Building the Archive of the Future

Horison Information Strategies
Fred Moore, President
www.horison.com

HORISON
Information Strategies

Abstract

Relentless data growth is inevitable as digital 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. Retention requirements vary widely based on the type of data, but archival data is rapidly piling up in every business. Given its potential value, modern data archiving has surpassed backup becoming a key strategy for enterprises and a required discipline hyperscale data centers.

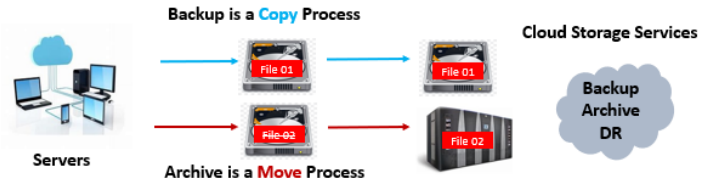
Many data types are being stored indefinitely anticipating that its value will eventually be realized. Industry surveys indicate nearly 60% of businesses plan to retain increasing amounts of data in some digital format from 50 to 100 years or more and much of this data will never be modified or deleted. 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. Fortunately, the required technologies and architectures are now available to manage the modern archival challenges.

Why Has Archival Data Become So Relevant?

Most archival data have never been monetized leaving its true value a mystery, but companies are just now realizing that untapped digital archives have great potential value. To realize the value of data, it must be put to work. Companies looking to be relevant between now and 2025 will need to understand the significant role archive data can play in their organization's success and how data archiving strategies will evolve during that period of time. The archival pandemic is rapidly spreading throughout the global datasphere. Given this trajectory, the archival storage paradigm will need to reinvent itself.

Backup and Archive Are Very Different Processes

Many businesses continue to confuse the backup and archive processes, often thinking they are the same thing. Backing up disk data has been the initial and primary use case for tape for over 50 years, but that mentality has transitioned to data archiving. 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. The more critical the data, the more critical the recovery speed becomes.



Backup (Copies Data)	Archive (Moves Data)	Active Archive (Fast Access to Archives)
Copies data for protection and recovery, source data left in place	Moves infrequently used data to more cost-effective storage, frees up space on source devices	A combined solution of intelligent software, SSD, HDD, and tape systems
Restores files to desired point in time in event of data loss, speed critical	Retrieves files for future reference and analysis, retrieval speed less critical	Uses HDDs or SDDs as a cache front-end for high capacity tape libraries
Cyclic process, overwrites itself when end of retention time is reached	Forever growing, typically unchanging and is not overwritten	Provides file and object level access to archival data all the time
Short-term duration 1-120 days	Protect permanent and long-term data from change	Improves workflows with better access to archive data

Source: Horison, Inc.

Archiving is the process of moving data that is no longer actively used but is required to be securely retained at a different physical location for long-term storage, freeing up more costly space on the source location. Most applications today treat archive data as read-only to protect it from modification.

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 unchanging data from the backup set to speed up the backup (and restore) process and free up costly disk storage capacity in the process. Though disk backup processes such as deduplication can help, the growing length of backup windows remains a major issue and is under constant pressure as enterprise data growth rates reach 25 - 30% annually. An active archive implementation provides faster access to archival data by using HDDs (Hard Disk Drives) or SSDs (Solid State Disk Drives) as a cache front-end for a robotic tape library. The larger the archive becomes, the more benefit an active archive provides.

How Much Data is Archival?

More data is created per hour now than in an entire year just two decades ago. The sum of data generated by 2025 is estimated to reach to 175 zettabytes though approximately 7.5 ZB will actually be stored due to the short-lived, transient lifespan of most data generated. Between 60 and 80% of all data is archival and much of it is stored in the wrong place, on HDDs and totals 4.5 – 6 ZB of stored archival data by 2025 making archive the largest classification category. Correctly allocating archival data to highly cost-effective tape storage systems presents a significant financial savings opportunity.

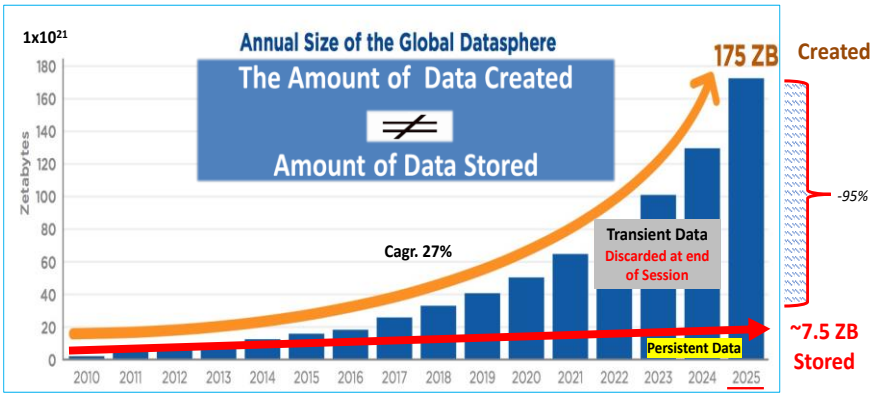
Correctly allocating archival data to highly cost-effective tape storage systems presents a significant financial savings opportunity.

The Modern Archive is Mostly Unstructured Data

Modern archive data is primarily unstructured (not as easily searchable) and includes office documents, video, audio, images, historical records, and basically anything not in a database. Object storage is becoming the preferred archive format as it is better for storing unstructured data, storing large data sets, and storing data with custom data

Global Datasphere Expansion is Never-ending

Nearly 80% of Data is Unstructured

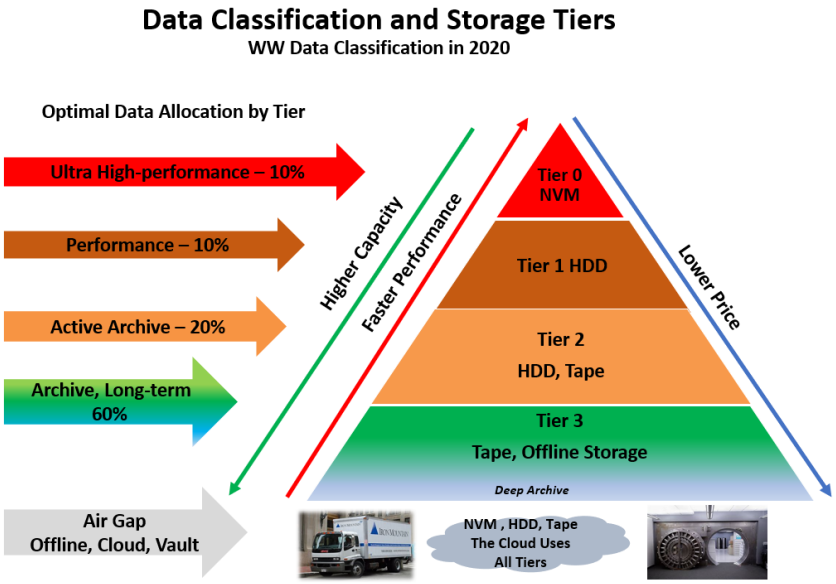


Source: https://www.storagenewsletter.com/2018/11/28/global-datasphere-from-33zb-in-2018-to-175zb-by-2025/

preservation, deletion, and retention policies. Big data is mostly unstructured archive 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, tags and catalogs to make search and access easier and much faster for archive data.

Data Classification Guidelines

The data classification process is critical to effectively manage data and becomes more critical as the storage pool grows. 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 which is the largest and fastest growing storage segment. Data classification enables the alignment and mapping of data characteristics with the optimal storage technology tier. Moving as much data as possible to the lowest cost storage tier is the core key ingredient for modern archiving strategies and returns the greatest economic value.

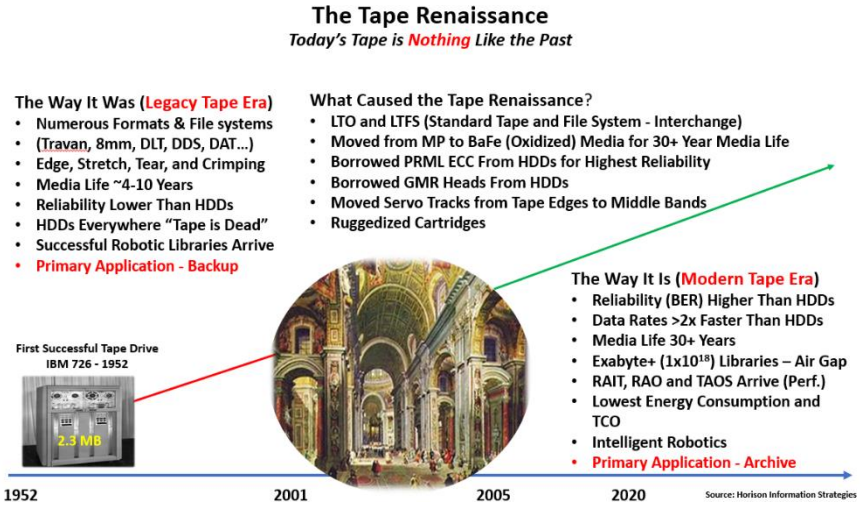


Moving as much data as possible to the lowest cost storage tier is the core key ingredient for modern archiving strategies and returns the greatest economic value.

Looking ahead, the ideal archive solution will classify data and create metadata upon ingest. The top four challenges to unlocking the value of archival data are: 1) making archival data accessible at ingest (classification, index, catalogs, metadata), 2) managing the long-term archival storage infrastructure, 3) ensuring that only the needed archive data is actually stored, 4) ensuring the security and availability of archival data.

The Tape Renaissance Transitioned to the Modern Tape Era

Since the first tape drives appeared in the early 1950s, tape has primarily served as a backup device for disk data. Troublesome tape issues of the past including edge damage, stretch, tear, loading problems, and media alignment with older (now obsolete) tape formats such as DAT, DDS, DLT, and 8MM tape were successfully addressed. By 2000, the Legacy Tape Era was ending and the initial use case for tape (backup) began to give way to archiving as *the tape renaissance* was underway as the tape industry was building a new foundation to address many new and emerging data intensive applications.



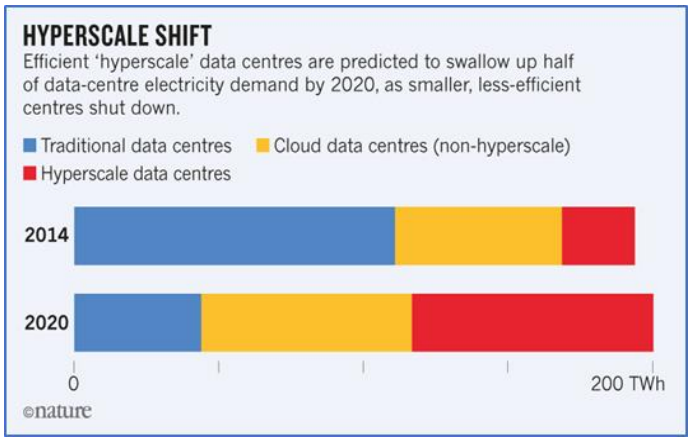
With the internet, hyperscale data centers, cloud, big data, tele-health, compliance, analytics and IoT waves all promising unprecedented data growth, the timing for advanced tape functionality couldn't be better. The era of modern tape era has arrived delivering the following capabilities:

- Tape is cheaper ($/TB) to acquire than disk
- Tape is less costly to own and operate (lower TCO) than disk by 5 - 8x
- Tape has surpassed HDD reliability by three orders of magnitude
- The media life for modern tape is 30 years or more for all new media
- Faster tape drive performance (throughput) with RAIT and higher availability with RAIL, and faster file access times with RAO and TAOS
- Intelligent, faster, and more efficient robotic movement for tape libraries
- Tape data compression ratios at 2:5 - 1 or higher
- Security capabilities with encryption, WORM, and air gapped storage
- LTFS a standard open tape file system with media partitions for faster "disk-like" access
- The 10-year LTO roadmap for tape technology is well defined with few foreseeable limits

Looking at the decade ahead, tape momentum should increase as data growth continues on an explosive path across many new applications and workloads, and in most of the largest Hyperscale Data Centers. Tape will not replace HDDs or SSDs, but it will be the de-facto standard archive architecture for the foreseeable future. A commonly stated objective for many data center managers today is that *"if data isn't used, it shouldn't consume energy*" clearly places tape as the greenest storage solution available. The arrival of many rich tape technology improvements has set the stage for tape to continue to be the most cost-effective storage solution for the enormous high capacity and archival challenges that lie ahead. Modern tape is nothing like tape of the past.

Cloud and Hyperscale Data Centers (HSDCs) Implement Tape Archive Solutions

There are projected to be as many as 570 HSDCs worldwide in 2020 representing the fastest growing type of data center. Data centers and information technology consume about 2% of the world's electricity currently and is expected to soar up to 8% by 2030. HSDCs may be the epicenter for modern archiving strategies. HSDCs face insurmountable growth of disk farms which are devouring operating budgets, energy sources and overcrowding data centers *forcing* archival and lower activity data migration to tape solutions. For the hyperscale world "adding disk is tactical – adding tape is strategic." Tape cartridges spend most of their life in a library slot or on a shelf and consume no energy when not mounted in a tape drive making tape the ideal archival storage choice. Advanced, easily scalable air-gapped tape architectures that can support erasure coding, geo-spreading with exascale capacities while providing the lowest TCO, highest reliability, and improved cybersecurity protection will play an increasing role to address the enormous HSDC storage challenges that lie ahead.



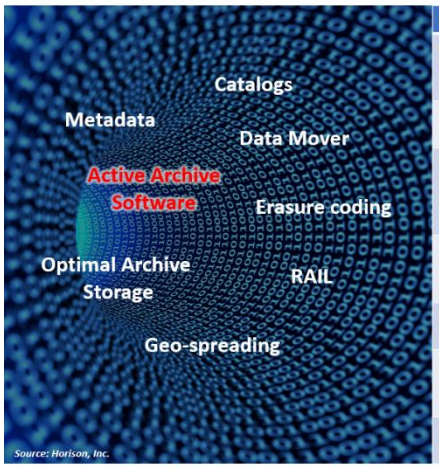
> For the hyperscale world "adding disk is tactical – adding tape is strategic."

Building the Archive of the Future

The successful archival architecture of the future will have the capability to store, protect, preserve, retrieve and easily scale into exabyte level capacities spread across multiple locations. For most enterprises and service providers, a single location may not be sufficient to deliver a high availability data protection strategy since the entire data center might go offline or its access be blocked by a natural or man-made disaster, cyber-attack, EMP or other catastrophic event.

Key components of the archive of the future will include active archive software with smart data movers, data classification and metadata capabilities, a highly scalable tape library technology, RAIL (Redundant Arrays of Intelligent Libraries) architectures, erasure coding and geo-spreading data across zones in different locations for higher fault-tolerance, redundancy and availability.

The Anatomy of the Modern Archive



Component	Function
Active Archive Software	Object storage software that scales out and geo-spreads unstructured and object data to manage and protect archival storage needs
Data Mover	Retrieves data from a storage device and makes it available to a network client
Classification - catalogs, metadata creation	Creates on ingest a detailed inventory of all data assets in an organization, to quickly place, protect and find the most appropriate data for any analytical or business purpose
Optimal archive storage	Modern tape for optimal TCO, reliability, media life, scalability, minimal re-mastering, lowest cost. Hi – capacity HDDs for active archive
RAIL	Stripes data across multiple tape libraries for redundancy and higher availability
Erasure coding	A parity-based protection scheme where data broken into shards, encoded then stored across multiple locations (nodes) for high availability
Geo - spreading	Distribution of data across multiple physical storage systems located in multiple locations

Conclusion

Businesses, governments, societies, and individuals worldwide increase their dependence on data making preservation and archiving has quickly become a critical IT practice. 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 storage cost savings while aligning data with storage tiers. Unless a new archival technology arrives, the numerous improvements in tape have made it the clear-cut optimal data archiving storage choice for the foreseeable future. The hardware, software and management components to implement a cost-effective, high availability archive infrastructure are now in place.

About the Sponsor

Quantum

Quantum focuses on creating innovative technology and solutions to help our customers get the most value from their data. With 40 years of storage know-how, Quantum's technology, solutions, and services help customers capture, create, and share digital content – and preserve and protect it for decades. Whether it's unlocking the potential of digital content, powering breakthrough innovations, creating entertainment that enriches lives, or keeping nations secure, Quantum works with customers and partners to make the world a happier, safer, and smarter place.

About the Author

HORISON
Information Strategies

Horison Information Strategies is a data storage industry analyst and consulting firm specializing in executive briefings, industry seminars, 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.

WP00260A-v01 Aug 2020