

On CHOW: [**10 winter beers to try now**](#)

- BNET Business Network:
- [BNET](#) |
- [TechRepublic](#) |
- [ZDNet](#)

Deduplication could be the answer to "too much data"

by [Guest Contributor](#) | Apr 18, 2007 7:00:00 AM

Takeaway: Deduplication could significantly streamline your data management operations.

By Mike Karp

I have yet to meet an IT manager who doesn't complain about having to manage too much data. Many, it seems, feel that accumulating data has become something like the weather -- everyone talks about it, but apparently there's just not a whole lot anybody can actually do.

The superabundance of data in corporate data centers results from many things, including:

- Substantial growth in the use of rich media (video on demand movies, for example)
- The digitizing of analog data to make it more rapidly and usefully accessible (making x-rays part of a patient's online medical records)
- The use of mirrored disks, clones and replicated volumes as part of corporate data protection schemes.

There are less valid reasons behind data buildup, like keeping the accounting department's football pool on corporate storage and the fact that stored data, whatever its irrelevance or lack of value, never seems to get discarded. Once things get stored, they tend to stay stored forever, and forever is a very long time. You can

solve some aspects of the problem through technology, such as using data deduplication.

E-mail and file attachments

The most blatant examples of redundant data are multiple copies of file attachments, a problem with which every Exchange administrator is all too familiar. A typical scenario for this might be as follows: an initial e-mail message with a 2MB attachment gets sent to 50 recipients, each of whom, perhaps for only a few days but maybe for many months, saves his own copy of the attachment in the Exchange database. The original 2MB of data now takes up 100 megabytes of storage, and impacts all services that are applied to the exchange database: backups take longer, data retrieval takes longer, and so forth. The second-order effects are more difficult to calculate, but extend well beyond storage and are hardly subtle: Traffic on the network increases as services are applied to the data, resulting in network brownouts that happen because of unplanned-for hot spots that have created constriction points in the data traffic.

The problem doesn't stop with Exchange, of course. Users have a habit of saving multiple instances of the data they create themselves. Users often maintain, for example, multiple iterations of a PowerPoint presentation they are working on, or keep several copies of a document that's in group review in order to capture its history. But is it really necessary to keep all of that data in order to maintain the same amount of information? The answer is no.

Data deduplication solves some of the problem

Data deduplication techniques ensure that only one instance of each significant piece of information is kept on the system; every other instance -- even gigabyte-sized objects -- will be replaced by a pointer to the initial instance. Using deduplication, an Exchange environment that needed 100MB of storage to accommodate 50 separate instances of a 2MB file would now only need 2MB of storage for the file itself, plus several additional bytes of storage to accommodate each of the file pointers. In an Exchange environment with 1,000 mailboxes, the potential for saving disk space is enormous.

What is equally interesting is that you can do many vendors' data deduplication at the "sub-file" level. Products deduplicating at the sub-file level can identify identical

"chunks" (that is, byte aggregations) of data. Once this "byte-level differencing" identifies those chunks, it replaces the byte strings with pointers. This is particularly useful during backups because less data has to be sent to the backup device. But it really proves its value during recoveries where it delivers substantial performance enhancements.

There are, of course, several methods you can use to implement, not only byte-level differencing, but every other aspect of data deduplication. **Data Domain**, **EMC Avamar**, **Falconstor**, and **Quantum** do deduplication at the block level, while **ExaGrid**, **Diligent**, and **Septon** do it at the byte level. Some people dedupe the original data, installing agents on servers to accomplish this, while others prefer to leave the original data alone and do their deduplication on a virtual tape or other secondary storage device. When you keep data at remote or branch offices, deduplication is less of a storage issue but more of a concern for network administrators. In fact, some of the earliest deduplication technology comes out of the wide area files services (WAFS) segment, where net admins emphasized reducing redundant data so as to improve bandwidth utilization of expensive WAN assets.

The value of deduping

Just about every form of data deduplication delivers value by reducing the total amount of data. Vendors claim data reductions of anywhere from 30 to 300%, and such claims are likely justified. Whether or not they are significant is another question. The significance depends on the vendor's product as well as on the environment to which the technologies are being applied. Not all data lends itself well to deduplication at the byte level -- MRIs and digitized x-rays clearly fall within this category.

Selecting a technology

I would suggest you use six parameters as a guideline in selecting any deduplication technology:

- Performance (different measures will apply depending on where the deduplication takes place)
- Capacity
- Scalability

- Deduplication ratio (only useful in making like-to-like comparisons, such as comparing VTLs to one another)
- Data types (what will be deduped)
- Data location (data center or remote location).

If you take these criteria into consideration, it is a pretty good bet that you can find a product that will significantly streamline your data management operations.

Mike Karp is a senior analyst with Boulder, Colo.-based Enterprise Management Associates (www.emausa.com), an industry research firm focused on IT management. Mike can be reached at mkarp@enterprisemanagement.com

Copyright © 2008 CNET Networks, Inc. All Rights Reserved. [Privacy Policy](#) | [Terms of Use](#)