OpenVMS Technical Journal V8

Information Lifecycle Management (ILM) Strategies and OpenVMS Storage Applications

Ted Saul - Offsite Support Consultant

Introduction

Information Lifecycle Management (ILM) planning is a new approach to data storage tracking and management. ILM encompasses the life of data from its creation and how and when it will be accessed, to its point of deletion. The total data life cycle should be viewed from two points -- its access value and its retention/restore value -- while answering the following two questions:

1. How fast is this data required to be available?

2. How long must the data be saved and restorable?

Data security must always be considered as well, not only when it's online, but also when it has been backed up and archived.

The following article will help you answer those two questions in order to properly manage the life of your data. Additionally, the article looks at two HP OpenVMS products currently available and covers how they can help manage data. This includes data that is located on disk for immediate access and data in need of regular backup and archiving cycles. These two products will be discussed as examples of how applications may make ILM achievable and more efficient.

Terms Related to ILM

The following terms related to data location are essential for the discussion of ILM:

- Online Storage – Data is kept on disk for immediate access by users. A directory command will show the files and all the information about the files.

- Near-line Storage – Data is not kept immediately online but considered shelved and located in a format either on another disk or on tape within an automatic loader device. To the user, the data will show on a directory command, but upon access there is a slight delay as the data is automatically "unshelved" from the near-line device. This is an optional step.

- Off-line Storage – This is the traditional backup scenario where data is backed up to tape. The tape may then have been removed from the library and stored in a safe location. Information about the files backed up on the tape may be kept by manual methods or by using an application cataloging facility. Users will not see the files in their directory and a strategy needs to be in place for retrieval of this data.

How Fast is Data Required to be Available?

The term access value may be defined as the weight data carries and corresponds to how quickly it needs to be accessed by the end user. Data with a high access value will require an immediate response. Current patient records, for example, have a high access value to doctors, nurses, and other medical technicians requiring immediate information during the diagnosis and treatment of the patient. Storing this data off-line on tape for retrieval would be unacceptable. The same would be true for data in any real-time transaction processing environments.

Retrieval times for data with a high access value is measured in seconds and tenths of seconds. Information with a lower access value might include statistical data such as weather data to help determine the record high and low temperature for a particular day. This data probably won't be considered critical and its retrieval could be scheduled and planned. Another good example of data that may fall into this category is historical phone charges by a cell phone provider. This data might be classified as medium-level access and kept near-line or off-line. Immediate access that is measured in seconds will not be as critical, allowing for an off-line data storage strategy. Contrast this with data from the current cycle required for billing, reporting, and even criminal investigation purposes.

It may also be helpful to sub-categorize high access values into some strategy such as the following.

- High demand access where the data is required to be available almost immediately. Any delay in response will be noticeable and may adversely affect this business it is supporting.
- Medium demand where a slight delay in retrieval may or may not be noticeable but still tolerable.
- Low access where a request for retrieval may be made to another organization and lag time may be measured in a day or so.

These additional access classes help determine which type of software to use to manage data at certain points in its life. Data with the highest access value always needs to be kept online and available. The use of large disk arrays along with database applications can help to manage this data and ensure that it is always available when needed. Online data protection such as RAID strategies and volume shadowing are available to protect the vulnerability of this data between backups.

How Long must Data be Saved and Restorable?

How long data is to be retained or its retention requirement is the second important point to consider in its lifecycle. The period of time that may be required to retrieve and restore will also fall into this category. It is always important to keep backups of your data, but a prioritization of backups should be put into place so it is easy to know what type of recovery is available. Some data may also reach a point when the retention/restore value is not as high. This typically happens at the end of the data's lifecycle.

Retention and restore values consider how long data must be available in one form or another for auditing purposes and how easily the data can be restored. As with access value, retention might be sub-categorized as:

- High -- where restores may be required within a few minutes. Backups may be sent to disk rather than tape or to libraries where the tapes can be quickly mounted to restore data.
- Medium -- where restores may take a day to complete. Typically this media is left onsite and stored in some type of secure area.
- Low -- where restores may take a day or two particularly if the data is located at an offsite facility.

Medical data is a prime example of data that has a long retention period. In most cases, patient health information by law must be retained for seven years or -- in the case of minors -- until they reach the age of twenty-one. Medical data may not require immediate access but must be restored upon request.

A system of date-based backups may be setup and retained with daily, weekly and monthly backups. A weekly backup may replace a set of daily backups and a monthly backup may replace a set of weekly backups. Yearly backups may also be captured for long-term storage. Duplicate copies of long-term backups may be kept as well for disaster recovery purposes.

When thinking long-term, the media used to store data must be taken into account as well. Consider asking questions such as:

- Is the life of the media (tape, etc) in use expected to last for the required life of the data?
- Will the hardware be available to handle your current media in the long-term future?
- Should a plan for migration of data be put into place should advanced technology become available?

During an actual disaster is not the ideal time to test your backup strategy. The process for retrieving media, installing software, and recovering data both onsite and at DR sites should be regularly tested to ensure the quickest recovery time possible. It will be important to be able to identify where each critical piece of data exists in its lifecycle to ensure that no transactions are lost or overlooked.

HP Storage Products

OpenVMS offers two storage-related products to help with ILM:

- Hierarchical Storage Management System (HSM)
- Archive Backup System (ABS) applications

HSM provides manageability of data that has a high access value while ABS assists with managing the restore and retention values of data. ABS as a backup application also provides protection of all data no matter where it resides in its lifecycle. Both products serve their own purpose but can work together to provide a complete ILM solution.

Data that is needed for immediate access needs to be backed up on a regular basis. A proven backup strategy should be in place during this time to ensure that any failure -- whether minor or catastrophic -- can be recovered from easily. Hardware functionality such as RAID or shadowing may be looked at to deepen the level of protection of this data. Once the access value of data begins to drop, HSM can be used to shelve data to near-line storage and free up disk space while allowing reasonable access times. This strategy is most powerful when large data files are required to be accessed online but only randomly. From the user point of view, the data is still on the disk and visible via a directory command. In reality though, only the file header is left behind with the bulk of the data located either on another storage disk or on an easily accessible tape within a library. HSM catalogs then keep track of the location of the data for easy retrieval. Database applications may also be included in the scheme with the use of export functions to move data from the main application to a backup location. In turn, this file is shelved to near-line storage for access when the need arises to import the data back to the application.

Once the access value is no longer an issue, archiving of the data via ABS can take place. At this point, you need to prioritize and categorize how long data must be retained overall. In order to save space and reduce backup windows, low access data may be archived from disk. Archiving typically refers to the backing up of the data and a subsequent deletion from disk. These tapes may be stored in a library but, for safety's sake, should be taken out of a library and shipped to a secure offsite location.

HSM Settings

In HSM, policies can be set up to manage how long data is kept and maintained before being shelved to a near-line device. These policies can do the following:

- Specify which files to move between primary storage and shelf storage.
- Specify which files are not to be moved from primary storage.
- Set a "high-water mark" on primary storage to automatically trigger shelving on dormant data to shelf storage. A high-water mark is a defined percentage of disk space used that, when exceeded, causes shelving to begin.
- Set a "low-water mark" as a space-recovered goal to limit the number of files that are moved to shelf storage. A low-water mark is a defined percentage of disk space used that, when reached, causes policy-defined shelving to stop.

You will want to set your policies so that data with a medium access value becomes a candidate for shelving. However, data with high access values or that for any other reason should not be moved, should be set with the "no-shelve" bit set. Databases are prime candidates for this type of setting. Current data will be located in the database and archived out of the database via an export-type command. This exported data may then become eligible to be shelved.

ABS Settings:

ABS has two policies that directly affect ILM: the SAVE and the ARCHIVE.

The SAVE policy identifies what files or disks are to be backed up, the schedule to do so, and specific information about the backup. This is typically unique data that only occurs once within the configuration. The ARCHIVE policy defines information about the backup usually found to be redundant with other backups. The ARCHIVE policy is then associated with one or more SAVE policies. For example, the catalog where data about each backup is to be stored is written to a field on the ARCHIVE record. There may be multiple SAVE policies storing their data to the same catalog. In this case, the one ARCHIVE policy will be associated with all these SAVE policies.

Between the ARCHIVE and SAVE there are seven policy settings that may affect ILM.

1. Retention – Found on the ARCHIVE, retention is the length of time in days to keep data in catalogs and available for lookup. For example, retention might be set to 7, 30, or 365 days. The longer the data is kept, the larger the catalog will grow. It is important to create the

correct system of catalogs to ensure efficient disk usage. For example, yearly backups should be kept in a catalog of their own. This data will need to be retained for a long time period and there is no use in having daily or monthly backups work around this data. The retention setting is the key in the ILM environment to ensure that metadata about backups is available as long as needed.

2. Scratch Date – The scratch date is the length of time to keep data on tape. Within ABS/MDMS, a separate volume database tracks information about tapes. The scratch date is stored on the volume record and is initially derived from the retention value. It is possible to manually change the scratch date found on the volume causing a tape to be retained longer than the information kept by retention in the catalog. This may be useful for low-priority-retention-valued volumes where some system of manually tracking these tapes has been put into place. This can help reduce the size of the catalogs. ILM will use the scratch date to preserve the data on tape for its appropriate lifetime. Once the scratch date is reached, the volume may be set to a free state and the data at risk to be overwritten.

3. Offsite Date – Recorded on the volume record as well, this field defines when to take the tape to an offsite location. Vaulting is a process that needs to be set up at each site to ensure that backup data is safe from physical harm. Moving backup volumes also reduces the threat of a single point of failure by having backups at multiple locations. The offsite date is the key to its implementation.

4. Onsite Date – Also stored on the volume record, this field defines when to bring the tape back onsite. Vault management settings, onsite, and offsite dates, play an important role in ILM and are set after a backup is completed. This may be done via an epilogue command or by using some manual process.

5. Consolidation – Stored on the ARCHIVE policy, ABS/MDMS uses this field to determine how long to make or keep a volume set. A volume set is one or more tapes tied together using the previous and next pointers. There are three different modifiers available for consolidation: interval, savesets, and volumes. The most commonly used is the interval qualifier that tells the number of days that volume set should be active. Once this date is reached, ABS/MDMS will retire the volume set and start a new one. ILM needs this setting to ensure that data is movable offsite in a timely manner and does not tie up large amounts of data in a volume set.

6. Expiration date – Expiration is also on the ARCHIVE and is the date the saved data expires. Expiration can be used as an alternative to retention.

7. Catalog – When implementing ABS/MDMS it is advisable to develop a system of catalogs to ensure time-efficient cleanups, ability to backup and restore, and proper amount of disk storage available to store the catalogs. Depending on the amount of data backed up and the length of the retention, catalogs can become very large and need to be managed carefully. Catalog is found on the ARCHIVE and may be used for multiple SAVEs. Catalog become key in initiating faster restores by allowing for the easy location of data as well as volumes.

ILM Data Zoning

To get started on setting up your site-specific Information Lifecycle Management strategy, it is a good practice to spend time reviewing the data on your systems, their security requirements, and current locations. With large arrays of disk with gigabytes and terabytes of data, it is even more important to get a handle on what files are on your system.

1. Define a list of policies and rules that affect the data on your system such as:

- Governmental Policy's (e.g., HIPAA, Sarbanes-Oxley)
- Corporate Mandates
- Application requirements
- Commonsense rules

With this information, a corporate backup policy can be initiated that includes how long data should be retained and the process for handling the data. Security issues should be addressed in the policy as well as requirements for scrubbing a tape after use.

2. Divide data farms into zones:

- System Files - System files may change with one of the following conditions.

  ECO installed which could affect drivers or other images on the system data files such as:

  – vms$audit_server.dat
  – vmsimages.dat
  – sysuaf.dat
  – rightslist.dat
  – qman$master.dat

  Configurations to the system such as new users, security changes, adding of queues, and DECNET and TCP/IP changes can affect these files.

- Application Files – Upgrades, updates, and ECOs for specific application may affect the images belonging to applications. These images may also affect who has access to what data.
- Data Files – The data for an application that may be contained in text, binary, database, and other type of files. Define whether they are high, medium, or low access.
- User Files – Those files kept by individual users for managing their day-to-day work. These files will vary from system to system but they should not fall into one of the above categories.

3. Apply the policies to each data zone:

Write into the backup policy and apply findings from step one to those in step two. For example, if dealing with patient information within the medical community, a retention value of seven years may need to be applied. Accounting data will have Sarbanes-Oxley requirements applied should any type of governmental audit take place.

The policy should also include a "bare metal restore" section in case of catastrophic outages. This will address how often application and system files need to be backed up. Perhaps ECOs or application updates will only be permitted after a system backup. Backups of original distribution binaries should also be covered in the policy to ensure their quick location.

4. Test recovery procedures:.

As mentioned earlier, an actual disaster recovery is not the time to test your procedures. Schedule specific times to restore tapes and rebuild systems to ensure they are valid and accessible.

5. Review on a regular basis:

Ensure that processes are kept up to date. When architectural changes take place review your backup policy to make sure that everything is covered. If using HSM and/or ABS, you may need to review your policies for them as well. Make sure all your data is being backed up appropriately and at the expected time. A formal change process that updates the backup policy during any software or hardware modification is also a good idea.

Information Life Cycle Stages

The following chart suggests the stages of life your data may travel through. Time intervals and depth to which backups must be retained may vary from site to site.

State	**Retention Requirement**	**Access Value**	**Description**
Data creation	High	High	Data is being created and may be subject to many changes. Getting a good backup may be difficult and the use of journaling may be useful.
Current	Regular backup	High	Data is accessed and read consistently. Modifications may be made during this point of time.
Not current – high probability access	Regular backup	High	The age of the data is getting older but it is being accessed regularly.
Not current – low probability access	Regular backup	Medium	Data is now being used less often but archiving to tape would not be efficient. A near-line solution may fit well at this point.
Stored	Send to tape – archive	Low	Data has been moved to tape and deleted off disk. Access requirements are low but policies will determine the retention.
Transit	Tape going offsite	Low	Data is moving to a secure location and unavailable during this timeframe.
Offsite	Tape secure	Low	Data is now located at a secure offsite location.
DR – transit	Tape coming onsite	High	An outage has occurred requiring disaster recovery procedures to be initiated. Data is now on its way back to the datacenter.
Transit – normal retire	Low	Low	In this case, rather than the DR Transit, the data is on its way back to the datacenter for its normal retirement. This date has been set by backup policy.
Volatile – tape in transition	Low	Low	Backup application may put the data in this state as a last chance to recover effort. Application would need to be updated should recovery be done at this point.
Overwritten	Tape reused	Data no longer	Tape is reinitialized and possibly reused at this point.

		available	

Summary

In the early years of computing, data was used to accomplish an organization's work on the computer and, in theory, make operations more efficient. In today's world, though, security and privacy issues along with the many levels of governmental control have made data a liability to the organization. It is imperative that it be controlled and protected. Managing the information lifecycle can be a time-consuming and tedious process but must remain a priority. Any file created and stored on a computer whether online, near-line, or off-line should be in its location by design with its movement carefully monitored.

For more information

Please contact the author with any questions or comments regarding this article. To get to the latest issue of the OpenVMS Technical Journal, go to:
http://www.hp.com/go/openvms/journal.