Solving the long term archiving and retrieval challenges with IBM Information Archive

White Paper

Nils Haustein, Consulting IT Specialist, IBM European Storage Competence Center

April 2011, Document Version 2.0

Table of Content

Acknowledgements:

I would like to thank Michael Murtagh, Sanjay Tripathi and Dinesh Patel from IBM as well as my good friend Nick Cvetkovic (Database consultant) for the thorough review and valuable hints in regard to proper English formulation, grammar and content.

Executive summary

In the backdrop of exponentially growing data, digital archiving is becoming more and more important. The ever growing volumes of data along with stringent regulatory and industry compliance requirements are the main drivers for long term archival solutions. Just using backup tapes to fulfill the requirements for archiving is not sufficient.

There are many unique challenges for digital archiving, the long retention periods being the most relevant. It causes technological obsolescence bearing the risk of loosing access to the archived data. Periodic technological refresh is required to assure that data and systems are migrated to newer technology while the authenticity and the readability of the data are being assured.

When it comes to the archive storage system, the durability of the storage medium is not the key question. For example, the value of a 20 year old tape containing archived data is very limited because there is no system available anymore that can read and interpret that data. However, the capability of the archive storage system to migrate data to newer storage technology helps to mitigate the risks associated with the long retention times and technology obsolescence.

Furthermore, storage systems with tiered storage capabilities are able to place the archived data on the most appropriate storage medium, which might change over time. For example, archived data is frequently accessed during the first two years where it is stored on a fast storage medium like disk. Later on the data is rarely accessed and can be migrated to a storage medium providing better operational cost such as tapes. Tiered storage capabilities help to optimize total cost of ownership.

There is no single solution which addresses all archiving requirements. Archiving solutions typically depend on the kind of data to be archived and the systems where this data is generated and managed. An archiving solution comprises a set of HW and SW components with numerous interfaces and protocols. Ensuring the interoperability of these mainly non-standardized interfaces and protocols is a key prerequisite for establishing a well designed archiving solution.

This paper gives a brief introduction to the motivations for digital archiving and describes the components of an archive system including Enterprise Content Management (ECM) system and archive storage. The reference architecture for an archive system is presented which makes clear that the intelligence of an archiving solution represented by functions for search and discovery, classification, retention management and business process management comes from Enterprise Content Management (ECM) system. The archive storage protects the archived data and manages the storage life cycle.

Archiving poses unique challenges to archive storage systems in terms of preventing modification and unauthorized deletion of data during the retention period. IBM Information Archive system serves as an example as to how these requirements can be met by virtue of using advanced innovative IT technologies.

Finally, the paper outlines how various IBM software and hardware components can be combined and optimized to achieve tailor-made arching solutions for different kind of data.

Digital archiving

The goal of archiving is to store data for long period of time and to provide an access to these data when needed within a reasonable amount of time. With the dramatic increase in digitization of information, digital archiving is rapidly gaining in importance and acceptance. People want to store digital photos, music and films for later generations. Companies are mandated to archive business records and documents for years, decades or in some cases even longer in order to comply with regulatory requirements

Digital archiving is motivated by three main reasons:

1. Regulatory requirements are a key reason for archiving data. They are typically country and industry specific and regulate how long data and information must be preserved. Companies doing business in certain countries and industries must adhere to the according regulations. For example, the US Securities and Exchange Commission (SEC) Regulations 17 CFR 240.17a-3 and 17 CFR 240.17a-4 stipulate the records retention requirements for the securities broker-dealer industry. According to these regulations, every member, broker and dealer is required to preserve records subject to rules 17a-3 for a period of up to 6 years.
2. Reducing costs by freeing expensive production storage and moving data which does not change anymore but still needs to be kept to less expensive archive storage is another important reason for archiving. Offloading data from production storage does not only reduce the amount of storage capacity but also decreases the efforts for maintaining and protecting production storage.
3. The preservation of information is another reason for archiving. The motivation to preserve information is manifold. For example, companies in the aerospace industry preserve design specifications to assist failure analysis even 20 years after an airplane has been developed. Patents are typically granted for period of 20 years and during this time the patent specification are preserved. Last but not least due to the increasing digitalization of information with digital cameras, intelligent mobile phone and other gadgets people have a growing desire to preserve digital information over a lifetime.

Archiving is a process in which information, and therefore data, which will usually not change and is to be retained for a long period of time, is transferred to and stored in an archive. The archiving process determines which data is to be archived when and where the best location is to store it. Essential for archiving is that the information and the data can be discovered, retrieved and made available for processing – otherwise the whole archiving process is a waste of time. In addition assurance is required that the data will be protected during the retention period according to the regulatory requirements.

An archive system is required to perform the archiving process and manage the archived data. In order to understand the complexity and the challenges associated with archiving the reference architecture for archive systems is introduced.

Reference architecture for archive systems

A three-layer architecture in figure 1, consisting of applications (layer 1), archive management (layer 2) and archive storage (layer 3), has proven to be effective in describing digital archive systems [2]. Each layer distinguishes itself by certain functions relevant to archiving.

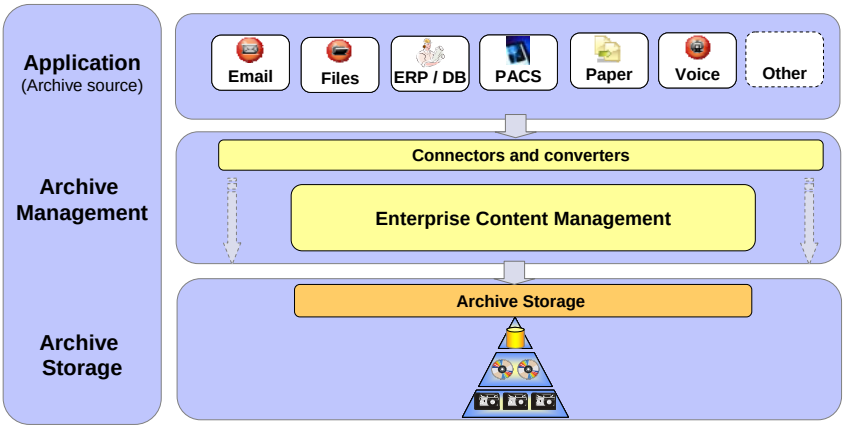

Figure 1: Reference architecture for archive systems

The top layer (layer 1) consists of applications that run on computer systems and generate, process and analyze information and then store this information in the form of data. These can be standard applications such as email systems or customized workplaces of users that generate, process and archive information, such as invoices, documents, drawings, damage reports and patent specifications. The applications in layer 1 communicate over an interface with archive management (layer 2).

The archive management in layer 2 is usually specialized software that runs on a separate computer system. Archive management is commonly referred to as 'Enterprise Content Management' system (ECM) and provides a number of functions which are important for archiving. The ECM typically includes connectors and converters which connect to the applications in layer 1, select the data to be archived based on rules and move it to the ECM system. Subsequently the ECM indexes, categorizes and classifies that data. The ECM also provides capabilities for search and discovery which enables the applications in layer 1 or ECM-specific discovery programs to find information even after many years. In addition the ECM can also link data and information from different applications and thus map complete business processes. For example, the ECM can combine and cross reference data from an Enterprise Resource Planning system (ERP system) with associated files from a file system. After processing the ECM transfers the data being archived to archive storage (layer 3).

Archive storage in layer 3 contains the storage media and provides functions for management of the data on the storage media. Data management functions ensure that the data is protected according to the requirements and that the data is stored on the most appropriate storage medium.

There are interfaces between each layer which are used for the communication between the components. The interfaces are represented by a protocol and a certain type of wiring. Unfortunately there are only a few standardized interfaces which make the implementation of archiving solutions more complex because it is dependent on the type of data and applications used in layer 1 as well as the ECM system used in layer 2.

The intelligence for the archiving process is represented by the archive management in layer 2 which provides policy based archiving, classification, indexing, discovery and search as well as business oriented workflows and processes. The archive storage in layer 3 is responsible for protecting the data and managing the data over its lifecycle.

Challenges for archiving

There are three key challenges for digital archiving:

- many country and industry specific regulatory requirements
- long retention periods
- unpredictable data growth

These challenges are discussed below.

Regulatory requirements

One key business challenge for archiving is the need to comply with laws and regulations. Failures to comply with laws and regulations are typically subject for large fines. For example a bank which is trading in the United States of America was unable to provide requested data to the SEC auditor because the data was stored on backup tapes and the index to the data on tape was not available. This bank received a fine amounting to more than 200 million US Dollar.

The first and perhaps hardest step is to figure out which laws and regulations are relevant for a company which does business in a certain industry and country. Laws and regulations are typically country specific, even in the European Union. For example, the US Securities and Exchange Commission (SEC) stipulate the records retention requirements for the securities broker-dealer industry. In addition there are industry-specific regulations, such as the 'Health Insurance Portability and Accountability Act (HIPAA)' which applies to pharmaceutical companies in the United States. A company in the United States – such as a pharmaceutical company traded on the New York stock exchange - may have to comply with multiple laws and regulations depending on the kind of data being dealt with. For example all the research data might have to be archived in accordance to HIPAA rules while all data relevant to securities trading and company stock has to be archived in compliance with the rules of the Securities and Exchange Commission (SEC).

Fortunately there is a set of common requirement which most of the laws and regulations specify:

- the type of data requiring archiving is specified, such as emails, files, invoices, report, etc
- the retention periods for different types of data
- data must be WORM (write once read many) protected to prevent manipulation or deletion during the retention period
- data must be accessible to an auditor at all times
- data access must only be allowed for authorized personnel
- copies of data must exist in order to prevent data loss in case of system failures
- upon expiration of the retention period data must be deleted
- systems and processes used for archiving data must be documented

The most important requirement is the last one: Thoroughly documented archive systems and processes allow discovering and closing gaps and provide proof for compliant archiving. Regulatory requirements must be implemented by the entire archive system as presented in the reference architecture in figure 1 (see section Reference architecture for archive systems).

Retention times

The legally required retention period for archived data is an eternity compared to how quickly technical, corporate and social changes take place. For example the retention period specified by the SEC is up to 6 years. In Germany companies have to retain trade and tax records between 6 and 10 years. In the pharmaceutical industry retention periods go beyond

50 years. The long retention periods challenge archive systems because the components in archive systems undergo technological progress and become obsolete over time.

The only practical way to cope with this challenge is the migration of systems and data to newer technologies [1]. Thus an archive system including all three layers must include migration capabilities.

Data growth and cost

The rapid growth of data combined with long retention periods creates another challenge to archive systems: costs. These costs apply not only to procurement but also to the operating costs of long periods of archiving. The general rule of thumb is that the procurement cost of an archive system amounts to 30% of the total cost for an operational period of 5–7 years. This means that 70% of the total cost must be applied towards the operation of the archive system. The operational costs include system administration and maintenance, power consumption, floor space and also costs for data and system migration.

Archive storage systems which address these challenges must implement specific functions which are discussed in the next section.

Archive storage requirements

In order to address the challenges associated with digital archiving the archive storage system must provide specific functions, such as:

- WORM protection of archived data
- disaster protection and recovery
- flexible placement of data based on age and access to data
- migration of data due to technological obsolescence
- standardized interfaces to the ECM system

WORM protection

One key regulatory requirement for archive storage systems is WORM protection (write once read many) which essentially demands that the data cannot be deleted or changed during the retention period which starts when the data is archived. WORM protection must not necessarily be enforced by the storage medium such as non-rewritable optical disks (CD, DVD). Instead it can be provided by control software (or microcode) embedded in an archive storage system. Such archive storage system can store the data on traditional hard disks which are rewritable and the control software provides logical WORM protection and prevents rewriting or deleting any data prior to the expiration date. One advantage of logical WORM protection is that the storage space can be reused once the old data has expired.

Data protection

The archive storage must provide capabilities for disaster protection and recovery to ensure that the data is readable during the retention period. One simple way to protect the data from a disaster is to backup the data. Backing up the data to tape is very cost efficient because the tape does not consume power as long as the data is not accessed. Another method for disaster protection in an archive storage environment is data replication to one or more remote archive storage systems. This additionally provides quick recovery times in case of a disaster but is more expensive.

Disaster protection and recovery must also maintain the authenticity of the archived data. Therefore it is necessary to provide ways for verification of the authenticity of the data, for example calculating and storing checksums when the data is archived.

Flexible data placement

The flexible placement of archived data on different types of storage media allows optimization of total cost of ownership. There is some archived data which is not accessed anymore but still has to be retained. Storing this data on disk for a long period of time increases the operational costs because disks consume power every second.

Tiered storage is one technique which allows moving aged data from an expensive but faster storage medium such as disk to a slower but less expensive storage medium such as tape. It facilitates quick access to archived data during the first couple of years after archiving. Later on, when the data is no longer accessed, it is migrated to tape which provides much better cost efficiency.

Data migration

The ability to migrate data when the storage medium becomes obsolete is another important requirement for archive storage systems. Just imagine the archived data is stored on old LTO-1 tapes which will eventually become obsolete. Again, the tiered storage functionality can be used for the automated and transparent migration of data from an old storage technology to a new storage technology. It is important that the migration does not influence the retention times and the authenticity of the data and that it is transparent allowing the ECM system to read the data at any time.

Standardized interfaces

Another important requirement for an archive storage system is the data interfaces used by the ECM in layer 2 to archive, query and retrieve data. Archive storage interfaces must be compatible with a wide range of ECM and they must be durable over a long period of time. Standardized interface meet these requirements because they typically last for very long time periods and are widely accepted and implemented by ECM systems. There are only a few standardized archive storage interfaces, for example NFS and XAM.

The Network File System (NFS) is over 20 years old and standardized by the Internet Engineering Task Force (IETF). NFS is a protocol for managing files in a remote file system. Thus an archive storage implementing the NFS interface acts as a NFS-Filer with some proprietary enhancements providing archiving specific functions such as WORM protection and retention periods. NFS is widely accepted in the market even though the archiving specific enhancements are not standardized.

The eXtensible Access Method (XAM) [6] is a newer standard which was published by the Storage Networking and Industry Association (SNIA) in 2008. XAM is an interface which is made for archive and incorporates archiving specific functions. XAM is new in the market and has not yet been widely accepted.

IBM Information Archive

This chapter gives an overview of the archive storage system IBM Information Archive and demonstrates how this innovative technology perfectly meets the requirements for long term archive storage [3].

IBM Information Archive (IBM IA), the next generation archive storage system, is designed as an archiving repository for structured and unstructured content. It helps organizations of any size to address complete information retention needs—business, legal, or regulatory. An IBM IA system is an appliance which can host up to three so-called document collections.

An IA document collection represents a physical partition for archived data and associated retention and storage management policies. An IA collection has a data interface providing either file system protocols such as NFS, CIFS and HTTP protocol or the TSM API protocol. Accordingly a collection providing the file system protocols is called a File Archive collection and the collection providing the TSM API protocol is called a SSAM collection. Each collection is comprised of dedicated storage system and one or more access nodes which provide the interface to the ECM application and perform the retention and storage management functions. Multiple collections within one IA appliance are configured as a high available cluster.

IA SSAM collection

An IA SSAM collection provides essentially the same retention and storage management functions as its predecessor product IBM DR550. The key component is the System Storage Archive Manager (SSAM) software which allows event-based and chronological retention policies, deletion hold and release and flexible storage management functions such as backup and migration. The SSAM server can also be configured for encryption, compression, data deduplication and data shredding.

An SSAM collection does not allow deletion or modification of any archived data during the retention period, when the retention period has expired the data will be automatically deleted from the storage system.

One particular highlight of the SSAM collection is that it supports the transparent data migration from a DR550 system. This allows DR550 customers to migrate data from the DR550 to the IA SSAM collection while the DR550 is concurrently used by the ECM application.

IA File Archive collection

The IA File Archive collection is a Network Attached Storage server (NAS) providing retention protection. Files can also be indexed based on the metadata and content of files. A search interface allows searching for files based on the index. An IA File Archive collection can export up to 500 file shares which can be accessed via the NFS, CIFS and HTTP protocol. File share users can be authenticated by a directory server such as LDAP or Microsoft® Active Directory™ or via a local user repository.

The File Archive collection can be configured in one of three protection modes:

- basic mode: allows deleting data and back-date retention periods
- intermediate mode: allows back-dating retention periods

- maximum mode: does not allow deleting data or back-dating retention periods, equivalent to the protection mode of an SSAM collection

A File Archive collection supports a variety of innovative retention management functions, such as:

- Auto-protect – files stored in the IA file share become automatically retention protected and the retention period is derived from preconfigured policies. Such policies associate the file based on its attributes such as name, size or owner to a retention period.
- Snaplock™ compatibility – the file attributes “last access time” and “read-only” are used by the application to manage the retention. The application sets the “last access time” to the date and time when the file should expire and commits the file subsequently to WORM status by setting the file to read-only.
- Metafiles – for each file which is stored in an IA file share the IA system automatically creates a metafile in a separate metafile share. The application can access these metafiles via the metafile share and manage the retention for each file using customized XML-tags. This allows very flexible retention management where the application can set fixed or variable retention periods on a per file basis, multiple deletion holds and releases and additional index information.

With the metafiles additional metadata can be added for each archived files. This metadata is protected and managed together with the file and can be leveraged with the optional indexing and search functionality.

The File Archive collection provides two pools for storing archived data which are linked by preconfigured policies. Both pools reside on the same physical storage system of the collection. The first pool is a file system where the files are placed initially by the user via the file system interface. Once a file becomes retention protected the integrated TSM HSM client automatically pre-migrates (copies) it to the second pool which is a storage pool in the integrated TSM HSM server. Depending on the TSM storage pool definition the file is stored on disk or on tape within the second pool in TSM.
After the file has been pre-migrated to the second pool two copies exist: one in the first pool and one in the second pool. The configurable migration threshold allows migrating files from the first pool to the second pool which will cause the file to be stubbed in the first pool. The stub provides transparent access for the user via the file system interface while the content of the file is stored in the second pool.
Within the second pool additional functions like compression, data deduplication and data shredding are supported. The key value-add of the integrated TSM HSM is server is that files can be placed on different storage technologies such as disk or tape. More precise files can be placed on tape initially or after a configurable amount of time. This enables the implementation of tiered storage and Information Lifecycle Management concepts.

Made for archiving

IBM IA is designed to help companies meet regulatory requirements, manage risk, reduce cost and improve operational efficiency.

IBM IA provides the following key functions:

- cost optimization through tiered storage function and data deduplication
- build in migration functions
- standardized interfaces for easy integration with ECM systems
- assessed compliance through logical WORM protection
- high level protection through integrated backup functions and optional replication feature

These key functions are now being discussed.

Compliance

Independent auditors have assessed IBM IA for regulatory compliance according to laws and regulations of multiple countries and industries [4], [5]. The flexible retention management functions allow seamless integration with retention models of ECM applications. Additionally IBM IA generates and retains audit logs which track all file-level operations. Audit logs can be accessed by distinct auditor user groups.

Protection

IBM IA provides different methods for data protection and disaster recovery. Data is protected by RAID-6 on the internal disk system allowing for multiple disk failures in one RAID-array. In addition data can be backed up to tape which provides additional data protection and data recovery in case the internal disk system experiences a failure. For enterprise-level protection IA offers consistent mirroring. Thereby the archived data is mirrored synchronously or asynchronously between two IA systems. If one IA system fails the other system can be seamlessly activated.

Additionally IBM IA can be configured to calculate checksums of the data being archived which can subsequently be used for verification of the authenticity of the data.

Cost optimization

Managing cost for the ever growing volumes of archived data is also addressed by IBM IA through tiered storage functions. Tiered storage allows flexible placement of data on the most appropriate storage medium. IBM IA allows setting up different tiers of storage, for example one tier on disk providing fast data access and one tier on tape providing cost efficiency. Tiered storage policies can be configured which cause automated migration of data from one tier to the other. For example a policy can be defined which migrates all data that is one year old or has not been accessed for one year from tier 1 to tier 2. At the same time backup copies of the data can be retained for disaster recovery.

Data deduplication is another technique which helps to optimize storage cost. The data deduplication function in IA identifies identical data blocks within archived data and stores these only once. Data de-duplication can be configured selectively on a storage pool basis allowing only subsets of data to be deduplicated.

Data migration

One key requirement for archived data which has to be kept for long periods of time is the periodic migration to new storage technologies. This function is integrated in IA whereby the data migration is transparent to the ECM application which improves operational efficiency. Additional checksums can be used to verify the authenticity of the data after the migration.

With the IA migration functions it is also possible to refresh the data by moving it from one storage medium to another. This way the readability of the data can be assured over a long period of time.

IBM IA also supports the migration of all data from the predecessor product IBM DR550. With this migration technique all the data archived in IBM DR550 system can be copied into IBM IA while the DR550 is used for archiving. When all data has been copied to IBM IA the ECM application can be switched to IBM IA and continue to work seamlessly.

Operational efficiency

Operational efficiency is important because an archive storage system is typically operated over a long period of time. IBM IA addresses this with an integrated graphical user interface (GUI) and a command line interface (CLI). The IA GUI and CLI allow centralized administration and monitoring of up to three collections of any type. The IA management software is connected to all intrinsic hardware and software components and generates customizable alerts when failures occur.

IA includes self-healing functions. All IA components are configured redundantly and if one component fails another component will take over. This dramatically decreases the length of outages.

Standardized interfaces

IBM IA provides the standardized interfaces such as NFS and CIFS with its File Archive collection. This allows easy integration with ECM applications because IBM IA is like a file system to the ECM system. The flexible IA retention management functions on top of the file system protocols satisfy the majority of retention management scenarios of ECM systems.

Indexing and search

The IA File Archive collection allows to index files based on content and metadata and provides a user-interface to search for files. Indexing is an optional IA feature and can be configured for file-content, file-metadata or both. File-metadata includes attributes of the file such as owner, time stamps and permission and IA metafiles.
The key value of indexing and search comes with archiving solutions where the application does not provide this functionality. For example when one or more users store files directly via the IA file system interface the index and search functionality can be used to find files even after many years.

Archiving solutions

Archiving solutions are typically designed according to the reference architecture presented in the section Reference architecture for archive systems. The design of archiving solutions is driven by the kind of data being archived and the application system (layer 1) where the data is generated and managed. There is not a single archiving solution which can be used for all kind of data such as e-mails, files, ERP data, database records and tables, voice streams and web-content. Instead the archiving capabilities of the application system in layer 1 have to be assessed in order to find a viable archiving solution.

One of the difficulties today is the interface of the applications systems in layer 1. There are no standards that are generally accepted. Thus an email-server has a different interface for archiving than an ERP system. It is even more complex: different email servers – such as Microsoft® Exchange® and Lotus Domino™ - have different interfaces for archiving. In order to cope with this challenge connectors and converters are required (layer 2) which adapt to the interface of the application and execute the archiving process by transferring the data to be archived to the ECM system or directly to the archive storage (see figure 1).

IBM has a wide suite of products which are capable of archiving any kind of data. The next sections describe three archiving solutions for email, ERP and file and highlight the functions of the different IBM software and hardware products.

Email archiving

Companies could not survive without email. On the other hand, email data often contain business-relevant information, such as quotations, invoices and contracts, which for regulatory reasons must be archived. Another motivation for archiving email is the constant data growth in email systems. Archiving old email reduces the amount of data in the email server, and, consequently, the administrative effort required.

The popular email systems today such as IBM Lotus Domino™ and Microsoft® Exchange must use archive systems to achieve the regulatory compliance required for the long-term storage of business-relevant information. IBM provides a complete suite of products for email archiving as shown in Figure 2.

IBM Infosphere Content Collector for email (ICC for email, top of layer 2) acts as a connector to the email server (layer 1) and captures email to be archived based on preconfigured rules such as age and size of the email. After moving email to the ECM system, ICC for email places a reference in the mailbox. This way an archived email consumes only a few bytes of capacity in the mailbox but the user can easily access the archived email via the reference. During the archiving process ICC for email provides additional value added functions - such as single instancing (which is a form of de-duplication for email and attachments), classification or encryption.

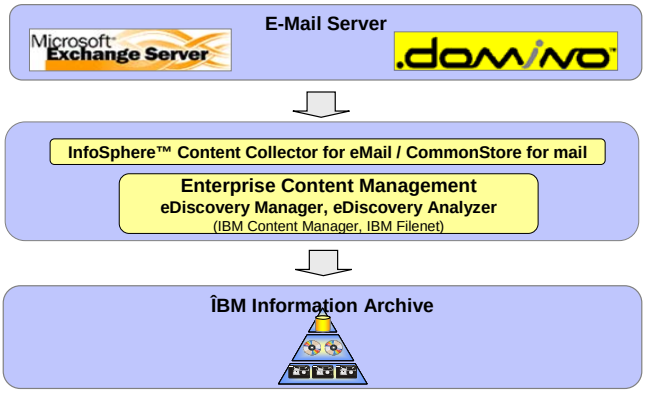

Figure 2: IBM email archiving solution

The IBM ECM system can be either IBM Content Manager or IBM Filenet P8. It indexes the email and provides subsequent search and discovery functions. After indexing email the IBM ECM system stores it in an IBM IA SSAM collection and subsequently manages the retention time.

The IA SSAM collection provides secure storage which does not allow deletions or modifications. The IBM ECM system controls the retention period and signals deletion via the event-based retention policy. Beside compliance IBM IA provides storage management functions providing backup and / or migration of the archived email to tape (see section IA – the perfect archive storage). IBM IA can also be configured in a mirrored environment allowing instant disaster protection and recovery.

ERP archiving

Enterprise Resource Planning (ERP) systems such as SAP® help companies to use and manage existing resources (capital, equipment and personnel) efficiently. These systems contain large quantities of data, such as invoices, correspondence and logistical information, that must be retained according to regulatory requirements. The huge growth in data quantities in these systems is another reason for transferring old information from production ERP systems to archive storage and releasing storage space in the production system.

SAP® is the most popular ERP system and provides its own interfaces for archiving. The older interface is ArchiveLink® and the newer one is based on WebDAV. IBM provides a complete suite of products for ERP archiving, including SAP, as shown in Figure 3.

IBM CommonStore for SAP (top of layer 2) acts as a connector to the SAP server (layer 1), captures the data to be archived and transfers it to the IBM ECM system. Alternatively CommonStore for SAP can be configured to directly interface to the IA SSAM collection. The SAP server keeps an index and reference to the archived data which can be used for retrieval.

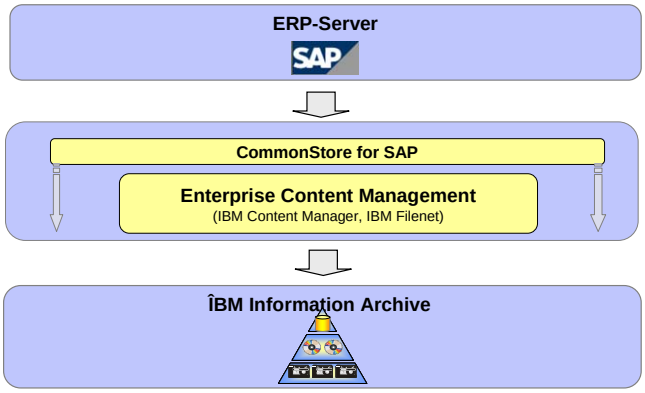

Figure 3: IBM SAP archiving solution

The IBM ECM system can be either IBM Content Manager or IBM Filenet P8 and can be configured to index the ERP data when this is required. Indexing enables subsequent search and discovery functions. The IBM ECM systems also provide records management and business process management functions which are seldom used in the context of ERP archiving. The archived ERP data is stored in the IBM IA SSAM collection.

Similar to email archiving (see section Email archiving) IBM IA provides secured, protected and tiered storage and allows the IBM ECM system or IBM CommonStore to manage the retention period and signals deletion via the event-based retention policy.

File archiving

Companies use files to store important information such as contracts, correspondence letters, product specifications, drawings and marketing presentations. Many types of files are therefore subject to regulatory requirements that make file archiving inevitable. The enormous data growth in file systems and the resulting effort required for management and backup are other important reasons for archiving files.

Files are normally stored in a file system that is located either locally on an applications computer or remotely over the network (in layer 1). An archiving solution for files must be able to identify files to be archived regardless where they reside and move selected files to the archive system while still allowing access to the file even after it has been archived. IBM provides a complete suite of products for file archiving as shown in Figure 4.

IBM Infosphere Content Collector for File (ICC for File, top of layer 2) connects to the file system (layer 1) and captures the files to be archived based on preconfigured policies such as age, size, date of last access or type of the file. When capturing the files to be archived ICC for File can be configured to leave a reference to the archived files in the file system. This way an archived file consumes only a few bytes of capacity in the file system but the user can still access the archived file via the reference. During the archiving process ICC for File provides additional value added functions - such as single instancing (which is a form of de-duplication), classification or encryption – before it transfers the file to the IBM ECM system.

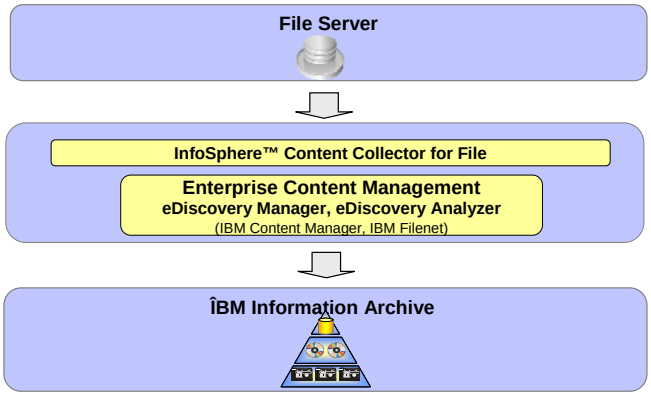

Figure 4: IBM file archiving solution

The IBM ECM system can be either IBM Content Manager or IBM Filenet P8. It indexes the files and provides subsequent search and discovery functions. The IBM ECM systems also support records management and business process management functions which are important for compliance archiving. After indexing the files the IBM ECM system stores it in an IBM IA SSAM collection and subsequently manages the retention time.

The IA SSAM collection provides compliant, protected and tiered storage which does not allow deletions or modifications. The IBM ECM system controls the retention period and signals deletion via the event-based retention policy.

With the indexing and search function provided by the IA File Archive collection simple file archiving solutions can established using tools available in many operating systems. Additional metadata can be added through the metafiles. Indexing and search can be used to find files even years after archiving them.

Conclusion

When designing archiving solutions all kind of data should be considered with the goal to establish one archive system for different kind of data instead of multiple archiving islands. One archive system may include different connectors for different sources of data, but the ECM system and the archive storage can be common. This saves operational and administrative costs and makes migration much easier.

Because of the challenges and the risks associated with long term archiving it might be appropriate look for a trusted and competent vendor when setting up archiving solutions. The archiving solutions which have been presented demonstrate that IBM can offer complete solutions for long term archiving independent of the kind of data including hardware, software and services.

Another option to minimize the risk associated with archiving is to engage experienced vendors take care for implementing and operating an archive system. Professional archive cloud offerings help to save costs and reduce the risks for a company.

Appendix

References

[1] “100 Year Archive Requirement Survey”; SNIA Data Management Forum; January 2007

[2] “Storage Networks Explained”; Troppens, Erkens, Mueller, Haustein, Wolafka; Wiley 2009

[3] IBM Information Archive web-site:
http://www-03.ibm.com/systems/storage/disk/archive/index.html

[4] “Bericht über die Prüfung der IBM Information Archive Speicherlösung Version 1.1”; KPMG AG Wirtschaftsprüfungsgesellschaft; März 2010

[5] “SEC 17a-4(f) Compliance Assessment for IBM Information Archive”; Cohasset Associates, Inc.; December 2009

[6] Extensible Access Method:
http://www.snia.org/tech_activities/standards/curr_standards/xam/

Disclaimer

This document reflects the understanding of author on many of the questions asked about archiving solutions with IBM hardware and software. This document is presented “As-Is” and IBM does not assume responsibility for the statements expressed herein. It reflects the opinions of the author. These opinions are based on several years of joint work with the IBM Systems group. If you have questions about the contents of this document, please direct them to the Author (nils_haustein@de.ibm.com).

The Techdocs information, tools and documentation ("Materials") are being provided to IBM Business Partners to assist them with customer installations. Such Materials are provided by IBM on an "as-is" basis. IBM makes no representations or warranties regarding these Materials and does not provide any guarantee or assurance that the use of such Materials will result in a successful customer installation. These Materials may only be used by authorized IBM Business Partners for installation of IBM products and otherwise in compliance with the IBM Business Partner Agreement.”

Trademarks

The following terms are trademarks or registered trademarks of the IBM Corporation in the United States or other countries or both: the e-business logo, IBM, system x, system p, System Storage.

SnapLock™ is a registered trademark of Network Appliance Corporation in the United States.
Microsoft is a registered trademark of Microsoft Inc. in the United States.
SAP is and SAP ArchiveLink are registered trademarks of SAP AG in Germany and other countries.
Other company, product, and service names may be trademarks or service marks of others.