THE CLIPPER GROUP
Navigator™
Navigating Information Technology Horizons℠

Published Since 1993 Report #TCG2010029 June 11, 2010

Simplified, Online Access to Archived Data — Turning History into an Active Archive

Analyst: David Reine

Management Summary

We are living in an increasingly complex world that we attempt to understand by collecting more and more data. In many cases, unfortunately, the true value of that information is not known immediately. That true value comes to fruition months, or even years, down the road when we have the opportunity to look at it in historical perspective, if we can find it. I think back to the TV series *Stargate SG-1* in which one Stargate, a portal to other galaxies, has been stored as an artifact in an unmarked crate in one of many nameless government warehouses. Once found, and activated, it was put to use as a gateway to obtain knowledge of the universe previously unavailable. The key was the ability to locate the forgotten Stargate and put it to good use.

In today's data center, we are collecting reams of information, such as medical diagnostics and research. Because of the costs and energy requirements associated with keeping all of this data online, for decades, medical data centers frequently archive the information on magnetic tape to preserve it for future use. Archiving data is not unique to medical facilities. Enterprises engaged in gas and oil exploration generate as much new data, if not more, annually, in the search for the natural resources necessary to run the power plants to power our homes and offices and the gasoline needed to drive our cars and enable our transportation systems. This information must also be archived in order to determine which discoveries might be the best value for drilling in the future. **Unfortunately, in the past, when data was archived to tape, it often became more difficult to access.**

However, with recent innovations in tape hardware and software, the data center can now use tape that acts like disk, in much the same way that Virtual Tape Libraries (VTLs) use disk to act like tape. With new file systems acting as a front end to tape, along with extensions to standard file systems, the lifecycle of tape comes full circle, with tape regaining prominence in the data center as an extension to disk. With a proven cost-effective storage media, the IT staff can once again get online access to active, fixed data, using tape as an *active archive* to make enterprise data archives searchable, accessible and available, re-connecting to lost or forgotten data, with a persistent and easy-to-use interface.

What, you might be asking, is an *active archive*? It is an archive that contains production data, fixed content or variable, no matter how old or infrequently accessed, that can still be retrieved online. Online access to this archive is achieved through a file system interface, to facilitate ease-of-use, enabling a search of all of the historical data. Some of the goals of an active archive would be to improve storage density and reduce footprint to lower cost, and to ensure media and data integrity in order to promote business continuity. An active archive should also enable the IT to manage media within the library better. To learn more about the advantages of an active archive and how the *Active Archive Alliance* is working to simplify this architecture, please read on.

IN THIS ISSUE

Proliferation of Data in the Enterprise

Over the past few years, data center storage has undergone an era of unprecedented growth. Between mergers and acquisitions, consolidations and virtualizations, industry regulations and government compliance, enterprise data centers are storing more data, and more kinds of data, than ever before, not only for backup and recovery, but for long-term archiving, as well, in order to take advantage of the long-term value of this data.

This data can be segmented into three basic categories: *structured data*, such as databases, *semi-structured data*, such as office documents and email, and *unstructured data* (often binary objects, such as research and medical imaging, digital photos, and broadcast media files). Furthermore, the data center is storing more copies of this data than ever before, in order to mitigate risk. In fact, the issues surrounding the backup of structured data have been addressed for quite some time and most data centers have this segment under control. The same cannot be said for semi-structured data, where new technologies (such as data deduplication) are currently being deployed to try to get a handle on the issue of backup and recovery. The pains surrounding the preservation and retrieval of binary objects, and the length of time this data needs to be preserved, remain outstanding and currently are attracting the most attention.

Structured Data

When we think of structured data, one of the largest categories that come to mind would be databases. Usually managed by applications at the block level, databases are growing by leaps and bounds, primarily because of replication, with these mission-critical backup copies typically being held for months. The IT staff can try to control the growth with replication management, based on approved best practices, and with the deployment of new data deduplication appliances, which can become expensive. The staff can also try to control the cost of storage for structured data by migrating these secondary copies from more expensive SSDs and FC disks to less expensive SATA and SAS disk arrays. Another means of establishing capacity control is via Hierarchical Storage Management (HSM) software to manage out-of-date versions.

Finding the right copy of data, by date and time, is usually managed by the indexing provided by the backup software. Security concerns typically are addressed by features built-in to the application software or by a separate encryption appliance deployed by the data center.

Semi-structured Data

The continuing growth in unstructured data is typically fed by the propagation of office documents, such as email, letters, spreadsheets, PowerPoint slides and PDFs, and the like, which often are text oriented and distributed on a one-to-many basis. These documents, and often many copies of these documents, are typically retained for years, often to protect the enterprise against litigation, sometimes to their detriment. The backup requirement for these files is very high to enable the data center for disaster recovery, or merely to restore files that were deleted accidentally. In order to control the backup window and the sheer volume of data being preserved, many data centers have deployed data deduplication engines in order to eliminate multiple copies of the same files, and have adopted an approved policy for the frequency of deletions. Some data centers also employ HSM software to migrate files to a tier of storage with a lower TCO in order to reduce the volume of backups. These files usually have multiple levels of security, from file systems with access control lists to encryption, in order to protect the data from theft or inadvertent disclosure.

In many cases, the data center will migrate these documents from an active storage arena into an archive in order to free up primary storage, reduce the backup window, and protect the enterprise from litigation. In fact, **backup and archiving are two different processes with two different end results**. Enterprises that have not built archiving infrastructures are missing opportunities to conserve storage, improve performance, and more intelligently manage one of their most important assets – their data.[1]

Unstructured Data and Binary Objects

The issues surrounding the preservation of binary objects are multiple and complex. What the data center has to deal with could range from the meteoric growth path of digital photos and music being stored and shared on any number of social media websites to the terabytes of digital data contained within a growing number of 3D movies, such as *Avatar* and *How to Train Your Dragon*. All of these objects share a common

[1] See the issue of Clipper Notes dated February 1, 2007, entitled *Archiving – Do You Need It?,* and available at http://www.clipper.com/research/TCG2007018.pdf.

characteristic – they are unchanging. In addition, they are typically undecipherable by humans and, thus, not easily indexable. The growth of this segment is very large, as the files tend to be large in size, growing daily in number, and are retained for long periods of time. Other categories that fall into this segment include large data files that feed HPC applications, such as those involved in oil and gas exploration, astronomy, physics, and biotechnology.

The urgency of retrieval demands that the long-term preservation of this data be managed in an *active archive*, in order to have rapid access to the information in a cost-efficient manner, with the cost of the storage media employed often being a determining factor. Migration to slower and less expensive media being employed as the data ages and the frequency of retrieval requests diminish. Let's take a look at medical imaging, for example. Medical science has used technology to advance our understanding of the human body from the simplest X-ray to increasingly complex MRIs and CAT scans. Every image that doctors take helps them to diagnose diseases earlier than ever before. As you might expect, each of these images generate megabytes (or gigabytes) of data, which (for medical and legal reasons) have to be studied, shared, and preserved for decades. That preservation can be migrated to less expensive media when there is no immediate requirement for it. However, if a patient is scheduled for surgery, or simply a checkup, that information needs to be found and retrieved in time for that next appointment.

The data created by these images, however, can be dwarfed by the information created by genome sequencing instruments, which can put out several terabytes of data per day. In laboratories with many of these devices, it doesn't take much to imagine petabytes of new data being generated per year.

Storage capacity requirements to cope with this demand typically are expanding by a magnitude of two every twelve-to-eighteen months. Floor space, energy, and administrative costs are taking a huge bite out of the IT budget, not to mention maintenance and other operational expenses. While the acquisition cost per terabyte for active archive storage may be fairly stable, or even coming down, these ancillary costs contribute heavily to the total cost of ownership (TCO) of the IT infrastructure.[2]

The Need for More Storage

For many enterprises, the archive storage capacity requirements for unstructured data and binary objects is greater than that for semi-structured data, which in turn is greater than that of structured files. Archive capacity, however, is only one of IT's stress points. While meeting capacity requirements does contribute to escalating cost of storage, the data center must find a means to lower the overall TCO of the storage environment, while at the same time getting a better handle on historical data, if it is to succeed in addressing the needs of the enterprise. This includes improving energy efficiency with techniques, such as MAID[3], increasing the density of backup and archive media, improving the reliability of the storage process, implementing faster access times, and most significantly, easily integrating with archive management applications to enable a simplified access to an active archive.

For most enterprises, primary storage will continue to find a home on a heterogeneous mix of high-performance disk devices, consisting of the highest-performing Tier-0 SSDs[4], high-availability Tier-1 Fibre Channel (FC), and high-capacity Tier-2 SAS/SATA, as will backup images for data with immediate recovery requirements. Enterprise RPO and RTO policies will dictate which backup information needs to reside on disk. However, whenever economics is a concern, best practices for data retention in the data center indicate that long-term storage of email and other compliance documents and archiving environments shall continue to reside on less expensive media as long as that media can satisfy a *timely* retrieval from the active archive. This will protect the enterprise and its officers from failure to comply with internal policies and government regulations. **In the end, it doesn't matter where the information came from; it must be protected and easily accessible.** In fact, some data centers that had evolved to a disk-to-disk (D2D) environment (i.e., without tape), now are returning to tape (D2D2T), in

[2] See the issue of Clipper Notes dated October 21, 2008, entitled *Disk and Tape Square Off Again – Tape Remains King of the Hill with LTO-4,* and available at http://www.clipper.com/research/TCG2008056.pdf.

[3] Massive Array of Idle Disks.

[4] See the issue of Clipper Notes dated January 26, 2009, entitled *A New Tier of Storage Appears – Faster Solid State Drives State Their Case,* and available at http://www.clipper.com/research/TCG2009006.pdf.

order to take advantage of its high capacity, portability, low-cost WORM[5], and encryption technologies. In many cases, storage management applications enable the IT staff to establish policies for the migration of data from disk to tape and back again as enterprise objectives demand.

With more data being collected, the need to preserve it in active archives continues to grow. There is now a need to accelerate retrieval of this data in order to facilitate a simplified access to these volumes of information. In order to access this archived data, the data center will need an easy way to index, query, and retrieve a high volume of data, whether from disk or tape, quickly, simply, and efficiently. They also need to ensure media and data integrity with a simplified management process. This is especially true for the storing of fixed content data for ongoing mission-critical business uses.

The Need for Better Archives

The greatest unmet needs in this arena are for the long-term archiving of unstructured data and binary objects, with the most significant technical and cost challenges on the latter because of the size and retention period requirements. Since the underlying storage devices continue to improve and, sometimes, change to incorporate new features, the storage targets for the long-term archive must be virtual, i.e., represented independently of any particular (technology-specific or proprietary) physical characteristics. **What this means is that the archiving solution should not be targeted at a particular tier of storage or storage product. To achieve this, a common, device-independent architecture is required. Practically, this means that software needs to be written to allow an archive to be written, wholly or partially, to a specific class of storage devices or to a specific storage device. To allow for transportability (between vendors of archiving solutions and/or devices) or interoperability among them, a standard interface and storage architecture are required.** (See Exhibit 1, above, for a list of requirements.)

The Need for Lowering the TCO of Archives

With the demands for more data and more archives growing exponentially, the Total Cost of Ownership (TCO) for archived data becomes the most significant challenge. While an all-disk archive solution may be required for some business uses, most archived data requires a less-expensive vehicle for storing the vast archives. Tape must be part of the answer. (More on tape, later in this bulletin.)

Exhibit 1 —
Active Archive Requirements

- Cost-effective storage of all data online, with standardized archive architecture and common simplified search and retrieval.
- Provide timely access to all archived data.
- Reduce backup window through the migration of rarely used data from disk to tape.
- Enables archives stored on tape to meet compliance and e-discovery requirements.
- Lowers TCO via assignment of data to the correct storage tier.
- Eliminates trade-off between online visibility and nearline cost benefits.
- Increases efficiency of the data center through savings in energy consumption and floor space from high-density tape libraries.

The Active Archive Alliance is Formed

As organizations collect more and more data, their archives continue to grow, and the challenge of accessing that information intensifies. The *Active Archive Alliance* is a multi-vendor industry group recently formed to help bring together the information that the data center needs to assemble, efficiently, into an active archive system, thus taking advantage of the existing data infrastructure investments and enabling a better long-term solution.

The Alliance will provide IT organizations with "the best practices, tools and information they need to achieve simplified access to the online storage of their archived data."[6] In this manner, organizations can turn offline archives residing on tape into persistent, visible, and accessible extensions of their online storage environments. Formed in April, Compellent Technologies, FileTek Inc., QStar Technologies, and

[5] Write Once, Read Many (times).

[6] Active Archive Alliance press release dated April 27, 2010. See http://www.activearchive.com/news.

Spectra Logic have joined forces as founding partners, with the express purpose of helping organizations, both public and private, which are "grappling with data growth, retention compliance, and the need to leverage their knowledge and information."[7]

Each of the founding vendors brings unique experience to the alliance.

- **Compellent** provides ultra-efficient and flexible disk-based solutions that fit seamlessly into active archives,
- **FileTek** provides intelligent storage virtualization and data management,
- **QStar** provides enterprise-class archive, data management, and disaster prevention software and solutions, and
- **Spectra Logic** provides high-density, feature-rich tape and disk storage, supporting high-performance active archives.

The availability of file system interfaces that span an entire pool of storage, including both online disk arrays and high-density tape libraries will enable the data center to leverage all of this valuable data resource.

Leveraging Archived Data

As a result of today's economic conditions and the awareness of energy limitations, every organization is looking at ways to reduce the TCO of their IT infrastructure, especially that pertaining to long-term archiving. Server consolidation in the data center has led to storage consolidation, increasing the amount of data being backed-up and archived. **As the size of these archives continues to grow, the challenge to provide continuous access, with the ability to know what has been stored, where it has been stored, with the capability for *timely* search and retrieval from the archives, grows proportionately.** In order to gain control over these archives, the IT staff needs to have an established list of best practices to gain access to these volumes of data, such as those proposed by the Alliance. They also require appropriate tools to gain rapid access to this archived information.

What types of data are experiencing the most growth in your data center? Quite clearly, the volume of e-mail, structured data, video files, and office documents has experienced the most significant growth in the past few years. These files are typically archived according to pre-determined enterprise policies, and, for the most part, rarely if ever accessed after a month or two or three. However, if litigation rears its ugly head, for example, the need to find specific information becomes urgent, and possibly quite expensive. The need to access quickly the correct records can become mission-critical overnight. Active archives provide the enterprise with an affordable online solution to store and access all of this newly consolidated data.

As a result of the Alliance, compatible active archive applications will enable the data center with the ability to see and access archived data throughout the enterprise, through an extended file system interface. They give the IT staff an easy and affordable way to view and search archived data files across both tape *and* disk archives.

Leveraging LTO-5

Tape libraries with *LTO-5*[8] drives and media appear to be an ideal vehicle to improve storage density and reduce storage footprint while implementing an active archive to enable the enterprise to retain *all* archive data online, searchable, and quickly accessible. With a long-term roadmap, LTO provides the data center with an energy efficient solution that can scale – reliably – in both capacity and throughput for years to come.

With the availability of LTO-5, the native *capacity* for archiving on cartridge tape has increased, to 1.5TBs of uncompressed data per cartridge, lowering the TCO on a cost/TB basis. This is a fifteen-fold increase in capacity in only 10 years. The native *throughput* also has increased, from 15MB/s with *LTO-1* to 140MB/s with LTO-5, almost ten-fold. Over that time span, we have seen the inclusion of an integrated WORM capability, for compliance, beginning with *LTO-3* and embedded encryption, for security, with *LTO-4*. With LTO-5, we now see the inclusion of media partitioning as part of the specification, allowing the data center to improve data management and access, which might allow the enablement of self-describing media containers and structured data on tape. In

[7] Ibid.

[8] See **The Clipper Group Navigator** dated January 29, 2010, entitled *LTO Program Announces Next Gen Tape – LTO-5 Raises the Bar for Tier-3 Storage*, available at http://www.clipper.com/research/TCG2010002.pdf.

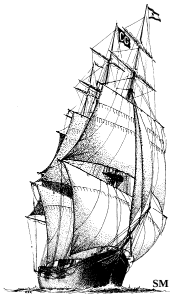
addition, LTO-5 contains reliability features to ensure media and data integrity that are an improvement over previous generations, including advancements in coating of tape film, read-after-write data verification, error correction codes, simplified tape paths and servo tracking systems. Moreover, there is an abundance of applications designed to improve media management within the library and to manage media for reliability.

In April, the LTO Program announced their roadmap for *LTO-6*, *LTO-7*, and *LTO-8*. These extensions increase native capacity from 3.2TB for Generation 6, to 6.4TB for Generation 7, to 12.8TB for Generation 8.[9] With a larger compression history buffer, this will create compressed capacities of 8TB, 16 TB, and 32TB, respectively. Native throughput also will increase for each new generation, from 210MB/s, to 315MB/s, to 472MB/s.

Conclusion

Many enterprises, both public and private, are struggling today with the problems created from massive data growth and the need to retain and protect that data while, at the same time, trying to leverage that information to create useful knowledge. The enterprise data center needs to be able to access that data online in order to increase its value, while at the same time enabling the dynamic migration of this valuable resource to the most economical media in terms of cost vs. availability. In addition, making data archives accessible and available will lower their TCO while, at the same time, accelerating compliance with industry and government regulations to ensure the safety and integrity of the information in their charge. Active archives deliver a simple and persistent interface for data access. The Active Archive Alliance will provide a common, compatible way to manage active archives.

Recent innovations in LTO-5 will enable the data center to realize lower TCO for active archives, in terms of power efficiency, data density, and throughput.[10] In addition, the scalability and performance available in current tape libraries enhances the information flow and meets the demands of facilities management for improved space efficiency. With the improved reliability of a well-managed active archive, the data center can ensure data access and improved profitability by evolving their environment to an active archive architecture deployed on both high-performance disk and low-cost tape.



[9] For LTO roadmap showing LTO-6 through LTO-8, see http://www.ultrium.com/technology/generations.html.

[10] Op. Cit., **The Clipper Group Navigator** cited in footnote #8, page 3, footnote #2.

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

David Reine **is a Senior Contributing Analyst for The Clipper Group.** Mr. Reine specializes in enterprise servers, storage, and software, strategic business solutions, and trends in open systems architectures. In 2002, he joined The Clipper Group after three decades in server and storage product marketing and program management for Groupe Bull, Zenith Data Systems, and Honeywell Information Systems. Mr. Reine earned a Bachelor of Arts degree from Tufts University, and an MBA from Northeastern University.

- ***Reach David Reine via e-mail at dave.reine@clipper.com or at 781-235-0085 Ext. 123. (Please dial "123" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group *Captain's Log***, **The Clipper Group Voyager**, **Clipper Notes**, and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, *"Navigating Information Technology Horizons"*, and *"teraproductivity"* are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosures

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

After publication of a bulletin on *clipper.com*, The Clipper Group offers all vendors and users the opportunity to license its publications for a fee, since linking to Clipper's web pages, posting of Clipper documents on other's websites, and printing of hard-copy reprints is not allowed without payment of related fee(s). Less than half of our publications are licensed in this way. In addition, analysts regularly receive briefings from many vendors. Occasionally, Clipper analysts' travel and/or lodging expenses and/or conference fees have been subsidized by a vendor, in order to participate in briefings. The Clipper Group does not charge any professional fees to participate in these information-gathering events. In addition, some vendors sometime provide binders, USB drives containing presentations, and other conference-related paraphernalia to Clipper's analysts.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.