Archival Data Storage

Planning for the Archive Avalanche

HORISON
Information Strategies

BY: Fred Moore President
www.horison.com

Abstract Relentless data growth is a given as data has become critical to all aspects of human life over the course of the past 30 years. Newly created worldwide digital data is expected to grow at 30% or more annually through 2025 and are now being generated by billions of people in addition to the large data centers as in the past, mandating the emergence of an ever smarter and more secure long-term storage infrastructure. Data is rapidly piling up in archives as a SNIA survey indicated 57% plan to retain data 50 years or more. Digital archiving is required by many industries to comply with government regulations for storing financial, customer, and patient information. As businesses, governments, societies, and individuals worldwide increase their dependence on data, archiving becomes more important.

Most data typically reach archival status in 90 days or less, and archival data is accumulating at over 50% compounded annually as many data types are being kept indefinitely in hope that potential value might be unlocked. For most organizations, facing terabytes, petabytes and even exabytes of archive data for the first time can force the redesign of their entire storage strategy and 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 long-term storage and archival data that lies ahead? *It's time to develop your game plan.*

What is Archival Data?

Simply stated, archival data is data that is infrequently used and seldom if ever changes - but potentially has significant value and needs to be kept indefinitely. Archival data is a collection of data objects, perhaps with associated metadata, in a storage system whose primary purpose is the long-term preservation and retention of that data. Data archiving is the set of processes and management of archival data over time to ensure its long-term preservation, accessibility and security.

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. Big Data uses analytics for large and complex data sets continually increasing the value of previously untouched archival data while adding pressure to improve the management and security requirements of the archive. Various industry studies indicate less than 10% of all stored digital data is ever analyzed (it may have an occasional reference) and that over 40% of all stored data hasn't been accessed at all in the past 6-12 months. 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. All this is changing as Big Data re-awakens the archives.

Much of today's archival data is created as unstructured data which typically is formatted as bitmaps, images, objects, plain text, photos, and video streams. Unlike structured data, unstructured data is not part of a database and has limited if any 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, new file systems and numerous search engines.

***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 continue to 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	Archive – A *Move* Process	Offline – A *Manual* Process
Creates copy(s) of data for recovery purposes	Moves infrequently used data to new, more cost-effective storage solutions	Requires a direct human action to physically access storage
Disk, Tape, Flash, Cloud	Tape libraries, Cloud, Low-cost disk, local and remote tape Vaults	Tape, Low-cost disk, paper, offsite manual access media

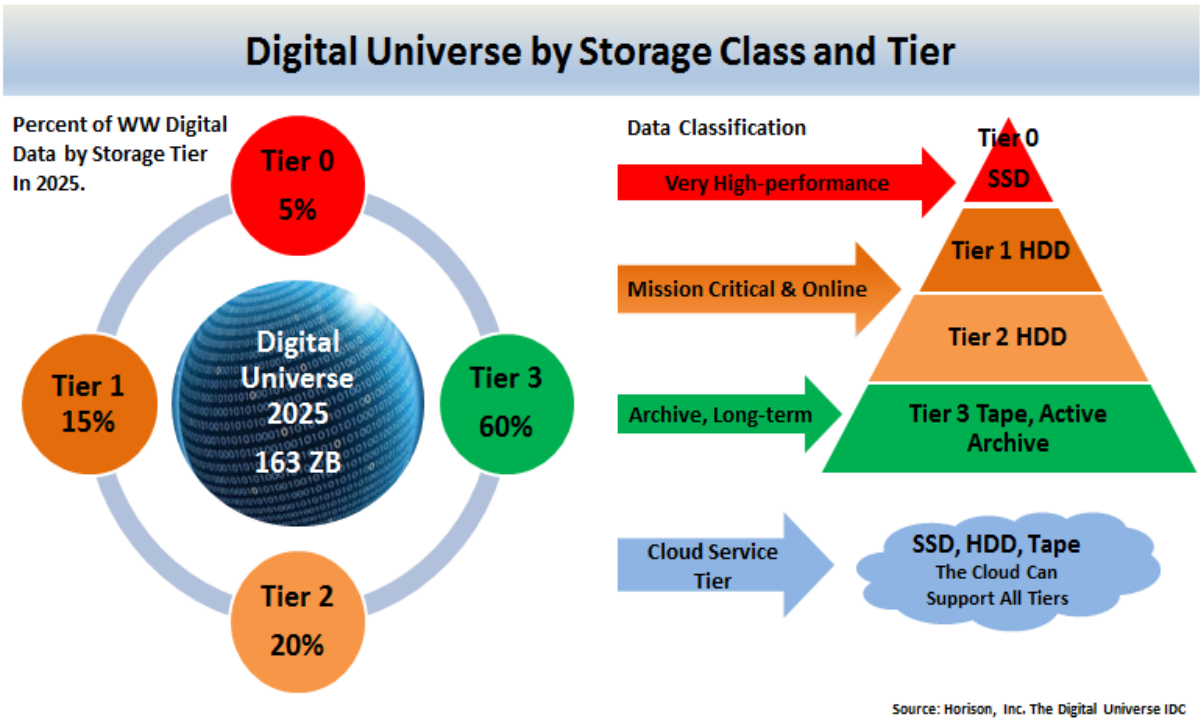
Archiving Reduces Pressure on the Backup Window

Though disk backup processes using compression or deduplication can help, the growing length of backup windows remains a major data center issue and is under constant pressure as growth rates exceed 30% annually as many data centers operate in 7x24x365 mode. There's no point in repeatedly backing up unchanged data – especially if it's seldom accessed – as this lengthens the backup cycle. Archiving can remove much of the low activity and unchanged data from the backup set to speed up the backup process.

Key points: ***Backup and archive are not the same. Backup occurs on your time – recovery occurs on company time. Archive moves the original data to 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. IDC's most recent report projects the digital universe in 2025 to total as much as 163 ZB (zettabytes) though a portion of this data will be short-lived reducing the net storage requirement. Given the magnitude of this projection, building an effective archive strategy along with an optimized tiered storage strategy can't begin soon enough for most businesses. By 2025, using industry data classification averages, it is anticipated that most all digital data will be stored on Tier 0 SSD (5%), Tier 1 and Tier 2 HDDs (35%) and Tier 3 tape or an Active Archive (60%). Other storage solutions may appear but need considerable progress to become relevant. Classifying your data by value, performance and capacity requirements will enable the right data to be in the right place at the right time.



Basic Steps for Building a Long-term, Secure and Scalable Data Archive 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. Most plans make provisions for more than one copy of archived data. Of course, if you don't want to deal with the growing amount of archival date, a cloud provider will be a viable option. Remember to keep it simple – let's get started.

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, future data value
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. The tape "air gap" prevents most cybercrime
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, HSM software, metadata management)	HSM (Hierarchical Storage Management) or policy driven archive software products monitor data reference patterns and metadata, applies user-defined policies to determine which data should be dynamically moved to archive status
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 facilities providing geographical redundancy for recovery and business resumption
Step 7	Set Rules for Who Can Access the Archives	Assign security codes, passwords, forensic IDs, for those in Charge! Identify each authorized person who can access the archive

Source: Horison Inc.

As many companies are painfully discovering, coping with rapid accumulation of archival data cannot be cost effectively achieved with a strategy of continually adding capacity with more costly disk drives. From a capital expense perspective, the cost of acquiring disk drives and keeping them functional can easily spiral out of control. From an operational expense perspective, the deployment of additional disk arrays increase spending on administration, data management, floor space and energy compared to more efficient tape solutions as the data repository increases in size. Unlike disk, tape scales capacity by adding more media, not more drives, making tape a more cost-effective archival solution.

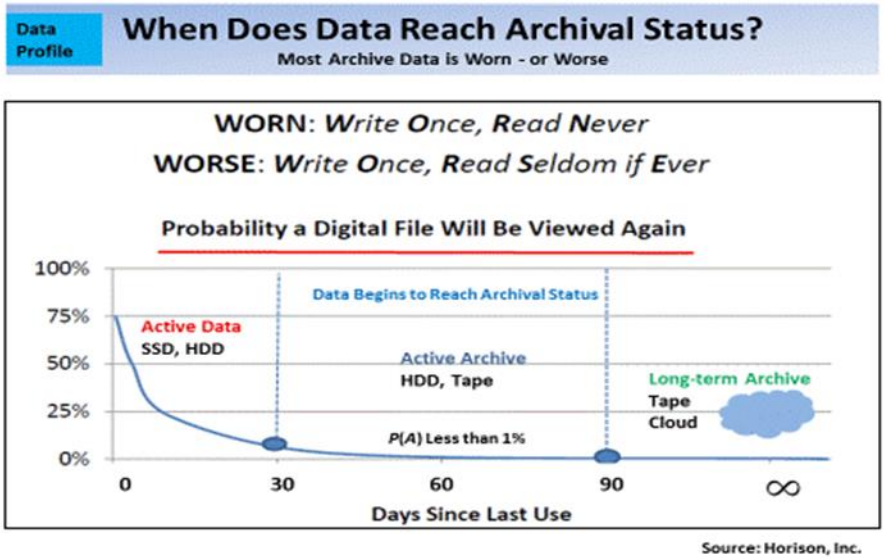
Key point: ***Data archiving is a comparatively simple process to understand, but can become a challenge to implement without a plan. It's time to get started before the pandemonium begins.***

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. 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 solutions using policies or metadata-based software from a variety of companies. All data is not created equal and determining data value is a key process for effective data management.

	**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 costlier enterprise class disk arrays and SSDs requiring multiple replication or backup copies that can be 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 certain business functions and often resides on enterprise and lower-cost disk subsystems.
	**Sensitive data** is information that might result in loss of an advantage or level of security if disclosed to others. Sensitive data comprises an average of 25 percent of all data stored but doesn't require immediate recovery capabilities. The RTO can take up to several hours without causing major operational impact. Sensitive data normally resides on low cost disk arrays and automated tape libraries.
	**Archive and Non-critical** data typically represents 40 percent or more of all digital data. Lost, corrupted or damaged data can be reconstructed with less complex recovery techniques requiring minimal effort, and acceptable recovery times can range from several 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 and 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 based on internal policies, customer and business partner requirements, and compliance data. As most data ages since its creation, the probability of reuse declines. Many files begin to reach archival status after the file has aged for a month or more, and whenever the *P(A)* (probability of access) falls below 1%, often after three months. 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 necessary retention periods. Archive policies can include last access date, age of data, frequency of access, pre-set capacity thresholds, deletion criteria and will address numerous specific legal retention regulations. These policies should be reviewed periodically as archival requirements can change based on a variety of circumstances or new 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. Tape is protected against most cybercrime attacks given the tape "air gap" prevents electronic access to most all tape storage.

Step 4 – Determine How Long to Retain Data

This step defines when data is no longer needed, when data reaches end of life, and its final disposition criteria. 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 pandemic data will be kept forever for future trend analysis. Most of today's video, television programming, sports events, movies, news items 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) data-mover type software. The HSM management system monitors access and usage patterns and makes user-defined, policy or metadata based decisions as to which data should 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 be deleted once its useful life has expired. Some HSM software products also provide backup and recovery functionality. Artificial intelligence (AI) will likely be used by these solutions to make even better decisions in the future.

Examples of HSM and Archive Products	Vendor
DFHSM, Tivoli Storage Mgr. (IBM Spectrum Protect), HPSS	IBM
StorNext	Quantum
SAM-QFS	Oracle
DMF	SGI
DiskXtender End of Life – Replaced by Seven10 Storfirst	EMC/Dell
NetBackup Storage Migrator	Veritas (Symantec)
HPE Storage Software	HP
CA-Disk	CA
Simpana	CommVault
Dternity	Fujifilm
Versity Storage Manager	Versity

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?

In addition to classifying data, storage platforms are classified based on criteria such as performance, capacity, access speed, reliability and cost. Several new and important technologies have been implemented for tape yielding numerous improvements including unprecedented cartridge capacity increases using BaFe media, vastly improved bit error rates compared to disk, and much 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 beginning with the LTO-5 tape drive, 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 positions tape to more effectively address archive requirements by enabling tagging of files with descriptive text, allowing for faster and more intuitive searches of cartridge and library content. LTFS has helped tape become the most viable, lowest cost, most reliable cloud archive solution.

Cloud Archiving Providers Using Tape

The inherent consolidation of data into large-scale cloud storage systems signals that another category of storage is emerging – tape in the cloud. 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 however it is a very expensive option. A disk drive can consume from 7 W to 21 W of electrical power every second to keep them spinning and even more energy is needed to cool them. The TCO advantage for tape is expected to become even more compelling with future technology developments. The chart below compares key archival considerations for tape compared to disk to implement an optimized archive infrastructure.

Tape and Disk Comparison for Building the Optimal Digital Archive

Archive Capability	Tape	Disk
TCO	Favors tape for backup (2-4:1) and archive (15:1) over disk	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) @ $1x10^{19}$ versus $1x10^{16}$ for 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 – encryption and WORM	Encryption and WORM capability available on all LTO and enterprise tape. The tape "air gap" prevents hacking	Becoming available but seldom used on selected disk products, PCs and personal appliances.
Capacity growth rates	Roadmaps favor tape over disk for foreseeable future with 220 TB demo by Fujifilm and IBM	Slowing capacity growth as roadmaps project disk capacity to lag tape for foreseeable future
Scale capacity	Tape scales by adding cartridges	Disk scales by adding more drives
Data access time	LTFS and the Active Archive improve tape access time	Disk is faster than tape for initial access and random-access applications
Data transfer rate	360 MB/sec for TS1155 tape, 300 MB/sec for LTO-7	Approx. 175 MB/sec for disk
Portability - Move media to different location for DR with or without electricity	Yes, tape media is completely removable and easily transported in absence of data center 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, encryption keys and forensic markers should only be assigned to those who have authorized access to the archives. Remember to appoint a team leader. Somebody must 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

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 most of this data will eventually wind up back in a large business, service provider or a cloud provider data center. The applications listed in the chart below all create significant volumes of data that become archival as the data ages.

Archive Applications Driving Storage Demand

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 (AI), Space Management, Integrated Tape, Disk and Flash	
Advanced LTFS partitioning for enhanced tape access, Recommended Access Order (RAO)	

Source: Horison Inc.

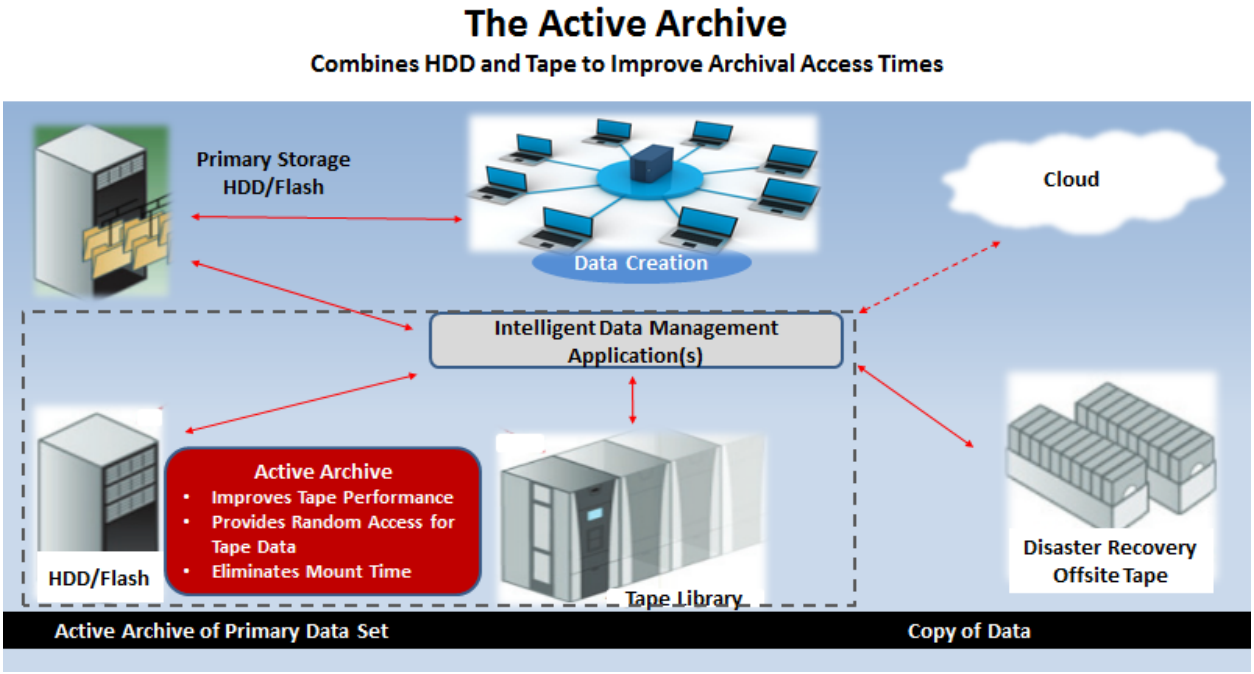
The size and value of archival data are increasing and researchers are quickly discovering the benefits of analyzing very large objects, files and datasets. For many data types, the lifetime for data preservation

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 1/6th to 1/15th that of disk for archival storage, and with reliability having surpassed disk drives, the pendulum has shifted to tape to address much of the tier 3 demand.**

The Active Archive Gains Momentum and Combines Disk and Tape

An Active Archive provides a persistent online view of archival data by integrating one or more storage technologies (flash, disk, tape *and* cloud storage) behind a file system that gives users a seamless means to manage their archive data in a single virtualized storage pool. Disk serves as a cache buffer for the archival data on tape and provides higher IOPs and random access to more active data in the large tape archive. Data in Active Archives are indexed and have search capabilities so files and parts of files can be more easily located and retrieved. Using LTFS and a 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 which combines a NAS with HDDs, LTFS, and a tape library. The Active Archive with 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 in conjunction with implementing an Active Archive. The Active Archive concept is supported by the Active Archive Alliance. See Active Archive conceptual view below.



Conclusion

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. 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. Each organization will have to justify their archive implementation to senior management in their own way, but the strategy to move low-activity data to the optimal storage tier for secure, long-term retention immediately yields significant cost savings with improved security. The bottom line is that your business-value case for data archiving will 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 seem to have no limits while tape technology continues to make tremendous strides – what timing! The future role of tape in archival storage cannot be denied and the sizeable cost savings of using tape compared to disk for archival storage promise to become even more compelling in the future. The Active Archive adds fast initial access times and performance to tape-only archives. Tape densities will continue to grow and tape costs will decline, while disk drive performance is expected to remain flat and capacity growth rates slow. 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 optimal choice for data archiving for the foreseeable future. Are you prepared to address these enormous archive challenges that lie ahead?

Summary: ***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.***