuai Li, Xiao Zhang
Department of Computer Science
Northwestern Polytechnical University
Xi'an, P.R. China
luluhe8848@hotmail.com

***Abstract*—With the information and network technology, rapid development, rapid increase in the size of the data center, energy consumption in the proportion of IT spending rising. In the great green environment many companies are eyeing the green store, hoping thereby to reduce the energy storage system. Data de-duplication technology to optimize the storage system can greatly reduce the amount of data, thereby reducing energy consumption and reduce heat emission. Data compression can reduce the number of disks used in the operation to reduce disk energy consumption costs. By studying the data de-duplication strategy, processes, and implementations for the following further lay the foundation of the work.**

Keywords-cloud storage; green storage ; data deduplication

I. INTRODUCTION

Now, green-saving business more seriously by people, especially in the international financial crisis has not yet cleared when, how cost always attached great importance to the issue of major companies. It is against this background, the green store data de-duplication technology to become a hot topic.

In the large-scale storage system environment, by setting the storage pool and share the resources, avoid the different users to prepare their free storage space. Data de-duplication can significantly reduce backup data storage, reducing storage capacity, space and energy consumption [1]. The major storage vendors are launching related products or services, such as IBM in the IBM System Storage TS7650G ProtecTIER products using data deduplication solution (Diligent) can demand up to the physical storage devices to 1 / 25. By reducing the demand for storage hardware, reduce overall costs and reduce energy consumption. NetApp V Series supports redundant data removed, at least 35% can reduce the amount of data. Deletion of the redundant data technology, now more of a secondary storage doing filing, copying the same data when multiple copies of the deletion. Trend of the future will be, in the main storage system, the removal of duplicate data in real time.

II. DATA DEDUPLICATION

Duplication[2][3] is very simple, encountered repeated duplication of data when data is not saved backup, instead, add a point to the first (and only one) index data. Duplication is not a new thing, in fact it is just data compression derivatives. Data compression within a single file to delete duplicate data, replacing the first data point to the index. Duplication extend this concept as follows:

- Within a single file (complete agreement with data compression)
- Cross-document
- Cross-Application
- Cross-client
- Across time

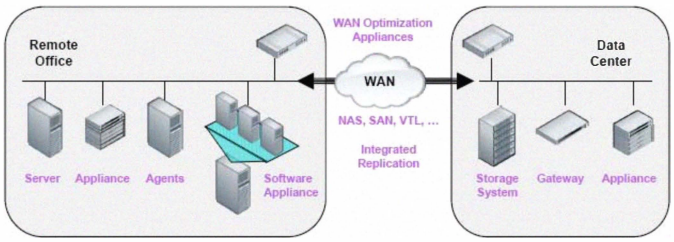


Figure 1. Where deduplication can happend

Duplication is a data reduction technique, commonly used in disk-based backup systems, storage systems designed to reduce the use of storage capacity. It works in a different time period to find duplicate files in different locations of variable size data blocks. Duplicate data blocks replaced with the indicator. Highly redundant data sets[4][5] (such as backup data) from the data de-duplication technology to benefit greatly; users can achieve 10 to 1 to 50 to 1 reduction ratio. Moreover, data deduplication technology can allow users to efficiently between the different sites, the economy back up data replication.

Compression through the compression algorithms to eliminate redundant data contained in a document to reduce file size, but duplication is distributed through the algorithm to eliminate the same file storage system or data block.

Data deduplication technology is different from the normal compression[13][14]. If you have two identical files, data compression, data will be repeated for each file and replace the exclusion of the first data point to the index; the duplicate data were excluded to distinguish the two documents are identical, so only Save the first file. Moreover, it with data compression, as the first file exclude duplication of data, thereby reducing the size of the stored data.

978-1-4244-9088-2/10/$26.00 ©2010 IEEE

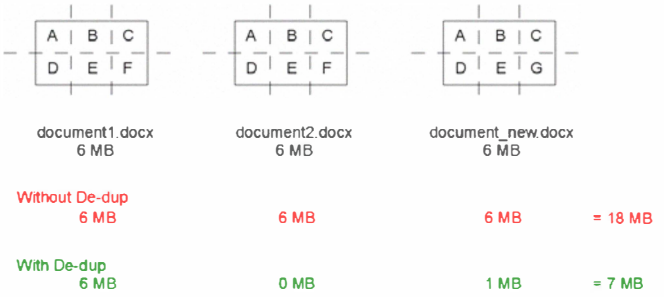


Figure 2. Data deduplication[6]

Duplication is also different from the normal incremental backup. The thrust of incremental backups back up only new data generated, and data de-duplication technology, the key is to retain only the only instance of the data, so data deduplication technology to reduce the amount of data storage has become more effective. Most manufacturers claim that their data deduplication products can be reduced to the normal capacity of 1/20. Data de-duplication technology, the basic principle is to filter the data block to find the same data block, and the only instance of a pointer to point to replace.

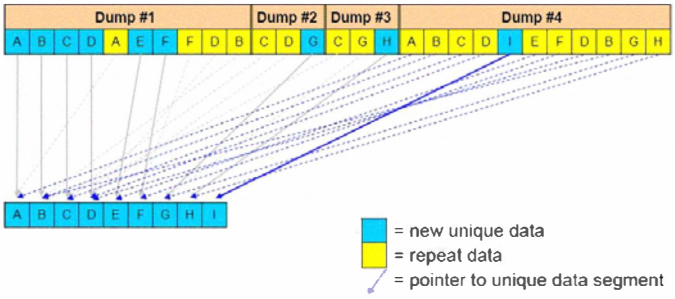
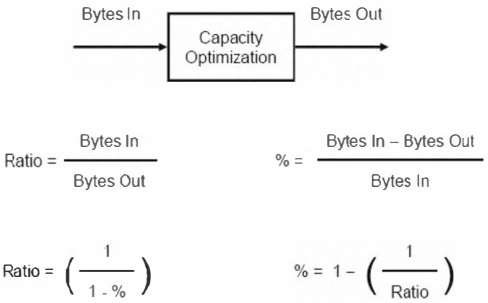


Figure 3. shows data deduplication technology

Basically, it can reduce the storage space occupied by the data. This will bring the following benefits[12]:

- IT savings funds (do not need the extra space needed to increase investment)
- Reduce the backup data, data snapshots of the size of the (cost-saving, saving time, etc.)
- Less power pressure (because of less hard, less tape, etc.)
- Save network bandwidth (because only fewer data)
- Saving time
- Because of the need less storage space, disk backup possible.

Backup equipment is always filled with a lot of redundant data. To solve this problem, save more space, "duplication" technology will be a matter of course become the focus of attention. Use "data deduplication" technology can store the original data reduced to 1/20, so that more backup space, not only can save the backup data on disk longer, but also can save offline storage a lot of bandwidth required.



III. DATA DEDUPLICATION STRATEGY

Data de-duplication technology to identify duplicate data, eliminate redundancy and reduce the need to transfer or store the data in the overall capacity [7][8]. Duplication to detect duplicate data elements, to judge a file, block or bit it and another file, block or bit the same. Data de-duplication technology to use mathematics for each data element, "hash" algorithms to deal with, And get a unique code called a hash authentication number.. Each number is compiled into a list, this list is often referred to as hash index.

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

A. File-level data deduplication strategy

File-level deduplication is often referred to as Single Instance Storage (SIS)[9], check the index back up or archive files need the attributes stored in the file with the comparison. If not the same file, it will store and update the index; Otherwise, the only deposit 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 original file.

B. Block-level data deduplication technology

Block-level data deduplication technology[10][11] to data stream divided into blocks, check the data block, and determine whether it met the same data before the block (usually on the implementation of the hash algorithm for each data block to form a digital signature or unique identifier).. If the block is unique and was written to disk, its identifier is also stored in the index; Otherwise, the only deposit pointer to store the same data block's original location. This method pointer with a small-capacity alternative to the duplication of data blocks, rather than storing duplicate data blocks again, thus saving disk storage space. Hash algorithm used to judge duplicate data, may lead to conflict between the hash error. MD5, SHA-1 hash algorithm, etc. are checked against the data blocks to form a unique code. Although there are potential conflicts and hash data corruption, but were less likely.

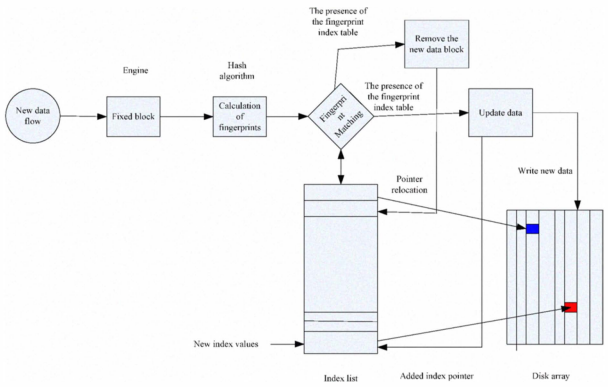


Figure 5. logical structure of data de-duplication strategy

1) Remove the efficiency of file-level technology than the case of block-level technology:

File internal changes, will cause the entire file need to store. PPT and other files may need to change some simple content, such as changing the page to display the new report or the dates, which can lead to re-store the entire document. Block-level data de-duplication technology stores only one version of the paper and the next part of the changes between versions. File-level technology, generally less than 5:1 compression ratio, while the block-level storage technology can compress the data capacity of 20:1 or even 50:1.

2) Remove file-level technology, more efficient than block-level technology scenarios:

File-level data de-duplication technology, the index is very small, the judge repeated the data only takes very little computing time. Therefore, the removal process has little impact on backup performance. Because the index is small, relatively low frequency, document-level processing load required to remove the technology low. Less impact on the recovery time. Remove the technical need to use block-level primary index matching block and the data block pointer to "reassemble" the data block. The file-level technology is a unique document storage and point to the file pointer, so little need to restructure.

C. Byte-level data deduplication

Analysis of data from the byte stream level data de-duplication is another way. The new data stream and have stored more bytes of data stream one by one, to achieve higher accuracy.

With byte-level technology products are usually able to "identify the content," In other words, the supplier of the backup process the data flow implementation of the reverse engineering to learn how to retrieve the file name, file type, date / time stamp and other information[14].

In determining duplicate data, this method can reduce the computational load. Warning? This method usually play a role in the post-processing stage - the backup is complete, the backup data to judge whether the repeat. Therefore, the need to back up the entire disk of data, must have the disk cache, to perform data deduplication process[15]. Moreover, the duplication process may be limited to a backup set of backup data stream, rather than applied to the backup group.

Completed the duplication process, the byte-level technology can recover disk space. In the recovery room before the consistency check should be implemented to ensure that duplicate data after deletion, the original data can still meet the goal. To retain the last full backup, so the recovery process need not rely on reconstructed data, speed up the recovery process.

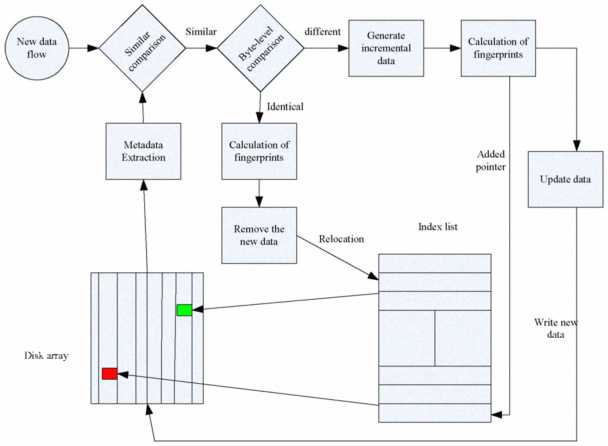


Figure 6. byte-level data de-duplication strategy for logical structure

IV. DATA DEDUPLICATION PROCESS

The basic steps to delete duplicate data consists of five stages[3][8][12][15]:

The first phase of data collection phase, by comparing the old and the new backup data backup, reducing the scope of the data.

Comparing the second phase of the process of identifying data, in bytes, of the data collection phase marks a similar data objects. If the first phase of the work sheet created the need for data identification, then we must use a specific algorithm to determine which data backup group is unique, what data is repeated. If the first phase identified from the meta-data level of data and backup group, the same as the previous backup, then the recognition stage in the data bytes of data will be compared.

The third phase of the data is re-assembled, new data is saved, the previous stage was marked duplicate data is saved data pointer replacement. The end result of this process is to produce a copy of the deleted after the backup group view

The fourth stage will actually remove all the duplicate data before performing a data integrity check efficacy.

Finally remove the redundant storage of data, the release of previously occupied disk space for other uses.

V. IMPLEMENTATIONS

In accordance with the duplication occurred in the location of business systems, implementation can be divided into foreground and background processing two types of treatment.

Foreground processing is done through means of pure software implementation. Software itself is a re-delete function of the backup software, use the Client / Server structure. Server to develop strategies and initiate a backup at the appointed time, scheduling Client on the backup data block, data block input Hash operation procedures, the results included in Hash Table. Block behind the operation result with the same value exists in the table to delete data, different results on the record and save the data block[12][14]. Data recovery and preservation of the value under the table data blocks to restore deleted data, data recovery to the restructuring of the system. Implementations advantage of data output to the network before the hosts in the business to achieve the re-delete operation, to reduce network traffic and reduce the amount of storage space.

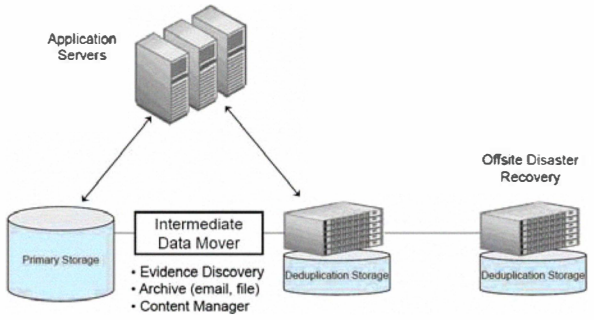


Figure 7. Realization front schematic processing functions

Background processing method used to achieve integrated software and hardware equipment, the overall advantage of the heavy equipment used to delete the CPU and memory, the operation of the business applications are not affected. Deletion in the integrated device in accordance with re-location took place is divided into In-line and Post-processing implemented in two ways.

In-line implementations also based on the Hash algorithm, data written to device memory, delete the buffer for re-operation, variable-length data block partition function of the first scan data, analysis can produce the maximum repetition rate of the split point and then split the data and form variable size data blocks, Hash algorithm based on principle, after re-delete processing[12][14][15].

Post-processing methods used to achieve differential algorithm Hash algorithm technology or technology[12][13]. The biggest difference with other implementations when the data is written, does not deal with being directly saved to the integration of storage devices, and then to re-delete operation. Based on 1 Byte units can scan the maximum found in the duplication of data, provides a maximum deletion ratio, according to dozens of different data types can achieve a ratio of 1 to 1000 than the compression utility. As the latest data is so well preserved that in accordance with the rules and regulations for the need to ensure the authenticity of the user data copy, meet compliance requirements.

VI. Conclusion And Future Work

With the information and network technology, rapid development, rapid increase in the size of the data center, energy consumption in IT spending in the increasing proportion of data deduplication to optimize storage system can greatly reduce the amount of data, thereby reducing energy consumption and reduce heat emissions. Data compression can reduce the number of disks used in the operation to reduce disk energy consumption costs. Remove duplicate data for the large data center information technology system backup system a comprehensive, mature, safe and reliable, More green save the backup data storage technology solutions, has a very high value and great academic value., with very high application value and important academic research value.

Acknowledgment

This work was supported by a grant from the National High Technology Research and Development Program of China (863 Program) (No. 2009AA01A404).

References

[1] McKnight J,Asaro T,Babineau B.Digital archiving:End-User survey and market forecast 2006-2010.2006.http://www.enterprisestrategygroup.com/ESGPublications/ReportDetail.asp?ReportID=591

[2] FalconStor Software, Inc. 2009. Demystifying Data Reduplication: Choosing the Best Solution. http://www.ipexpo.co.uk/content/download/20646/353747/file/DemystifyingDataDedupe_WP.pdf, White Paper, 2009-10-14, 1-4.

[3] Mark W. Storer Kevin Greenan Darrell D. E. Long Ethan L. Miller. 2008. Secure Data Deduplication. StorageSS'08, October 31, 2008, Fairfax, Virginia, USA. 2008, 1-10.

[4] A. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicaterecord detection: A survey. Knowledge and Data Engineering,IEEE Transactions on, 19:1–16, 2007.

[5] Y. Wang and S. Madnick. The Inter-Database Instance Identification Problem in Integrating Autonomous Systems. In Proceedings of the Fifth International Conference on Data Engineering, pages 46–55, Washington, DC, USA, 1989. IEEE Computer Society.

[6] http://www.linux-mag.com/id/7535

[7] Medha Bhadkamkar, Jorge Guerra, Luis Useche, Sam Burnett, Jason Liptak, Raju Rangaswami, and Vagelis Hristidis. BORG: Block-reORGanization for Selfoptimizing Storage Systems. In Proc. of the USENIX File and Storage Technologies, February 2009.

[8] Austin Clements, Irfan Ahmad, Murali Vilayannur, and Jinyuan Li. Decentralized deduplication in san cluster file systems. In Proc. of the USENIX Annual Technical Conference, June 2009.

[9] Bolosky WJ,Corbin S,Goebel D,Douceur JR.Single instance storage in Windows 2000.In:Proc.of the 4th Usenix Windows System Symp.Berkeley:USENIX Association,2000.13-24.

[10] Jorge Guerra, Luis Useche, Medha Bhadkamkar, Ricardo Koller, and Raju Rangaswami. The Case for Active Block Layer Extensions. ACM Operating Systems Review, 42(6), October 2008.

[11] Austin Clements, Irfan Ahmad, Murali Vilayannur, and Jinyuan Li. Decentralized deduplication in san cluster file systems. In Proc. of the USENIX Annual Technical Conference, June 2009.

[12] http://www.snia.org/search?cx=001200299847728093177%3A3rwmjfdm8ae&cof=FORID%3A11&q=data+deduplication&sa=Go#994

[13] http://bbs.chinabyte.com/thread-393434-1-1.html

[14] http://storage.chinaunix.net/stor/c/

[15] Austin Clements, Irfan Ahmad, Murali Vilayannur, and Jinyuan Li. Decentralized deduplication in san cluster file systems. In Proc. of the USENIX Annual Technical Conference, June 2009.