Pund-IT

PUND-IT, INC.

Weekly Review

NOVEMBER 23, 2005

Space Savers: The Benefits of Freeze-Drying Your Storage

By David G. Hill, Mesabi Group

Typically, files and larger pools of data have a lot of redundancy. With the cost of both disk and tape continuing to drop, finding ways to save space by squeezing out the excess data water due to inherent redundancies might not seem to matter. But it does. Redundant patterns of data can be "freeze-dried" to save space and then later rebuilt into the original information. In addition:

- Saving disk space defers acquiring more storage thus freeing up IT funds for other uses.
- Backup performance can be improved. Since there is less data that needs to be backed up, the process can be completed faster; an especially important issue if an organization is "running out of night" for its backup.
- Network bandwidth can be allocated more efficiently when transmitting data, accomplishing more with less bandwidth and avoiding extra IT investments.
- Electronic vaulting for remote backup and replication is more likely to be affordable as bandwidth and storage requirements are smaller and less costly.

There are four techniques companies typically use to save space or bandwidth:

Compression — an algorithm (such as the Lempel-Ziv algorithm for textual information) looks at the redundancy found in a stream of bits within a single file in order to condense it; for business information a ratio of 2:1 reduction is considered reasonable; information stored on a tape cartridges is typically compressed to increase capacity and to enhance the ability to restore more data more rapidly.

Single instancing — this approach stores only a single copy of a file in a pool of storage, for example, content addressable storage (CAS) systems take a unique "signature" of each file and delete extra copies with the same signatures; the reduction ratio is different for each organization, but can be quite significant; single instancing can be used to increase efficiency in fixed content repositories and for managing that information.

File differencing — this approach notices small changes in files via a byte level scan and sends only the changes over a network from a target to a source repository; this approach improves the transmission of files over a network since only the changes are sent; this approach is useful in decreasing the time required to backup files.

Data reduction — this approach determines common sequences of data at a sub-file level across a large volume of data; the key is the ability to reassemble the

constituent parts that make up unique files; claims for how much data can be compacted are quite high.

Data Reduction: The New Kid on the Block

All four techniques have their place. However, a key principle is that the wider a solution's scope (i.e., beyond single files) the greater is its ability to eliminate redundancies. Single instancing and data reduction look at redundancies across a pool of storage. Because of its unique signature approach, single instancing records each version of a file as a separate object whereas data reduction enables time-based recovery of a document (such as a Microsoft Word document) at each stage of its evolution. Secondly, single instancing does not distinguish the commonality within files, such as the Microsoft Word overhead necessary for using the application but unnecessary for storing the file. That would seem to make data reduction useful for active archiving repositories of data, as CAS products do by using single instancing. However, data reduction suppliers have tended to focus on eliminating redundancy in the backup process.

Data reduction techniques are useful in two other ways for backup. First, since multiple servers may use the same pool of backup storage, operating system and application files (in addition to data files) that have some commonality should benefit from data reduction. Second, traditional backup processes involving full and incremental backups are likely to generate large redundancies over time. Data reduction techniques can not only save space, but also enable the time-based recovery of information as it was at an earlier stage.

Vendor Offerings/Customer Needs

The advocates of data reduction (though this is not typically the term that vendors use) report fabulous success with a 20X reduction of data, if not much more! That may be the case in certain situations, but enterprises should check such claims closely as their "mileage" may vary greatly. Structured information (databases) is different from semi-structured information (files such as word processing documents), and both are different from unstructured information (bitmapped data such as audio, video, and medical imaging).

The inclusion of metadata management capabilities is essential to providing efficient, dependable data reduction since the key result is getting data back together in a useful form. Since enterprises are dependent upon those files for their business well-being, unordered bits of data are useless.

Early vendor advocates of data reduction tend to be smaller companies. However, some larger vendors are working on as yet unannounced products.

Table 1: A Sampler of Products That Include Data Reduction

Vendor	Product	Product Focus	Technology Foundation
Asigra	Asigra Televaulting	Agentless backup and restore solutions for distributed environments	Uses a technique that it refers to as delta blocking that only backs up the changed blocks of data

Avamar	Axion family	Primarily appliance-based backup and restore solution although it can be standalone software	Uses a patented technique called commonality factor filtering
Data Domain	DD400, DD200	Onsite disk-based backup	Uses the phrase Capacity Optimized Storage (COS) to describe its approach
Tacit Networks	I[Shared] family	Eliminate branch office servers and storage as well as backup	Have a technique that looks at all of the data

Mission Accomplished?

Saving space is not the only benefit of freeze-drying storage. Improving backup performance is a goal that many organizations ardently want to achieve. In addition, more efficient use of network bandwidth not only is a boon to the pocket book, but may also enable the use of electronic vaulting for transmitting data over a network for data protection purposes. So IT enterprises should pay attention to space saving technologies; data reduction is a promising newcomer to the storage technology arsenal.

Data reduction is still a nascent technology. Organizations with specific requirements can use products that include data reduction features today, but the data reduction market still needs an IT "Good Housekeeping Seal of Approval" from one of the major players. That said, data reduction is a technology that holds a lot of potential for both business customers and IT vendors.

David G. Hill is principal of the Mesabi Group (www.mesabigroup.com). The Mesabi Group focuses on the revolutions in Storage Networking and Storage Management, and helps clients make the best and most efficient use of information for business value.

Contact:
Pund-IT, Inc.
Phone: 510-909-0750
E-mail: charles@pund-it.com
Web: www.pund-it.com