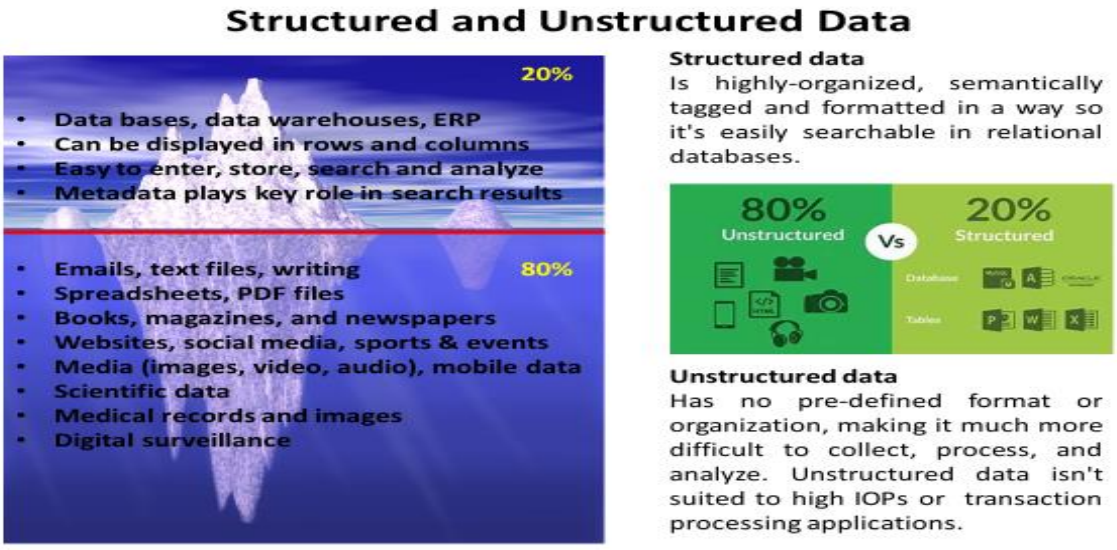
Abstract

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.

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 archival data 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. As businesses, governments, societies, and individuals worldwide increase their dependence on data, archiving and data preservation become a critical practice. *It's time to develop your game plan!*

What is Archival Data?

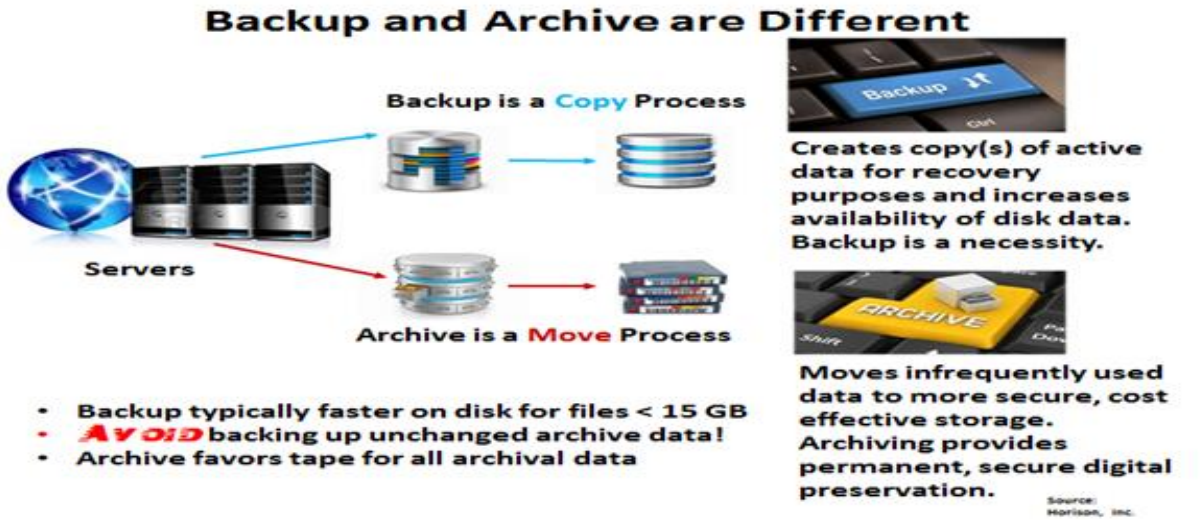
Simply stated, archival data is data that is infrequently used and seldom if ever changes - but potentially has significant value and needs to be securely stored and accessible indefinitely. A key benefit of data archiving is that it reduces the cost of primary storage and also reduces the volume of data that must be backed up. Big data refers to extremely large datasets that are difficult to analyze with traditional tools. The term big data is closely associated with unstructured data and most of this data soon becomes archival. Most estimates suggest 80% or more of all digital data is unstructured. Many of the tools being designed to analyze big data must address mountains of unstructured archival data to make it useful.



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.



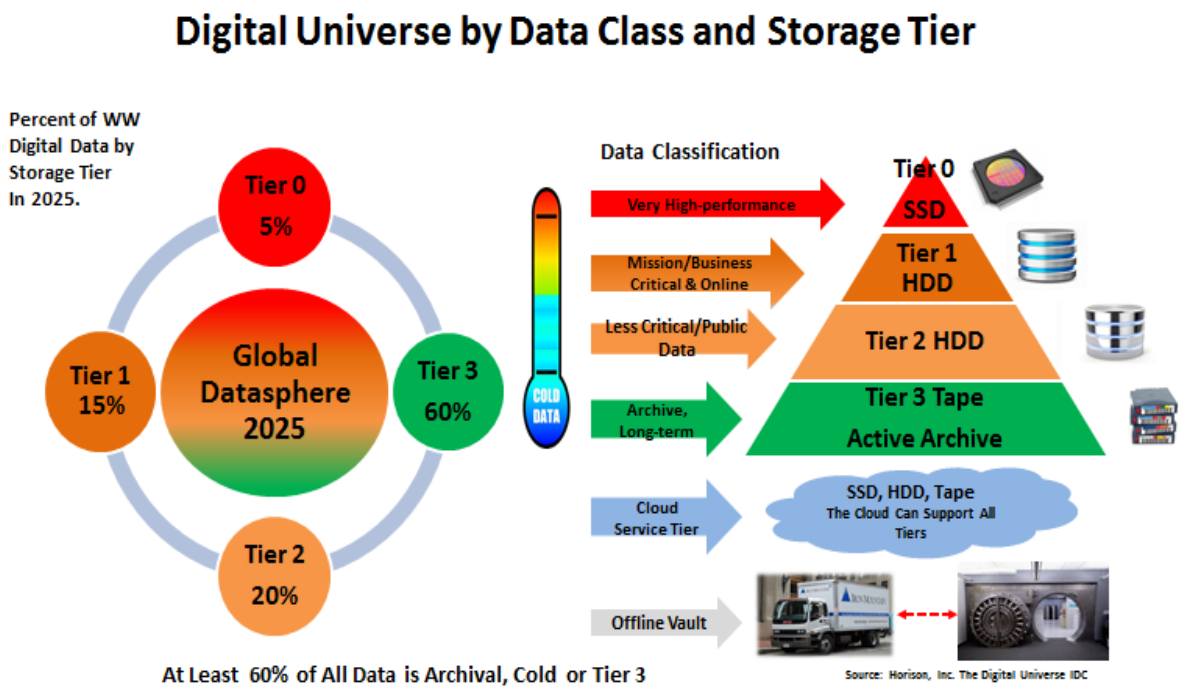
Archiving Reduces Pressure on the Backup Window

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 time and the amount of data to manage in the backup cycle. Archiving can remove much of the low activity and unchanged data from the backup set to speed up the backup (and restore) process and free up costly storage capacity in the process. Though disk backup processes using compression or deduplication can help, the growing length of backup windows remains a major issue and is under constant pressure as data growth rates exceed 30% annually.

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

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) not resulting in any net storage requirements. Most estimates indicate that over 80% of the retained data is unstructured. 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.



Basic Steps for Building an 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 and will 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 performance critical, mission-critical, business critical, 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 Data Retention Requirements and 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 software 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 effort, floor space and energy compared to more efficient tape solutions as the data repository increases in size. Unlike disk, tape capacity scales 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 Guidelines

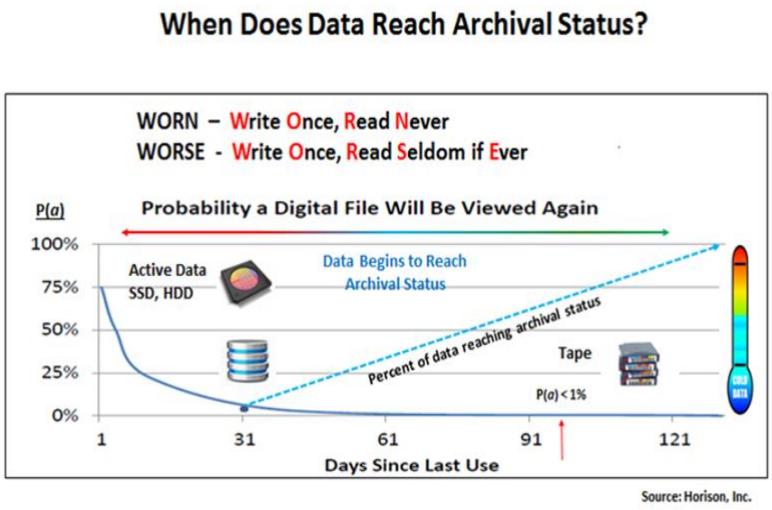
All data is not created equal and classifying data value is a key process for effective data management and to protect data throughout its lifetime. Though you may define as many levels as you want, four de-facto standard levels of classifying data are commonly used: mission-critical data, vital data, sensitive data and **archival data**. Data classification also aligns data with the optimal storage tiers and services based on the changing value of data over time.

Archival data presently represents over 60 % of all data stored making it the largest and fastest growing data classification segment. Setting the right data retention policies is a necessity for archival data and in particular for governance and legal compliance. At best, many businesses are fulfilling their industry's compliance and regulatory requirements, but not yet looking deeper into the strategic value of their stored information. Metadata is used to organize information so it's easier to discover and use**.** The use of descriptive metadata for unstructured files is becoming key archives as it describes a resource for purposes such as discovery and identification. For archival data, lost, corrupted or damaged archival data can be reconstructed with minimal effort, and acceptable recovery times can range from hours to days. The primary technology for storing archival data is automated tape libraries used in local, cloud and remote locations. Tier 3 archival data represents the prime growth opportunity for cloud providers.

When Does Data Reach Archival Status?

Archival data presently represents over 60 % of all data stored making it the largest and fastest growing data classification segment. Lost, corrupted or damaged archival data can be reconstructed with minimal effort, and acceptable recovery times can range from hours to days. At best, many businesses are fulfilling their industry's compliance and regulatory requirements, but not yet looking deeper into the untapped strategic value of their stored information. The optimal storage solution for archival data is tape technology used in robotic libraries, local, cloud and remote locations.

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 *P(A)* (probability of access) begins to fall after one month and typically falls below 1%, often after three months. See adjacent chart.



Software Solutions Automate the Archive Process

Archives are best managed by Hierarchical Storage Management (HSM) data-mover or similar types of archiving software. These management systems monitor access and usage patterns and make 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 adds to archiving many capabilities to manage storage devices efficiently, especially in large-scale user environments where storage costs can mount rapidly. 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. The process of moving files from one storage medium to another is known as *migration*. The apparently available files are known as *stubs* and are pointers to the real location of the migrated file in backup storage. Several HSM software products also provide backup and recovery functions. AI will likely be added to these solutions to make even more dynamic 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
StrongLink	StrongBox Data
Fujifilm Data Management Solutions	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 Used for Archive

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 taking advantage of the air gap 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. Some storage clouds are optimized to handle archiving (the green cloud). Amazon Glacier and Microsoft Azure are examples of large-scale cloud storage services designed for data archiving and backup relying on tape. Tier 3 archival data represents the prime growth opportunity for cloud providers.

Comparing Disk and Tape for Data Archiving

Disk *can* be used for archival storage however it is an expensive option compared to tape. 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. Tape also provides WORM (Write-Once-Read-Many) and encryption capabilities enabling a secure storage medium for compliance, legal and any types of valuable archival files. The Tape Air Gap adds significant protection against cybercrime and helps prevent ransomware attacks. The chart below compares key archival considerations for tape compared to disk to implement an optimized archive infrastructure.

Archive Functionality	Tape	Disk
TCO	Favors tape for archive as much as 6-15x over disk and cloud	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	~4-5 years for most HDDs before upgrade or replacement, 7-8 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. 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 has 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, TAOS and RAO improve tape access time	Disk is much faster (ms) than tape (secs) for initial and random-access
Data transfer rate	400 MB/sec for TS1160, 360 MB/sec for LTO-8, RAIT multiplies tape data rates	Approx. 160-220 MB/sec for typical HDD
Portability - Move media 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
Cloud Storage Archives	Tape Improves Cloud Reliability and Security, Lowers Archival Storage Costs, Unlimited Capacity Scaling	HDDs Become Expensive as Cloud Providers and Hyperscale data centers grow

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 Reawaken the Archives

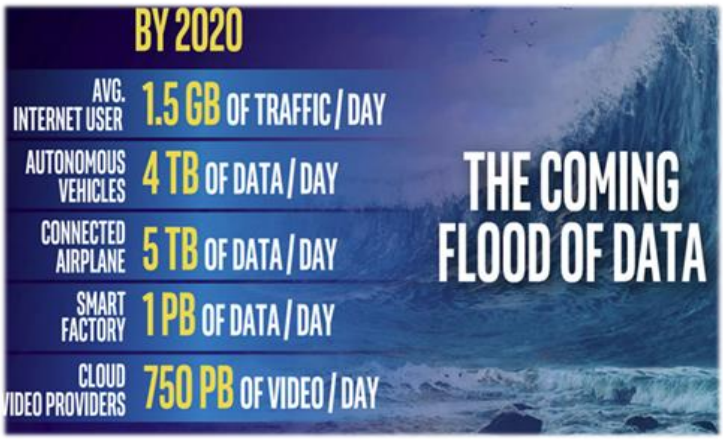
At the beginning of this century, large businesses generated roughly 90% of the world's digital data. Today it is estimated over 80% of all digital data is generated by individuals - not by large businesses – however most of this data will eventually wind up back in a large data center or cloud service provider's data center. Organizations are quickly learning the value of analyzing vast amounts of previously untapped archival data. For example, Big Data applications use analytics and data mining techniques on 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.

Several industries are adopting AI technology to improve managing their storage growth. Leading examples include autonomous vehicles (AV) development, media and entertainment (M&E), and video surveillance and security. All three industries generate massive data that they retain for research, reuse, and investigations.

AVs generate lots of data, which is approaching 100 TB a day on high-end AVs. The data is both sensor information that allows the AV to react to driving conditions, and statistics on AV safety. For example, proving that an AV model is as safe as a competent human driver takes 275 million miles of driving and testing per model.

M&E (Media and Entertainment) is traditionally a large tape user market segment because of tape's large capacity and data rate for massive broadcast files. M&E uses and reuses digital content to reach more customers and increase profits. The M&E industry has some of the most extensive experience with film archiving and data migration as preserving digital content indefinitely has become critical to their survival. The Sandvine "2018 Internet Phenomena Report" indicates that Netflix alone accounted for 15 percent of all global downstream traffic across the entire internet in 2017! Streaming services easily store and move thousands of petabytes of compressed data daily.

Video surveillance is rapidly evolving. Cameras generate richer, higher density images in response to advances in facial recognition and other data-intensive types of video. This does not require tape if retention periods are 30 days or less. However, many retention periods are required well past 30 days, especially in the intelligence and law enforcement communities leaving tape as the cost-effective and valuable surveillance solution.



Additional examples of large-scale archive applications include compliance data, GDPR, medical records, photos and images, e-mail history, unstructured file data, scientific, video, movies, audio, documents, collaboration, social media history, archive cloud applications, security system history and archives, off-site media storage, remote data vaults, and BC/DR.

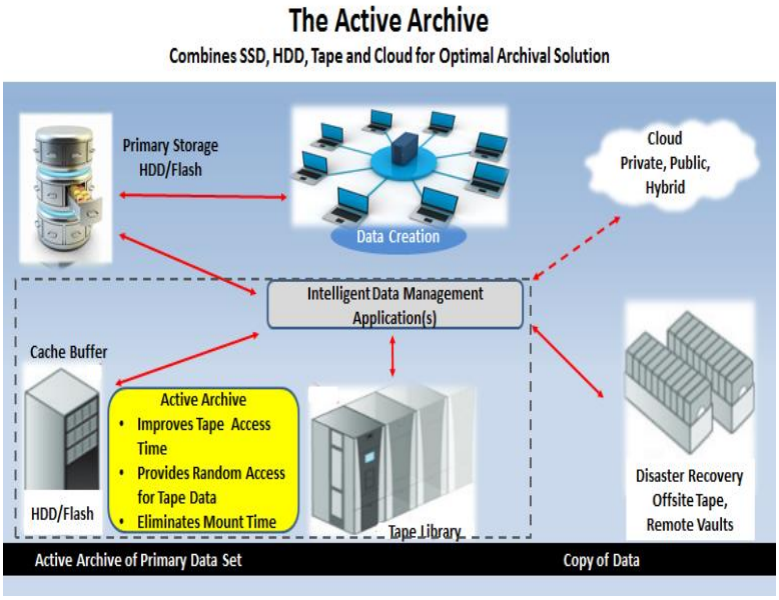
For many data types and files, the lifetime for data preservation has become "infinite" and will constantly stress the limits of the archive infrastructure as much data will never be deleted. The size of preserving digital archives are now reaching the order of petascale ($1x10^{15}$), exascale ($1x10^{18}$) 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 enormous tier 3 demand on the horizon.**

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 has 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 the adjacent Active Archive conceptual view.



Conclusion

A strategy to move low activity, but potentially valuable archival 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, added security, more efficient searches 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! 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.

Summary: ***The components to implement a cost-effective archive are now in place– 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.***