

Optimizing Data Retention and Archiving

November 2007

*Calvin Braunstein, Executive Director of Research
Mimi Ho, Senior Research Analyst
Robert Frances Group*



120 Post Road West, Suite 201
Westport, CT 06880
<http://www.rfgonline.com>

Executive Summary

Archive data needs to be managed, retained, and protected effectively, and then disposed of properly when it is no longer needed. To be effective, IT executives must ensure that a combination of data retention policies and a clear media strategy is developed. In addition, long-term retention data needs to be migrated to newer media and associated technologies, an important consideration seldom considered in total cost of ownership (TCO) analyses. A world-class strategy for archiving encompasses all forms and needs across the enterprise, with a scalable set of technologies that provide a common solution that addresses both data retention and space management needs at a low cost. Archiving indeed encompasses both data retention and space management goals, and the choice of whether to remove archived data from operational data stores is an important one that can reduce cost and minimize staff overheads.

Within the strategy, there are two key decision points that represent the focus of this paper. First, there is a choice of whether tape should play a role within the storage solution, and second, a basic choice of how aggressively to archive data. The Robert Frances Group recommends that clients proceed aggressively in that tape has great value in a storage hierarchy and aggressive archiving has significant benefits and reduces costs. For example, the more data is moved to archive, the less the amount of daily, weekly, and monthly operational backup processing (assuming it is moved from operational data stores), an issue for many in the context of space management and overnight processing windows. Use of tape can represent a significant cost savings over disk-only solutions, and has potential other benefits as well.

The major reason for pursuing an aggressive archive strategy is drawn from research showing that 75 to 99 percent of all archived data will never (or seldom) be read, varying by industry of course. Although only a small amount of stored data will be accessed after the first few months of its life, customers do not know which records will be retrieved and which will not be retrieved. As such,

all data tends to be kept on the floor "just in case" (JIC) someone or some authority might need it.

Storing JIC data on expensive storage, such as high-speed online storage, can be cost prohibitive. Enterprise executives should therefore ask themselves how much they should be spending to manage and store for long periods of time large amounts of data that may never be used again. In other words, what is the appropriate amount to spend to have an effective archive solution that meets business requirements, but provides reasonable recall capability when the data needs to be accessed? Enterprise executives must also ask themselves how exactly they will achieve this. Should they employ disk, optical, or tape media, or a combination of all three?

What is needed by most companies today is a cohesive data media strategy that offers a full range of price/performance options. These options must match the diverse range of business requirements that span the full data life cycle. By developing a strategy, technology integration, scalability, performance, and cost requirements begin to drive proper decision-making when it comes to putting data on the most cost-effective media, and building in technology refresh and data migration capabilities as needs and technology evolve.

The Robert Frances Group performed an analysis to assist clients in their decision-making. Cost models included in this study reflect the cost differences between two key things: the cost of a disk-only archive approach versus a disk/tape approach, and the cost differential recall rate increase (the typical assumption is 1 percent). The results show conclusively that leveraging tape can be significantly less expensive, and that the IBM System Storage DR550 selected for this analysis offers a built-in migration and technology refresh capability that mitigates those cost issues. (No media migration costs were included in the models). In the analysis of the recall rate, a 10 percent recall rate was compared with the baseline 1 percent model, representing what would be a more aggressive archive strategy that might result in higher recall rates. The result showed a minimal increase in costs, warranting consideration of a

more aggressive data archive approach and strategy.

Introduction

Although the terms backup and archive are used interchangeably by some, they describe different processes with different goals and results. Generally speaking, backups protect against hardware outages, independent of the reasons for archiving, and store data for future use in the event that the original data is erased or corrupted. As databases continue to grow, this not only increases the amount of disk storage that needs to be assigned to the database, but also slows the backup and restore process. Moreover, it increases the risk of erasure or modification to business-critical data.

Archived data is different in that it is not a backup or recovery copy. This data can be both regulated and non-regulated, and is stored for multiple years, not weeks or months like backup data. The other factor to be considered is reference data, like e-mails that need to be moved to archive once they are processed. In such cases, there is no need to “age” such data, and it may or may not be appropriate to leave it in the live data stores. User policies will determine whether data is kept on the production floor or not, but compliance-based data retention and space management are both opportunities when the technology can rapidly find and recall needed data. These two aspects of data protection and management, including peripheral issues like data integrity, security, performance, and others represent a major opportunity for most clients to reduce costs and simplify operations by leveraging technology to address them all in a cohesive manner. Indeed, the more aggressive the archival policies, the more backup, security, integrity, and performance issues are improved.

Enterprise executives must also keep in mind that archived data typically outlives the media on which it is stored. Depending on the length of time involved for saving data (usually 7 years for most types of data in the U.S. and 10 years in Europe, Middle East, and Africa (EMEA)), there may be several migrations for each piece of data.

However, some data must be kept for 30 years or more, even if never touched again.

This means that an organization will have to repurchase terabytes of storage at least two to three times during the life of that data. Tape drive technology generally lasts about six years, while disk has a three-year technology refresh rate. In addition, each time the media is replaced, a data migration is required. Thus, organizations need to factor in the costs of storing data over the life of the data retention period, and not just the initial hardware and software costs of implementing the data archive. And, the more media turns, the more important it is for enterprises to make sure that the technology they are purchasing has built in migration capabilities.

The Value of a Holistic Data Management/Protection/Archival Strategy

Data has become central to the functioning and survival of the enterprise. As enterprises grow, much of the data they generate comes to reside on various storage devices. Threats and attacks against enterprise storage facilities can affect customer confidence and, ultimately, enterprise survival. And, as the importance of data grows, regulations governing its use have continued to proliferate in an attempt to help protect the privacy of sensitive data residing on such systems. The sheer size of live data has become a major problem for many clients, stressing the ability of storage administrators to perform the daily, weekly, monthly, quarterly backups before the next business day. An aggressive archival approach can help reduce the size and complexity of live data, once it is archived and removed from operational data stores.

Despite the criticality of enterprise data, many organizations still have multiple yet separate ongoing initiatives around such data. These include disparate strategies for backup, archive, media migration, and compliance with local, regional, or federal regulations. In the context of archive, most companies have multiple levels of requirement, based both on regulations and

unique requirements within each line of business (LOB). Enterprise executives should strive to build a singular storage strategy that ties all of these pieces together into one composite plan. In addition, they should leverage technologies that facilitate and support this consolidation effort.

By bringing backup, archive, media migration, security, data integrity, performance, and compliance processes together, the enterprise can realize numerous benefits, including the following.

- Cost savings related to both the administration of storage systems and any legal liabilities resulting from improper data retention policies. Elimination of older, outdated, long-term storage technologies such as optical and jukebox
- The ability to leverage data across different business units, thereby streamlining and improving operations
- The connection of disparate devices, allowing the enterprise to architect systems independent of where the data is stored or even the type of media being used
- The foundation for deployment of a service-oriented architecture (SOA) within the enterprise

Enterprise executives should also consider building a data and data media center of excellence (COE) that is tasked with developing this cohesive strategy for data management to ensure the realization of these benefits. A COE provides a formal business structure necessary for aligning enterprise processes and operations with business strategies. The COE also enables enterprise subject matter experts to develop and maintain best practices and operational excellence for the enterprise. The COE can be viewed similarly to a franchise model that leverages standard, repeatable practices/processes to drive consistent, high quality brand equity throughout enterprise organizations. Establishing a COE, especially in highly decentralized and/or global companies, can help deliver enterprise standards and consistency across the organization for developing efficient products and services to business constituents. For decentralized organizations, the COE would be virtual,

implying that (at least some) staff have dotted line reporting into the COE.

Whether formalized as an on-going part of the IT organization or not, those in charge of developing the data management strategy should consider not just the day-to-day functions of managing storage. They should also address how specific tactics can coalesce and contribute to more comprehensive, enterprise-wide storage initiatives. Although the COE is responsible for creating and documenting the standards for data management, it is also accountable for ensuring that established standards and policies are set across the enterprise, and adhered to by LOBs. Executives should carefully evaluate the enterprise environment to determine the most appropriate placing of the COE. The COE must have proper authority for defining processes and directions, and must necessarily be supported by senior executives.

The COE should also be responsible for determining how to improve and facilitate the flow of information to aid service oriented architecture (SOA) efforts. Although data and information are frequently used interchangeably, there is an important distinction between the two. Data itself, or the formatted 1s and 0s, has limited value if it cannot be interpreted into meaningful values, or information. Organizations that save data just to save it will not realize the true value of that data. The enterprise must have the decryption keys or applications that make sense of the 1s and 0s, and the incumbent operating system and tape drive must recognize the format in which the data was saved.

Thus, being able to render data back into information is a critical part of data retention solutions and is vital to SOA operation. SOA aims to link resources throughout the network, and to make such resources available to users as services with standardized access. Yet, such a goal cannot be achieved without current storage systems being properly linked, data being consolidated, and information being accessible.

The DR550: IBM's Solution to the Data Archive Quandary

The IBM System Storage DR550 solution is designed to provide a fast access platform for archiving and retrieving data using policy based data retention of data and options for retaining data in a non-erasable, non-rewriteable technology. The benefit of such a solution is that it can help organizations manage and secure regulated and non-regulated data from accidental or intentional erasure or modification. The DR550's tiered storage architecture also enables customers to set policies that allow the migration of data to significantly lower cost media, such as tape, as the need to access that data decreases over time. Additionally, customers can use the DR550 as a primarily disk-only solution (like competitive offerings), with the option to use tape for data or files that make the most sense. Hence, IBM has made it affordable for companies to archive JIC data, and then dispose of the data when a retention period has ended.

IBM DR550 and DR550 Express solutions are designed specifically to help address the requirements of effective long-term data retention and protection in a cost-conscious manner, according to IBM. Both offerings provide the same functionality, although the DR550 Express has less storage and a lower entry price. The DR550 Express is intended for small to medium-sized businesses (SMBs), and the DR550 is intended for mid to large enterprises that have higher capacity archive requirements.

The components of both solutions are standard IBM hardware offerings that have been integrated within a lockable rack, and pre-installed program code with customization to provide additional security. The hardware is pre-installed, pre-configured, and pre-cabled into one or two standard racks. As for the program product components, they are preloaded to minimize the customer's installation time.

The following charts provide an overview of the DR550 and DR550 Express elements.

IBM System Storage DR550 at A Glance	
Nodes	<ul style="list-style-type: none"> • Single- and dual-server configuration options • Based on IBM dual-core System p5 POWER5+ 2.1-GHz processor platform <ul style="list-style-type: none"> ○ Fibre Channel adapters ○ Ethernet adapters • Optional DR550 File System Gateway Hardware <ul style="list-style-type: none"> ○ Provides NFS / CIFS interface to applications and stores data in DR550
Storage controller	<ul style="list-style-type: none"> • One or two IBM System Storage DS4700 SATA storage system controllers (each controller comes with 12 TB of storage and can support up to 72 TB of additional storage) • 0 to 12 IBM EXP810 Storage Expansion Units, each with 16 500 GB or 750 GB SATA disk drives • Optional Metro Mirror or Global Mirror for replication • 2005-B16 Fibre-Channel switches • IBM 7014 rack model T00 <ul style="list-style-type: none"> ○ Rack locking feature ○ Additional power distribution units (PDUs)
Software	<ul style="list-style-type: none"> • DS4000 Storage Manager, • IBM AIX 5.3 • IBM HACMP (dual-server configuration)

	<ul style="list-style-type: none"> • IBM System Storage Archive Manager (SSAM) • IBM System Storage DR550 File System Gateway Software (optional)
--	---

Source: Robert Frances Group and IBM Corp.

IBM System Storage DR550 Express at A Glance	
Nodes	<ul style="list-style-type: none"> • Single server configuration • Based on IBM dual-core System p5 POWER5+ 2.1-GHz processor platform <ul style="list-style-type: none"> ○ Ethernet connections ○ Flat panel monitor • Optional DR550 File System Gateway Hardware <ul style="list-style-type: none"> ○ Provides NFS / CIFS interface to applications and stores data in DR550
Storage controller	<ul style="list-style-type: none"> • Eight 146 GB, 10,000 RPM, Ultra-320 SCSI drives with just over 1 TB of capacity • One optional IBM System Storage DS4700 SATA storage system controller with up to 12 TB of additional storage
Software	<ul style="list-style-type: none"> • DS4000 Storage Manager • IBM AIX 5.3 • IBM System Storage Archive Manager (SSAM) • IBM System Storage DR550 File System Gateway Software (optional)

Source: Robert Frances Group and IBM Corp.

Prices for the DR550 and DR550 Express include the noted pre-installed hardware and software. The list price for the DR550 Express is starting at approximately (U.S.) \$25,500. Organizations can order the DR550 with 8TB, 12TB, 16TB, 24TB, 32TB, and 48TB of "raw" disk storage, with upgrades being done in 8TB or 12TB increments of up to 168TB of raw storage capacity. The DR550 Express can support up to 13.1 TB. Both the DR550 and DR550 Express can support petabytes of tape storage. Prices for the DR550 start at approximately (U.S.) \$94,000 (list). Here again, a cohesive strategy enables planning for a wide variety of needs using singular or a few strategic technologies to accomplish it. In this regard, it is important to remember that IBM hardware and software become cheaper as volume increases, unlike some of its competitors. Thus, a lower unit cost of storage (either disk or tape) can be achieved.

A clear advantage that IBM has over its competitors lies in the way that the DR550 is designed. Specifically, the design enables DR550 to use available modular components for upgrade

flexibility. This is critical with archives of 10 to 20 years or more. This means that upgrading to newer technologies will be relatively seamless, and coupled with the built-in media migration capabilities, provides a fast and easy way to manage data as it moves from one technology to another. In addition, the built-in servers that drive the storage are also driving the storage management software, a real plus for simplifying operations and leveraging that capacity.

With support for over 400 secondary and tertiary storage devices, the DR550 enables enterprises to leverage their existing storage investments. For example, an organization could attach to the DR550 supported free-standing optical libraries or a tape library that can be enabled for re-writable tape media or "Write-Once-Read-Many" (WORM) tape media. Thus, the DR550 offers enterprises significant flexibility in choices for archive technology, and at a lower cost than traditional optical and disk storage.

Having data on disk, optical, or tape, has little relevance to the ability of an organization to

respond to a legal request to data. Although disk storage providers will argue that the enterprise needs disk storage for such rapid data retrieval, lawyers can usually sometimes negotiate the scope of such a request. Data is not turned over to authorities in 24 hours; it is turned over 24 hours from the agreed-to discovery time. Even with the latest legislation that steps up the response requirements, the real difference in retrieving data from tape in lieu of disk is usually only a matter of a few hours at most. This realization is extremely important. Previous IBM calculations have shown that a tape solution costs approximately one-sixth that of a disk solution. Thus, if low cost is a high priority, it is logical for the organization to invest in tape. Without a doubt, disk offers quicker access to data than tape. IBM offers a solution that blends the best of both these worlds – fast retrieval from disk and cost-effective storage on tape. With the DR550, customers have a hybrid disk/tape offering. This solution also includes the management software costs in the overall price.

Yet, it should be noted that the DR550 is a component of a full, end-to-end archiving and compliance solution. Key to managing the data stored on disk, optical, and tape are applications for e-mail archiving, content and records archiving and management, and database archiving. IBM has forged relationships with several independent software vendor (ISV) technology providers to bring these capabilities to customers. Through these relationships, and coupled with acquisitions such as the recent FileNet Corp. acquisition, IBM can offer its clients a storage solution for retention that leverages first-rate software to find and sift through data to help interpret it into valuable information

Backup versus Archive: Understanding the Differences and Relationships

While the DR550 is not intended to be a backup solution, a cohesive storage strategy helps identify relationships of the various processes around data retention, management, and protection. Backup is

a process that enables recovery of lost or inaccessible live data. Backups are a major effort for most, especially in the context of mapping data to applications and maintaining proper versioning. With many companies experiencing data growth rates in excess of 50-percent compound annual growth rate (CAGR), the ability to perform the needed backups within the allotted window has become very difficult indeed. The simple point here is that backup does not have to be performed for archived data, since it is no longer live and has already been stored and protected on a long-term basis. By taking a more aggressive approach to archiving, the amount of data requiring backup decreases, thus simplifying operations and reducing the amount of backup work (and associated processing capacity and personnel) required.

Data loss prevention encompasses both archive and backup processes, and is a top priority for both corporate and IT executives, directors, and managers. Backup and archive operations are often performed by the same staff, and sometimes in the same job execution. Reinventing the technology approach to archiving should at least be considerate of how the backup technology will evolve. Rapid technology advancements are a major enabler to addressing the issue, which means that IT executives must plan on frequent "refresh cycles" for the associated backup and archive storage technologies.

From a process standpoint, introducing double-checks and verification/sign-off checklists for both backup and archive work can greatly help as well. The human factors will always be a vulnerable area, but minimizing the amount of active (versus archived) data, and categorizing data so that additional precautionary steps are taken to protect what is most critical, are other best practice approaches for reducing risk. IT executives should view the menial tasks of backup and archive (including such things as transport to off-site facilities) as critical to their success, since data loss can severely impact both an individual's and the corporation's brand image.

The problem with aggressive archive, however, is the lack of understanding about data access requirements. Few IT organizations, or even their

customers in the LOBs, really know the life cycle characteristics of data, which is necessarily unique to each LOB and application. This lack of understanding leads to a reluctance to archive, especially when the technologies employed are cumbersome to manage, use, and support. If data is archived and then requires access, it can take an extended period to find and restore on older technologies that do not offer object-based restore capability. It can also add significant cost if data is moved too early into archive. However, TCO data shows that the DR550 can mitigate both issues.

With technologies such as the DR550, object-based storage and retrieval capabilities can help simplify data management and use, and enable a more aggressive archival strategy. Many clients have been looking for alternatives to the outdated jukebox and optical technologies, which can effectively be consolidated within the DR550 or like technologies. As this data is combined into a single box with common management software that is policy-based, the issues of scheduling archival processes are also mitigated. The ability to define policies for archive, coupled with the ability to restore objects that require access after the data has been archived, suggest that a more aggressive archival policy is in order.

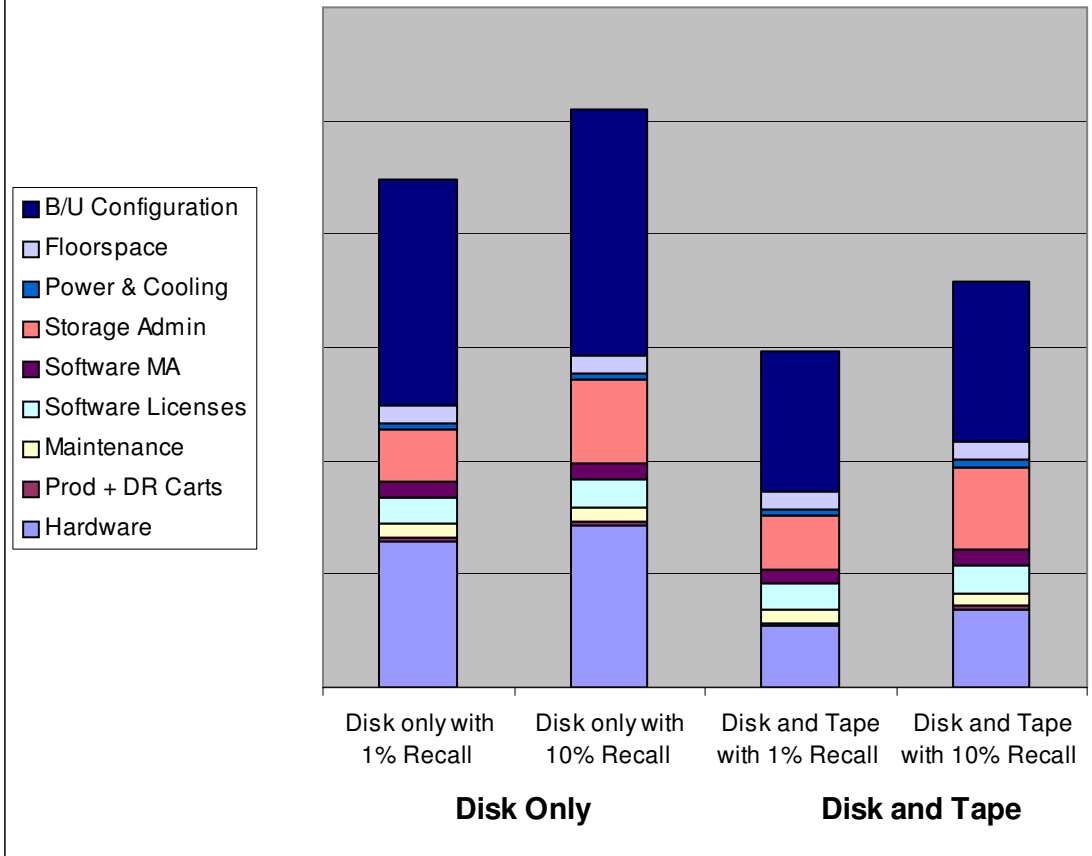
TCO Study

The IBM DR550 TCO Model allows customers to perform meaningful calculations as it examines the life of the data to be stored and the cost of its retention over long periods. Various user parameters are included to reflect real-world implementation requirements. There were four model iterations used in this study, two with 1 percent recall rates and two with 10 percent recall rates. The 10 percent recall is intended to reflect the “price” of aggressively archiving data that later must be recalled. The point here is that the additional costs of recall are lower than most might think, suggesting that premature archive can reap benefits that outweigh the higher cost. The key variables in the model include:

- Recall rate – The amount of data that is expected to be recalled after archive. The standard recall rate assumption is 1 percent, but a run was made using 10 percent recall rates for both the disk-only and disk/tape scenarios to reflect the cost of higher post-archive access needs (inherent in a more aggressive archival strategy)
- Data retention period – Assumed to be 7 years in this analysis
- High performance access duration – How long data will stay on high performance disk (assumed to be six months)
- Contingency capacity – The amount of excess capacity margin in place to address unplanned and sudden requirements (assumed to be 10 percent in the model iterations)
- Disk and tape cartridge utilization – Assumed to be 70 percent and 85 percent, respectively. This represents the maximum usable capacity (versus raw) for the technology in question
- Initial load requirements – This reflects the total space requirement for the data initially being archived. This is assumed to be 5TB, which represents both new data and conversion of older archive data from other media. To accurately reflect the aging for older data, there is a sub-parameter that defines how old the data is at load time.
- Technology refresh rates (and costs) – Defines how often technology is replaced. For disk, this is assumed to be 3 years, and for tape, six years
- Tape technology choice is another option, and LTO4 tape was used for costing
- Storage administration staffing – Reflects the costs and utilization of staff in supporting archive operations. For this model, it was assumed that a single administrator can manage a maximum of 5TB of data

Numerous other parameters exist, including technology price/performance improvements and future unit cost projections. The analysis period is another key parameter, which was set to 12 years in this analysis to reflect full life cycle costs throughout the data life cycle.

IBM DR550 Scenario Comparison



Source: Robert Frances Group

As you can see, the model reflects costs for all aspects of acquisition, maintenance, usage, and support. The model also addresses environmental factors such as power and cooling issues, which increasingly drive total costs, as well as nearly 40 percent of costs addressing disaster recovery backup configurations. Technologically, as the size of the DR550 hard drive drawer increases, power and cooling variables will remain the same. Thus, as the size of the physical footprint requirement shrinks, costs will decrease. Use of tape in an integrated data retention system has obvious pluses.

A TCO comparison of IBM versus EMC (or other competitors) was not performed in this analysis,

but instead, the IBM disk-only solution is meant to reflect typical pricing for such offerings in contrast to a hybrid disk/tape solution. Notably, beyond IBM's DR550 offering, the competition is featuring disk-only solutions, with the notion that tape is becoming a niche solution. Note also that actual dollars are not displayed in the model, but were run using list prices to show the relative differences between the four scenarios.

The results are quite compelling. Using the disk-only scenario to represent the base case (100 percent), the differences are as follows:

- The disk/tape scenario (contrasting tape vs. non-tape) costs only 66.2 percent of base case.

- The disk/tape scenario with higher recall rates (10 percent, representing a more aggressive archive strategy) is still cheaper than the disk-only base case – 79.8 percent.
- The disk-only scenario with 10 percent recall adds only 13.7 percent to the disk-only base case, a modest increase that suggests adoption of more aggressive archive policies.

Additionally, the storage requirements tied to archiving must be understood in the context of archival process maturity. Lack of a clear set of archive policies, and ultimately, lack of a clear archive strategy, can lead to wasted capacity, unplanned crises to meet legal or other archive requirements, and misuse of staff and infrastructure resources. Companies that are just starting to archive more aggressively usually require large growth increments during the first two years, then experience a more predictable growth rate. Many clients assume a 15- to 25-percent growth rate in steady state mode, which is often low in the context of new opportunities and requirements tied to legislative regulations. The combinatory effect of growth, recall, and multiple copy requirements can mean an additional 20 to 30 percent more capacity than originally planned.

Summary

Regulatory and other retention requirements continue to increase, leaving many clients in a quandary over how to address them. Because of the tactical orientation among most storage managers and administrators, however, each requirement is often addressed on a one-off basis. This leads to non-standard, expensive, and hard to maintain solutions. In contrast, combining these needs and providing a common solution will clearly provide a better solution that can be leveraged and will encompass all data protection, management, and archival needs. Failure to do so

compounds the already formidable issues of backup, archive, and improved access to data.

Developing comprehensive life cycle data models for every file, database, and application is the ideal, but requires that IT and business collaborate on defining and refining how data is used, accessed, and stored. This is unlikely to happen in the foreseeable future. Thus, some assumptions must be made about data life cycles to press forward with an aggressive archival strategy, which will be increasingly mitigated by newer technologies such as the DR550.

Ongoing legal, audit, and regulatory requirements will continue to drive IT groups to improve archive policies, processes, strategy, and efficiency. The choice of which technologies to use will have a profound impact on the success of such efforts, since technologies like the DR550 embody many aspects of the strategy, processes, and policies that must be decided upon. When it comes to tape, IBM's DR550 is unique in providing that support. Competitors tout disk-only solutions as the wave of the future, but research indicates otherwise. The most basic benefits are cost and mobility, and despite the various vendor proclamations to the contrary, tape is still only a fraction of the cost of disk and will remain so in the foreseeable future.

When it comes to more aggressively archiving, the fear of having a "just in case – JIC" requirement has prevented many companies from doing so. With technology that can quickly restore data on an object-oriented basis, and with mounted cartridges to provide reasonably fast access to tape media, both aggressive archive and use of tape are sound business practices. Enterprises that architect a scalable disk/tape archival solution that incorporates the appropriate policies and processes will go a long way towards implementing a world-class storage strategy that reduces cost, improves efficiency, and meets compliance requirements.

Copyright © 2007 Robert Frances Group, Inc. All rights reserved. 120 Post Road West, Suite 201, Westport, CT 06880. Telephone 203/429-8950. Facsimile 203/429-8930 www.rfgonline.com. This publication and all publications may not be reproduced in any form or by any electronic or mechanical means without prior written permission. The information and materials presented herein represent to the best of our knowledge true and accurate information as of date of publication. It nevertheless is being provided on an "as is" basis. Reprints are available for a nominal fee.