Active Archive and the State of the Industry

Taking Data Archiving to the Next Level

Abstract
This report describes the state of the active archive market.

New Applications Fuel Digital Archive Market

IDC projects that by the year 2025 the digital universe will reach 163 zettabytes ($1x10^{21}$) bytes representing a tenfold increase from the amount of data created in 2016. Where is all this data going to be stored long-term? Much of it will be archival storage, which while infrequently accessed must be maintained for the possibility of future mining, analytics or historical research purposes. Businesses need to store this data in the most efficient, reliable and economical manner possible, while maintaining access and protection of the content. In most cases, tape will provide excellent archival capabilities. In other cases, a low-cost object storage solution can provide an excellent add-on to archive workflows with faster and more random access requirements.

The top external factors driving archival and long-term retention requirements include government compliance regulations, security and surveillance systems, the relentless growth of Big Data analytics and the emerging IoT (Internet of Things). The IoT is the network of physical objects or "things" embedded with electronics, software, sensors and network connectivity that enables these objects to collect, exchange, and store unknown amounts of data, much of which will quickly become archival.

Big Data generates vast volumes of data types that require a long-term, cost-effective means of preservation to produce insights that can vastly improve decision-making capabilities from individuals to large enterprises. These enormous data volumes are often collections of new types of data that have emerged with the advent of social media platforms, high definition video, surveillance, the advanced 3D and 4D media capabilities, and the yet unknown impact of the IoT. Add to this the complexity of regulatory compliance and laws associated with long-term data retention and you can see how the sheer size of archival data is becoming a costly strain on storage resources in organizations across the globe.

Backup and Archive Are Different Processes

It's important to distinguish between backup and archive as these core processes are often not well understood. Backup and archive are entirely different processes and have different objectives.

| Function	Operation	Description
Backup *For data protection*	Copy data	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.
Archive *For long term retention*	Move data	Archiving moves unchanging data to a new location(s) and refers to data specifically selected for long-term retention. Archival data is typically secure and is not overwritten. Archiving is the ideal solution for regulatory compliance or data with historic value and it is easier to search.
Active Archive *Faster access for long term retention*	Move data	An active archive provides easy access and a persistent view of all your archival data, all the time. An active archive is a combined solution of open systems software, flash, disk, and tape hardware that gives users an effortless means to store and manage all their data. An active archive can serve as an "express lane" into large amounts of archival data.

Archival Data Characteristics

Other than system files, catalogs, indices, transactional databases and directories, the probability of reuse, or future access, for most data types declines as the data ages and this data typically reaches archival status in 60 days or less. Archival data is accumulating faster than ever as many data types *must* be securely stored and protected for indefinite time frames.

Active Archive Implementations

Active archive data is typically fixed content and unstructured data that still must be available online, but has lower frequency of access and can tolerate longer response time latencies. Exabyte (1×10^{18}) sized archives now loom on the horizon and much of this data will be required to "come alive" as analytics are set to proliferate. Leveraging innovative and integrated solutions combining flash, disk, tape, and the cloud will be required to fully enable the active archive experience. The widespread usage of flash, lower cost capacity optimized HDDs coupled with tape's highly favorable economics and archive characteristics have propelled the emergence of active archives.

The active archive is a method of tiered storage which gives the user unified online access to data across multiple storage systems and media types including solid-state drive/flash, hard disk drives, magnetic tape, and cloud. This software provides faster online random access, search and retrieval capability for archival data in a single virtualized storage pool and migrates data between storage tiers.

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

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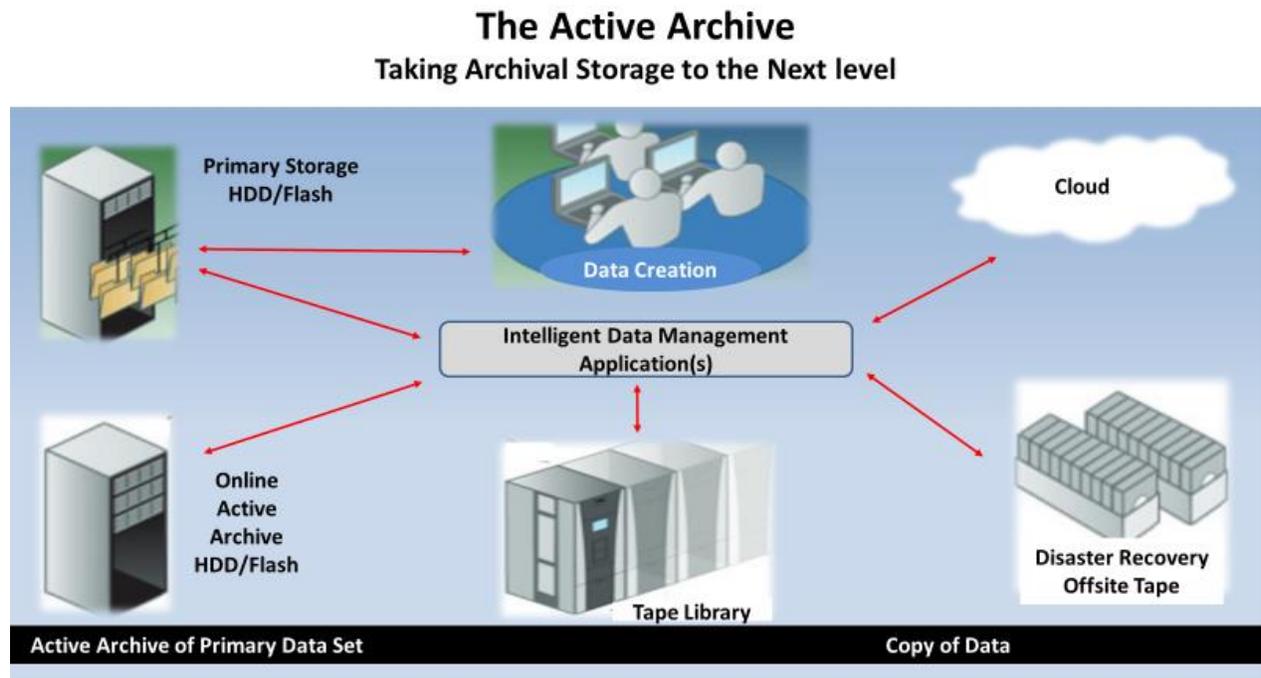
Active archiving implementations can use existing storage devices to build an integrated hardware and software solution and can incorporate enhanced file systems such as LTFS (Linear Tape File System) or TAR (Tape Archive). For those who do not want to build their own repository using existing equipment,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

The increased popularity of cloud storage for offsite backup and large-scale archiving is further driving active archive demand. Today, many organizations are benefiting from a hybrid cloud storage model in which some data is stored in the cloud and some is stored using on premise disk and tape solutions.

Archives are Evolving
Archives are undertaking a radical transformation, driven by much larger file sizes and new access requirements. Active archive solutions support file, block and now object storage systems, which makes them extremely versatile for any data type. Object storage enables IT managers to organize archive content into containers and to easily allow retention of massive amounts of unstructured archival data. Object storage software or appliances can transform an archive into an active archive that is positioned between high-performance storage and either tape, high capacity HDD or cloud. A flexible active archive fabric can scale performance by adding HDD, flash or cloud, and unlimited capacity by adding tape media to accommodate these diverse requirements within one common environment.

Benefits of an Active Archive

- **Ease of Use.** Provide end users with file-level access to all their data, all the time – without IT intervention. Whether an organization stores active archive data in a privately managed object storage cloud, on tape or in an offsite public cloud, an active archive gives the IT department the flexibility to select the most appropriate storage media while not inhibiting user and application access to the data.

- **Energy Savings.** Archiving less active data from HDDs to tape can significantly reduce energy consumption. An active archive can expand the scope of an existing archive by adding a higher performance capability.

- **Improved Archival Storage Performance.** Archive access time is significantly improved by using HDDs to serve as a cache front-end for tape libraries offering faster access for most files because data can live for extended periods of time on HDD before it is moved from the HDD cache to tape.

- **Improved Primary Storage Performance.** Primary storage infrastructures contain a high proportion of infrequently accessed, or cool/cold data. By reducing the total amount of data to manage on primary storage, or partitioning inactive data from active data, organizations may see substantial improvement in system performance.

- **Knowledge Discovery.** With the advent of Big Data, organizations are learning the value of accessing and analyzing vast amounts of previously untouched archival data to gain a competitive edge in the marketplace. Big Data analytics is creating demand for active archive solutions that provide faster access times.

- **Less Media Conversion.** By archiving to tape, the need to migrate enormous amounts of archive data to a new media format is less frequent than disk. Modern tape drives can read the current version and usually the two prior format versions, which makes upgrades to a new tape technology and media less frequent. The typical HDD will last 4 to 5 years before replacement while the average tape drive lasts 8 years or more before replacement.

- **Lower Cost.** Reduce TCO by matching media type to SLAs by moving archival data to lower cost storage such as high capacity tape.

- **Reduced Backup Window.** Even with backup to HDDs using data compression and data deduplication, backup windows face constant pressure from data growth rates that can exceed 40% annually. There is no benefit in repeatedly backing up unchanged archival data. Archiving can remove terabytes of data from the backup set on primary storage greatly reducing the duration of the backup window as archiving streamlines the backup process.

- **Scalability.** An active archive can easily add capacity and scale to petabytes and exabytes and a billion or more files, while providing cache access to more frequently accessed archival data. Remember, when using tape, scaling capacity has virtually no limits.

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 (HDD, flash, tape, software *and* cloud storage – 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 archive function. The tiered storage concept allows a system

administrator to define policies for 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. The cloud is now embracing both disk and tape as storing archival data quickly gains momentum for cloud service providers. Active archiving brings the same benefits to the cloud that it does to the data center.

Select Case Studies That Highlight the Value of Active Archive Solutions
- Australia's NCI Implements Active Archive
- CrossFit Accelerates Media Workflows
- Imperial War Museums Digitally Preserve Modern War and Wartime History Indefinitely
- Minnesota Supercomputing Institute Manages High Volume Data Growth with Active Archive
- NERSC Exceeds Reliability Standards with Tape-Based Active Archive
- ORNL Enhances Data Integrity and Accessibility with Active Archive Solutions
- University of Oslo Implements Active Archive to Support Storage Needs of NorStore

Future Archival Solutions Will be Active
The active archive concept is reawakening the archives. An active archive implementation using an open file system to expand over disk, tape and cloud technologies allows organizations to build a continuous view of their retained data, making it easier and faster than ever to access archival files.

The bottom line is that the business case for active archives will become increasingly compelling and likely include cost containment (free up HDDs, lower energy usage), risk reduction to ensure regulatory compliance and improved productivity by getting inactive data out of the path of the backup window. They will also deliver more efficient searches and retrieval, and yield improved storage administrator efficiency. 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.

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
4. Reduce risk of non-compliance and data loss

 Note: *The Active Archive Alliance includes representatives of DDN Storage, FUJIFILM, HGST, MT-C, Quantum, Spectra Logic and StrongBox Data Solutions.*

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