Abstract

Relentless digital data growth is inevitable 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 mandating the emergence of an ever smarter and more secure long-term storage infrastructure. Data retention requirements vary widely, but archival data is rapidly piling up. Digital archiving is now a required discipline to comply with government regulations for storing financial, customer, legal and patient information. As businesses, governments, societies, and individuals worldwide increase their dependence on data, archiving and data preservation become highly critical.

Most data typically reach archival status in 90 days or less, and archival data is accumulating at over 50% compounded annually. Many data types are being stored indefinitely anticipating that eventually its potential value might be unlocked. Industry surveys indicate nearly 60% of businesses plan to retain data in some digital format 50 years or more and a growing amount of data, much for historical preservation, will never be modified or deleted. 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". *It's time to develop your game plan!*

What is Archival Data?

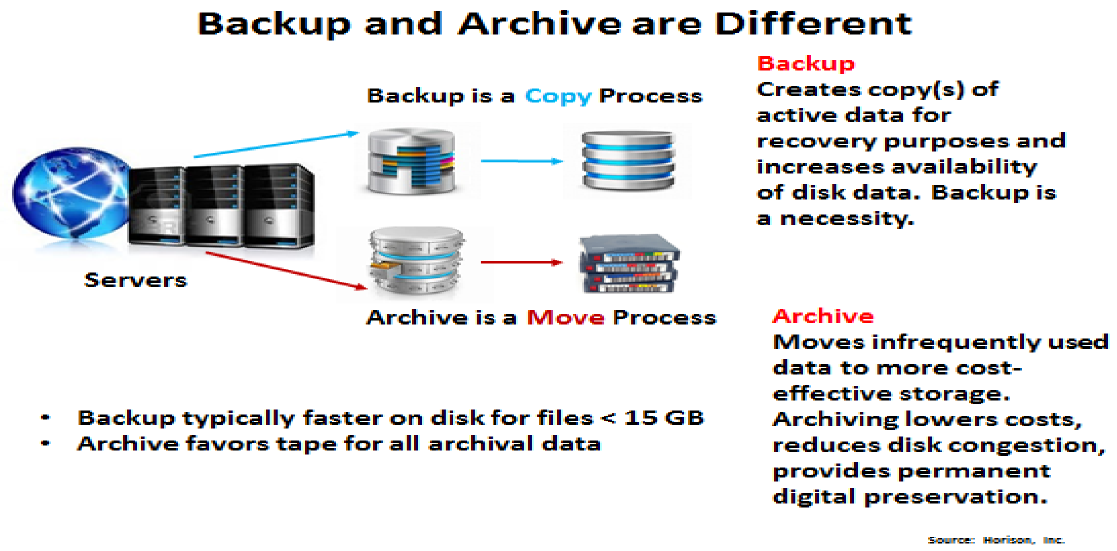
Simply stated, archival data is data that is infrequently used and seldom if ever changes - but potentially has significant value and needs to be securely stored and accessible indefinitely. Data archiving is the set of processes and management of archival data over time to ensure its long-term preservation, accessibility and security. A key benefit of data archiving is that it reduces the cost of primary storage and also reduces the volume of data that must be backed up. Removing infrequently accessed data from the backup data set improves backup and restore performance and frees up a lot of storage capacity. Do it!

***Key point:* Archives are no longer a repository for low-value data. Effectively managing the fast-growing 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 the backup and archive processes – 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. Some archives treat archive data as read-only to protect it from modification, while other data archiving products treat data as read and write capable. Data archiving is most suitable for data that must be retained for historical, future data mining and regulatory requirements.



Backup and Archive are Different

Backup
Creates copy(s) of active data for recovery purposes and increases availability of disk data. Backup is a necessity.

Archive
Moves infrequently used data to more cost-effective storage. Archiving lowers costs, reduces disk congestion, provides permanent digital preservation.

- Backup typically faster on disk for files < 15 GB
- Archive favors tape for all archival data

Source: Horison, Inc.

Archiving Reduces Pressure on the Backup Window

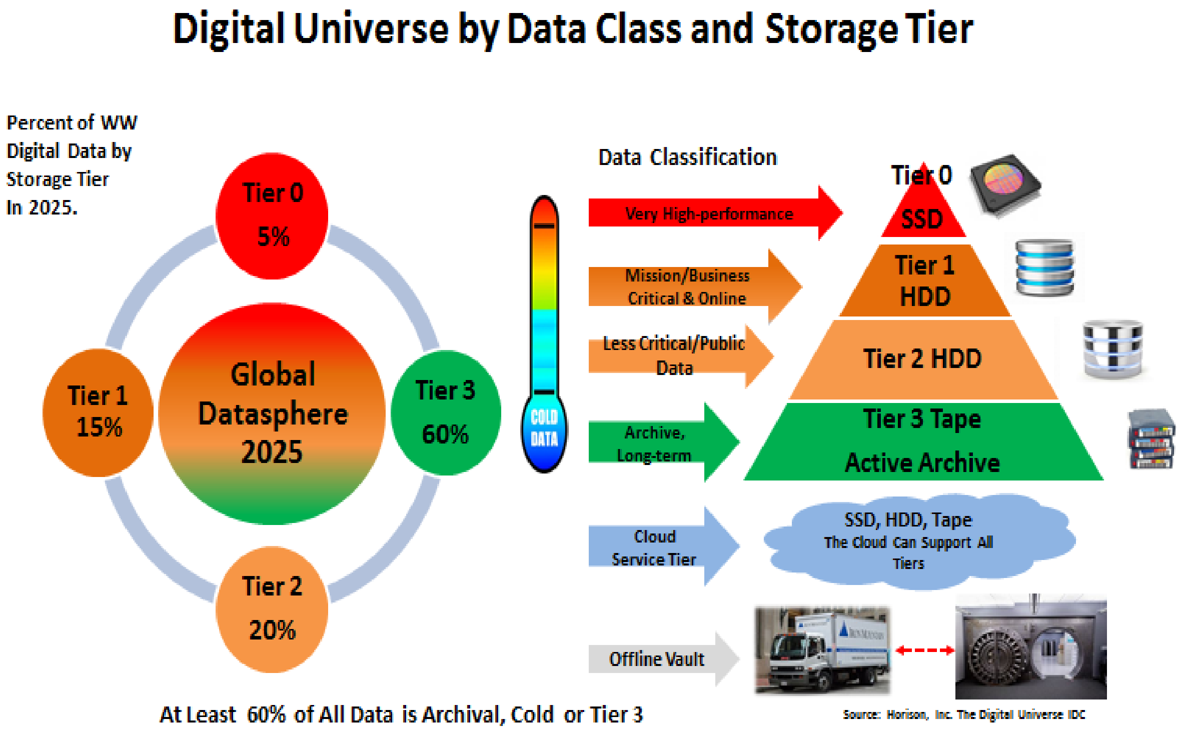
Studies indicate that as much as 85% of an organization's data is historically valuable, rarely accessed and cannot be deleted and as much as 60% of that data typically resides on disk drives. There's no point in repeatedly backing up unchanged data – especially if it's seldom accessed – as this lengthens the backup cycle. Archiving can remove much of the low activity and unchanged data from the backup set to speed up the backup process and free up storage capacity in the process.

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 data growth rates exceed 30% annually and many data centers now operate in 7x24x365 mode.

***Key points:* Backup and archive are not the same. Backup occurs on your time – recovery occurs on company time. Archiving moves the original data to more cost-effective location for long-term storage.**

How Much Data is Archival?

IDC's most recent digital universe report projects by 2025 the Global Datasphere will total as much as 163 ZB (zettabytes - 1x10^{21} bytes) though most of this data will be transient (short-lived) and not result in any net storage requirements. By 2025, using standard industry-wide data classification averages, it is anticipated that most all digital data *should* optimally be stored on Tier 0 SSD (5%), Tier 1 and Tier 2 HDDs (35%) and Tier 3 tape or an Active Archive (60%). Note: tier 3 is referred to as the tape tier or archive tier.


Digital Universe by Data Class and Storage Tier
Percent of WW Digital Data by Storage Tier In 2025.
Tier 0 5%
Tier 1 15%
Global Datasphere 2025
Tier 3 60%
Tier 2 20%
COLD DATA
Data Classification
Very High-performance
Mission/Business Critical & Online
Less Critical/Public Data
Archive, Long-term
Cloud Service Tier
Offline Vault
Tier 0 SSD
Tier 1 HDD
Tier 2 HDD
Tier 3 Tape Active Archive
SSD, HDD, Tape
The Cloud Can Support All Tiers
At Least 60% of All Data is Archival, Cold or Tier 3

Source: Horison, Inc. The Digital Universe IDC

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

Are you prepared to manage the avalanche of archival and permanent data that lies ahead? 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. New solutions are steadily appearing and will include Artificial Intelligence (AI). AI will go mainstream in the enterprise, transforming business with intuitive, out-of-the-box AI experiences and provide a huge assist to the entire data management discipline in the not-too-distant future.

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 data, a cloud provider can be a viable option. Remember to keep it simple.

Steps	Archive Strategy	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, 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, forever? 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, AI on the horizon)	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 or deleted
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 businesses 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 increases spending (TCO) 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 and scalable 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 arrives.**

Data Classification by Value and Criticality

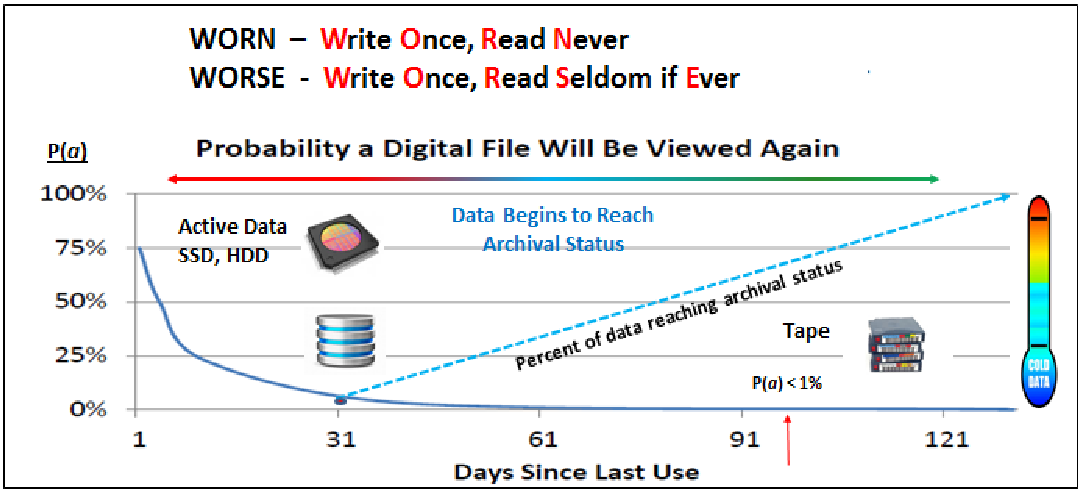
All data is not created equal and classifying data value is a key process for effective data management and to protect data throughout its lifetime. Though you may define as many levels as you want, four de-facto standard levels of classifying data are commonly used: mission-critical data, vital data, 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 been labor intensive but will greatly benefit from AI (Artificial Intelligence) in the near future. De-facto standard data classification guidelines are in the chart below.

	Mission-critical data defines the most important revenue generating business processes, customer facing applications and typically accounts for about 15 percent of all stored data. 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. Social security and credit card numbers are in this category.
	Vital data averages up to 20 percent of all stored data; however, vital data doesn't require "instantaneous" recovery to remain in operation. Vital data is critical to certain business functions and often resides on enterprise and lower-cost disk. If lost or disclosed could negatively affect operations.
	Sensitive data is information that might result in loss of an advantage or level of security if disclosed to others and is not meant for public disclosure. Sensitive data comprises an average of 25 percent of all data stored but doesn't require immediate recovery capabilities. Sensitive data normally resides on low cost disk arrays and automated tape libraries.
	Archive and less-critical data typically represents 40 percent or more of all digital data. Lost or damaged data can be recovered requiring minimal effort, and acceptable recovery times can range from several hours to days since this data is normally not critical for business survival. However less-critical data doesn't mean it isn't valuable. Less-critical data may suddenly become highly valuable based on unknown circumstances and, because of this characteristic, is most cost-effectively stored on tape.

Determine When Data Should Be Archived

Establish the criteria for what types of data and when 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.

When Does Data Reach Archival Status?



Source: Horison, Inc.

Software Solutions to Activate the Archive Process

Archives are best managed by Hierarchical Storage Management (HSM) data-mover or similar 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. Several HSM software products also provide backup and recovery functions. AI will likely be added to these solutions to make even better and less labor-intensive decisions in the future.

Examples of HSM and Archive Software	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	HPE

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.

Online, Offline and Cloud Storage

Data archives can take a number of different forms. Some systems may intentionally use online storage, which places archive data onto disk systems where it is readily accessible. Other archival systems use offline data storage (no electrical connection) in which archive data is normally stored on tape rather than being kept online. Storing archival data on tape in the cloud represents a significant growth opportunity for tape providers and a much lower cost, more secure archive alternative than disk for cloud providers; a win-win. Because tape media can be removed and stored offline, tape-based archives consume far less power than disk systems and offer the added benefit of cybercrime protection with its "air gap". Amazon Glacier and Microsoft Azure are examples of large-scale cloud storage services designed for data archiving and backup that use tape.

Digital Archives Embrace Object Storage

Archiving was the initial enterprise use case for object storage providing scalable, long-term data preservation. Object storage is popular with cloud providers and enables IT managers to organize archival content with its associated metadata into containers to easily allow retention of massive amounts of unstructured data. In July, 2017 IBM Spectrum Archive™ Enterprise Edition V1.2.4 which uses LTFS, announced a connection with OpenStack Swift to enable the movement of cold (archive) data from object storage to more economical tape and cloud storage for long-term retention. LTFS now provides a back-end connector for open source SwiftHLM (Swift High Latency Media), a high-latency storage back end that makes it easier to perform bulk operations using tape within a Swift data ring.

Comparing Disk or Tape for Archiving

Disk *can* be used for archival storage however it is an 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. Cloud storage uses disk and tape and is relatively inexpensive, but cloud data retrieval/transfer costs can soar as the amount of data transferred increases. The chart below compares key archival considerations for tape compared to disk to implement an optimized archive infrastructure.

Archive Capability	Tape	Disk
TCO	Favors tape for archive as much as 6-15x over disk	Much higher TCO, more frequent conversions and upgrades
Long-life media	30 years or more on all new	~4 years for most HDDs before

	enterprise and LTO media favoring archive requirements	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, WORM	Encryption and WORM 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 - native 200+ TB cartridge have been demonstrated	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, the Active Archive and RAO improve tape access time	Disk is faster than tape for initial access and random-access apps
Data transfer rate	360 MB/sec for TS1155 tape, 360 MB/sec for LTO-8, RAIT multiplies tape data rates	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:* *The tape industry continues 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.

Storage Intensive Applications Reawakening the Archives

At the beginning of this century, 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 data center, service provider or a cloud provider data center. Organizations are quickly learning the value of analyzing vast amounts of previously untapped archival data. For example, Big Data uses analytics and data mining for very large and complex data sets continually increasing the value of previously untouched archival data while adding pressure to improve the management and security capability of the archive. Various industry studies indicate less than 10% of all stored digital data has actually been 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 reaching the order of petabytes ($1x10^{15}$), exabytes ($1x10^{18}$) and will approach zettabytes ($1x10^{21}$) of data in the foreseeable future. The applications listed below all create significant volumes of data that become archival as it ages.

Applications Driving Archive Storage Demand	
Digital Assets (Fixed Content), Documents, Printed Materials, Books, Magazines	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
Government Compliance, GDPR, Legal Materials	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.
Internet, Social Media, Static Images, Digital Photo Repositories, IoT	Digital Surveillance/Security, Motion Sensors, Forensics
Archival Storage Futures…	
Automated Tiered Storage for SSD, HDD and Tape (AI and advanced software)	
Intelligent Active Archive – Pre-staging (AI), Space Management, Integrated Tape, Disk and SSD	
Advanced LTFS partitioning and Recommended Access Order (RAO) for faster tape access	

Source: Horison Inc.

For many data types, the lifetime for data preservation has become "infinite" and will constantly stress the limits of the archive infrastructure as much data will never be deleted. The size of preserving 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.

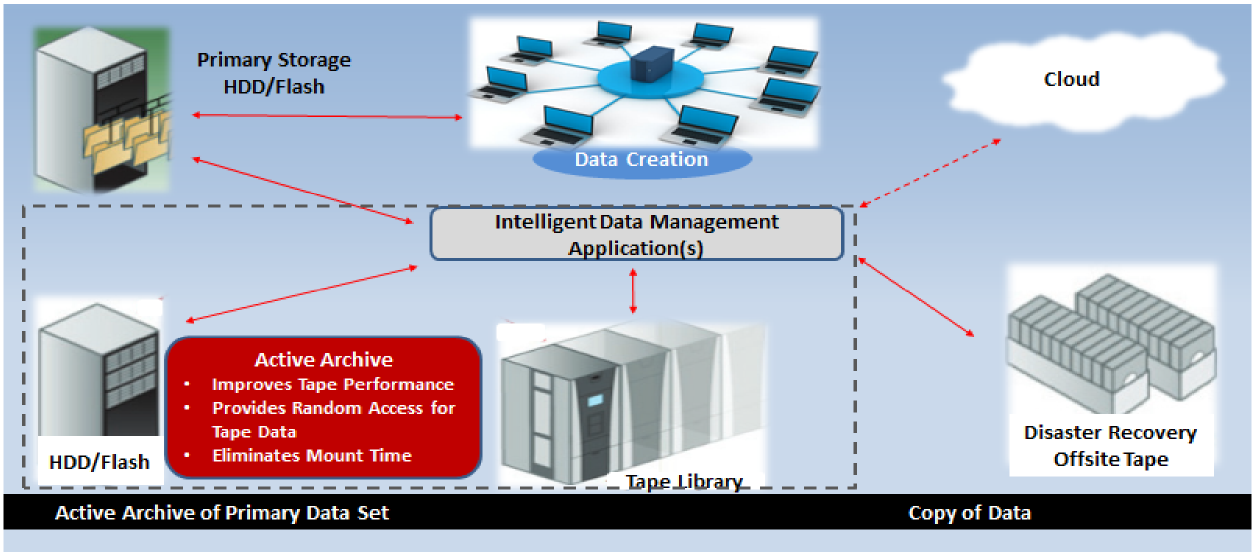
***Key point:* With tape now having a TCO of 1/6th to 1/15th of disk for archival storage, and with reliability having surpassed disk drives, the pendulum has shifted to tape to address much of the tier 3 demand.**

The Active Archive Combines Disk and Tape for Even Better Performance

The Active Archive provides a persistent online view of archival data by integrating one or more storage technologies (SSD, 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. Using LTFS, a data mover software solution and a disk array or NAS in front of a tape library creates an Active Archive. The Active Archive with LTFS and tape partitioning have barely scratched the surface of their potential and has yet to introduce AI to its functionality. Expect an increasing number of ISVs (Independent Software Vendors) to exploit LTFS in the future in conjunction with implementing Active Archive solutions. The Active Archive concept is supported by the Active Archive Alliance. See Active Archive conceptual view below.

The Active Archive

Combines HDD and Tape to Improve Archival Access Times



Conclusion

A strategy to move low-activity, but potentially valuable 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 for 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 growth and 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 promises to become even more compelling in the future. Tape densities will continue to grow, and tape costs will steadily decline, while disk drive performance is expected to remain flat and capacity growth rates have slowed. 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. The time has come 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.**