Data Management Forum

Archiving and Tiered Storage for Email

A DMF ILMI Best Practices White Paper

October 1, 2004

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION

Author:
Aloke Guha, Chief Technology Officer, COPAN Systems
Co-Chair, SNIA ILM Technical Work Group

Table of Contents

Table of Figures

1. Introduction

The SNIA definition for ILM:

> The policies, processes, practices, services and tools used to align the business value of information with the most appropriate and cost-effective infrastructure from the time information is created through its final disposition. Information is aligned with business requirements through management policies and service levels associated with applications, metadata and data.

This definition includes how data is archived and maintained within the IT infrastructure. Specifically, this paper considers archiving for a horizontal application, email.

This white paper is intended to provide examples of archiving application data on different tiers of storage so as to ensure the most appropriate and cost-effective infrastructure at any time of existence of the data. It generalizes on how application data should be archived using best practices so that data that is meant for long-term preservation is still accessible when required by the application.

This paper is one in a series of white papers that include Data Recovery [1], Security [2] and Archiving, produced by the SNIA Data Management Forum's ILM Initiative specifically for IT Administrators as the intended audience. It is written with the intent of providing IT Administrators with usable application-specific guidelines on deploying ILM solutions using today's technologies

Each white paper defines best practices for a specific dimension of ILM solutions. In this paper, the topic is data archiving for general email applications such as Microsoft Exchange, Lotus Notes, and Unix Sendmail.

1.1. Purpose and Scope

This paper represents an attempt to formalize the general definition of archive as well as its management. Because there have been many confusing definitions of what constitutes an archiving process, different approaches to creating archives and how archived data is accessed are included. Historically, hierarchical storage management (HSM) techniques have been used to create archives on tiered storage. This paper therefore contrasts archive and hierarchically managed storage

With increasing recent interest in maintaining corporate and business data for regulatory reasons, new requirements are being imposed on archived data. This paper therefore elaborates on the storage, data and information management needs of archives. In addition, new emerging storage technologies and approaches to managing storage, and appropriate best practices using these technologies are also presented.

1.1.1. Storage Environment

This paper outlines how archive uses the storage infrastructure below the application level and the capabilities required for purposes of archiving on the storage infrastructure. Much of today's archive solutions consider different storage subsystems, such as to DAS, SAN, NAS and CAS This paper focuses on solutions leveraging the benefits of networked storage. Unless otherwise noted, these apply equally to SAN and NAS-based solutions.

Given the nature and access of archival storage, specifically, scale and cost, many archive storage solutions are comprised of a combination of different storage media, both disk and tape. As will be explained, the creation and management of the archive

requires data management practices that extend beyond simple storage space management.

1.1.2. Email

This paper provides archive best practices that may be applied to several different email server implementations. While this paper addresses general approaches to email archiving, approaches that are specific to certain email applications are covered elsewhere in application notes. For purposes of illustration, examples are cited from Microsoft Exchange 5.5, 2000, and 2003[1], Lotus Domino Server version 5, 6 and 6.5[2], and SMTP mail servers such as BSD Unix Sendmail[3][4].

1.2. Defining Data Archive and Archiving

1.2.1. Data Archive

An archive is a collection of data that is maintained as a long-term record of a business, an application, or an information state. Archives are typically kept for auditing, regulatory, analysis or reference purposes rather than for application or data recovery.

1.2.1.1. Characterization of Archives

All archives are not created equal.

The storage infrastructure and the management of an archive depend on the activity level on the archive, i.e., the frequency of access to the archive. For example, if the archived data is not expected to be read except in a contingency situation, such as tax records that are referenced only in the case of an audit, it can be maintained on a storage device such as removable media. However, data integrity on the media even for such infrequently accessed archived data is still important.

Active Archives: unlike deep vault storage, many data archives have need for frequent access to data, perhaps many times a day. We refer to these archives as “Active Archives”. Examples of active archives include reference library data, physical process simulation results, multimedia content, seismic data processing, etc. Applications using active archives cannot tolerate the long access times, and therefore, require storage solutions with fast response times and data integrity that satisfies cost constraints.

Deep Archives: deep archives are long-term vaulted data that have very infrequent access. Unlike active archives, they are more tolerant of long access delays.

Retention: another distinguishing characteristic of archives is the lifetime of the archive. This can vary greatly, from a few years to, theoretically, infinity.

Implications of Long Term Archives: one of the most serious challenges of maintaining and managing a long term archive, say a 100-year archive, is how the data is reliably maintained far beyond the useful life of the original IT infrastructure that created and supported it. This infrastructure includes the storage and the access components that encompass the network and processing hardware and software. In fact, the archive may span 20 or more turns of technology changes.

[1] http://www.microsoft.com/exchange/
[2] http://www.lotus.com/products/product4.nsf/wdocs/dominohomepage
[3] http://www.sendmail.org/faq/
[4] http://www.milter.org/

1.2.2. Archiving Process

The description that follows applies to general archiving and not to email archiving alone. It is provided as essential background material for discussion of email archiving.

Historically, the archive process consists of creating a data archive through a copy or move operation of the data for purposes of retention. In the copy-based approach, the data is copied to the target, usually a lower performance lower-cost, secondary storage system[5], and then maybe deleted from the primary storage location. In the move-based approach, the data is moved to the secondary storage target and links or references to the data are either maintained or deleted. Both options are possible in practice but result in requiring different access mechanisms to the archived data, and are therefore worth further discussion and analysis.

1.2.2.1. Referential Links to Moved Data: Hierarchical Storage Management of Archive Data

When the referential links to the moved data are not deleted, then the original application *can* access the data transparently, whether it is on the primary storage system or on the secondary storage system. In this case, a hierarchical storage management (HSM) system has been created for the original application.

By itself, the HSM system does not provide an archive solution because it does not guarantee retention of the data and because the application or its users can modify or delete the data independent of any external control.

An archive is created, if and only if, there is a mechanism to control the retention of the data independent of the application. Thus, an archive generation requires *an archive management function that guarantees the existence and integrity of the archive independent of the application*. Both the application and the archive management access the same data in their respective namespaces but with different access control rights to the data.

1.2.2.2. No Referential Links to Moved Data: Application-Independent Access to Archive Data

When the referential links to the moved data are deleted, then the application *cannot* access the data without the aid of the archive management function. In this case, all access to the archived data requires requests to the archive management. The application does not have visibility to the archive data.

1.3. Comparison of Archive Management and HSM

It is important to note that the two archive approaches described earlier have been and are in use. Since an archive can reside on a single storage platform, it is not necessary that an archive must use tiered storage. However, for economic reasons, it is common to move the data when archived to a more cost-effective storage platform distinct from the primary storage on which it was first created. This is common between practical archiving and HSM.

The goal of an archive is to retain data with integrity independent of then application that created it. HSM is used to leverage the cost benefits of lower cost storage platforms. The distinctions between archive management and HSM data are summarized in the

[5] When we refer to a storage system that is used in conjunction with the primary storage systems and is typically lower performance and lower cost than the primary storage, then we will refer to is as the secondary storage system.

table in Figure 1**.** Note that while there are HSM products that may provide some archive management functions, this paper discussion and refers to HSM as a practice.

Attributes	Archive Management	HSM
Access Method	Application can access data directly or indirectly through Archive Management	Application can access data directly. HSM is transparent to Application
Access Control	Only read access by application; application cannot modify or delete records	No limitations
Data Immutability	Guaranteed during the retention period	Not guaranteed
State of Data	Data is not in operational state but used for reference	Data is in operational state
Data Copies	Archive Management may maintain a second copy of the application data. The data archived may exist under Application control	Only one instance of data is maintained under control of the Application
Use of Tiered Storage	Can use tiered storage for cost-effective retention but not necessary	Uses tiered storage
Management Function	Manages retention, access and integrity of data, usually on tiered storage, usually set by policy	Manages transparent migration of data between tiers, usually set by policy

Figure 1. Comparison of Archive Management and HSM

The archive management function can be provided as an independent software function, within and in conjunction with the archive storage system, or as extensions to the original application.

In the remainder of this paper, we will focus on the archive management system, how archives are created and managed, recommendations on what attributes of archives need to be supported, and how email archiving is accomplished. Further, since most archives are maintained on a tier of storage distinct from where it was created, we also discuss the tiered storage architectures that are used for archiving.

1.4. The Role of Archiving in ILM

Archiving is an important aspect of an overall ILM strategy within the data center. Here we consider the ILM perspectives on data archiving for email applications.

Because ILM advocates the use of the most appropriate and cost-effective infrastructure, the area of archiving is a direct manifestation of ILM in practice.

Archiving specific email records[6] implies that a certain set of records are not expected to be accessed frequently, and therefore, it is appropriate to locate those records, by

[6] In this paper, for simplicity, email records and email messages and the associated attachments are used interchangeably.

whatever means is most effective, on lower cost and lower access performance storage devices when different tiered storage is available. ILM is usually associated with the use of tiered storage, and therefore relies on migration techniques and tools used in HSM to move data between the storage tiers. By moving a portion of the email records to lower performance storage, better utilization of primary storage is possible across all active data applications. Another concomitant benefit is that by reducing the size of the email records under the email application's native database, in the case of Microsoft Exchange and Lotus Domino, the performance of the email server will also be improved.

There are therefore a number of areas of consideration in the archiving process within the context of ILM:

1) The use of tiered storage that provides variable performance, scale of storage and cost options.
2) The archival process that includes the selection of email records, set by policy, that need to be moved from one storage tier to another, and the process that moves email records between tiers of storage
3) The management of the archived email, including the retention, access, security and integrity.

We note that HSM and Archive Management share the first two properties listed above. However, as noted earlier, data that is moved for creation of an archive may actually be a copy of the data still under application control, unlike in the HSM case where there is only one copy of the data.

Data migration for archiving can be driven by some of the same criteria as HSM. However, archiving may also be driven by additional criteria, such as those specified by compliance needs.

Archiving is not intended for protection of primary operational data. In many cases, the archive may be the only instance of the original data. Therefore, it is a requirement to protect the archive data and metadata. More discussion on protecting the archive is provided in Section 4 on Archive Management Attributes.

2. Storage Architectures for Archiving

A key aspect to archiving for long-term data is the use of tiered storage. Tiering refers to use of storage systems of different performance, scale and cost. It is not necessary that archive data must use tiered storage. However, since archive data is a growing data asset and not frequently accessed as operational data, use of tiers of storage is usually preferred to reduce cost of storing and managing the archive. This aspect of using tiered storage is common with HSM (Section 1.3). Beyond HSM, archive management has many other issues to consider.

The following characterize archive storage architectures:

- The first tier uses a high performance disk storage system for operational data access
- A secondary tier is used to create a lower overall cost per unit data with different performance characteristics, as well as highly scalable storage capacity. A tertiary tier may also be used if better performance versus capacity are possible
- Archive management that includes retention, compliance, and security must be maintained for the archive across all tiers

Given the plethora of storage systems, there are many different storage tier combinations that can be created. These include primary disk (e.g., Fibre Channel), secondary disk that uses lower cost lower-performance (e.g., ATA) disk systems of different scale, size and cost, and tape systems such as automated tape libraries.

Different combinations of tiered storage are required when considering different criteria which include cost, scale, performance, compliance support, etc. From a connectivity perspective, we broadly classify tiered storage into two categories: two-tier and three-tier architectures

As mentioned earlier, these tiered architectures apply to both archiving as well as HSM.

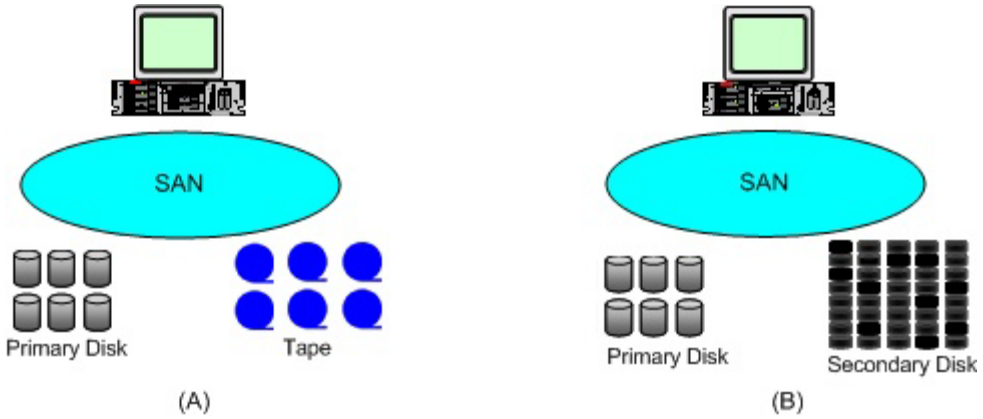

Figure 2. Two-Tier Storage Architecture

2.1. Two-Tier Storage Architecture

Figure 2 shows two instances of two-tiered storage architecture: disk to tape (D2T), and disk to disk (D2D) where the second disk can be implemented using different technologies such a disk arrays, NAS, CAS (content addressable storage) or MAID (massive array of idle disks)..

Disk to Tape (D2T) is the traditional tiered storage architecture, as in historical HSM (Figure 1A). The bulk of the data is maintained on the tape, which might comprise an automated tape library and vaulted tape media. The D2T architecture is considered most suitable where long access times to retrieve archive data are acceptable, i.e., very low activity archives. Tape also provides a high density of storage in footprint terms and is usually the lowest cost media compared to traditional disk storage.

Disk to Disk (D2D) is an emerging storage architecture (Figure 1B). It allows much higher performance than D2T. Its cost is usually higher than tape but lower than enterprise disk. There are a number of possible disk-based appliances that can be used as the secondary disk tier. These include:

- Using standard RAID array for fast access at the block level
- Using NAS for file-based access
- Using CAS to access data by content
- Using MAID storage [4,5] that can be accessed in any of the above presentations, disk, NAS or CAS, as well as virtual tape

Besides differences in presentation, each of the disk-based appliances provides different performance, cost, scale and accessibility features. Regardless of implementation, D2D is the preferred tiered storage architecture for active archives because of inherent performance advantages over tape.

Best Practice Considerations

Two-tiered storage architectures provide a simpler data management architecture, compared to the three-tiered model, requiring the fewest data movements. The choice of whether to use tape or disk is dictated by needs of cost and performance.

Archive management for retention and compliance is a feature orthogonal to the tiered storage architecture, and needs to be provided by an archive management function that can be either embedded external to the storage or can be associated with the tiered storage system. More detail on the functionality of the Archive Manager is provided in Section 5.

2.2. Three-Tier Storage Architecture

Figure 3 shows a typical instance of the three-tiered storage architecture: disk to staging disk to tape (D2D2T). As before, the first tier enterprise disk is for operational data. The motivation for using two secondary storage tiers is to more cost-effectively scale storage capacity.

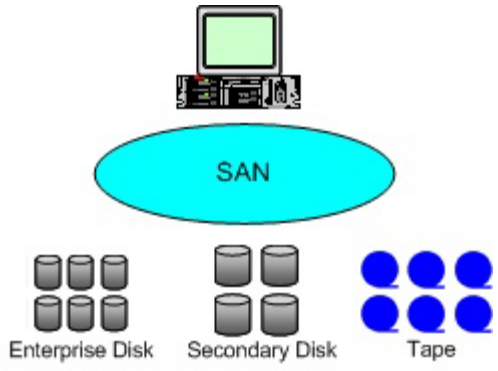

Figure 3. Three-Tier Storage Architecture

The most common three-tier architecture uses secondary disk for an intermediate stage to tape. Data placed on the secondary disk can be accessed faster than from tape, if the retrieval is expected to be from the most recently archived data. This configuration is also typically used for backup and recovery storage where there is evidence of temporal locality in restores. Because the scale, capacity and cost of the staging disk are not competitive with that of tape, when cost is a consideration, bulk of the archive is kept on tape. Periodic migration of data from the staging disk to the tape is therefore required. The duration of the data residence on the staging disk is dictated by the application, and can range from a few weeks few months. In the case of active archives, the temporal locality in data retrieval can be low and therefore limit the effectiveness of staging disk [7].

Best Practice Considerations

Three-tiered storage is recommended when large scale archives cannot be implemented cost-effectively with two-tiered storage architectures. The use of staging disk storage is effective if there is evidence that retrieval from the archive has reasonable temporal locality.

As in the D2D two-tier architecture, the secondary disk can be one of many appliances such as storage arrays or NAS. CAS is not used as a staging disk but rather as the last tier of storage given its specific support for fixed content. MAID storage is not used as a staging disk since its acquisition cost and scale is competitive with tape [6].

As in the two-tier storage case, archive management for retention and compliance is orthogonal to the tiered storage architecture, and is provided by an Archive Manager described in Section 5.

3. Archival Process

The decision to archive will require identifying what email records need to be archived and when.

The selection of what email records to archive can be usually set by policy [6]. The policy can be specified by:

1) Archive elements: only specified mailboxes, email messages, attachments, folders or directories, or PST files (in the case of Microsoft Exchange) are archived. The specified data that is to be archived will be defined by importance assigned by the organization. This will include journal-based archiving where all incoming and outgoing emails are journaled by the email server locally, especially in case of Microsoft Exchange that can automatically journal all emails. In this case, the journal is moved to a designated storage tier.
2) Age: email messages that are older than a prescribed age are archived
3) Frequency: periodic archive process initiation such as running the archive process on a weekly basis
4) Size: the archive policy might be to only archive email records above a certain size

The process used to archive email records can use either a move or a copy and delete process described earlier in Section 1.2.2.

3.1.1. Archive Management

Once the email archive has been created, there are number of elements of archive management that must be supported.

3.1.1.1. Archive Storage Management

Storage management for archives on tiered storage results in the complexity of continuously monitoring the usage and capacity planning of multiple storage systems that comprise the archive. When email storage in the first tier exceeds a specified threshold, emails satisfying archiving criteria are moved off to secondary storage tiers. The capacity for the lower tier storage is continually increased.

The total storage capacity is a function of the retention period for the email, which is dictated by regulatory specifications or policies set by the organization.

The biggest challenge in the storage management of email archives is that the storage technology used may be obsolete before the expiration of the retention period.

Best practices in archiving must therefore ensure that the storage system is available and reliable during the duration of the archive life. All retained copies of the data must be tracked and managed until the archive is disposed or deleted. There are a number of considerations to meet this requirement. These are described in Section 4.1.7 on Storage and Data Resource Management.

3.1.1.2. Archive Data Management

There are two key requirements in archive data management:

1) Movement of the data between storage tiers, and

2) Protecting the archive

These requirements are critical since as the email storage grows daily, selected email records need to be continually removed, or copied if the records are required to be in operational use, from primary storage and added to the secondary storage tiers. We note that if the operational copy of the email is changed after it has been created and it has been archived the first time, then changes made to old email messages in individual mailboxes cannot be captured in an archive unless they are maintained in the email server for the next archiving pass.

Conversely, when an archived email record(s) is (are) requested, the email data has to be moved from the secondary tiers to the Email Server. It needs to be imported by the email application in case the access to the archived email is only through the archive management system.

The latter requirement of protecting the archive is also non-trivial. Since the archive is the primary copy of the email, data protection requires creating a backup or replicated copy of the archived email.

Another aspect of the data protection of the email archive is to ensure that the data integrity of the email is maintained.

3.1.1.3. Archive Information Management

Archive information management extends the capabilities of archive data management to enable as well as support content management. There are two areas that may be provided:

1) Content-driven movement of data between storage tiers, specifically from the production storage to the archive storage tiers. This is usually driven by compliance needs that may dictate that email messages on specific topics need to be archived.

2) Support for fast and efficient search of email messages using automated indexing or content-based keywords.

In terms of best practices of archive information management, use of disk systems is preferred since indexing and searching on the content on disk archives, especially with the use of CAS, will be far superior to that on sequential media such as tape. The same guideline applies to content-driven movement of data between storage tiers, i.e., disk based storage is preferred over tape.

More details on different requirements for managing the email archive are provided next.

4. Archive Management Attributes

Archiving is driven by many needs that include internal corporate records maintenance and governance, audit trails for corporate transactions, regulatory needs, as well as for reference purposes, for example, data kept for long-term public access such as in the Library of Congress[7].

There are two characteristics of archived data as described earlier:

1) Long-term retention – the retention period can vary significantly depending on the nature of the archive data. More importantly, the data needs to be maintained after it is produced for future reference with assurance that the original data has not been changed.

2) Infrequent access – typical archive data is accessed less frequently than transactional or operational data, i.e., the data is not used for day to day business operations. The level of access or activity on the archive can vary with the nature of the archive.

These and a number of other archive management attributes such as performance of access, data integrity, compliance and security, described earlier in Section 1.3.3 define specific needs of archives. Support for managing these attributes constitutes best practices in archiving.

4.1.1. Retention and Disposition

Retention specifies the duration for which the email records are maintained. The specific challenges that this imposes on storage and data management are:

1) Retention beyond storage device life: the life of the archive can easily surpass the life of the storage media. This requires not only that the email records are guaranteed to exist but also maintained with data integrity, even as the storage devices fail or suffer obsolescence when the underlying technology is no longer supported by the vendor.

2) Retention beyond application compatibility: the life of the archive can also surpass compatibility with the original application; thus, there may be many versions of the email application that may span the life of the archive. The archive management should be able to guarantee that the current email application version can read the email records created from a version many generations old.

In terms of best practices, the above requirements place serious constraints on maintaining the archive. Historically, non-electronic media, such as paper or film have been used to store archive data. While lacking automation in management and access, they had two fundamental advantages: i) paper media can last hundreds of years under reasonable storage while lifetime of tape is twenty years or less (disk has a even shorter life), ii) image or printed text can be read without reliance on any software.

The primary requirement for archive retention is the need to refresh the data without loss of data integrity on new storage technology, when the retention period exceeds the life of the existing storage media.

[7] http://www.loc.gov

The above requirement assumes that the application that created the archive is available and supported during the lifetime of the archive. It does not guarantee that the archive is readable by the end user. There are few alternative approaches that can be used to solve such a long-term retention problem, although they face practical limitations as well:

1) The operating environment, comprising the application, the operating system and the computing platform, are escrowed so that the archive is still accessible to end users.
2) The archive data is maintained in application-independent standard formats, such as CSV and XML, that are accessible and readable over the life of the archive. This would be advised if the original application program cannot be guaranteed to exist over the lifetime of the archive.

Compliance requirements specify how the archived email is to be disposed of at the end of retention period. Some regulations demand that all copies of the email must be verifiably destroyed. Disposition therefore imposes requirements on archive storage and data management, in terms of ensuring all copies of the email are tracked and monitored and the media is erased or destroyed, especially in the case of removable media and copies.

4.1.2. Performance

There are two dimensions of performance: access time for retrieval and data rate at which the archived email is retrieved.

Although most archive data is not accessed often, i.e., a small fraction of the archive is accessed at any one time, there may be a wide variation in the performance needs. Here we provide examples of performance parameter ranges.

1) Time of Retrieval: the retrieval time depends on the business demands. When an email record has to be accessed from the archive for purposes of responding to customer queries on audit trails, minute to hour delays may not be acceptable. However, a legal request for email correspondence as part of a deposition may allow for a few hours to a multi-day response.
2) Data Rate: unlike multimedia archives, email archives are not usually data rate driven. However, if the email record includes large attachments, or large mailboxes are to be retrieved, the time to retrieve the data will be dependent on the data rate of the archival storage system.

Since email archives may contain many messages or records, performance in retrieval will improve if there is support for the indexing of email records.

In terms of best practice, if the archive is considered an active archive that has constraints on retrieval time performance, it would be required to locate the archive on disk-based storage within cost constraints.

4.1.3. Data Protection and Availability

Archived data refers to the primary instance of the email records. As in the case of all primary data, it must be protected from incidence of operational failure or from site disaster.

Locating the bulk of the archive on disk storage, as in the case of D2D tiered storage has the advantage of having email records on highly-available storage. If tape is chosen as the end target as in D2T tiered storage, then duplexed copies of tape have to be used to ensure availability of the data. The tradeoff that needs to be considered is the cost of using disk versus tape in solutions that provide both high-availability and retrieval performance.

Best practices in archive data protection needs to pay special attention to how archive data and its metadata are protected. Since archive metadata is usually maintained external to the archive storage, the data protection scheme needs to ensure that both sets of information are copied within the recovery time objective (RTO). Thus, if a backup technique is used and the metadata is located on primary disk with the archive on secondary disk, then both must be copied (whether incremental or full) to the backup storage target within the RTO.

Additionally, because archive data is generated from operational data, the data protection of the operational data and the archive data need to be coordinated. Specifically, operational data must be copied first followed by that of the archive within the RTO so that changes to the archive, usually, additions to the archive from the operational data, are captured to maintain consistency.

For purposes of disaster recovery (D/R), a copy of the archive may be located offsite. In that case, network-based replication schemes can be used to create the offsite copy. Besides network-based replication, making a copy of the archive on tape and using physical transport of the tapes to the D/R site is another option. There are regulatory guidelines, e.g., SEC 17a-4, that state that an exact copy of the archive data must be maintained separately from the original copy. Thus, availability of the archive copy may be a requirement beyond just data protection for operational recovery at the primary archive site.

More details on approaches on best practices for data protection of email records are provided in the SNIA DMF ILM Best Practices White Paper: ILM Best Practices in Data Recovery for Email [1].

4.1.4. WORM (write once read many) and Data Integrity

A common characteristic of archive data is immutability driven by regulatory needs. The archive management must then ensure that stored records cannot be changed, whether accidentally or intentionally (tampering). For email archives that fall under compliance requirements, such as Sarbanes Oxley, the records must be maintained unchanged. All archived records have to be read-only until the date of expiration at which point in time it can be deleted.

Ensuring data immutability of the stored archive can be solved by two approaches. First, WORM devices such as optical disks can be used. Second, storage-level firmware can be used with rewritable storage media, such as disk, to ensure that data once written to the archive is neither modified nor deleted during the specified retention period.

4.1.5. Security

Because email records are increasingly dictated by regulatory guidelines, access to the archive will be restricted to authenticated users with specific access rights.

Specific security needs include:

1) Confidentiality and access controls: any request made to the email archive may be controlled by previously set access controls. The requested has to be authenticated and access for any operations, such as read, copy or delete, are allowed.

   The archive may also be encrypted for ensuring confidentiality. This may be especially important if remote copies are made for D/R purposes. All emails sent to the remote archive outside of the data center may have to be encrypted.

2) Audit Trails on Access: the email archive may have sensitive information and all accesses to the archive may have to be logged in an audit trail.

3) Data Integrity: this is an aspect of security especially in the context of malicious change to the email archive, and has been covered earlier.

While much of the security needs are inherited from the email application security metadata [2], the archive data imposes its own access controls as well as restricted rights of access. For example, unlike users of the operational email application, users of the archive cannot delete data.

More details on compliance-related security issues for email archives are covered in the SNIA DMF ILM Best Practices White Paper Securing Email - Best Practices in Security for ILM [2].

4.1.6. Compliance

There are different regulations such as SEC-17a-4, Sarbanes Oxley, and DoD 5015.2-STD, which specify compliance requirements on email archive. While these are covered as broad requirements [3], independent of the compliance reasons, it is worth noting what requirements are typically specified by the regulations. These include

1) Retention period – period for which the archive must be maintained;

2) Disposition – at the expiration of the retention period, some regulation explicitly state that all copies of archived data must be destroyed. In terms of best practices, creating archive copies on removable media such as tape causes need for more management oversight for disposition because of all portable media need to be located and destroyed.

3) Data integrity – assurance that archived email records have not been changed, and that the quality and accuracy of the storage media can be verified.

4) Security – this includes controlled access to the email archive, audit trail of access, and assuring data integrity (Section 2.1.6)

5) Data Protection – ensuring that the email archive can be accessed even in the event of any failure, whether in the storage hardware or software of the archive storage or due to human errors.

4.1.7. Storage and Data Resource Management

As mentioned earlier in Section 3.1.1, long term retention of archives creates special challenges in addressing storage technology obsolescence. In terms of best practices, management of the archives should automate management of storage growth of the archive, as well as ensure the retention of the archive.

Best practices in policy-based storage resource management should be used for managing total storage capacity across the tiers. For retention, the archive data needs to be rewritten on new storage systems before data integrity of existing storage systems is compromised. Policy-based data refresh therefore needs to be instituted so that archived data is moved to new storage systems when existing storage systems exhibit proclivity to failures.

5. Methods of Archiving Email

5.1. General Architecture

Commercial email applications have different internal structures and therefore require a variety of approaches to archiving. The generalized approach to archiving email requires an Archive Manager that provides four archive management functions:

1) Extraction: extracts specified email records from email server's database or databases
2) Mover: moves extracted email records to the targeted archive storage tiers under the control of the Archive Manager
3) Search: provides search mechanisms to retrieve requested email record from the archive using indexing or based on content. This is required for corporate auditing as well as for litigation support.
4) Retrieval: the retrieval of requested email records based on direct access by name or by an indexed search through an API.

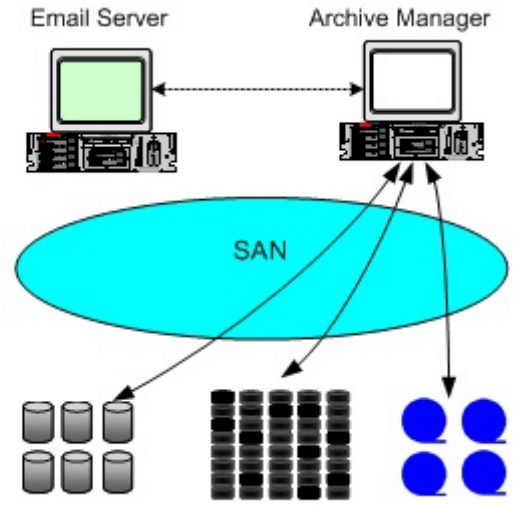

Figure 4: Functional View of Archive Process for Email

5.1.1. Extraction

The key step for the Archive Manager is extracting email records from the email server, usually from its databases, as in the case of Microsoft Exchange and Lotus Domino. Because these email applications have proprietary databases, the Extraction module relies on an API specific to the email program to access email records, folders, etc. In such a case, the Extraction module acts as a client of the Email Server.

5.1.2. Mover

Once selected email records can be accessed by the Archive Manager, it can locate the records on the primary storage, and thereafter initiate a movement so that the email records are migrated to the desired storage tier (Figure 4).

There are two approaches to the data movement: manually by the administrator managing the archive or by policy. In both cases, the data is removed from the Email Server namespace and entered into the Archive Manager namespace.

1) Manual Move: the move is initiated by the administrator, and email records are moved from one storage tier controlled by the Email Server to the storage tier controlled by the Archive Manager.
2) Policy Based Move: the same move is automated via policy. The policy can dictated by various parameters, including age, size, specific mailboxes, specific content, etc.

As a matter of best practice, it would be preferred to control the movement of the email records by policy since that ensures automation resulting in lower cost of management and better scalability.

5.1.3. Search

The Archive Manager maintains a catalog of the email records it has moved so that it can provide indexed base searches for different email records. Since archived emails are maintained independent of the Email Server, the retrieval access is through the Archive Manager, and a subsequent import process into the email application

Use of this approach may allow access to the archive independent of the original email application. This is important in the case where the current version of the email application is not backward compatible with the version that created the archived records.

5.1.4. Retrieval

The Retrieval process is for user access to the email records using the API provided by the Archive Manager. The retrieval can be based on a search request by the user or a request of the record by name.

In cases where the Archive can be accessed directly through the email application, the application can be used to retrieve, e.g., through an import process and a subsequent search, archived email records without using the retrieval API.

5.2. Fit with ILM

ILM architecture will assume a business-driven goals management capability that imposes service level objectives (SLOs) on the email archive as defined by the business goals of the organization.

The goals management will be assumed to be tied to the email specific archive needs. These goals can be dictated by corporate governance, legal or compliance needs, etc. The goals will drive the SLOs on different attributes of the archive management such as retention, performance, data protection, and data integrity, as described earlier in Section 4.

5.3. Microsoft Exchange

Implementations of the Extract for Microsoft Exchange Server will require working with the MAPI API from Microsoft[8].

The Mover and Search modules are less specific and depend on the implementation approach by the Archive Manager vendor.

[8] http://msdn.microsoft.com/library/default.asp?url=/library/en-us/e2k3/e2k3/_exch2k3_mapi_access.asp

More detail on specific approaches to Microsoft Exchange Server archiving can be found in the Application Notes for Microsoft Exchange Archiving in preparation.

5.4. Lotus Notes Domino Server

Implementations of the Extract module for Lotus Domino will require working with the Lotus Notes databases [9]. As in the case of Microsoft Exchange, the Mover and Search modules are less specific and depend on the implementation approach by the Archive Manager vendor.

More detail on specific approaches to Lotus Domino archiving can be found in the Application Notes for Lotus Domino Archiving in preparation.

5.5. SMTP (Unix)

There are few well-established options for archiving a UNIX-based SMTP mail: However, because SMTP based mail servers do not rely on proprietary databases, the Extract module in the case of these email servers are based on file access.

As in the previous cases, the Mover and Search modules are less specific and depend on the implementation approach by the Archive Manager vendor.

6. Cost

A key consideration in determining the archive architecture is cost, which has three elements: hardware, software, and management.

6.1. Hardware

The cost of the hardware is the total cost of the tiered storage assets, both acquisition and maintenance. This is driven by the total capacity and the per-unit cost of different tiered storage. Unlike primary storage applications, the acquisition cost of archived storage can be a larger fraction of the total cost of ownership (TCO) especially when considering large retention periods and the cumulative volume.

As stated earlier, what is required is balancing cost of meeting SLO goals of desired management attributes that depend on the storage type, for example performance, protection, longevity of the media, etc.

6.2. Software

The cost of software will be dictated by the cost of the Archive Manager for the archive, both license and maintenance. Other costs would include that of software modules specific to vertical markets, such as financial or manufacturing.

6.3. Management

The recurring cost of managing the archive is cost from information, data and storage perspectives.

While cost of storage management may be tied to the storage tier and its properties, the cost of data management is more subtle. This includes cost of data movement, data protection, data integrity, etc. A review of the storage tiers from Section 2 will reveal that when more than two tiers are present and data is moved from the second to the third tier, the cost of data management increases. Consider the case of the D2D2T. From a performance perspective, it is desired that the archive be kept on staging disk. However, from a cost and scale perspective, it is best to move the bulk of the archive to tape. Frequent movement of data increases cost from storage resource utilization and performance. Therefore, an option that avoids frequent data movement but still provides high performance at low cost would be the preferred archive data management approach.

The cost of information management will be dictated by the cost, efficiency and scale of search and retrieval of email records. This is a function of both the Archive Manager capabilities as well the capabilities of rapid indexed search and the performance of the storage tier where the archive resides.

7. Available Products for Email Archiving Solutions

The following DMF member companies provide solutions that support email archiving best practices described in this paper.

The following URLs are provided for additional information.

Archivas: **http://www.archivas.com**

Archivas is a leading archive object management company, focused on developing software that enables customers to store, protect, and manage fixed-content data indefinitely. The Archivas Cluster (ArC) is an object-based archive management system that is designed for long-term storage of fixed-content digital assets. ArC also enables companies to meet regulatory requirements for data retention, authentication, and availability.

COPAN Systems: **http://www.copansys.com**

COPAN Systems' Revolution 200T uses patent-pending Power Managed RAID™ and Disk Aerobics™ technologies to provide 224 TB of optimized MAID storage in a single footprint with 2.4 TB/hour throughput. The Revolution 200T is ideal for backup/recovery and active archive applications offering the performance and reliability of disk, at the cost and scale of tape.

EMC^2: **http://www.emc.com/legato; http://www.emc.com/centera**

EMC offers complete solutions to meet both long-term and cost-optimized multi-tier storage requirements for email. Legato EmailXtender makes administration of multiple tiers of email storage simple with a robust set of data migration policies. EMC Centera is designed to solve long-term record storage challenges at a lower TCO. Centera is integrated with over 100 industry partners.

Hitachi Data Systems: **http://www.HDS.com/solutions**

Hitachi Message Archive for E-mail Solution helps you increase or eliminate e-mail inbox size limits. Administrators and end-users can establish archival rules that move messages from primary to secondary storage yet preserve the client interface—making archived e-mail readily accessible, searchable, and retrievable.

IBM: http://www-306.ibm.com/software/data/commonstore; http://www.storage.ibm.com/disk/dr/index.html

DB2 CommonStore for Exchange Server and DB2 CommonStore for Lotus Domino archive e-mail and attachments for a secure enterprise solution. These products combine automated policy and user driven archive capabilities with easy retrieval. The IBM TotalStorage® Data Retention 450 is designed to help businesses meet the growing challenge of managing and securing retention managed data.

iLumin: http://www.ilumin.com/

iLumin's flagship product, Assentor Enterprise, delivers mail storage management, archiving and retention management, regulatory compliance, corporate supervision, discovery and litigation support solutions for commercial industry and government organizations. iLumin is a leading provider of intelligent content solutions that maximize the business value of corporate messaging systems, provide an immediate ROI by lowering total cost of ownership and effectively manage the risks inherent with these systems.

Kasten Chase: http://www.kastenchase.com/

Kasten Chase's Assurency(tm) SecureData storage security solution provides security and encryption for stored data throughout the data lifecycle including archives and backups, enhancing regulatory compliance. SecureData's Lifecycle Key Management provides policy-based, audited key creation, protection and deletion, ensuring efficient data compartmentalization and assured data destruction wherever data may reside.

Network Appliance: **http://www.netapp.com/solutions**

NetApp's online reference and archival solutions are typically built around the NearStore ATA disk based nearline storage system. NearStore delivers differentiated long term integrity via RAID-DP. Solutions addressing structured data (content management systems and databases) and semi-structured data (Email applications) are provided in conjunction with best of breed partnerships.

Permabit: http://www.permabit.com/

Permabit's Permeon software enables companies to efficiently and cost-effectively store archived electronic content, including email, for compliance and reference purposes. With Permeon, enterprises can easily implement industry best practices, and comply with even the most stringent government regulations for record retention and verification.

Sun Microsystems Inc.: http://www.sun.com

Sun's end-to-end email archiving solution meets corporate compliance as well as operational needs. Sun's partnership with AXS-One provides the archiving and access to messages via the original email system and legal discovery and searches. Policy-based archival and immediate retrieval onto tiered storage meets legal as well as corporate requirements.

8. References

1. ILM Best Practices in Data Recovery for Email, SNIA Data Management Forum ILM Best Practices White Paper, October 2004. http://www.snia.org/tech_activities/dmf/docs/Email_Data_Recovery.pdf

2. Securing Email - Best Practices in Security for ILM, SNIA Data Management Forum ILM Best Practices White Paper, October 2004. http://www.snia.org/tech_activities/dmf/docs/Email_Security.pdf

3. Managing Email for Compliance and Litigation Support – An Overview, SNIA Data Management Forum, October 2004. http://www.snia.org/tech_activities/dmf/docs/Email_Compliance_and_Litigation_Support.pdf

4. A Dictionary of Storage Networking Terminology, SNIA, to be published November 2004. http://www.snia.org/education/dictionary

5. Dennis Colarelli and Dirk Grunwald**,** Massive Arrays of Idle Disks for Storage Archives, 2002 ACM/IEEE conference on Supercomputing, November 2002, Baltimore, Maryland. http://sc-2002.org/paperpdfs/pap.pap312.pdf

6. Aloke Guha, A New Approach to Disk-Based Mass Storage Systems, 12th NASA Goddard - 21st IEEE Conference on Mass Storage Systems and Technologies, April 2004, College Park, Maryland. http://storageconference.org/2004/MSST2004-Agenda.pdf

7. Theodore Johnson and Jean-Jacques Bedet, “Analysis of the access patterns at GSFC Distributed Active Archive Center,” International Symposium on Microarchitecture, 1996.

8. Carolyn DiCenzo and David Mario Smith, Vendors Respond to New E-Mail Active-Archiving, Gartner Perspective, 26 November 2003

9. Archiving Lotus® Domino™ Data using IBM® Content Manager CommonStore, December 2002, IBM. - http://www.106.ibm.com/developerworks/db2/library/techarticle/0212martin/0212martin.html

10. The Impact of Regulations on Email Archiving Requirements, Osterman Research, Inc. 2003.

11. Requirements for Managing Electronic Messages as Records, ANSI/ARMA Draft Standard for Records and Information Management, 2003.

About the Author:

Aloke Guha is the Chief Technology Officer of COPAN Systems. He has more than 20 years of experience in R&D and senior management, creating both new technologies and new companies. He was the founder and CEO of Datavail where he developed content storage management prior to Datavail's acquisition by CreekPath. Aloke was Vice President and Chief Architect at StorageTek, responsible for technical strategy across business units, including tape, disk arrays, software and storage networking. He initiated and oversaw the development of the industry's first intelligent storage networking switch technology. Prior to StorageTek, he was the CTO of Network Systems, an early pioneer of channel and secure networking products. Aloke has authored over 15 patent applications (6 issued) and over 65 publications in storage, networking, switching, security and parallel and distributed processing. He is a senior member of IEEE and continues to be active in leadership positions in technical committees and industry working groups in storage and networking. He holds a Bachelor of Technology degree from the Indian Institute of Technology, Kanpur, and a Ph.D. in Electrical Engineering from the University of Minnesota.

About the Data Management Forum:

The SNIA Data Management Forum is a cooperative initiative of Information Technology Professionals, Vendors, Integrators, and Service Providers formed to define, implement, qualify, and teach improved and reliable methods for the protection, retention, and lifecycle management of electronic data and information.

About the SNIA:

The Storage Networking Industry Association is a not-for-profit organization made up of more than 300 companies and individuals worldwide spanning virtually the entire storage industry. SNIA members share a common goal: to set the pace of the industry by ensuring that storage networks become efficient, complete and trusted solutions across the IT community. To this end, the SNIA is uniquely committed to delivering standards, education and services that will propel open storage networking solutions into the broader market.

