FederalComputerWeek

How many copies do you need?

- By Karen D. Schwartz
- Aug 06, 2007

Data deduplication wipes redundant data so you can reclaim storage and backup time

Like many government organizations, the Army's Information Management Support Center (IMCEN) stores hundreds of terabytes of data. A big challenge of owning a data collection that large is finding enough time after work hours to make a backup copy of it, not to mention the cost of secondary storage systems to house it all.

With newly created data piling up all the time and adding to the problem, IMCEN officials finally reached a tipping point last year.

The answer for IMCEN was data deduplication — a way of reducing storage needs by eliminating redundant data. With this method, storage systems retain only changed blocks of data, instead of automatically backing up everything, including duplicate and unchanged files, as traditional backup solutions do. The process indexes files, identifies duplicates and eliminates them based on factors such as having the same name, the same modified dates or the exact same file size.

IMCEN officials replaced the disk-to-disk backup and recovery system they had been using with a deduplication system from Data Domain. The move enabled the organization to increase performance dramatically and free storage space.

"Now we can push these enormous backups down to one box, and the depth of the recovery data increases," said Bob Dixon, senior architect at IMCEN on contract from NetCentrics. "Instead of just going back two or three [nightly] backups, you can go back two or three months' worth of backups. It's a game-changing technology."

The primary function of deduplication is to make data management more efficient, reduce the amount of space needed to retain data and improve storage economics. Most deduplication vendors position their solutions so that they can work with an organization's existing backup software and secondary storage systems, whether they use disks, magnetic tape or disks that emulate tape. The latter are sold as virtual tape libraries (VTL).

Both EMC's Avamar and Data Domain apply deduplication to both file storage and VTL interfaces, while Sepaton and Diligent Technologies focus on the VTL interface.

Network Appliance has taken another route. In addition to offering deduplication on secondary storage, it is among the first to offer it on primary storage. That's a new development, but applying deduplication technology to primary storage is the wave of the future, said Arun Taneja, president at consultant Taneja Group.

Deduplication works by examining the data stream as it heads towards the storage media, checking for blocks of data that are identical and eliminating redundant copies. That method all but ensures that important data won't be

permanently deleted accidentally, said Dale Wickizer, chief technology officer at NetApp's federal systems division.

Deduplication is attracting interest among organizations grappling with growing volumes of e-mail messages, documents and other electronic files. Many government agencies certainly fit that bill and, like their private-sector counterparts, can expect the amount of data they manage to double every 18 to 24 months, if not faster, industry experts say.

"Government agencies also have multiple backup and archive requirements that are mandated in many cases, [such as] off-premise replication of data files for disaster recovery and encryption of sensitive data for storage purposes," said Rick Marcotte, president of DLT Solutions, a solution provider focusing on government. "To the extent that deduplication can be deployed, it can generate a nice savings in terms of operational costs, leading to more efficient use of taxpayer dollars."

Agencies at all levels of government are starting to discover the benefits of deduplication. For example, the Sacramento County, Calif., Department of Human Assistance uses a NetApp solution that includes deduplication and an integrated disk-to-disk solution from Veritas — now part of Symantec — to manage its data growth. With deduplication, the department was able to reduce storage requirements by 80 percent while performing backups 16 times faster and getting systems back on line six times faster, said Keith Scott, an information technology analyst in the department.

The Energy Department's Western Area Power Administration, which markets hydroelectric power from dams throughout the western United States, is another recent convert to deduplication.

"We had moved to disk-based backup some time ago and used a storage-area network, but it was difficult to find enough space to store all of our data," said John Pyle, an IT specialist at the Sacramento-based organization. "We would back it up to disk, stage it to tape and then wipe out the disk, all on a daily basis. The process took a lot of time, and we didn't have enough space to save it and keep it on there."

The organization made the switch to deduplication and reduced the amount of data it backs up by a factor of nearly 14 to 1.

In terms of space savings, "deduplication has far exceeded my expectations," Pyle said. "I expected it to handle maybe six months' worth of data, but it handles a year's worth.

"And now I don't have to take it off disk," Pyle said. "I can leave it on disk and copy it to two tapes: one that goes off-site and one that stays in the library. That way if I need to do a restore, and do a restore from disk, and if that were to fail me, I can go to a tape that's on-site."

Deduplication benefits

As Pyle discovered, the benefits of deduplication can be significant. The biggest perk by far is space savings. Organizations using deduplication can expect to reduce their storage capacity requirements by a factor of at least 20 to 1. IMCEN experienced a 23-to-1 reduction rate during its testing phase, a number Dixon said exceeded even Data Domain's promises.

Some industry officials speak about the technology's ripple effect. "Deduplication has a multiplicative effect through the entire data life cycle process, so you save space on the amount of data you have to back up, save space on your backup systems and save on any archive applications you have," said Patrick Rogers, vice president of solutions at NetApp's federal systems division.

And if you're saving space, you're freeing up other resources, too. "You'll save power and cooling, real estate and management time," Rogers said. "Typically organizations have a certain number of administrators per terabyte of

storage. If you can save space, it increases the productivity of everybody associated with managing your storage infrastructure."

Another benefit of deduplication is the ability to retain more data for longer periods of time. The technology can help organizations meet retention requirements without a need to continually add storage capacity to keep up with additional data. That's particularly true in the government arena, which has strict retention policies, particularly for e-mail messages, said Beth White, vice president of marketing at Data Domain.

Performance and backup speed are other benefits. If you transfer less data to a backup device, you speed the entire process. That's particularly important when nightly backups take more hours than are available. Using deduplication, organizations can reduce the volume of daily or nightly backups by as much as 90 percent, proponents say.

Finally, storage economics can greatly improve with deduplication. Although the technology isn't cheap — it starts at about $19,000 for a half-terabyte of data and runs into the hundreds of thousands of dollars — it can pay for itself.

"The economics are directly related to the amount of capacity required," White said. "When you compare it to tape, the capital expenditure to purchase an appliance like ours might seem similar price-wise, but over time system maintenance and the disaster recovery process of taking tapes off-site, storing them and recovering data are costly."

In a market driven by those advantages, experts predict significant growth in the use of the technology. Taneja Group, for example, predicts that data deduplication for secondary disk storage alone will grow from $87 million in 2006 to $1.62 billion in 2010.

"Data deduplication will become an integral part of both primary and secondary storage solutions within 24 months," Taneja said. "Look at it this way: If I can allow you to keep a certain amount of information with complete integrity and guaranteed access and use 1/20th the amount of storage, why would I ever buy anything that doesn't have that feature?"

That kind of return on investment is why organizations that already have deployed data deduplication are continuing to think of ways to gain more benefits.

IMCEN officials, for example, are considering depositing an archive of several years' worth of e-mail messages into a repository that they can later access and index.

"Then it will get deduplicated along with the rest of the [current] e-mail that gets backed up," Dixon said. "The deduplication capability offers benefits we didn't even think about when we started."

Upcoming advances in the technology are expected to increase the deduplication ratios. "There is still a lot of innovation left," Taneja said. "We are babes in the woods right now."

Schwartz is a Washington writer specializing in business and technology issues. She can be reached at karendschwartz@gmail.com.

Quick studyWhat it is: Sold as software or integrated with hardware, deduplication solutions analyze data before it is stored and retain only changed blocks of data on storage media, greatly reducing the amount of disk or tape needed to store data.

Benefits: Frees disk capacity and saves data center space, increases retention time for storage, improves storage economics, reduces amount of real estate and power needed, and decreases management requirements.

Cost: Entry-level systems start around $20,000, but systems can cost hundreds of thousands of dollars,

depending on the amount of data being backed up, the level of redundancy and the length of time the data must be kept accessible.

— Karen D. Schwartz