OSTA.org
Optical Storage Technology Association

Critical Data Retention: Archiving Reference Data and Business Critical Information

Optical Storage Technology Association
19925 Stevens Creek Blvd.
Cupertino, CA 95014 USA
Phone: (408) 253-3695
Fax: (408) 253-9938

Enterprise data center operations focus not only on present and future concerns, but on the past as well. More and more business drivers are demanding that the data center devote time, planning and resources to the creation and maintenance of information archives.

An archive in the data center is a digital repository meant for business records that either should or must be kept over time. Relevant data objects (files, databases, graphic images) are migrated outright from primary storage to that repository.

Archiving describes the process of consolidating and migrating data, typically from Fibre Channel disk arrays or enterprise Serial Advanced Technology Attachment (SATA) arrays, to less-costly nearline or offline storage media. In some cases, such as when archiving for compliance, archiving emphasizes data longevity and authenticity, especially for databases, emails, IMs, document files, or other semi-structured or unstructured data.

Preserving Business Records

The information retained in archives is, ideally, useful business records. Even before making the technology decisions, policies must be in place that defines what is to be kept and what is not. There is always a temptation to say, "Archive everything" to meet business demands. But this deceptively simple approach is not the best way, since it wastes storage space and forces IT managers to spend extra time sorting through irrelevant file data to find the required business records.

What is a business record? Probably the best definition is the one drafted in 2001 and set in ISO 15489[1]. In the standard, a record is "information created, received and maintained as evidence and information by an organization or person, in pursuance of legal obligations or in the transaction of business."

The definition can be applied to the whole range of data types: structured (such as databases), semi-structured (such as email) and unstructured (text files, graphics files). Each data object must be weighed against a policy and either is preserved or discarded.

Another temptation in the data center is to proclaim "But we already maintain backups, on which we've spent time, money and manpower to protect data." This is, of course, true, but backup is not archiving, and the two should not be confused.

Backup is Not Archiving

A Byte and Switch Insider report on email archiving addressed the differences between backup and archiving[2]. According to that report, backup involves making point-in-time copies of data to protect against hardware failures or catastrophic data loss. Backup typically reaches not only transactional or operational data, but operating systems and applications packages as well. The life expectancy of backup volumes is only a matter of days, at which point they are replace by new incremental volumes.

Archiving, in contrast, requires fast file-level access to data and should involve search and retrieval software and indexed repositories. Where

[1] *www.iso.org/iso*
[2] *www.byteandswitch.com/insider*

backup volumes are usually kept for days before being replaced by new backup volumes, archived files can be kept for years or even decades. The contrast between the handling of archived data and backed-up data is even simpler. In backup, the IT staffer is copying data. In an archive, the data is being moved, perhaps leaving a stub file or an index reference behind.

Although the data looks like it is on a disk, it is actually elsewhere on less expensive, longer-term media. In part due to the confusion between backup and archiving, some users tend to use archiving only in a niche in their environments and take less advantage of it than they should.

Why Archive At All?

The business community is increasingly operating in an environment of increased accountability, both public and private. This accountability, demanded by boards of directors, customers, vendors, regulatory authorities and the courts, demands a more structured, well-considered approach to the management of reference data. Reference data is the term applied to business information that, although sought and restored less frequently than transactional or operational information, retains business or evidentiary value over time.

Effective data retention through archiving benefits a company or a business unit by storing and promptly recalling information needed for corporate governance and decision support, cost-efficiency in storage management, regulatory compliance, and electronic discovery in the event of litigation. These drivers are examined later.

The Storage Hierarchy for Archiving

The heart and soul of archiving is the storage repository. Using the right technologies, usually in a tiered format, should yield cost effectiveness, good performance and long-term data integrity.

Much has been made of the continuous struggle between vendors of tape technology and magnetic disk technology for domination of the data lifecycle. But when weighing various technology deployments, diligent consideration should be given to optical disc technology. While optical disc dominates in consumer formats such as CD and DVD, optical disc has a significant role to play in the data center and the information lifecycle.

Magnetic disk drives are commonplace in primary storage, handling transactional and operational data that changes frequently. The information on magnetic disk drives is frequently accessed and experiences the most change in a short time. Installation is fairly easy, and the raw cost per gigabyte is comparatively low.

In the enterprise and in the midrange as well, individual drives are typically only an element of a disk array purchase. High-availability disk array subsystems, running one of many RAID levels is much more common in requests for quotation. Arrays of magnetic disks can be costly, depending on the controllers and other components. Experienced manpower is important to manage RAID arrays, providing soft cost considerations to the investment.

Hard disk drives (HDD)are the greatest power consumers in any storage hierarchy. The power demand of constantly spinning drives can send the cost of spindles soaring. Steadily rising energy costs have been the rule for quite a while, and

customers are paying more for power every year. IDC suggests that by this year (2008), 50% of IT managers will not have enough power[3]. This statistic is saying more than power bills will be high; it goes to keeping the data center up and running to meet essential duty cycles.

Greg Schultz, founder and analyst at StorageIO, reports that, in the August 2007 EPA report to Congress on energy usage in US data centers, during 2006 IT data centers consumed about 61 billion kilowatt hours of electricity at an approximate cost of about $4.5 billion. Also reported is that IT data centers, on average, consume 15-20 more times the energy per square foot than a comparable office building.

StorageIO's Schultz also observes that the major power draws in the data center, according to his research for commonly deployed storage systems, are spinning HDDs and their enclosures, which account for on average 66-75%. The balance is power draw by controllers.

Magnetic disk drive vendors have lately claimed that disk drives have a manifest destiny to dominate the information lifecycle from primary creation to the reference data archive. The vendors accurately point to low costs per gigabyte and high performance in recovery of data in support of their position that all roads lead to the disk drive. But HDDs are intricate electromechanical products. They incorporate diverse technologies, including precision motors, spinning platters on spindles, head positioning electronics, advanced read/write heads, slider assemblies and more. With so many building blocks, there are many, many ways that a hard disk drive can fail. There are also the non-

---

[3] *The Diverse and Exploding Digital Universe, IDC, www.emc.com/collateral/analyst-reports*

component threats of thermal buildup and rotational vibration.

But the real concern in using magnetic disk drives as archives is the very real danger of drive failures. It has become widely accepted in enterprise computing that disk drive failures are a matter of 'when' rather than 'if.' A study from Carnegie Mellon presented at the 5th USENIX Conference on File and Storage Technologies (February 2007)[4] points out that customers are replacing disk drives at rates far higher than estimated MTTF would justify.

The Carnegie Mellon study reported on larger high-transaction systems, including high-performance computing sites and Internet services sites running SCSI, FC and SATA drives. The data sheets for those drives listed MTTF between 1 million to 1.5 million hours, which the study translated to annual failure rates "of at most 0.88%." However, the study showed typical annual replacement rates of between 2% and 4%, "and up to 13% observed on some systems."

Although data centers seek return authorizations on drives for a variety of reasons, valid and otherwise, the study noted that a harsh environment at the customer site and intensive, random read/write operations that cause premature wear to the mechanical components in the drive were cited frequently.

Since data archives are designed for longer-term data retention then the service life of the hard disk drive, the use of HDD technology as a long-term data repository is less useful.

Before the advent of optical technology, what archives that have been kept were put in digital

[4] *www.usenix.org*

repositories using tape technology. The technology is a 50+-year veteran in the data center, best known as a backup and archival medium. Tape is a lower-cost alternative that offers high capacities…the popular LTO-4 tape cartridge holds 800 GB in native format. It also offers the advantage of portability; tapes may be taken from place to place for off-site storage as an element of a disaster recovery or data migration plan.

In spite of vigorous efforts by disk drive marketers, tape continues to be looked upon as the backup technology of choice. With an aggressive HDD industry envying their market share, though, tape drive manufacturers are betting their futures on the archiving requirements that are growing consistently.

This may not be the best possible bet. The major criticism of tape technology is in the performance space. As a sequential access medium, mean time to data (MTTD) will always lag behind random access alternatives like hard disk and optical. Before the demands of regulatory agencies and e-discover, a significantly slower time to data could be tolerated for reference data. But modern archives are more frequently used now, and require faster access. Additionally, data integrity has been a long-time issue in the tape industry. Some users have complained vocally about corrupted tapes discovered only after an effort to recover data.

Optical technology represents a realistic, tested and recognized alternative to disk and tape technologies in the storage and recovery of reference data. Since its first commercial debut in the late '70s, optical technology has grown in technical sophistication, reliability in the data center and business usefulness—especially in

records management in the enterprise and departmental data centers.

Optical disc drives make use of laser light or electromagnetic waves near the light spectrum to read and write data. The technology is best known as the dominating technology in commercial audio and video recording and playback. Indeed, optical disc use is too frequently stereotyped as CDs and DVDs, well publicized by the format wars that have been waged in motion picture and recording studio hallways.

Those in the know, however, realize that optical disc is at home in the data centers and the departmental computing environments. And a leading application for optical technology is data archiving, where optical is the compelling compromise between tape and magnetic disk that makes business sense.

The first important business benefit of optical technology is the media life. As an exemplar, 5.25" UDO media is calculated as 50+ years by the manufacturer. This should be sufficient for the most exacting regulatory requirement or e-discovery demand. Tape's longevity is estimated at about 30 years, but the media is subject to wear from continual contact with the tape heads. Optical platters never touch the read/write assembly, adding service life to the platter.

The National Archives and Records Association (NARA) has weighed in regarding service life issues.[5] The organization recommends that data be migrated from a hard disk every three years and from tape media every 5 years. This may be a higher maintenance schedule than some data centers care to shoulder, regulations notwithstanding.

---

[5] *www.archives.gov*

Removability is another business advantage of optical disc---one that cannot be matched by conventional hard disk drives, like tape media. Optical media is easily transported from place to place in fulfillment of a data migration, data security or disaster recovery strategy.

Drivers for Archiving

Enterprises have four primary reasons for archiving:

- *Storage management*
- *Records management*
- *Regulatory compliance and litigation support*
- *Corporate governance and decision support*

Storage management is the most universal business driver for archiving. Archiving allows a user to shrink the storage footprint and enjoy a positive effect on reducing backup and recovery time. There is a significant business ROI created by archiving for these reasons alone. There are also performance and reliability issues as well. Email programs tend to become unstable when archive files become too large. As the volume of stored messages grows, message databases are more likely to become corrupted. After a certain point, response times can also slow down dramatically.

If storage management is about bringing expanding storage requirements under some kind of control, records management is about handling databases, email, IM messages and unstructured files as business records, with all the baggage of business records. This includes protecting their availability and protection for corporate

governance requirements, for litigation support, and for regulatory compliance.

Once policies are in place defining business records for the enterprise, it is a given that at least some of a business' stored data constitute business records, and the company needs to manage and archive them.

The exercise of identifying given information on line as business records is a reflection of the effort to use stored information as an asset rather than as a burden to be managed and tolerated. When viewed as an asset, information is preserved based on content rather than on the age of the file or other data object, the space available to store it, or other factors used to reduce or limit the volume of information to be managed. As a managed asset, organizations prepare themselves to be more compliant than if they did nothing to the information they have.

Compliance involves storing and tracking information because an organization is legally required to do so. Of the many reasons for archiving reference data, compliance is the one that is evolving most rapidly, gets the most media attention and is the one with the most severe consequences for failure. While poorly implemented storage management or records management is expensive, mishandled or neglected compliance could send someone to jail. {See the Chart below.}

If a business is governed by any of the following regulations, it may require stricter archiving policies for storage and record management:

- *Sarbanes-Oxley Act: Makes any public company's senior management individually accountable for the accuracy of its financial reporting.*

Management must demonstrate the integrity of systems used to generate financial reports and monitor or correct lapses in controls.

- *Healthcare Information Portability and Accountability Act(HIPAA): Requires procedures to prevent, detect, contain, and correct security violations of a patient's records. Companies must ensure the privacy of protected health information.*
- *Gramm-Leach Bliley Financial Services Modernization Act: Requires that financial institutions ensure security and confidentiality of customer personal information against "reasonably foreseeable" internal or external events. Finance companies must implement processes that assess and monitor the threat environment as well as tools and policies to counter threats.*
- *Federal Information Security Management Act: Requires Federal agencies to develop, document, and implement programs to secure data and information systems.*
- *California Security Breach Information Act: Organizations maintaining information on California residents must inform those individuals if information is compromised.*

The Federal regulations can be especially stiff for those who decide that compliance is too expensive, too time consuming or too prodigal of resources. The Chart illustrates some of the regulations and the consequences of breach.

Regulation	Penalty	Possible Fine
Sarbanes-Oxley	20 years	$15 million
GLBA	10 years	$1 million
USA Patriot Act	20 years	$1 million
HIPAA	10 years	$100/violation, annual cap $25.000
SEC 17a-14	Suspension, expulsion	$1 million

Litigation support also justifies the development of an archiving solution. A recent study by the legal firm of Fulbright & Jaworski LLC supports the need for effective archiving & discovery tools. It suggests[6]:

- *Nearly 90 percent of all U.S. companies are engaged in some sort of litigation*
- *A corporation with $1.5 billion in revenues averages more than $8 million per year in corporate litigation costs*
- *The average $1 billon per year company faces more than 140 legal cases in the U.S. at any given time*

In late 2006, the Federal Rules of Civil Procedure were amended to reflect the growing importance of a new class of evidence[7]: electronically stored information (ESI). Rules 16 and 20 are amended to provide the court early notice of e-discovery matters. Specifically, 16b states that the scheduling order must include "provisions for

[6] *www.fulbright.com/index.cfm?fuseaction=home.494*

[7] *A good discussion of this may be found at cyber.law.harvard.edu/digitaldiscovery/digdisc_library_4.html*

disclosure or discovery of electronically stored information." Rule 26f requires that parties discuss any issues relating to preserving discoverable information and develop a proposed discovery plan.

These rules accelerate matters in litigation. A party to litigation literally needs to have ESI available for assessment and analysis earlier in litigation than ever before…even before the decision is made to fight the case or settle it. It is also clear that the number of cases subject to rapid case assessment, "litigation holds" on information otherwise scheduled for retirement, evidence preservation and collection will increase.

To a business, all of this means that a fully planned and executed management program is necessary to preserve, protect and track data throughout its lifecycle. Management, financial management, IT and legal counsel will have to interoperate to install a data management and security infrastructure, and reliable archiving is a must..

Corporate Governance

The last ten years or so has seen changes in the philosophy of corporate governance. Shareholders, customers, vendors, government, and the general public are increasingly concerned about how businesses allocate resources, provide incentives, and manage the various transactions and operations involved in doing business.

Corporate governance is the set of processes, customs, policies, laws and institutions affecting the way a corporation is directed, administered or controlled. Corporate governance also includes the relationships among the many players involved and

the goals for which the corporation is governed. The principal players are the shareholders, management and the board of directors. Other stakeholders include employees, suppliers, customers, banks and other lenders, regulators, the environment and the community at large.

There has been an embarrassment of corporate governance failures, including Enron, WorldCom, Tyco and many others. These failures have resulted in newfound public awareness and support for both the theory and practice of governance. One of the byproducts of these corporate failings has been the daunting number of regulatory efforts discussed earlier. But it is important not to confuse corporate governance with regulatory compliance. Governance is the superset; compliance is the subset. In their paper ***The Eternal Charter: Improving Corporate Governance through Compliance and Assured Records Management,*** *document archiving experts at Cohassett Associates have stated:*

"Cohasset Associates believes that governance is controlling and directing the way a business operates to assure that its activities comply with all relevant legal rules and regulations and conform to prevailing "best practice" ethical standards that frequently can go beyond mere legal compliance. Cohasset further believes that good governance manifests the achievement of standards over time and therefore imbues the company with enhanced credibility with regard to future expectations of performance."[8]

Present day business decisions should not be made without a strong foundation in past actions and policies. These can be accessed through a robust and reliable archival program. In the same document, Cohasset Associates points out that the

[8] *www.cohasset.com*

various building blocks of corporate governance are:

"… information-dependent. Accordingly, none of the functions can be performed effectively without access to the records and information assets of the company. And, in the digital society, these "building blocks" also are technology-dependent since these functions must be performed using software and hardware technologies that are indigenous to the creation and lifecycle management of electronic records. The information and technology dependency aspects of governance are overlooked due to the way organizations are run."

Conclusions

Fifteen years ago, archiving meant hauling endless stacks of heavy boxes of paper files to the basement or seeing off the truck transporting paper records to a secure storage repository. Some "high-tech" companies even had their paper records converted to microfiche. Nowadays, though, archives are digital, and optical technology is a big part of some companies' archival solutions.

The appreciation of optical in the archival marketplace can be tracked in part to cost-effective new technologies, such as blue laser, which have increased optical storage capacity and reliability. Another big advantage of optical is that there is no magnetic recording that can fade over time.

In selecting an optical solution, the data center can point to essential differentiators between optical disc and competing technologies.

Optical technology has been proven in the field. There is actually 20-year-old optical media out

there that continues to be read. Contrast that to HDDs and tape, which, as noted earlier, should migrate their data every three to five years, so optical scores very high on both longevity and reliability.

Optical technology is standards-based, in contrast to many tape and disk archive efforts. The Optical Storage Technology Association (OSTA) has endeavored to bring standardization to the technology.[9] This means, for example, that media written by one UDO drive should be able to be read by any UDO drive from any platform vendor. This is in marked contrast to some disk solutions that are very much a proprietary commitment and can lead to vendor lock-in.

Finally, optical products are vendor agnostic as regards compatibility with disk array solutions. Consultant and author Jon William Toigo said in a feature article in Enterprise Systems Journal[10]: "What this means is that Plasmon couldn't care if you originally captured data onto an EMC DMX, Clarion, IBM Shark, Network Appliance Filer, HDS TagmaStore, or any other array. They can work and play well with everything out there and can compliment infrastructure you already have." The same can be said for the other platform vendors in optical technology.

Unfortunately, optical in the data center faces an uphill battle when it comes to market acceptance — mostly due to the reputation it inherits from optical formats such as CD and DVD, which are associated with consumer applications such as gaming, audio, and video applications. While it is true that consumer applications of storage technologies usually lack the robustness and error

[9] *www.osta.org*
[10] *www.esj.com*

handling characteristics required in the data center, the array of products in the optical disc drive field possess a full range of data-grade quality controls. The assumption that all optical disc storage is created equal is quite wrong, and could lead to poor decisions on technology deployment.

As the data center moves from generation to generation, business complexity will continually challenge businesses to handle their information better. The effective archiving and availability of a company's business records will undoubtedly influence how companies perform in terms of storage management, compliance and corporate governance. The value of reference data often influences the value of the business itself. Archiving, then, is a must in the business, and the data center will most effectively safeguard that value in optical subsystems.

Optical Storage Technology Association
19925 Stevens Creek Blvd.
Cupertino, CA 95014 USA
Phone: (408) 253-3695
Fax: (408) 253-9938