HORISON
Information Strategies

Fred Moore, President
Horison.com

2020 Technology Update Series

ARCHIVAL DATA STORAGE

Managing the Archival Upheaval

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 and it promises to play a much greater role over the next 30 years. Much of this data will be stored forever mandating the emergence of a more intelligent and highly secure long-term storage infrastructure. Data retention requirements vary widely based on the type of data, but archival data is rapidly piling up everywhere. Digital archiving is now a key strategy for larger enterprises and has become a required discipline for hyperscale data centers.

Many data types are being stored indefinitely anticipating that its potential value will eventually be unlocked. Industry surveys indicate nearly 60% of businesses plan to retain data in some digital format 50 years or more and much of this data will never be modified or deleted. For many organizations, facing terabytes, petabytes and potentially 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, data preservation and archiving has become a critical IT practice. *Fortunately, the required technologies are now available to manage the archival upheaval.*

WHAT IS ARCHIVAL DATA?

Archiving involves moving data that is no longer frequently accessed off primary systems to lower cost storage for long-term retention and protection. Archive retention requirements of 100 years or more are data is unstructured and includes office documents, video, audio, images, and basically anything not in a database. Big data is mostly unstructured data which is difficult to search and analyze with traditional methods. Fortunately, many of the tools that analyze big data are beginning to exploit metadata and global namespaces to make search and access capabilities much easier.

HOW MUCH DATA IS ARCHIVAL?

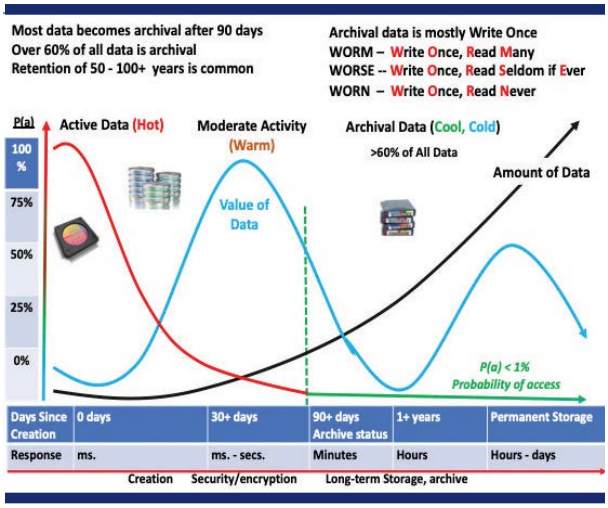
Industry estimates vary but the amount of data projected to be actually stored in 2025 is believed to be ~7.5 ZB according to IDC's 2018 Data Age report. The effects of the global Covid-19 pandemic on storage demand remain unclear.

What we do know is that approximately 1.1 ZB of total storage capacity was shipped in 2019 across Non-Volatile Memory devices (SSDs), HDDs, and magnetic tape media with HDDs making up the majority of the shipped capacity. Today at least 60% of all data can be classified as archival and it could reach 80% or more by 2025, making it by far the largest and fastest growing storage class while presenting the next great storage challenge.

Most archival data have never been monetized as the value of data remains unknown, but companies are just now realizing that digital archives have great potential value. Companies looking to be relevant between now and 2025 will need to understand the role archive data can play in their organization's success and how data archiving strategies will evolve during that period. Given this trajectory, the archival storage paradigm will need to reinvent itself.

WHEN DOES DATA REACH ARCHIVAL STATUS?

Archival data will continue to be the largest and fastest growing data classification segment. As data ages after its creation, the probability P(a) (probability of access) begins to fall after one month and typically falls below 1%, most often at 90 - 120 days. Some data becomes archival upon creation and can wait years for access or further analysis adding to the archival pile-up. Today the most cost-effective solutions for archival data are tape robotic libraries used in local, cloud and remote locations. See adjacent data lifecycle chart.



KEY POINT

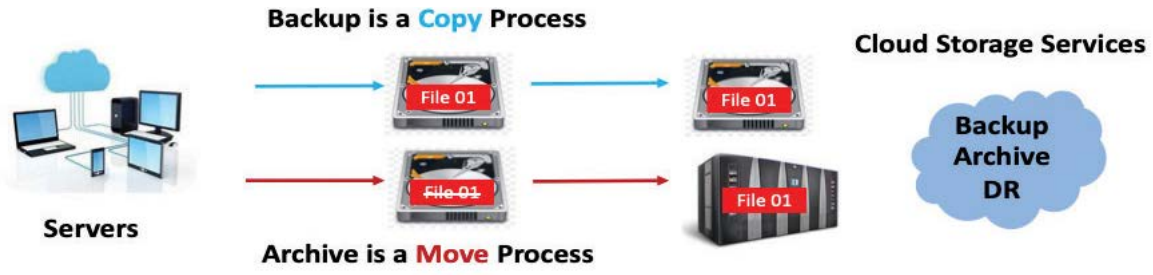
The true value of discovering what's behind the great wall of archival data remains unknown but offers the potential keys to alter the future. Effectively managing enormous digital archives is attainable and now requires a multi-tiered strategy.

DID YOU KNOW?

Backup and Archive Are Very Different Processes?

Many businesses continue to confuse the backup and archive processes often thinking they are 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 freeing up space on the source location. Most archive applications treat archive data as read-only to protect it from modification, while others treat archive data as read and write capable.

BACKUP AND ARCHIVE ARE DIFFERENT PROCESSES



BACKUP (Copies Data)	ARCHIVE (Moves Data)	ACTIVE ARCHIVE (Fast Access to Data)
Copies data for protection and recovery. Leaves source data in place.	Moves infrequently used data to more cost-effective storage. Frees up space on source devices.	A ***combined*** scalable solution based on intelligent active archive software, tiered media, and/or the cloud.
Restores lost files to desired point in time. Speed is critical factor.	***Retrieves*** files for reference and analysis. Retrieval speed is not a critical factor.	Provides rapid file- or object-level access to archived data.
Short-term duration averages 1-120 days.	Protects permanent and long-term data from modification or deletion.	Protects data and provides efficient data access. Typically operates without manual intervention.

ARCHIVING REDUCES PRESSURE ON THE BACKUP WINDOW

Studies indicate that as much as 85% of an organization's data is historically valuable, infrequently if ever accessed and seldom deleted. As much as 60% of that data typically resides on expensive to operate disk drives. 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 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. An active archive adds performance to an archive by using HDDs or SSDs as a cache front-end for a robotic tape library and is discussed later.

KEY POINT

Backup and archive are not the same. Backing up unchanging archive data is time consuming and very 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.

BASIC STEPS FOR BUILDING AN ARCHIVE STRATEGY

Data archiving is a relatively simple process to understand and can be successfully implemented given the advanced hardware and software solutions available today. The basic steps listed below provide logical guidelines to build a sustainable archive capability. You may choose to add additional steps to the process based on specific business and workflow needs. If you don't want to manage the growing amount of archival data, a CSP (Cloud Service Provider) can be a viable option. Remember the goal is to get the right data in the right place and keep it simple. This often means using hybrid solutions that strike a balance between on-premises and cloud-based storage.

STEPS	ARCHIVE STRATEGY	WHAT IT MEANS
Step 1	**Classify**	Determine if data is performance critical, mission-critical, business critical, or non-critical
Step 2	**Determine**	Which data to archive, how many copies needed, define what data needs to be stored for future reference
Step 3	**Determine**	When to archive data, set archive and security policies, identify last access date, age of data, and frequency of access, assign encryption and worm
Step 4	**Choose**	Data retention and deletion requirements for archive data - months, years, forever?
Step 5	**Select**	A software solution to automate and manage the archive process. (A policy-based data mover, HSM software, metadata management, AI software, use global namespaces)
Step 6	**Select**	The optimal archive storage platform, remote vault, local, hybrid or cloud options
Step 7	**Set Rules**	For who can access the archives, assign security codes, passwords, forensic Ids

Source: Horison, Inc.

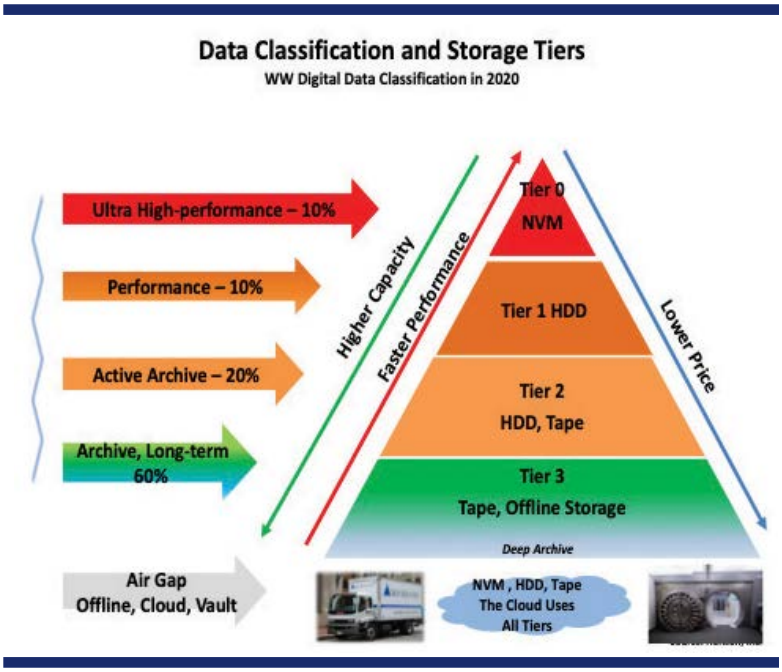
High capacity disk and tape store most of the world's archive data. However, coping with non-stop rapid accumulation of archival data cannot be cost effectively achieved using a strategy of just increasing capacity with more costly disk drives. From a capital expense perspective, the cost of acquiring disk drives and keeping them operational can quickly skyrocket. The deployment of additional disk arrays increases spending (TCO) on administration, data management effort, floor space and energy consumption compared to more cost-efficient tape solutions as storage demand grows. Unlike disk, tape capacity scales by adding more media, not more drives, making tape currently the most cost-effective and scalable archival solution.

KEY POINT

Data archiving provides significant economic benefits and is a comparatively simple process to understand but can become a challenge to implement without a plan. Doing nothing is a strategy - just not a very good one..

DATA CLASSIFICATION GUIDELINES

All data is not created equal making data classification the key process to managing data throughout its lifetime. Though you may define as many storage tiers as you want, four de-facto standard tiers of classifying data are commonly used: Ultra-high-performance data (OLTP), Performance Data (Mission critical), Active archive (lower activity data) and Long-term Archive. Data classification enables the alignment of data characteristics with the optimal storage technology tier.



A NEW STORAGE TIER?

Moving towards 2025, the momentum is building for the emergence of a new deep archive storage tier that provides a secure, infinite-life storage infrastructure for enormous amounts of data that is more price competitive than traditional archival solutions. Deep archives must be easily scalable and require minimal remastering cycles. Access times of hours or even days can be acceptable. The modern data tape roadmap with no fundamental areal density limitations is best positioned to addresses this tier. The emerging deep archive tier will offer great appeal to the hyperscale and CSP market.

CLOUD AND HYPERSCALE DATA CENTERS (HSDCS) PROVIDE SPECIALIZED ARCHIVAL SERVICES

Storing archival data in the cloud represents a significant growth opportunity for tape system providers and a much lower cost alternative (than disk) for cloud providers. HSDCs face insurmountable growth of disk farms which are devouring budgets and overcrowding data centers forcing data migration to tape solutions. Advanced tape architectures can now scale beyond an exabyte enabling HSDCs to shrink costs by providing the lowest TCO, highest reliability, the fastest throughput, and improved cybersecurity protection via the tape air gap. Cloud archive services are fairly inexpensive, but cloud data retrieval/transfer (bandwidth) costs can quickly soar as the amount of data transferred increases. Amazon Glacier, Amazon Glacier Deep Archive, and Microsoft Azure are examples of hyperscale (green) cloud storage services for data archiving relying heavily on tape to control costs.

SOFTWARE SOLUTIONS AUTOMATE THE ARCHIVE PROCESS

Archives are best managed by Hierarchical Storage Management (HSM), data-movers, or similar types of archiving software. These systems continually monitor access and usage patterns and make user-defined, policy-based decisions as to which data should be moved to archival status and which data should stay on primary storage, and which data can be deleted. The process of moving files from one storage medium to another is known as *migration*. The apparently available file(s) appear as *stubs* and are pointers to the real location of the migrated file in archival storage. The stubs are referenced to *recall* archived data for later usage. Several HSM software and data mover products also provide backup and recovery functions and have metadata assignment capabilities. Look for AI to play a role with these solutions in the future.

KEY POINT

Effective and intelligent archival software solutions are available to determine when data reaches archival status, where it should be stored, and how long it should be kept.

COMPARING DISK AND TAPE FOR DATA ARCHIVING

Tape and disk are today's primary options for large-scale data center archiving and share the archive market. A disk drive can consume from 7 W to 21 W of electrical power every second to keep them spinning and to cool them making energy costs a major component of overall disk TCO. Tape also provides WORM (Write-Once-Read-Many) and encryption capabilities enabling an immutable, secure storage medium for valuable archival files. The Tape Air Gap adds significant protection against cybercrime and cybercrime attacks by providing an electronically disconnected copy of data that hackers can't access. The table below compares tape and disk for key archival functions used to implement an optimized archive.

HSM, DATA MOVER AND ARCHIVE SOFTWARE PRODUCTS/VENDORS

DFHSM, Tivoli Storage Mgr. (IBM Spectrum Protect), HPSS
IBM

Fujifilm Data Management Solutions
Fujifilm

Fujifilm Object Archive Software
Fujifilm

Storcycle
Spectra

StorNext
Quantum

SAM-QFS
Oracle

NetBackup Storage Migrator
Veritas (Symantec)

HPE DMF (Data Management Framework)
HPE/SGI

CA-Disk
CA Technologies (Broadcom)

Hedvig Sofware (Replaces Simpana)
CommVault

Seven10 Storfirst (Replaces EMC Disk Extender)
EMC/Dell

StrongLink
StrongBox Data

Versity Storage Manager
Versity

ARCHIVE FUNCTIONALITY	TAPE	DISK
TCO	Favors tape for archive as much as 5-8x 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 tape	~4-5 years for most HDDs before upgrade or replacement
Reliability	Tape BER (Bit Error Rate) @ $1x10^{19}$ versus $1x10^{16}$ for disk	Disk BER has fallen behind tape by three orders of magnitude
Inactive Data Does Not Consume Energy	Most tape data doesn't consume energy. "If the data isn't being used, it shouldn't consume energy"	**TCO studies** indicate that disk is 90x more expensive for energy than tape and produces 89x more Co2
Highest Security Levels	Encryption and WORM available on all tape, "air gap" prevents hacking	Encryption and WORM are available, not frequently used on disk
Capacity Growth Rates	Roadmaps favor tape over disk for foreseeable future – 300-400 TB cartridge have been demonstrated	Slowing capacity growth as roadmaps project disk capacity to lag tape for foreseeable future
Scale Capacity	Tape can scale by adding cartridges	Disk scales by adding more drives
Data Access Time	LTFS, the Active Archive, TAOS and RAO improve tape file access time	Disk is much faster (ms) than tape (secs/mins) for initial access and provides random-access capability
Data Transfer Rate	400 MB/sec for TS1160, 360 MB/sec for LTO-8, RAIT multiplies data rates	Approx. 160-220 MB/sec for typical HDDs
Portability - Move Media for DR Without Electricity	Tape media is removable and easily transported to another location 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 Storage Costs	HDDs become very expensive as CSPs and Hyperscale data centers grow

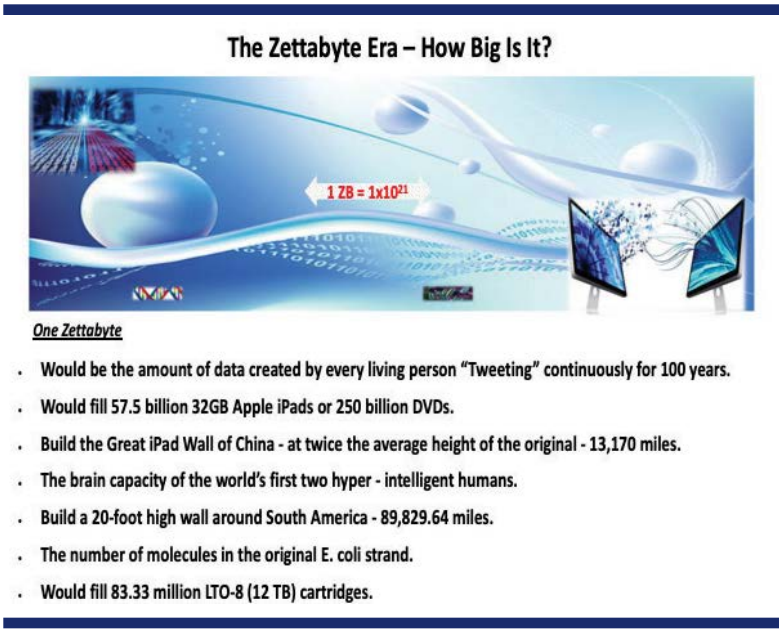
Source: Horison, Inc.

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 most cost-effective choice for archiving as well as playing a larger role for backup, business resumption and disaster recovery.

RUSHING INTO THE ZETTABYTE ERA—
Storage Intensive Applications Flood the Archives

For modern enterprise data centers with billions of files and petabytes of data, an effective archive strategy must deliver efficient data movement to move high volumes of data from primary storage to archive and for fast recall. Large-scale archive applications are now part of day-to-day business operations and include compliance data, GDPR, medical records, photos and images, e-mail history, scientific, video, audio, documents, collaboration, social media, archive cloud applications, off-site media storage, surveillance footage, remote data vaults, and BC/DR.



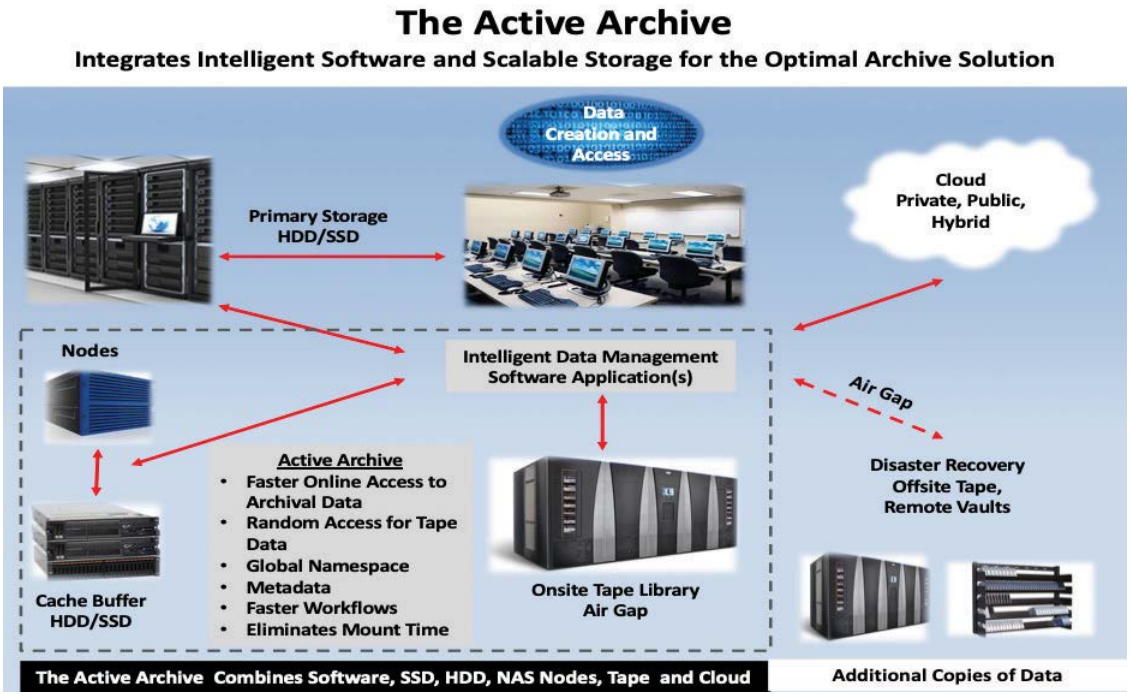
Some specific archive application and workload examples are highlighted below:

- **Financial** – Online banking transaction archives, ATM history, POS, audit logs, and communication logs.
- **Health Care and Life Sciences** – Electronic medical records, images (X-Ray, MRI, CT), genome sequences, pharmaceutical development data, and telemedicine will drive many new use cases. LTO with LTFS allows clinicians and administrators to quickly retrieve and share EMR, PACS, DICOM and other medical data that is typically only stored for a short time without workflow impact to any other department in the practice.
- **HPC** – Archives feed compute intensive applications for pattern recognition and simulate future consequences and predict outcomes. When the study is complete, the data becomes archival again.
- **Insurance** – Long-term accident records and images, health claims, disputes, payment history.
- **Media & Entertainment** – The M&E industry relies heavily on tape and digital archives to provide raw production footage. Most M&E content is never deleted and uses and re -purposes archival content to reach more customers and create new revenue streams. The M&E industry has extensive experience with film archiving and data migration as preserving digital content is a M&E mission critical function.
- **Online Advertising** – Clickstreams and ad delivery logs access legacy archives for new build sequences.
- **Physical Security and Surveillance** – Raw camera footage typically becomes archival after 7 days, surveillance retention periods are quickly increasing.
- **Science / Research / Education** – Archives provide research input and potentially new results, including data for seismic tests for oil & gas exploration, atmospheric science and predictive weather modelling.
- **Sports Archives** – The MLB Network archives over 1.2 million hours of content, which is indexed and stored with infinite retention periods and makes it available to the production team via proxy video.
- **Technology** – The IoT, mobile apps, autonomous vehicles, video, RADAR, LIDAR and sensor data generate data much faster than it can be analyzed creating enormous archives for future use.

THE ACTIVE ARCHIVE COMBINES DISK AND TAPE FOR IMPROVED ACCESS TIME

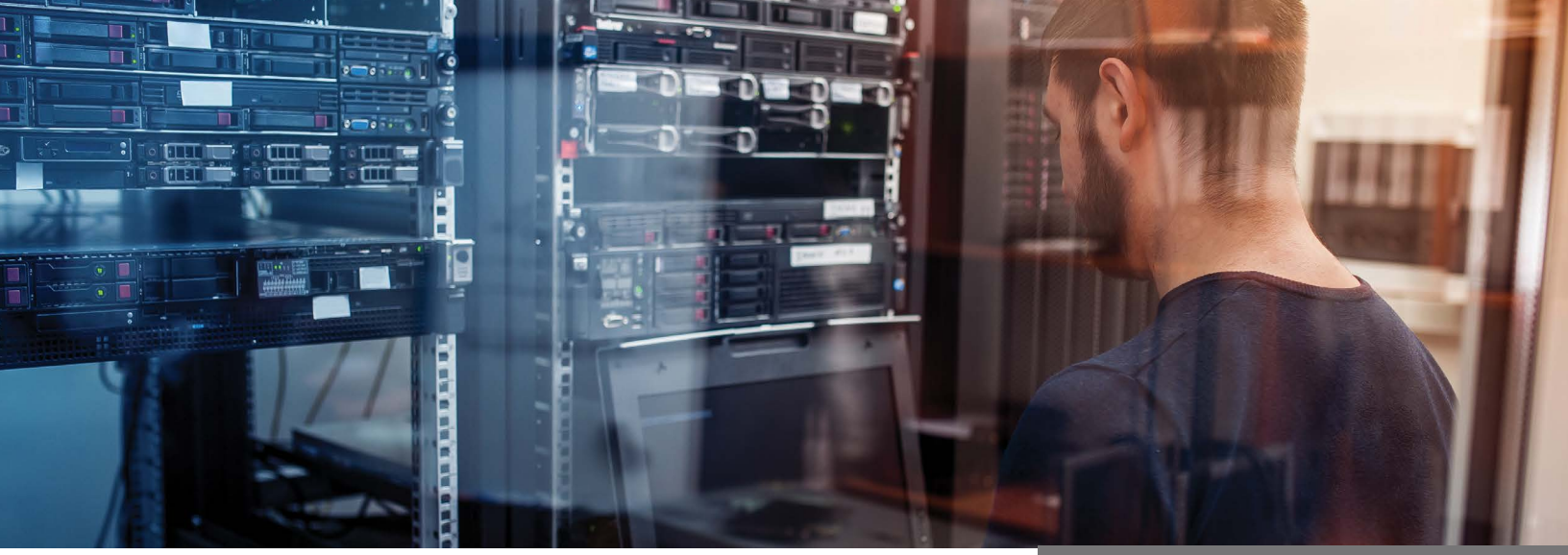
The Active Archive adds performance capabilities to archive storage systems. One or more storage technologies (SSD, disk, tape *and* cloud storage) are integrated behind a file system that gives users a seamless means to manage their archive data in a single virtualized storage pool. Disk serves as a cache buffer for the archival data on tape and provides higher IOPs, faster access to first byte of data, and random access for more active data in the large tape archive. Using LTFS, data mover software (HSM), and a disk array or NAS in front of a tape library creates an Active Archive.

The Active Archive with LTFS uses tape partitioning to improve access times and has barely scratched the surface of its potential. LTFS currently has 38 implementers and expect an increasing number of ISVs (Independent Software Vendors) to exploit LTFS in the future. The Active Archive concept is supported by the **Active Archive Alliance**. See the Active Archive conceptual view below.



KEY POINT

Large capacity disk will continue to play a key role by adding a performance layer for archives. However, the pendulum has shifted to tape to address the archival requirements that lie ahead.



CONCLUSION

The size of preserving digital archives is now reaching the order of petascale ($1x10^{15}$), exascale ($1x10^{18}$) and will approach zettascale ($1x10^{21}$) capacities in the foreseeable future. A strategy to move low activity, but potentially valuable archival data to the optimal storage tier immediately yields the greatest cost savings. Archive storage growth and requirements have no foreseeable limits and could demand a new *deep archive* storage tier in the next few years. Unless a new archival technology arrives, the numerous improvements in tape have made it the clear-cut optimal data archiving choice for the foreseeable future.

SUMMARY

The hardware, software and management 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. Not implementing an effective archive strategy is a missed opportunity. Now is the time to address the archival avalanche.

Horison Information Strategies is a data storage industry analyst and consulting firm specializing in executive briefings, market strategy development, whitepapers and research reports encompassing current and future storage technologies. Horison identifies disruptive and emerging data storage trends and growth opportunities for end-users, storage industry providers, and startup ventures.