SearchStorage.co.UK E-Guide

How to Get Started With Data Deduplication

Data deduplication has certainly generated quite a buzz among storage professionals in the UK, and while there's a lot of curiosity and interest, many UK storage professionals have yet to take off with this increasingly popular backup trend. In this guide, you'll gain insight on the benefits of deduplication, as well as best practices to follow as you prepare to launch, or enhance, your data deduplication strategy.

TechTarget
The IT Media ROI Experts

SearchStorage.co.UK E-Guide

How to Get Started With Data Deduplication

Table of Contents:

So Far, Backup is Killer App for Data Deduplication

October 6, 2008 | Antony Adshead, UK bureau chief, SearchStorage.co.UK

More and more storage professionals are looking to implement disk-based data deduplication, driven chiefly by the need to keep backups within windows. Increasingly, the inability to stream to tape quickly enough is leaving businesses exposed or is eroding network bandwidth as data protection creeps further into the working day.

However, data deduplication has still to make real headway among the vast bulk of the user community. It has made a lot of headlines, but IT departments are bound by budgetary buying cycles and do not rush new technology implementations without serious forethought of the implications.

Consequently, most UK storage and backup professionals are at a look-and-evaluate stage where they are coming to grips with what data deduplication can do and to explore differences between data deduplication products. That's the assessment of Tony Lock, programmer director with analysts Freeform Dynamics.

According to Lock, businesses that have implemented data deduplication already are not confined to any particular sector. "It depends on need," Lock says, "although clearly some verticals, such as financial services, have the resources to begin early research into the technology. In general the UK market lags a little behind the US as that is a market more constrained by external factors such as compliance and, because most of the vendors are there, they start selling there sooner."

The convenience of writes and reads to disk and the power of data reduction offered by deduplication mean that UK businesses can now remove tape from the equation as nearline storage, says Clive Longbottom, service director, business process analysis, with analysts Quocirca.

"The main reason for the increasing rejection of tape is the low cost of disk and the use of virtualization, making logical tape based on disk a far better bet," says Longbottom. "Backing up to disk rather than tape is far faster, and so can manage far greater volumes in less time. Combine it with dedupe, and all of a sudden, backups that were taking more than a day now take an hour or so."

Backup is data deduplication's killer app. Because data deduplication can bring data reduction ratios of 10 times, 20 times, 30 times or more—depending on the types of data being processed—it can speed backup to disk hugely, with the added bonus of allowing far more information to be kept on disk and so shortening recovery time and stretching recovery point further back in time.

That was the experience of surfwear manufacturer O'Neill Europe, which cut backup times from 14 hours to two, slashed restore times, extended on-site retention to a year and made possible full backups of data where previously it had to select the most important. It made these gains by implementing a disk-to-disk-to-tape strategy with one Data Domain DD565 at the firm's headquarters plus four DD510s at other sites across Europe. O'Neill is backing up 57 TB of data to just 3 TB of disk space and getting a dedupe ratio of 18:1.

Peter Maljaars, global IT service and infrastructure manager at O'Neill Europe says, "We're actually now backing up more data than we did before—5.3 TB instead of 1.4 TB. Before, we just couldn't back everything up—it would have taken 20 or 30 hours if we included all the image files, which are important but we can do without. We used to have to decide what was most important. Now we can back everything up and don't have to make that decision."

Besides the core benefits of reducing backup times and enhancing RPOs and RTOs, data deduplication also reduces dependence on tape. This not only means that less has to be spent on tape equipment and media but also removes the need for the human hand in management of tapes. This is a common source of data retention anguish—we've all heard scary stories about the security guard, admin person or salesperson at the branch office forgetting to take the tape out or being off sick.

Because the amount of data that has to be moved is drastically reduced by data deduplication this also lowers the potential load on the WAN link, meaning opportunities for off-site backup, replication and disaster recovery come into view.

But can you benefit from data deduplication? To get an answer to that question it pays to look under the bonnet and see how data deduplication works. Results vary a lot depending on your environment and the product used.

Essentially, data deduplication is a form of data reduction, with duplicate blocks of data removed and replaced with a pointer to the original instance generated by a mathematical algorithm during inspection of files and their component parts. For that reason data deduplication is, at its most basic level, suited to data types that contain a lot of repeated patterns, such as database files and email data.

Conversely, data formats that contain little in the way of repeated information offer little scope for impressive deduplication ratios because there is not much redundancy to be eliminated. Ratios achievable vary from 10 times to 50 times or more, dependent on the data being processed.

Deduplication achieves higher ratios over time. The technology relies on being able to spot repeated patterns, so—assuming some homogeneity of data type—it will shrink backups far more after several weeks than it will after only a few days. Over a longer period it simply doesn't have the opportunity to point to already existing identical patterns of zeros and ones. So, if your data types don't vary too much, you will achieve good ratios. If not, then it may not be for you. With data deduplication it's a case of your-mileage-may-vary, Freeform Dynamics' Lock says.

"It all comes down to the state of the data you have," Lock says. "Some organisations will have lots of duplicate data – 47 instances of one PowerPoint presentation emailed to 47 team members, for example—and deduplication can cut this to one with 46 stubs pointing to the original. The length of time over which data is deduplicated will also bring variations in results, especially as deduplicated data will be able to be kept for longer on disk and so will make the process of eliminating redundancy more efficient."

If your data profile means the technology is something you can profitably use, there are then further questions to ask. These are mainly about what type of deduplication product is best for you.

For now, nearly all data deduplication products are either backup applications, virtual tape libraries or NAS boxes.

These are also isolated and limited examples of deduplication in primary storage. Most experts think that primary storage for high-transactional data simply cannot stand the invasive processing load that deduplication entails.

Data deduplication products come as hardware appliances and as software. While software data deduplication products are less expensive than data dedupe hardware appliances, they put a heavier load on host CPUs and are more difficult to maintain, given their likely multiple dependencies on other software, such as backup applications, as well as changes in their immediate environment. Software products are generally located at the backup server, which means they can cut bandwidth requirements.

Another key product differentiator is between inline and post-process (or out-of-band) data deduplication. The former processes data as it passes through the dedupe product before it reaches the target disk—so can cut down on bandwidth used by backup traffic—while the latter takes in the data and then deduplicates it, and so requires more disk.

"It's a simple choice of speed versus space," says Chris Reid, managing consultant with integrator Morse. "Inline deduplication requires the least space because excess data is stripped away as it arrives on the system, but this is a slower process, which impacts backup windows. Post process deduplication is faster but requires much more space. By taking a disk backup and then deduplicating the data it places less strain on the system during the task and also ensures that there is at least one full copy of the last backup on disk, which can help with restoration."

Testing data deduplication products is vital. Since deduplication products are affected by data type and environment, realistic testing of the types of backups and restores that your environment is likely to bring is the only way to assess products, Lock says.

"You need to assess whether the product will do what you want it to in your environment and with your data," Lock says. "So, you need to test with data sets that are analogous to those that it would encounter in a live situation and over a period of time that can allow the deduplication process to begin to gain traction. Also, when comparing products you need to ensure you're running horses over the same course, so subject all products to the same tests."

Data Deduplication Technology Review

October 6, 2008 | Antony Adshead, UK bureau chief, SearchStorage.co.UK

What is data deduplication?

Data deduplication reduces the amount of data that needs to be physically stored by eliminating redundant information and replacing subsequent iterations of it with a pointer to the original.

Data deduplication products inspect data down to block- and bit-level and, after the initial occurrence, only the changed data they find is saved. The rest is discarded and replaced with a pointer to the previously saved information. Block- and bit-level deduplication methods are able to achieve compression ratios of 20x to 60x, or even higher, under the right conditions.

There is also file-level deduplication, called single instance storage In file-level deduplication, if two files are identical, one copy of the file is kept while subsequent iterations are not. File-level deduplication is not as efficient as block- and bit-level storage because even a single changed bit results in a new copy of the whole file being stored. For the purposes of this Special Report, data deduplication is defined as operating at block and bit level.

What practical benefits does data deduplication have?

Data deduplication's killer app is in backup. It demands too much processor power to be used in primary storage applications.

Data deduplication reduces the amount of data that has to be stored. This means that less media has to be bought and it takes longer to fill up disk and tape. Data can be backed up more quickly to disk, which means shorter back-up windows and quicker restores. A reduction in the amount of space taken up in disk systems, VTLs for example, means longer retention periods are possible, bringing quicker restores to end users direct from disk and reducing dependence on tape and its management. Less data also means less bandwidth taken up, which means data deduplication can also speed up remote backup, replication and disaster recovery processes.

What deduplication ratios can be achieved?

Deduplication ratios vary greatly, according to the type of data being processed and over what period. Data that contains lots of repeated information, such as databases or email, will bring the highest levels of deduplication, with in excess of 30 times, 40 times or 50x times deduplication ratios possible in the right circumstances. By the same token, data that contains lots of unique information, such as image files or financial ticker tape, will not contain a great deal of redundancy that can be eliminated.

What are the advantages of hardware-based deduplication versus software dedupe?

Purpose-built deduplication appliances relieve the processing burden associated with software-based data deduplica-

tion products. The hardware-based deduplication offerings can also incorporate deduplication into other types of data protection hardware, such as backup appliances, VTLs and NAS.

While software-based deduplication usually eliminates redundancy in data at its source, hardware-based deduplication emphasises data reduction at the storage subsystem. For this reason, hardware-based deduplication may not bring the bandwidth savings that might be gained by deduplicating at source, but compression levels are generally better.

Hardware-based data deduplication brings high performance, scalability and relatively nondisruptive deployment. It is best suited to enterprise-class deployments rather than SME or remote office applications.

Software-based deduplication is typically less expensive to deploy than dedicated hardware and should require no significant changes to the physical network. But software-based deduplication can be more disruptive to install and more difficult to maintain. Lightweight agents are sometimes required on each host system to be backed up, allowing it to communicate with a backup server running the same software. The software will need updating as new versions become available or as each host's operating environment changes over time. Deduplication at the source is also processing-intensive so the host backup server must be configured for the job.

How does inline differ from post-process?

Data deduplication can be carried out inline or post process. Inline (or in-band) data deduplication removes redundant data as it is being written to media. Inline can be more efficient because data is taken in and digested simultaneously, although the additional processing power needed to handle the process may extend the backup window. The advantage to the inline method is that data passes through only once, but because it is being processed as it does, it can slow throughput.

Post-process (or out-of-band) data deduplication is carried out after data has been written to disk. This method does not affect the backup window and sidesteps CPU processing that might create a bottleneck between the backup server and the storage. Post-process deduplication uses more disk space during the data deduplication process because data is ingested then deduplicated. Disk contention is another possible issue with disk performance potentially affected as users attempt to access storage during the deduplication process.

It is recommended that you not only test the different deduplication methods to determine how they work in your environment, but also test them against backups of differing size, data types and numbers of streams.
How do deduplication products eliminate redundant data?

Deduplication systems use a variety of methods to eliminate redundant data by inspecting data down to bit level and determining whether they have been stored before.

- **Hash-based algorithms.** Hash-based methods of redundancy elimination process each piece of data using a hash algorithm, such as SHA-1 or MD5. This method generates a unique number for each piece of data which is compared to an index of other existing hash numbers. If that hash number already exists on the

index, the data need not be stored again. Otherwise, the new hash number is added to the index and the data stored.

SHA-1 was originally devised to create cryptographic signatures for security applications. SHA-1 creates a 160-bit value that is statistically unique for each piece of data.

MD5 is a 128-bit hash that was also designed for cryptographic purposes.

Hash collisions occur when two different chunks produce the same hash. The chances of this are very slim indeed, but SHA-1 is considered the more secure of the two algorithms.

- **Bit-level comparison.** The best way to compare two chunks of data is to perform a bit-level comparison on the two blocks. The cost involved in doing this is the I/O required to read and compare them.
- **Custom methods.** Some vendors use custom methods to identify duplicate data, such as their own hash algorithm combined with other methods. For instance, Diligent and Sepaton use a custom method to identify redundancy and follow that with bit-level comparison.

What is the difference between source deduplication and target deduplication?

Data can be deduplicated at the target or source. Deduplicating at the target means you can use your current backup software and the backup system operates as usual. The target identifies and eliminates redundant data sent by the backup system.

Deduplication at the source involves must installing backup client software from the deduplication vendor. The client communicates with a backup server running the same software and if the client and server agree that data has already been stored it is not sent, saving disk space and network bandwidth.

How does a deduplication device record the existence of redundant data?

Once a deduplication device has identified a redundant piece of data, it has to decide how to record its existence. There are two ways it can do so.

1. Reverse referencing, which creates a pointer to the original instance of the data when additional identical pieces of data occur.
2. Forward referencing, which writes the latest version of the piece of data to the system, then makes the previous occurrence a pointer to the most recent.

There are arguments that there is a difference in restore times possible between the two methods. For example, Sepaton claims its forward referencing method provides quicker restores.

How does encryption affect data deduplication?

Deduplication works by eliminating redundant files, blocks or bits, and encryption turns data into a data stream that is random by its nature. Therefore, if you encrypt data first—that is, effectively randomise it and remove similar

patterns—it may be impossible to deduplicate it. So you may find that data should be deduplicated first and then encrypted.

Table: Data deduplication product review

Vendor	h/w or s/w?	VTL, NAS etc?	Algorithm used?	Inline or post-process?	Source or target?
Copan	H/w	VTL and NAS	SHA-1	Post-process	Target
Data Domain	H/w	VTL and NAS	SHA-1	Inline	Target
Dell/Equallogic	See Exagrid	-	-	-	-
EMC	H/w	VTL, NAS, SAN attached	SHA-1 and MD5	Post-process	Target
EMC/Avamar	S/w	-	SHA-1 and MD5	Inline	Source
ExaGrid	H/w	NAS	-	Post-process	Target
FalconStor	both	VTL and NAS	SHA-1 with optional MD5	Post-process	Target
Fujitsu	See Avamar	-	-	-	-
HP	H/w	VTL	SHA-1	Inline	Target
Hitachi Data Systems (HDS)	See Diligent and Exagrid	-	-	-	-
IBM/Diligent	S/w	VTL	Custom	Inline	Target
NetApp	S/w (in OS)	NAS/SAN	Custom	Both	Both
Overland Storage	H/w	VTL	Custom	Inline	Target
Pillar Data Systems	See Data Domain, Diligent, Falconstor, Symantec	-	-	-	-
Quantum/ADIC	Both	VTL and NAS	MD5	Both	Target
Sepaton	S/w	VTL	Custom	Post-process	Target
Spectra Logic	See Falconstor	-	-	-	-
Sun/StorageTek	See Falconstor	-	-	-	-
Symantec	S/w	-	SHA-1	Inline	Source

Resources from Data Domain

datadomain

Data Domain—Data Deduplication Center

Deduplication: The Post—Snapshot Revolution in Storage

Webcast: The ROI and TCO Benefits of Data Deduplication for Data Protection in the Enterprise

About Data Domain

Data Domain is the leading provider of deduplication storage systems for disk backup and network-based disaster recovery. Companies worldwide have deployed Data Domain's storage systems to reduce costs and simplify data management. Data Domain delivers the performance, reliability and scalability to address the data protection needs of enterprises of all sizes. Data Domain's products integrate into existing customer infrastructures and are compatible with leading enterprise backup software products. To find out more about Data Domain, visit www.datadomain.com. Data Domain is headquartered at 2421 Mission College Blvd., Santa Clara, CA 95054 and can be contacted by phone at 1-866-933-3873 or by e-mail at sales@datadomain.com.