Data Archiving: Foundation Capabilities for Compliance and Cost Optimization

Introduction to Archiving

This paper introduces concepts and techniques for data archiving.

Compliance can touch so many aspects of an organization—legal, security, and records retention, among others; and this is one reason that many point products exist to attempt to solve the need. Compliance involves, amongst others, asset management, identity access, storage, server, file systems and database monitoring, and especially change management and configuration management. All of the compliance requirements are addressed by enterprise archive solutions. These solutions are complex, costly to build and maintain, and frequently need to be re-designed to incorporate new technologies or address expanding scope and new regulations.

Based on IBM experience, enterprise archive solutions built within the framework of sound reference architecture provide cost savings—for a company with 1 PB storage environment, the savings on average exceed $10M annually as a result of optimized utilization of resources.

What's driving the need for a comprehensive approach to Enterprise Archive and Information Management?

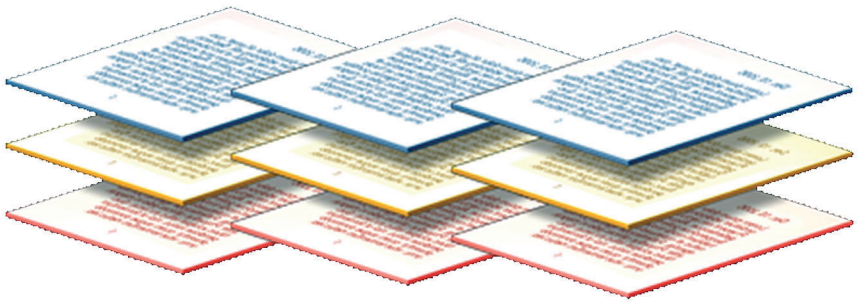
Compliance mandates have brought about increased regulatory oversight with a multitude of new regulations and enforcement initiatives. For many organizations, compliance management has become a fact of life, and not addressing compliance issues can potentially result in heavy fines, corporate sanctions or expensive class-action lawsuits. Today without clear understanding of how or what information relates to regulations, many customers are simply taking a save everything forever approach. This is driving data growth well above normal organic and business growth. In many cases, this is 50 - 60% annually. For most organizations, this growth cannot be sustained forever.

Organizations need a flexible approach that helps them make smart, cost-effective decisions. While each mandate has its own specific requirements, all have a common denominator: the need for greater visibility and control of information and processes to lower risk.

Despite these growing market needs, most vendor offerings remain focused only on archiving point solutions with limited simplistic functionality. These offerings typically are limited to solutions that collect and archive only one type of content in inflexible, siloed environments that have no common access. Organizations need an integrated, on demand architecture that uniquely addresses their compliance needs for compliant information management.

Definition of Archiving

Archiving is an intelligent process **for placing on the right tier** inactive or infrequently accessed data that still has value while providing the ability to **preserve, search, and retrieve** the data during a specified **retention period**.



Scope

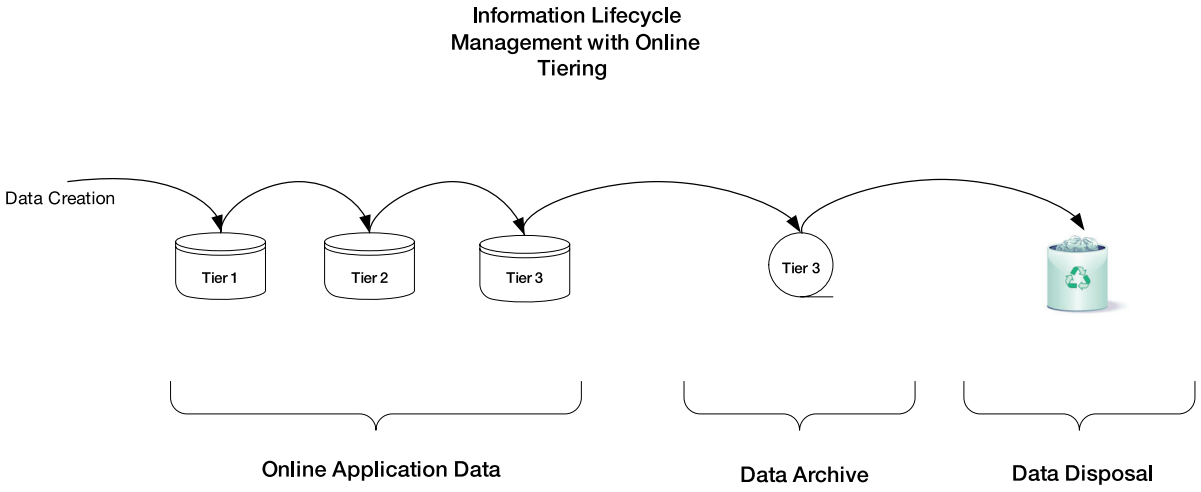
Archive goes beyond simply the infrastructure that makes up the archive environment. Before an archival architecture can be defined, it is necessary to understand the types and requirements of an organization's information. Archive starts with the development of well defined information management policies that clearly identify each information object, how long to retain it and when it becomes static. Once policies are defined, it is important to understand the data characteristics such as: is it structured, unstructured, or semi-structured data etc.; what is the data volume; what are the retrieval requirements, etc.

It is important to define what and when to archive information but it is just as important to define when to purge the information. Not managing the disposal of information can cause an explosion of retained data and make management and searching the archive difficult.

In addition to understanding the requirements of information management, it is also important to develop a long-term strategy and architecture for the archive environment. Today, most information's retention requirements outlast the technology that it resides on. It is important to build an architecture strategy that is flexible and can adjust to changing requirements and technologies.

Archive in its most basic form is simply the migration of information/data from online application accessible storage to an offline long-term storage repository. However, in order to leverage the efficiencies of the use of an archive, it is necessary to look at an enterprise level. At this level, we introduce the process of Information Lifecycle Management (ILM). At its simplest form, ILM manages the disposition of information/data from its creation to its placement on initial online storage to its migration to archive long-term storage to is final deletion.

In today's IT environments, most organizations leverage tiering of online storage to optimize the efficiency of their environment. The illustration below depicts how ILM manages the process of data mobility from inception to final deletion.



Pain Points addressed by Archiving

Examples of common pain points felt in the IT industry:

- *"Save Everything" attitude*
- *Excessive Storage Costs (physical and electronic)*
- *Litigation (underway or impending)*
- *Government Audit/Investigation (underway or impending)*
- *Excess time to locate materials/records*
- *Lack of records management policies and/or procedures*
- *System Changes (especially to electronic document/records system)*
- *E-mail/filing e-mail*
- *Vital Records Protection*

- *Long-term preservation*
- *Outdated retention schedules*
- *Security of Confidential Information*
- *Meeting regulatory/compliance requirements*

Archive should address both Business and IT requirements

Ultimately the goal of all IT services is to help meet the business objectives of the organization and do it as efficiently as possible. Enterprise Archive should be positioned to meet these objectives. The business challenges for information management focus on three main areas: Compliance and Risk, Document and Process Automation and Information Innovation aligned with Business Value. IT challenges focus on System Efficiency and User Productivity.

Archiving Benefits

The benefits of an enterprise archive strategy align with challenges of the business and IT. The three basic areas include Managing Risk and compliance, Improve Efficiency and cost savings.

Manage Risk and Streamline Compliance

- *Improve controls for records management, e-mail, and auditing*
- *Comply with government/industry regulations*
- *Reduce costs and risk for legal discovery*
- *Support litigation requirements*

Improve Efficiency

- *Enhance systems/e-mail performance*
- *Leverage existing information to meet business objectives (service model strategy)*
- *Improve and sustain efficiency by leveraging people, processes, and technologies used to deliver information services to the business*
- *Define and implement the appropriate archiving strategy to address current and future business requirements*

Reduce Cost and Simplify

- *Control demand for storage*
- *Improve asset utilization*
- *Reduce HW/SW/personnel costs*
- *Reduce data migration effort*

Drivers for Information Growth

There are many drivers that contribute to the information growth in the IT industry, such as:

Regulatory/Industry Compliance

- *Basel II*
- *Sarbanes-Oxley, Euro-SOX, and J-SOX*
- *FRCP*

Existing applications

- *E-mail*
- *E-commerce*
- *ERP/CRM*

New applications

- *Web 2.0*
- *Digital video, voice, and audio*

Data multiplier effect

- *Backup and Disaster Recovery*
- *Test and Development*

Mergers and acquisitions

Building an Enterprise Archive

The strategy should include all aspects of the storage environment, including technology, process, organization, and governance.

Identify Application information management requirements

Meet with application owners and IT support staff to understand each applications true requirements and what service IT is currently providing to meet those requirements.

Identify the information objects

Identify the key related components that make up a set of information of the application.

Identify Compliance requirements

Regulatory requirements are the requirements established by the Government to have data accessible. These include how long the data must be kept before purging and how quickly the data needs to be presented back to the systems.

Organizational requirements are just as it states, the requirements of the different business units for having the data accessible.

Determine business value of information over time

All data is created equal but over time loses its value to the business. Establish the timeline that the data should be on Tier 1 storage and when it can move to lower, more cost-effective tiers later in its lifecycle.

Establish Archive policies

Establish the archive polices for application data based on Regulatory and Organizational requirements. When can this data move to a more cost-effective tier of storage, disk, online tape, or offline tape, etc.

Develop Enterprise storage architecture

Technology

- *Hardware*
  Design the hardware requirements necessary to support the environment. Develop Tiers of storage that can include different disk sizes and speed, VTS libraries to keep data online but off more expense disk, and ultimately offline tape architecture for longer term requirements.

- *Software*
  Standardize on the software platform to support the environment. Software architecture should address backup and recovery requirements, archiving capabilities, recall capabilities and ultimately data purging/deletion.

Process

Establish the processes for the management of information within the archive. One of the key processes of an archive is Information Management.

Organization

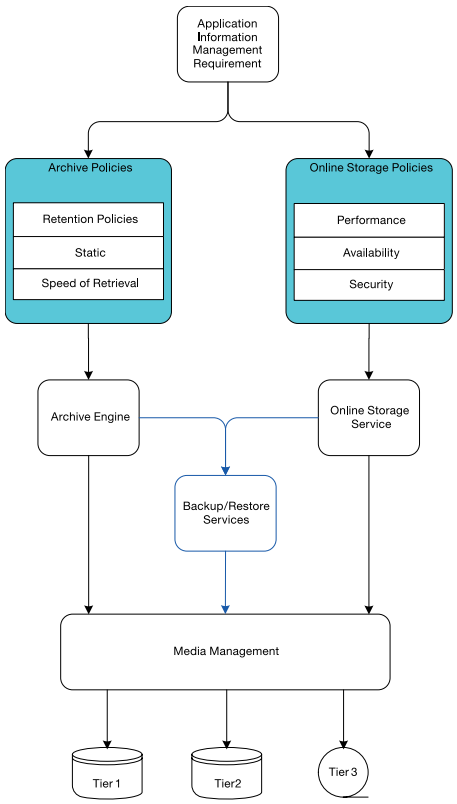
Define the key roles and responsibilities for management of the archive.

Governance

Establish the governance for managing the information such as monitoring and enforcement compliance for information management.

Enterprise Architecture

High level View of an Archiving Solution



The first step in building an archiving strategy is to identify the application information requirements. Once the requirements are defined, the archiving polices can be established along with the storage services required to support them. Information is then fed into the Archiving Engine, which places the information on the appropriate tier of storage.

What is a Reference Architecture?

A reference architecture is a set of architectural best practices for use by everyone in an organization. A reference architecture does not simply span technology. It also encompasses process, organization and governance. The key to a reference architecture is standardization across an organization. Standardization of processes, organizational roles and enforcement through governance drives better management and use of the technology.

IBM Reference Architecture for Archive

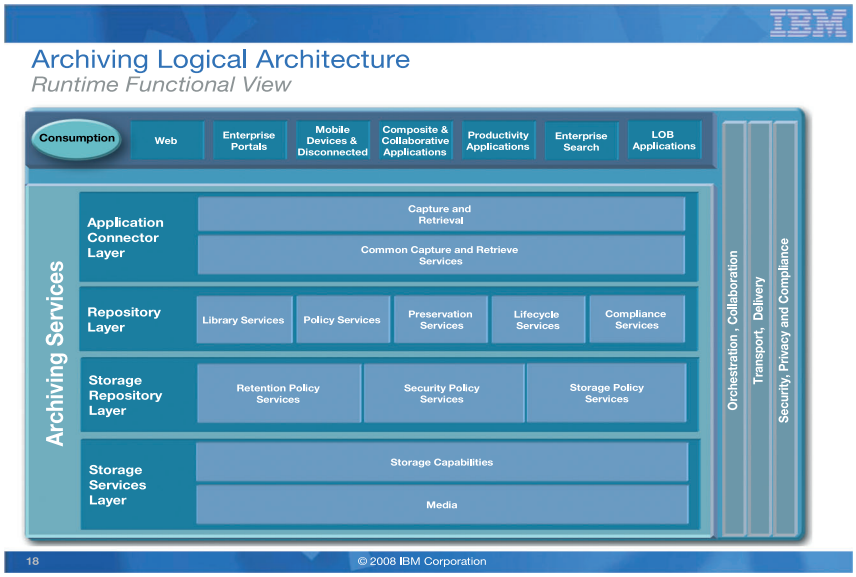
In 2008, IBM announced its Reference Architecture for Archive. It is the first and only archive reference in the world today. IBM's approach to Archive is an Object Oriented Architecture. With components within the architecture that perform specific related functions and standard interface to other components. Through this methodology, related functions are independent of others and can be modified or upgraded as requirements and technologies change.

The building principles of the reference architecture reflect the need for information agenda for the enterprises:

- *The need to determine primary business objectives that can be impacted through the use of information*
- *The need to drive consistency around how information is defined and used across the enterprise*
- *The need to identify technology components and capabilities to address information requirements*
- *The need to establish a plan for projects that deliver both long and short-term returns on investment*

The Archiving Reference Architecture is one of the numerous assets and accelerators that are available to apply throughout the entire Engagement Process to reduce risk and improve our delivery execution. Other assets are IBM's IT Optimization Reference Architecture, IBM's Information on Demand Reference Architecture, and IBM's Service Oriented Modeling & Architecture.

The following illustration depicts the **Archiving Offerings Framework**. This is a high level view of Archiving Reference Architecture intended for use with non-technical, business audiences. The diagram uses terminology with which most business people are familiar and is designed to show IBM Archiving offerings at a high level and to frame the discussion; while not specifically an architecture diagram, it provides a starting point for a discussion of the actual reference architecture diagrams to follow.



The above slide depicts the set of Archiving Technical Reference Architecture diagrams. This level of the Architecture spans multiple diagrams in order to show the necessary detail behind each major box on the Archiving Logical Architecture diagram. In other words, these diagrams represent a decomposition of the Archiving Logical Architecture. The diagrams show implementable components (i.e., software or hardware) of the architecture, and do not attempt to illustrate business concepts.

Application Connector Layer:

This is the interface between the Archive Service and the producers and consumers of archive information. This includes functions like extraction of information from the producer such as applications and the retrieval and presentation of archived information back to the consumer.

This layer provides the methods and mechanisms for the interfacing with the producers and consumers of information. It takes direction from Archive Management, such as what to capture, and performs the action. It also is the interface for consumers to request information from the archive. It also includes basic interface functions such as decomposition, formatting and compression of the information for storage and recomposition of the information into a format easily interpreted by the consumer.

Capture Management

This is the process for extracting information from the producer based on policies defined in the Policies Services process. It determines what to extract and when to extract the information.

Retrieval Management

The process for handling requests from the consumer of archive information, executing the request and presenting the results to the requester.

Common Capture and Retrieve Services

These are the services for formatting information and preparing it for either storage in the archive or presenting it to a requester.

Rules and Management Layer:

The Rules and Management Component Layer is the central component of the archive services. It is where rules for how to manage the information within the archive are documented and maintained. It directs the activities of the process in all other component layers. It determines what and when information is captured from producers, how the information is stored. It also directs how information is secured, the length it is retained, and when information is to be disposed.

Library services

Library services provides for and directs the Storage Component Layer on the handling of archived information. This includes placement of data on the initial tier, the movement of data between tiers and the destruction of the data when it becomes obsolete as determined by the policies defined in policy services and directed through Lifecycle services.

Policy services

Policy services defines policies for handling what information to extract, security, confidentiality, retention and disposition of information objects. It interfaces with Library Services, Capture Management, and Retrieval Management. Policies are gathered by Preservation Services.

Preservation services

Preservation services gathers information long-term storage requirements and establishes how information is to be handled within the archive. It is the interface to all the external processes such as Compliance and Information Management, Consumer and Producer Management.

Lifecycle services

Lifecycle services directs the Library services on how information objects are to be handled within the archive, such as where to store the information objects, how long to retain them, when they should be moved from one tier to another and how to dispose of them when they become obsolete. Lifecycle policies originate from Preservation Services and maintained in Policy Services.

Compliance services

Compliance Services ensures the rules documented in policy services are enforced. It monitors and logs the activities of the other services as it relates to the custodianship of the information objects stored in the archive.

Storage Management Layer:

This layer is the perform layer for managing the data that represents the archived information objects. Management of data includes the retention, security and storing of the data on the appropriate media.

Retention Policy services

As directed by the Library services, Retention Policy services manages when data representing information objects are deleted from the archive media.

Security policy services

The security and confidentiality of the data is fundamental to the custodianship of information objects within the archive. Security Policy Services maintains and ensures that the policies for security and confidentiality are followed. This includes processes such as user access validation, encryption and security logging.

Storage policy services

Storage Policy Services communicates to the storage Service Layer on how to handle the placement, movement and purging of data from the storage media.

Storage Services Layer:

The storage services component layer is the workhorse of the archive services. It takes direction from Archive Management and performs the actions for maintaining the data that constitute the information objects. It handles such actions as storing the data on the appropriate tier, moving it between tiers, and disposing of the data when the retention policy directs.

Storage capabilities

This is the process by which all activities for the storing and accessing of the data as well as managing business continuity and disaster recovery. This includes remote data replication, backup and restore, and data placement.

Media

Media constitutes the hardware and software for the storing of the data. This includes disk and disk management, tape and tape media management.

Interface Requirements

IBM's Reference Architecture for Archive identifies the requirements for interfacing between the different service layers.

Application Connector Layer to Repository

- *FS (Meta + Retention)*
- *Standard: CMIS (IBM internal standard today)*

Repository to Storage Repository

- *FS (Simple/Meta + Retention)*
  - *CRUD + Retention*
  - *Management/Search via other mechanisms*
- *TSM/SSAM (Compatibility)*
  - *CRUD + Query- + Retention++*
- *Standard: XAM (Rich/Future)*
  - *CRUD + Query + Retention++*

Storage Repository to Storage

- *Block*
- *File*
- *OSD*

The interface requirements provide assurance for longevity of the enterprise architecture—the clients can replace existing technologies with new ones as long as the replacements meet the interface requirements.

Enterprise Archive Software Enablers

The data handling of structured and unstructured data differs in the methods used for extraction, storing and retrieval.

Structured data management is supported by SQL based Data Base Management Systems. The challenge in archiving structured data lays in the federation of related information objects. Many tables can make up a single information object to be archived. Understanding, relating and capturing them takes careful planning and data management.

IBM's Optim™ simplifies the process of managing the data from capture to retrieval. It interfaces with multiple database types such as Oracle®, DB2®, IMS®, and others.

Unstructured and semi-structured data such as e-mail, documents, images, and other files impose their own challenges to archiving. Individual files may or may not have relationships to other objects and their context may not be clearly understood in the future. Having a tool to help manage the capture and catalog of these data types is an important part of an Enterprise Archive. Since 80% of an organization's critical business information is in the form of documents and e-mail, enterprise content collection and archiving is the first step to taking back control. Content Collection and Archiving is the process and technologies for collecting, enhancing and managing virtually all types of content, regardless of creation type and storage mechanism. Today, many organizations' needs have moved beyond the limited business value of e-mail archiving to include file systems, Quickr™, SharePoint, instant messages and other information; all are growing exponentially. And, organizations cannot archive all of this content. They need more advanced functionality to intelligently manage content that meets new legal eDiscovery and records management requirements. IBM's Enterprise Content Management software can handle nearly any data type and situation. IBM's Content Collector is an example of software suited to handling e-mail as well as other file types; IBM's Commonstore helps archive information from SAP® databases.

As part of an object oriented, enterprise archive component architecture, management of media and data placement and disposal are handled and managed in concert with, but separate from, other components in the architecture. These functions can be handled either at a software level using tools like Tivoli® Storage Manager or as an appliance like IBM's DR550, which is a WORM (Write Once Read Many times) device with bundled SSAM (Systems Storage Archive Manager) software.

As a software solution to policy management, both structured and unstructured data needs to be moved into appropriate tiers of storage supported by the right classes of service. The software assisting with this is the policy management software; like IBM's Tivoli Storage Management—TSM—and its archive instance Systems Storage Archive Manager—SSAM.

Enterprise Archive Hardware Enablers

As demand rises to manage the exploding growth of data, more advanced technologies continue to be developed by suppliers to manage the movement, storing and disposal of data. IBM has a number of products well positioned to manage all aspects of the information lifecycle.

IBM TotalStorage DR550

DR550 is a WORM (Write Once Read Many times) device with the following characteristics:

- *Designed to meet the requirements of efficient long-term data retention and protection*
- *Scalable to 224 TB*
- *Offers low TCO by using multiple storage tiers (disk, tape, and optical) within an archive*

Its benefits include:

- *Addressing customer data growth challenges by retaining inactive business data in a cost-effective, secure repository*
- *Managing data throughout its lifecycle, supporting data retention, data security and end-of-life destruction*
- *Reducing the cost of storing data over the long-term through the usage of multiple low-cost storage tiers (disk, tape, and optical)*
- *Addressing regulatory and non-regulatory data retention requirements*

Emerging HW Technologies

Several technologies for secondary storage systems are emerging to help customers store and manage the increasing amount of data retained in active archives.

These secondary storage systems present different types of interfaces for use in different contexts. For example, these systems can act as Virtual Tape Libraries, storing backup/archive data from applications such as IBM Tivoli Storage Manager (TSM). They can also present themselves as file servers via an NFS or CIFS interface, or as block storage subsystems via an iSCSI, FCP or other SCSI interface.

Massive Arrays of Idle Disks (MAID)

MAID is a system using hundreds to thousands of hard drives for near-line data storage. MAID is designed for Write Once, Read Occasionally (WORO) applications. In a MAID each array is only spun up on demand as needed to access the data stored on that drive. Some implementations cap the percentage of arrays that are spun up to ensure minimal energy usage, and because of the low energy usage pack the disks extremely densely to achieve minimum footprint as well. MAID can supplement or replace tape libraries in hierarchical storage management.

Deduplication Gateways and Systems

While generic compression technologies can be expected to achieve compression ratios of 2x – 4x, deduplication technologies are emerging which have been shown in real environments to achieve compression ratios of 10x and sometimes much more.

Whereas generic compression technologies such as Lempel-ziv look for fine-grained repetition in data, deduplication technologies search for larger (data sector, data block, file) sized duplications, and with much less meta-data overhead, store only the unique chunks of data.

Deduplication has been shown to be especially effective in the context of backup/restore data streams, where a deduplicating system can capitalize on deduplicating the same file from multiple sources as well as slightly changed versions of the same file from the same source.

Deduplication is an effective technology for minimizing the amount of data stored in an archive or backup repository, thereby saving on capacity, energy and physical footprint. In addition, deduplication technology can minimize bandwidth requirements for transmission of the data into a repository. This particularly enables remote storage of backup or archive data by reducing the WAN bandwidth requirements enough to make remote storage practical and cost-effective.

Tape Libraries

Tape and tape library technologies, while they have been around for a long time, continue to evolve very quickly, and they provide data storage at significant cost and energy advantage over disk-based subsystems. A recent study showed that tape libraries can hold twice the amount of data in a given space at much lower cost while using a fraction of the energy and providing similar streaming performance. Of course tape storage has limitations due to the time required to mount and load the tape in a tape drive and seek the tape prior to retrieval.

Current tape technologies such as LTO are highly reliable, and tape is still preferred for applications where physical movement of bulk data (e.g. for offsite archiving) is required.

Integrated Systems

Evaluating, deploying and optimizing use of all these technologies can and will be a daunting task. Enterprises looking to invest in archiving functionality can expect that system vendors such as IBM will soon offer solutions that integrate many of the archiving software and hardware components across multiple layers into a simple-to-deploy package.

Roadmap

IBM's roadmap to a successful Enterprise Archive Architecture is based on three basic phases—Archiving Briefing, Archiving Roadmap, and Archiving Solution. This end to end approach will assist clients in adopting IBM's Reference Architecture for Archive into a flexible Enterprise Architecture.

Archive Briefing

In the briefing phase, the client is guided to understand key business and IT imperatives. Each briefing is tailored to a particular customer audience (e.g., LOB, CIO) and creates tangible results to guide the success of the next phases. Objectives are: understand the customer's business and IT strategy and key initiatives, identify the business value proposition of information agenda to support the strategy and key initiatives, and identify the initial scope for archive agenda development. Key activities are Customer Tailored Briefing by Industry, Discuss Industry Archive Agenda, and Enterprise Archive Methodology; and key deliverables are Industry Records Management and Retention guidelines, Customer case studies, and White papers

Archiving Roadmap

During the Archiving Roadmap, clients are assisted to define Strategy approach and high-level planning with the help from key stakeholders from LOB's, IT, and process and strategy attend. The objective is to build a vision and high-level roadmap for an Archiving journey. Key activities are Archive Vision Presentation, Enterprise Archive Assessment, Transition Plan Development, and Business Benefits Evaluation. Key deliverables are Enterprise Archive Assessment, Transition Plan, Key Archive Initiatives, Business Benefits, and High level Reference Architecture.

Archiving Solution

During this phase, clients are guided to define the Archive Solution during a Workshop—a customer facing iterative engagement that identifies, defines and scopes an Archiving project; attendees are LOB/IT SME's and LOB/executive sponsors. Key activities define the solution project, model future state solution architecture, evaluate and select software, technical exploration, and business value assessment; and key deliverables are Solution Architecture, software demo, technical architecture, business value assessment, architecture plan, and a business proposal.

Use Cases

ABC company's CIO has been challenged by his CEO to reduce IT cost and improve information management regulatory compliance. The CIO engaged IBM Global Technology Services (IBM-GTS) to assist in developing an archive strategy to address the uncontrolled growth of data and improve ABC Company's regulatory compliance issues. ABC Company has had major issues in the past with complying with regulator auditor requests for information, either unable to locate the information in a timely manner or in not within the specified time required. Storage growth is outpacing business growth by more than 40%.

IBM partnered with ABC with a two part approach to address the challenges. During part one—briefing and roadmap—IBM assessed the current archive state and created the desired state or the Archive Strategy for ABC. As part of the Assess phase, IBM consultants worked with ABC's IT and Application developers to review and document the current environment. A data rationalization study was performed to identify valid data versus invalid data. Reviewed and document retention policies of their key largest storage consumer applications and identified data objects or "keys" that the customer will reference when retrieving the data from archive. Key deliverables were the retention policies by record/information type, the trigger conditions for archiving (or when the information type becomes static), and the speed of retrieval requirements. IBM then documented the recommended archive architecture to support the ILM requirements.

In the second phase—Solution—IBM used the newly defined Archive Architecture and designed the Archive Service including identifying hardware and software components and developing the processes and organizational roles and responsibilities in support of the enterprise architecture. IBM assisted ABC in establishing the ILM governance to enforce the policies previously defined.

The design included a comprehensive solution that addressed all types of data, structure, unstructured and e-mail. The design leveraged IBM's data management tools, Optim for structured, and Content Collector for unstructured data. It also leveraged the data storage and mobility capabilities of IBM's XIV Storage System and low-cost and content management capabilities of the DR550.

As a result of this partnership, ABC was able to realize a 50% reduction in online storage consumption, increase online application operations efficiency, reduce IT costs and improve regulatory compliance.

Conclusion

Today it is critical that data is archived in accordance with laws and regulations and is stored in a way that allows it to be retrieved effectively should it be needed.

In the current environments driven by strict regulations, ensuring the availability, retention, and the access to business information needs to be supported reliably and cost-effectively, which continues to increase the importance of enterprise archive solutions.

IBM's Reference Architecture for Archive enables the creation of long lasting enterprise architecture protecting the enterprise investment in software, hardware, and process and organizational support.

For more information

To learn more about the Data Archiving: Foundation Capabilities for Compliance and Cost Optimization, please contact your IBM marketing representative or IBM Business Partner, or visit the following Web site:
http://www-01.ibm.com/software/data/data-management/optim-solutions/data-growth.html

About the authors

Mike Cholewa
Managing Consultant, IT Storage Optimization
Mike is a Managing Consultant specializing in storage services within ITS. Mike has over 30 years of experience in IT and was a contributing member on the teams that developed IBM's ILM and TCO offerings. Mike has been with IBM for 9 years.

Niraj Desai
Strategy and Service Products Executive
Niraj Desai has 17+ years of industry experience, with a focus on developing and deploying services for enterprise customers.

Drew Peterson
Managing Consultant
Drew has 15 years experience in IT assisting customer's improve operational efficiencies. He is a key contributor to the development of IBM's Information Lifecycle Management service offerings.

IBM Systems and Technology Group
Route 100
Somers, NY 10589

Produced in the United States of America
April 2009

OIW03026-USEN-01