Demystifying de-duplication

http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9011622

By Ray Lucchesi

Source de-duplication is set to be "more disruptive" than previous technologies

February 21, 2007 (Computerworld) -- Data de-duplication technology has emerged as a key technology in the effort to reduce the amount of data backed up on a daily basis, which in many enterprises is growing at more than 100% every year.

For example, John Thomas, IT manager at Atlanta-based law firm Troutman Sanders LLP, was able to use data de-duplication technology to reduce the amount of data streamed from more than a dozen remote offices and thereby cut his backup window from 11 hours to 50 minutes. Thomas says his compression ratio for his backups run as high as 55:1.

Vendors have taken different approaches to the technology resulting in multiple distinct products that users should become familiar with in order to choose the flavor that best suits their environments.

Data de-duplication uses commonality factoring to reduce the amount of data either at the backup server or at the target storage device. As a result of enormous compression ratios achieved by data de-duplication technology, disk is becoming more attractive as a viable, online alternative to traditional tape-based backup. For example, people working at remote and branch offices need instant access to all the data and applications available at their company's headquarters. So IT shops typically set up remote mini data centers, with application servers, block and file data storage, backup tape and report printers, sacrificing administrative control. By utilizing data de-dupe technology, backups can be performed over the WAN using spare nighttime bandwidth, eliminating the need for tape at remote sites.

Greg Schulz, senior analyst at The StorageIO Group, says de-duplication technology mainly resides in the backup space, complementing traditional tape libraries with the purpose of lowering costs and reducing data.

The main benefit to de-dupe technology is that you're not seeing your virtual tape library fill up, and you're "not seeing your backup targets fill up as fast as it normally would," Schulz says.

De-dupe can be done at the target of a backup stream (the storage array or tape drive) or at the source of the data being backed up (the application server). Traditionally, de-dupe products had been used as a target for backup data, but Schulz says there is "a growing emphasis de-duping back on the server."

Target de-dupe products are generally used as part of a final repository for backup data. Most backup software today supports tape volumes, files or raw disk as targets. Target de-dupe products mimic a tape library and support virtual tape libraries (VTL), or they can act as a network-attached storage file server supporting network file system (NFS) or Common Internet File System files. Target de-dupe technology can also work on raw disk supporting Internet SCSI or Fibre Channel logical unit numbers (LUN). Prominent target based de-dupe products are sold by Data Domain Inc., Diligent Technologies Corp., ExaGrid Systems Inc., FalconStor Software Inc., Quantum Corp., and Sepaton Inc.

Today, de-duping data at the target is the leading method, but de-duping data at the source where the data is coming from is even more disruptive, and the benefits are far greater, Schulz says.

Source de-dupe products replace backup software used in a client/server configuration, where remote clients de-dupe data being backed up and only transmit unique data to the central server. This reduces bandwidth requirements considerably, according to Schulz. Some prominent source-based de-dupe products include Asigra Inc.'s Televaulting for Enterprises, EMC Corp.'s Avamar, Network Appliance Inc.'s SnapVault, and Symantic Corp.'s NetBackup Pure Disk.

In band or out of band

Another characteristic used to discriminate target de-dupe products is when data de-duplication processing occurs. Data de-duplication takes time to compute and find commonality in the data being backed up. To minimize the effect on backup performance, some vendors de-dupe data in the background. These de-dupe products buffer the backup stream to disk and then after the fact reduce its size via de-duplication. ExaGrid, FalconStor and Quantum provide target de-dupe products that do background data de-duplication

Other products can handle the backup stream and de-dupe in band, in real time. Target vendors that de-dupe in band include Data Domain and Diligent. All source vendors de-dupe in band as well. Paradoxically, the in-band vendors are able to sustain full backup stream performance.

The unit of de-dupe granularity, called chunk size, further differentiates de-dupe products. Only NetApp touts a fixed chunk size equal to data block size, according to Schulz. Most de-dupe products claim variable chunk size from the file level down to sub-block level. By using variable chunk size for data inserted into a file, you need only show the changed data as being different, and the rest of the file would be the same. Even more impressively, most de-dupe products can not only reduce data from generations of the same file but also eliminate data copies across files.

Yet another de-dupe difference is file-type sensitivity. Some target products

open the backup stream to determine data type and invoke file-type-specific policies to provide better de-duplication. Sepaton claims to have the most file-specific policies. Other target vendors claim that their products perform file-specific de-dupe to a lesser extent, including Data Domain, Diligent and Quantum, which all claim they are completely agnostic regarding backup streaming.

The numbers for de-dupe compression rates range anywhere from 3:1 to 500:1. De-dupe products can sustain these high compression rates because backups generate duplicate data every time a full backup is run. Moreover, beneath the file level, most data is not unique even though a file is modified. But, some data does not de-dupe well, including audio, photo, movie and other media files that simply don't have excess white space or duplicate data.

Troutman Sanders' Thomas manages 18 Data Domain boxes throughout the U.S. They are being used as NAS targets for Symantec Backup Exec. The Data Domain boxes have a capacity of 1TB to 15TB. But, each of Troutman Sanders' NAS boxes backs up the equivilant of 400TB to 500TB of data.

The Data Domain arrays have been in production for more than eight months. Each office does a local backup to its on-site Data Domain box. That data is replicated off-site to headquarters in Atlanta and then again replicated to the law firm's secondary site. Before Data Domain, Thomas says his company had DLT-4 tape changers and LTO-1 tape. Backups at some offices used to take more than 11 hours, but now with Data Domain, the backups are under 50 minutes. Troutman Sanders backs up one site with 163 TB of data onto 4.6TB of Data Domain for a compression ratio of 34.6:1. Smaller sites have experienced a compression ratio as high as 55:1.

"The [Data Domain] hardware was comparable to buying all-new tape hardware, but the speed and the ability to not have to manage the tape is what really got us … the replication is an added bonus," Thomas says.

Grahame McKenzie, manager of IT at Crawford Adjusters Canada Inc. in Hamilton, Ontario, had performed local backups at each of the company's 80 remote sites. He now manages 80 remote and 20 local Asigra clients backing up to a single 2TB Asigra TeleVault. They mirror the central TeleVault to an off-site hardened data center and copy their monthly full backups off onto tape moving these off-site to another location.

McKenzie says that before, "we weren't even doing it [backups] on the smaller XP share points -- there was no backup. It just wasn't economical; if it was an NT server, it would [be] backed up to tape." This process was time-consuming and not subject to easy validation. He says that "with Asigra, backups all take place, and they all come here, … and I get a report on everything the next day." McKenzie says that in some cases, it was difficult to quantify a return on his investment because no backups were being performed, but he did say "we don't invest in tape on our new branch servers, and we have one tape library

here."

Schulz says data de-duplication allows users to retire most of their tape infrastructure at remote and local sites. In some environments, it may not be possible to eliminate all tape processing, but it can be reduced considerably. At essentially the purchase price of replacing tape, you can get all the benefits of tape with the convenience of disk -- mainly quicker, less error-prone and less operator-intensive backups and restores.

De-dupe product comparison table

Products	Purchase	Inline or Offline	Chunk Size	Person-ality	File type special processing	De-duplication performed
Asigra	Bundled appliance	Inline	Variable	N/A	N/A	Source
Data Domain	Bundled appliance	Inline	Variable	NAS & VTL	None	Target
Diligent	Software license	Inline	Variable	VTL	None	Target
ExaGrid	Bundled appliance	Offline	Variable	NAS	CA's ArcServe, EMC's Networker, Symantec's NetBackup, Symantec's Backup Exec, CommVault's Galaxy, IBM's TSM	Target
EMC Avamar	Software license	Inline	Variable	N/A	N/A	Source
FalconStor	Bundled appliance	Offline	Variable	VTL	EMC's Networker, Symantec's NetBackup, IBM's TSM	Target
Network Appliance SnapVault	Bundled appliance	Inline	Fixed	N/A	N/A	Source
Quantum DXI	Bundled appliance	Offline	Variable	NAS & VTL	None	Target
Symantec NetBackup Pure Disk	Software license	Inline	Variable	N/A	N/A	Source
Sepaton	Bundled appliance	Offline	Variable	VTL	Symantec's NetBackup, IBM's TSM, HP's Data Protector	Target

About the author:

Ray Lucchesi *is president of Silverton Consulting and since 2004 has helped startups to Fortune 500 companies improve storage product development and marketing. Ray has also helped end-users better select and use storage subsystems.*

mailto:info@silvertonconsulting.com
http://www.silvertonconsulting.com