Active Archive and the State of the Industry

Innovative Solutions Drive New Markets and Use Cases

Abstract
The Active Archive Alliance launched on April 27, 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 Active Archive Alliance goal is to align the education and technologies needed to meet the rapidly increasing requirements for archival data. Alliance members strive to extend solutions to the greater general IT audience that needs advanced online data archive options. This report describes the current state of the active archive market.

The State of the Digital Archiving Market

According to the latest Digital Universe report approximately 90% of the data in the world today was created within the past two years and the vast majority of it is not changed once created reaching archival status in a relatively short period of time. Overall, IT budgets are relatively flat, yet newly created digital data is growing at over 40% annually, and is now being generated by billions of people, not just by large data centers as in the past, mandating the emergence of an ever smarter and a more cost-effective and secure long-term storage infrastructure. Top external factors driving long-term retention requirements include compliance regulations, business risk, security exposure and the rapid emergence of Big Data analytics. For most organizations, facing hundreds of terabytes or several petabytes of archive data for the first time can trigger a need to redesign their entire archival storage infrastructure.

Exabyte ($1x10^{18}$) sized archives are now on the horizon. In the era of Big Data analytics, stored long-term data needs to be accessible for large-scale data queries long into the future. Redefining the term “archive” to become an online, accessible, affordable data management platform is necessary to solve data growth, improve performance, and to meet retention challenges going forward. Leveraging innovative and integrated solutions for disk, tape, and the cloud will be required to fully enable the active archive experience. An active archive can meet the challenges that lie ahead for data archiving.

Backup and Archive Are Different Processes

It is important to understand the differences between backup and archive as they are entirely different processes and have different objectives. Surprisingly, the differences between these processes are not well understood. Backup is best utilized during an emergency, when data

needs to be quickly restored, and copies of data are kept to mitigate short-term risks. Archives are most efficient for data management, compliance and information re-use purposes.

Backup (A Copy process for protection) The backup process creates a 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. Backup data is typically overwritten and is difficult to search. Backup data has little if any historic relevance.

Archive (A Move process for longer-term retention) The appropriate process of archiving moves infrequently used data to a new location(s) and refers to data specifically selected for long-term retention. Ideally more than one copy of archive data should exist since having a single copy of any meaningful data presents an exposure should the only copy become inaccessible. Archival data is typically secure and is not overwritten. Archiving is an ideal solution for regulatory compliance or data with historic value and it is easier to search.

Archive Characteristics Other than system files, catalogs, indices, OLTP and directories, the probability of reuse, or future access, of most data types declines as the data ages and typically reaches archival status in 60 days or less. Archival data is accumulating faster than ever, as many data types *must* be securely kept indefinitely. A customer survey from the 2013 SNIA Storage Developer Conference indicated a retention period of 20 years or more is required by 70% of respondents and 50 years or more retention is required by 57% of respondents making data preservation strategies critical.

Active Archive Implementations

The lower cost of capacity optimized HDDs coupled with tape's highly favorable economics have fueled the emergence of active archives. An active archive provides a persistent online view of archival data by integrating one or more storage technologies (flash, disk, tape *and* cloud storage) behind a file system that gives users a seamless means to store and manage their archive data in a single virtualized storage pool. Data archives are indexed and have search capabilities so files and parts of files can be easily located and retrieved. An active archive combines disk, serving as a cache for higher performance, with tape providing online access and easy retrieval of vast amounts of archival data. The increasing value of archival data demands faster access to archive storage systems with improved management and security solutions.

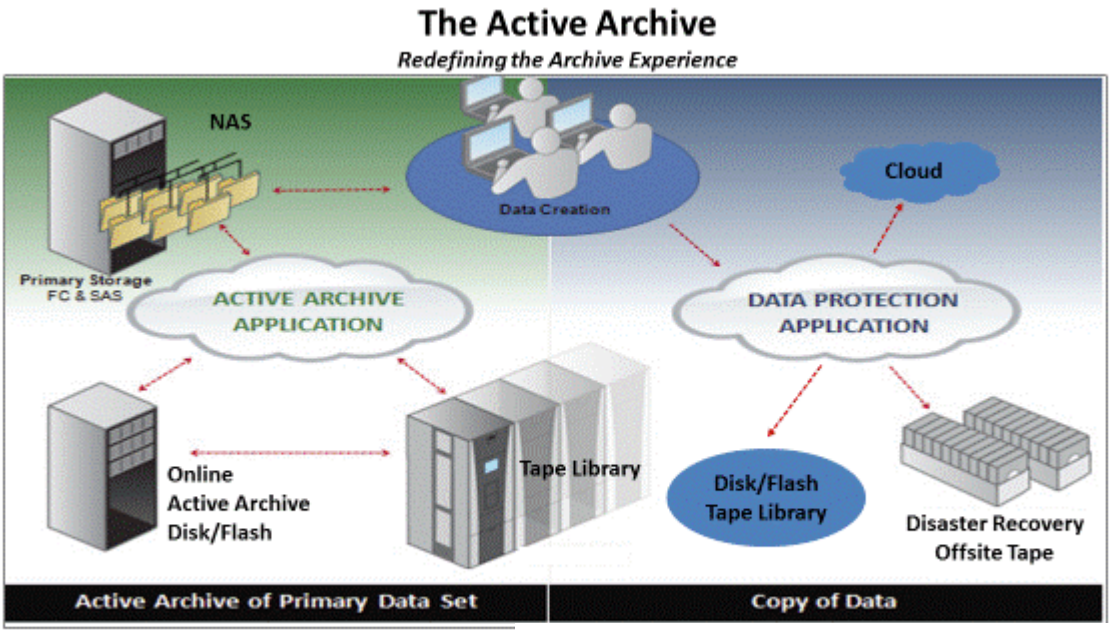
Active archive implementations can use your existing storage equipment to build an integrated hardware and software solution and can incorporate enhanced file systems such as LTFS (Linear Tape File System). For those who do not want to put together their own repository, vendors are offering preconfigured active archive appliances or "NAS heads" with various file systems that work with most any tape library back-end. Active archive solutions support file, block and object storage systems making them extremely versatile for any data type.

Archives are undertaking a radical transformation, driven by much larger file sizes and new access requirements. Object storage enables IT managers to organize archive content sooner.

Object storage software or appliances can transform an archive into an 'active' archive that is positioned between high-performance storage and low-cost disk or tape.

- An active archive combines the simplicity *and performance* of disk with the economics of lower cost storage such as tape in a highly scalable solution.
- An active archive can scale to a billion or more files and holds more frequently accessed archival data.
- The archive of today combines flash, disk, tape, in various configurations with data management software, and will deliver enhanced cloud services.

The diagram below illustrates how you can easily implement an active archive.



Source: Active Archive Alliance, Horison, Inc.

Benefits of an Active Archive Can Include:

- **Ease of Use**. Provide end users with file-level access to all of your data, all the time – without IT intervention. Whether your organization stores active archive data in a privately managed object storage cloud, on tape or in an off-site public cloud, an active archive gives the IT department the flexibility to select the correct storage media while not inhibiting user and application access to the data.
- **Improved Performance**. By reducing the total amount of data to manage, or partitioning inactive data from active data, organizations may see substantial improvement in system performance. Performance is further improved by using disk drives to serve as a cache for tape libraries.
- **Knowledge discovery**. With the advent of Big Data, organizations are learning the value of analyzing vast amounts of archival data to gain a competitive edge in the marketplace. Big Data analytics demand active archive solutions.

- **Lower Cost**. Reduce TCO by matching media type to SLA's by moving archival data to lower cost storage such as high density tape, SMR HDDs or object storage.
- **Reduced Backup Window**. Even with backup to disk using data compression and data deduplication, backup windows face constant pressure from data growth rates that often exceed 40% annually. There is no benefit in repeatedly backing up unchanged data. Archiving can remove terabytes of data from the backup set shrinking the backup window.
- **Scalability**. An active archive can easily add capacity and scale to petabytes of storage.

The Active Archive, Tiered Storage and the Cloud

By providing a persistent online view of archival data by integrating one or more archive technologies (disk, tape *and* cloud storage) 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 as the cloud is now embracing both disk and tape tiers and archival cloud services have momentum. Active archiving brings the same benefits to the cloud that it does to the data center.

Select Case Studies Highlight the Value of Active Archive Solutions:

- CyArk Looks to Active Archive for Preserving the Past and Protecting the Future
- MLB Network Hits Home Run with Active Archive
- NASA Ames Selects Tape Active Archive to Manage and Store High-Volume Data
- NCSA Employs Active Archive for World's Largest Active File Repository

Recent Active Archive Announcements

- SAN JOSE, Calif., December 10, 2015 Western Digital Corporation Joins Active Archive Alliance to Push for Always-on Data Access

Looking Ahead at Active Archives

As the complexity of managing rapidly accumulating amounts of archival data increases, businesses are seeking to optimize their existing or implement a new archival infrastructure. Implementing an active archive process 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-value case for active archives will become increasingly compelling and likely include cost containment (free up disk space), risk reduction to ensure regulatory compliance, improved productivity by getting inactive data out of the path of the backup window, deliver more efficient searches and retrieval, and yield improved storage administrator efficiency. Thoughtful deployment of today's advanced active archive solutions can yield a highly cost-effective and sustainable archive strategy. Isn't it about time to take advantage of the many benefits offered by active archives?

Want to learn more about active archives? Stay tuned for a new webinar series from the Active Archive Alliance coming later this spring featuring Best Practices and Case Studies of Active Archives in action. Save these dates and times on your calendar:

Leveraging Active Archives to Solve Data Protection and Cloud Requirements
Tuesday, May 10, 2016 at 8:00 a.m. PDT/11:00 a.m. EDT

The Role of Object Storage in Active Archive
Tuesday, May 24, 2016 at 8:00 a.m. PDT/11:00 a.m. EDT

MLB Network Hits Home Run with Active Archive
Tuesday, June 7, 2016 at 8:00 a.m. PDT/11:00 a.m. EDT

Note: The Active Archive Alliance includes representatives of DDN Storage, FUJIFILM, Hewlett Packard Enterprise, HGST, 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.