Clipper Notes™
VENDOR-AGNOSTIC EXPLANATIONS AND ADVICE FOR THE INFORMATION TECHNOLOGY BUYER
Navigating Information Technology Horizons℠

New in 2007 Report #TCG2007016 February 1, 2007

The Evolution of Backups — Part Two – Improving Capacity

Analyst: Dianne McAdam

Management Summary

Many enterprises are reporting that the amount of data that they need to manage and store grows every year - with no end in sight. It is not unusual to have enterprises report annual growth rates of fifty percent or more. Disk storage vendors have addressed this data growth problem by developing smaller devices that can hold larger amounts of data. Remember the 5.25-inch floppy disk drives that were the only storage devices for original PCs in the 1970s? They held a whopping 1.2 MBs of data. Those floppy disk drives could not store an average PowerPoint presentation today. Memory sticks today have replaced floppy disk drives as the portable PC media of choice. They can hold 16 or more GBs of data and easily be carried in a pocket. Middle-of-the-pack hard disk drives, the basic storage components of today's intelligent storage systems, have increased from 2 GB of capacity in 1995 to 500 GB today, with next-generation 750 GB drives just beginning to appear.

Tape drive vendors, like their disk vendor partners, have continued to expand the capacities of tape storage. IBM's round reel model 3420 tape drive, popular in the 1970s, held about 180 MB of data. Compare that to today's drives that can store up to one TB or more on a single cartridge, with a target of 8 TBs.

While disk and tape vendors have continued to improve the capacities of their storage devices, IT organizations continue to gobble up more and more storage to store primary data and its backups. Storage is like closet space in a home – you never seem to have enough of it and when you build more closet space, it seems to fill up quicker than you thought was possible. The need of IT for more and more storage is good news for storage vendors, but not such good news for customers. Little can be done to stop the growth of primary data. One large consequence to this increase in primary data is that all of this data must be backed up to ensure that when corruption or accidental deletion occurs, the data can be quickly restored. As the amount of application data grows, the size of the backups that every IT organization must manage continues to grow.

However, new technologies can help stem the tide of the rising storage requirements for backups. Read on to find out how.

IN THIS ISSUE

The Evolution of Backups

In the first part of this paper[1], we examined various improvements in backup technology that have improved the performance of regularly-scheduled backups. Backup software vendors have added support for differential and incremental backups, to improve backup times. Backup vendors have also added support for disk-based targets, allowing enterprises to choose either tape-based or disk-based backups. Hardware vendors have combined lower-cost SATA disk with tape emulation to produce virtual tape solutions.

All of these enhancements have improved the performance of backups continually, and a few have reduced storage requirements through features, such as compression or incremental backups, which only back up changed files. Nevertheless, none of these solutions has significantly reduced storage requirements. That requires intelligent software that can detect and eliminate duplication.

Reducing Duplication

Reducing duplication is the next evolutionary step in backup technology. Several vendors are now delivering products (and others are promising products) that can dramatically reduce storage requirements for backup.

This technology is called by many names, such as *single instance store*, *data reduction*, *data deduplication*, *global compression,* or *commonality factoring*. The premise is simple – we back up the same data over and over again. Alternatively, we back up data that has changed slightly - repeatedly - with the same result. **We waste storage capacity.**

Consider the many iterations of a 4 MB *PowerPoint* presentation. Changing one line on one slide means that the file has been altered. Differential or incremental backup applications would back up the entire 4 MB PowerPoint file every time a line or graphic was altered. Changing the PowerPoint presentation twenty times results in the file being stored twenty times – 80 MB of storage for one 4 MB PowerPoint presentation. However, imagine if software was intelligent enough to back up only the small line that was changed, storing only bytes of storage and not a 4 MB file.

This technology already exists or has been announced from vendors such as Asigra, Avamar, DataDomain, Diligent, Exagrid, FalconStor, RockSoft (recently acquired by ADIC/Quantum), Sepaton, and Symantec. Data-reduction software examines backup data for repeated data patterns or sequences and stores only one version of the same sequence. All other repeated sequences are discarded and replaced with pointers.

Not all data reduction works the same. Some examine backup streams as they are received to eliminate redundancies. Others examine backup data after it is stored on disk. And, there are different techniques for reducing the data.

Compression – An Oldie But Goodie

There are many different ways to reduce the amount of data stored on devices. One of the older, but still viable, techniques that deserves mention is *data compression*. All tape drives today have hardware-based compression that examines the streams of data and removes repeating characters, such as blanks. The results of hardware compression vary. Some enterprises see data compression of 5 to 1, while others see 1.5-to-2 to 1. Some disk-based backup appliances also support hardware compression.

Backup software vendors use similar compression algorithms that are available on tape drives to compress backup data. However, these implementations are software-, not hardware-based, and can affect performance. Enterprises looking to implement compression are better served by the more efficient hardware-based compression solutions.

Advantages of Data Compression

- Hardware data compression is very efficient and can reduce repetitive characters. It is a standard feature in many tape drives.

Disadvantages of Data Compression

- Software-based compression can be very inefficient and is seldom used.
- Some data, such as JPEG files, do not compress well at all.
- Data already encrypted does not compress well. Enterprises that want to encrypt and compress data need to compress data first, and then encrypt it.

SIS- S.O.S for Duplicate Files

Single instance store (SIS) schemes examine all files and eliminate multiple occurrences of the **same file**. This technique is available from several vendors. For example, *Microsoft Storage Server 2003 R2* runs SIS as a background process and examines all files, removes any duplicates and replaces those duplicates with pointers to the single copy of the file. Many email archiving solutions review messages sent to multiple readers and store only one copy of the email (and one copy of each attachment). If

[1] See the **Clipper Note** dated February 1, 2007, entitled *The Evolution of Backups – Part One - Improving Performance*, available at http://www.clipper.com/research/TCG2007015.pdf.

someone changes the file, email or attachment, a new copy of the data must be stored - in its entirety.

Advantages of SIS

- SIS can reduce storage by eliminating duplicate files. This can be very storage efficient when backing up and archiving applications such as email.

Disadvantages of SIS

- Any change to a file requires that the entire changed file be stored. Frequently changed files would not benefit from SIS.

Data Deduplication – Getting More Granular

Single instance store removes duplicate files and can save space in enterprises where the same email is sent to numerous recipients. This technology does not save space for constantly changing files, but data reduction software does! Vendors with products in this category use different terms, such as data coalescence, commonality factoring or simply *dedupe* to describe the function. One way to implement this technology, usually referred to as data deduplication, examines data within files and looks for common segment patterns. If I change the title of a 2 MB Microsoft *Word* document, SIS would retain the first copy of the Word document and store the entire copy of the modified document. However, data deduplication would recognize that only the title had changed – and only store the new title, with pointers to the rest of the document.

How Does It Work?

Data deduplication software has to split each file into segments or chunks; the segment size varies from vendor to vendor. How big should the segment size be? The answer depends on the data. If the segment size is very large, then fewer segment matches will occur, resulting in smaller storage savings. If the segment size is very small, then the information to manage all of the segments becomes very large and can limit scalability.

These vendors use hashing algorithms to create an ID for each new segment. These IDs are compared to existing IDs to determine if the new segment is identical to one stored previously. One problem with using hashing algorithms to generate 'unique' IDs is that in rare occurrences, two different segments can generate the same ID, which is called a *false match* or *hashing collision*. This condition is rare – vendors can check for false matches through further checks such as comparing the segments of the two matched IDs byte by byte.

Vendors also differ on how to split up the files. Some vendors split files into fixed-length segments, while others use variable-length segments. Each has its benefits and disadvantages.

Let's look at a simplified example that will help to explain the fixed segment process. Remember the old typing exercise?

Now is the time for all good men to come to the aid of their party.

Let's assume that we break up this sentence into fixed, five-character segments. It would look like this.

NOW I	S THE	TIME
FOR	ALL G	OOD M
EN TO	COME	TO T
HE AI	D OF	THEIR
PART	Y .	

The software would assign each segment an identification (like an ID number) and initially store each segment. Now let's change this sentence to:

Today is the time for all good men to come to the aid of their party.

Now, the segmentation would look like this.

TODAY	IS T	HE TI
ME FO	R ALL	GOOD
MEN	TO CO	ME TO
THE	AID O	F THE
IR PA	RTY .	

Now, none of the new segments matches any of the previous segments – the result – all new segments must be stored. Of course, this is a simplified example with a very few segments to examine.

Let's try the same approach using variable-length segmentation. Here the software creates variable length segments based on data dependent positions. The segmentation might look like this.

NOW	IS	THE	TIME	
FOR	ALL	GOOD	M	
EN	TO	COME	TO	T
HE	AID	OF	THE	IR
PARTY	.			

Variable-length segmentation does not always segment on word boundaries, as shown in the above simplistic example. Note that a word can be segmented into several parts. For example, the characters **THE** in the word **THEIR** are placed in its own segment to maximize the number of segment that match a high-frequency object (in this case, **THE**).

With this example, when the sentence is modified by **TODAY** instead of the **NOW**, all segments, except for the first segment match the segmentation shown in the third table above.

T	O	D	A	Y		I	S		T	H	E		T	I
M	E		F	O	R		A	L	L		G	O	O	D
	M	E	N		T	O		C	O	M	E		T	O
	T	H	E		A	I	D		O	F		T	H	E
I	R		P	A	R	T	Y	.						

Only the first segment will need to be saved as a change. Remember, this is a very simplistic example. Variable length segmentation, when implemented with some knowledge about the nature boundaries of the data, can provide greater space savings and it is less sensitive to insertions at the beginning of the file.

Advantages of Fixed-Length Segmentation

- Fixed-length segmentation always segments data in the same length segments. All data is treated equally. This approach works well with structured files, such as databases.

Disadvantages of Fixed-Length Segmentation

- Insertions or deletions within a file may require that the majority of the changed file be stored.

Advantages of Variable-Length Segmentation

- Variable-length segmentation can provide more storage savings over fixed length segmentation if it segments on data dependent positions. It is more effective than fixed length segmentation when reducing unstructured files.

Disadvantages of Variable-Length Segmentation

- Variable length segmentation requires more processing power than fixed-length segmentation.

Byte-Level Delta Reduction

Another form of data reduction is called *byte-level delta reduction*. The designers of this software understand the format of backup files and know that different versions of a backup contain a large amount of similar data. Let's start the process with Sunday's backup file, which is stored in its entirety. The next day, Monday's backup image is stored in its entirety. Then a comparison is made between the data in Sunday's and Monday's backup, and only the changes are stored (Monday minus Sunday) and the full image of Sunday's image is removed. So we now have a full Monday backup, and the delta changes for Sunday. The process continues on Tuesday, where the full Tuesday image is stored and the delta changes between Tuesday and Monday are stored along with the Monday-to-Sunday delta. The combination of one full image and its corresponding deltas is called a *version chain*. The most current backup is always stored in a non-reduced format, which means that files do not have to be reconstructed from different segments before the restore operation can be started. If Sunday's backup is required, it is recreated from Tuesday's image and the deltas from the previous two days. This approach can be very effective for backup data that changes over time. However, it is not very effective with fixed-content data.

Advantages of Byte-Level Delta Reduction

- The latest version of the backup is stored in non-reduced form that can speed up the restore process.
- Version chains are self contained and can be migrated to larger systems without losing the cumulative effects of data reduction.

Disadvantages of Byte-Level Delta Reduction

- Byte level delta reduction, by itself, does not reduce fixed content data since it compares changing files against each other. However, it can be combined with other technologies to reduce fixed content data. (See *Hybrids* section, below.)

Hybrids

Some solutions combine two or more techniques to gain greater storage reduction. For example, a hybrid solution may segment the data into large chunks and analyze the data for identical segments. These solutions can also detect nearly identical segments, called *near duplicates* and then use byte level delta reduction on near duplicates to provide further storage savings. Detecting near duplicates is not new – it is the same technique that has been in use to detect authors that attempt to plagiarize the works of others by changing a few words in the hope of avoiding detection.

Other solutions may first eliminate duplicates by single instance store techniques, and then further reduce the data through byte level delta reduction.

When to Process the Data

Vendors differ in not only how they reduce the files, but also when they do it. Some vendors that use fixed- or variable-length segmentation examine each segment as it is received, compare the segment to existing segments, and determine if the new segment must be stored. **This takes processing cycles and the solution must have sufficient power to prevent the incoming data stream from slowing.** Others, such as those that use byte-level

data reduction, process the data after they have been saved on storage at preset intervals and eliminate duplication at that time. **This approach does not slow down the incoming data rate but requires more storage to store the incoming data initially**.

Data Reduction, In General

All data reduction software can result in significant storage savings. It allows enterprises to keep more backup copies on disk than was previously feasible which means more recoveries can occur from disk. Many enterprises have long-term retention policies for backups and have procedures to migrate older backups from disk to tape. These enterprises need to understand how they can continue to create tapes using each data reduction solution.

Advantages of Data Reduction

- Data reduction software dramatically reduces backup storage requirements. Some enterprises report a reduction of 20- to-1 or more. Of course, your mileage may vary.
- Data reduction software is not a replacement for backup applications but can be used with traditional backup applications[2]. Enterprises can implement data reduction software in their environments ***without changing existing backup applications***. Incremental and differential backups can reduce backup time, while data reduction software reduces the storage requirements.

Disadvantages of Data Reduction

- Data reduction is computer intensive. Ensure the solution has sufficient processing power to examine backup data.
- Data reduction solutions add cost to the existing backup infrastructure. However, the reduction in storage requirements can significantly offset the cost of the solution.

Data reduction solutions should not be viewed as a disaster recovery solution. Provisions must be made to store copies of data off site. Many data reduction vendors leverage their technology to reduce bandwidth requirements between two or more sites dramatically. For example, a one TB backup would take about 60 days to be transmitted across a T1 line. But if only 1% of that data changes, it would only take 15 hours to transmit those changes across the same line. Data reduction can make disaster recovery solutions affordable and feasible for enterprises, without breaking the bank on communications costs.

One important note! Encryption "scrambles" data to make it unreadable. Data reduction software is not very effective when given encrypted data as input since previous patterns are now scrambled. We expect that some vendors will offer solutions that can encrypt data *after* the reduction process is complete.

[2] Avamar's implementation does not work with traditional backup applications.

Evaluating Data Reduction Solutions

Which data reduction solution is the right one for your environment? Answering the following questions may help to narrow down the choices.

1. **Does the solution work with your existing backup applications?** What changes, if any, are required to the backup process? Has the backup vendor certified the solution?
2. **Is the data reduction solution an *appliance* – that is, software pre-installed with a disk system?** Or, is the solution a software only solution? Software-only solutions allow enterprises to re-use older disk systems, while appliances simplify installation.
3. **What is the maximum performance of the solution?** Can the data reduction process be monitored and throttled to minimize performance impact during heavy backup processing? Can additional processing power be added to keep up with data growth or to improve backup and restore speeds?
4. **How much storage is initially required to hold the first several cycles of backups? (**It takes several backup cycles to achieve large reductions in storage.) How long will it take to achieve the maximum de-duplication effect? What is the expected storage savings with full backups? What is the expected savings with a combination of full and incremental backups?
5. **How much storage can the data reduction solution manage?** What is the maximum capacity of the system? Is this expressed as compressed (reduced) or uncompressed (native capacity) in size? If the maximum capacity is expressed as a compressed number, what is the assumed compression/reduction ratio? Does the solution work with existing compression algorithms (such as Lempel-Ziv)? Solutions that work with existing compression algorithms allow enterprises to achieve greater storage savings. Is there a maximum backup set size that the solution supports? If the amount of backup data increases, can additional storage be added? Is this upgrade nondisruptive? If additional solutions must be added to keep up with data growths, can these multiple systems be managed as one logical system?
6. **What level of data-integrity checking has been implemented?** How are the pointers to the various segments protected? The loss of these

pointers means that the backups can no longer be retrieved.

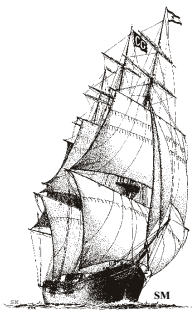
7. **How is the solution priced?** Capacity-based pricing? By number or breadth of solutions? What is the warranty period and maintenance pricing? What levels of support are available? What are the electric costs to power and cool the equipment? How much footprint is required? How do these compare to your existing backup infrastructure costs? In spite of all of these costs, you may be pleasantly surprised to see significant cost savings when you implement data reduction technology.
8. **How many solutions have been installed?** How long has the vendor been in business? What are enterprises' experiences for data reduction and performance? What backup performance do enterprises typically see? How does restore performance differ from backup performance?
9. **What is the vendor's roadmap?** What enhancements are planned and when will they be available?
10. **Does the solution support creating tapes for long-term retention?** Enterprises that use tape for long term data retention need to understand if the data reduction solution supports tape creation. If the solution does not support tape, what processes must be put in place to support tape creation?
11. **Does the solution support remote replication for disaster recovery?** Data reduction not only reduces the amount of storage required locally but also reduces the amount of network traffic during replication, which can result in significant bandwidth savings.

Conclusion

Backup technology continues to evolve since its early days. Many improvements have been made in both hardware and software to improve performance, but have not produced dramatic storage savings to keep pace with the growing amount of data that needs to be backed up and protected. Data reduction solves addresses that problem.

IT administrators can continue to run backups as they have in the past and keep their storage vendors happy by purchasing more and more storage. Or, they can implement data reduction solutions, and stop storing the same data over and over again.

The bottom line – any data reduction software can result in significant storage savings and should be a part of every enterprise's backup plan.



About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Dianne McAdam **is Director of Enterprise Information Assurance for the Clipper Group**. She brings over three decades of experience as a data center director, educator, technical programmer, systems engineer, and manager for industry-leading vendors. Dianne has held the position of senior analyst at Data Mobility Group and at Illuminata. Before that, she was a technical presentation specialist at EMC's Executive Briefing Center. At Hitachi Data Systems, she served as performance and capacity planning systems engineer and as a systems engineering manager. She also worked at StorageTek as a virtual tape and disk specialist; at Sun Microsystems, as an enterprise storage specialist; and at several large corporations as technical services directors. Dianne earned a Bachelor's and Master's degree in mathematics from Hofstra University in New York.

- ***Reach Dianne McAdam via e-mail at dianne.mcadam@clipper.com or at 781-235-0085 Ext. 212. (Please dial "212" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group Captain's Log**, **The Clipper Group Voyager**, **Clipper Notes**, and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, *"Navigating Information Technology Horizons"*, and *"teraproductivity"* are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.