the Availability Digest

www.availabilitydigest.com
@availabilitydig

Data Deduplication in the Cloud

March 2018

Several years ago, February 2011 to be exact, I wrote an extensive article on data deduplication.[1] This was long before the advent of cloud computing that now encompasses so much of our IT activities. Is data deduplication applicable to the cloud? Absolutely!

What Is Data Deduplication?

So what is data deduplication? In simple terms, data deduplication is a method in which a specific block of data in a large database is stored only once. If it appears again, only a pointer to the first occurrence of the block is stored. Since pointers are very small compared to the data blocks, significant reductions in the amount of data stored can be achieved.

The amount of storage savings is very dependent upon the characteristics of the data. For instance, full database backups contain mostly data that is present in earlier backups. Email systems often contain many copies of the same email message and/or attachments that have been sent to several people. These examples can benefit significantly from data deduplication. However, incremental backups contain mostly unique data and will not benefit very much.

Deduplication requires that an initial full copy of the data be present. Thereafter, if further updated copies of that data must be made, only the changes are stored.

Deduplication proceeds as follows. The file to be copied is broken up into *chunks*. A *hash value* is then calculated for each chunk. The hash values for previously stored chunks (including those in the original copy) are stored in a hash table with pointers to the data chunks they represent.

If the hash value for an incoming chunk is found in the hash table, the incoming chunk of data is replaced with a pointer to the previously stored chunk. If the hash value for the incoming chunk is not found in the hash table, the chunk is preserved in the file copy; and its hash value with pointer is stored in the hash table. The result is that a copied file is a series of pointers to pre-existing chunks interspersed with new unique chunks.

As deduplicated files are deleted, chunks that are pointed to by hash keys cannot be deleted. However, when there are no more pointers to a chunk, it may be deleted and its hash value removed from the hash table.

The calculation of hash values is processing-intensive. Therefore, the deduplication of a file requires significant processing resources. Depending upon the implementation, deduplication may be done either

[1] Data Deduplication, *Availability Digest*: February 2011. http://www.availabilitydigest.com/public_articles/0602/deduplication.pdf

as *inline processing* or as *post processing*. An inline-processing implementation deduplicates as the file is being received. Post processing stores the file first and then deduplicates it. Inline processing may slow down the file transfer. Post-processing is faster but requires additional storage to hold the full file before it can be deduplicated.

When a file is to be restored, it is read a chunk at a time and is delivered to the requesting application. Whenever a pointer is encountered, the chunk to which it is pointing is delivered instead. Unlike deduplication, file restoration imposes very little overhead and is almost as fast as full-file reading. Both deduplication and restoration are transparent to the end users and applications.

Note that deduplication is not performed just within a file. The hash table can point to chunks anywhere in the entire database. Therefore, if one file contains data that is identical to that contained in other files, the deduplicated file contains pointers to the common data no matter where that data is found.

Deduplicating Data Stored in the Cloud

Storing an enterprise's data in a cloud is becoming a common practice. Cloud providers such as Amazon with their S3 Simple Storage Services store objects, files, and data blocks in their clouds. The customer pays for the amount of cloud storage used.

The cloud service customer is therefore incentivized to minimize the amount of data storage that he consumes. Deduplicating his data is a powerful approach to achieving this goal.

There are several deduplication methods for data stored in the cloud:

- Inline deduplication duplicates the data in real time as it is stored.
- Post-process deduplication deduplicates data after it is stored.
- Client-side deduplication deduplicates data at the source.
- Target-based deduplication deduplicates data after sending it to the target.
- Network attached storage (NAS) deduplication sends deduplicated data to the target over an IP network.
- Storage area network (SAN) deduplication sends deduplicated data to the target over a Fibre Channel.
- Global deduplication communicates only metadata. The target node asks the sending node for only the deduped objects that it needs.

Let us look at some deduplicating strategies.

Deduplicating Before Storing Data in the Cloud

With this method, the data to be stored in the cloud is deduplicated before transmitting it to the cloud. This has the advantage of minimizing the amount of cloud storage required, the network load imposed by sending the data to the cloud, and the amount of time required to transmit the data.

However, deduplication is a processing-intensive activity. Deduplicating it before sending it to the cloud imposes a heavy processing load on the client system.

Deduplication While Moving Data to the Cloud

While moving data to the cloud, each chunk of data is being handled. It is a simple matter to calculate the hash for that chunk to see if it is already represented in the hash table. If so, the chunk can be replaced with a pointer to its entry in the hash table, with only the hash table pointer being replicated. With this technique, the time required to transmit the data to the cloud may be significantly reduced.

In addition, this technique lends itself to the use of hardware deduplication appliances such as HPE's StoreOnce. Appliances such as these read the raw data from the source system and deduplicate the data on-the-fly as it is being sent to the target system. The use of such an appliance can speed up the deduplication process and remove significant processing load from the source system and the cloud systems.

Deduplication After Storing in the Cloud

Another approach is to move the entire data set to the cloud and then deduplicate the data while it is stored in the cloud. Since cloud systems often can allocate processing resources as needed and then can recover these resources to use for other purposes once they are no longer needed, significant processing power can be focused on the deduplication activity, thus speeding it up.

The cloud system can offer this as a service to the customer. Alternatively, the cloud system may not make the fact that it is deduplicating data known to the customer. It can charge the customer for the totality of data that it is storing prior to deduplication. In this case, deduplication is more valuable to the cloud provider than it is to the customer.

Deduplication Applied to Backups

Deduplication is especially effective when making backup copies of a database. This is because most of the data in a backup is identical to that in the previous backup. Therefore, only changed chunks must be stored. The backup is primarily made up of pointers to previous data chunks in earlier backups. Only the chunks that have changed since the last backup need to be stored.

The shrinking difference in cost between tape and disk is pushing many people to the cloud. Disk storage in general, and the cloud in particular, allow features such as deduplication, replication, linked copies, recovery directly from backups, and recovery of an entire data center in the cloud.

Real-World Examples

Consider an email server that contains 100 instances of the same 1 megabyte sales presentation that was sent to everyone on the sales staff. Without deduplication, the saved emails would consume 100 megabytes of storage. With deduplication, only one copy of the sales presentation needs to be stored, reducing the storage requirement to one megabyte.

Another example is governance. With deduplication techniques, massive amounts of data can be stored in a cloud and made available to management to address compliance and regulatory issues. Additionally, being able to analyze data among a set of users helps IT understand data usage patterns and be able to further optimize data redundancies across users in distributed environments.

Summary

Deduplication is a powerful tool for handling the growing amounts of data with which an organization must contend. It is especially important to apply deduplication to data stored in a cloud, as more and more use is made of the advantages that cloud storage provides. In a cloud, there is almost no limit to the extent of processing resources that can be brought to bear to process large data sets. When the processing is complete, these resources can be released to work on other needs.

Acknowledgements

Information for this article was taken from the following sources:

Data deduplication in the cloud explained, parts one, two, and three, *Computerworld*; August 20, 2013, October 8, 2013, October 31, 2013.
Understanding Data Deduplication – and Why It's Critical for Moving Data to the Cloud, *Druva*; March 24, 2015.
How does deduplication in cloud computing work and is it beneficial?, *Tech Target*; undated.
Why is tape declining in the backup world?, *Network World*; February 23, 2018.