BY: Fred Moore President
www.horison.com

The Archival Upheaval

Planning for Petabyte Pandemonium

2015 Tape Technology Update Series

Abstract Did you realize 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? Overall, newly created digital data is growing at over 40% annually and are now being generated by billions of people, not just by large data centers as in the past, mandating the emergence of an ever smarter and more cost-effective and secure long-term storage infrastructure. Most data typically reaches archival status in 90 days or less, and archival data is accumulating even faster at over 50% compounded annually as many data types are being kept indefinitely. For most organizations, facing hundreds of terabytes or several petabytes of archive data for the first time can trigger a need to redesign their entire storage infrastructure. Archiving is now a required storage discipline and is quickly becoming a critical "Best Practice". Are you prepared to manage the tremendous growth of archival data that lies ahead? *It's time to develop your game plan.*

What is Data Archiving?

Simply stated, archival data is data that is infrequently used - but needs to be kept indefinitely. Data archiving is the set of processes, activities and the management of archival data over time to ensure its long term accessibility and security. The requirement for an advanced data archiving capability is now widespread as the need to move enormous amounts of data to a secure repository that provides easy access while optimizing costs is rapidly increasing. In addition, the rapid growth in business analytics for large and complex data sets – also known as Big Data - is continually increasing the size and value of archival data, adding pressure to the management and security requirements of the digital archive.

Much of today's archival data is created as unstructured data which typically is formatted as bitmap images, objects, plain 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 for easier access, hence the name unstructured. The need to effectively search for and retrieve large amounts of unstructured content has resulted in the deployment of naming conventions, tags, indices and numerous search engines. Additional archive considerations and factors are listed below.

Archive Data Growth > 50% CAGR Makes it the Fastest Growing Data Category

The capacity needed to store unstructured data (~70% of all stored data) continues to escalate far beyond the capacity required for structured data normally stored in databases (~20% of all stored data). This requires using a highly scalable (scale-out) technology for the archive platform.

Reduced Pressure on the Backup Window

Even with disk backup processes using compression or deduplication, backup windows face constant pressure from data growth rates that can exceed a 40% CAGR. There's no point in repeatedly backing up unchanged data – especially if it's seldom accessed. Archiving can remove much of the low activity and unchanged data from the backup set to speed up the process.

Big Data, Data Mining, Analytics, Derive Value from Archives

Not all data is created equal in value but, in the Big Data era, organizations are quickly learning the value of analyzing vast amounts of previously untapped archival data. Various industry studies indicate less than 5% of all stored digital data is ever analyzed and that over 40% of all stored data hasn't been accessed in the past 6-12 months. All this is changing in the Big Data era.

Compliance and Legal Requirements

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 often extending the need to maintain archival data for indefinite, if not infinite periods of time.

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

Did You Know - Backup and Archive are Very Different Processes?

Many people still confuse backup and archive – some even think it's the same thing. Backup is the process of making copies of data which may be used to *restore* the original copy if the original copy is damaged, corrupted, or after a data loss event. Archiving is the process of moving data that is no longer actively used, but is required to be retained, to a new location for long-term storage.

Backup (A Copy process): The back up process creates copy(s) of data for recovery purposes which may be used to restore the original copy after a data loss or data corruption event. Backups are cycled and updated frequently to account for and protect the latest versions of important data assets.
Primary Solutions: Disk, Tape, Flash, and DVDs for PCs and personal appliances

Archive (A Move process): The process of archiving moves infrequently used data to a new location(s) and refers to data specifically selected for long-term retention. Ideally more than one copy of archive data should exist since having a single copy of any meaningful data presents an exposure should the only copy become inaccessible. Note: Tier 3 storage is commonly referred to as the archive layer or the tape layer of the storage hierarchy.
Primary Solutions: Tape, Low-cost Disk, Local and Remote Data Vaults

Active Archive: The Active Archive concept first appeared in 1997 for mainframe automated tape library systems and provides a performance enhancement for the tape-based archives. An Active Archive combines disk - serving as a cache - with tape providing online access, search capability and easy retrieval of long-term data. Active Archiving implementations can use your existing storage equipment to build an integrated hardware and software solution and may also incorporate enhanced file systems such as LTFS (Linear Tape File System). For example, in the case of another outbreak of the Ebola virus, past records of historical and clinical data on the virus can become important and quite active for a period of time as researchers try to find connections with this outbreak and prior outbreaks. In this case, archival data becomes active archive data and resides in the disk buffer for an extended period of time.

To help address the challenges with the archival upheaval, The Active Archive Alliance was launched on April 27, 2010 as a collaborative industry association formed to educate end user organizations on the evolving new technologies that enable the most effective access to their archived data. See http://activearchive.com/

Offline Storage: Offline storage, also called cold storage, requires some direct human action to make access to the storage media physically possible. Off-premise vaults are often used for offline storage and include warehouses, highly secure physical facilities, and even digital vaults built into solid granite.
Primary Storage Solutions: Tape, Low-cost Disk, Paper, DVDs, Film

***Key points:* Backup and archive are not the same. Backup creates additional copy(s) of the original data for the purpose of data recovery. Archive moves the original data to another more cost-effective location for long-term storage.**

Building an Archive Strategy – Getting Started

Data archiving is a relatively simple process to understand, and can be successfully implemented given the more effective, advanced hardware and software that is available today. The requirements for the optimal tier 3 storage solutions now heavily favor tape over disk for archiving and long-term storage. The basic steps listed below provide realistic guidelines to build a sustainable archive capability. You may choose to add additional steps to the process based on specific business needs. Remember to keep it simple! Many plans make provisions for more than one copy of archived data.

Basic Steps for Building a Long-term, Secure and Scalable Data Archive

Steps	Archive Planning	What it Means
Step 1	Classify Your Data by Value and Criticality	Understand your data to determine if it is mission-critical, vital, sensitive, or non-critical
Step 2	Determine Which Data to Archive, How Many Copies Needed	Includes defining archiving parameters such as legal regulations, what data is no longer needed, when data reaches end of life, internal company rules
Step 3	Determine When to Archive, Set Archive Thresholds and Security Policies	These often include last access date, age of data, space limitations, and frequency of access. Assign Encryption and WORM capabilities to prevent data from being altered, stolen, or destroyed
Step 4	Determine How Long Data Will Remain in the Archive	Months, years, infinity? These include internal policies, B2B, B2C and legal requirements - review periodically
Step 5	Select a Software Solution to Automate the Archive Process (A policy-based data mover or HSM software)	HSM (Hierarchical Storage Management) or archive type software products monitor data reference patterns and applies user-defined policies to determine which data should be moved to archive status and when – and then moves it dynamically
Step 6	Select the Optimal Archive and Active Archive Storage Platform, Remote Vault, Local or Cloud Options	Implement the most cost-effective type of storage for archival purposes. This heavily favors tape along with offsite solutions providing geographical redundancy for data protection, recovery and business resumption
Step 7	Set Rules for Who Can Access the Archives – Insure Someone is in Charge!	Assign security codes, passwords, forensic IDs, for each authorized person who can access the archive

Source: Horison Inc.

Businesses face numerous challenges in managing scalable archives such as controlling costs, meeting regulatory requirements, data ingest, data security, and providing timely retrieval. The Big Data era has

increased the value of archival data as the benefits of analyzing very large datasets are invaluable. Presenting an ever-moving challenge, 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.

Coping with rapid archival data growth and accumulation, as many companies are painfully discovering, cannot be cost effectively achieved with a strategy of continually adding more capacity with costly disk arrays. From a capital expense perspective, the cost of acquiring disk arrays and keeping them functional can easily spiral out of control. From an operational expense perspective, increasing the deployment of additional disk arrays increases spending on administration, data management, floor space and energy compared to more efficient tape solutions as the data repository increases in size.

Key point: ***Data archiving is a comparatively simple process to understand, but can become a challenge to implement without a plan. It's time to get started before the pandemonium begins.***

Step 1 - Classify Your Data by Value and Criticality

Classifying data is a critical IT activity for the purposes of implementing the optimal solution to store and protect data throughout its lifetime. This process works best with a small team knowledgeable of the applications and storage infrastructure. Empower a team leader to oversee the process! Though you may define as many levels as you want, four de-facto standard levels of classifying data are commonly used: mission-critical data, vital data, sensitive data and non-critical data.

Determining criticality and retention levels also reveals which data protection technology is best suited to meet the RTO (Recovery Time Objective) requirements. Data classification also aligns data with the optimal storage tiers and services based on the changing value of data over time. Defining policies to map application requirements to storage tiers has historically been time-consuming, but has improved considerably with the help of several advanced classification and policy-based software management solutions from a variety of companies. All data is not created equal.

Mission-critical data is used in the most important business processes, revenue generating or customer facing applications and typically averages as much as 15 percent of all stored data. Mission-critical applications normally have a RTO (Recovery Time Objective) requirement of a few minutes or less to quickly resume business after a disruption. Losing access to mission-critical data means a rapid loss of revenue, potential loss of customers and can place the survival of the business at risk in a relatively short period of time. Ideally, mission-critical data resides on highly functional, highly available, and more costly enterprise class disk arrays or SSDs requiring multiple backup copies that are often stored at geographically separate locations.

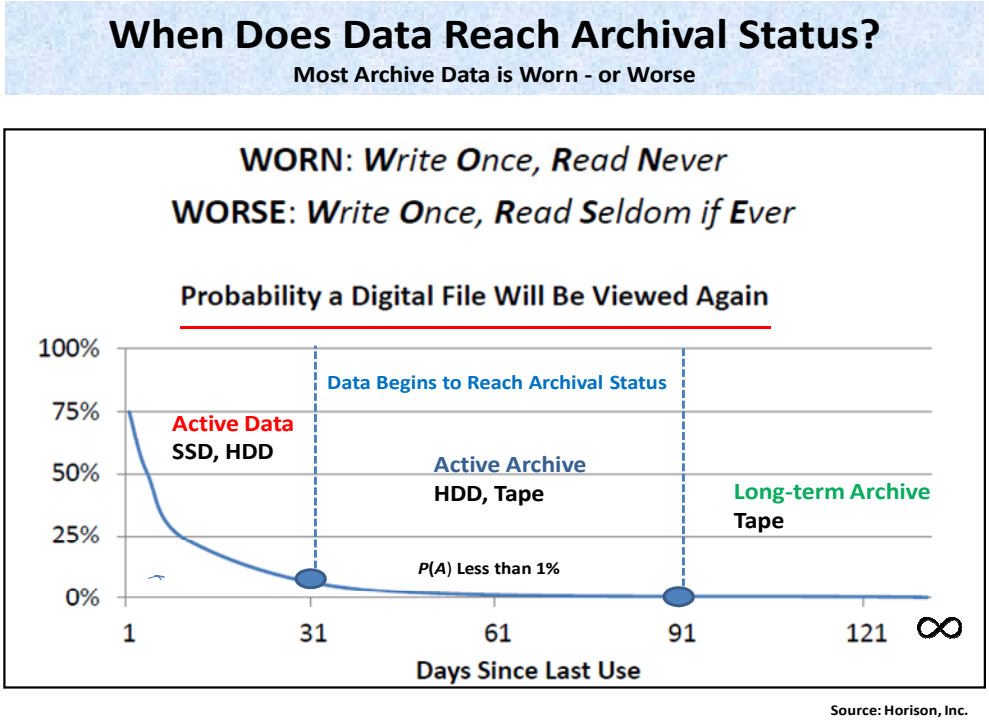
Vital data averages up to 20 percent of all stored data; however, vital data doesn't require "instantaneous" recovery for the business to remain in operation. Data recovery times - the RTO ranging from a few minutes to a few hours or more, are typically acceptable. Vital data is critical to the business and often resides on enterprise and mid-range disk subsystems.

Sensitive data comprises an average of 25 percent of all data stored online and is important but doesn't require immediate recovery capabilities. The RTO can take up to several hours without causing major operational impact. With sensitive data, alternative sources exist for accessing or reconstructing the data in case of permanent data loss. Sensitive data normally resides on low cost mid-range disk arrays and automated tape libraries.

Non-critical data typically represents 40 percent or more of all digital data, making it the largest and the fastest growing data classification level. Lost, corrupted or damaged data can be reconstructed with less complex recovery techniques requiring minimal effort, and acceptable recovery times can range from hours to days since this data is not critical for immediate business survival. *However non-critical data doesn't mean it isn't valuable.* Non-critical data may suddenly become highly valuable based on unknown circumstances. E-mail archives, legal records, medical information, entertainment, historical data and scientific recordings often fit this profile and this category typically provides much of the large-scale content driving Big Data analytics. Non-critical data is most cost-effectively stored on tape.

Step 2 – Determine Which Data Types Should be Archived

Establish the criteria for what types of data to archive such as internal policies, customer and business partner requirements, and compliance data. Many files begin to reach archival status after the file has aged for a month or more, or whenever the *P(A)* (probability of access) falls below 1%. See chart below.



Step 3 - Determine Archive and Security Policies

Setting the archive and security policies are the rules that will govern what gets archived, when and where it gets archived and retention periods. Archive policies can include last access date, age of data, frequency of access, pre-set capacity thresholds, and will address any specific legal retention regulations. These policies should be reviewed periodically as archival requirements can change based on a variety of circumstances or laws. Encryption and WORM security requirements can also be assigned to the appropriate files, data sets, and objects to prevent them from being altered or destroyed.

Step 4 – Determine How Long Data Will Remain in the Archive

This step defines when data is no longer needed, when data reaches end of life, and its final disposition. For many applications and data files, the lifetime requirement for data preservation has become indefinite - or infinite - as this data may *never be deleted*. As an example, the retention period for certain medical records such as X-rays and MRI images may need to be kept for the lifetime of the patient while epidemic data will be kept forever. Most of today's video, television programming, sports events, movies, and scientific data will be kept in a digital archive forever and for most of this data, frequency of access will steadily decline over time.

Step 5 - Select Software Solution to Enable Archive Process

Archives are best managed by Hierarchical Storage Management (HSM) type software. The HSM management system monitors access and usage patterns and makes user-defined, policy-based decisions as to which data can be moved to archival status and which data should stay on primary storage. HSM can help to identify candidate data for inclusion in a deep or active archive and can identify temporary data that can simply be deleted once its useful life has expired. Some HSM software products also provide backup and recovery functionality.

Selected HSM and Archive Products	Vendor
DFHSM, Tivoli Storage Mgr., HPSS (HPC)	IBM
StorNext	Quantum
SAM-QFS	Oracle
DMF	SGI
DiskXtender	EMC
NetBackup Storage Migrator	Veritas (Symantec)
HP HSM	HP
CA-Disk	CA
Simpana	CommVault
StrongBox Data Manager	Crossroads
FileStor/HSM	Fujifilm

Key point: ***Several effective archival software solutions are available to determine when data reaches archival status, where it should be stored, and how long it should be kept.***

Step 6 - Determine the Optimal Archive Platform - Disk or Tape for Archives?

In addition to classifying data, storage platforms are classified based on criteria such as performance, capacity, access speed, reliability and cost. Tape has been shifting from its historical role as primarily a backup solution to addresses a much broader set of storage requirements specifically including data archive and disaster recovery services. Several new and important technologies were implemented for tape yielding numerous improvements including unprecedented cartridge capacity increases using Barium Ferrite (BaFe) media, vastly improved bit error rates compared to disk, much longer media life, and faster data transfer rates than any previous tape - or disk - technology. The media life for all new LTO and enterprise class tape now reaches 30 years or more making tape the most highly secure, long-term archive digital storage medium available.

With the announcement of LTFS (Linear Tape File System) in 2010, the long standing rules of tape access were changed as the traditional longer, sequential search times for tape have given way to more disk-like access using familiar drag and drop techniques. The LTFS partitioning capability, first made available with LTO-5, positions tape to more effectively address archive requirements by enabling tagging of files with descriptive text, allowing for more intuitive searches of cartridge and library content. Using LTFS and NAS in front of a tape library to create an active archive is sometimes referred to as “tape NAS”. A good example of “tape NAS” can be found in Fujifilm’s Dternity NAS solution. LTFS and tape partitioning have barely scratched the surface of their potential. Expect an increasing number of ISVs (Independent Software Vendors) to exploit LTFS functionality in the future.

Tape is quickly becoming the most viable, lowest cost, cloud archive solution. 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 has the potential to significantly improve the economic model for cloud archival services. Storing archival data on tape in the cloud represents a future growth opportunity for tape providers and a much lower cost, more secure archive alternative than disk for cloud providers.

Disk *can* be used for archival storage and serves the active archive concept well; however using disk for long-term archives has become the most costly and least desirable option. A disk drive can consume from 7 W to 21 W of electrical power every second to keep them spinning. A comprehensive TCO study by ESG (Enterprise Strategies Group) comparing an LTO-5 tape library system with a low-cost SATA disk system for backup using de-duplication (best case for disk) shows that disk deduplication has a 2-4x higher TCO than the tape system for backup over a 5 year period. The study also concluded that disk has a TCO of 15x higher than tape for long-term data archiving. The TCO advantage for tape is expected to become even more compelling with future technology developments. See A Comparative TCO Study: VTLs and Physical Tape, by ESG. Clearly disk technology has been advancing, but the progress for tape has been even greater over the past 10 years.

Archives can become challenging and more complex as storage requirements grow emphasizing that a carefully designed digital archive strategy which yields improved operational efficiencies and sizeable cost savings is increasingly important. The chart below compares key archival storage capabilities which are best addressed by tape or disk to implement an optimized archive infrastructure.

Tape and Disk Considerations for Building the Optimal Digital Archive

Archive Capability	Tape	Disk
TCO	Favors tape for backup (2-4:1) and archive (15:1)	Much higher TCO, more frequent conversions and upgrades
Long-life media	30 years or more on all new enterprise and LTO media favoring archive requirements	~4 years for most HDDs before upgrade or replacement, 7 years or more is typical for tape drives
Reliability	Tape BER (Bit Error Rate) 3 orders of magnitude better than disk	Disk BER falling behind - not improving as fast as tape
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 all midrange and enterprise tape	Becoming available on selected disk products, PCs and personal appliances
Capacity growth rates	Roadmaps favor tape over disk with 154 TB demonstrated by Fujifilm and IBM	Slowing capacity growth as roadmaps project disk capacity to lag tape for foreseeable future
Data access time	LTFS has improved tape access time with disk-like "drag and drop" capability for files	Disk is faster than tape for initial access and random access applications
Portability - Move media to different location for DR with or without electricity	Yes, tape media completely removable and easily transported in absence of electricity	Disks are difficult to physically remove and to safely transport

Source: Horison, Inc.

Key point: ***Tape vendors continue to innovate and deliver compelling new features with lower economics and the highest reliability levels. This has established tape as the optimal tier 3 choice for archiving as well as playing a larger role for backup, business resumption and disaster recovery.***

Step 7 - Set Rules for Who Can Access the Archives

Archival data may contain original content, confidential, classified and regulated legal files, and may need to be encrypted and kept in a highly secured facility with restricted access. Security codes, passwords, and forensic markers should only be assigned to those who have authorized access to the archives. Remember – appoint a team leader. Somebody has to be in charge!

Key point: ***The successful archive strategy requires people to agree on the relative value of specific applications and data to the survival of the business. Then follow the steps above.***

Numerous Applications are Driving Sustained Archival Data Growth

Tape has been expanding its historical role as a backup solution to a much broader set of requirements including data archives and disaster recovery services. Just ten years ago, large businesses generated roughly 90% of the world's digital data. Today an estimated 75-80% of all digital data is generated by individuals - not by large businesses – however the majority of this data will eventually wind up back in a large business, service provider or cloud provider data center. The applications listed in the chart below all create significant volumes of data that become archival as the data ages.

Big Data and High Capacity Applications Driving Future Tape and Archive Storage Demand (Tier 3)

Digital Assets (Fixed Content)	**Rich Media (Motion, 3D, Multi-dimensional)**
E-mail archives, Compliance & Litigation with long-term storage requirements	Digital Audio & Streaming Video, YouTube
Big Data - capture, storage, search, sharing, transfer, analysis and visualization	Intelligence Gathering - Satellite, Drone, Remote Sensing
Documents, Printed Materials, Books, Magazines	Entertainment - TV, Sporting Events, Digital Games, Music, Movies, Shopping
Medical Patient Data, Archived Files/Fixed Images	Medical Images (3D MRIs, Digital Scans, Ultra-sound, Facial Recognition)
Insurance Claims, Financial Transactions/Data, Banking Records, Contracts	Scientific, Atmospheric, Geophysical, Geospatial, GIS, etc.
Web Content, Social Networking and Media, Static Images, Digital Photo Repositories	Digital Surveillance/Security, Motion Sensors, Forensics
Archival Storage Futures…	
Automated Tiered Storage for Flash, HDD and Tape (Advanced HSM-like policy-based software)	
Intelligent Active Archive – Pre-staging, Space Management, Integrated Tape and Disk (Flash?)	
Advanced LTFS partitioning for enhanced tape access	

Source: Horison Inc.

The size and value of archival data is increasing and researchers are quickly discovering the benefits of analyzing very large objects, files and datasets. For many data types, the lifetime for data preservation has become "infinite" and will constantly stress the limits of the archive infrastructure as the data will never be deleted. Presenting an ever-moving target, the size of preserving large-scale digital archives are now reaching the order of petascale ($1x10^{15}$), exascale ($1x10^{18}$) and will approach zettascale ($1x10^{21}$) capacities in the foreseeable future requiring highly scalable storage systems.

Key point: ***With tape now having a TCO of 1/15th that of disk for archival storage, and reliability having surpassed disk drives, the pendulum has shifted to tape to address much of the tier 3 demand.***

Conclusion

Archival storage systems can be implemented today to effectively address the archival upheaval. The old process of keeping inactive data on disk storage for extended periods of time is essentially obsolete and not only creates security risks and performance problems, but significantly increases operational expenses. Thoughtful deployment of today's much improved hardware and software solutions can yield an effective and sustainable archive strategy. Each organization will have to justify their archive implementation to senior management in their own way, but the compelling reason to move low-activity data to the optimal storage tier for long-term retention yields significant cost savings with improved security capabilities. The bottom line is that your business-value case for data archiving will be quite compelling and likely include cost containment (free up disk space), risk reduction to ensure regulatory compliance, improved productivity by getting inactive data out of the path of the backup window, more efficient searches and retrieval, and improved storage administrator efficiency.

Archive storage requirements are mounting as tape technology has made tremendous strides – what timing! The future role of tape in archival storage cannot be denied and the cost savings of using tape compared to disk for archive are most significant. Tape densities will continue to grow and costs will decline, while disk drive performance is expected to remain flat, and capacity growth will increase but at a slower pace. 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 clear cut and optimal choice for archiving for the foreseeable future. Are you prepared to address these enormous archive challenges that lie ahead?

Designing a cost-effective archive is attainable and the components are in place to do so – sooner or later the chances are high that you will be forced to implement a solid and sustainable archival plan. Now is the time to get started.

End of Report