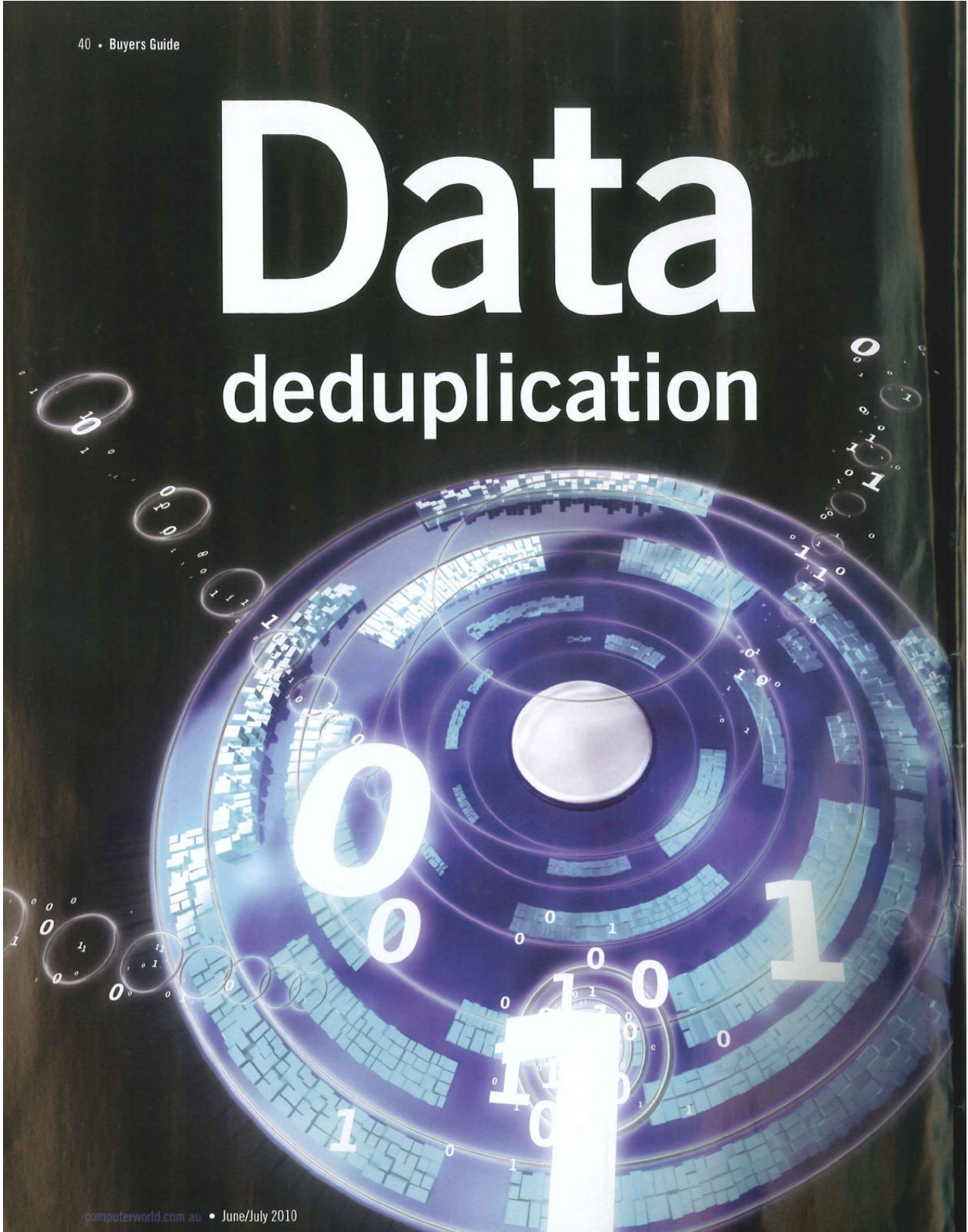


BCAP	June/July 2010	Circulation: 9,783	Page: 40
------	----------------	--------------------	----------

40 • Buyers Guide

Data deduplication



Exponential data growth is emerging as arguably the biggest issue IT managers face. But, handily, data deduplication is one available tool to help tackle the problem. **Tim Lohman** reports.

You wouldn't call TechnologyOne your average Australian SME but the software developer is in many ways very typical when it comes to the issue of managing rampant data growth.

The company, with around 700 staff and an information growth rate of 30 per cent — in a quiet year — was until recently trapped in a cycle of throwing ever more disk at the problem in the hopes of keeping near-exponential data multiplication under control.

The company's backup window — the period of time when backups are permitted to run on a system — was also starting to grow by about half a day per year, meaning the organisation was in danger of breaching its own backup timeframe requirements.

However, as IT manager, Andrew Bauer, explains, the company has joined a growing number of organisations turning to data deduplication technology (in this instance, NetApp 'powered' by Commvault) as its preferred method of coming to grips with one of the great IT challenges of the moment.

As Bauer tells it, while the company had been on a consolidation path for some time — adopting virtualisation and moving to blade servers as a way to reduce cooling and power requirements — it had struggled with managing its storage.

"Disk hasn't gotten smaller physically and its power and cooling haven't gotten smaller, so there hasn't been much we could do about consolidating storage," he says.

"So, when dedupe became available we saw that as a real way to reduce our requirements by avoiding more storage purchases, going forward, and making better use of the storage we have now."

Data deduplication 101

Gartner defines deduplication — also referred to as 'dedupe' — as technology which uses identification and comparison algorithms in selecting data so that only unique "chunks" are stored, thus eliminating redundancy and reducing space requirements. Reduction rates, depending on the type of data and method used, can be anywhere between three and 25 times, sometimes even more.

Dedupe typically comes in two forms: Pre-processing or post-processing. In pre-processing, also

Data deduplication products

Hitachi Data Protection Suite: Includes Data deduplication which provides users the choice of when and where to use deduplication — remote sites, local file and databases, or in reducing an organisation's data footprint when transferred to tape. The wider suite incorporates backup and recovery, point-in-time replication, archiving, and storage resource management under a single graphical user interface. tinyurl.com/qt6wlz

Double-Take Atom Deduplication: A file-level deduplication feature of the Double-Take Backup, Atom Deduplication scans the Double-Take Backup repository in real time to locate duplicate files among all the data. Replicates and stores only byte-level changes to the protected data, minimising any duplication caused by the backup solution itself. tinyurl.com/368k2zv

EMC Avamar: Enables data reduction and secure backup for VMware environments, remote offices, LAN/NAS servers, and desktop/laptop systems, and reduces backup time, growth of secondary storage, and network utilisation. EMC claims to reduce daily backup data up to 500x, backup times up to 10x, and total storage up to 50x. tinyurl.com/bkgokp

Quantum DXi series: Hardware-based DXi-series claims an average 125 per cent increase in backup performance, 87 per cent fewer failed backup jobs, and typical disk capacity reductions of 90 per cent or more and 95 per cent in virtual environments. tinyurl.com/2vmoyog

NetApp Deduplication: Part of NetApp's ONTAP architecture, NetApp deduplication can be used across primary data, backup data, and archival data. Users can schedule deduplication to occur during off-peak times, select which datasets to deduplicate, and perform a full byte-for-byte validation before removing any duplicate data. tinyurl.com/mjprw8

Symantec Backup Exec 2010: Integrated data deduplication allowing dedupe at the source or remote server, at the media server, deduplication appliance-level or on data travelling from remote offices to headquarters. tinyurl.com/6hx22w

known as in-line, the dedupe technology sits between the server and the storage media deduplicating data as it travels from the server to the storage. In post-processing, data is allowed to go straight from the server to the storage during the day and is run against the stored data at a later time — typically overnight.

Dedupe solutions can also operate at the file, block or bit level. With file level deduplication, only a single instance of a 5MB PowerPoint presentation emailed to 10 people in the same organisation would be saved. In this way, what would have been a 50MB storage requirement is cut down.

Block and bit level deduplication are useful for their ability to ignore file types and examine changes which occur within data at the file or block level. If the same 5MB PowerPoint file mentioned above has 10 different drafts, then 10 different files roughly equalling 50MB won't be saved. Instead, only the original 5MB file plus the bit or block-level data relating to specific changes made to that original file are kept, cutting down on the amount of data stored.

Dedupe benefits

On top of postponing the point at which an organisation must invest in more storage, dedupe can increase data availability, cut costs associated with power and cooling, along with helping to increase network bandwidth through lower data throughput.

In the case of Curtin University of Technology, it also dramatically improved the organisation's disaster recovery ability. As CIO, Peter Nikolettatos, explains, initiatives around digitising lectures and server virtualisation have contributed to exponential growth in storage demand, resulting in issues around incomplete cycles and decreased reliability in its tape-based backups.

"It was apparent that we were backing up data that was clearly duplicated," he says. "The introduction of [dedupe] has resulted in considerable deduplication as well as removing the need to use tapes and increasing the ability to recover data from days to minutes."

According to Nikolettatos, moving to a dedupe solution — in Curtin's case EMC's Avamar product — has seen a doubling in the number of virtual machines being backed up and an average deduplication rate of close to

99 per cent. The University's backup window has been reduced from 24 hours to under three. As an example of the technology's power, Nikolettatos says a Microsoft Exchange email server was recently recovered from disk in just 30 minutes. Recovering from tape would have taken between one to two days.

Assessing a Dedupe Solution

So, it's clear that dedupe can be a pretty powerful technology, but given the breadth of offerings in the marketplace — either as part of a wider suite of information management tools or as a standalone solution — consideration needs to be given to a few things before deciding on a solution.

The first step in this process, Forrester analyst, Tim Sheedy, advises, is to take a step back and ask whether the dedupe solution you are acquiring is to address a single point problem or whether there is a broader organisational requirement for dedupe.

"Is it a single requirement to get rid of excess customer records or do you need to do broader data quality, cleansing or enhancement, as that's when you'll begin looking at an external third party provider as opposed to buying a piece of software and managing it yourself," Sheedy says.

Assessing whether you need to do the deduping in-line or after-the-fact will also drive you to a different set of vendors, as will assessing how well a potential vendor sits within your broader architecture. As mentioned above, you'll also want to consider whether a broader suite of information tools, rather than a stand-alone data dedupe solution is appropriate.

"The problem with the whole information management landscape over the past few years is that you have a single piece of software for data quality, another for dedupe, another for data loading and ETL [extract, transform, and load] so that you end up managing 15 or 20 vendors and you have to integrate it all yourself or pay a lot of money to an IBM, CSC or Accenture to come in and do it for you," Sheedy says.

"The advantage of the big companies is that they integrate all this themselves so you know your BI and ETL and other applications will work together."

Another crucial consideration is to determine whether post- or pre-processing

"The advantage of the big companies is that they integrate all this themselves so you know your BI and ETL and other applications will work together."

TIM SHEEDY

dedupe is right for you. In the post-processing camp, TechnologyOne's Bauer says his organisation decided on its particular solution for reasons of speed and effectiveness.

"We found [post-processing] dedupe generates better dedupe than the in-line as it has more time and available resources to process the deduping and assess the blocks against the dedupe pattern it has already created," he says.

"The advantage of block-level dedupe is that, while we initially used it against CIFS [Common Internet File System] data, we can also use it against block level storage LUNs [Logical Unit Number] that have virtual machines mounted against them or raw disk served up out of the SAN [Storage Area Network], we can dedupe that as well."

The results, Bauer says, were a 50 per cent-plus reduction on file server data, and in some instances, up to 90 per cent reduction on source code repository data.

In the pre-processing camp, Brad Jansons, IT Manager at Australian Motoring Services (AMS), argues that in-line — in its case the Hitachi Data Protection Suite from Commvault — was the way to reduce storage costs, store more data more efficiently, and decrease backup and restore times.

"In-line has the advantage over post-processing, as with the latter you have to have the space available there in the first place — to store the excess data — which partly

defeats the purpose of dedupe," he says. "As the dedupe is running to the local SAN there's no issue with performance and the amount of data we are deduping means that in-line-related performance issues aren't an issue. We are also not a 24/7 business so backups are done outside of hours."

With its in-line solution, AMS has seen up to 50 per cent reduction in disk volumes and has been able to move to disk, instead of tape, for its backup and recovery. Moving to dedupe also saved the organisation from investing in another tray of high performance disk for its SAN.

Lastly, in devising a business case, Gartner advises that it is also worth considering whether the bandwidth savings created by the use of dedupe make disaster recovery, centralised backup of remote offices, and application deployment methodologies operationally and economically viable.

Dedupe Tricks and Tips

When it comes to dedupe, one of the more important tips is to understand that not all data can be deduplicated.

Compressed data, such as pictures, voice, and video files, and encrypted content are examples which dedupe doesn't necessarily lend itself to. However, dedupe can be an effective data reduction method for data sets that contain common bit strings. Clearly it's important to choose a dedupe solution that is right for your business and needs.

But another tip is to keep an eye out for the creation of hot spots in your storage environment. As TechnologyOne's Bauer explains, dedupe effectively takes a large amount of data and reduces it down to a smaller amount stored on a reduced number of disks. As a result, the increased number of data requests against a smaller number of disks could lead to the I/O capability of those disks being overloaded.

"You will have an end-user impact and they will find it is slower to access data from that set of disks," Bauer says. "Instead of having 100 users accessing files across 100 disks, you will have 100 users accessing files on one disk."

The software developer managed the issue by making sure that deduped data was separated out across a number of higher performance disks.

"If we knew there would be a higher I/O load then we made sure it would be on higher performance disk," Bauer says. "What it means is that you can have a small number of high performance disks for high-load data, and cheaper lower performance disk for the low-load data. All up you end with less disks and less cost." **CW**

