Enterprise Information Archiving Software Evaluation Guide

December 2019

Evaluator Group

Enabling you to make the best technology decisions

Table of Contents

Evaluator Group

Evaluation Guide Overview

Evaluator Group's Enterprise Information Archiving Software Evaluation Guide is part of a series of guides designed to help IT professionals evaluate storage technology alternatives. This Evaluation Guide and the accompanying workbook are designed to assist potential buyers understand the options and products available and to help match requirements to the available technology choices.

What sets Evaluator Group's Evaluation Guide series apart from vendor sponsored whitepapers is the lack of vendor bias. Our Evaluation Guides are not sponsored by vendors and are written for IT managers seeking a vendor neutral discussion of the design considerations behind new products, technologies, and trends.

What is Enterprise Information Archiving Software

Archiving utilizes several key components to accomplish the task of preserving and managing information for a long period of time. The components can be broken down into elements that can be described in general terms for better understanding. The following figure gives one representation of the common elements in an archiving system. The diagram depicts that an archiving "application" or software will move information from primary storage to the archiving system. The application may be a specific archiving software program such as file archiving and migration or Enterprise Content Management or may be a part of another solution such as email archiving. Other software products may have archiving as a basic function such as PACS (Patient Archiving Communication System) software used in healthcare. Some archiving systems support using system utilities (move, copy, drag-and-drop, etc.) as a way to get information on the archiving system.

The archiving system will take custody of the information after the ingestion is complete and manage the information according to the rules or business regulations established. The information in many archiving systems may be available from an active archive and stored on removable media for long-term protection. The use of removable media meets many of the requirements around long-term preservation, disaster protection and uniquely serialized copies of information. Multiple copies are made based on disaster protection requirements and business or regulatory practice.

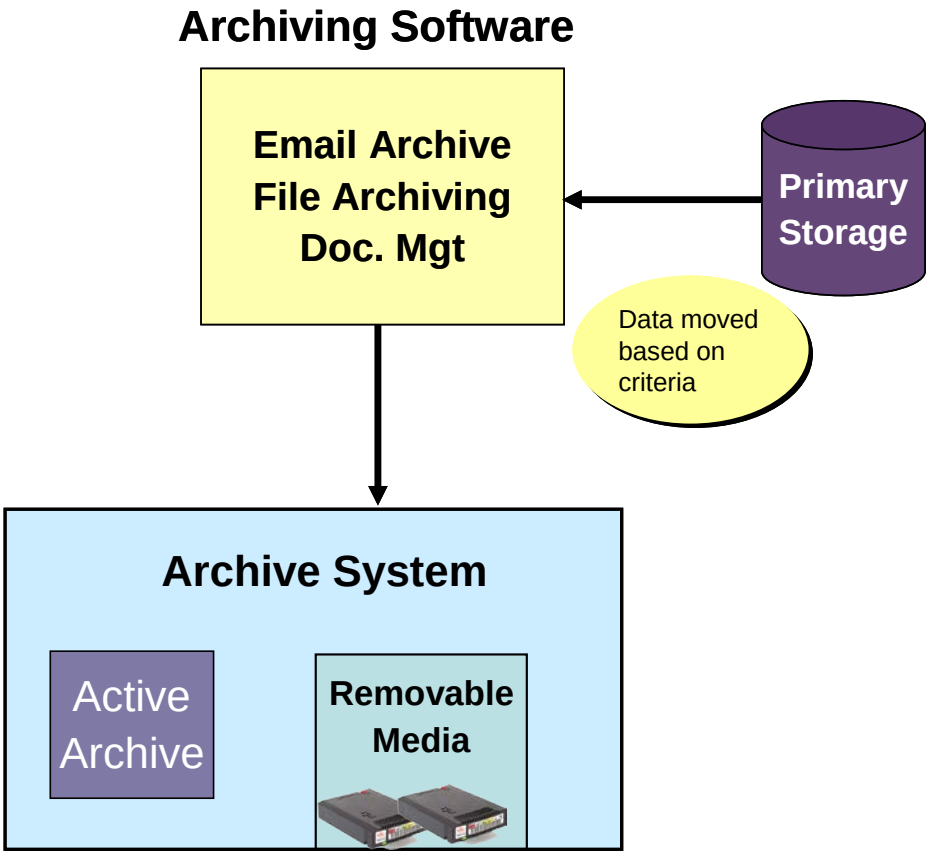


Figure 1: Archiving Overview

Archiving Software Applications

Looking in more detail at the archive elements, the first area to consider is the archiving software application. The application will run on a client or server and will move information to the archive system. There are a number of different types of archiving applications, each with their own uses and characteristics.

File Archiving or Migration

File Archiving or migration software moves data from one storage system to another based on a set of rules or controls. Hierarchical Storage Management (HSM) is one type of software that does this archiving but other types of migration software exist. HSM software has existed for over 30 years, first in the form of products that execute on IBM mainframe systems. These products were relatively complex with experienced administrators required to install and manage them. The investment made in

an HSM infrastructure required training and long-term commitments to the methods of operation and the types of storage systems being used.

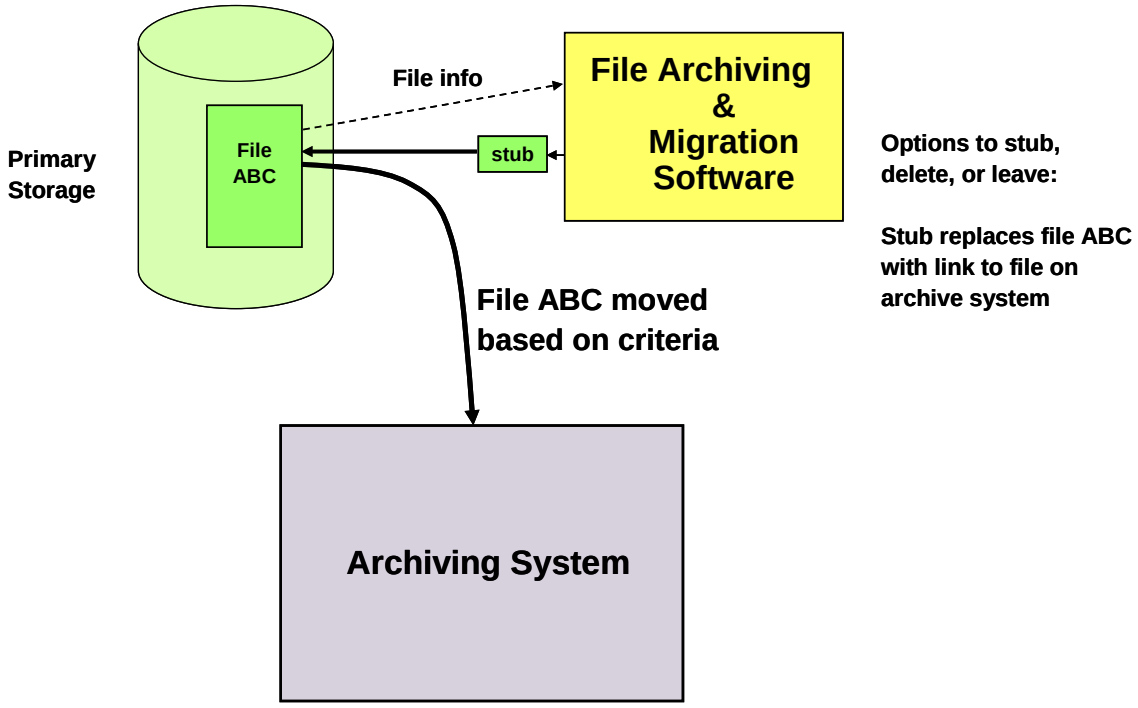


Figure 2: File Archiving Typical Operation

File archiving or migration software has been gaining in popularity for usage in open systems environments. Some of the solutions are complicated and require trained administrators to manage the usage. There are many products with varying capabilities and complexity. A simpler type of product moves or migrates data but does not manage the generations of the data moved and is usually called Migration Software. There are a number of products in this category that have varying capabilities as well. There is not a taxonomy that defines what capabilities belong in what category so each product must be analyzed based on its own functionality. As part of the file migration, files may also be stored as objects on object storage systems.

Note that most archiving is in regards to files, primarily because unstructured data in the form of files is the largest and fastest growing amount of data in most environments. Data in databases is typically managed by the database software, which may include an archiving capability. Recently, the ability to

archive or move data to another tier for object storage has begun to be available but mostly as an integrated function of the object storage system. As time progresses, archiving software may encompass files, objects, and event structured data, but the focus currently is on files.

Whether the complex functionality of a file archiving product compared to the simplicity of a migration software solution, from an archiving standpoint, the goal is to move data from primary storage to archive storage based on a set of criteria. The criteria may be something as simple as move all the data or a more complex set of rules that may be customized to particular operational needs or business requirements. Typical criteria that may be selected for choosing data to be moved to an archive system by file archiving or migration software include:

- Date of last access of the data – Using the last access time of the data gives an indication of where the information lies on the probability of access curve. Using business-specified data criteria such as the fact that the data has not been accessed in the last 90 days for example, as a trigger to allow the data to be archived is a very common criteria for archiving.

- Date of the last time the data was modified – Rather than an access time which seems to be relevant with the probability of access curve, some of the archiving software allows the selection criteria to be based on the last time the file was modified. There may be some types of applications where the modification date is really the trigger for when it is no longer needed.

- Creation date of the data – There is another approach for controlling the archiving of some types of data. The control is rooted in the types of business and information created and used in the business. It may make sense for that specific information to be archived at a specific point after it was created. This does not seem to be as universally applicable as the date of last access but has certain business reasons where it may be the selection method.

- Size of the data – There may be business rules that dictate that files of a certain sizes will be archived at some point in the cycle of when choices are made. This may seem like short-sighted or relatively arbitrary criteria but has historically been used quite often.

- Type of data – Another business rule that is often used is to select data to be archived based on the type of data. Usually the type is indicated by the filetype extension of the name but there may be other means for determination. The type selection represents a decision as to the default value of that type of information. The fact that it may be archived (and not deleted) indicates that while it is not exactly worthless, it is not part of an ongoing business process.

- Utilization – Establish some utilization criteria as to when to begin archiving is another option used in many of the criteria. A watermark type of system is very common where archiving (based on other criteria) is started when particular capacity utilization is reached and continues

until a lower threshold is achieved. Some of the archiving software will use the utilization of a volume as the trigger as to start scanning for data eligible to archive. The controls are usually set up for beginning the scan when the free space on the volume is less than some percentage of the total space and end the scan when the volume has free space to another percentage (the bracket of watermarks).

- Combinations – It is obvious that combinations of other criteria may be used to aid in the selectivity. Reaching certain goals for what gets archived by time and type based archiving seems logical but the complexity of the combination may lead to some unexpected archiving. An example of this may be where specific sizes of files become eligible for archiving after a capacity utilization watermark is reached. The file archived may be critical to processing and should remain on primary storage for a longer period.

- Exclusions – Individual files or specific types of files may be excluded from eligibility for archiving. Reasons for exclusions may include very fundamental practices such as executable files in certain directories are never archived or there are some files containing information may be needed at a known point and should not be archived. For whatever the reason, most of the archiving applications provide some mechanism to do exclusions.

Archiving is usually about moving data from primary storage to archive storage. The file archiving and migration software is used for archiving but there are several choices regarding the disposition of the data on the primary storage system. The choices of actions are usually based on operational requirements. The different software products have different capabilities and unfortunately use different terminology but the choices for the actions of file archiving and migration software generally fit into the following categories:

- Archive and leave – In this case the data is copied from the primary storage system to the archiving system and the original data is left in place on the primary system.

- Archive and stub – The data is copied from the primary storage system to the archiving system and the primary data replaced with a "stub" (or dynamic or symbolic link) that represents the original data and the location of where the data can be retrieved. Some implementations will redirect access to the archiving system while others will actually retrieve the data on access from the archiving system and restore it to the primary storage system.

- Archive and delete – The data is copied from the primary storage system to the archiving system and then is removed (deleted) from the primary storage system. An alternate path to the data may be established for the application to access the data directly from the archiving system but the primary storage system is not involved.

The stubbing or deletion action may be delayed for a specified period of time by some of the archiving applications. The delay serves to provide assurances that the archiving operation was completed successfully and provides that "last chance access" that some administrators may be worried about. The action that occurs after the delay is sometimes called "pruning" and usually is a configurable time when setting up the archiving application software.

The following diagram illustrates the typical archiving actions that were just described.

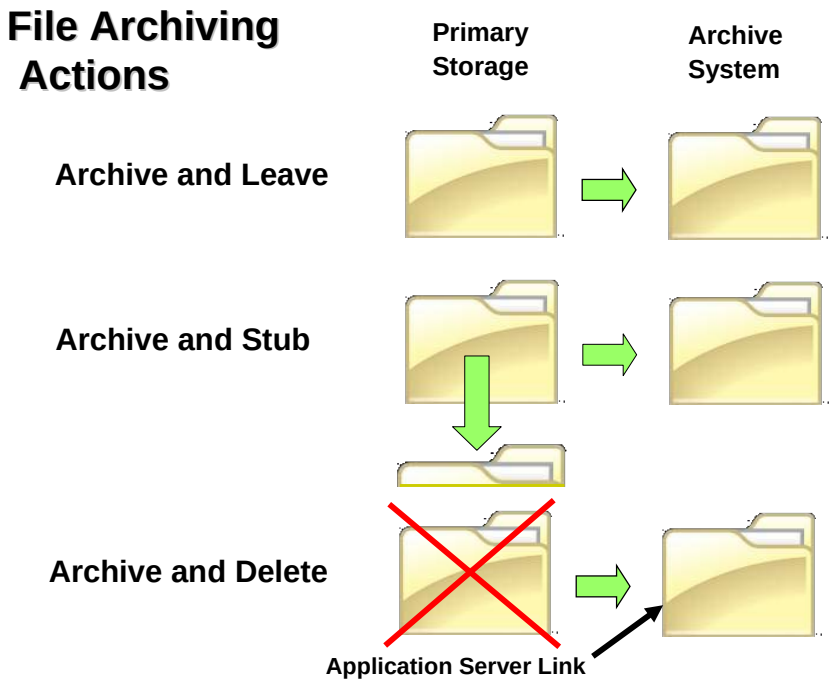
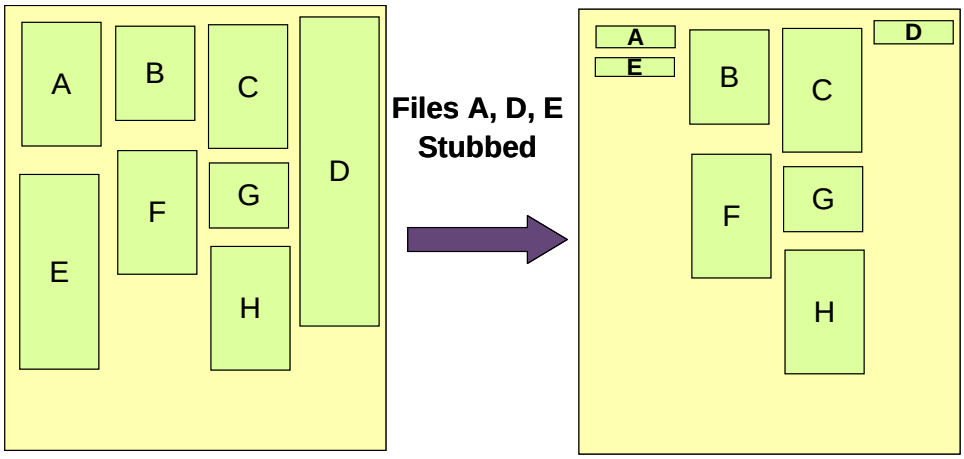


Figure 3: File Archiving Actions

The stubbing of files has a dramatic effect on reducing the use of space on primary storage systems. Stubbing will reduce any size file to a minimum representation of metadata about the file. The effect is illustrated in the following diagram:



Stubbing also called parse points, tags, sparse files, sparse mount, etc.

Figure 4: File Stubbing Effect

There are many variations of file stubbing and different names are used.

- Stub – A stub is the generic descriptive name of a file structure that is left in place of an archived file that contains information about how to access the file from the archived location.
- Shortcut – A shortcut is a pointer to a file (or program) in the Windows environment.
- Symbolic/Dynamic Link – A symbolic or dynamic link is a special type of file that contains a reference to another file or directory. An access to a symbolic link will behave like an access to file itself – the indirection will occur without interrupting the access. Symbolic links are typically used in archiving with NFS systems.
- Reparse point – A reparse point is an object that has information that a file system filter can interpret to determine how to access the data. In the case of NTFS a reparse point is a type of NTFS file system object. This provides a means to create a symbolic link that a filter driver can use to access the actual location of a file. This allows access to information that has been archived but leaves the symbolic link in place. Reparse points are used by file archiving software to transparently migrate files to the archiving system. When the file archiving software migrates a file, the contents of the file are removed and a reparse point is created that contains the information for the filter driver to locate the file on the archival system when subsequently accessed.

- Tag – A tag is a type of file that is left in place of an archived file that when accessed, will cause the archived file to be restored to the original location. The tag contains the information necessary to retrieve and replace the file.
- Link – A link is a general name for a relocation pointer that is created on the source volume to provide access to the archived data. In general, NFS systems use what is called a symbolic or dynamic link and CIFS/SMB systems use a Windows shortcut.
- Sparse Mount – A sparse mount is where a small subset of the information in an archive is in a mounted file system and is visible for applications or system access but the complete information is only available on the archive device. This allows a minimal amount of information to be stored (typically in an active archive or on a primary system) but yet identifying information about the archived data. The mounted information appears to the accessing system as an ordinary mounted file system.
- Sparse Files – Sparse files are files that only have minimal information stored. Space that has been allocated but not filled is not written to the file system and when accessed, zeroes will be filled in for that data. Another use for sparse files is when data has been removed when archived to reduce that amount of primary storage space used. In the case where a sparse file is used for archiving an API is called for access to a sparse file which will provide information about the files but actual data in the files only exists on the archive device. Included information in the sparse files will allow access or retrieval from the archiving system.

In general, there are two primary actions that may be performed depending on the implementation when a file stub on a primary storage system is accessed. The actions depend on the specific implementation of the software so selection of which software product meets the business need requires understanding of how the stubbing (by whatever name used) operates.

The most common implementation is on a reference to a stub, an agent of the file system is invoked to redirect the access to the file from its location on an archiving system. There may be a slight delay in the access but the reference to the file is fulfilled by a redirection of access.

The second most common implementation is when an access to a stubbed file is made; an agent for the file system will use the information from the stub and copy the file back from the archiving system to the primary storage system. The last access date will be changed thus potentially starting the archiving selection criteria process from a new point. Usually, the file is left on the archiving system as well. The following diagram illustrates this process.

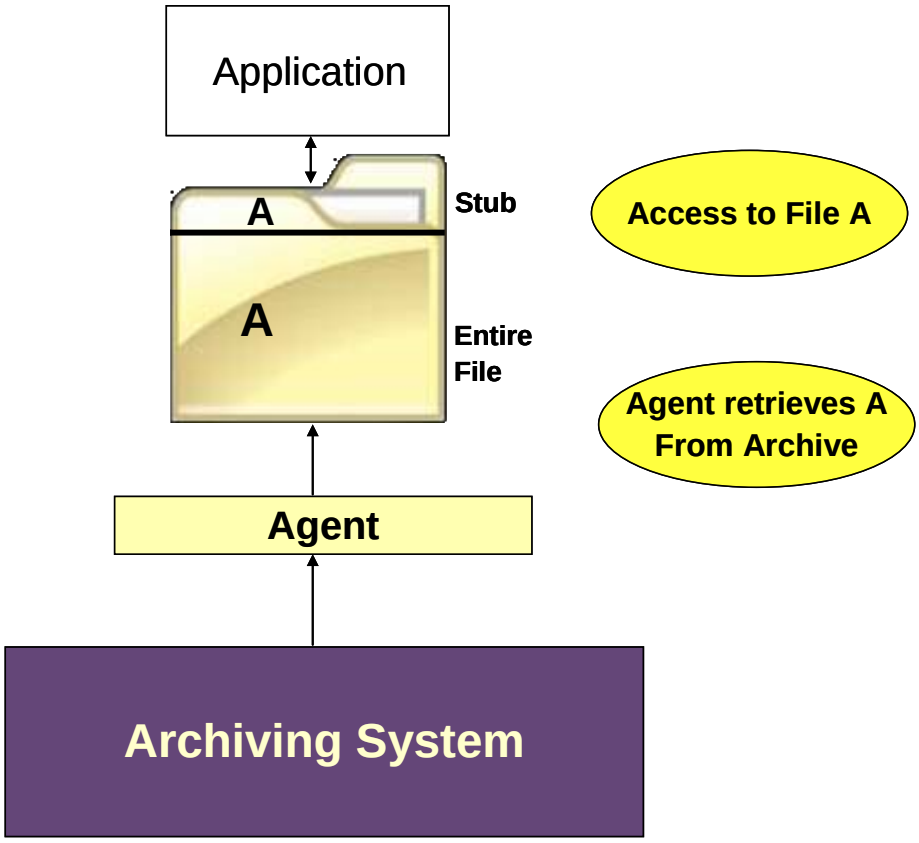


Figure 5: Stubbed File Access

The file archiving and migration software products are varied as to their capabilities but also vary as to the different operating systems and hardware platforms they support. Operationally, some of the software products may be installed as a single instance on every server that has control over primary storage or may be installed on a single server with agents or communication mechanisms to other servers for centralization of the file archiving or migration function.

The file archiving and migration software may also be integrated into other application specific solutions to perform the archiving function as controlled by the application. An example of this is the many different PACS (Patient Archiving and Communication System) where a file archiving product is integrated to work directly with the PACS software.

Analyzing the effect of using file archiving or migration software is useful for operations staff and for the vendors of the software in selling their products. Consequently, many of them have analysis software that can be downloaded from their web sites. Running them will yield results showing the data that is eligible to be archived and the amount of savings in capacity. Many of these attempt to add Total Cost of Ownership (TCO) calculations along with the analysis to put the effect of archiving on an economic basis.

Considerations for Archiving Software

The features and functions of the various software applications that may be used to archive data are important to consider. In addition, several pervasive questions must be answered as well.

Where does the software execute?

The following diagram gives a few possibilities.

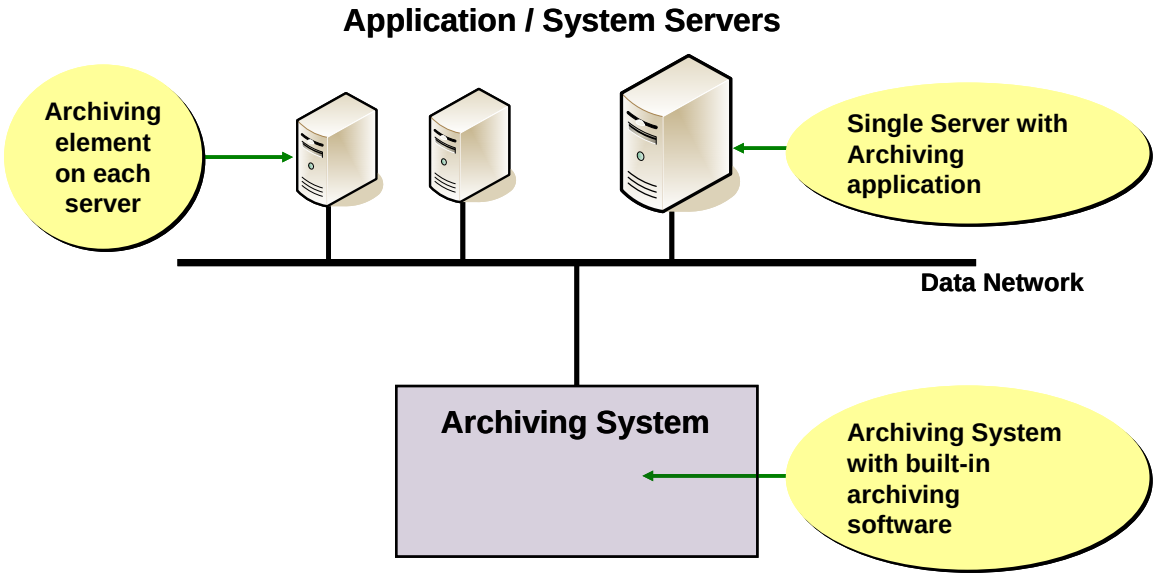


Figure 6: Where Archiving Software Executes

This has implications in many areas. There may be a requirement for some software to be installed on every server such as an agent or driver. This may result in some additional costs and certainly some additional administrative concerns. A single server may be the central control point for all operations but this may necessitate a separate server just to perform this function. In some cases, an archiving system may be used that has the archiving software integrated into it. This may limit functionality and future expansion or adoption of new technologies and may result in more special purpose systems.

What are the performance impacts of archiving software?

Moving data has the potential to impact the device involved, networks, and the servers that are executing the archiving software. The different archiving software may perform differently, may take different amounts of processor resources, and have different efficiencies in transferring data. Many of the archiving software applications will use databases for managing the history of what has been

archived. The performance of databases can change over time based on the size and number of entries and organization.

Can the operation be automated?

Many of the archiving software products have the ability to set some automation around the archiving process including selection and movement. The amount of automation and the complexity can relieve much of the administrative overhead.

Evaluation Questions

Deciding on which archiving software to choose is complicated because there are so many demands based on the usage environment. Evaluator Group offers an eBook on its website called *Information Archiving – Economics and Compliance* that should prove to be useful in understanding the requirements and choices. There are a number of questions to think about that can get the discussion started:

- What type of information is being considered for archiving?
- What are the access requirements?
- What applications are in used where the information is a candidate for archiving?
  - Enterprise Content Management
  - Email
  - User home directories
- Will users need transparent access to the information that was archived?
- What are the access rights to the archived data?
- What compliance requirements are there?
- Are there defined retention periods for the data?
- What is the operational environment – operating systems and networks?
- Can there be a separate server to control archiving?
- Do the different business owners need to manage their own archive process?
- How much data will this take out of the backup process?

Evaluator Group *EvaluScale*™ Requirements

Working with many IT clients, Evaluator Group has developed a list of the most important criteria for making product selections. These criteria and the associated requirements comprise the EvaluScale. For each product, Evaluator Group publishes a Product Brief that includes an EvaluScale showing how the product measures up.

Requirements do vary depending on usage and IT environments, variations that generally follow a segmentation of high-end enterprise, mid-range or entry-level. The EvaluScale incorporates these differences into each requirement and orders the criteria based on Evaluator Group's opinion and information gathered from IT client engagements.

Criteria for Enterprise Information Archiving Software Solution

	Criteria	Description	Requirement
1	File archiving capability with Archive Functions: archive & delete, archive & leave, archive & stub	Ability to archive files and manage the deletion or stubbing actions.	Archiving of files with the set of archive functions is the basic requirement. Additional email archiving or SharePoint are additive.
2	Detailed criteria for selection of files to archive	Set of criteria for administrator to create policies for file selection.	Criteria such as last time accessed, owner, last time modified, etc. are necessary for selection and a requirement. Must also have an exclude capability.
3	Support for target systems – NAS, object storage, archive systems	Target systems where files can be moved	Archiving targets are often NAS systems. A target could also be an object storage system using a native object protocol. There are archive systems that manage access that may be targets as well.
4	Access protocols supported – CIFS/SMB, NFS, S3	Support for file or object protocols to target device.	The requirement is to support access to the target device through file and object protocols.
5	Compliance capabilities	Features to support regulatory compliance	Much of data has either regulatory compliance or business governance rules. The compliance features to meet this requirement include WORM setting, retention controls, audit trails, security access, and permissions.
6	Leave a symbolic link or stub in NAS or filesystem	Support transparent access by leaving link or stub in place of file.	Automatic retrieval of an archived file requires a link or stub to replace the file and then cause the retrieve action. Transparent access through this method is the basic requirement.
7	Systems supported – Windows and Linux	Usage environments include Windows and Linux.	The primary requirement is to support both Windows and Linux environments. Other types of Unix systems would be a positive as well.
8	Versioning support	Store duplicate copies as unique versions and manage version access.	As part of support for WORM mode and self-protecting storage, versioning capability is required. Assigning trimming controls by groups is an advantage as well.
9	Data reduction	Techniques to reduce the amount of duplicate data stored.	The requirement is for data reduction, which could be any of compression, single instancing, or deduplication.
10	Encryption	Encryption of data at rest	The minimum requirement is to encrypt data as it is archived and provide the keys for key management. Additional function of encrypting data during transfer add advantage.

Summary

The usage of archive software is a part of any information management strategy. It can be one of the best means to manage the continued capacity growth needs for IT environments. The software works with target storage systems or cloud locations for the archive repository. The archiving software must also be concerned with meeting s regulatory compliance and business governance requirements. There are many solutions and the best procedure in evaluating the different solutions is to have a clear understanding of the requirements. The existing IT environment dictates many of the choices but the various requirements will change the product list of possible solutions.

About Evaluator Group

*Evaluator Group Inc. is dedicated to helping **IT professionals** and vendors create and implement strategies that make the most of the value of their storage and digital information. Evaluator Group services deliver **in-depth, unbiased analysis** on storage architectures, infrastructures and management for IT professionals. Since 1997 Evaluator Group has provided services for thousands of end users and vendor professionals through product and market evaluations, competitive analysis and **education**. **www.evaluatorgroup.com** Follow us on Twitter @evaluator_group*