A CMI Business Brief White Paper:

The Case for Hierarchical Storage Management

By Jeff Guenthner, CMI

By automatically moving data between different storage devices based on policy and budget requirements, Hierarchical Storage Management (HSM) drives costs out of the data center with minimal impact on business operations.

In an ideal world, companies could instantly store and access all of their data on high-speed devices at all times. In the real world – a world where companies are experiencing data growth of 40% to 60% per year,[1] and where high-performance storage media cost as much as $100 per gigabyte – even Harry Potter and his Hogwarts companions couldn't conjure up such a miracle.

Enter Hierarchical Storage Management (HSM) – a near miracle also known as "Tiered Storage." An HSM system automatically moves data between high-cost and low-cost storage devices with minimal impact on business operations. Using HSM, massive amounts of information can be "migrated" between different logical and physical locations based on policies, performance expectations and cost requirements established by the organization. IT managers can store much of the company's data on slower, low-cost devices, and copy it back to faster, high-cost devices only when it's needed.

Let's assume that you manage a financial services company with 100 terabytes of data. You could have all 100 terabytes sitting on high-performance disk drives, keeping all that data right at your fingertips with sub-second response times. If money were no object, you wouldn't mind spending twice as much (or more) for an all-disk storage system than a multi-tiered system, which stores infrequently accessed data on tape or in the cloud. If money were an issue, however, you might decide to examine your storage requirements on a file-by-file basis. If you did, you'd probably discover that huge amounts of data weren't being accessed after 60 or 90 days – and that some data wasn't being accessed at all. (When did someone last view the company's fixed asset report from March 2002?)

In an HSM environment, information is prioritized and policies developed to assign information to different storage tiers based on how often it's likely to be needed in the future. For example, management and IT might decide that any file that isn't accessed after 60 days will be pushed from high-performance spinning disks to low-cost tape in "near online" storage.

When a program looks for a file on a high-performance disk drive, it instantly reads it. However, if the HSM software has been told to move that file from the primary to the secondary level, the software leaves a "marker" on the disk drive, indicating that the file

was moved to another storage medium. To access the file, it is simply retrieved from the other device and made available to the program. The same thing occurs at the tertiary level. If a program looks for the file, and sees a marker indicating it's been moved to the tertiary level, a request is made to retrieve the information from the other storage device so it can be accessed by the program.

At the average enterprise:

- Less than 20% of all IT files are accessed again within 30 days of creation.

- A large amount of duplicate files exist in the storage environment.

- Several terabytes of archived files gobble up expensive storage "real estate."

- Using HSM, the total amount of stored data can be much larger than the capacity of the high-speed disk. Since only rarely-used files are stored on slower and less accessible devices, most users won't notice any slowdown.[2]

Continuous, automated migration of data produces minimal disruption to operations; it can save millions of dollars annually, since devices such as SATA (Serial Advanced Technology Attachment) hard drives, tape and cloud-based services store data for a fraction of what high-performance disk drives cost.

Legacy Solutions Can't Cope

By continuously and automatically monitoring for ways to push data into least-cost storage media (tiering), companies can: reduce the cost of data storage and administration, reduce the amount of network bandwidth needed, simplify resource sharing, reduce costs and time for data backup, and ensure that any data needed to comply with government regulations can be quickly located and retrieved – when needed.

Legacy backup/restore solutions can't cope with today's data storage challenges – much less tomorrow's challenges. According to a recent Gartner Research report, at least 30% of organizations will change backup vendors by 2014 out of frustration over costs, complexity and/or inadequate capability.[3]   What's more, for every dollar an organization spends on additional storage, it can spend five times that amount (or more) on data management. And in general, more data means longer backup and restore times, which means more downtime and less profit. Worse, inadequate backup puts

critical data at risk, which can prove even more costly in the long run.

The Sum of All Tiers

In essence, HSM exploits the performance and cost characteristics of various storage media. It was first implemented by IBM on mainframe computers to reduce the cost of storage and to simplify data retrieval from slower media. The beauty of the system was that users didn't need to know (or even care) where data was being stored, and how to get it back, since the computer retrieved it automatically.

Later, the development of Serial ATA (SATA) disks created a market for three-stage HSM. Files are migrated from high-performance Fibre Channel Storage Area Network devices to slower but cheaper SATA disks arrays, and eventually moved from SATA to tape. An even higher performance tier of storage is based on Flash or DRAM memory in servers. This increases performance, as well as access to "hot data," and creates a fourth tier of storage:

Tier 0 uses solid state drives (SSDs), which get their high performance from high-speed Flash memory instead of spinning discs. In an automated tiered storage environment, frequently accessed data can be kept on SSDs while the rest is moved to other tiers. *The ASP (Average Selling Price) range for Tier 0 is $50-$100 per GB.*

Tier 1 includes high-speed hard drives, typically those with SAS or Fibre Channel interfaces. *The ASP range for Tier 1 is $7-$20 per GB.*

Tier 2 usually includes lower-speed SATA hard drives. SAS or Fibre Channel drives feature higher performance than SATA drives, but SATA drives are much less expensive on a per-GB basis, and they feature much higher capacities. *The ASP range for Tier 2 is $1-$8 per GB.*

Tier 3 storage includes low-cost tape and optical disc media. Their large capacities and slow performance make them useful for backing up or archiving data that isn't likely to be accessed for normal business purposes. *The ASP range for Tier 3 is 20 cents to $2 per GB.* (For small and mid-sized companies, the cloud may be another option for Tier 3 storage. However, certain challenges must be overcome to ensure its viability. To start, IT managers should review service level agreements of cloud providers to gain clarity on the security and information

governance policies; work to secure annual audit rights; and determine the degree of difficulty in retrieving the migrated data.)

As most data ages, requests for access drop off quickly, and it moves to an archival status. Without HSM and tiered storage, this data would hog expensive Tier 1 and Tier 2 disk systems. (Tier 3 storage is often labeled "green storage" because it has the lowest energy consumption of any tier.)

Important considerations when implementing tiered storage is data deduplication and its related concept "intelligent data compression." (For purposes of our discussion here, we will presume the two concepts are essentially the same.) In deduplication, unique or repeating chunks of data, (i.e. byte patterns), are identified and stored during a data examination pre-process. As the examination/analysis continues, other chunks/patterns are compared to the stored version and when there is a match, the redundant chunk/pattern is replaced with a small reference that points to the stored chunk/pattern. Since a given chunk/byte pattern may occur dozens, hundreds, or even thousands of times, depending on the nature of the data in question, the amount of data to be stored/transmitted can be, and often is, greatly reduced. Net result: a 'compressed' (read: smaller/lesser) amount of data to be stored, faster data loads/restores, and lower storage costs.

One final note: cloud storage is playing a larger role in storage in general, and tiered storage specifically, and for good reason. Particularly for Tier 3 and archive data storage, as just one example, cloud technology fits well (by serving as a thin provisioned long-term storage pool) which also eliminates the need for tape, tape drives, Iron Mountain-type storage, and so on. Today's organizations might be well served by considering cloud storage as an option when considering/implementing HSM.

Tivoli Storage Manager

HSM solutions vary by vendor, but to best illustrate how this technology works in the real world, let's examine IBM's Tivoli Storage Manager. It contains most of the features and benefits that characterize these policy-based data backup, storage and recovery products.

Tivoli Storage Manager is designed to provide centralized, automated data protection that reduces the risk of data loss while helping IT managers control costs, reduce complexity and comply with regulatory data-retention requirements.

including check-in/check-out procedures.The solution also takes advantage of Storage Manager's ability to store data in sequential volumes on disk, which produces a winning combination of high-performance and manageability on the disk subsystem. Best of all, IT staff no longer has to wait for tape mounts, which has dramatically streamlined restore times.

Thanks to Storage Manager's ability to store data long term on disk, and its progressive backup methodology, this company lowered the amount of data it backs up each night from 2.5 TB to 200 GB and slashed the amount of storage space needed.

While many storage management products produce massive amounts of duplicate data during their periodic full backups, Storage Manager backs up only new or changed versions of files, slashing redundancy, network bandwidth requirements and storage pool consumption.

The key benefits and features include:

Data life-cycle management: Tivoli products help companies reduce costs by moving data to the most cost-effective storage tiers while still meeting service requirements, ensuring data recovery goals and enabling transparent data access. Automated data archiving also helps to ensure compliance with data retention policies and to reduce the costs associated with compliance.

Data Reduction: Storage Manager uses data reduction technologies such as progressive-incremental backup, data deduplication and data compression to let organizations reduce backup storage capacity by up to 95 percent. It also provides advanced tape management and efficient tape utilization, which can further reduce storage capacity requirements. Reducing backup storage requirements lowers capital expenses, network bandwidth requirements and the operational impact of backups. By creating only an initial full backup and then capturing only new and changed data, Storage Manager avoids creating duplicate data in the first place.

Unified Recovery Management: URM gives IT groups the ability to manage data protection with a single user interface. From a single point, IT can support different applications, data types, operating systems and locations, as well as the different policies governing them. This reduces the costs and complexity associated with deploying and managing multiple data protection and recovery tools.

Scalability: Tivoli Storage Manager grew its storage capacity by 100 percent a year from 2009-2011, and now supports four billion data objects in a single server. This increase in capacity has greatly outpaced the 40% to 60% average data growth cited by IDC.

Ease of Use: The single-server footprint of Storage Manager makes it easier to manage, and automatic software updates keep IT administrators from spending valuable time making manual updates to systems across the infrastructure. These administrative time savings can dramatically reduce the cost of operations.

Policy-Based Migration: Storage Manager has a granular policy engine designed to automate its capabilities. Administrators can set policies as simple defaults or as customized options to reduce costs and still meet specific service-level agreements. Once configured, Storage Manager can manage data protection for thousands of computers with minimal administrator assistance.

High Performance: Storage Manager provides high-performance backup and recovery technologies that minimize the impact to business operations and, therefore, to revenues, reducing the potential for downtime caused by routine backup and recovery processes. For example, Tivoli Storage Manager FastBack can get applications and users up and running within minutes of any data loss, whether in the data center or in remote or branch offices, while full data recovery is performed in the background.

Support for Virtualized Environments: Although virtualization technology has helped organizations reduce their costs by improving server utilization and reducing application provisioning times, these savings can evaporate in the face of virtual machine sprawl. For example, backing up and restoring data for a dozen virtual machines residing on one physical server can bring all other operations on that server to a complete halt. Storage Manager for Virtual Environments solves this problem by eliminating the need to run backups on a virtual machine. This improves the frequency of backups, and enables faster recovery of data, which increases the business value of virtualization.

Assessing Storage Management Needs

Before attempting to drive costs out of your business by driving data to less expensive storage media, it's important to first assess how often different files are accessed and used. Only then can you make educated decisions about when and where the files should be stored. Only then can you create intelligent policies that match data to the best storage media, reducing costs while maintaining the integrity of your data.

Identifying which files to retain and which to migrate or delete can be a daunting task. Among the other challenges, there are often several copies of the same file in different locations. A storage management assessment should involve both the business management and IT teams, and include the following:

- A discovery audit of the environment (file systems, SAN storage, SAN fabric, library (archiving), metadata, databases, applications)

- An evaluation of the power, cooling and storage requirements

- A determination of the environment's scalability as the amount of work grows

As the data classification process proceeds, the assessment team must formulate prioritized retention policies that assign data to different tiers based on frequency of access and the data's importance to the business. These formulas will assist you in understanding the types of storage required and how and when to use them.

Many organizations struggle with how to manage data storage. HSM can help you cost-effectively manage, store, back up and recover business-critical data without noticeably affecting the average user's experience with the system.

---

[1] Amatruda, Robert and Laura DuBois, "Market Analysis: Worldwide Data Protection and Recovery Software 2011 – 2015 Forecast," IDC, July 2011.

[2] Laura Paoletti, "The Benefits of Hierarchical Storage Management." Tom's ITPro, April 12, 2012.

[3] IRussell, Dave, Sheila Childs, and Alan Dayley, "Gartner Magic Quadrant for Enterprise Disk-Based Backup/Recovery," Gartner Research, January 28, 2011.

Status Check: By Kris Neely, CMI's Chief Technology Officer

I'll admit up front that I'm a huge fan of Hierarchical Storage Management (HSM). Having juggled my fair share of IT budgets in my time, I know the incredible growth in both the cost and volume of storing data today – and the value tiered storage architectures such as HSM provide.

Left to their own devices, every IT department end-user will tenaciously fight for keeping all the data and files they own on high speed/high cost online disks. These well-meaning folks will argue all day that they don't just 'want' access to their data/files but they 'need' it. Yet, as this White Paper underlined, less than 20% of all IT files are accessed again within 30 days of creation. And with the passage of 30 days comes another "month-end" process and that first set of data gets supplanted by more current archived data, and the process continues. The result is lots of data just, well, sitting there taking up valuable high-speed space – and spending lots of money doing so.

Don't get me wrong: I know that end-users will eventually access some/many of those files – especially in today's world of intense and detailed data analytics. That means the data has to be available to the end-users, I get that. But at what cost?

In addition – losing data has become a criminal offense. Seems like every month brings a new story of some company or organization or government entity looking for the email(s) or other data that explains who-knew-what and when-did-they-know-it. Having data go 'missing' in such a scenario is an invitation to do the "perp walk". Add in the requirements of laws such as Sarbanes-Oxley (SOX) and losing any data whatsoever (even, so-to-speak "show your homework" interim data used to show how a company got from Point A to Point D, for example, and all the interim steps/levels) becomes huge.

If your firm hasn't explored HSM, I urge you to undertake an exploratory effort. That said – don't make the mistake of implementing HSM across your entire organization at one time or in one project. As this White Paper showcased, there are operational and performance issues associated with HSM, and those changes need to be absorbed and processed by the organization at a pace it can deal with.

Start small – prove the concept and get buy-in from the organization. Also build a financial model that demonstrates the costs involved. HSM is a great technology that has a place in organizations large and small.

See you next time,
Kris Neely
CTO, CMI

Reality Check: A CFO Comments on This White Paper

As technology has evolved in the past few years, with companies becoming increasingly mobile, moving to the cloud and increasing their data footprints, Hierarchical Storage Management is an important part of any IT and data infrastructure strategy.

In my own experience in the Visual Effects Industry, HSM became an essential part of the IT architecture. As my company expanded into multiple locations and started using server co-location facilities (essentially a dedicated cloud), the access and synchronization of data became a big headache. We were dealing with enormous data sets, and having an artist in Vancouver wait 60 minutes for a scene file to download was not an option.

The solution was to use large–format cache storage devices at each location. Based on policy, high frequency files are downloaded to each location to enable fast access by the local artists. This is exactly the same as HSM - data storage was intelligently applied to enable each user in each location to access the right files as quickly as possible. The local cache devices are akin to a Solid State Drive with the cloud being the SATA storage in a typical company. For us, the alternative of having large amounts of storage at each location was not financially viable.

This solution of high speed local storage combined with a cloud was much cheaper not only as we set up each location, but also in the long term. We estimated a saving on the initial capital equipment of over 25%. It was also more scalable - additional storage could be purchased at the cloud and did not need to be replicated at each location – a large ongoing saving. For the average company, applying this hierarchical approach to the devices that store data can lead to comparable savings.

While the Visual Effects Industry has unique and extreme data requirements, it really is analogous to what many companies are facing – increasing data sets, geography, increasing mobile device use, virtualization and moving to the cloud. Any combination of these factors means that a "one size fits all" approach to IT and storage will result in inefficiencies, which means higher cost. As your organization and technology infrastructure evolves, an HSM strategy is an important part of the overall plan which can create significant and material financial and operational savings.

Best,
Kevin

Mr. Kevin Weston, is the former CFO of Digital Domain Productions, as well as former VP Finance and Operations of LucasArts, and was CFO/SVP Finance and Operations at Eidos Interactive Inc.