Active Archive and the State of the Industry

Active Archives Offer Modern Strategies to Solve Relentless Data Growth Challenges
July 2019

ABSTRACT

Archival data is piling up faster than ever as organizations are quickly learning the value of analyzing vast amounts of previously untapped digital data. Industry studies consistently find that the vast majority of all digital data is rarely, if ever, accessed again after it is stored. However, this is changing now with the emergence of big data analytics made possible by **Machine Learning** (ML) and **Artificial Intelligence** (AI) tools that bring data back to life and tap its enormous value for improved efficiency and competitive advantage. The need to securely store, search for, retrieve and analyze massive volumes of archival content is fueling new and more effective advancements in archive solutions. These trends are further compounded as an increasing number of businesses are approaching hyperscale levels with significant archival capacity requirements. This report describes the state of the archive market and the expanding role the active archive plays.

THE DIGITAL ARCHIVE MARKET FACES RELENTLESS DATA GROWTH

IDC projects the Global Datasphere will grow at 27% annually from 33 ZB in 2018 to 175 ZB by 2025, while a smaller sub-set of this data, but still a massive amount of data (~7.5 ZB in 2025, up from ~1.0 ZB in 2018) will actually need to be stored for some period of time on SSD, HDD, and tape. Data retention requirements of 100 years or more are now common and necessitate a more intelligent and secure archival storage infrastructure. The top external factors driving archival and long-term retention requirements include government compliance regulations, growing dependency on security and surveillance systems, advanced 3D and 4D video capabilities, scientific capture, rich content produced in the media and entertainment (M&E) sector, the relentless growth of big data analytics, and the rapid emergence of the Internet of Things (IoT), which is generating huge volumes of data. The impact of the Global Datasphere appears unstoppable!

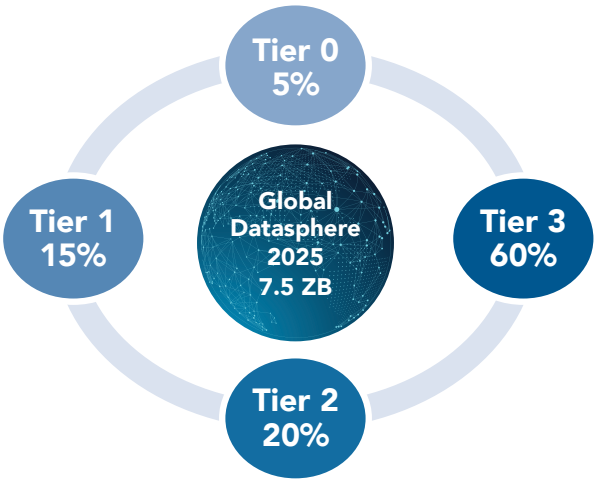
A key emerging trend for new data creation indicates that even though transactional (high IOPs) data and high-performance applications are demanding higher performance, the amount of reference and archival data is growing even faster. This accumulation is primarily due to the sheer amount of data "yet to be analyzed" sometime in the future in anticipation that potential value might be discovered with new ML and AI tools. Most data typically reach archival status in 90 days or less but does not get classified as archival data. The 2019 Active Archive Alliance survey of 122 respondents indicates that 46% are not prepared to either implement an archive strategy or modify their existing strategy within the next year. Given the growth in archival data and use cases that depend on archival data, the survey response reveals a huge, unaddressed and growing archive data challenge for end users and an opportunity for the storage industry to address at the same time.

The magnitude of data volume will require data centers to optimize SSD, HDD, and tape with intelligent tiered storage solutions. The greatest economic advantages of tiered storage occur when tape is included in the stack as the price and TCO differentiation between multiple disk tiers is far less significant. Classifying data as it's created by its value, performance, and capacity requirements enables the right data to be in the right place at the right time. Using industry averages for data classification, it is projected that the Global Datasphere, if optimally stored by data classification, would be Tier 0 SSD (5%), Tier 1 and Tier 2 HDDs (35%) and Tier 3 tape or an active archive (60%). Unfortunately, data allocation is seldom optimized. Depending on the accuracy of these projections, the data suggests 4.5 ZB of Tier 3 archive capacity worldwide would be needed in 2025 if optimally stored.

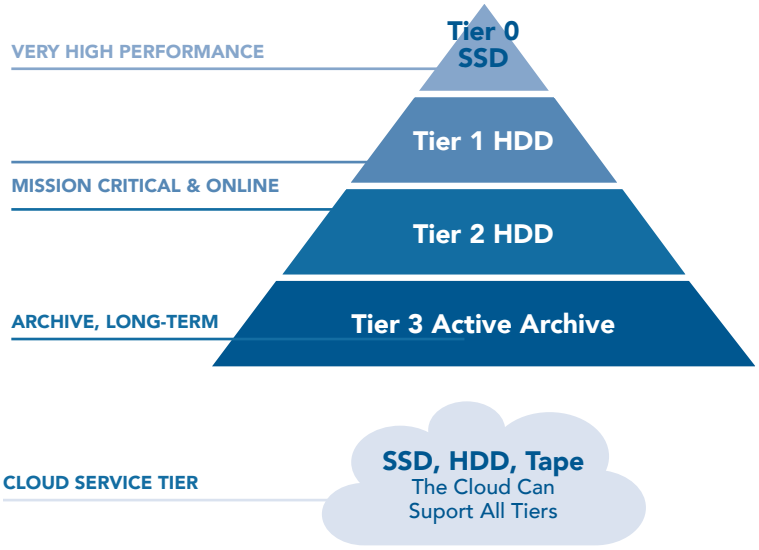
See following chart.

GLOBAL DATASPHERE
By Class and Tier in 2025

Percent of WW Digital Data Stored if Optimally Classified by Tier in 2025



Data Classification



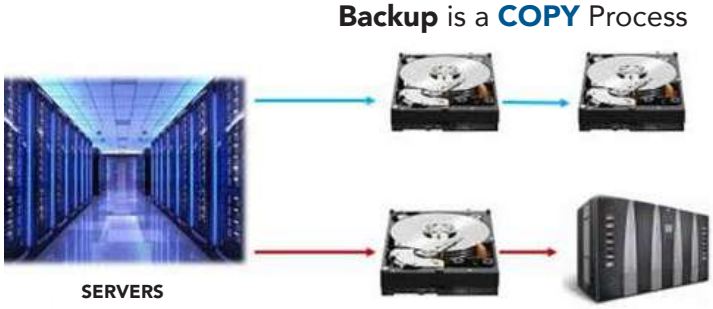
Source: Horison, Inc. The Digital Universe IDC

BACKUP DOES NOT EQUAL ARCHIVE

It's becoming critical to distinguish between backup and archive as these core IT processes are not the same and are too often misunderstood. The 2019 Active Archive Alliance survey indicates 57% of respondents still use their backup system for storing archive data. Using backup copies to store archival data and repeatedly backing up unchanging archives lengthens the backup window and wastes costly HDD space. Backup and archive are entirely different processes and have different objectives.

The backup process creates copy(s) of data for recovery purposes which may be used to restore the original copy after a data loss or data corruption event. Backups are cycled and updated frequently to account for and protect the latest versions of important data assets. Archiving frees up disk capacity and moves unchanging and less frequently used data to a new location(s) and refers to data specifically selected for long-term retention. Archival data is typically unchanging and is rarely overwritten. The failure to distinguish between backup and archive can become very costly resulting in poor resource utilization.

- Archiving heavily favors tape for all archival data
- Modern tape has a media life of 30 years or more
- Tape has the lowest energy consumption and TCO



Backup is a **COPY** Process

Archive is a **MOVE** Process

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.

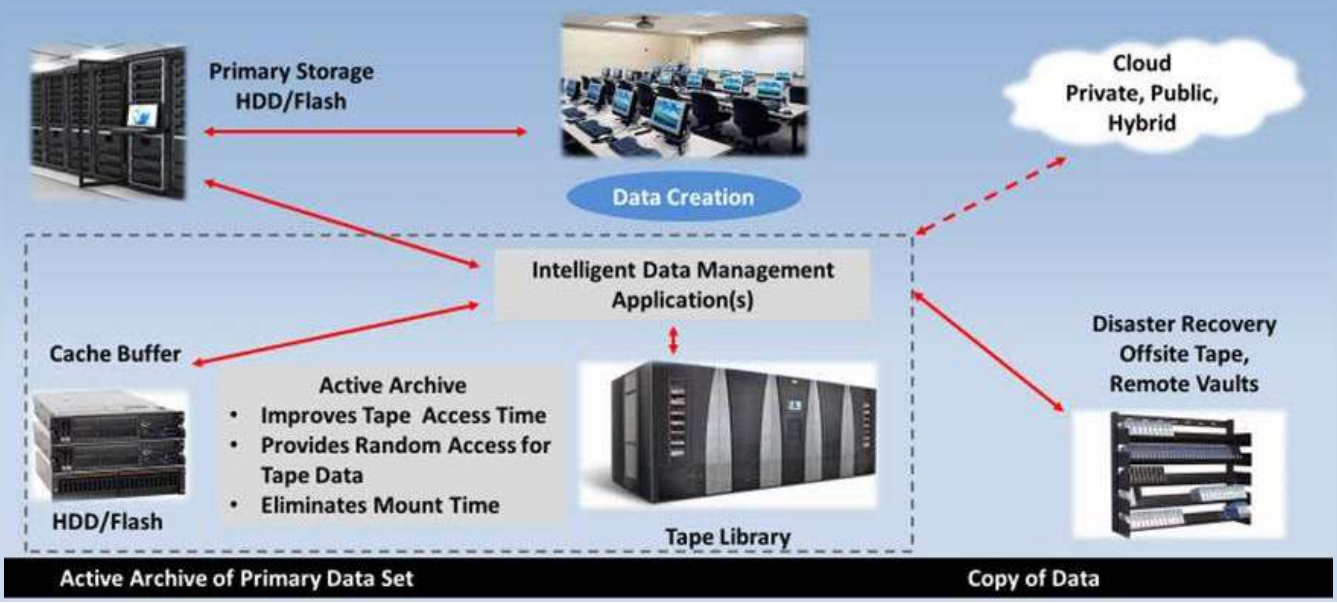
THE ACTIVE ARCHIVE BRINGS NEW LIFE TO ARCHIVES

An active archive is a high-value storage architecture that gives the user access to data across a virtual file system and migrates data between one or more storage systems and media types based on user defined policies. Most active archive implementations use high capacity HDDs and tape or the cloud. The active archive supports file, block or object storage systems using advanced data management software to maintain accessibility to archival data.

Intelligent data management software has made great strides and provides faster online random access, search and retrieval capability for archival data in a single virtualized storage pool, and automatically migrates (moves) data between storage tiers. SSDs or high capacity HDDs integrated with tape's highly favorable economics, security capabilities, reduced energy consumption, and archive characteristics define the growing benefits of active archives. As the archive gets larger, the value of the active archive increases. The diagram below illustrates the key components of an active archive implementation.

THE ACTIVE ARCHIVE

Combines SSD, HDD, Tape and Cloud for Optimal Archival Solution



An active archive resolves complexity by leveraging the benefits of an intelligent data management layer, and an increasing number of effective software products that address this are now available. Access and management of data is getting more complex and is requiring modern strategies with intelligent data management techniques. These strategies and techniques are now being enhanced by AI to further improve and automate data management.

The 2019 Active Archive Alliance survey indicates that 40% of respondents have plans to deploy an active archive where data is available online and always accessible. Active archive implementations can use existing storage devices to build an integrated hardware and software solution and can incorporate enhanced file systems such as the open standard **LTFS** (Linear Tape File System) for Linux, Apple and Windows, or **TAR** (Tape Archive) for Unix systems. For those who do not want to build their own storage repository using existing equipment, several vendors offer preconfigured active archive appliances and file systems that work with most any tape library back-end.

ACTIVE ARCHIVE KEY BENEFITS

Ease of Use	Provides end users with file-level access to all their data, all the time without IT intervention.
Improved Primary Storage Performance	Primary storage contains vast amounts of infrequently accessed, archive or cold data. By reducing the amount of data to manage on primary storage via moving inactive data to an active archive, organizations should see substantial improvement in system performance.
Free Disk Space	Moving low activity data from disk to tape frees costly space for future growth. As much as 60% of data on disk is better suited for archives.
Improved Archival Storage Performance	Archive access time is significantly improved by using a cache (SSD, HDD) front-end for tape libraries offering faster access for most files as data can live for extended periods of time in cache before it is migrated from the cache to tape.
Reduced Backup Window	Even with backup using compression and deduplication, backup windows face constant pressure to take less time. Archiving streamlines the backup process by removing terabytes of data from the backup set on primary storage while greatly reducing the duration of the backup window. There is no benefit in repeatedly backing up unchanged archival data.
Cybercrime Protection	The "tape air gap" prevents cyber-attacks since data stored offline - without an electronic access - cannot be hacked.
Lower Cost	Reduces TCO by moving archival data from HDDs to lower cost tape.
Less Media Conversion	By archiving to tape, the time consuming need to migrate enormous amounts of archive data to a new media format occurs less frequently than disk. The typical HDD will last 4-5 years before replacement while the average tape drive lasts 7-10 years or more before replacement.
Scalability	An active archive can easily add storage capacity and cache to scale to petabytes and exabytes. Remember tape adds capacity by adding media while HDDs add capacity by adding drives.
Cloud Usage	An active archive can combine onsite, offsite, and cloud environments. Most, if not all, cloud providers are offering archival data services including active archiving. Active archive brings the same benefits to public clouds as it does to on-premises solutions.

AN ACTIVE ARCHIVE ECOSYSTEM INTEGRATES STORAGE, SOFTWARE AND THE CLOUD

An active archive can be implemented on-premises, in the cloud, or in both places as offering archival data services is gaining significant momentum for cloud providers. Active archiving brings the same benefits to the cloud (private or public) that it does to the data center. The active archive provides a persistent online view of archival data by integrating SSD, HDD, tape, and the cloud with intelligent software (the active archive ecosystem) behind a standard file system such as LTFS. The active archive allows a system administrator to define policies for automatic data migration and retention to control the movement of petabytes of data from more expensive to less expensive storage systems.

CASE STUDIES AND VIDEOS HIGHLIGHT THE VALUE OF ACTIVE ARCHIVE SOLUTIONS

CONCLUSION AND OUTLOOK

Enterprises archive their data because they either want to or because they have to, but either way, the magnitude of this requirement is quickly becoming overwhelming. With the amount of archival data exploding and no end in sight, active archives will play a vital and increasing role in optimizing storage to reduce costs and will ensure archived data is accessible and protected. Future SSD, HDD, and tape technology roadmaps promise continued innovation, reliability, and capacity increases with lower costs, which make the active archive concept even more appealing.

New techniques, applications and use cases such as AI, ML, big data, IoT, and video surveillance will drive increased demand for active archives by organizations that need to effectively manage data from terabytes to exabytes across multiple storage pools throughout its life cycle. Hardware and software technologies have made significant advancements that improve data and storage management capabilities for all types of data including archives. Why not take advantage of these capabilities and optimize your storage infrastructure before it's too late?

ABOUT THE ACTIVE ARCHIVE ALLIANCE

The **Active Archive Alliance** is a vendor-neutral, trusted source providing end users with technical expertise and guidance to design and implement active archive solutions for intelligent data management. The goal of the Alliance is to encourage a multi-vendor effort to promote and align the awareness and technologies needed to meet the rapidly increasing requirements for archival data in the following ways:

1. Active archives enable efficient access to data throughout its life
2. Active archives are compatible with flash, disk, tape, or cloud (public or private) block or object storage systems
3. Active archives help move data to the appropriate storage tiers to minimize cost while maintaining ease of user accessibility

The Active Archive Alliance includes representatives of:

SPECTRA
FUJIFILM
STRONGBOX DATA SOLUTIONS
IRON MOUNTAIN®
Qumulo
komprise
Quantum®

More information on active archives and the Active Archive Alliance can be found at **www.ActiveArchive.com** and on **Twitter**, **LinkedIn**, and **Facebook**.