



## Research Study

# The Rise of Active Archiving: Hitachi Content Archive Platform Version 2.0

This report provides an overview of the information management challenges that are driving organizations to adopt active archiving solutions and examines the features, capabilities, and architecture of one such solution, the Hitachi Content Archive Platform. The newly released Version 2.0 of the Hitachi Content Archive Platform adds several significant enhancements to an already strong product.

The Hitachi Content Archive Platform is a relatively new but notable entrant in the market for active archiving solutions. Many of the information management challenges that active archiving is intended to address are especially well-addressed by Hitachi's Content Archive Platform. In order to appreciate the strengths and benefits of Hitachi's solution, it is helpful to have an understanding of the information management challenges that organizations face.

Among the forces driving enterprises to implement active archiving solutions, the three most important are: regulatory compliance and information governance; electronic discovery and litigation support; the explosive growth of electronic information.

## Compliance and governance requirements

Businesses and other kinds of organizations have always had internal policies and processes for retaining important business records. Large organizations typically employ full-time professionals to manage their record-keeping operations. When the records in question are physical objects (for example, papers or microfilm), records management professionals can draw on a set of

Copyright © 2002-2007 Data Mobility Group, LLC. All Rights Reserved. Reproduction of this publication without prior written permission is forbidden. Data Mobility Group believes the statements contained herein are based on accurate and reliable information. However, because information is provided to Data Mobility Group from various sources, we cannot warrant that this publication is complete and error-free. Data Mobility Group disclaims all implied warranties, including warranties of merchantability or fitness for a particular purpose. Data Mobility Group shall have no liability for any direct, incidental, special, or consequential damages or lost profits. The opinions expressed herein are subject to change without notice.

long-established and highly effective best practices to ensure proper governance of these records. However, best practices for managing electronic records (i.e., data generated or captured by computers) are only just beginning to be established, and relatively few organizations have effective policies and processes in place.

In addition to internal policies, organizations must comply with a growing number of external regulations. These regulations, imposed by government and industry bodies, variously dictate the kind of information that must be retained, the length of time to retain it, and even the manner in which it must be stored (for example, there may be explicit requirements related to access control, encryption, immutability of data, etcetera). Records and information management professionals can recite a litany of such regulations: Sarbanes-Oxley, HIPAA, NASD 3010, 21 CFR Part 11, Basel II, SEC 17a-4, and so on, ad infinitum. While many regulations only apply to particular industries or in particular geographies, there are literally thousands of regulations and nearly all enterprises will be subject to at least some of them.

## **Electronic discovery**

Electronic discovery, or e-discovery, refers to the process of locating all relevant information in the event of a lawsuit, audit, or other type of investigation, and then providing that information to lawyers or investigators. Unfortunately, with the information management systems in place in the vast majority of organizations today, gathering “all relevant information” for a discovery request is well-nigh impossible. It is almost always a shockingly expensive and time-consuming undertaking, and the results are usually fraught with oversights and omissions. The cost of responding to an electronic discovery request can easily run into the hundreds of thousands of dollars, and not infrequently into the millions. And if the information that is eventually produced is incomplete or is not produced quickly enough, the consequences can be severe: monetary damages—in the form of fines, settlements, or judgments—can reach millions of dollars.

Amendments to the Federal Rules of Civil Procedure (FRCP) in December 2006 have added to the pressure. The FRCP revisions make clear that organizations must be able to produce electronic information in a timely manner. The new FRCP also formalize the concept of a “litigation hold” for electronic information. Whenever litigation can be reasonably anticipated, an organization is obligated to make sure that no potentially relevant information gets deleted; whatever retention and disposition policies that an organization might observe under normal operating conditions are to be overridden.

Moreover, it isn't just lawsuits that organizations must worry about. When faced with a regulatory investigation or an audit, an organization may be required to demonstrate compliance or produce requested information in a very short period of time.

## **A flood of information, a dearth of management**

Anyone who works in IT is aware that the amount of data that must be stored and managed has been growing by leaps and bounds, and the already astounding rate of growth appears to still be accelerating. A large portion of this rising tide of data is comprised of semi- and unstructured content (Microsoft Office documents, images, audio, video, etcetera). Much of this content could either be deleted or moved from primary storage to a less expensive alternative. This would yield benefits such as shortened backup and restore times, improved application performance, reduced costs, etcetera. One reason to implement an archive tier of storage is simply to obtain such benefits, or avoid the inverse problems (i.e., unacceptably long backup and restore times, degraded application performance, spiraling costs, and so on).

But it's not just the sheer volume of semi- and unstructured content that is creating problems, it's also the number of places where this information is stored—departmental file servers, individuals' laptops, Tier 1 disks on the SAN, application-specific (i.e., stove-piped) repositories, sitting on backup tapes in a remote warehouse, etcetera—and it's the fact that organizations typically have no idea what data is contained in all those files and have no way of searching across all the different files in all the different locations.

Having files scattered all over the place exposes an organization to unnecessary risks. Those files may contain confidential data about customers or employees that is required to be kept private, or may contain intellectual property or other valuable business information that should not be allowed to fall into the wrong hands. Most importantly, with files stored in a multitude of locations and systems, and with the content of those files unknown and invisible to IT administrators and business users, it is difficult at best to enforce compliance and governance policies, and eDiscovery becomes either fantastically expensive or just plain impossible.

## **A perfect storm**

The forces described above are interrelated in such a way that each serves to compound or amplify the others. We are still in the early days of this gathering storm, but it has already become painfully clear that organizations are going to need new technology—both software and hardware—to help them achieve acceptable levels of information governance, avoid eDiscovery fiascos, and deal with ever-growing amounts of content. There is no “silver bullet” solution, no single product or process that will dispel the storm, but it is possible to identify foundational technologies that will inevitably be part of the solution. Active archiving is one of those foundational technologies.

## **Active archiving: a technology borne of necessity**

Archived data has traditionally been stored on tape, or some other offline medium, and warehoused in an off-site facility. That approach made sense in a world where regulations didn't specify the details of how data should be stored, and where archived data didn't necessarily need to be accessed quickly or frequently, but that is no longer the world we live in. In today's world, organizations are subject to numerous regulations governing the storage of data, and organizations must be able to quickly locate particular data in their archives. In addition, many organizations are interested in leveraging the information contained in documents and other types of semi- and unstructured data. These new requirements led to the creation of a new type of electronic archiving solution: the active archive.

In an active archive, data is not stored offline on tape, but instead is kept online on disk arrays and is usually indexed for rapid searching. This allows organizations to rapidly find and retrieve data in response to eDiscovery requests, customer inquiries, new business initiatives, and so on. An active archive solution should also have mechanisms to ensure compliance with regulations that govern content retention, authenticity, security, and so on.<sup>1</sup>

EMC was an early entrant into the active archiving market, releasing a content addressed storage system called Centera in 2002. Because of EMC's first-mover advantage, Centera became the market leader in terms of installed base, but IBM, HP, NetApp, and several smaller companies have developed competing products.

One such company was a vendor named Archivas. Hitachi Data Systems entered into an OEM agreement to use the Archivas software as the foundation for a new active archiving solution, the Hitachi Content Archive Platform, which was released in June 2006. The initial release of the Hitachi Content Archive Platform was praised by Data Mobility Group and other analysts for its forward-thinking architecture and openness, and HDS signaled its commitment to becoming a leader in the active archiving market when it announced its intention to acquire Archivas outright. That acquisition closed in February 2007, and in June 2007 HDS unveiled a new version of the Hitachi Content Archive Platform with a number of significant enhancements and new features.

<sup>1</sup> Active archiving is sometimes called fixed-content archiving because archived data usually consists of objects, like emails or medical diagnostic images, whose content does not (and should not) change after they are created. Active archiving is also sometimes called content addressed storage, or CAS, because some active archiving solutions compute a unique hash code for every archived object and then later use that hash code to locate objects within the archive (i.e., the hash codes are used as the addresses of archived content). However, not all active archiving systems use content addressed storage. The Hitachi Content Archive Platform, for example, is an active archive solution for fixed content, but it is not a content addressed storage system.



## Hitachi's Content Archive Platform

The Hitachi Content Archive Platform (HCAP) is designed to serve as an enterprise-wide repository for fixed content. HCAP manages each content asset as an object comprised of three parts: the content itself (for example, an image or a Microsoft Word document); metadata that describes the content; the policies that should govern the content (for example, a retention policy directs HCAP to ensure that the content is retained for a specified length of time). All objects exist in a single global namespace. Administrators can carve up the namespace in whatever way works best for their organization, but users and developers can still search for objects across the entire archive if they wish to, a feature that is important for eDiscovery, compliance, and governance, as well as other applications.

HDS sells HCAP in two basic configurations. One bundles the archiving software together with HDS' WMS100 storage in a self-contained, fully-integrated appliance. HDS offers a range of initial storage capacities—the entry level model has 5TB—and customers can essentially just wheel it into the data center and plug it in to get started. The second configuration is a diskless appliance called HCAP- DL (the DL stands for “disk-less”) that can use the whole range of HDS storage platforms, including the Universal Storage Platform V, Network Storage Controller, Adaptable Modular Storage (the AMS200/500/1000 arrays), and Workgroup Modular Storage. If deployed with USP V, the HCAP-DL configuration can also utilize third party storage that's been virtualized by the USP V.

HCAP provides a rich set of services for managing and protecting archived content. Some of the features and capabilities of HCAP include:

**Content authenticity and immutability.** Many regulations require archived data to be stored in a demonstrably immutable form. HCAP provides true “write once, read many” (WORM) storage that meets these requirements. HCAP computes a hash code (sometimes called a digital signature or fingerprint) for every object and this code is later used to verify objects' integrity and authenticity. Administrators can choose from several different hashing algorithms, including: MD5, SHA1, SHA256, SHA384, SHA512, or RIPEMD-160.

**Retention and destruction.** Retention periods can be set on a per object basis, ensuring that no item is deleted prematurely. In addition, ad hoc retention holds can be placed on individual objects or groups of objects. When it comes time to delete an object, HCAP can securely shred the content file such that no trace of it will be recoverable.

**Custom metadata.** Individual objects can be annotated with whatever additional information developers deem useful; this capability can greatly facilitate eDiscovery and other applications.

**Indexing and searching.** HCAP supports the automatic ingestion of content metadata as data is loaded into the archive. In addition, HCAP builds a full-text index of ingested content. A familiar web-based, Google-like, interface enables users to run searches, based on objects' metadata and the full-text index, across the whole archive, and developers can execute searches using HCAP's HTTP-based API. Search results can be exported, and a group of found objects can be placed on retention hold (e.g., for litigation) with one click. HCAP can index 370 file formats and 77 languages.

**Elimination of duplicate files to reduce storage consumption.** HCAP can identify and delete duplicate content. HCAP decouples objects' names and metadata descriptions from the objects' content; multiple objects, each with a different name and different metadata, can share the same underlying content. HCAP uses a two-step process to detect duplicates. First, it identifies potential duplicates by finding objects with the same content hash code as other objects. It then does a binary comparison to determine if potential duplicates are in fact exactly the same.

**Content encryption with embedded key management.** HCAP provides an innovative (patent-pending, in fact) encryption mechanism based on a "secret sharing" approach. HCAP encrypts content, metadata, and search indexes, and stores the encryption key internally, but it splits the key up and stores the pieces on several different nodes. This prevents data from being recovered from a stolen disk or server, because no single disk or server has the full key needed to decrypt the data; the data can only be decrypted in a fully functioning HCAP cluster with all the key-sharing nodes attached. The fact that key management is handled internally within HCAP should appeal to customers who are justifiably wary of having their archive dependent on an external key management system.

**Data protection.** To ensure that data can survive multiple device failures, HCAP can be configured to store two, three, or four copies of an object. The creation and maintenance of these copies is independent from the RAID protection offered by the underlying storage arrays. Taken together, the multiple copies and the RAID protection should ensure that objects will never be permanently lost, albeit at the cost of extra storage capacity.

**Object replication across multiple sites.** For still more data protection, HCAP supports object-level replication from one archive to another. Actually, HCAP supports bi-directional, one-to-many, and many-to-one replication between multiple archives. Full objects—content, metadata, and policies—are replicated,



and the data can be compressed and encrypted for faster and safer transport across the WAN. Object replication allows an organization to keep synchronized archives in multiple sites for disaster recovery and business continuity.

In addition to the above features and capabilities, two other noteworthy aspects of the Hitachi Content Archive Platform deserve to be highlighted: the innovativeness of the architecture and the openness of the platform.

### **An innovative architecture, designed for extreme scalability and reliability**

One of the strengths of the Hitachi Content Archive Platform is its unique architecture, which Hitachi has dubbed a SAIN, or “SAN-attached Array of Independent Nodes.” HCAP is a self-configuring, self-healing clustered system. A cluster is comprised of archive nodes, search nodes, and storage. Archive nodes run the software that performs core archiving services, while the search nodes maintain the full-text index and process queries. New archive nodes, search nodes, or storage can be added at any time and HCAP will adjust its load balancing to incorporate them. If a node goes offline or develops some other kind of problem, HCAP is able to continue operations using its fully functioning nodes while the failed node is repaired or replaced. With Version 2.0, HCAP has implemented a new node selection (i.e., data writing) strategy that optimizes data protection and minimizes the impact of a double node failure (the latter is meant to appeal to customers concerned about the fact that EMC’s Centera can lose data in the event of a double node failure). Given the underlying SAN storage’s RAID protection, neither node failures nor disk failures will cause the HCAP to become unavailable to users or result in permanent data loss.

HCAP’s high availability is matched by its high scalability. A single node can support up to 400 million objects and 256 TB of storage. In an 80-node configuration, the HCAP-DL deployed with the USP V can scale to 32 billion objects and 20 petabytes of data. Performance can scale near-linearly by adding archive, search, or storage nodes as needed.

Moreover, because the SAIN architecture decouples the front-end processing nodes from the back-end SAN storage, HCAP customers can scale their archive in a flexible manner. For example, if a customer has a relatively small number of objects, yet the content of the objects tends to be quite large (tens of megabytes each, say), that customer might need only a few archive nodes to go along with massive storage capacity. Conversely, a customer might have a very large number of objects, but the content of the objects could be small (a few kilobytes each, say); that customer would need more archiving nodes, but would perhaps need less back-end storage capacity. The SAIN architecture enables customers to deploy an optimum ratio of front-end archiving nodes to back-end storage capacity for their particular needs.

The decoupling of the front-end processing nodes from the back-end storage also gives HCAP customers flexibility when it comes to upgrading. As more powerful or more energy efficient servers come along, customers can upgrade their archiving and search nodes independently of their storage. Likewise, customers can migrate data to newer and better storage devices independently of their archiving and search nodes. A digital archive must be able to preserve content for decades, and over that length of time storage technology will obviously continue to evolve and need to be refreshed several times, so this is an important architectural feature.

## **An open platform**

Hitachi would like customers to use HCAP as their primary enterprise-wide content repository, and for that to happen HCAP must be easy to integrate with a multitude of different applications and with existing IT infrastructures. In addition, Hitachi's strategy for HCAP places a high priority on enticing third-party software vendors to integrate their products with HCAP, and there again ease of integration is key.

Accordingly, Hitachi designed HCAP to be easily accessible using a variety of standard protocols. The default access protocol, smartly enough, is HTTP. Any client application that can be made to talk HTTP—and that more or less means all client applications—can work with HCAP. Alternatively, applications can use WebDAV, another standard protocol (codified in RFC 2518) layered on top of HTTP. HCAP can also be accessed natively as an NFS or CIFS/SMB file share, allowing users and applications to work with archived objects just as if they were files on an ordinary file server. Client applications can even submit objects to be archived via email, using the standard SMTP protocol. HCAP also supports other standard IT protocols, such as NDMP, for integrating with backup products, and SNMP, for integrating with IT management systems, and supports the SNIA standards SMI-S and XAM<sup>2</sup>.

Another important element of HCAP's open design is that content files retain their original file names and remain in their original file formats after they are ingested into HCAP. This makes it easier to migrate content to other systems or other formats if the need arises—and over a sufficiently long period of time, it is a near certainty that such a need will arise. The fact that content files are stored with their original names and formats intact, in conjunction with the ability to work with archived content using an HTTP-based API and the ability to browse through archived content using NFS/CIFS, means that HCAP is a highly “future-proof” solution—that is, customers will likely never have any difficulty accessing the content they have stored in HCAP.

<sup>2</sup> XAM is still a work in progress, but HDS has been involved in the XAM development effort and has pledged to support it once it's finalized.



Because HCAP has been designed with an emphasis on openness, archived objects can be accessed by many different applications and used for many different purposes, including business analytics, knowledge management, etcetera. Both in-house developers as well as commercial software vendors will find it to be an easy and straightforward process to use HCAP as the content repository for a wide variety of applications.

## Final Thoughts

Market demand for active archiving solutions is poised to take off. Based on the genuine strengths of the Hitachi Content Archive Platform, Hitachi deserves to seize a large share of that market. Enterprises today confront a number of information management challenges that call for the use of an active archive. Many organizations have an acute need for a solution such as the Hitachi Content Archive Platform. Because HCAP has been designed to be an open, flexible, and long-term platform, organizations that purchase HCAP to address urgent tactical problems will find themselves in the fortunate position of owning a system that can support their long-term strategic requirements as well.

Hitachi is positioning HCAP in an interesting way. Hitachi is not attempting to provide a complete records management system, or content management system, or information classification solution, or email archiving solution, etcetera. Instead, Hitachi is focused on providing the core archiving and storage services that all such systems require. Those core HCAP services also happen to be required by innumerable other applications, and also happen to be services that are essential for any compliance, governance, or eDiscovery initiative. Somewhat paradoxically, because Hitachi has chosen to limit the functional domain of HCAP, it has greatly increased the range of solutions and environments in which HCAP can play a useful role. Customers will find HCAP useful now for easing immediate pain-points *and* will find HCAP useful in the future for purposes as yet unimagined, for applications not yet even in existence. Similarly, software vendors and other partners have every reason to build solutions that use HCAP to provide secure, reliable, scalable, compliant, and open content archiving services.

Hitachi has a clear strategic vision for HCAP, and HCAP itself is solid and innovative technology. Any organization that is considering the purchase of an active archiving solution should be sure to evaluate the Hitachi Content Archive Platform. 