Active Archive and the State of the Industry 2018

The Active Archive Becomes Integral to Your Game Plan

Abstract
Organizations are quickly learning the value of analyzing vast amounts of previously untapped archival data. Industry studies suggest that only about 20% of all digital data is ever accessed or used again after it is stored, underscoring the archival challenge. The need to effectively store, search for and retrieve enormous volumes of archival content is fueling new advancements in archive solutions. This report describes the state of the archive market and the role the active archive plays.

Relentless Data Growth Fuels Digital Archive Market
Newly created worldwide digital data is expected to grow at 30% or more annually reaching 163 zettabytes ($1x10^{21}$ bytes) by 2025, according to IDC. Data is rapidly piling up in archives as retention requirements of up to 100 years to forever are now common. 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, content producers in Media & Entertainment, the relentless growth of Big Data analytics, and the emerging IoT (Internet of Things).

A key emerging trend for new data creation indicates that even though transactional (high IOPs) data and high-performance applications are steadily growing, the amount of reference and archival data is growing faster. This is primarily due to countless regulations and the sheer amount of data "yet to be analyzed" in the future in the anticipation that some potential value might be discovered. Most data typically reach archival status in 90 days or less. IDC estimates that by the end of 2025, only 15% of the data in the global data-sphere of 163 zettabytes will be tagged and only 20% of that will be analyzed. Therefore, if 80% of the data created is never analyzed, that data will likely reach archival status upon creation. Recent analyst surveys indicate that less than 40% of corporations have a dedicated archive strategy in place. This reveals a huge, unaddressed and growing archive challenge that is ready to embrace modern archiving concepts.

These enormous data volumes will require that businesses build their storage architectures by optimizing SSD, HDD and tape in tiered storage solutions. The greatest economic advantages of tiered storage occur when tape is included. Classifying your data upon its creation by its value, performance and capacity requirements will enable the right data to be in the right place at the right time. By 2025, using industry averages for data classification, it is projected that 5% of digital data will be stored on Tier 0 SSD, 35% on Tier 1 and Tier 2 HDDs and 60% on Tier 3 tape. Tier 3 is typically referred to as the archive tier with an average of 60% of data classified as archival upon creation. See following chart.

Digital Universe by Storage Class and Tier

Source: Horison, Inc. The Digital Universe IDC
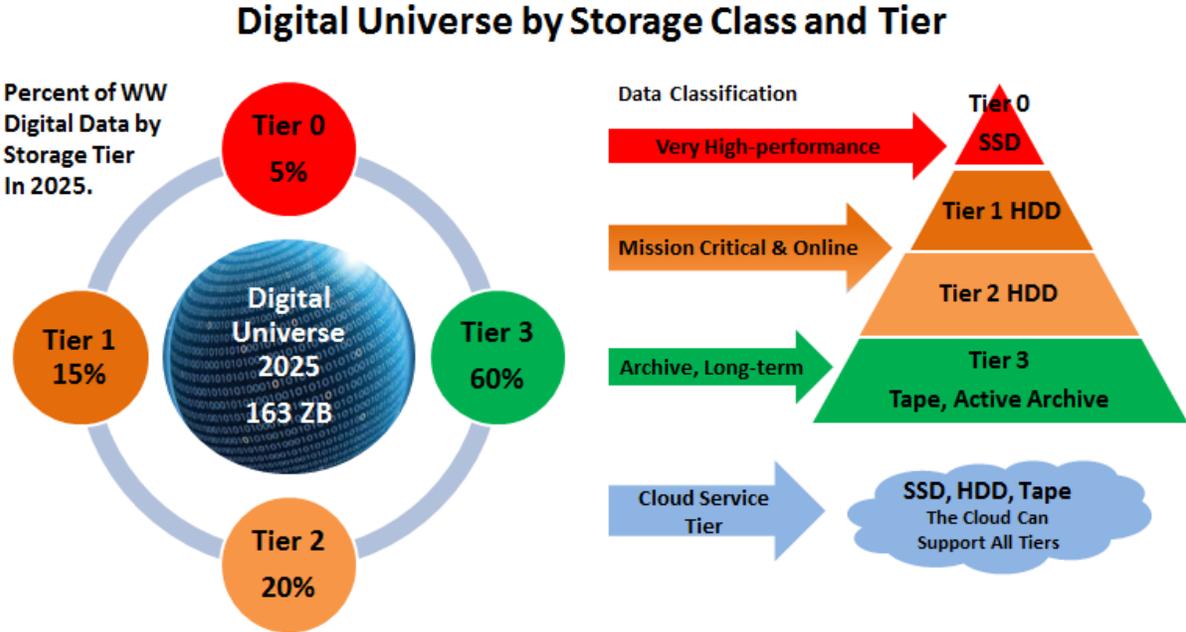
Backup and Archive Are Entirely Different Processes

It's important to distinguish between backup and archive as these core IT processes are not the same and are often misunderstood. Many businesses still use backup copies to store archival data and repeatedly back up unchanging archives wasting HDD space. Backup and archive are entirely different processes and have different objectives.

The backup process creates copies 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 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 not overwritten.

Backup (*Copy data*)	Archive (*Move data*)	Active Archive (*Fast access to data archives*)
Creates copy(s) of data for recovery purposes and data protection.	Moves unchanging data specifically selected for long-term retention to a more cost-effective location(s).	A combined solution of open systems software, SSD, HDD, and tape storage systems.
Used to restore the original copy after a data loss or data corruption event.	Archiving frees up storage by removing data from backup set to achieve greater efficiency and cost savings.	Provides faster online random access for data archives by using HDD or SSD to serve as a cache front-end for high capacity tape libraries.
Backups are cycled and updated frequently.	Archival data is typically unchanging and is not overwritten.	Provides end users with file-level access to archival data all the time without IT intervention.

The Active Archive Re-awakens the Archives

An active archive solution integrates SSD, HDD, tape, and the (public or private) cloud, leveraging the popular tiered storage model to be focused on the archive function. The active archive supports file, block or object storage systems using advanced data management software to maintain end user accessibility to archival data regardless of the storage device it is residing on. Intelligent data management software provides faster online random access, search and retrieval capability for archival data in a single virtualized storage pool, and automatically migrates data between storage tiers based on user policy. The widespread usage of SSDs and high capacity HDDs coupled with tape's highly favorable economics, security, and archive characteristics have propelled the successful emergence of active archives.

The diagram below illustrates the key components of an active archive implementation.

The Active Archive
Taking Archival Storage to the Next level
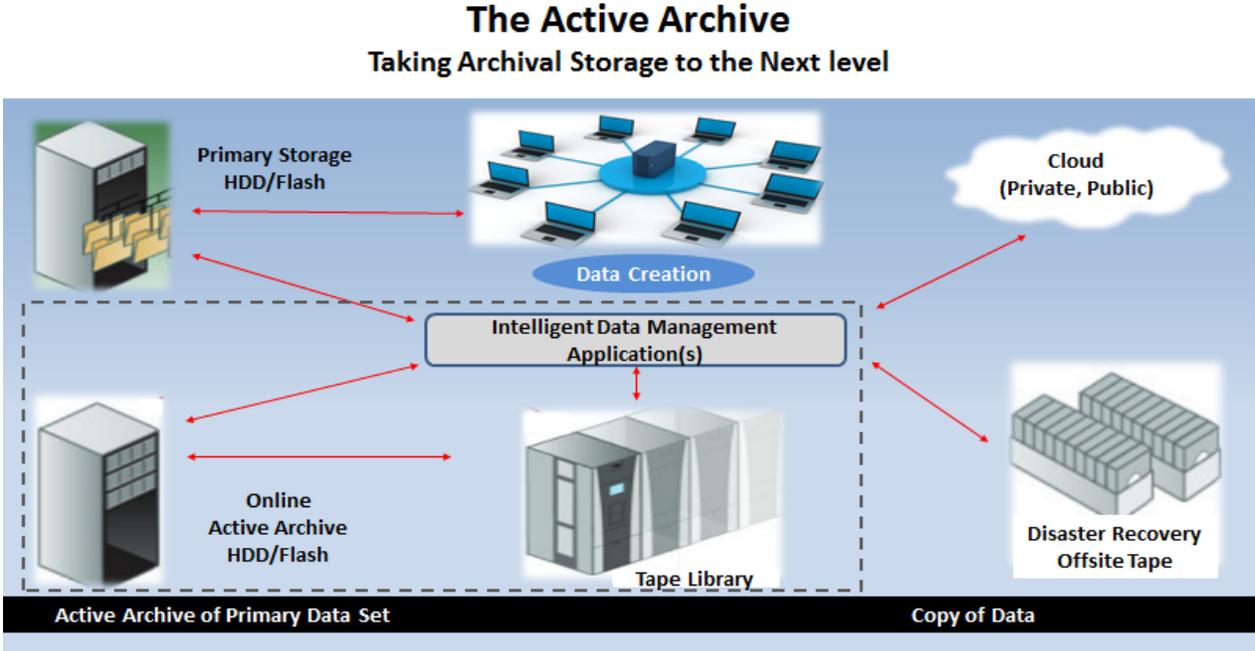
Active archiving implementations can use existing storage devices to build an integrated hardware and software solution and can incorporate enhanced file systems such as the open standard LTFS (Linear Tape File System) for Linux, Apple and Windows, or TAR (Tape Archive) for Unix systems. For those who do not want to build their own repository using existing equipment, several vendors offer preconfigured active archive appliances and file systems which work with most any tape library back-end.

Digital Archives Embrace Object Storage
Archiving is the earliest enterprise use case for object storage, having been used for over a decade providing scalable, long-term data preservation. Object storage 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 connection with OpenStack Swift to enable 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 much easier to perform bulk operations using tape within a Swift data ring. Cloud storage is the most prominent use case for object storage.

Active Archive	Benefits Summary
Ease of Use	Provides end users with file-level access to all their data, all the time without IT intervention.
Less Media Conversion	By archiving to tape, the 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 - HDDs add capacity by adding drives.
Improved Primary Storage Performance	Primary storage infrastructures contain a large amount of infrequently accessed, archive or cool/cold data. By reducing the total amount of data to manage on primary storage by partitioning inactive data from active data, organizations should see substantial improvement in system performance.
Improved Archival Storage Performance	Archive access time is significantly improved by using a cache front-end for tape libraries offering faster access for most files because data can live for extended periods of time in cache before it is moved from the cache to tape. Tape continues to be the most reliable, cost-effective archive solution available.
Reduced Backup Window	Even with backup using data compression and data deduplication, backup windows face constant pressure from data growth rates that typically exceed 40% annually. 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.
Lower Cost	Moving archival data to lower cost storage such as high capacity tape can provide a significant TCO reduction.
Energy Savings	Moving less active archival data from HDDs to tape will significantly reduce energy consumption. Energy costs for tape are typically < 5% of the equivalent amount of disk capacity.

An Active Archive Ecosystem Can Include Storage, Software and the Cloud

By providing a persistent online view of archival data by integrating one or more archive technologies (SSD, HDD, tape, and software – the active archive ecosystem) behind a file system, the active archive is a form of the widely used tiered storage concept specifically targeted for the data archive function. The tiered storage concept 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. An active archive can be implemented on-premises, in the cloud, or in both places as offering archival data services quickly gains momentum for cloud providers. Active archiving brings the same benefits to the cloud (private or public) that it does to the data center.

Select Case Studies Highlight the Value of Active Archive Solutions

ACRI-ST Advances the Understanding of Earth with Satellite Data

Civil Engineering Firm Archives Legacy File Share Data

EPFL Selects Active Archive System for 50 Years of Music Archives

Oak Ridge National Laboratory Enhances Data Integrity and Accessibility with Active Archive

Pacific Biosciences' Journey to Intelligent Data Management

Conclusion and Outlook

Enterprises archive their data because they either <u>want</u> to or because they <u>have</u> to, but either way, the magnitude of this requirement can quickly become overwhelming. Future SSD, HDD, and tape technology roadmaps promise continued innovation, reliability, and capacity increases with lower costs providing archiving the necessary infrastructure. With the amount of archival data soaring and no end in sight, active archiving is poised to play an increasingly important role in tomorrow's data centers as it reawakens the archives. Are you prepared to manage the relentless growth of archival data that lies ahead? It's time to develop your game plan.

About the Active Archive Alliance

The Active Archive Alliance launched in 2010 as a collaborative industry association formed to educate end user organizations on the evolving new technologies that enable reliable, online and efficient access to their archived data. The goal of the Active Archive Alliance is to encourage a

multi-vendor effort to align the education and technologies needed to meet the rapidly increasing requirements for archival data by addressing the following:

1. Reduce the complexity of long-term data storage
2. Provide scalable storage solutions
3. Reduce total cost of ownership for long-term data storage
4. Reduce risk of non-compliance and data loss

Note: *The Active Archive Alliance includes representatives of FUJIFILM, HGST, Komprise, Quantum, Spectra Logic, StorageDNA, and StrongBox Data Solutions.*

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