The Data Archive Challenge

What's Your Game Plan?

Fred Moore, President
Horison, Inc.
www.horison.com

Abstract: *Did you know that 90% of the data in the world today was created within the past two years and the vast majority of it reaches archival status in a relatively short period of time? New digital data is growing at over 40% annually and are now being generated by billions of people, not just by data centers as in the past, mandating the emergence of an ever smarter and more cost-effective storage infrastructure. For some organizations, facing hundreds of terabytes or several petabytes of archive data for the first time may trigger a need to redesign their entire storage infrastructure. Archiving is a discipline and is quickly becoming a critical "Best Practice". Do you have a game plan for managing the tremendous growth of archival data that lies ahead?*

The Challenge of Data Archiving

Data archiving is the set of processes, activities and the management of digital data over time to ensure its long term accessibility and security. The requirement for a significantly advanced data archiving capability is becoming widespread as the need to migrate enormous amounts of data to a repository that makes data easier to find while lowering costs is rapidly increasing. In addition, the rapid growth in business analytics is continually increasing the value of archival data, further pushing the security, size and management requirements of the digital archive.

Much of today's archival data is created as unstructured data which typically is formatted as bitmap images, objects, text, photos, and video. Unlike structured data, unstructured data is not part of a database and has little or no metadata or naming tags to describe its contents, hence the name unstructured. Also, a growing list of global compliance, government and legal regulations now describe the way data should be managed, protected and how long it should be stored throughout its lifetime extending the need to archive data for indefinite if not infinite periods of time.

Bottom line: *Archives are no longer a repository for low-value data. Effectively managing the digital archive is attainable and now requires a multi-faceted strategy.*

Backup and Archive are Different Processes
Many people still confuse backup and archive. The process of backing up is making copies of data which may be used to *restore* the original copy after a data loss event. Data archiving is the process of moving data that is no longer actively used to a separate data storage system for long-term retention. Archive data itself normally needs to be backed up since having a single copy of any meaningful data presents an exposure should the only copy become inaccessible.

Backup (A Copy): The back up process creates copies of data which may be used to restore the original copy after a data loss or data corruption event.
Primary Solutions: Disk, Tape, Flash, DVDs, for PCs and personal appliances

Archive (A Move): The process of archiving moves data to a new location and refers to data specifically selected for long-term retention. Archives are data which is infrequently used that was removed from its initial location and stored elsewhere for long-term retention. Archive data should also be backed up as having only one copy of any meaningful data is risky.
Primary Solutions: Tape, Disk

Active Archive: An Active Archive is a solution combining archive applications along with disk and tape hardware allowing users to preserve, protect and access all of their archival or tier 3 data. In an active archive, more active archival data often resides on disk serving as a cache buffer while the less active archival data resides on tape. An Active Archive may contain production data, no matter how old or infrequently accessed, that can still be retrieved online if required. Active Archiving can use your existing storage equipment to build an integrated hardware and software solution.

Note: The Active Archive Alliance is a collaborative industry alliance formed to educate end user organizations on the evolving new technologies that enable reliable, online and efficient access to their archived data. See → http://activearchive.com/

Offline storage: Offline storage requires some direct human action in order to make access to the storage media physically possible.
Primary Solutions: Tape, paper, DVD, Flash, film

Bottom line: *Backup and archive are not the same. Archives need to be backed up – something many businesses fail to consider when building the archive.*

Archive Considerations for End Users
Coping with archival data growth, as many companies are discovering, cannot cost effectively be achieved by deploying more storage capacity in the form of high cost disk arrays. From a CAPEX (Capital Expense) perspective, the cost of acquiring storage arrays and keeping them functional can spiral out of control as the data repository increases in size. From an OPEX (Operational Expense) perspective, increasing the deployment of additional disk arrays increases spending on administrative personnel, data management, high availability requirements, security, footprint and utility power compared to more efficient solutions.

Data archiving is a comparatively simple process to understand, but can be a more difficult one to implement, as many companies are finding. Tier 3 storage refers to the archive layer for data storage. The requirements of tier 3 storage solutions now favor tape which provides numerous advantages over disk for archiving. The basic components listed below highlight several key challenges to address in order to build a sustainable archive capability.

Hierarchical Storage Management (**HSM**) is a data storage technique which automatically moves data between high-cost and low-cost storage media based on pre-defined policies. While it would be ideal to have all data available on high-speed devices all the time, this is prohibitively expensive for most every organization particularly as the size of the archives gets larger. Instead, HSM systems store the bulk of the enterprise's data on slower devices, and then copy data to faster disk drives when needed. In an active archive, HSM uses faster disk drives as caches for the mass storage devices. The HSM system monitors the way data is used and makes decisions as to which data can be moved to archival devices and which data should stay on the disk storage. In many cases, HSM is the dedicated software that intelligently moves data into the archive.
Frequently used HSM products include IBM Tivoli Storage Manager, CommVault Simpana Archive, VERITAS Enterprise Vault, Sun Microsystems SAMFS/QFS, Quantum StorNext, HP HSM, and EMC Legato OTG DiskXtender.

Bottom line: HSM software improves the capability of archiving to manage storage devices efficiently and determine when data reaches archival status, especially in large-scale storage environments where storage costs can mount rapidly.

Basic Components for Building a Long-term and Scalable Data Archive

Archive Planning	What it Means
Classify data for archiving	Classification defines the user policies, processes, and software tools to dynamically determine when data reaches archival status
Build a cost-effective storage platform that delivers the required access, security and throughput for archival data	Implementing the most cost-effective type of storage for archival purposes usually combines disk and tape along with offsite solutions enabling geographical redundancy for recovery and business resumption
Establish archive policies	Users need to agree to expiration/deletion/retirement dates, retention periods, migration and backup frequency, and security controls to move data assets in and out of the archive platform from the time data is conceived through its final disposition
A policy-based data mover	A HSM (Hierarchical Storage Management) system monitors the way data is used and applies user policies to determine which data should be moved to archive storage and which data should stay on the faster and more expensive devices
Provide lifetime data protection	Includes backup of the archive, encryption, and WORM capabilities needed to prevent data from being lost, altered, or destroyed

Deliver the most cost-effective archive solution over time	The disk TCO is ~ 15x greater than tape for archiving and ~4x greater for backup, this trend heavily favors tape for archive and is expected to continue for the foreseeable future – move archives to tape whenever possible
Design the archive with acceptable search and retrieval times	Building an Active Archive combining disk and tape offers optimal retrieval times. In addition, beginning with LTO-5, the LTFS tape partitioning capability makes tape behave more like disk with "drag and drop" capability improving tape access times.
Determine whether to archive on-premise or in the cloud?	Consider the cloud for archiving as it can provide cost savings for your data without having to own and operate the actual hardware and software that makes up the archive

Source: Horison Inc.

Businesses face growing challenges in managing archives, controlling costs, and meeting regulation requirements. Difficulties include data capture (ingest), storage, protection, and providing timely retrieval. The value of an archive is increasing as the benefits of working with larger and larger datasets enable analysts to project business trends, prevent diseases, and improve security and national defense. Presenting an ever-moving target, the limits of archives are now on the order of petabytes ($1x10^{15}$), exabytes ($1x10^{18}$) and will approach zettabytes ($1x10^{21}$) of data in the foreseeable future.

Bottom line: *Data archiving is a comparatively simple process to understand, but can be a more difficult one to implement requiring careful planning. The recent improvements in tape technology have greatly enhanced the end-user's ability to implement a highly available, scalable and cost-effective archive solution.*

The Era of Colossal Content is Near

E-Discovery Benefits from the Latest Tape Developments

Electronic discovery (E-discovery) refers to any process in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a civil or criminal legal case. Archival data is quickly becoming a key and necessary data source for the E-discovery process. The nature of digital data makes it extremely well-suited to investigation since digital data can be electronically searched with relative ease, whereas paper documents must be scrutinized manually. E-discovery applications and the latest storage solutions enable organizations to pull information and records from massive volumes of data spanning an enterprise.

Much of the data used for E-discovery is physically stored as archival data, normally used infrequently, but used more frequently during the E-discovery search process making it an ideal archival application. The costs of E-discovery actions can be lowered and the process managed more effectively than ever before given the recent developments in tape. Access time improvements with tape partitioning using LTFS and by implementing an active archive now enable a highly cost-effective E-discovery infrastructure compared to a disk-only archive. E-discovery is quickly becoming a primary application for digital tape archives.

Compliance and Regulatory Issues Soar

Compliance and regulatory issues are increasing E-discovery activity and expanding the size of archives. Regulations often prohibit data deletion meaning that data will need to be kept indefinitely. For example, the United States Securities and Exchange Commission (SEC) require broker-dealers to digitally preserve and produce communications with their clients, along with most other business documents. While regulations such as the Sarbanes-Oxley Act and the Health Insurance Portability Accountability Act (HIPAA) are widely recognized as covering publicly traded companies and healthcare organizations, most industries today must comply with a much longer list of electronic information laws and standards. Financial institutions are required to abide by the Check Clearing Act for the 21st Century, and merchants that accept credit card purchases are required to comply with the Payment Card Industry Data Security Standard (PCI DSS). Being able to audit older (archival) records often forms the foundation of these regulations and laws. Businesses in any industry can be required to produce historical tax records, financial data and supporting information by state, local or federal tax agencies at any given time.

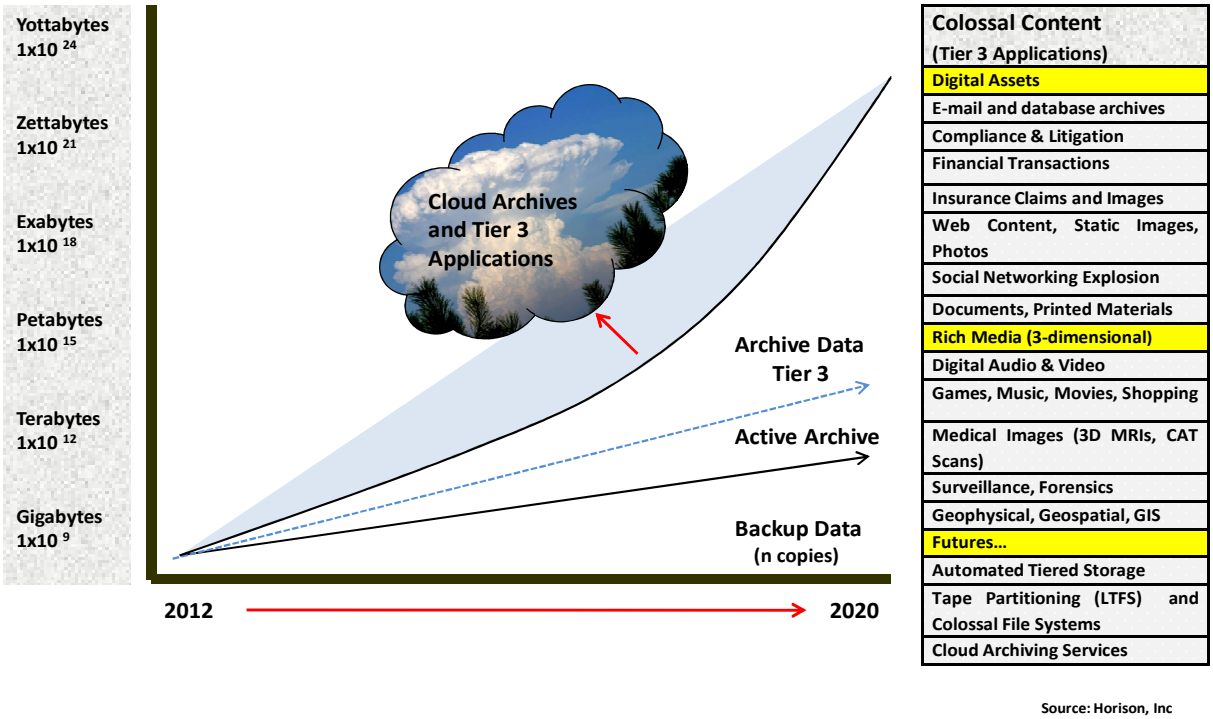
Social Networking Generates Archive Data

Another relatively new driver of long-term storage requirements is social networking. The social data revolution has defined a significant shift in human communication patterns increasing personal information sharing, made possible by the arrival and rapid rise of social networks in early 2000s. While social networks were used initially to share photos and private messages, the subsequent trend towards people passively and actively sharing personal information on a broad scale has resulted in unprecedented amounts of public data and sets the stage for the *n to n* hyper-exponential spread of data. The real value of social network data remains basically untapped. This enormous and continually updated data source is destined to become a new type of analytic tool for the social sciences. Researchers can use social data to forecast trends such as unemployment, travel and entertainment patterns, spending, security threats and political opinions in a way that is faster, more accurate and cheaper than any prior reports, surveys or polls. A characteristic of social data is that it becomes inactive and quickly reaches archival status after creation.

E-discovery, compliance and social networks are just a few examples of many applications that are quickly expanding data archive requirements. Clearly, archived information is no longer limited to e-mails, word processing documents, spreadsheets, PDF documents, medical and graphic images, and electronic check images. Ultimately, all of this data not only has to be archived, but must be both accessible and recoverable over indefinite periods of time.

The Era of Colossal Content

Tier 3 Applications Driving Unprecedented Demand



Source: Horison, Inc

Bottom line: *Archive, fixed content, E-discovery, compliance, entertainment, scientific, social networks, and unstructured data requirements are soaring defining the era of colossal content and have become the primary drivers for future archival storage demand. Much of this growth will reside on tape technology.*

Storage is *the* Essential Component

Tape has been shifting from its historical role as a backup solution to a technology that addresses a much broader set of data storage goals specifically including data archive and disaster recovery services. Previously various non-standard application formats did not always allow customers to easily find their data on tape. However, the recent advent and improvements of software that supports both disk and tape by presenting a file system image to the user, such as the relatively new LTFS tape partitioning capability first made available with LTO-5 tape, enables tape to even more effectively address the archive market. With LTFS, the traditional longer, sequential search times for tape have given way to more disk-like access using familiar drag and drop techniques. The advent of the cloud, and the inherent consolidation of data into large-scale storage systems that cloud storage implies, signals that another category of storage is emerging – tape in the cloud –which should further improve the economic model for archiving data.

The chart below compares key archival storage requirements that are addressed by tape and disk to yield an optimized infrastructure. Though challenging and becoming increasingly complex as storage requirements grow, carefully designing the digital archive yields much improved operational efficiencies and sizeable cost savings. The time has arrived for businesses to begin re-architecting and optimizing their archives before the task becomes overwhelming.

Tape and Disk Tradeoffs for Building an Effective Digital Archive

Tier 3 Capability	Tape	Disk
TCO	✓ Favors tape for backup (4:1) and archive (15:1)	Much higher TCO, more frequent conversions and upgrades
Long-life media	✓ 30 years or more on all new tape media	~4 years for most HDDs before upgrade or replacement, 7 years or more typical for tape drives
Reliability	✓ Tape BER (Bit Error Rate) has surpassed disk since 2005	Disk BER not improving as fast as tape
Move data to remote location for DR with or without electricity	✓ Yes, can move data remotely with or without electricity. Natural disasters can force physical media movement	Difficult to move disk data to a remote location for DR purposes without requiring electricity
Inactive data does not consume energy	✓ Yes, this is becoming a goal for most data centers. "If the data isn't being used, it shouldn't consume energy"	Rarely for disk, potentially in the case of "spin-up spin-down" disks *Note: data striping in arrays often negates the spin-down function*
Provide the highest security levels	✓ Yes, encryption and WORM capability available on essentially all midrange and enterprise tape drives	Becoming available on selected disk products, PCs and personal appliances, not yet widespread
Capacity growth rates	✓ Roadmaps favor tape over disk with 35 TB capability jointly demonstrated by Fujifilm and IBM	Continued steady capacity growth but roadmaps project disk to lag tape
Data access time	✓ LTFS has improved tape access with "drag and drop" capability for files	✓ Disk is faster than tape for initial access and random access applications
Portability	✓ Yes, media completely removable and easily transported	Disks are difficult to physically remove and to safely transport
Active archives	✓ Effective when combined with disk	✓ Effective when combined with tape

Source: Horison, Inc.

Bottom line: *Tape vendors continue to innovate and deliver compelling new features with lower economics and higher reliability which have positioned tape as the optimal choice for long-term archiving as well as continuing to play a key role for backup.*

Conclusion

Are you ready to address the archive challenges that lie ahead? With growing volumes of data, increasing data security breaches, and complex application-performance issues, most enterprises today face continually many new data management challenges. The old process of keeping inactive data online for extended periods of time not only creates security risks but significantly increases infrastructure cost. Some enterprises are aggressively throwing away low-value, no-value data to minimize complexity and lower these costs. Data archiving supports many requirements, including E-discovery, regulatory compliance, managing test data, data analytics, and historical preservation. Businesses should consider building an enterprise-wide data archiving strategy to cost-effectively address the growing amounts of unstructured data in addition to increasingly larger databases and structured applications.

Tape densities will continue to grow and costs will decline, while disk drive performance is flat and capacity growth will slow. In the near future, solid archive strategies will carefully evaluate deploying a "tape in the cloud capability" for the optimal archiving solution. The opportunities for tape storage solutions on-premise or in the cloud have grown considerably and are being fueled by a plethora of significant technology advancements positioning tape to address much of the inevitable colossal content explosion. It really shouldn't matter which technology is the best for digital archiving, it just happens that the numerous improvements in tape have made it the optimal choice for archiving for the foreseeable future. Designing a cost-effective archive is attainable – now is the time to develop a solid and sustainable game plan.

End of report