



vaultscape

The Dangers of Data De-duplication

Data Deduplication and Data Compression are not the panacea for storage

● Summary

Data compression is a very well established methodology for reducing the needs for storage. There are several flavors including Lossy (MPEG for example), Lossless (ZIP for example) and more recently Deduplication. Lossy is useful for movies where data size is critical if the entire movie is to be stored on a relatively small device; Lossless is vital where the data must be restored exactly as it was prior to compression. Deduplication is useful where more real time compression / decompression are required with the understanding that data may be lost in the process. People are looking to deduplication for the future of data storage in order to reduce the amount of drives spinning which in turn reduces data center footprint for storage and, of course, reduced power needs. The problem with deduplication is, however, the real potential for data loss.

● Overview of Compression Choices

Lossless Data Compression

Lossless data compression is composed of data compression algorithms that will allow the identical original data to be extracted from the compressed data. Lossless data compression is commonly achieved by an adaptive compression algorithm. Here the compression model is dynamically updated as the data is traversed and compressed. This means the compression and decompression engine have the same initial trivial view of the data which initially results in relatively poor compression. As more data is compressed, however, the algorithm learns about the data and continues to improve the data compression performance. In theory adaptive lossless data compression should be the best lossless compression methodology. Unfortunately, it has a several deficiencies:

- Data that is already compressed data does not compress well, in fact the resulting data can actually be larger than the original data.
- Lossless compression acts best when the file data is comprehended. For example text data will respond best when using a lossless compression algorithm that is designed for text; similarly image data will compress to a greater density if the compression algorithm is designed for image data. Transposing the two compression algorithms will not result in ideal compression.
- The resulting length of compressed data is not known until the compression is completed.



Lossy Data Compression

Lossy data compression is based upon data compression algorithms that when compressing and then decompressing the data will restore data that will most likely be different from the original data, however the resulting data is still accurate enough to be useful. The most common uses for Lossy Data Compression include video and audio files. Here the resulting data in the case of video may not be as sharp or as rich but in a moving image this will not be easily perceptible to the eye. Similarly for audio files, the resulting sound will not have the range or depth of the original but will sound acceptable to the ear. Lossy technologies are commonly used in audio players such as the iPod as well as in cable and satellite TV, DVD and small video players. There are two main types of lossy compression algorithm: a) lossy transform where portions of image or sound are chopped into small segments then each segment is individually compressed; and b) predictive compression where the data prior and/or following the data being compressed is used to predict the compression within the current data being analyzed. Since lossy data is useful for entertainment content and knowing that humans are excellent at filling in missing data, it is ideal for its purpose. For data storage, however, lossy compression is not a viable methodology. Other lossy data compression deficiencies include:

- Data that is already compressed will undergo significant degradation in the subsequent compression. At some point, the compressed data will no longer be useful.
- Much like lossless compression, lossy compression behaves best when the data type is comprehended. Results will be far from ideal if an inappropriate compression algorithm is used on the wrong type of data.

Deduplication

Deduplication is sometimes referred to as single-instance storage or capacity optimization. Deduplication algorithms work in two typical ways, the first by searching for redundancy within large sequences of data across large comparison windows. For example, a file of 32MB could be broken into pieces of 16KB segments, each segment will then be hashed using a cryptographic hash function and then that result compared to a history of other such segments. Where matches are found, the first identified and stored version of a segment is referenced as part of this new file and the newly identified segment is not stored. The second methodology is to deduplicate the entire data file. This uses essentially the same methodology as the first via some form of hashing technique and adds to that some other file analysis such as CRC (Cyclic Redundancy Check) or file length comparison. A unique problem exists with deduplication when the hash matches an existing data segment erroneously – in this case the new data segment is discarded in favor of the existing data segment. The result is the irretrievable loss of the original data from the new segment. There are several methods that can be used to reduce the potential for an erroneous match, for example the entire new matching segment could be compared byte-for-byte with the existing matching segment but this would be costly in time and CPU use.



In a recent study: http://www.swissbib.org/wiki/Deduplication_study the following findings on a deduplication test using 80,000 data records in Library database. This specialized data algorithm accuracy was as follows:

Accuracy of algorithm

It produces approximately:

- 5.3% of false duplicates
- 2.2% of false non-duplicates
- 9.2% of probable duplicates
- 3.9% of probable non-duplicates

If each false duplicate was not subjected to a secondary test (byte-by-byte compare for example), the result would have been the unrecoverable loss of a significant amount of data.

I. Why Deduplicate or Compress at all?

The reason for deduplication and/or compression on storage systems is to save on disk space. In the case of lossy Compression, its need is obvious when memory or bandwidth constraints exist. Since the human brain is adept at “filling in the gaps” the use of lossy compression will probably continue to make sense going forward.

The author has significant history with lossless compression – several years ago he was part of a major push to increase the amount of disk storage in the PC by adding a hardware lossless compression board to the computer and also provided software lossless compression (slower, but utilizing the same algorithms). These solutions were wildly popular for a couple of years – at the end of which Moore’s Law caused hard drive densities to double (which they continue to do every 2.5-3 years) and the need for these lossless compression tools diminished and then vanished. The cost of the new double density drive was identical to that of the now smaller drive when it was the biggest in town. As a result, the need for lossless data compression for storage systems simply went away.

The use of deduplication has too many inherent risks, if the algorithm is working well data loss is still likely and if it is poorly implemented the risk of data loss is significant.

Indeed, if the data is compressed or deduplicated, the process of data analysis will be slower and in the case of a partially corrupted file, it will not be recoverable at all.

In the Cloud storage industry and the VTL systems, data deduplication is the norm. The Cloud storage vendors use it to reduce their cost per GB and generally do not pass on the savings to the user resulting in higher margins. In the case of the VTL vendor, it allows them to increase their margins by selling a bell and whistle.



Vaultscape has taken a different position: we do not compress or deduplicate our customer's data. We store data exactly as delivered to us – we believe this is the only way to ensure 100% data reliability. 100% data reliability is, of course, the Vaultscape SLA.

For a Free Consultation about your archival needs, please contact us at: **(858) 217- 4848** or email info@Vaultscape.com or visit us on the web at <http://www.Vaultscape.com>.

169 Saxony Road, Suite 114
Encinitas, CA 92024
858.217.4848 (Tel)
858.876.0783 (Fax)
info@vaultscape.com