InterSect360 RESEARCH

Actionable Market Intelligence for High Productivity Computing

Active Archives Help Organizations See What They Already Knew

Addison Snell

April 2010

White paper

EXECUTIVE SUMMARY

Active archives are addressing a growing challenge across multiple industries – how to still leverage information once it is stored for good. By applying new file system and tape library technologies, organizations can now build a persistent view of the data in their archives, making it easier to access again if it ever becomes needed.

With a virtualized file system in front of it, a tape library can now become a visible extension of a global storage pool, so the data on tape can be conveniently accessed and administered. This potentially allows an organization to confidently move more of its data to tape, which can have significant advantages over disk in cost per byte, cost per watt, and byte per floor tile.

In addition to the file system capabilities that enable active archiving, significant other enhancements make it practical for many applications, including increased scalability and throughput over previous generations, and enterprise-level data management and reliability features. In particular, tape libraries can now offer predictive failure of media, so administrators can manage their active archives over time.

Many trends favor the adoption of active archives, and additional support will come from the Active Archive Alliance, a multi-vendor consortium with a common interest in promoting the new paradigm. This combination of supply-side and demand-side support will carry active archive adoption ahead in the coming years.

MARKET DYNAMICS

Lost in the Archive: If You Can’t Find It, Do You Really Have It?

The Archive. Is there any other location that evokes both confidence and hopelessness in such equal magnitude? Consider the fate of the Ark of the Covenant at the end of *Indiana Jones and the Raiders of the Lost Ark*, being stowed in a vast government warehouse in an indifferently numbered crate, about to become at least as lost as it was when it was buried in the desert and guarded by an improbable number of subterranean snakes. On a smaller but still frustrating scale, this is the same problem we deal with when we search for the fondue forks amidst the clutter in the back of our kitchen cabinets, the bottom drawer, the guest room closet, and the three boxes in the basement that are marked only “KEEP!!” We know we have them somewhere, we just don’t know where.

Data archives suffer the same problem. Never sure which data is important to KEEP and which can safely be deleted, organizations are writing exponentially increasing amounts of data to their archives. The majority of this data conforms to an access model known as WORN – “write once, read never.” It doesn’t matter if you can’t ever find it again; you just have to know it’s there.

This works great, right up until the time you need the data equivalent of the fondue forks, and we begin a storage archive version of “Where’s Waldo?” Imagine a geologist attempting to prove a new theory that major earthquakes can be predicted by a particular pattern of low-level temblors in the preceding days. Finding the seismic data that preceded major earthquakes should be easy enough, but what about finding times when the same pattern occurred *without* a major quake afterwards? That previously unremarkable data exists, but where?

This pattern is repeated in multiple industries that need to keep incredible quantities of old data, in case they become important again, or because of regulatory mandate to preserve it. Consider the following possibilities:

- An oil company develops a technology for analyzing sub-salt formations in areas it previously decided narrowly not to drill. Is it worth it now?
- The state wants to implement photo-based tickets for red-light violations. Can they store and access the digital images efficiently and conveniently?
- A new theory suggests that thicker ventricular walls, visible on CT scans, may correspond to lower incidence of stroke later in life among men who are also at least 20 pounds overweight. Does existing data support the theory?
- Due to a lawsuit disputing ownership rights, a bank is asked to find a scanned image of a check from seven years ago. How quickly can it be located?
- A woman suddenly is thrust into the news for committing or suffering some heinous crime. She was a Laker Girl dancer for two years in the early 2000s. Can ESPN find footage of her dancing at a Lakers game?

These examples are hypothetical, but they also feel plausible. If the questions that we eventually do ask of our data aren’t these exact questions, they will be equally challenging from a data retrieval aspect. Also, each case comes with a high degree of confidence that the data exists, somewhere, if only we could lay our hands on it. There is a philosophical question in play: If you can’t find the data, do you really have it?

Active Archives: See How Much You Know

This challenge, persistent across many data-intensive industries, is being met in a new way, brought about by recent, simultaneous data management technology improvements. The concept is generally known as “active archival,” and the concept is simple. Active archives use advancements in data management software to turn offline archives into visible, accessible extensions of online storage systems through a file system interface. Rather than losing data in the archive, organizations can maintain persistent views of all their storage, thereby increasing the likelihood that they can find and retrieve critical pieces of information when they want them.

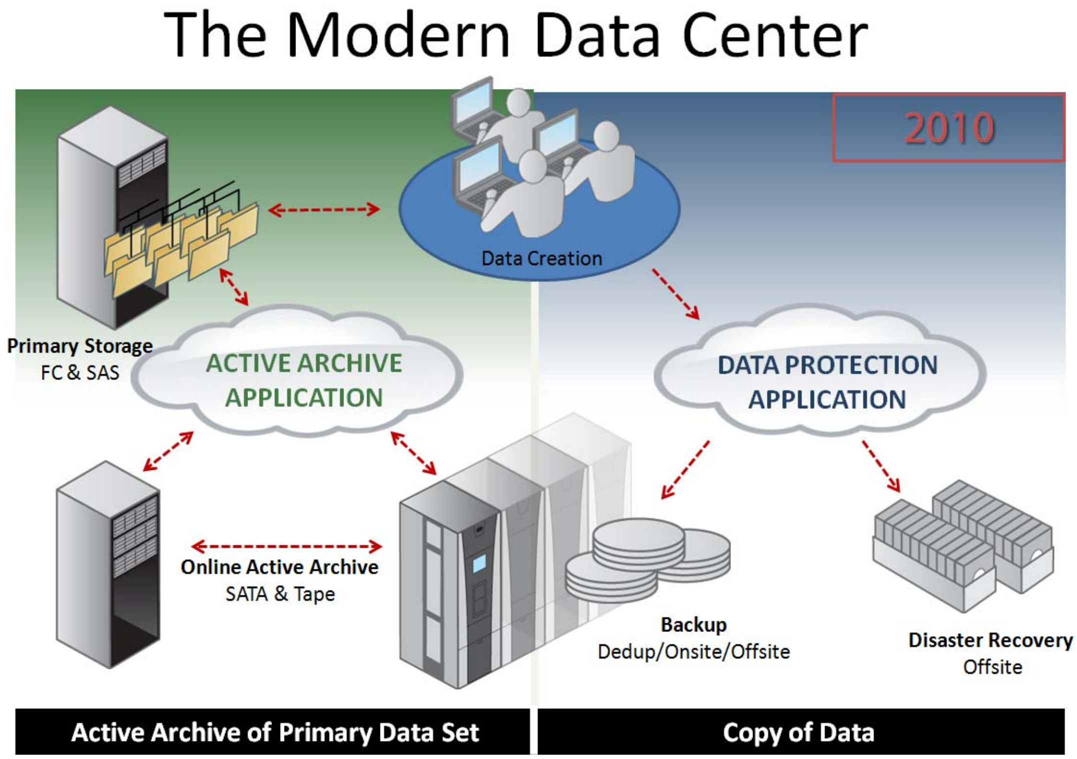
The appeal of active archives is so readily apparent that it raises the question of why they weren’t used widely before now. For one, the need to access hidden treasures in vast oceans of stored data is appearing in more and more industries, and the exponential increases in the amount of data created is continuously compounding the problem. For another, the technologies necessary for active archives either didn’t exist or were not cost-effective until recently. Organizations could choose between keeping data visible in online disk arrays and moving the data to more cost-effective offline tape. The tape archive is where you would put the data once you didn’t think you’d need it again.

There are several technologies necessary to build an active archive. Perhaps the most critical (and newest) of these is the ability to see and access data on tape through a file system interface. File systems keep data organized on disk. File systems make it relatively easy to view directories of existing data files, and they allow end users to search those directories for particular files. Most importantly, file systems are the map that links the name of a file to its location in storage.

Tape archives have not previously benefited from this degree of organization. The data on disk was visible under a file system; the archives on tape were not. This made unexpected searches for information on tape an onerous task. If you knew what you would be looking for, you could plan for it, but the unanticipated queries would present a challenge. Many organizations would compensate for this by holding more data on disk, but even less-expensive, lower-performing disks tend not to be as cost-effective as tape, either in terms of capacity (dollars per byte) or facilities (space and power consumed).

Recent evolutions in both file system and tape technology have allowed the possibility of viewing a tape archive as a normal extension of the storage infrastructure, making files on tape just as visible as files on disk. This new ability to view a tape library through a file system interface allows organizations to keep track of critical data after it has been archived. This is an important innovation, because it improves the likelihood of finding data again when needed. Rather than a final resting place for information, an active archive is a connected repository of institutional knowledge, retrievable when the situation suddenly arises.

Figure 1: View of an Active Archive Architecture in 2010
Source: Spectra Logic



Active archiving enables simplified, online access to an organization's primary data set archives on both SATA and tape storage. A copy of the data can be stored as an onsite backup or offsite disaster-recovery image.

This essential advancement, access to tape through a file system, is driving the trend in active data archival, along with a few other essential technology elements. Other necessary advancements – all recently enhanced – are:

- *Scalability:* The tape archive must scale to a level that accommodates the information flow.
- *Performance:* The data must be retrievable in a timeframe that makes the archive useful.
- *Facilities efficiency:* The archive must fit within the facilities budget for floor tiles covered and watts consumed.
- *Data management:* There must be software tools for efficiently migrating files between disk and tape.
- *Data verification:* The user needs to know when data is written and to have confidence that it is still good data years down the road.

Fundamentally, active archives are meant to help organizations deliver new insights on top of what they know. With this combination of technologies in play, organizations with increasing libraries of institutional knowledge can turn to active archives as a way to continue to take advantage of the data at their disposal.

Active Archives in Practice

Despite recent advancements that make active archives practical for a wider range of applications, they are not an entirely new concept. Many media and entertainment companies have been doing some form of active archiving for several years, using proprietary software to provide the interface to the file structure in the tape library. The internal investment required for such a solution has been driven by that industry's persistent need to leverage its historical data, a trend that has gained prevalence across multiple industries.

More recently, organizations like NASA's Ames Research Laboratories have taken active archiving to the next level, using standard file system technologies to view their tape libraries. The trend toward open systems development – incorporating standard applications into active archives, rather than relying on proprietary development – is another driving factor in active archive adoption.

These early adopters have paved the way for active archive as an industry trend, and the need to access large volumes of data exists across many markets. The maturation of the requisite technologies – the file systems, the scalability, and the reliability of tape libraries – creates an opportunity for active archives to become a notable trend in the near term.

File System Interfaces for Active Archives

Of all the technologies associated with large-scale datacenters and high performance computing, file systems are going through some of the most significant transitions. Once one of the most fragmented technology areas – a 2007 qualitative InterSect360 Research study found 18 different file systems in use among only 39 organizations[1] – recent years have seen consolidation due to industry acquisitions and advancements in parallel clustered file systems.

The most common file systems used in datacenters are NFS (predominantly in UNIX and Linux environments) and CIFS (predominantly in Windows environments). As the scale of their data management infrastructure increases, many organizations turn to parallel clustered file systems, which offer greater scalability and throughput than NFS. So far the most adopted of these has been GPFS[2], and there is also considerable market interest in pNFS, an upcoming, multi-vendor consortium-driven standard for parallel NFS.

Active archives do not require a specific file system; rather, file system choices will be driven by the application software used in the active archive itself. The consolidation in the file systems market, coupled with a trend toward open systems

[1] Comprehensive Research Study: "Data Management Requirements in High Performance Computing," 2007. Published as Tabor Research.
[2] InterSect360 Research, High Performance Computing User Site Census data, 2009 – 2010.

development, will help create active archive environments that are easily adoptable into existing data management infrastructures.

Specifically, many active archive applications – such as HPSS from IBM or StorHouse from FileTek – function through a virtualization of the file system. Virtual file systems are mounted file systems that can be either local or remote, but they appear as local file systems to an end user, who can switch between mounted file system views. Through a virtualized file system, a tape library can appear as part of a global storage pool. It is this interface that gives organizations a persistent view of their archived data. Further developments of applications compatible with parallel clustered file systems will provide increased opportunity for industry-standard adoption of scalable active archive solutions.

INTERSECT360 RESEARCH ANALYSIS

Adoption of Active Archive

InterSect360 Research studies have consistently shown industry trends that would support the adoption of active archive. Despite increasing densities and falling prices for storage, the growth in data requirements is driving storage to take up an increasing percentage of end users' budgets[3], as well as a climbing share of product and services spending over the next five years.[4] These trends will continue to pressure organizations to be efficient with their data management strategies, including the consideration of how to optimally include tape archives as a productive part of their infrastructures.

Although "active archive" is a relatively new term, the concepts have been incubating for years. Under one name or another, we expect the concepts behind active archive to continue to grow. (Names have evolved for other tape concepts in the past, as "hierarchical storage management" begat "information lifecycle management," which is now more commonly referred to as "tiered storage.") Whether active archives persist as a phenomenon by that name may be tied to how many companies hitch their marketing wagons to that horse.

Active Archive Alliance

For active archives to gain widespread acceptance and long-term success, it will be critical to have broad adoption involving multiple technology and end-user partners. In support of that goal, several leading technology companies have announced the formation of the Active Archive Alliance, an industry group dedicated to the propagation of technologies to ensure the enduring accessibility of archive data. Some of the founding members are:

- Spectra Logic, provider of the highly scalable T-Series line of tape libraries for applications in big-data environments in both research and production
- FileTek, a storage management solution provider and the supplier of StorHouse data virtualization software, which interfaces with either NFS or CIFS for active archive applications
- QStar Technologies, which provides enterprise archival storage solutions across multiple industries
- Compellent Technologies, a provider of high-density SAS and SATA drives, performance-optimized RAID, and Compellent's Fluid Data Architecture for the automation of data movement and management

This technological ecosystem will help ensure that adopters of active archives are able to fully realize the benefits in a way that is adoptable into their existing infrastructures, and we expect other companies to join the Active Archive Alliance

[3] InterSect360 Research, End User Site Budget Map: Economic Sector, 2009.

[4] InterSect360 Research, "Traditional HPC Market Model and Forecast, 2009 – 2013," 2009.

with their technologies and solutions soon. Although active archives are still in their early phases, a broad base of industry participation increases the likelihood that the concept will take hold.

Future Outlook

The trends that have been driving the need for active archival are intensifying rather than lessening, and the technologies that enable the solution have begun to mature. The coming years could therefore present a “tipping point” of adoption, provided that awareness catches on, both in research-oriented HPC applications and in the broad enterprise realm of persistent data sets. That awareness will require a willingness for organizations to update their thinking about tape as a static – even bygone – technology, as innovations challenge their established perspectives. But if marketing a tape-based solution seems like a tall order, at least we can be reasonably confident that potential buyers are in search of new technologies to address their data archival and retrieval challenges.

The new mindset that Spectra and other active archive solution providers will need to create is that tape – through these new innovations – is the new technology being sought. But even here the metaphor supports the claim. The whole point of active archiving is to continue to drive new insights by applying novel thinking to established knowledge. When organizations are shown how they can remember what they already knew, they might also realize new benefits to storing data in a visible tape archive.