

WHITE PAPER

Active Archiving: Hitachi Content Archive Platform

Sponsored by: Hitachi Data Systems

Laura DuBois

June 2006

EXECUTIVE SUMMARY

Content or information runs today's businesses. Content describes information such as text, sound, or images that resides in different types of structured and unstructured data. Business-relevant content can exist in unstructured data, such as invoices, contracts, email, instant messages, spreadsheets, and Web pages, as well as structured sources, such as applications and databases. Content, regardless of its structure, can have significant legal, compliance, or business relevance to a firm. These dynamics are driving a convergence of content and information technology.

Unique, active archiving solutions are emerging to dynamically manage content across heterogeneous application sources according to business and technology policies. Active archiving solutions support policy-based ingestion of content from multiple sources to a fast, online disk-based content archive platform for the purposes of secure retention, preservation, retrieval, or verifiable destruction. Older archive approaches, either paper-based or electronic, acted as physically separate file cabinets. They did not provide centralized search, policy-based retention, authentication or protection, or adequate levels of accessibility, performance, or scalability across multiple content sources. New active archive solutions address these historical problems and provide time-critical access to this content when required in response to discovery, audit, or business need.

CONVERGENCE OF CONTENT AND TECHNOLOGY

Business content, such as account information, patient medical images, or legal correspondence, can reside in formats that are deemed critical electronic records by a firm. Electronic records serve as digital evidence of a firm's activities, events, or transactions and are managed and retained due to their operational, business, legal, regulatory, or historical value to a firm. While once the management of records, largely paper-based, was the exclusive purview of records managers, today a firm's electronic records and their associated content are under the supervision of both business and technology officers. The effective management of content, and thus electronic records, is required in a competitive business environment, and the challenge for a firm is in locating, organizing, storing, and disposing of content based on both the information technology and business policies of a firm.

Below are some examples of specific industries with business content contained in electronic records:

- ☒ **Life sciences/healthcare.** A life sciences firm maintains new drug applications, which include Serious Adverse Events (SAE), patient diaries and clinical trial information, in which to document drug ingredients, results from clinical trials, and processes for the drug manufacturing, processing, and packaging. Healthcare providers, payers, and business associates need to maintain electronic patient health information (ePHI).
- ☒ **Transportation.** An airline or air freight firm retains service and maintenance records, including maintenance manuals, for aircraft it operates to track service history of the aircraft and its repairs over the life of the asset. An automotive rental or leasing company maintains lease, rental, and insurance contracts as well as copies of a client's driver's license to document the legal terms of a client agreement and maintain customer information.
- ☒ **Entertainment.** A television network preserves its digital video and photographic content and tracks these assets and their associated contractual ownership in the metadata for the digital asset. Movie studios need to track and maintain digital content at various stages of production.
- ☒ **Financial services.** Brokerage firms maintain electronic correspondence between brokers and clients to document broker-client communications and/or any trade transactions. Insurance firms retain images of insurance claims, enrollment forms, policy applications, policy claims, and payment checks; banking institutions retain account statements in order to retain history on customer accounts and correspondence.
- ☒ **Manufacturing.** Automotive as well as consumer product manufacturers must retain product recall case records, product non-conformity records, and manufacturing change forms for quality assurance and ISO certification.

Business forces require that the right content be available, when required, by customers, partners, employees, auditors, attorneys, shareholders or internal business units. The value of content and electronic records that have a guarantee of security with accessibility as well as integrity and reliability is paramount to a business's livelihood. If the integrity of the content is compromised, if privacy or security policies are violated, or if content is not accessible within the required window — the result can include damage to company reputation, loss of revenue, stock price declines, personal executive liability, court sanctions, or claims of spoliation of electronic evidence.

As a result, there is an unparalleled convergence between content, data, and archive services that are being instantiated in new active archiving solutions. This convergence is organizationally aligning IT professionals with the business through the formation of compliance and ediscovery liaisons that bridge the worlds of technology, legal, records management, compliance, and business units. These liaisons understand not only a firm's business policies required to satisfy compliance and legal requirements but also the complexity, limitations, and locations of different file structures across a firm's IT environment.

Increased Competitive Advantage

Historically, through data reuse and business analytics, organizations sought to capitalize on the value of electronic information resident in structured databases for revenue generation, customer satisfaction, and strategic advantage. However, increased levels of advantage can be realized in the reuse of unstructured data. As firms need to comply with regulations that stipulate the retention of content contained in unstructured repositories for specific periods of time — new usage patterns for that data emerge that can yield a competitive advantage.

For example, at a credit card processing firm, historical customer account information is retained. The disk-based archive and retention of these records can be done in a manner that provides fast access in minutes versus days. Since the records are accessible they can be made available through information portals to internal and external customers for self-service use. The provision of these records in an online and self-service manner saves on employee costs and allows the credit card processing firm to charge for portal services. These extra services create additional opportunities for revenue generation and higher customer satisfaction. This manner of data reuse is difficult to achieve with older, traditional archive technologies because the information has not been indexed, is not easily accessible, or, if stored on removable media, runs the risk of being lost.

TODAY'S BUSINESS CHALLENGES

The convergence of content with technology is being driven by increased regulatory compliance, electronic discovery, and content security within a firm. The risks due to the compromise of business content can include regulatory fines, court sanctions, rulings of inverse inferences, poor public publicity, corporate brand dilution, declines in stock valuation, and the like. These significant pressures are causing firms to increase their attention to the management of content and corresponding electronic records.

Regulatory Compliance

Long-standing federal, state, and local regulations stipulate how companies manage and retain content found in records. However, more information is created electronically today in different systems and formats, and regulatory bodies, such as the SEC, DHHS, EPA, FDA and FAA, are more strictly monitoring and enforcing these regulations.

These regulatory mandates subject companies to requirements that specific business content is captured and retained in a way that ensures information integrity, security, and accessibility. This legislation is concerned with the content and not the format in which it is stored. It is the firm's responsibility to identify the file formats, applications, or media that contain this regulated content and ensure that the corresponding electronic records are managed according to requirements. The problem is compounded as most firms must comply with a complex matrix of tens to hundreds of regulations that impact content found in different electronic records. Moreover,

frequently the same electronic records are subject to multiple requirements and retention requirements.

Regulations may stipulate that related content contained in electronic records be retained for specified time periods or outline rules to secure the privacy, security, and lack of compromise to sensitive information. Central to satisfying regulatory compliance requirements is the fast provisioning of electronic records in response to a regulatory audit, typically within a 24- to 48-hour time period. As a result, electronic records subject to regulatory compliance should be indexed for easy search, retrieval, and production. Equally as important as retention is the proper disposition and destruction of electronic records, absent any legal requirements to preserve, once they no longer have any regulatory or business value.

Litigation and Electronic Discovery

The discovery of electronic information that may be relevant to or act as evidence during corporate litigation is increasing. The largest Fortune 500 firms as well as firms in highly litigious industries, such as manufacturing, consumer products, construction, medical supply, healthcare, and pharmaceutical, face frequent litigation. In 2000 alone, Wal-Mart was sued 4,851 times — or nearly once every two hours, every day of the year, and it is not uncommon for large institutions to face hundreds to thousands of legal matters at any one time.

Firms have a duty to preserve electronic information for the purposes of discovery before litigation starts, at the point when litigation can be reasonably anticipated. Prior case law, such as *Zubalake v. UBS Warburg LLC*, in combination with pending December 2006 amendments to the U.S. Federal Rules of Civil Procedure is applying more definition to what is considered reasonable and what constitutes anticipated. Once litigation is reasonably anticipated, firms have a "duty to preserve" electronic records, and current records retention and rotation policies must be halted. State and federal court rulings have become more stringent regarding corporate information spoliation, which can be called if proved that documents were not preserved once litigation was anticipated.

During the discovery phase, the challenge of collecting the data begins to mount as IT must locate and/or restore data in response to discovery. The collection effort can be very complex and time consuming and must be done in a manner that preserves the chain of custody. The time and cost of retrieving emails and other documents with traditional archive technologies can mean significant time and cost to the IT department as well as lost productivity. However, just producing the information is only a piece of the cost. Even more costly than physical retrieval is the legal review and production process, and courts have fined companies millions of dollars for not producing information in a timely manner.

Mitigate Content Vulnerability

Driven by regulatory compliance and sound corporate policies and practices, firms increasingly need to mitigate the risks of sensitive information leaks of content within unstructured data. To alleviate potential exposure of sensitive information, the discovery, indexing, and documentation of content within distributed volumes of networked file data are required. Once content within the data is understood, the risk of exposure can be mitigated by the proper tagging of data based on corporate classification, role-based access, or, in some cases, data can go through disposition.

Firms are placing an increased focus on the discovery and indexing of unstructured data and the content-based classification of data in order to understand any content exposures and then to apply any information management policies, such as retention, access restriction, or disposition. Once discovery, indexing, and classification are done, search and retrieval of unstructured network files can be completed to optimize their security, storage, and disposition.

Consider the risks for a firm in the exposure and visibility to sensitive information, such as company information (financials, intellectual property, engineering data, sales data, business strategy) and customer information (name, social security numbers, account numbers, patient health information, medical history, or credit history), by the wrong people. Content-aware discovery and classification can help a firm address the challenges with litigation, regulatory compliance, and optimization of storage capacity, while it helps to protect sensitive information from unwanted leaks or exposures.

Continued Data Growth

In addition to the business problems associated with unstructured data, firms must manage their explosive growth, and effective classification techniques can aid in the archiving of data to lower-cost tiers of storage and reduce the costs associated with growing unstructured data sets. Storage capacity is growing at a rate of between 50–100% year over year. A central problem for large firms is the sheer growth of primary data, while the number of IT resources to manage that increasing capacity remains flat. The drivers for this corporate data growth are sources of new or newly managed content, which is coming from the following:

- ☒ Business use of RFID to capture data and store more information about products. This information is used by retailers and manufacturers and enables real-time supply chain management.
- ☒ The continued growth of mobile devices and increased levels of intelligence and content in remote devices such as automobiles, cell phones, and PDAs will require increasing storage capacities.
- ☒ The convergence from a predominantly paper-based to digital form of working in specific industries such as healthcare and government. The science of records management, used for years to store and delete paper records, has advanced into the electronic world for paper-based industries.

- ☒ Use of new business communications such as the use of the instant messaging and Internet to conduct internal and external rich media Webinars and the continued proliferation of online, social network communities.
- ☒ Insatiable desire for information via the Internet and quickly advancing tools to locate, filter, and/or present information in specific formats, including across an increased number of languages.

The implication of this data growth on the ability to archive increasing volumes of content and electronic records is significant. For example, as the number of emails being archived by a firm continues to grow between 70–100% annually, archive solutions must be able to keep up with the capacity in ingestion of the email files into the active archive. An active archive solution that can scale at pace with the primary content growth can also help with reduction of cost and primary storage space in the production email server. The same is true in the archive of database content, which entails moving less frequently accessed database records or tables to an archive database, while still preserving the referential integrity of the data. In moving out less used data, the performance gains on the production database can be significant.

ACTIVE ARCHIVING OF CONTENT

Active archiving of content is a requirement in today's business and technology environment. An active archiving solution provides the ingestion of file and metadata from different content sources for the purpose of secure, integrity, and authenticity-based retention, preservation, and disposition of electronic records. An active archive solution makes use of a metadata index of content attributes in which to support information management policies and provides for horizontal policy-based management, search, and retrieval across many content repositories. These information management policies take into consideration both the legal, regulatory, and business value of the content and the characteristics of the technology on which the content resides.

ACTIVE ARCHIVING CONSIDERATIONS

Multiple Content Sources

A firm should deploy an active archiving solution that is able to support the archive of content from different applications, both commercial and homegrown systems, and structured and unstructured data into a single active archive architecture. It is not uncommon for a large firm to have hundreds of different application and file system repositories that must be actively archived for compliance, legal, or business purposes. These application repositories can include multiple different content sources, such as content or document management, email or instant messaging, document or medical imaging, databases or homegrown applications, and distributed file systems. The operating and capital cost savings in management, training, and technology in leveraging a common active archive solution for multiple content sources can be profound and allows for centralized search/retrieval and data security, authentication, and integrity.

Support For Business Policies

The single, most important criterion for legal council, compliance officers, risk managers, records officers, or business units to consider in an active archive solution is their support for a firm's business policies. While research shows that these business policies are best established by those that know and understand the business — the actual planning and implementation of the policies is a complex process involving participation by a group of functions within an organization. Business policies are being developed in a consortium of business, technology, and records offices, and increasingly these business policies are a consolidation of legal requirements, compliance requirements, and overall corporate policies about how a firm wants to manage its information assets. A sample of some questions related to a firm's business policies include:

- ☒ Which data sources, based on their content, constitute official electronic records for a firm? Why is the data considered an electronic record? What regulations is the content subject to? Are certain electronic records subject to multiple regulatory requirements? Who is responsible for mapping a firm's internal systems to its records, and how are they managed and retained?
- ☒ How long does a record type need to be retained? Is a "second copy" of an electronic record required? How or in what format and with what level of authenticity, integrity, or permanence do the records need to be retained? What level of accessibility and response time is required for the records in response to audit or discovery?
- ☒ With what categorization or taxonomy should corporate data be classified? Should a published, standards-based taxonomy be used? Is a simple file system metadata classification satisfactory, or is content-based keyword, expression, or concept-based classification required? How are electronic records searched and delivered as part of a legal discovery or regulatory audit process?
- ☒ During litigation, how is a "records hold" implemented, and how can it be authenticated? How are multiple hold orders on the same record managed? How are records released from a hold order and their retention periods reinitialized? How is the controlled deletion of electronic records assured, absent any legal or regulatory requirement to retain the record? Can your records hold order be audited and verified?
- ☒ How can unauthorized duplication and/or distribution of sensitive corporate information, such as intellectual property or financial data, be prevented? How does the firm ensure that policy-driven access controls are applied? Does the firm require a means to audit or monitor and verify sensitive information leaks and for whom?

Common Archive Services

Once an active archive solution that supports many different content sources is in place, a series of common, horizontal services, such as index/search, data classification, access controls, retention, preservation, or disposal, can be performed. These common management services can be applied centrally across multiple content types based upon the defined business policies.

A firm facing litigation related to an employee lawsuit may need to search for and retrieve records related to a set of named employees across their email accounts, user file shares, and any applications that they had access to. In some cases, the employee may still be with the firm, and the company may want to monitor or audit employee activities. Similarly, a securities broker-dealer may need to retrieve all records related to a particular broker's correspondence with a client, and this may be done using email, instant messaging, voice communications, and the like.

The alternative is for a firm to manage policies, taxonomies, access control, and search and retrieval on an application-by-application basis — across many different repositories. Consider a large financial services firm has over 180 different commercial and homegrown application sources, each individually managed from a records compliance and technical perspective. The economies of scale realized in archiving that data in a common, active archive — while not disrupting the logical application access or data structures — can streamline both business and IT processes.

Data Authenticity, Integrity, and Preservation

An active archive should preserve the authenticity and integrity of the electronic records it is storing and ensure these records are secure and safe from unauthorized access, tampering, and will not change over time while they are being stored. Some regulations require that electronic records be stored in a way that ensures they cannot be changed, modified, or deleted during the lifetime of the record. Additionally, the active archive should provide some reasonable means of auditing and verification that the data being preserved is authentic. The consequences to a firm if the integrity, authenticity, or security of electronic records is compromised can include financial, reputation, or legal exposure.

Companies must conduct analysis on which regulations, both at the federal and state level, they must comply with that impact how they manage their electronic records. Some examples of federal or state regulations that impact electronic records retention, accessibility integrity, security, privacy, or safeguarding include:

- ☒ 17 CFR 240.17a-4 (also known as SEC 17a-4)
- ☒ Electronic Signatures in Global & National Commerce Act (also know as e-SIGN)
- ☒ Gramm Leach-Bliley Act (also known as GLBA or the Financial Services Modernization Act)
- ☒ Health Insurance Portability and Accountability Act (also known as HIPAA)
- ☒ 21 Code of Federal Regulations Part 11 (also known as 21 CFR part 11)
- ☒ California Database Protection Act
- ☒ Sarbanes Oxley Act of 2002 (also known as SOX)

Data Longevity

An active archive solution must preserve the longevity of the data. In many cases, content within electronic records must be retained for long periods of time. For example, some healthcare institutions must retain records for close to 50 years, and during that time, the underlying technology storing the electronic records becomes obsolete. Because the content will likely outlive the media on which it is stored — an active archival solution must provide for data longevity by easily and transparently migrating the content to the technology of the day. Data must be migrated to newer technology in such a way as to maintain data authenticity and integrity while preserving the chain of custody. In addition, the ability to migrate to new technology in a manner which is nondisruptive to application or user access is required.

Another level of complexity associated with data longevity is the format in which the content is stored. An active archive solution must support the ability to store data in standard formats for later retrieval. However, even standard file and application formats change and evolve over time, and an active archive solution must support the ability to store data in a format that can be retrieved with the common file or application format of the day. This requires that the active archive solution provide a continual means of updating or migrating content to newer standard formats as the formats are updated.

Easy Application Integration

From a technical perspective, an active archive solution in supporting ingestion of content from multiple heterogeneous repositories must provide for easy application integration. Easy integration is achieved by active archive solutions that use open, standards-based interfaces and avoid proprietary API integration development or training costs. The points of integration between content-producing applications and the active archive solution must be standards-based interfaces, in particular, to support internally developed or legacy applications.

Open, standards-based interfaces such as NFS, CIFS, HTTP, and WebDAV allow for commercial and custom application integration between the application and the active archive solution. Using open, standards-based interfaces eliminates the cost, risk, and burden of writing to and becoming locked into a specific vendor's API. Firms should evaluate which standards-based interface is the right implementation based upon the characteristics of the application and operating system and the underlying performance requirements.

Support For Storage and IT Policies

Storage and IT policies must be established in concert with business policies to ensure that the technical implementation of business policies satisfies, to the best of the firm's ability, the nature of the requirements of the firm. As a result, storage and IT policies are being reviewed by business, legal, risk, and records offices within a firm. A sample of some questions related to a firm's storage and IT policies include:

- Do you know which applications or servers, across multiple locations, have content that needs to be archived according to defined business policies? How much capacity is required in an active archive solution, both today and in five years? What is the current file volume, and what ingestion rates are required from the archive solution today and based on the volume of files in five years?
- With what frequency, access pattern, or capacity watermarks should content be moved or migrated from primary storage to an active solution? What response times need to be provided to retrieve content from an archive? Do these migration policies and response times vary by application or system?
- Which applications require or are best suited for implementation of duplicate elimination or single-instancing technology? What is the firm's goal in cost and storage savings related to this technology? Is all the proper metadata associated with duplicate records preserved so as to eliminate legal or audit risk during retrieval and production?
- What records can be deleted and when? How much storage savings will the firm realize with controlled deletion policies? What conditions will trigger a deletion? Will the deletion be automatic or manual? Is destruction of the content to a particular standard, such as DOD 5015, required? What is the firm's policy regarding destruction of content on drives that are being retired or decommissioned? How are old employee accounts and information handled?
- Can you identify which IT assets are being used by each employee? Can you identify which applications an employee has access to? How can SPAM or inappropriate content be blocked from viewing or storage? Do you know where content sensitive information lives, and are there employees who have access to this information and those who shouldn't?
- Does the archive solution need to withstand multiple points of and types of failure, such as hardware, network, application, and site disaster? Does the primary archive need to be protected and to what recovery level? Does the primary archive require high availability and automatic failover?
- Can you identify when backup tapes contain data from a selected server or application? When you migrate from older technology to new technology, what happens to the old media? How often do you perform a physical media audit to ensure that media is accounted for or disposed of according to company policy?

Storage Tiers and Integrated Management

Given the business drivers to control the preservation or disposition of electronic records in conjunction with continued data growth, firms need to deploy and leverage tiered storage architectures not only to meet backend service levels but to provide increased levels of storage economy for the firm. An active archive solution provides a second or archive tier of storage where primary data on primary storage can be moved to. This is important because it is estimated that less than 20% of the data a company possesses needs to be stored on high-performance enterprise disk storage. The offload of data from expensive enterprise disk to less-expensive ATA, SATA archival storage improves enterprise, back-office, and mission-critical application performance, and also reduces the time required to back up the primary data. However, a prerequisite to the effective use of tiered storage and active archival of content is understanding the value of the content to the business — so that information can be archived appropriately.

Making use of tiered storage architectures allows the right content to be placed on the right tier of storage, based on business and IT policies. However, a central requirement is the ability to monitor, report on, and control that tiered storage architecture, of which an active archive solution is included from a single management interface. Active archive solutions must be integrated with best-of-breed storage management standards, such as SMI-S, and fit into heterogeneous storage management architectures and products.

Ease of Management

Sole reliance on higher-capacity, lower-cost SATA disk media to address active archiving requirements is not sufficient. With continual price declines of the storage media, increasingly the costs associated with storage are in administration and management. Active archive solutions must address the requirement for less administration and provide the ability to be self-managing — thus relieving storage administrators from cumbersome and routine tasks, such as zoning, partitioning, and provisioning of storage, thereby freeing up these critical human resources for other important, more productive business functions.

To truly reduce administration, an active archive solution that houses terabytes of data must provide capacity and workload balancing, automatic capacity expansion, automatic reliability, and data rebuild and/or failover on error detection and automated data migration/ technology refresh. When additional capacity is added, it should be instantly available with no drives to format and LUNS or file systems to create. Conversely, the removal or failure of a component within an active archive solution should trigger the system to redistribute data redundancy, load balancing, and data migration within the archive system. Additionally, with large volumes of data being stored, an active archive solution that can eliminate the storage of duplicate records decreases storage capacity costs and further reduces administrative time.

Technology Refresh

As technology continues to evolve and improve, faster, higher performance, cheaper, and denser disk storage options become available. The challenge that a firm has to face is how to preserve and migrate data from older, obsolete technology to newer technology when they refresh their technology every three to five or even seven years. Historically, archival solutions have not provided an automated, nondisruptive way in which to accomplish underlying technology refreshes. Manual and labor-intensive migrations were planned for and conducted over the span of months.

Active archive solutions provide for online, nondisruptive migration of data from older media to newer media, thus removing the manual steps in the migration process. In addition, active archive solutions can support a mix of different media technologies within the same active archive architecture so that new technology can be deployed over time and older technology removed from the system over time. This reduces the pain and cost associated with forklift upgrades or migration services. Data is accessible during the migration process.

Interoperability

An archive solution must interoperate within a firm's existing environment and work across heterogeneous servers, both mainframe and open systems as well as across a broad mix of applications. The solution should not require any host-level software to be deployed and not require proprietary APIs to be tested, deployed, or upgraded. The solutions should work within an existing data protection schema and not require specialized products or tools for data protection or disaster recovery of the active archive environment.

An active archive solution should fit into a firm's larger tiered storage and storage management environment, and higher level storage management frameworks should be able to discover, monitor, and report on an active archive's storage properties, such as performance, capacity, utilization, events, and the like.

HITACHI DATA SYSTEMS: PROVIDING A FRAMEWORK FOR ACTIVE ARCHIVING

Hitachi Data Systems (HDS), a wholly owned subsidiary of Hitachi Ltd. (NYSE:HIT), is a leading supplier of storage systems, software, and services to enterprises around the globe. HDS conducts business through direct sales and more than 800 resellers in the public, government, and private sectors in over 170 countries. Its customers include more than 50% of Fortune 100 companies.

HDS' Application Optimized Storage provides an excellent framework on which to deploy an active archive approach to the long-term retention, preservation, retrieval, and disposition of critical business content. HDS' Application Optimized Storage is a strategy to align business and IT by optimizing storage infrastructure and management with application requirements based upon price, performance, availability, and functionality. Hitachi Data Systems' best-of-breed tiered storage solutions also offer an excellent foundation. The solutions use common storage

services and a heterogeneous and standards-based approach to storage management, providing the necessary storage infrastructure for active archiving.

Hitachi has introduced the Hitachi Content Archive Platform. The Hitachi Content Archive Platform is designed to address the limitations of first-generation content-addressed storage solutions. These early solutions did not provide adequate levels of interoperability, scalability, or availability, and they had limited ingestion of content from multiple content repositories. Previous generation solutions showed limited scalability in capacity requirements and did not provide sufficient retrieval performance or make use of open standards.

With the Hitachi Content Archive Platform, HDS extends its tiered storage offerings to include content-based archival storage solutions as part of the portfolio. The Hitachi Content Archive Platform can make use of HDS' tiered storage architecture and its complementary storage services to help customers achieve the security, retention, preservation, classification, retrieval, or disposition of business-relevant content. Over time, common storage services can be leveraged to perform tasks such as heterogeneous data replication and migration between storage products, locations, and environments. Additionally, the ability to monitor, report on, and control the entire HDS-tiered storage infrastructure, including the Hitachi Content Archive Platform, from a single management interface reduces operating expenses and improves internal service levels.

Hitachi Content Archive Platform

The Hitachi Content Archive Platform is an active archiving solution made up of both software and hardware, which support policy-based ingestion and archive of content from many distributed or centralized repositories, such as email, file systems, databases, applications, and content or document management systems. The Hitachi Content Archive Platform ensures the secure archival-quality retention, preservation, or verifiable destruction of content. Through the use of the Hitachi Content Archive Platform, users can leverage a set of common and unified archive services, such as centralized search, policy-based retention, authentication, and protection.

The Hitachi Content Archive Platform provides the following advantages:

- ☒ **Future proof.** Storing data in standard file formats for future access and the ability to nondisruptively migrate data from one platform to another
- ☒ **Self managing.** Support for truly enterprise-class content provisioning by automatically adding storage capacity to the Hitachi Content Archive Platform
- ☒ **Multiple application support.** Support multiple applications using a standard file system interface versus proprietary APIs
- ☒ **Scalability, performance, and reliability.** Due to inherent data protection capabilities, SAN back-end storage and distributed object-based architecture objects are secure, protected, and available
- ☒ **Secure and compliant.** Retention, authentication, and immutability at the object level ensures compliance and integrity

- ☒ **Easy implementation.** Use of open interfaces for easy application integration and delivered as an integrated appliance for seamless installation and configuration

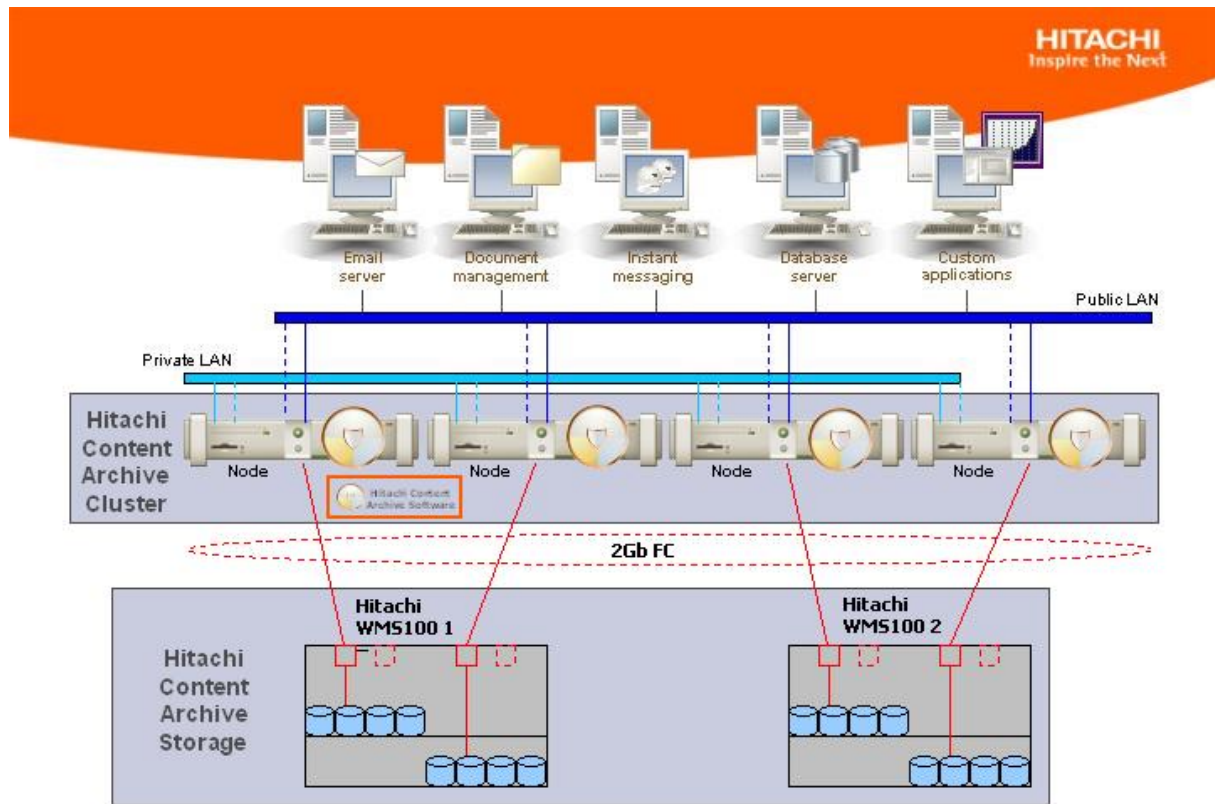
Product Architecture

The Hitachi Content Archive Platform implements a SAN plus Array of Independent Node architecture that is an integrated solution of software and storage. It is deployed as a preconfigured package at the firm's location. The following components, illustrated in Figure 1, all work together within the Hitachi Content Archive distributed architecture.

- ☒ **Hitachi Content Archiver.** The Hitachi Content Archiver, powered by Archivas, is the software component that provides the intelligence, policy-based control, authentication, preservation, and protection of the Hitachi Content Archive Platform. The Hitachi Content Archiver software runs on a preconfigured operating system that runs industry-standard servers. The Hitachi Content Archiver software communicates with external content applications via a series of standard, IP-based protocols, such as NFS, CIFS, HTTP, and WebDAV.
- ☒ **Hitachi Content Archive Platform Cluster.** The Hitachi Content Archiver software runs on minimum of four standard, Intel-based server nodes that make up a cluster. Each node in the cluster comes with a preconfigured operating system and the Hitachi Content Archiver software. All nodes in the Hitachi Content Archive Platform cluster communicate across a distributed architecture to share information, processing, and indexing responsibilities. Each node supports front-end communications over the IP network via standard protocols and back-end communications with the Hitachi Content Archive Platform Storage via Fibre Channel.
- ☒ **Hitachi Content Archive Platform Storage.** The Hitachi Content Archive Platform makes use of the SATA-based Hitachi TagmaStore Workgroup Modular Storage WMS100 platform. Each Hitachi Content Archive Platform Cluster communicates with the TagmaStore WMS100, which supports a minimum of 5TB (useable capacity), via Fibre Channel protocols. The Hitachi TagmaStore Workgroup Modular Storage model WMS100 offers high capacity, easy-to-install data storage ideal for active archive environments. The WMS100 matches requirements for the cost-effective scalability, easy upgrade and performance in a small footprint while maintaining reliability. Over time, the Hitachi Content Archive Platform will add support for the Hitachi TagmaStore Adaptable Modular Storage (AMS) line as well as the family of intelligent virtual storage controllers — the Universal Storage Platform (USP) and the Network Storage Controller (NSC). With USP/NSC support, customers will benefit from the virtualization inherent in these products to create a common storage environment that includes the archive storage tier.

FIGURE 1

Hitachi Content Archive Platform Architecture



Note: Hitachi Content Archive Cluster nodes and Hitachi WMS 100 Storage all run in a standard rackmounted configuration.

Source: Hitachi Data Systems, 2006

Content Flow

The Hitachi Content Archive Platform ingests content from many different content repositories simultaneously. The content is ingested or acquired as files via standard network protocols, and policy-based controls such as retention are applied to the content during the preservation phase. Content is stored in an open, standard-based format to allow for unified indexing, search, and retrieval of content at a later date. The following phases of the flow of content through the Hitachi Content Archive Platform are described in more detail below:

- ☒ **Creation.** Data is generated by a creating application, such as an email application, a database, various office applications such as Word, Excel, etc. The data is then stored on primary disk storage and is under the control and purview of the associated application. Based upon a firm's business policies, the content within the data may need to be archived for prescribed periods of time or the firm may desire to migrate the data out to a secondary tier of disk storage to optimize performance and/or storage capacity.

- ☒ **Movement.** The copy or migration of data from its primary storage location to an archive location is done by data movement middleware software, such as email, database or file system archiving software or hierarchical storage management (HSM) software. These software tools are responsible for the identification of individual files, folders, or email for migration to an archive based on either standard file system metadata or application-specific metadata policies.
- ☒ **Ingestion.** Ingestion defines how content gets added or archived to the system. Ingestion occurs when files are copied or migrated from a primary storage location to the Hitachi Content Archive Platform. The Hitachi Content Archive Platform uses standard network protocols, such as NFS, CIFS, HTTP and WebDAV, to ingest files from multiple content sources. Files ingested into the Hitachi Content Archive Platform are associated with policies and metadata for automated maintenance and easy retrieval.
- ☒ **Preservation.** The preservation policies define how data is stored, protected and managed over time. This occurs through retention, authentication, and protection policies, which are either enforced by or defined in the Hitachi Content Archiver software layer. Retention defines how long a period to archive a file for and prevents modification or deletion prior to expiration of its retention. Authentication ensures that content stored in the Hitachi Content Archive Platform matches a digital signature or hash value of its content and verifies that the content has not and will not change over time while it is being preserved. Protection ensures the integrity of the files by protecting them via RAID in order to support up to three simultaneous points of failure.
- ☒ **Access.** Access occurs when the business needs information out of the Hitachi Content Archive Platform due to regulatory audit, legal discovery, or general operational requirements. Access to the content within the Hitachi Content Archive Platform can occur through the controlling application (which provides for search of its own content) or directly through the use of the Hitachi Content Archiver software, which allows for a unified search of content based upon a keyword or series of keywords, across applications which have their data stored in the Hitachi Content Archive Platform. This unified view and search into the content reduces discovery times, shortens audit windows, and produces results in a more efficient and accurate manner than conducting searches on an application-by-application basis.

Content Policies

The Hitachi Content Archive Platform makes use of a combination of data, metadata, and application policies in conjunction with its own active archive metadata and policies when establishing its retention, authentication, protection policies. Policies represent a unique type of metadata that provide for setting a retention period for an archive file or collection, providing assurance of authenticity using configurable hashing algorithms (MD5, SHA256, etc.), or defining the level of protection to ensure integrity and availability. These policies are wrapped together as part of the content for transfer to another system (e.g., for replication) or for backup purposes. Doing this maintains the integrity between the content itself and its associated metadata and policies.

The Hitachi Content Archive Platform manages content as active archive objects within the system. Each active archive object comprises the following components:

- ☒ Data that contains content such as email archive files.
- ☒ Metadata that includes all the application, file system, and active archive metadata. Application metadata includes metadata provided by the application such as with email header information (to, cc, from, subject, date). File system metadata is provided by the file system and includes attributes such as file name, file size, m time, c time, etc. The Hitachi Content Archive Platform also applies archive metadata.
- ☒ The Hitachi Content Archive Platform archive objects can have their own policies or policies that are inherited from parent directories or file systems. Hitachi Content Archive policy attributes include retention periods (time-based, event-based, or infinite), authentication (to ensure content of a file matches its digital signature or hash value), and protection (such as RAID levels to support simultaneous points of failure).

Multiple Simultaneous Content Sources

The Hitachi Content Archive Platform can support ingestion of content from many different content sources. The operating and capital cost savings in management, training, and technology in leveraging a common active archive solution for multiple content sources can be profound and allows for centralized search/retrieval and data security, authentication, and integrity. The Hitachi Content Archive Platform supports ingestion of content from the following:

- ☒ Multiple, line-of-business Enterprise Content Management (ECM) or document management, document imaging, or check imaging systems
- ☒ Divisional or enterprisewide email or instant message archive systems
- ☒ Departmental or facility-based Electronic Medical Record (EMR), Computerized Physician Order Entry (CPOE), or medical imaging systems
- ☒ Networked, distributed heterogeneous file systems across departments, locations, and operating platforms
- ☒ Specialized databases or application systems, either commercial or homegrown

Open, Standards-Based Design

The Hitachi Content Archive Platform uses open, standards-based interfaces, which eliminate requirements for proprietary API integration development or training costs. The points of integration between content-producing applications and the Hitachi Content Archive Platform must be standards-based interfaces to support internally developed or legacy applications. The solution uses standards-based, open interfaces, such as NFS, CIFS, WebDAV and HTTP, as well as storage management standards, such as SMI-S, and are a natural extension of HDS commitment to a standards-based approach to storage management.

- ☒ NFS is used predominantly in Unix environments and is supported by all leading applications as well as internally developed systems but does have some associated protocol overhead. The active archive solution will present a NFS file system view to the application source and will handle and recompile objects that come in out of order due to the nature of the NFS protocol.
- ☒ CIFS is used for Windows systems and is used to map a Windows network drive to an active archive folder or directory. Windows administrators can set up Windows folders or drives, which they map to an active archive directory location where the files are stored.
- ☒ WebDAV is a platform independent, high-performance interface, which is an extension to HTTP and supports RFC 2518 compatible clients. The mount point to an active archive file system or directory is presented as part of the HTTP URL.
- ☒ HTTP is a high-performance interface that defines a set of rules for exchanging content between a Web-enabled application source and the active archive solution.

Additionally, the Hitachi Content Archive Platform stores files in their native form with original names to allow for data longevity and to ensure easy access to and retrieval of content over time.

Scalability, Performance and Reliability

The Hitachi Content Archive Platform has been tested to scale in capacity from 5TB to over 300TB in a single Hitachi Content Archive Platform repository. Additionally, the product can support a huge volume of files (up to 350 million per archive tested to date and growing) ingested from multiple application sources simultaneously and can scale linearly with additional clusters. Architecturally, the Hitachi Content Archive Platform can scale to more than 2.5PB in a single repository and support up to 2 billion files per archive. The product makes use of native RAID capabilities within the TagmaStore product in a SAN + array of independent node architecture (SAIN) to ensure data protection and can withstand simultaneous points of failure. Additionally, through the use of remote replication to a second Hitachi Content Archive Platform, the Hitachi Content Archive Platform can withstand site disasters. The Hitachi Content Archive Platform makes use of a symmetric, parallel processing, distributed architecture combined with a backend SAN design to allow users to add capacity quickly and without reconfiguration. Lastly, the performance of retrieval of objects

from the Hitachi Content Archive Platform occurs in seconds and has been shown to be 4 to 5 times faster than first-generation content-based archive approaches.

CHALLENGES FOR HITACHI DATA SYSTEMS

HDS' foundation of hardware, software, and professional services offerings and its proven expertise and excellence in delivering best-of-breed technology will prove invaluable in building out a long-term active archive approach. HDS expertise in developing best-of-breed storage-tiered solutions on which a set of common storage services can be used while making use of heterogeneous and standards-based approaches to storage management provides an excellent foundation for active archiving. The challenges for HDS delivering on an active archive solution are largely factors that are external and include environmental complexity and the partner ecosystem.

Environmental Complexity

One of the challenges ahead for HDS involves the complexity of the customer environment from both a business and technology perspective. The business complexity involves the changing legal landscape as it relates to electronic discovery as well as vague or currently evolving legislation that is not prescriptive in nature and can cause confusion within the customer environment. HDS can use its network of business and professional services partners to help keep up to date with changing legal, regulatory, and business environments. The technology complexity evolves around the lack of certainty regarding technical requirements to satisfy regulatory requirements and the scope of the technical environment from a data, systems, application, and network capacity perspective. HDS manages some of the most complex computing environments today, and it is one of the few vendors that can provide an integrated, consistent, and reliable approach to active archival and long-term protection of business critical content.

Partner Ecosystem

The second challenge for HDS is that an active archive solution must interoperate and work within a large system or ecosystem of application, database archiving, enterprise content management, and file system vendors. This ecosystem can quickly become a complex Web or matrix environment that must be tested, validated, and managed to ensure that the customer environment, and its critical business content, is secure and protected not only at deployment, but over the course of several decades.

HDS quality assurance and testing process in combination with its partner programs ensures a comprehensive, integrated portfolio of best-of-breed hardware and software components from the industry's leading vendors. This broad portfolio provides firms with a choice of selecting the combination of components that best meet their specific business objectives or interoperability requirements. HDS partner programs, in conjunction with Hitachi Data Systems' iLab, assure that customers will have full interoperability testing and support, while also offering cost-effective and efficient solutions for the management and administration of storage.

SUMMARY

To address the business needs of today's regulatory, legal, and competitive business environment, firms must consider active archival solutions that ensure that content integrity, accessibility, and reliability are maintained. The value of content integrity is immense, and the method for and timeliness of retrieving this content are critical in today's competitive environment.

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2006 IDC. Reproduction without written permission is completely forbidden.