

 SearchStorage.com All-in-One Guide

Advanced Storage

Chapter 5:

Data/Storage Management

It's relatively easy to add more disks or arrays to meet storage growth. Disks are also getting cheaper, so today's storage is often more cost effective than years past. But storage itself isn't the problem — the challenge lies in managing storage volumes that escalate at annual rates frequently exceeding 60%. While there's no universal management "solution" today, the products are evolving and practices are constantly improving.



Advanced Storage Chapter 5: Data/Storage Management

Table of Contents

[Overview](#)

[Data management tools: Is storage automation right for you?](#)

[Data management tools: Buyer's Guide — Data classification tools](#)

[Data management tools: SRM packages — A closer look](#)

[Data management process: Managing data migration among storage tiers](#)

[Data management process: The rising cost of cheap storage](#)

[Data management process: Archive or backup?](#)

[Data management process: Tech Report — Content-addressed storage preferred for fixed-content storage](#)

[Data management process Expert: E-mail archiving with CAS](#)

[Data management process: Out of capacity — What now?](#)

[Data management process: Thin provisioning slims storage volumes](#)

Overview

By Stephen J. Bigelow, Features Writer
September 1, 2006 | SearchStorage.com

It's relatively easy to add more [disks](#) or arrays to meet storage growth. Disks are also getting cheaper, so today's storage is often more cost effective than years past. But storage itself isn't the problem — the challenge lies in managing storage volumes that escalate at annual rates frequently exceeding 60%.

Administrators need a versatile and robust suite of tools to report utilization and availability, streamline storage allocation, move data between disks and platforms, and automate many of the mundane but essential practices that today's IT organizations rely on. While there's no universal management "solution" today, the products are evolving and practices are constantly improving.

Data management tools

Administrators can leverage a rich suite of software tools to manage tasks like configuration, [provisioning](#), [migration](#) and [archiving](#). Additional tools monitor availability, measure performance and report timely data to decision makers.

Configuration tools are usually tailored to specific storage systems, usually supporting setup and operational characteristics. As one example, the Hitachi Resource Manager Utility Package can be used to configure storage systems from Hitachi Data Systems Inc. (HDS). The utility allows users to define, change and reassign [logical unit numbers](#) (LUN) without rebooting, handle virtual LUNs, manage cache and maintain security. Hewlett-Packard Co.'s (HP) StorageWorks LUN Configuration and Security Manager XP software is another notable product. The biggest areas of growth in configuration tools are improved heterogeneity to support an increasing range of systems and superior security to prevent unauthorized changes to critical configurations.

Provisioning tools are used to allocate storage resources to specific users or applications. More recently, provisioning tools are adding removal (deallocation) features. Today, there are many tools to choose from, including Storage Manager 2.5 from Crosswalk Inc., WysDM for Fileservers from WysDM Software Inc., adaptive provisioning software from 3PARdata Inc. and automation software from Opsware Inc. Larger players in the industry are also important. These include HP's Storage Essentials Enterprise Edition software (using AppIQ technology), the Hitachi HiCommand Path Provisioning module (part of Hitachi HiCommand Storage Services Manager), Storage Provisioning Services in EMC Corp.'s ControlCenter software, Tivoli software from IBM and Veritas Provisioning Manager from Symantec Corp. As with configuration tools, provisioning tools are expanding their compatibility across multiple platforms — a major element of that compatibility is the adoption of [SMI-S](#) standards. The tools are also getting easier, using better intelligence and wizards to automate more of the provisioning process.

Migration tools handle data movement from one storage platform to another, but it's important to select a product that will move data in a fashion that is best suited for your environment. For example, the Transparent Data Migration Facility (TDMF) from Softek Storage Solutions Corp. moves data between disk

storage devices and a local server. The Veritas Volume Replicator from Symantec moves data between servers using TCP/IP. The Topio Data Protection suite (TDPS) will do both. All three tools move data in blocks.

Archiving tools also move data like migration tools. But since archives are typically intended for long-term data retention, archiving tools often focus on single-instance storage (e.g., data deduplication to reduce disk use) and security features that prevent unauthorized data alteration or deletion. There are numerous hardware products, including EMC's Centera, the SnapLock from Network Appliance Inc. (NetApp), IBM's DR500 and HP's Reference Information Storage System. Hardware products are usually categorized as [content addressed storage](#) (CAS) devices. Some archiving software is application specific, focusing on email, database or other applications. These products include EMC's Email Xtender, Message Manager from CA and archiving products from Zantaz Inc. Archival and search tools are increasingly used to meet compliance requirements.

Finally, administrators use monitoring, measurement and reporting tools to gain insights into the behaviors and usage of the storage infrastructure, helping to identify problem performance areas and plan for future storage expansions or upgrades. Notable products in this area include Symantec's Veritas CommandCentral Storage product, Vantage from CA, StorageScope from EMC, StorageEssentials SRM (storage resource management) from HP and Storage Manager from Softek. The key attributes of these tools are heterogeneity, and support for a diverse array of storage products within the user's environment; the product should be able to see and report on everything in order to be useful.

Data process tools

While tools can add tremendous value to a storage organization, they are useless without a suite of internal policies and procedures that address your specific business requirements. Processes are frequently broken up into change management, performance and capacity planning, and [tiered storage](#), sometimes broadened to [information lifecycle management](#) (ILM).

A [change management](#) process defines the steps needed to add or change storage for a user or application. It may start with a change request and then wind through management authorization, budgeting and storage administration before finally being implemented. This may seem like unnecessary bureaucracy, but the goal is to prioritize and coordinate changes to get the most value from a storage investment. Companies that choose to tackle change management in-house often employ tools like SANscreen from Onaro Inc. or the Redcell Network Change Management Solution from Dorado Software Inc. Independent consultants, like Protiviti Inc. and NaviSite Inc., offer change management services for IT organizations that prefer to outsource change functions.

Storage organizations must also monitor performance and prepare for growth at the storage and network levels. This ensures that each application or user receives the resources necessary to meet service level agreements, but prevents excess spending on unnecessary storage or infrastructure upgrades. Capacity planning and performance management (CPM) tools are often used to streamline this management, including PerfMan from Information Systems Manager Inc., TeamQuest View from TeamQuest Corp. or nGenius from NetScout Systems Inc. Dedicated appliances, like the Avalanche 2200 from Spirent Communications,

can simulate real-world conditions that stress test a network infrastructure. While many tools are intended for enterprise-wide analysis, some CPM tools are optimized for demanding applications, like SQL Server or Oracle. Some examples of these include BEZProphet for Oracle, BEZPlus for IBM DB2 UDB and BEZPlus for NCR Teradata — all from Bez Systems Inc.

Finally, ILM manages stored data from the time it is created until it is destroyed, saving storage costs by placing data on a corresponding storage tier and migrating data between tiers as the data's value changes over time. This is a difficult business process because [data must be classified manually](#) by individuals that understand its value to the organization. Then, rules for handling and disposal must be applied to each classification. Organizations frequently leverage tools for data classification, search, migration and retention. Typical tools include HP StorageWorks, ArC from Archivas Inc., or SnapLock from NetApp. There are also dedicated hardware platforms for data classification, such as the IS-1200 family of dedicated information servers from Kazeon Systems Inc. and the ICM 5000 platform from StoredIQ Corp.

Compliance

Storage is now further [complicated by a proliferation of regulations](#) that govern data retention, integrity and security. Meeting the demands of these regulations is called compliance. Some of the most notable regulations include the [Sarbanes-Oxley Act](#) (SOX) and the [Health Insurance Portability and Accountability Act](#) (HIPAA), but there are now more than 10,000 state and federal regulations that impact data storage. With so many rules affecting key industries like finance and healthcare, IT professionals often rely on tools to meet their own compliance efforts.

Compliance typically relies on the use of storage platforms like EMC's Centera or Clariion. NetApp is also a notable player with SnapLock and LockVault software running on NetApp FAS and NearStore storage platforms. The Axion storage system from Avamar Technologies Inc. handles legal discovery and support for regulatory compliance. IBM offers the DR500, while HP provides the Reference Information Storage System (RISS). Email archiving and workflow automation software is also used to help meet compliance requirements. Encryption is playing a greater role in security — especially for data at rest within the data center or to protect data sent off site.

Data management tools: Is storage automation right for you?

Christopher Poelker
May 11, 2006

What you will learn from this tip: You will become more familiar with what storage automation is, and what it is not, and what it really means to you. Also, you will learn why you should, or should not implement a storage automation product.

There has been a lot of buzz around the term "storage automation" lately. It sounds nice, but what does it really mean, and do you really need to spend the money to make it a reality for your organization? Depending on who you talk to, the term storage automation can mean a number of things:

- Automated [provisioning](#)
- Capacity on demand
- Host transparent data movement
- Hot spot elimination
- Transparent data migration
- [ILM](#)
- Utility storage

Storage automation is the ability to encapsulate time consuming or repetitive tasks into a best practices policy that can be initiated on command or triggered via an event with little or no human intervention; you can think of it as being like a script.

Automation is a good thing for routine tasks that take a lot of time. Automating routine tasks can decrease administrative overhead and save money by allowing fewer people to do more, and can eliminate the human error factor. Storage automation products are available from software and hardware vendors as software modules or hardware microcode that can:

- Provision storage to hosts based on your policies for [switch zoning](#), [LUN](#) masking and performance criteria.
- Move application data from low-performance to high-performance disks (or visa versa) within an [array](#) based on application [I/O](#) metrics or statistics.
- Migrate data between different classes of storage based on data age or frequency of access.
- Automatically expand LUN sizes to hosts based on policies for how much capacity is left.
- And a host of other cool capabilities that are outside the scope of this tip.

Storage automation is not a panacea for IT administrators, so they can just forget about the day-to-day administration of storage resources to focus on more important stuff.

But let's face it: Stuff happens. Even though routine tasks can be automated, administrators still need to be on their toes and monitor the environment to make sure everything is humming along properly. An automation process may fail, or hardware may fail in the middle of a data move, which may cause data corruption. Automating routine provisioning is a great place to start implementation, since failures during that process present much less risk to critical data.

Do you really need storage automation? That depends on your environment. If your shop is small, and your employees are familiar with API scripting, buying off-the-shelf automation tools may not be required, since you can roll your own. If your shop is large, and your staff is continually fire fighting or dealing with outages, implementing vendor tools to automate most of the day-to-day tasks will allow your staff to be more productive. In either case, make sure that the product you put in place includes enterprise-class monitoring and reporting capabilities, so your administrative staff will have the proper controls and problem notifications should something go wrong.

Data management tools: Buyer's Guide — Data classification tools

By Stephen J. Bigelow, Features Writer

May 8, 2006 | SearchStorage.com

As the volume of corporate data continues to grow, storage administrators are faced with two distinct problems. First, all data cannot be treated the same, so administrators must match the cost and performance of each tier to the value of each data type across the enterprise. Second, growing data volumes make it increasingly difficult to find specific files once they're stored. Just imagine searching through millions of files to locate a missing memo or important e-mail. The practice of [data classification](#) allows a corporation to organize its data according to its relative value so that it can be stored to the appropriate tier and more easily retrieved in the future.

Automated tools play an important role in the data classification process. Many tools allow administrators to discover the data resources that are available, apply uniform classification rules to data across the entire enterprise, create and manage a searchable index of detailed metadata, move data to the corresponding tier and later comb the metadata to conduct detailed searches. This article explains the essential concepts of data classification tools and their role in the enterprise, highlights the leading vendors in the marketplace and offers some advice to help ease purchasing and implementation issues.

Understanding data classification tools

Data classification tools generally cover four main areas to some extent: discovery, classification, search and migration. The discovery process identifies the files and data types available in your infrastructure — it tells you what you have. Classification works on the discovered data, applying metadata to each file and file type based on a defined set of rules. Metadata is then stored in a database that can be searched and referenced later. The rules themselves may be developed internally within the enterprise or may be imported into the tool from a third-party source. Once implemented, rules can be changed and tweaked over time as business and technical needs dictate.

Search capabilities are a natural extension to classification, utilizing the metadata created in the classification process to locate files based on criteria that goes well beyond conventional metadata, like filenames or creation dates. Search features are particularly important when data is being classified for archival or compliance purposes. Since data classification is usually coupled to a tiered storage strategy, data migration features (sometimes called policy management) can help to move files across the storage infrastructure. For example, non-critical files can be moved from Fibre Channel (FC) disk to a SATA storage array, or infrequently accessed data can be moved off to a content-addressed storage (CAS) platform. It's important to note that not all data classification tools provide search and migration capability.

Analysts are quick to note that data classification tools are becoming more robust and thorough, often able to examine files and documents for keyword sequences and make contextual decisions about the data. "Now it's more about getting in and looking at the data," says Greg Schulz, founder and senior analyst at

Storage IO Group. The trick is in giving the tool enough information so it can draw inferences and make intelligent decisions about the data it is examining.

Hardware vs. software

Data classification tools may be implemented as hardware or software. Software-based tools are installed on at least one server in the enterprise, though multiple servers may be aggregated together to improve discovery and classification performance. In fact, multiple servers may be essential for larger organizations managing hundreds or millions (even billions) of files, or requiring hefty classification rates (e.g., 1,000 files per second).

Tools may also be implemented as hardware appliances — essentially dedicated servers running data classification software. Although more expensive than software-based tools, hardware appliances are generally easier to integrate and configure, especially when clustering appliances together, and support a wider range of enterprise operating systems.

Vendors and product selection

The data classification arena is broad — most vendors have a unique take on the scope, utility and scalability of their own data classification tools. Recognized vendors like Kazeon Systems Inc. take an all-encompassing approach. Kazeon's Information Server IS1200 appliance promises to catalog and classify all files on the network, providing detailed reports intended to help improve storage efficiency. StoredIQ takes an even broader view, building on discovery, classification and migration features to include retention policies and maintaining an audit trail of classified data activity. This general-purpose approach is consistent with the general definition of [data classification](#).

Abrevity Inc. also takes an all-in-one approach with its FileData Classifier software, intended to offer discovery, classification, policy management, security, backup and archiving features for small and midsized businesses (SMB). Comprehensive search capability is provided by Abrevity's separate FileData Manager utility. Even emerging products like Destiny, from startup Scentric Destiny, touts a universal product to data classification, allowing cataloging, classification and control of structured and unstructured data. Arkivio Inc., Network Appliance Inc. and Hewlett-Packard Co. also offer general data classification/information life-cycle management (ILM) products.

But some companies take a more narrow view of data classification, catering to specific applications within the enterprise, such as Exchange. One example is Exchange E-mail Indexing Appliance from Index Engines Inc. The appliance interfaces to the SAN, indexing e-mail and documents during the backup process. Intradyn Inc. focuses on the SMB market, offering the ComplianceVault06 appliance designed for e-mail archiving and retrieval with applications like Exchange and Lotus Notes. NearPoint from Mimosa Systems Inc. also deals specifically with Exchange, providing archive, discovery, recovery and storage management features through a software-based product.

And of course, EMC Corp. touts numerous hardware and software products designed to address the various aspects of ILM technology.

Selecting the right product

Selecting a data classification product can present some unique challenges for an organization. Each tool is different — often focusing on a particular strength, such as data migration or searching. As a rule, determine what functionality you need from a data classification tool in advance, and then weed out tools that do not provide the desired feature set. Once you narrow the field, a few potential candidates can be thoroughly tested in-house. Analysts suggest the following points that can help you identify the best product for your own production environment.

Consider the product's versatility. Any data classification tool must be compatible with the types of data that you work with. Since the majority of company data is unstructured, global data classification initiatives should use tools that fully support structured and unstructured data. Tools that handle only structured or unstructured data, or are only intended for certain applications, may not meet your objectives.

Consider the product's scalability. Data classification products generally have a practical limit to the number of files that they support. Make sure to select a product that can accommodate your current and anticipated future data volume. Understand the upgrade path so that you can estimate the cost and effort needed to expand the data classification platform later on.

Evaluate the support for external rules. All data classification products rely on a set of rules that drive the classification engine. Early data classification tools relied almost entirely on rule sets created in-house, but many of today's tools can import established rule sets — often to support medical or legal industries. Also determine if imported rule sets can be modified or adapted to your specific needs.

Consider the impact of hold capabilities. If your primary concern is locating and protecting specific data involved with litigation, consider a data classification tool with litigation-hold (or file-hold) support. That is, when a search is conducted, the data involved in the search can be frozen to prevent modification or deletion — even if deletion had been previously approved.

Evaluate compatibility with outside tools. Although some data classification tools can manage policies or move data to the appropriate tier natively, many tools look to outside policy managers and data movers to handle those tasks. See that your tool can interface with any external policy managers, migration applications or storage platforms currently in your environment. For example, a data classification tool might identify financial or Health Insurance Portability and Accountability Act data, and then move that data to an existing EMC Centera or another CAS device.

Evaluate the performance characteristics. Understand the time required to discover and work with enterprise data, and determine the maximum amount of data that the data classification tool can support. Also understand how the data classification platform handles data in terms of files and size. "If a vendor tells me they can classify 1 gigabyte [GB] per hour, that might be interesting," Schulz says. "But how many files is that [per hour]?" For example, an organization with a large numbers of small files may opt for a data classification tool that favors such behavior. An organization with a lower number of large files may do better to select a product that focuses on overall throughput.

Best practices for implementation

Regardless of how you implement a data classification tool, analysts suggest keeping a close eye on performance figures during the classification or search process, such as files per hour or GB per hour) and verify that you are receiving acceptable performance. Make sure that the product does not become bogged down under significant classification processing tasks. Poor figures, or performance that falters when the environment scales, may suggest a need to reconfigure the data classification infrastructure. Some other general policies can help you get the most from any data classification tool.

No substitute for human intervention. No tool can determine the value of your corporate data, so corporate leaders must be involved in any data classification initiative. Tools are improving, and prefabricated rule sets are increasingly common. This often eliminates the need to develop classification guidelines from scratch, but even the most comprehensive rules must be tweaked and refined for your specific business.

Avoid the urge to over classify. When properly implemented, data classification can enable efficient and cost-effective storage, but it's sometimes hard to know when to stop. Many organizations only support up to three storage tiers or service levels; usually high-performance FC SAN, some form of low-cost, high-volume SATA storage and a tertiary tier that is often tape. As a rule, classification schemes typically reflect these tiers. Applying finer levels of classification than tiers allow yields little benefit.

Don't be afraid to get help. If there isn't enough in-house expertise to address your data classification initiative, consider contracting the services of a consultant that specializes in your industry — particularly legal and financial industries. An outside consultant can sometimes help mitigate the effects of internal politics and bring a focus to the classification process that might not otherwise be possible.

Start small and build out. Companies can find data classification to be a daunting exercise, so analysts suggest focusing their efforts on a specific objective to start and then expanding the initiative in phases over time. "Do a pilot (a prototype) to address a particular business need or pain point," Schulz says. "Use it to build up support."

Data management tools: SRM packages — A closer look

Rick Cook

January 19, 2006

What you will learn from this tip: Storage resource management (SRM) is central to the storage administration in the modern enterprise. But like a lot of terms, SRM tends to conceal as much as it reveals. There are many SRM products available, with at least as many different combinations of features. This tip offers guidelines on finding the right product for your needs.

When comparing [storage resources](#) management (SRM) products you have to keep in mind which features really matter to you. While this is true of most applications, it is especially important with SRM products because SRM is a fairly fuzzy term and vendors often add features that are outside what most would consider SRM's core function.

For example, managing [e-mail archiving](#) is arguably a storage management function and a number of vendors, such as IBM (Tivoli Storage Manager), Veritas (NetBackup) and EMC (Legato NetWorker) have e-mail management features. But an SRM product with an e-mail archiving and management feature doesn't do you much good if you use a third-party e-mail archiving and management product, such as Mimosa Systems Inc. NearPoint.

Information versus management

SRM started with glorified report generators, which gathered and organized statistics on storage utilization, performance and such. Later, SRM products added the ability to actually manage storage as well as report on it. However, different SRM packages tend to have very different mixtures of information and management features.

Some packages will let you drill down as deeply as the individual spindle on an application but may not offer the sophisticated management features found in other SRM products.

In part, this is because there is a difference of opinion among storage administrators about what they want SRM to do. Some administrators want to be able to manage their storage resources from the SRM package. Others prefer to use SRM strictly for information gathering and reporting and do the actual management with other kinds of storage management tools.

Handling heterogeneous environments

SRM products vary widely in their ability to handle environments with multiple operating systems and products from multiple vendors. Products tend to trade granularity and control for openness. In other words, products tend to trade detailed reports and fine control for the ability to work in a [heterogeneous](#) storage environment.

Detail versus ease

Another continuum in the SRM world is between detail of reporting and degree of control and ease of installation and use. Some products, such as EMC Corp.'s Control Center and Veritas' Command Central, try to provide broad, deep control over storage. That tends to make them harder to install, configure and learn than products with more modest goals.

How much information do you really need?

While information is generally a good thing when it comes to systems management, there is such a thing as too much information. With the constantly falling price of storage devices and other hardware, at some point the cost of gathering the information exceeds the benefit you can derive from having it. At the present time, that is a highly enterprise-specific and subjective decision. In making it, keep in mind that the cost of gathering information is far more than the raw cost of the SRM system. It includes the configuration cost and analysis costs as well.

Data management process: Managing data migration among storage tiers

Rick Cook

August 3, 2006

What you will learn from this tip: Rick Cook outlines a number of approaches to tiered data migration.

As storage becomes more sophisticated, and [information lifecycle management](#) (ILM) becomes a fact of life for more enterprises, the data migration among storage tiers becomes more complex.

Ultimately, at least some of the data will be migrated to tape, [magneto-optical](#) (MO) or some other form of [archival](#) storage, a function normally handled by the [backup](#) software. However, before then, larger amounts of data may need to be shifted three or four times to progressively cheaper storage.

One goal in all of this is migrating data with minimum disruption. Ideally, the applications using the data should require either minimal or no changes as the data is migrated until, perhaps, it is finally archived. However, this isn't easy, especially if you want to do it automatically in response to established business rules.

There are a number of approaches to the process of tiered migration, depending on budgets and specific needs.

A number of storage management systems, including Hewlett-Packard Co.'s StorageWorks, offer file migration agents — either built-in or as add-ons — that are designed to work closely with the main management application.

For many vendors, tiered migration is part of [storage virtualization](#). Companies such as EMC Corp., (which bought Rainfinity) and Neopath Networks Inc., offer NAS virtualization front ends which can shuffle data between storage devices, and hence storage tiers, automatically. Ideally, the process is invisible to users and applications alike. Likewise, companies like Incipient Inc., offer migration as part of their virtualization offerings for storage area networks (SAN). Incipient's NSP software runs on the SAN switches to handle moving data.

A potentially cheaper approach, especially for small and medium-sized enterprises, is to keep the data on the same storage array, but shift it among [RAID](#) levels. Something like RAID-5 requires less redundancy, and hence makes cheaper storage, than RAID-10, which requires 2 [megabytes](#) (MB) of storage capacity for each MB of data. This saves the expense of additional storage devices, but usually means that the system can't take advantage of lower-cost technologies such as [SATA](#).

Data management process: The rising cost of cheap storage

Pierre Dorion
September 7, 2006

Storage budgeting tip: The commoditization of storage has been driving acquisition cost down for quite some time. However, hidden costs are now starting to surface as organizations struggle to manage hundreds of terabytes of data. This tip discusses the some of the pitfalls of “cheap storage” and offers some advice for keeping costs in check.

If I had to pick one term that comes to mind when talking about cheap storage, it would have to be procrastination. Wikipedia defines **procrastination** as “*the deferment or avoidance of an action or task which requires completion by focusing on some other action or task.*” Avoidance is exactly what a lot of organizations are opting for when faced with the growing problem of data explosion combined with the absence of data retention policies. The declining cost of storage has allowed companies to address the problem by throwing more hardware at it — much like we give a spoiled kid more candy to appease repeated tantrums. Eventually, there is no candy sweet enough and concrete actions are required... but enough server room psychology.

Declining data storage costs have allowed organizations to put off having to make decisions regarding [data lifecycle management](#) and data categorization. Some of the major resulting issues include:

- Massive email servers that are difficult to restore
- Multi-[terabyte](#) file servers that can hardly be backed up over a weekend (forget [restoring](#))
- Storage management challenges (complex environment and shortage of skills)
- Inability to meet growing data availability requirements
- Backup storage capacity requirements that are multiples of the production data
- A daunting data inventory task
- Uncertain records discovery and retrieval capabilities in the event of litigation
- Costly data and storage management solutions

It becomes clear that declining storage costs can quickly be overshadowed by other rising costs such as those outlined above. This is easily noted if you consider that overall IT spending doesn't follow the same curve as storage capacity cost.

Reducing the costs

Without trying to trivialize the effort, the first step is to start identifying what data the organization has in storage. This is an inventory effort, in which all functional areas of the business must participate. The IT department cannot identify data much beyond reporting on file types, size, the systems or applications that access the data and the last access date.

Once the data is inventoried (and this is no small task), it is up to the data users/owners to indicate how critical that data is to their daily activities or if it is still used/needed. If the data is no longer used and can be disposed of, it should be deleted now. If the data must be retained (for whatever reason), it should be taken out of the costly daily data management loop ([mirroring](#), [backups](#), monitoring, virtualization, etc.).

There are a number of products available that provide data archival or [hierarchical storage management](#) capabilities while making use of [storage tiers](#) but they all have one thing in common — products will not make the decisions for you. It must be noted that the object is not necessarily to reduce the amount of data stored, but rather to reduce the amount of data that is subject to costly storage-related processes.

Organizations cannot prevent the creation of new data or transactional records although some are taking steps to control it. For example, some companies are adopting policies to encourage use of the telephone (remember that device you use to actually talk to people) to reduce the volume of email messages. However, the true gains come from storage decisions about data that is no longer needed or used on a regular basis.

Of course, this cannot be considered a one-time exercise. To make this a cost-effective effort, the decisions made about existing data must also be applied to new data as it is being created to avoid having to repeat the same exercise down the road. It is also the foundation for [information lifecycle management](#).

Data management process: Archive or backup?

Pierre Dorion
July 18, 2006

What you will learn from this tip: Archiving, backup and their differences.

[Backups](#) and [archives](#) consist of copies of data kept for a certain amount of time for future access (at a very high level). In essence, both are very similar in nature except for their respective lifespan. Generally, most view a backup as a short-term retention copy of a file or record in case the original is lost or damaged beyond repair. Conversely, an archive is typically viewed as the means to meet a requirement to retain a record for future reference.

Theoretically, you could take a backup copy of specific data and retain the copy for many months or even years and you would have yourself an archive. In that particular context, the first question that comes to mind is: When does a backup copy become an archive? This is where things get a little muddy. While many IT practitioners associate the term archive with long-term retention, it is not just a question of time. This is mostly due to the way most traditional backup products handle data retention. Keeping track of which “backup jobs” to delete and which ones to keep can become a daunting data management task. Traditional backups are usually part of a sequence, which is typically a series of weekly full backups followed by daily incremental backups that are kept for a predetermined amount of time (i.e., 30 days). In order to keep a copy for a longer period than usual, an out-of-sequence copy must be created. That is, a copy that is not associated with the 30-day retention in our example. This is where the attributes of an archive start to take shape. We can think of an archive as an out of

sequence copy; a copy that is not associated with other copies for retention purposes (i.e., full and incremental). Let's look at other attributes that should differentiate an archive from a backup object:

- Archives should not be retained simply based on the number of existing copies. Each archive should be a unique object bearing a time stamp, descriptor and a retention parameter.
- We typically backup data to protect it from being lost or altered and because it must remain readily available; it would therefore go against the rules to delete a file after backing it up. Conversely, data is often archived so it can be deleted from its original location because immediate access is no longer required.
- Archived data can be extracted from its original context and catalogued or indexed for later retrieval. This is the case for [CAS](#) or [email archiving](#) products where a message or attachment is taken out of its usual structure and stored elsewhere.

As a general rule, we can go back to the days of paper records and draw a parallel with today's backups and archives. Back then, records were typed or handwritten and carbon copies or photocopies were used for backups. When a document lost some of its daily business relevance but still had to be retained, it was taken out of the filing cabinet, put into a cube-box and sent to some basement or warehouse to be kept as an archive. That said, this is pretty much where the similarities end. We don't have a problem reading a paper document that was archived 50 years ago — the same cannot be said about electronic archives.

In closing, and without trying to oversimplify things, if a record is copied for protection, we can probably call it a backup. If the same record is stored on some media with particular concern with immediate access, it's probably safe to call it an archive.

Data management process: Tech Report — Content-addressed storage preferred for fixed-content storage

By Jerome M. Wendt

June 19, 2006 | SearchStorage.com

[Content-addressed storage](#) (CAS) safeguards retention data and prevents its alteration.

For most companies, fixed-content storage requirements are simple: Store the data securely, do it cheaply and provide fast access. With more data subject to external and internal audits, CAS products are becoming the preferred storage medium for the long-term protection of fixed content.

CAS products come in four different architectures:

1. The [RAIN](#) architecture is the predominant way vendors offer CAS hardware. Inexpensive servers or nodes with high-capacity disk drives are clustered together; software locks the data stored on each node. As growth occurs, more nodes are added to the RAIN cluster.
2. Network Appliance Inc. (NetApp) presents a network file system over an Ethernet connection on the front end while using [WORM](#) technology to lock the data down and data deduplication to optimize its capacity. The system accommodates growth by adding more disk capacity to NAS head configurations or allowing upgrades to higher capacity NAS filers. There's no way to move data from the NetApp disk to tape or optical media.
3. The [HSM](#) architecture offered by IBM Corp. allows applications to archive and retrieve data from the CAS system using APIs provided by the CAS software. IBM requires users to deploy its Tivoli Storage Manager (TSM) 5.3 for Data Retention software, which comes with its TotalStorage DR550. IBM's CAS product differs from the other CAS architectures because it doesn't use data deduplication or single-instance storage (SIS) by default, although users can deploy these technologies and use TSM to manage the data.
4. Nexsan Technologies Ltd. offers a networked storage array architecture that includes CAS software as part of the array to manage data retention and ways to move data between disk, tape and optical.

There are trade-offs with each of these designs. Each one requires some type of software to classify and then move the data to and from the CAS device. Products using RAIN architectures—including Archivis Inc.'s Archiving Cluster (ArC), EMC Corp.'s Centera, Hewlett-Packard Co.'s (HP) StorageWorks Reference Information Storage System (RISS) and Permabit Inc.'s Permeon Compliance Store — store files as objects. This introduces a new format for storing files whose long-term management costs and liabilities aren't yet well understood. File system approaches don't support SIS and require third-party products to manage file meta data. HSM architectures are based on a model that may not respond well with large data stores, while the network storage array approach has limitations on total disk capacity and the disk it will support.

CAS features

All CAS products deliver the following fixed-content storage requirements:

- Data is accessible by content, not storage location
- Scales economically
- Manages large amounts of data
- Guarantees data authenticity and security
- Manages data-retention periods
- Facilitates rapid data recall

To deliver these basic requirements, many CAS vendors use the RAIN grid architecture. A RAIN device is usually a commodity server (called a node) with internal [SATA](#) hard drives and vendor-supplied CAS software. The nodes in EMC's Centera and Permabit's Permeon RAIN architecture support two personalities: an access or portal node and a storage node. The access nodes are clustered and connected to the Ethernet

network to receive and process either incoming files or requests for data using a number of network protocols. The access node identifies where the object is to be stored or where it resides, and then stores or retrieves the object from the storage node.

You can also preserve the integrity of data or objects by balancing data across multiple storage nodes. Archivas Inc.' ArC and Permabit's Permeon Compliance Store use this approach to enable users to deploy nodes one at a time; this allows data to be distributed evenly across storage nodes. Each object is copied and stored on at least two different nodes to prevent data loss due to a node hardware failure.

Another consideration when evaluating each vendor's RAIN architecture implementations is the type of hardware to be used. Although each vendor uses off-the-shelf Intel server hardware to host its software, Archivas' ArC and Permabit's Permeon Compliance Store allow users to choose any vendor's brand of server, while CAS vendors such as EMC and HP require users to purchase server and storage hardware from them. HP only sells and certifies its ProLiant DL380 servers as nodes to support its StorageWorks RISS software.

Users with existing server hardware or server agreements may opt for Archivas' ArC or Permabit's Permeon Compliance Store because they run on any server vendor's hardware. For firms more concerned with deploying an end-to-end configuration sold and supported by a single vendor, choosing EMC or HP for the hardware and software in a preconfigured CAS product may be a better option.

The hashing algorithms a CAS product uses to create digital identifiers for each object is also important. Some hashing algorithms may be cracked or hacked over time; having the ability to upgrade the digital signature may therefore become more important. Caringo Inc.'s CASTor, a new CAS software product, lets users upgrade the hashing algorithm and digital signature as new ones become available.

Most RAIN architectures support only nodes with internal disk drives. Only Bycast Inc.'s StorageGrid and Caringo's CASTor let users deploy nodes that support different types of external storage and manage the placement of data on these different tiers of storage based on policies set by users.

A final concern is the protocols used to access the RAIN nodes. One way RAIN vendors circumvent the [API](#) problem is by presenting a mountable file system to the operating system (OS) and allowing apps to use the more common [NFS](#) and [CIFS](#) protocols to store and retrieve data. Most RAIN vendors, including Archivas, Bycast, HP and Permabit, support this configuration, and even EMC is jumping on the bandwagon.

NetApp's NAS products use file systems, but they support CAS in a slightly different manner. By using SnapLock (an optional WORM feature) with the Data Ontap OS that comes standard with all NetApp filers, and its new Advanced Single Instance Storage (ASIS) feature, users can lock down data and optimize storage capacity on filers. The main drawback of file-system architectures is that they require either a separate appliance or third-party software such as Open Text or FileNet to classify each file, create and store meta data, and manage the file's data-retention periods and user access permissions.

IBM prepackages its TotalStorage DR550 with TSM for Data Retention software to enable apps to classify and manage data. (Shops already using TSM can host the Data Retention component on an existing TSM

server.) For small- to medium-sized firms, IBM offers DR550 Express, which also ships with the TSM software, but supports only internal disk drives with an option for tape vs. the DR550 that supports external disk and tape and is available in clustered configurations. TSM is required to manage data placement, retention and security policies; all host apps will need to support TSM's APIs to store and retrieve data.

IBM prepackages its TotalStorage DR550 with TSM for Data Retention software to enable apps to classify and manage data. (Shops already using TSM can host the Data Retention component on an existing TSM server.) For small- and to medium-sized firms, IBM offers DR550 Express, which also ships with the TSM software, but supports only internal disk drives with an option for tape vs. the DR550 that supports external disk and tape and is available in clustered configurations. TSM is required to manage data placement, retention and security policies; all host apps will need to support TSM's APIs to store and retrieve data.

Data management process Expert: E-mail archiving with CAS

Bill Tolson

April 24, 2006

I'm considering designing an e-mail archiving infrastructure using CAS. Could you outline some of the benefits of this approach?

CAS, or [content-addressable storage](#), is an object-oriented storage model which provides the user the ability to store objects completely separate from the logical location. One of the main benefits of CAS is its ability to store a "single instance" of a given object. This capability reduces the need for disk space.

For example, say an HR department sends out an e-mail to all 1,000 employees announcing the company picnic and attaching a 1 megabyte (MB) jpeg map of the location. Depending on the e-mail system, when this e-mail and attachment is saved by employees, you potentially could have tens to hundreds of copies if this 1 MB e-mail consuming hundreds of MBs of storage. A CAS storage device would ensure only one copy of the e-mail is saved.

Many [e-mail archiving](#) products already in the ability to store single instances of messages in their archive. Before purchasing a CAS storage device for e-mail archive storage, you should double check with your vendors to see if they already offer it in their product.

Data management process: Out of capacity — What now?

Greg Schulz

April 6, 2006

What you will learn from this tip: This tip discusses how to avoid running out of storage capacity by evaluating your options, investigating re-allocation and knowing where to turn if you can't solve the problem yourself.

Running out of storage capacity is similar to running out of gas in your car — both are impacted by usage and are caused by ignoring indicators and gauges.

Storage and performance thresholds and monitoring tools are equivalent to your cars' dashboard indicator lights. If you ignore them long enough you'll find yourself running on borrowed time and risk running out of resources and being stranded. Unlike your car where you can call AAA for emergency roadside assistance, I'm not aware of an on-the-spot, 24-hour data center service for emergency storage capacity replenishment. Granted, some storage vendors and product providers have creative pre-positioning and marketing programs where you can access storage on-demand; however, what happens when those are exhausted?

One of the first things you need to do is identify if you are completely out of storage capacity, or if you have reached a quota for your particular storage allocation. You may not be completely out of storage capacity; you may be running on reserve capacity. It is important for you to understand how your system's performance and stability are impacted when using reserve capacity. (Warning: If you ignore the indicators and use up your reserve capacity, you may be stranded. Last I checked, AAA did not have spare gigabytes and terabytes for emergency assistance.)

In the short term, investigate whether reconfiguring your storage allocation will help. If you can, remove previously backed up and unused files. Reconfiguring and removing files may be disruptive and you have to identify what files can be removed. Also look to see what databases can be pruned, purged and compressed to free up space or if any temporary and maintenance space can be freed up. Look to see if you can borrow storage capacity from other systems or applications or from your storage vendor. Some vendors offer [storage-on-demand](#) programs where storage is physically pre-configured onsite and you are charged when you use it.

Do you know how long the storage capacity shortage is expected to last and is it expected to get worse? If there is no more physical storage capacity onsite, how quickly can a vendor ship and install new storage to you? Keep in mind that once the storage is on your premise, the storage needs to be setup, configured and allocated to applications, all of which take time and can include possible disruptions.

Do you know how quickly you can obtain (purchase, rent or lease) more storage capacity and do you have a budget that will allow you to acquire storage? Will your environment support the addition of more storage, and do you have capacity in your existing storage systems to add disk drives, available switch ports or adapter ports to attach storage and what will the cost to upgrade applicable software license be?

If you do not have performance monitoring and resource usage tools (also known as [SRM](#)), you'll need to get some, as you will need these to be proactive and to react to future storage capacity shortages. EMC, HP, IBM, Abreivity, StoredIQ, SofTek, Tek-Tools, 3PAR and others have tools that can help identify what resources are being used. These tools can be used from a planning perspective and to react to capacity shortages, helping to identify and move data to free up capacity.

Start your spring cleaning to free up disk space and establish capacity and performance threshold alert indicators. When it comes to storage capacity threshold rules of thumbs, you will get plenty of different responses — some based on old myths that range from keeping usage under 50% to 90% of capacity.

To avoid future capacity shortages and outages, put a storage capacity plan together. A storage capacity plan can consist of a simple outline of roughly when and how much storage you will need. If you are not sure how to put a forecast together, ask your storage vendor or product provider for help, or talk to third-party analysts and consults for advice. Check out the [Computer Measurement Group](#) (CMG) as a source of information for capacity planning. Also refer to the SearchStorage.com tip [Balance costs and demands for proper storage allocation](#)."

Data management process: Thin provisioning slims storage volumes

Alex Barrett, Trends Editor
January 26, 2005

At issue: Tired of listening to your storage and applications people fight over application capacity? NetApp, DataCore and 3PAR are offering auto-provisioning features that might help.

It's an old story: Manual [provisioning](#) of storage is a source of tension between storage and application folks, and contributes to the poor capacity utilization that persists in many [SANs](#).

Those on the application side usually ask for more capacity than they need, says Craig Nunes, senior director of marketing at SAN array maker 3PAR, Fremont, Calif.

"The application guy thinks, 'If I get 500 GB, I'll never need to take down my application,'" says Nunes. Storage admins, meanwhile, try to put off unnecessary disk purchases and push to allocate less than they're being asked for. In the end, the application usually wins out. In Nunes' experience, applications typically get 75% more capacity than they use.

To this end, 3PAR users can license Thin Provisioning, which presents an application with a virtual volume of arbitrary size, but only allocates physical capacity to it as data is actually written. LeftHand Networks also supports this concept, says Tom Major, the firm's VP of marketing, although it doesn't have a name for it.

Last month, DataCore announced an auto-provisioning feature for users of its SANmelody software, which works across DataCore Networked Managed Volumes on heterogeneous [Fibre Channel](#) and IP-based SAN arrays.

Network Appliance's new FlexVol feature, part of its new Data Ontap 7G operating system announced in November, is also a thin provisioning product, says Mike Fisch, director of storage and networking at Wellesley, Mass.-based Clipper Group. Thin provisioning, he says, is "anything that presents a volume bigger than the capacity it uses and can then fill that capacity."

For its part, EMC is working on what it calls "oversubscribed volumes." A company spokesperson says customers "are interested in this feature because it has the potential to address two issues: reducing storage costs by increasing storage utilization, and minimizing the re-provisioning effort." EMC plans to include the technology across its portfolio, including Symmetrix, Clariion, Celerra and Storage Router.