Reinvent your storage infrastructure for e-Business

Paul Wang
SolutionSoft Systems, Inc.
2345 North First Street, Suite 210
San Jose, CA 95131
pwang@solution-soft.com
408.346.1400

Abstract

As the data explosion continues to increase exponentially, concerns with data backup and storage grow. With the limited storage of disk devices, environments with hundreds of gigabytes, terabytes and more become overwhelmed. The data explosion is not new, and methods for managing huge data farms continue to change. Throughout the '90s, hardware solutions that used a disk subsystem and were controlled by a disk operating system became very popular.

However, while these hardware devices handled the data availability issues, they could not handle the backup, restore and disk-device installation issues. Data backup and restore continue to be large problems, and have actually increased with more and larger disk devices.

The presentation will first provide a general discussion of current offline and nearline archiving methods, their shortcomings and the issues that system administrators face with these types of archiving. Then a basic tutorial will be presented on a revolutionary solution for policy-based online archiving for UNIX and Windows NT platforms. A solution that stores data online, so it is available 24 by 7. Keeps more data online for security, regulatory compliance, customer service, and historical analysis. And allows access to the data immediately through automatic decompression.

The selectable policy parameters will be explained: File extension, size, name, last-modified date, or owner. Issues of the solution that will be discussed include availability of data, reliability, scalability, manageability, and performance. Advantages will be discussed for easily juggling disk space online to make room for new files, emergency tasks, or testing databases. Advantages highlighted will also include the delay of moving files to HSM tape or offline storage and reduction of time to manually archive and retrieve files.

Introduction

Companies around the world rely more and more on information systems to run their operations. This has led to an explosion of the data storage on computer systems. There is no avoiding it: the demand for disk space rarely keeps up with the requirement for it.

On one hand, new technologies such as Data Warehousing, Workflow Automation, Document Imaging, and graphics and multimedia all use large amounts of disk space and ensure that storage requirements will continue to grow at unprecedented rates. On the other hand, the price-per-megabyte of disk space continues to drop, providing more disk space for the same price and leading applications to use more data and keep it around longer. Finally, government and agency auditing requirements often mandate 5, 7 or 15 years worth of operational data must be kept available. This further magnifies the data explosion problem.

Adding more disk space is just a temporary solution for a permanent problem. Disk space will continue to grow, consume hardware budgets, and become more difficult to manage and more time-consuming to back them up.

A big part of the solution is to better manage your disk space automatically. Once those automated procedures are setup, you can be sure your MPE/iX disk space would continue to be saved and better utilized.

System administrators have historically relied on offline archiving for data backup and storage. In a typical scenario, offline archiving is a manual process for moving data to a media that is no longer connected to the system environment. When it becomes necessary to retrieve that particular data, then another similar manual process must be performed to bring the data back onto the system environment so that the data can be used. Other than the intensive time required in this manual archiving method, other drawbacks exist, including these:

1. It impacts user productivity, while he or she waits for an operator to locate the right tape and load it.
2. Locating the desired files can be a challenge with potentially hundreds or even thousands of on-site and off-site tapes.
3. When found, the data on the tapes may be corrupted due to an indefinite shelf life of tape medium caused by oxidation.
4. It increases management and helpdesk costs, as it is a very manpower intensive operation.

The other traditional data storage, nearline archiving involves moving the data to a slower media such as robotic tape and laser or magnetic optical jukeboxes. Nearline archiving is also referred to as hierarchical storage management (HSM). Retrieving data

from nearline archiving devices is slow, but is much faster than doing it from offline archiving, since it is not a manual process.

An HSM system selects files through a policy procedure and archives them. The archiving is a multi-step process including data compression and then moving the files to the nearline storage device. Additionally, when a user or application attempts to access an archived file, a time lag occurs. The HSM will find the device and media where the file is located, and then inform the device to load the appropriate media. Once the file media is loaded, the HSM will retrieve the file from the media, and decompress it, at which time the file will be available.

Issues the system administrator faces in nearline archiving include the configuration requirements for optimum storage: archiving of the least-needed data. Additionally, the HSM system must operate as desired without adversely affecting performance on a regular basis.

For example, let's say an HSM system is configured, and files are migrated to nearline devices. A "performance hit" or lag time is required to access a particular file and bring it back to the online system. If the HSM system is not properly configured, one of two situations can occur. First, the system administrator is not archiving enough data because he or she is not sure whether it will be needed or whether the performance lag time is acceptable. Or, second, too much data is archived and each time the file is accessed, lag time results.

A case in point is an application that requires a nearline-archived file every three months. On each occasion, this file is retrieved from a tape robotics system, brought back into the system, and lag time is incurred. Here's how this scenario plays out. In 60 days, this particular file is moved off the system, and 30 days later, it is moved back on the system. As a result of this highly unproductive movement, most system administrators generally opt for the first extreme of not archiving enough data due to the lag time issue.

Then, there is the cost of nearline archiving, because it is a highly complex system. Both the hardware and the software is expensive.

However, the highest cost incurred with nearline archiving, or HSM, is management. HSM is complex to configure and to manage well. Without archiving data, system administrators will definitely run out of disk space. Each time this occurs, the system is brought down, and new hardware is installed. Then, it is configured, and the data is reloaded. The downtime and management is very expensive. (This scenario assumes that the hardware was already purchased and delivered. If not, the cost of managing this system skyrockets.)

Further, the more pieces of hardware, the greater the opportunity for failure. For example, if the disk drive mean-time-between-failure (MTBF) is five years for 1 disk, then with 60 disks, MTBF is one month, and with 180 disks, it is only 10 days!

Online Archiving

The amount of disk space allocated for infrequently-used files is enormous. Just take a look how many files have not been accessed for a month and you will know what I mean. It is the 80/20 rule at work here: 80 percent of the processing is done on only 20 percent of the data. This means that 80 percent of your data is less often used or not at all (see figure 1).

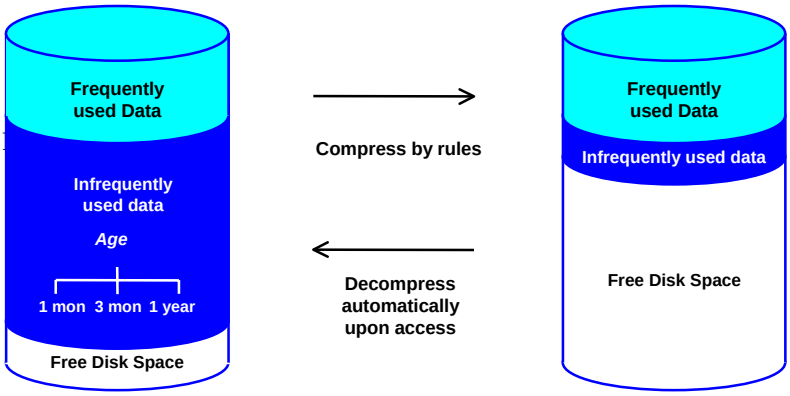

Figure 1. Compress infrequently used data by rules save disk space

Online archiving for UNIX and Windows NT environments is now making its entrance to resolve these storage and backup issues that are plaguing system administrators. Online archiving refers to taking data not being used on a regular basis and storing it efficiently on direct access systems -- disk drives or enterprise storage systems connected via SCSI, fiber, or other cabling. Additional hardware is not required in an online archiving environment. But more importantly, in addition to efficient data storage, the hallmark of online archiving is high-speed access when the data is needed. Key benefits to the system administrator are reduced backup time and reduced hard-drive requirements, which in turn, translates into reduced management, maintenance, and support expenditures.

Online archiving comes at a time when cost of ownership continues to escalate dramatically. Take for example a $10,000 hardware investment. Industry experts say the cost of running a piece of hardware like a disk drive is $5 to $7 annually for every dollar spent on hardware. Therefore, for a $10,000 investment, annual cost is $50,000 to $70,000. A five-year cost of ownership for that $10,000 hardware investment is about a quarter million dollars, including the cost of the hardware. However, online archiving helps the system administrator to save those high-level expenses by providing a more efficient way to store data. Files continue to reside on the direct access disks. Hence data availability is increased and access time is greatly reduced, compared to offline and nearline archiving.

There is also the key benefit of performance gain during normal backup procedures due to the following aspects:

1. Compressed files remain compressed on the backup tape. This reduces the time and resource required in moving the data back online.

2. There is less data to travel through disks into computer memory and then to tape devices. A probable benefit is the decrease of network bottlenecks during network backup or utilizing network- attached storage (NAS).

Set Policy and Forget It

When online archiving is used, the user sets the storage policy within seconds by specifying the filing characteristics: For example, file extension, size, name, last-modified date, or owner. Then, he or she can forget about it. Online archiving dynamically compresses the data according to the preset policy. The file remains online for immediate access and is transparent to users and applications. When the user accesses the file, online archiving retrieves that data twice as fast as it was compressed.

The user can easily juggle disk space online to make room for new files, emergency tasks, or testing databases. Also, files can be compressed to delay moving them to HSM tape storage. The user has at his or her command, compression ratios up to 99 percent so that more files can be stored without adding disks and thus, remain under budget and keep a safe lead ahead of today's dramatically growing data.

In effect, an online archiving system is automatic and all its key operations are transparent to the user. When data is compressed automatically, all filing characteristics remain exactly the same. And when users access those compressed files, they are automatically decompressed. When they finish with this data, the file is left uncompressed for better performance. Then, when they archive it again, based on the file meeting the policy, the file is re-compressed at that time. The rationale behind this is that users accessing a file will likely use it multiple times: Appending to it, updating it, or just reviewing it.

System administrators can also tune compression to trade off speed versus compressibility. When the archiving policy is set up, part of that procedure is deciding whether to optimize speed or compressibility. The user can set his or her compression policy characteristics to uphold specific performance levels.

As far as performance is concerned, files compressed at any ratio consume less time to transfer to directories such as NFS drives or to write to tape. Faster restore reduces the time spent on system reloads and disaster recovery.

Online Archiving and HSM

Online archiving is complementary to nearline archiving. For instance, a system administrator may currently have 500 gigabytes of disk space. He or she knows that one day soon, a nearline system will be needed. But he or she does not want to engage it yet because system administration resources aren't available.

Here's where online archiving provides system administrators a stepping-stone to archiving capability without the difficulties involved in purchasing and installing a hierarchical storage management system. Plus, an online archiving system can serve as an educational tool for understanding what archiving is really about before moving into the complications of an HSM.

Online archiving works with an HSM system as a first-line archiving step. For example, a file that is not used after 30 days becomes online archived. If, after six, nine, or 12 months that particular file hasn't been touched, then it is moved to the nearline device. The system administrator immediately gains the archiving efficiencies with online archiving before that file is moved off. This way, system administrators gain that additional step that they don't have with an HSM system.

If they already have an HSM system in use, system administrators would naturally question having online archiving. The most significant advantage of adding online archiving is to be able to efficiently store and quickly access valuable files that are periodically needed, without incurring costly performance penalties. On the other hand, if they don't have an HSM system, but are rapidly accumulating data, adding enormous numbers of disk drives, and are worried about the eventual HSM purchase and installation, then they can opt for online archiving. This involves easy software installation without hardware considerations.

This approach alleviates system administrators from being overly concerned about disk space usage or worrying over the problems of archiving the right amount of data or of incurring major performance lag time. The major benefit of online archiving is gaining disk space savings without taking performance hits.

Conclusion

In summary, online archiving provides the following benefits:

1. Reduced system administration costs through the reduction of disk space for archived files. This saves $50,000 to $70,000, annually, per $10,000 saved in hardware.

2. Reduced backup media and time, since fewer bytes are used.

3. Reduced occurrence of out-of-disk space failures and future disaster-recovery time to increase up time.

4. Reduced number of disk drives to decrease risk of hardware failures.

5. Increased archiving capacity and performance through the addition of a stepping stone before needing to archive nearline.

Biographic

Paul Wang is the president of SolutionSoft Systems, Inc. He has more than sixteen years of industry experience and is a specialist for storage management and system performance. Previously, he was an internal architect at HP.