Active Archive
A Blueprint for Long-term Preservation of Business-critical Digital Data
All around the world, data retention regulations are becoming more stringent, and penalties for non-compliance can damage an organization’s balance sheet. For example, the U.S. Securities and Exchange Commission (SEC) has found in recent years that several financial services firms were in violation of the content preservation requirements of SEC Rule 17a-4. For failing to preserve e-mail communications and other business documents, the companies were fined a total of US$8.25 million.

As another example, the United States District Court in Baltimore granted a government discovery motion that required Microsoft to produce old e-mail messages in an antitrust action against the company. Although the cost and inconvenience of searching 25,000 e-mail backup tapes was substantial, Microsoft had to bear the burden, knowing that, if the e-mail messages were not produced, the judge could instruct the jury to make assumptions unfavorable to the company. In addition to e-mail messages, many other types of digital documents are now receiving significant attention from litigators, government and industry regulators, and, through necessity, corporate executives. Maintaining these fixed-content documents in their original form, and providing quick retrieval when needed, demands a new approach to long-term data storage.

To satisfy electronic discovery and other regulatory data preservation requirements, both structured and unstructured digital content...
must be protected from harm and safeguarded against unauthorized access. This type of authenticated content preservation requires a combination of hardware, software, and enforced business processes to guarantee that digital content is securely preserved and readily available.

**Long-term Data Preservation**

Most traditional types of data are transaction based, and have a relatively short period of usefulness. The ones and zeroes that record the rise and fall of a stock price or the value of goods in a digital database are subject to constant modification by applications, and/or deletion at the click of a mouse. But not all data can be allowed to change or disappear so easily. For many years now, certain types of organizations have been required to preserve transaction logs for “dynamically changing digital data” for much longer than the data is useful to the business.

Only in the past five years have regulations emerged that also require the long-term preservation of “fixed-content digital data,” such as Web pages, e-mail messages, corporate presentations, medical images, presentations, spreadsheets, digital audio and video, and check images that don’t change over time. As a general trend, this fixed-content data is increasingly being required to be preserved over long periods of time.

**Fixed-content Data**

Fixed-content data is best defined as data that is not intended to change. Fixed-content data includes digital records of real-world events that have happened at specific points in time. Examples would include an X-ray image of a broken arm, an e-mail message, a completed digital video, or a filing made to a government agency. For the fixed-content data to remain valuable in the future, it must remain fixed, to accurately reflect its original state. For most organizations, fixed-content data comprises the bulk of all storage needs. Although analyst estimates vary, fixed-content data is generally thought to represent between 80 percent and 90 percent of all storage capacity. It is also generally accepted that fixed-content data is being created much faster than dynamically changing transaction data. Thus, fixed-content digital data growth is driving demand for long-term archive storage capacity (see Figure 1).

**Why Retain Fixed-content Data?**

Prior to the advent of digital recording media, physical fixed-content data was indexed, cataloged, and filed in formats such as paper, microfiche, video, audio, or film, to provide a permanent archive that could be accessed when needed. However, certain types of digital fixed-content files, such as e-mail and instant messages, for example, may not have a physical counterpart since many e-mail and instant messages are never printed and are simply left on digital recording media. Since government regulations cover fixed-content digital data such as e-mail and instant messages, and since these types of digital data have no physical counterpart, maintaining a permanent record of digital fixed-content data
is becoming a priority for organizations seeking to achieve regulatory compliance.

The list of reasons for retaining fixed-content digital data grows continually. This data provides a historical record not only for the organization, but also for the auditors and regulators who enforce compliance with data retention statutes, and work with their clients to ensure best archiving practices for a wide range of industries. Organizations ignoring these requirements can face stiff penalties and legal consequences. E-mail is now considered a standard source of evidence in legal proceedings. Searches through e-mail archives are a routine first step in any legal discovery process. Organizations that do not have ready access to historical e-mail messages can not only be handicapped during litigation, but also saddled with the disruption and expense of manually searching through tape backups to satisfy discovery. Sometimes discovery costs and their disruption to operations can effectively force a business to settle a dispute that they could easily have won with timely fixed-content data production.

The Challenge: Disk-based "Active Archiving" of Fixed Content

For some industries the growing demand for disk-based “active archive” digital data storage is forcing IT to rethink conventional storage architectures, since storing digital archival data on top-of-the-line storage systems is usually not cost-effective.

Not all of today’s enterprise architectures were designed with long-term digital data archives in mind. Information on conventional disk-storage systems is usually considered working data, and often is characteristically archived to very inexpensive media such as tape or optical media as part of routine IT operations. Unfortunately, tape and optical media do not always satisfy the requirements of today’s regulatory environment, since tape and optical media are often manpower intensive to access and slow to search.

Storing archived fixed-content information in online SATA disk-based “active archive” storage systems offers an alternative to traditional offline archival storage using tape, microfiche, optical disk, or paper as archival media. An online disk-based “active archive” can not only improve customer service responsiveness by making infrequently needed data available faster in response to a customer question, but it can also improve online relational database performance by taking the infrequently needed data out of the relational database. An “active archive” can also reduce the disruptive (business and IT) impact of all types of auditing (including Sarbanes-Oxley) and litigation-based discovery.

What Would a Modern Disk-based “Active Archive” Look Like?

Rather than reinvent the wheel, IT planners can refer to substantial existing research on the records management process for archiving, whether the archived digital data is to be stored in a passive archive on tape or optical media, or in an “active archive” on disk. In the physical world, the long-term preservation of important records and artifacts is the domain of library sciences and archiving. Practitioners in these disciplines have given much thought to the problem of data preservation and have developed several general theories and best practices. Traditional archives provide a repository for records the organization has selected to preserve. The archive serves two fundamental goals: records must be preserved unchanged, and records must be easily accessed. These properties apply equally well to archives containing fixed content. Of course the choice of digital media selection for the archive will determine the ease of archival data access. These questions can only be answered on an industry-by-industry and enterprise-by-enterprise basis.

Ingesting Information to the Archive

Fixed-content digital data are created by a wide variety of enterprise applications. The digital archive ingestion process allows multiple applications to stream data into the archive simultaneously for long-term preservation and storage. Individual items become archival objects in a metadata software repository for each fixed-content record. In this way, the metadata can be quickly searched to retrieve the archived items that are needed for customer service, audit, or compliance purposes. The ingest process also associates a customer-defined retention period with the content. Retention periods ensure that content cannot be changed or deleted from the archive until a predetermined period of time has passed. In a traditional archive, the ingestion of records into the archive represents
The Hitachi Content Archive Platform supports a common set of archival functionality outlined in the ISO Reference Model for Open Archival Information Systems (OAIS). The OAIS model has been used as a foundation for some of the largest archives ever designed, including those managed by the U.S. National Archive and Records Administration and the gargantuan Planetary Data System Archive managed by the National Aeronautics and Space Administration (NASA).

The OAIS model is also extensible, allowing the specialized needs of individual implementations to be satisfied. The Hitachi Data Systems implementation of the OAIS model defines the processes for managing an "active archive" that supports the information lifecycle (see Figure 2).

Preservation Planning
During preservation planning, policies that control the management of digital data within the archive environment are defined. For example, the value and policies associated with a piece of information determine whether it is maintained on primary disk and then migrated to second- or third-tier storage.

Producer
A producer of data assets is any entity that submits data to the archive. Examples include applications, e-mail servers, and medical imaging systems.

Ingest
The ingestion mechanism indexes relevant metadata from the unstructured data and establishes authenticity by generating a unique identifier or signature, which is sometimes referred to as a hash.

Data Management
Metadata associated with the archived content is maintained in a data management system. The system allows users to manage policies associated with the archived content and to query and retrieve data using searchable indexes.

Archival Storage
The original data is maintained in the storage infrastructure of the archive. This can mean storing data on online disk, nearline disk, tape, or any combination of these media.

Consumer
The OAIS model defines the consumer as an end user or an application capable of searching the index and requesting archived data for retrieval.

Added Value
The ability to repurpose and reuse archived content in new and innovative applications can enhance the value of the content archive.
and encryption routines without jeopardizing the chain of custody of fixed-content data. This also means that before-and-after copies of each data transformation must be kept as an audit trail, so that subsequent researchers can determine if any unauthorized changes took place during conversion.

Search and Retrieval
The ability to successfully search and retrieve stored records is a key function of an archive. If records cannot be located they are effectively lost, and all attempts to preserve the content are wasted. A digital archive is almost certain to contain many millions of records, and, therefore, efficient, high-performance search and retrieval of information is critical.

Whether an archive is conventional or digital, the curators and managers of the information face a common problem: it is almost impossible to anticipate how future users of the archive will want to query the information. Some users will know exactly what information they are looking for and will be able to use specific indexes to locate items. Other users will need to browse the archive looking for relevant information. Digital archives can also support data-mining technology, allowing researchers to gain insights into archived content unavailable through other search and retrieval mechanisms.

Metadata provides a source for indexes that allow the flexible search for content in a digital fixed-content archive. Support for multiple access protocols also allows different applications to act as a front end to the search and retrieval process. The use of open protocols guarantees that records in the archive will always be accessible, regardless of whether search and retrieval technology changes.

Tiered Storage and Archiving
The recent escalation in demand for enterprise storage capacity has led many IT organizations to consolidate and centralize storage resources. Storage area network (SAN) and network attached storage (NAS) technologies have produced significant efficiency improvements, allowing administrators to manage much more capacity using a common set of storage management tools and procedures. In these new consolidated environments, IT planners favor storage solutions that integrate seamlessly with the existing tiered storage infrastructures and with existing Enterprise Content Management (ECM), Business Process Management (BPM) and compliance software platforms. Archival solutions that require a unique management and configuration approach lower administrator productivity, reduce efficiency, and raise the overall cost of storage.

Considering a digital archive’s potential to consume enormous amounts of storage capacity, it is imperative that the archive integrates seamlessly into an existing IT infrastructure. In this way, IT management can ensure that the archive remains cost-effective as it scales up for growth. Support for storage networking and common storage management tools will allow an archive to be treated as simply another tier in a multitiered pool of efficiently managed storage.

For an optimal enterprise architecture, a tiered storage approach to archival data can be as important to consider as a tiered storage approach to online operational data. Disk-based fixed-content digital data archiving and tape and/or optical fixed-content digital data archiving, therefore, are examples of various archival storage tiers. The data to be placed on each are determined by what makes the most financial sense for the business.

Hitachi Content Archive Platform
The Hitachi Content Archive Platform is a robust “active archive” that enables effective long-term, fixed-content data preservation for businesses of all sizes. Designed to seamlessly integrate into an existing enterprise storage infrastructure, the high-performance, high-availability, highly scalable archiving solution satisfies an organization’s regulatory compliance requirement by ensuring the secure, long-term preservation and fast search and retrieval of valuable business records (see Figure 3).

A first in SAN-based digital archival solutions, the Hitachi Content Archive Platform uses world-class Hitachi storage systems to provide scalability, availability, and performance, satisfying the growing demand for long-term fixed-content storage management. With built-in authentication, protection, and retention capabilities, the highly available platform guarantees archived content will be continually available for access for years to come.

The Hitachi Content Archive Platform provides:

- A WORM file system and time-base retention at the object level
- Authenticated content preservation with a user choice of digital signature or hash algorithms
- Embedded full-text index, search, and retrieval for content discovery
Automated object-level remote replication
Custom metadata support
Content privacy (encryption of data at rest)
Standards-based interfaces, including NFS, CIFS/SMB, HTTP, https, WebDAV, SMTP, NDMP
Duplicate data elimination
Support for Hitachi Universal Storage Platform™ V, Hitachi Universal Storage Platform, Hitachi Network Storage Controller™, Hitachi Adaptable Modular Storage, and Hitachi Workgroup Modular Storage
Support for Hitachi Device Manager, Hitachi Tiered Storage Manager, and Hitachi Tuning Manager software for reporting of Content Archive Platform metrics

Built around the Hitachi SAIN (SAN-attached Array of Independent Nodes) architecture, the Content Archive Platform’s architecture dramatically reduces the burden of managing archival fixed-content storage. The solution leverages both existing storage management and business continuity processes from Hitachi Data Systems. Using management software common to the entire storage environment lessens the training burden, cuts deployment time, streamlines information flow, and improves administrator productivity. The Content Archive Platform supports a wide variety industry-standard document and file formats, to enable archiving and future retrieval of all types of fixed-content digital data objects and associated metadata, whether the data is structured, semi-structured, or unstructured.

The Hitachi Content Platform delivers clear business benefits:
- Regulatory compliance
- Standards-based architectural model
- Simplified data storage and retrieval
- Integrated archive management within the storage environment

Services Oriented Storage Solutions Align Business and IT Objectives
Services Oriented Storage Solutions from Hitachi Data Systems provide an integrated approach to developing a storage infrastructure based on business requirements rather than technology features. The storage services can be applied to data as-needed in a multitier, heterogeneous storage infrastructure supported by application, content, data, and storage services. Services Oriented Storage Solutions allow the storage infrastructure to respond to the specific performance, availability, functionality, and cost requirements of each application (see Figure 4). By delivering capacity and services based on the specific needs of the business, Services Oriented Storage Solutions can substantially lower both operating expense (OpEx) and capital expense (CapEx) for the enterprise.

Since Services Oriented Storage Solutions can be applied on an “as-needed” basis to various storage tiers, the Content Archive Platform can seamlessly be integrated into almost any existing tiered storage infrastructure or enterprise software architecture.

Implementing the Hitachi Content Archive Platform
The Hitachi Data Systems Global Solution Services (GSS) team can help organizations of all sizes design and implement the optimal Content Archive Platform configuration to meet customer specific business and application requirements.

Basic Implementation Service
- Provides for the installation and configuration of the Content Archive Platform
- Includes all necessary configuration activities needed to incorporate the Content Archive Platform into the customer’s environment
- Presents the Content Archive Platform solution to the various archive applications
- Tests ingestion of archive data from the applications utilizing the Content Archive Platform

Cell Capacity Upgrade Service
- Enables the installation and configuration of capacity upgrade cells to an existing Content Archive Platform
- Provides for all activities necessary to implement this upgrade and make it available to the Content Archive Platform

Search Capacity Upgrade Service
- Enables the installation and configuration of Search upgrade cells to an existing Content Archive Platform
- Provides for all activities necessary to implement this upgrade and make it available to the Content Archive Platform

Figure 4. Each Services Oriented Storage Solution is built on a framework for aligning business and IT objectives.