Information Lifecycle Management technical overview

Executive overview

The importance of Information Lifecycle Management (ILM) is rising on end-users' priority lists. Reasons for the increased importance of ILM include that enterprises have a greater need to more efficiently refine business processes to be more competitive as well as understanding how the changing usage of data as it goes through its life cycle affects these business processes.

This white paper will discuss what ILM is, why ILM is important, the technologies utilized in an ILM solution, what is the HP ILM strategy, an overview of ILM architecture, and an example of how the HP ILM solution solves a typical enterprise customer problem.

The HP ILM approach encompasses the adaptive storage infrastructure, providing information management that ensures that data is placed, moved, or copied to the right storage media; provided with the right storage functions; and is protected adequately throughout its life cycle. This sort of information management is vital to an adaptive enterprise—where business and IT are perfectly synchronized and ILM solutions ensure that rapid change does not interfere with corporate governance objectives.

A Horison Information Strategies report on ILM states the following questions must be answered so that enterprises "can understand how data should be managed and where data should—ideally—reside during its existence. In particular, the probability of reuse of data has historically been one of the most meaningful metrics for understanding optimal data placement. Understanding what happens to data throughout its lifetime is becoming an increasingly important aspect of effective data management."[1]

What happens to data as it ages?

Does usage decline as data ages?

Does the value of data change—increase or decrease—as it ages?

Why are we keeping more data longer than ever before?

What conditions indicate when data should be retired?

Do storage management requirements change as data goes through its life cycle?

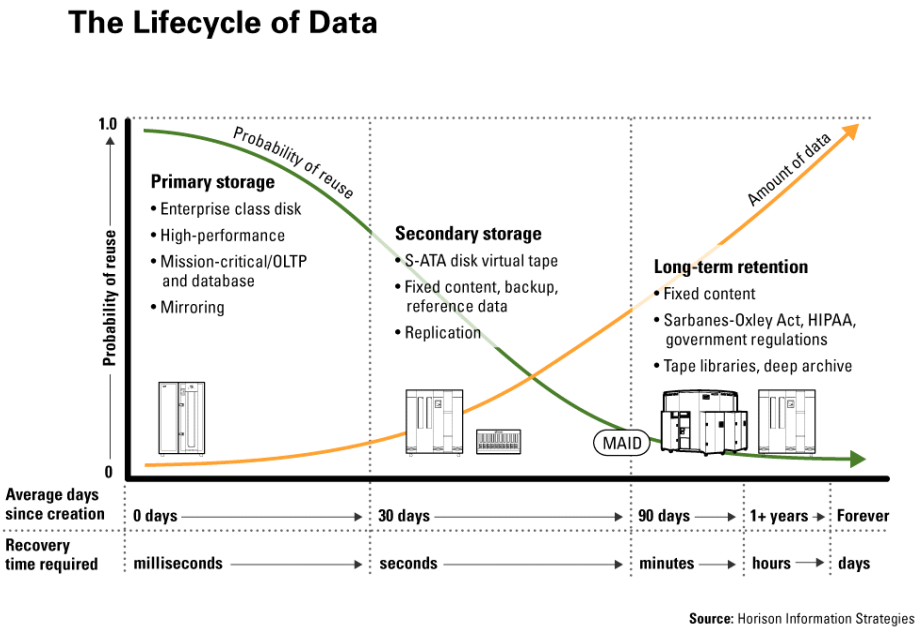
If data is the most valuable asset of so many businesses, why do we know so little about it?

The following figure from that report shows an interesting view on the life cycle of data as it relates to when the data was created, its probability of reuse, and the recovery time required.

However, there are two additional views that HP deems critical in an overall ILM solution that are not included in this illustration. These are the value of the information as is relates to the business, not just the association of the value of the data as it changes over time as is previously stated. The ability to find and retrieve information when needed—with the appropriate performance and availability attributes—is just as important. It also is important to think of the data in terms of the information value. Data by itself does not benefit the business, but the intelligent use of the data provides information, which the business can act upon to make informed decisions. In other words, information is data that is made useful.

[1] Horison Information Strategies, "Information Lifecycle Management," October 2003

Figure 1.



Source: Horison Information Strategies

The concept of ILM revolves not just around policies to manage data placement on the most appropriate storage media during it life cycle—from creation to deletion for cost efficiency reasons. As importantly, it must be placed based on the objectives of the business, how the data is used, and taking into account how this usage can change over time. This may encompass content identification, backup and recovery, replication, archiving, data migration, and data distribution across multiple storage technologies as well as robust indexing and search functions.

An ILM solution must link the adaptive storage infrastructure and associated storage management applications to the enterprise applications and business processes. This linkage is defined by policies that are determined by how the data is created, stored, and accessed when needed and meets the pre-defined service-level requirements. These service-level requirements should automatically change depending on how the data is used and how this usage can change over time.

In summary, an ILM solution must consider:

- An adaptive storage infrastructure that supports different classes of storage based on how the data is used during its life cycle. The storage hardware and software utilized as part of an overall ILM solution must support various attributes such as availability, performance, data protection and recovery, security, and cost requirements.
- Enterprise storage applications that are designed to simplify the management of complex storage infrastructures.
- The linking of the business-critical applications and the overall business processes to the adaptive storage infrastructure to ensure that the right data is available anywhere at anytime according to its business relevance, that is, its usage and the value the business derives from it at any point in time.

In addition to using greater amounts of information in new ways, the increase in compliance and regulatory requirements is driving the need for ILM solutions. There are more than 10,000 federal and state laws and regulations in the United States, and other countries have passed laws that require electronic information be stored in order that it can be retrieved, unaltered, at some future date. A common thread running through these regulations is that they address "records" information. These

regulations also address the process by which records must be created, stored, accessed, maintained, and retained over increasingly long periods of time.[2]

For example:

- Health Insurance Portability and Accountability Act (HIPAA) imposes requirements on the healthcare industry for the management of patient records.
- SEC 17a-4 regulations require that broker-dealers, as well as many multi-line financial firms, capture, index, archive, search, and retrieve their email communications.
- Sarbanes-Oxley impacts all publicly traded companies, affecting corporate governance, financial disclosure, and the practice of public accounting.
- Compliance for the drug industry is addressed in part by regulation 21 CFR Part 11. This rule affects all pharmaceutical, biotech, and laboratory device companies. This rule not only focuses on ensuring product quality exists and helps minimize risks during drug manufacturing but also covers security and electronic records storage.
- EU Directive on Protection of Personal Data is concerned with all companies that store information on European citizens and the criteria for ensuring the privacy of personal identification.

These new compliance requirements can drastically alter the methods by which companies manage data through its life cycle. As a result, records management, an important ingredient of our ILM solutions, is key in allowing enterprises to meet these requirements by maintaining the original data while copies (not backups) can be made and distributed for other purposes.

ILM overview

ILM is the process of managing the placement and movement of data on storage devices as it is generated, replicated, electronically distributed, protected, archived, and ultimately retired. ILM stores, indexes, searches, retrieves, and copies data according to the way it is used during its life span, including its ultimate destruction or deletion. To do this, ILM is composed of an interconnected set of processes, storage components, and data and storage management applications. HP provides an overall architecture and a growing portfolio of hardware, software, and services to help enterprises deploy an ILM solution.

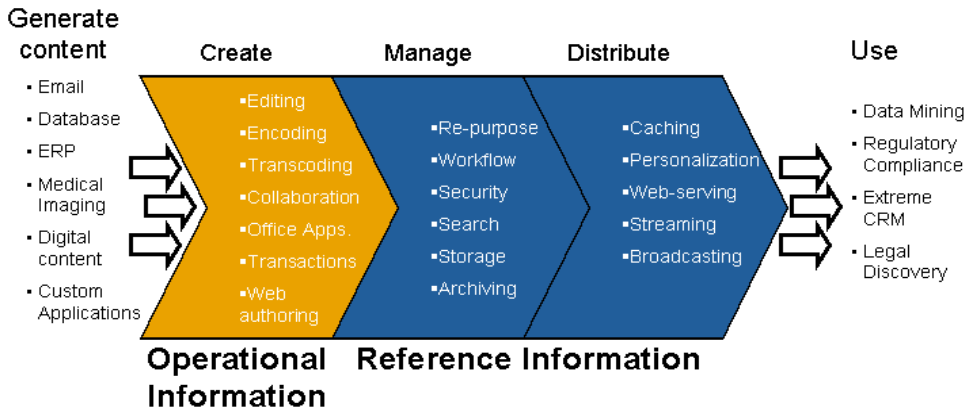
HP believes that an adaptive enterprise requires that information be managed throughout its life cycle. This means administering policy-based storage services to automate the management of information based on how it is used through all stages of its life cycle. It is important to not only manage the information where and when it is needed but also at costs and with data delivery attributes that accommodate the usage and retrieval needs of that information. Every piece of data has service-level requirements, such as performance and availability, to which attributes can be assigned, but the ability to truly take advantage of this important information—to store and manage it in ways that support its usage over time—is a critical next step to run businesses efficiently and agilely.

A key goal of ILM is to ensure that data is always stored on media that have the capabilities required to deliver the Quality of Service (QoS) and other attributes required at each stage of its life cycle. By doing so, ILM helps organizations optimize their systems and storage infrastructures. Not all information created has equal importance to an enterprise, nor are access requirements for all information the same. For these and other reasons, ILM includes the capability to classify the information according to its importance to the organizations that use it, and can then manage its placement on storage based on its classification. For example, critical data used to run the business must be accessible at all times and requires the highest level of protection (with the fastest recovery capabilities) and performance. ILM might place such data on mirrored volumes contained in high performance, highly available disk arrays.

[2] Enterprise Storage Group research, June 2003

The underlying goal of an ILM is to manage the data life cycle. To understand ILM, it is important to understand how HP views the data life cycle, which is illustrated in Figure 2.

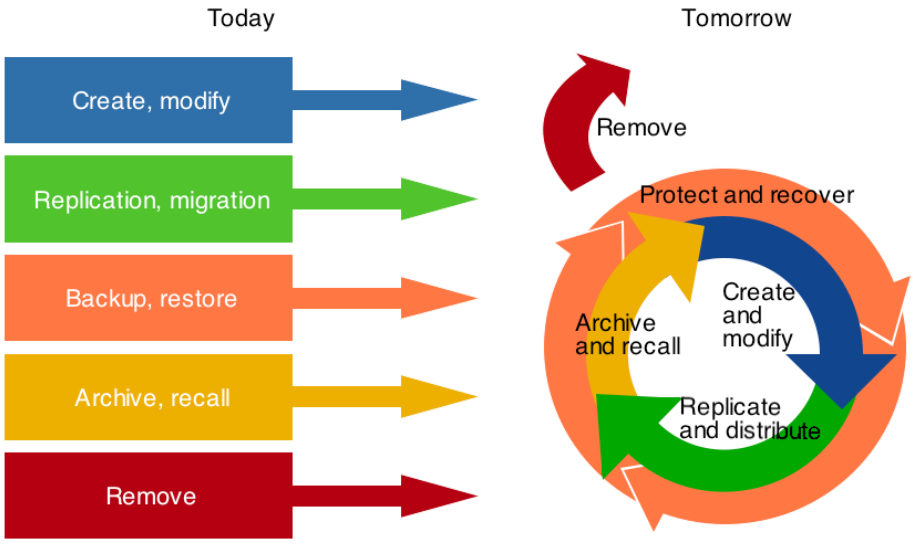
Figure 2.



The vast amount of content created, dynamically accessed, and modified regularly is shown at the left in Figure 2. It is classified for ILM purposes as *operational information* since it is created by, modified, and used for day-to-day operations. Over time, operational information becomes stable, and is not modified. In addition, the frequency with which information is accessed and modified changes over time. When this occurs, ILM manages and distributes the data to appropriate storage, based on policies that are derived from associated usage and business value attributes, to reflect the changed use patterns of the information When information becomes static (no longer modified), it is reclassified as *reference information.* Reference information is typically used for data mining, regulatory compliance, legal, and other purposes. Another aspect that ILM must consider is the fact that information used by different applications spends different amounts of time in different parts of the life cycle. For example, email spends very little time in the "create phase." After the email is sent, it enters the "manage" and "distribute" phases. On the other hand, an On Line Transaction Processing (OLTP) database application may spend much of its life in the "create" phase, as records are continually being created and modified.

ILM is not really about storage; it is about intelligent workflow and business process management. It is as much about automating the searching and indexing, categorizing, and managing of the information as it is about the storing and archiving of it. Importantly, ILM is about creating policies that can fit the information flow into the business environment as part of an overall architecture. By its nature, ILM is most effectively implemented using an application-specific solution approach that includes generalized management components such as policy engines, schedulers, and application-specific components like data movement directors. The HP strategy takes advantage of significant partnering efforts in concert with its storage and IT management expertise to create the close, application-specific linkages needed for ILM to be successful. This approach provides key HP developed storage management functionality and networked storage repositories. It adds partner- and HP developed application interfaces, and provides the option of unifying the whole ILM solution through HP or partner services, or by the customer.

Figure 3.



Today, the total process of managing the complete information life cycle is typically a manual one, as it relates to the various steps shown on the left in Figure 3. Each step to provision storage, replicate data, back up data, archive, and remove data is accomplished using dedicated and separate applications with no integration between these storage applications. Restoration of backup data and recall of archived data is an even more manually intensive effort. For example, a storage administrator must manually create scripts to run the backup application at pre-defined times—such as "create a full backup each Sunday at 10:00 PM" and "run incremental backups every night at midnight." Then, after a week, these tapes are physically archived to an off-site location or the data is replicated to another location. As a result of all this manual intervention, it is very difficult for IT to guarantee high (or even consistent) levels of QoS and conformance to service level agreements (SLAs). In addition, this rote approach does not granularly consider the usage of data, but rather is focused on users, servers, applications, and other fairly generalized IT-based criteria.

In tomorrow's IT world, ILM will provide the management capabilities to automate the data services component of the IT infrastructure. ILM is the management of information based on its changing business relevance and usage requirements. The cycle begins when data is created, and ILM governs its access, distribution, retention, and disposal according to customer-defined policies. These policies specify service levels, such as availability, protection, recovery speed, performance, security, geographic location, and cost. The key values offered by the HP approach to ILM are:

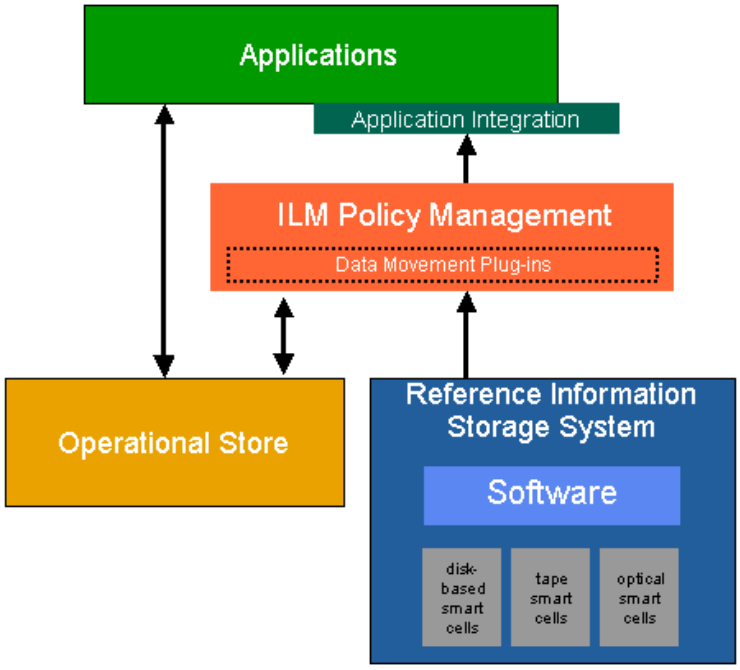
- Ensuring the appropriate availability of information to the business while minimizing the total cost of infrastructure and ownership.
- Utilizing a service-level management approach to maximize efficiency and minimize the administrative staff burden in maintaining this storage environment.

In addition, ILM can assist enterprises in aligning IT with business objectives, enabling data (email, financial data, digital images, documents, presentations, and so on) to be managed dynamically from the time of creation, through migration, to archiving and eventual removal, while adhering to business, regulatory, and legal requirements.

Information made available to the business, when the business needs it, helps enterprises meet changing legal requirements, mitigate risks, realize operational cost savings, improve data and application availability, and increase business agility. The HP ILM solution, by placing data in

managed pools of storage, enables enterprises to contain and reduce costs through the use of services, improved and increased utilization of storage, and reduced storage complexity, and by leveraging and protecting past hardware investments. All these are aimed at increasing application availability, reducing downtime, and improving performance by dynamically managing data throughout its life cycle.

Figure 4.



The HP ILM architecture at a high level is designed to understand the relationships, policies, and QoS requirements of the applications to the data as depicted in Figure 4. The basic components of the HP ILM architecture are:

Information stores are persistent repositories that contain either new, dynamic, frequently accessed data (Operational Information Store), or stable, static data (Reference Information Storage System). The latter can be a relatively complex system containing a number of types of storage subsystems and management software.

ILM Policy Management provides the automation framework that applies business-centric policies to control the data movement functions that are explicitly responsible for moving data between the information stores and among classes of storage within information stores.

Application-specific integration components connect ILM management functions to the applications they serve. These do things like quiescing applications to enable application-consistent replication to occur, or extracting data objects that are ready to be archived. The black connecting arrows represent the data flow through the total ILM system. Business applications connect directly to the operational data, which contains current and newly created data. The reference data storage system contains stable data that is in either its original or updated state; it may be accessed as frequently as operational data, or it may be accessed relatively infrequently. The HP ILM solution moves, and manages the movement, of data between stores and among storage systems within the stores. The

storage systems are classified according to QoS (protection level, recovery speed, performance) and other attributes, such as server or network connectivity, index/search requirements, immutability, and retention management.

Classes of storage reside on managed storage systems. Today, active data is placed on disk arrays such as the HP StorageWorks XP family, HP StorageWorks Enterprise Virtual Array (EVA) family, and the HP StorageWorks Modular SAN Array (MSA) family. HP also offers a range of ESL and MSL tape systems and optical jukebox systems, used for protecting and archiving data. Over time, HP will add other classes of disk storage using Serial ATA disks, iSCSI-connected arrays, and other technologies, and will expand its tape and optical offerings. HP will add intelligence and software to create and manage broad classes of storage for reference information. These are required for a complete set of storage classes for the data stores, as well as for automation, the integration into seamless ILM solutions, and the assurance of SLAs.

For more information on the HP ILM architecture, see the "HP ILM architecture overview" section of this white paper.

Why ILM

Today's enterprises face the daunting challenges of managing phenomenally increasing quantities of information so that its value to the business is fully realized, while minimizing the cost of maintaining and managing the IT infrastructure. It has become painfully clear that critical enterprise data must be continuously available. Automated processes to manage information through its life cycle are becoming essential.

According to a University of California–Berkeley study[3], the amount of new information stored on paper, film, magnetic, and optical media has about doubled in the last three years. The study goes on to state that some 92% of new information is stored on magnetic media, primarily hard drives. As a result, storage capacity requirements have nearly doubled over the same period, due in part to compliance and regulatory requirements. To meet the requirements for data retention, protection, security, and accessibility, enterprises must have plans in place to effectively and efficiently manage and protect this growing data.

It is also important to recognize that the way information is created and used has changed during recent years. Years ago, most content was created as printed media. Today's trend is to create more content—documents and other data—in digital (electronic) formats. HP believes that digital source content will rise to be the majority of content produced over the next several years. In addition, the Enterprise Storage Group (ESG), an independent storage analyst firm located in Milford, Massachusetts, states that reference information is growing at a 90% CAGR and predicts that by 2004 more than 50% of all stored data will be reference information, which is information that will never be altered[4]. For example, HIPAA states that a CT scan must be kept for the lifetime of the patient plus two years and must be stored in its original form, which is reference information. ILM solves problems related to document retention, automated data management (managing availability, performance, and other service-level delivery attributes) during its stages of use (Figure 5), and storage and accessibility management of the significant amounts of valuable reference (historical) data to facilitate data mining and other subsequent analyses. In addition, the way ILM chunks and manages data by Operational and Reference Information Stores, it speeds backup and reduces total storage requirements (less data is backed up); ultimately, it makes it possible to store every version of every file ever created.

HP believes that ILM must address the three major information management challenges that face enterprises today: Retention Management, Data Management, and Reference Information

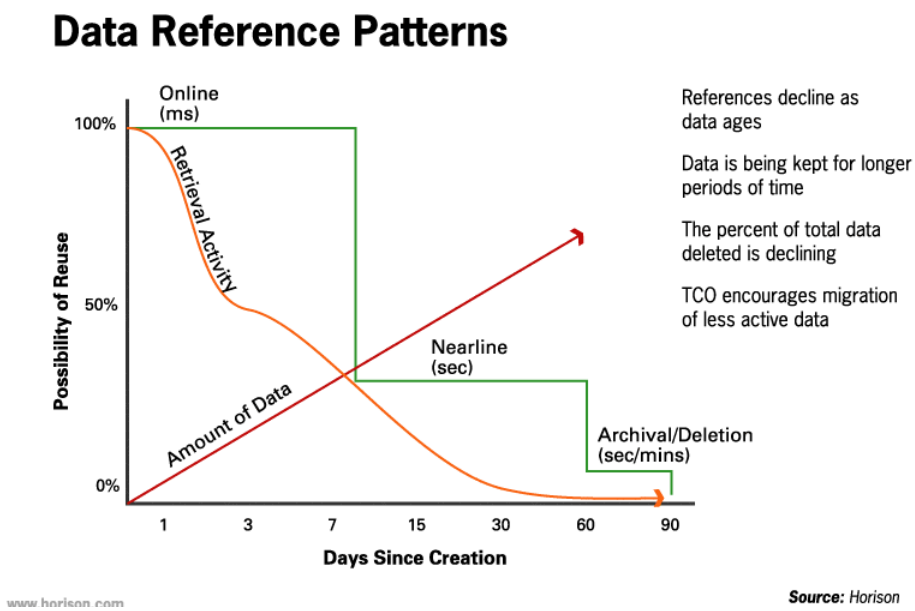
[3] How Much Information 2003, University of California–Berkeley Study, October 2003
[4] Enterprise Storage Group, May 2003, Network World Storage Technology Tour presentation

Management. In addition, data protection still is an overall issue that must be considered since an important aspect of ILM is that data must be protected appropriately at all stages of its life cycle.

This approach to ILM by HP is recognized by Horison Information Strategies in the same ILM report previously referenced. Figure 6[5] illustrates that the amount of data is being kept for longer periods of time, but with age the data is referenced less frequently so the retrieval time is not as important as it is for recently created data, although being able to retrieve information in usable forms is vitally important. As a result a policy-based ILM solution is an important criteria to simplify this process. The HP policy management approach is discussed in more detail later in this white paper.

As previously mentioned, the business use of data imparts value to the data, turning the data into information that is useful for informed business decisions and other purposes. This does not alter the impact of the Data Reference Patterns concepts, but the HP ILM perspective just carries this point further.

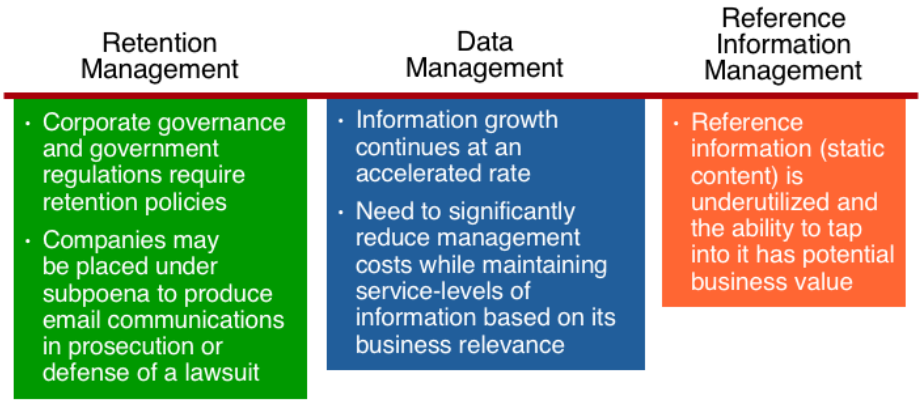
Figure 6.



The HP ILM approach addresses the issues of Retention Management, Data Management, and Reference Information Management by automating information management through its life cycle, based on its relevance, usage, and QoS objectives including data protection.

[5] Horison Information Strategies, October 2003

Figure 7.



Retention Management focuses on minimizing risk. Companies must comply with the many regulatory requirements affecting electronic record keeping. As mentioned previously, heavily regulated vertical industries such as financial services, insurance, healthcare, pharmaceuticals, and government agencies have specific rules and regulations regarding the preservation of all electronic documents including email and instant messages. Any company may be required to produce documentary evidence of email communications in connection with a lawsuit (either as plaintiff or defendant). Companies therefore must efficiently manage their email environment to ensure that email is retained in alignment with appropriate legislation and that "expired" records are destroyed at the end of the required retention period.

Data Management focuses on lowering costs and meeting service-level objectives required by applications and business needs. The growth of data has been well documented over the last few years. For example, email users are major consumers of storage capacity. It is estimated that by 2006 there will be over 600 million corporate mailboxes in use worldwide each generating an average of 8.6MB of new data per day. Enterprises must manage this increase, which has created significant challenges for IT managers maintaining service-level objectives such as performance, availability, and recovery time, while reducing data management costs.

Reference Information Management focuses on managing information as a business asset. For example, information stored in email and office documents may be seen as an extension of a company's knowledge assets. More companies recognize the need to extract maximum value from these information assets by mining archives for market intelligence and Customer Relationship Management (CRM). This need drives the requirement for powerful indexing and archive search tools and online accessibility.

HP ILM strategy

The underlying philosophy of the HP ILM strategy is to take advantage of current solutions and architectures that are proven, reliable, and extensible. HP will augment these capabilities as needed to provide complete, application-focused results. The goal is to take a solution approach, rather than just arbitrarily combining management software and hardware components. This provides customers with application transparency while limiting the extent to which applications might need to be modified to work with the ILM solution. The HP ILM strategy delivers significant elements of the HP Adaptive Enterprise and builds upon key architectural elements of the HP ENSAextended strategy.

ENSAextended is the blueprint for HP storage development over the next several years that delivers storage for the Adaptive Enterprise. As the name suggests, ENSAextended builds upon the work HP accomplished over the last half-decade, bringing us to the next phase of network storage—the adaptive storage infrastructure. ENSAextended puts businesses in control of their storage environment, allowing them to control complexity, uncertainty, and risk. With this control, they gain efficiency, confidence, effectiveness, and—ultimately—business agility. All of these are key elements of a robust ILM solution.

At the core of the adaptive storage infrastructure are three foundational elements: networked storage, virtualization, and data services.

Networked storage includes storage arrays, network attached storage (NAS), tape libraries, network infrastructure, and connections to hosts, which together are the physical components of the adaptive storage infrastructure. HP has a comprehensive offering of these network storage solutions today. In the future, these foundational elements will be enhanced with self-diagnosing and self-healing capabilities that will further enhance their resiliency. Self-diagnosing and self-healing hardware and software ensure high availability with minimal manual intervention.

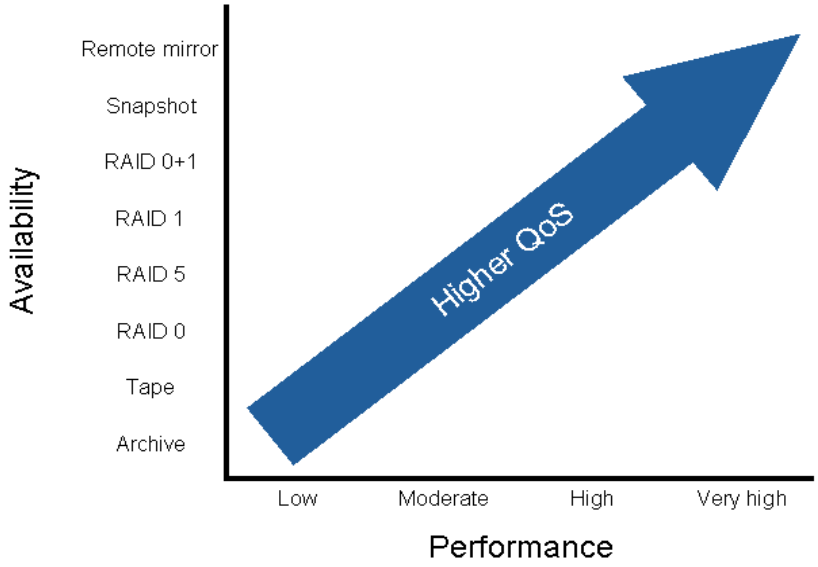
Storage virtualization, which provides storage abstraction, vastly improves storage capacity utilization and simplifies storage management. It consolidates capacity from heterogeneous networked storage arrays into a manageable, sharable logical pool. The pool is accessible by all servers and applications that are authorized to do so, enabling higher utilization of available disk space and offering the flexibility to choose and deliver QoS parameters, such as performance, availability, and protection. Today, HP offers a comprehensive family of products that use storage virtualization technology. These products are server-based, network-based, and storage system-based. This is part of the HP Adaptive Enterprise vision.

Data services contribute significantly to overall resiliency and data service-level delivery, as well as providing the core data movement capabilities that ILM manages. Today, HP offers a leading portfolio of solutions in this area, including data replication and migration, data protection and archiving, and multi-pathing failover, as well as on-demand provisioning solutions.

ILM places data onto storage that is classified according to QoS and other attributes. Data is classified according to its use during its life cycle. The combination of QoS requirements at each life cycle stage, plus the available classes of storage defined for ILM, allows the ILM software to properly place data on appropriate media and move data among classes of storage as needed. Figure 8 shows that the higher the QoS or usage need associated with the information increases, the more the availability and performance requirements also must increase.

Note that different classes of QoS attributes could reside on the same storage system. For example, different RAID configurations produce different QoS attributes, and multiple RAID sets can coexist on a single storage system such as an EVA. Similarly, a storage system might contain various types of both disk and tape devices. In summary, overall system policies manage data based on QoS and other application relevant attributes. And storage systems are configured to deliver these QoS attributes.

Figure 8.



ILM is intimately linked with business process, practices, and applications. To help customers implement successful ILM solutions, HP offers a broad range of ILM-focused services. These include services to assess a customer's current situation and to identify gaps in their ILM implementation, as well as traditional architecture, deployment, maintenance, and other services. In addition, HP ILM solutions support business and regulatory compliance requirements in terms of security, data integrity, data retrieval, and other criteria.

ILM technologies

This section provides a comparison of the various traditional data replication and recovery technologies including disk mirroring, snapshot, and point-in-time copies.

A key feature of the HP approach is to combine virtualization, multi-level data protection, and advanced storage management technologies in meaningful ways to efficiently manage data placement throughout the data life cycle. This section describes the key technologies HP takes advantage of, and the reasons they are important to customers.

Customers are looking to eliminate the protection gaps in how various data protection and recovery solutions are deployed. The key technologies in the market today include disk-to-disk backup, snapshot, and replication. Current solutions seek to eliminate the backup window, continuously copy and protect data, and enable instant recovery in case of an event that causes a disruption to the normal processing.

It is becoming increasingly clear that more than one approach will be required for protecting and recovering data. As mentioned, some of these technologies include:

- Disk-to-disk backups — As the price of hard drives falls and the reliability of low cost disk drives improves (particularly ATA and SATA drives), disk arrays can be used economically to back up directly to another disk rather than to tape. Disks offer different performance characteristics than tapes for the backup and recovery processes, and can be incorporated into a system where data can later be transferred to tape for offsite data storage without impacting ongoing operations.

- Snapshots — Rather than performing a complete backup, a snapshot can be executed that instantaneously captures changes made to source files or volumes. This can reduce the amount of data that must be backed up at any point in time and can reduce the amount of data traveling over the network. Snapshots do not replace full backups, but they do provide a means to protect data more frequently without as much CPU and network resource as a full backup. Snapshots also offer an escape from the dilemma of ever-shrinking backup windows; and, because of their small capacity requirements, many snapshots can be maintained online—a feature that can be used to provide many "roll-back" points.
- Replication over distance — Copying or backing up data to a secondary data center provides more complete protection than backing it up on site within single storage systems. Although Fibre Channel provides high data transfer speeds, a Fibre Channel capable infrastructure is not universally available and solutions to transport data over extended distances add cost and complexity. For these reasons, HP remote replication solutions take advantage of a selection of technologies that include Fibre Channel, Dense Wave Division Multiplexing (DWDM), SONET, and TCP/IP.

Content identification

Content identification involves identifying application-relevant data objects and applying metadata tags that can be used by ILM software to catalog, index, search, migrate, recover, and otherwise manage them.

Content in most cases is application-specific. Examples are Microsoft Exchange messages and mailboxes; PowerPoint files and the slides and other objects that comprise them; Microsoft Office documents; database records; video streams (the collection of blocks that are played back on-demand to a consumer); and so on.

One way to identify data content is through tags that provide cues to demarcate specific kinds of data objects or logical boundaries between objects. Data and content information can be stored on a content storage system, as compared to an online storage system that is designed to retrieve data fast when needed. A difference between a content-based storage system and an online storage system is the location-based addressing schemes found in NAS, storage area network (SAN), direct attached storage (DAS), tape, and optical solutions, which track information based on its physical location. A content storage system may also add management functions like indexing, search, and retrieval to basic data placement capabilities.

Compliance-focused content-storage systems are a part of an overall ILM solution portfolio that can assist enterprises in preserving content to comply with regulatory rules, including files that must remain unchanged over time to protect data integrity. Examples of content include:

- Email messages
- Microsoft Office files
- Patient medical records
- Digital x rays
- Blueprints
- Mechanical/engineering drawings

More and more organizations are establishing data-retention policies and are interested in storage systems that include archive/index/search/retrieval capabilities to achieve them. In some cases, these systems store fixed, read-only data with unique identification mechanisms to determine age, owner, and other attributes.

Data distribution

Data distribution can be defined as the capability to get information closer to where it is needed and when it is needed. Information that companies need to widely disseminate cannot be tied to one server or location without causing bottlenecks; the information must be made available on multiple servers to increase the number of simultaneous accesses it can accommodate. For example, web servers provide multiple copies of data that is accessed frequently, such as copies of product brochures or other static content.

For maximum benefit, data distribution should allow the creation copies of data without slowing down application server. This might entail the use of snapshots coupled with a write history log disk and commands that suspend and resume application operation to allow the preparation of copies during off hours. Write history logging, while good for gathering initiator updates, requires a considerable amount of time to merge the updates. During this merge time, the initiator controller may not be able to fully respond to server requests, so good time utilization of the server is necessary when using write history logging.

Copies of data can be distributed using remote copy sets from an initiator pointing to a different target subsystem. An initiator subsystem can have only one target subsystem, but the target can be redirected to another location. Then, normalization can occur between the remote copy set units to provide a data copy at the new location. After the copy is distributed to the desired location, the old target subsystem link can be re-established, or a new remote copy set to another location can be created.

Data distribution can be accomplished in a disaster tolerant environment, or if this is not a concern, it can be performed more easily in a non-disaster tolerant environment.

Data migration

Data migration is the moving of data from one storage device to another. The more traditional methods for accomplishing this either require application downtime, or severely impact the performance of applications. The primary objectives of managed data migration are to accomplish the data migration without data loss and minimal interruptions to ongoing business operations.

Typical data migrations include:

- Upgrading from legacy storage systems
- Migrating off end of lease storage systems
- Migrating from an alternate vendor's or alternate model of storage device
- Consolidating resources (server or storage)
- Transitioning to a networked storage environment
- Moving data to a different kind of media, such as from disk to tape

ILM moves data among classes of storage, which are described at the pool level rather than the device level, with minimal impact on running applications. The total ILM solution manages the movement of data among classes (pools) of storage, as well as placing data on devices within pools.

Disk mirroring

Mirroring, or RAID Level 1, involves the creation of two or more physically independent but logically dependent identical copies of data, usually on separate media. Mirroring is typically implemented at the logical volume level. It may be implemented within a single storage array, across multiple arrays, or extending to mirroring over distance for business continuity purposes.

Disk mirroring replicates logical units (LUNs), or volumes, for data availability reasons and requires the same amount of disk space for each target whether in the same array or in different arrays.

While disk mirroring provides sound protection against disk and some other mechanical failures, it is important to realize that it does not protect against all problems. For example, mirroring replicates everything—even corrupt data. A mirror of the data allows the user to recover from system or hardware failures; however, in the event of data corruption or inadvertent deletion, the problem is propagated and other protection is required.

HP offers mirroring capabilities in all of the HP StorageWorks disk arrays, as well as the HP OpenView Continuous Access Storage Appliance (CASA) and host-based solutions.

Snapshots, snapclones, and other point-in-time copies

Point-in-time copies can be used in conjunction with mirrors and sometimes instead of mirrors. They provide physically dependent but logically independent copies of source data. Generally speaking, snapshot technology allows the scheduling of periodic point-in-time copies of a live volume, requiring only a fraction of the capacity of the volume being protected. For example, it is possible to take a snapshot every hour, enabling the administrator to quickly restore the disk content back to a known good state, with hourly granularity.

A point-in-time copy, as defined by the Storage Networking Industry Association (SNIA), is a fully usable copy of a defined collection of data that contains an image of the data as it appeared at a single point in time. The copy is considered to have logically occurred at that point in time, but implementations may perform part or all of the copy at other times (for example, through database log replay or rollback), as long as the result is a consistent copy of the data as it appeared at that point in time. Implementations may restrict point-in-time copies to be read-only, or may permit subsequent writes to the copy. Three important classes of point-in-time copies are: split mirror, changed block, and concurrent. *Pointer remapping* and *copy on write* are implementation techniques often used for the latter two classes.

A snapshot may be either a duplicate or a replicate of the data it represents. Using the SNIA definition, a duplicate is generally referred to as a copy of a collection of data, including point-in-time copies, while a replica or replicate is a general term for a copy of a collection of data that could be a complete duplicate (that is, a copy that requires the same amount of storage capacity as the original), or a point-in-time copy or snapshot (which requires significantly less storage capacity than the source). In general, snapshots are file system based, subsystem based, or volume based. HP creates snapshots using both snapshot and snapclone, and cloning technologies.

Often, multiple point-in-time copies are maintained online, such as by the regular and frequent creation of snapshots. A key benefit of this approach is providing protection against operator error and data corruption by maintaining easily recoverable copies of data made before corruption or file deletion.

Our replication management controls the provisioning of storage and the creation of replicas but does not change the basic uses of disk and tape storage. Rather, it enables zero-downtime backup and virtually instantaneous data recovery. HP has long pursued a multi-level data protection strategy wherein disk-based replication includes snapshot, snapclone, clone, mirroring, and remote mirroring applications, and these technologies are used harmoniously with tape-based protection to create a complete spectrum of protection and recovery capabilities.

Replication management maintains indices of where the copies are and what they represent such as a LUN at a point in time, a collection of files at a point in time, a complete copy of a LUN, an incremental change in data from a copy made at previous point in time, and so on. In addition, it provides information about how to use the replicas to restore lost data or create a completely new copy of a LUN or other collection of data for new purposes (such as data analysis).

HP provides point-in-time copies through snapshot, snapclone, and other capabilities in the HP StorageWorks XP array family, EVA3000, EVA5000, MSA1000, HP OpenView Storage Virtual Replicator, and the CASA.

Remote replication

In the context of networked storage, replication is the movement of redundant copies of objects such as volumes, files, and tables over TCP/IP, among networked storage systems. Replication is often deployed for geographic redundancy purposes and for use in disaster recovery, although it is also commonly used to protect against a storage array failure. Normally, there is a copy on one site that is replicated to a copy on another site. Usually, the last available copy is used for the replication. This approach is suitable for site failover. Replicas created this way are exact replicas. While they protect against storage system failure, they do not protect against corruption caused by a virus attack, malicious corruption of data, or non-storage system corruption.

HP provides remote replication today through HP StorageWorks Business Copy HP StorageWorks Business Copy EVA, HP StorageWorks Continuous Access EVA, HP StorageWorks Continuous Access Virtual Array (VA), and CASA.

Copy synchronization

Copy synchronization, an important capability that is part of data replication, ensures the consistency of multiple copies of data, whether they are local or remote, or if they are full mirrors or replicas. This is critical if the copies are to be used for offline backups for data protection and data recovery purposes, or for online recovery in a case of a site disaster. In addition, the copy and synchronization process can be either host based or server based. Regardless of the implementation type, ILM must manage and monitor the entire copy and copy synchronization process.

After the synchronization begins, only the blocks (or other granular replication units) that were changed on the primary copy are written to the secondary copy. After the synchronization is completed, both the primary and secondary copies should be consistent.

HP has many products for the replication or copying of data. CASA provides cost-effective data replication over multiple transport protocols such as Fibre Channel or Internet Protocol (IP). CASA enables remote and local, synchronous and asynchronous replication and optional point-in-time snapshot copy capability across a wide variety of storage and server platforms.

In terms of host-based replication, Storage Virtual Replicator does point-in-time replication (snapshots) and online volume growth, while being cluster aware.

From an array-based remote replication perspective, two HP products enable remote array-to-array replication of data: HP StorageWorks Continuous Access and HP StorageWorks Data Replication Manager MA/EMA array families.

Array-based local replication—that is, local cloning and snapshots—is accomplished with Business Copy and VA and with HP StorageWorks Business Copy upgrade (EVM) for HP MA, EMA, and EVA disk arrays.

In addition, a disaster tolerant storage capability is delivered by our disaster tolerant SAN solution consisting of Data Replication Manager, cluster extensions, and Continuous Access, to name a few. These can be combined with server clustering and failover capabilities to ensure instantaneous resilience to many kinds of failures.

Continuous backup/instant restore

Continuous protection (also called continuous backup and instant/rapid restore) solutions attempt to solve all the data protection and recovery problems and issues in the other data protection technologies because not only do you have the latest data available all the time, you can store it locally or replicate it remotely. In addition, it can provide the full history of information at any point in time. This can be at the file level, directory level, application level, or the full system.

These solutions are intended to provide these customer benefits:

- Continuous protection for total data availability
- Guaranteed recovery to any point in time so that there is the elimination of protection gaps
- Assurance of 100% data integrity
- Instant recovery
- Protection that is fully automated and always on
- Built-in redundancy functionality

HP provides continuous data protection through instant recovery solutions that enable disk-based recovery, making it possible for customers to recover critical data in minutes rather than hours. Components of the HP instant recovery solutions include: HP OpenView Storage Data Protector software for ease of backup management by way of scheduling and copy recycling; HP StorageWorks software such as Business Copy, VA, and Continuous Access; HP StorageWorks disk arrays and tape libraries; HP servers; and HP consulting and support services.

Tape-based backup

A tape-based backup is defined as a collection of data stored on tape media for the purpose of recovery in case the original copy of the data is lost or becomes inaccessible for any number of reasons. In order for a tape backup to be useful for recovery, the backup copy must be made when the source data image is in a consistent state. For example, in the case of a database it must be in a stable state, usually achieved by bringing the database to a quiescent state, before making the backup or replica copy.

One of the drawbacks of utilizing a tape-based solution for business-critical data is that it requires a backup window due to the issue previously mentioned—although this can be reduced when used in combination with point-in-time images. The backup window is the period of time available for performing the backup. The IT administrator typically defines the backup windows by operational necessity. In some cases, the data and applications are unavailable during the backup window time.

The rate of recovery has been receiving a lot of attention lately. According to Enterprise Management Associates, "99 percent of IT sites in North America don't know how much of their backup data is actually recoverable."[6] This has become a major concern for enterprises. Along with the importance of disaster recovery, more and more companies have come to realize that storing backup tapes is no longer enough. The ability to recover that data is what is now critically important.

Backup creates copies of data for short- and intermediate-term data protection. These copies can protect against human error and system (hardware and application) failures. Rolling back through multiple backups can provide recovery from data corruption if a pre-corruption backup copy is available. Since backup is one of the key operations that cause copies of data to be created and retained, our ILM software will control backup operations. The backup copies, which in the parlance of ILM represent reference data because they will never be modified, are placed into a Reference Store, which is explained in more detail in the HP ILM architecture section.

In the area of tape-based backup, HP provides HP StorageWorks MSL libraries and the enterprise class HP StorageWorks ESL9000 Tape Libraries to meet the various needs of enterprise customers. The ESL9000 Series Tape Libraries provide component-level redundancy, high availability, and enterprise-level capacity for direct SCSI and SAN environments. The ESL9000 Series also provides for fully automated operation for backup, save and restore, of critical data when used with a variety of qualified industry-standard backup application solutions.

[6] eSecurityPlanet.com article "Backup Emerging from the Shadows of Storage," August 11, 2003

Archived copy

According to the SNIA, an archive is a consistent copy of a collection of data, usually taken for the purpose of maintaining a long-term durable record of a business or application state. Archives are normally used for auditing or analysis rather than for application recovery; therefore they are usually created on unalterable (read-only or not rewritable) media. After files are archived, online copies of them are typically deleted, and must be restored by explicit action. Thus, archiving is the act or process of maintaining this logically consistent copy of data.

An archiving solution must be able to track and identify information that must be deleted, from either the online environment or the archived environment after the information retention period has expired. In fact, an archiving solution might go so far as to securely purge the expired data, and log the fact that it has done so.

Archiving moves data to a durable archival store for the purpose of long-term data retention. Archiving—a key element of the information life cycle—is a complex subject that requires a systematic analysis and process for the development and deployment of a solution. Because of its business importance to customers, HP initial ILM solutions deal with archiving data for specific applications. Over time, the range of applications for which ILM solutions are available, as well as the range of ILM capabilities offered, will increase significantly.

HP data archiving solutions focus on our optical jukeboxes—the high-end HP StorageWorks 1200MX and 2200 MX and the mid-range HP StorageWorks 300MX, 600M, and 700MX. In addition to optical media, tape is also useful for archiving.

The recent acquisition of PERSIST Technologies, Inc., a leading provider of software designed for long-term storage and access of reference information, enhances the ability of HP to deliver complete ILM solutions. With PERSIST's active archiving software and disk-based archiving, HP will deliver enhanced archiving solutions to assist customers in complying with emerging and stringent data retention regulations and extract business value from large amounts of reference information. Email is one of the applications that this technology supports and is used as an ILM scenario later in this white paper.

Hierarchical Storage Management

The SNIA defines Hierarchical Storage Management (HSM) as automated migration of data objects among storage devices, usually based on inactivity. HSM is based on the concept of a cost-performance storage hierarchy. By accepting lower access performance (higher access times), one can store objects less expensively. By automatically moving less frequently accessed objects to lower levels in the hierarchy, higher cost storage is freed for more active objects, and a better overall cost performance ratio is achieved.

Traditionally with HSM there have been management tools that allow a user to define movement policies for the data so the movement would take place automatically, typically from "expensive disk" to "inexpensive tape or optical" media. This automation allowed IT to manage significantly more storage per administrator. While HSM-like tools have been available in the open systems market for decades, the ever-declining cost of disks made the purchase of expensive software unattractive. In addition, the need to integrate with the file system has made heterogeneous systems difficult.

Due to this, rather than manage storage resources with HSM-like functionality, IT bought more online storage. As a result, today's storage management costs can significantly outweigh the storage costs. Now the concept of HSM is making its comeback. The new HSM architectures typically provide lower-cost disk caches behind high-end storage and often in front of tape to provide a lower-cost, higher-performing alternative to backup and recovery, as well as file archive.

The current target applications for these new HSM-like capabilities are backup/recovery and archiving. These applications are ideal for an HSM architecture because as the data ages it becomes

less relevant to everyday business needs, although regulations may require the data to be retained for long periods.

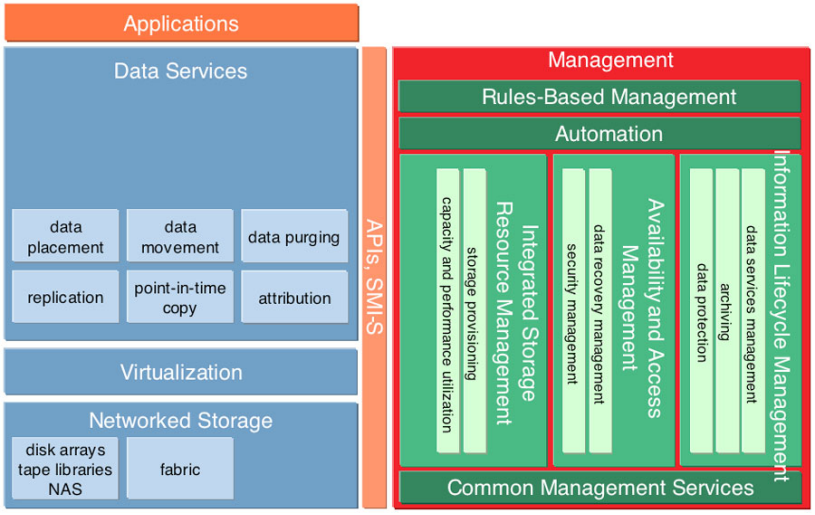
ILM refocuses HSM on application-relevant attribute classes of storage, and manages data placement and movement more granularly and completely than traditional HSM solutions. The HP HSM automatically moves data among classes of storage (containers). Storage classes are defined in terms of QoS (performance availability, speed of recovery, and so on), price, and other attributes rather than on price alone. This is possible in part because of the rapidly dropping storage costs and the large number of attribute classes of disk storage (such as different RAID levels and types of storage arrays such as the HP StorageWorks MSA, EVA, and XP families). There are also now more ways to move data among classes of storage (such as the CASA and remote replication applications). These options create opportunities to define in more granularly the storage hierarchy, and to accurately place data onto media that more closely match application needs at more stages of the data life cycle. In addition, ILM provides a complete system that includes the critical linkages between HSM and data protection and recovery capabilities.

HP ILM architecture overview

The HP ILM approach is an example of the vision described in the ENSAextended technical white paper (published January, 2003). ENSAextended is the vision for HP storage evolution for the next several years, providing an infrastructure fully supports the HP Adaptive Enterprise.

Figure 9 shows how ILM takes advantage of ENSAextended data infrastructure services and networked storage, and adds management components that are described in that white paper. ENSAextended puts businesses in control of their storage environment, allowing them to control complexity, uncertainty, and risk with the ultimate goal of achieving business agility.

Figure 9.



ENSAextended continues to evolve along specific directions and ILM represents a key evolutionary component of the overall strategy.

The left side of the figure depicts the various data services or storage applications, virtualization technology, and network storage components provided by the ENSA reference architecture. Our ILM approach is mapped to these components.

On the right side of the illustration are the common management services provided by HP including ILM, Availability and Access Management, and Integrated Storage Resource Management. Vendor-specific APIs and industry standards such as SMI-S provide the integration piece enabling the common management services to communicate seamlessly with all the applications and storage technologies on left side of the figure.

HP ILM Architecture

Along with being integrated with the ENSA reference architecture, HP also leverages its Adaptive Enterprise strategy and maps into the HP Darwin reference architecture. Figure 10 depicts the essential elements of our ILM architecture and maps them into the HP Darwin Reference Architecture.

Figure 10.

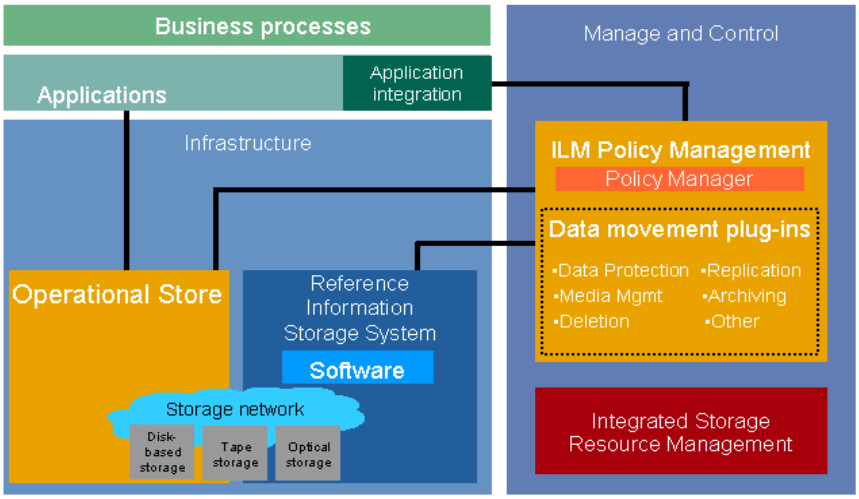


Figure 10 shows an overview of the HP ILM architecture. The left-hand side of the architecture depicts the logical data path, while the right-hand side shows the management path.

Operational data is the active, dynamic, often mission-critical data used by applications. Note that there is nothing placed in the data path between the operational data and the applications. As the data ages and becomes "static" or fixed based on ILM policies, it will manage the movement, including archiving, of this data to the more cost-effective reference information storage system. The system also assigns search and retrieval attributes.

The HP ILM approach works in concert with Integrated Storage Resource Management, which forms the structural underpinnings for subsystem discovery, monitoring, configuring, reporting, and other capabilities. Integrated storage resource management, delivered through HP OpenView Storage Area Manager (OpenView SAM) and other products, provides the management tools for managing the HP storage infrastructure, as well as storage and storage networking devices from other vendors. Much like the dashboard in a car serves as the car's command center, these tools serve as the dashboard for the storage network, performing fundamentally important tasks such as device discovery and mapping, monitoring and event management, and reporting on resource utilization.

The main architectural elements of HP ILM architecture consists of:

- Storage network devices (disk, tape, optical)
- Operational Information Store (software and storage components)

- Reference Information Store (software and storage components)
- Software for application integration
- ILM Policy Management
  - Policy Manager
  - HP OpenView Storage Builder
  - Data Movement plug-ins

Information Stores

The HP ILM architecture is founded on two fundamental *information stores*. The Operational Store contains dynamic data and the Reference Information Storage System holds stable (static) information.

Physically, these stores reside on networked storage systems (disk, tape, optical) that are organized into federated elements, called *SmartCells*. A SmartCell is a self-contained modular unit of "processor embedded" storage capacity that becomes the fundamental architectural element of HP ILM fabric and provides the atomic units that enable ILM scalability. Storage capabilities are scaled by adding SmartCells to form a grid, which contain the required attributes of availability, performance, cost, and so on. SmartCells can use disk tape or optical technology and includes a CPU to perform indexing, search, and retrieval functions within these cells. Because of the index/search/retrieve capability, various business applications can access the data through queries made to the index/retrieval application. In other words (and this is particularly important for archiving applications), the application that created or stored the data need not be used to retrieve the information. Any appropriately authorized application could access the data. For example, even years after a set of financial records have been stored by a now-defunct application, a current financial analysis application could access and use the data.

Operational Information Store

The Operational Information Store contains data that is the active, changing (undergoing modification), and often times mission-critical data being used by the business. In addition to being dynamic and frequently accessed, this data typically has a high immediate value. Therefore, the Operational Information Store uses classes of storage with the highest performance and availability as defined by the ILM policy. In addition, this store utilizes the fastest and easiest data recovery technologies to assure the highest quality of service.

Reference Information Storage System

Most information used by companies is relatively static, meaning it does not change. In our ILM context, this is called "reference information" and it is placed in an ILM Reference Information Storage System. Because this represents a vast and often rapidly growing body of information, the Reference Information Storage System is a highly scalable and reliable environment that can incorporate a broad range of devices, including disk arrays, magneto-optical, WORM, and tape devices.

The Reference Information Storage System is composed of two fundamental parts: the storage units or SmartCells and the software. Each SmartCell has its own indexing search capability, allowing a grid containing many SmartCells to execute content searches in parallel. This approach is more efficient and scalable than centralizing this functionality in an external database. This provides for a single reference storage repository for all of the reference data of a large corporation that is collected over many years.

Software is responsible for tying the modular storage units together into a single system. The front-end access can be scaled to handle more requests by adding more portal blades in the system, without adding new management software capabilities. This enables a system to be fundamentally self-managing and as a result, there is no need for separate software to configure and control it.

Application integration software

As previously mentioned in the ENSAextended discussion, vendor-specific APIs and industry standards such as the SNIA SMI-S provide the application integration piece that enables the storage applications to communicate with manage and control capabilities of the architecture through ILM policy management, which is discussed later. Application integration is a key requirement for enterprises since business applications create, manage, and associate value to data. It is in this context that ILM becomes important, as this data must be associated with the right application at the right time with the right attributes through user-defined policies. The bottom line is that an ILM solution must be completely transparent to applications and to the users of the applications. It should not matter where the data is being stored as long as it is accessible based on the required business-relevant attributes.

HP believes that a standards-based approach to managing data, and the infrastructure that supports it, is beneficial to its customers. We continue to be a leading proponent of industry standards and are active on many of these standards bodies such as the SNIA. Regarding SMI-S specifically, since the specification is designed for the integration of heterogeneous storage network components and storage management applications, HP has created a SMI-S Developers Program (SMI-S DP) designed to assist the independent software vendor (ISV), independent hardware vendor (IHV), and system integrator (SI) in developing their storage management software or hardware solution—using the industry-standard SMI-S interface—for HP storage system customers.

In addition, HP fully supports the Simple Object Access Protocol (SOAP), which is a key standard Web services technology. Web services are a major HP initiative that is supported through HP Darwin Reference Architecture. SOAP enables a program running in one kind of operating system to communicate with a program in the same or another kind of an operating system by using the World Wide Web's Hypertext Transfer Protocol (HTTP) and its Extensible Markup Language (XML) as the mechanisms for information exchange. SOAP and its related technologies (WSDL, UDDI, and so on) form the foundation for a new type of middleware for application-to-application integration that allows application-to-application integration over the Web. Over time, this will eventually allow other management front-ends and back-ends to interface with HP ILM solutions.

Policy management

ILM policy management makes use of a policy manager and data movement plug-ins as shown in the architecture diagram. For example, Storage Builder has a policy manager that provides:

- A centralized view of users and their capacity usage by host, storage device, LUN, volume, directory, or end user
- Historical trending and future extrapolations
- Reclaiming of old and junk files or old and junk directories (it identifies access characteristics of files meaning it can identify files that have not accessed in a long time)
- Capability to predict storage demands, enabling just-in-time capacity acquisition
- Full integration with other modules of HP OpenView Storage Area Manager suite

Longer term as the architecture evolves, the ILM Policy Manager will provide the focal point for a common data management architecture that enables a common framework for control of both operational and reference data. In addition, a central console is available to specify data quality of service and data policies. The Policy Manager's role is to analyze data in the context of applications and users and help the administrator answer basic questions about the data. For example, the Policy Manager is responsible for understanding the data under its control, the applications it is associated with, which storage container it is in, where it is in its life cycle, and where the data should be at that stage of the life cycle. The Policy Manager also automates data movement actions by setting policies that instruct data movement plug-ins to perform actions automatically based on where data is in its life cycle.

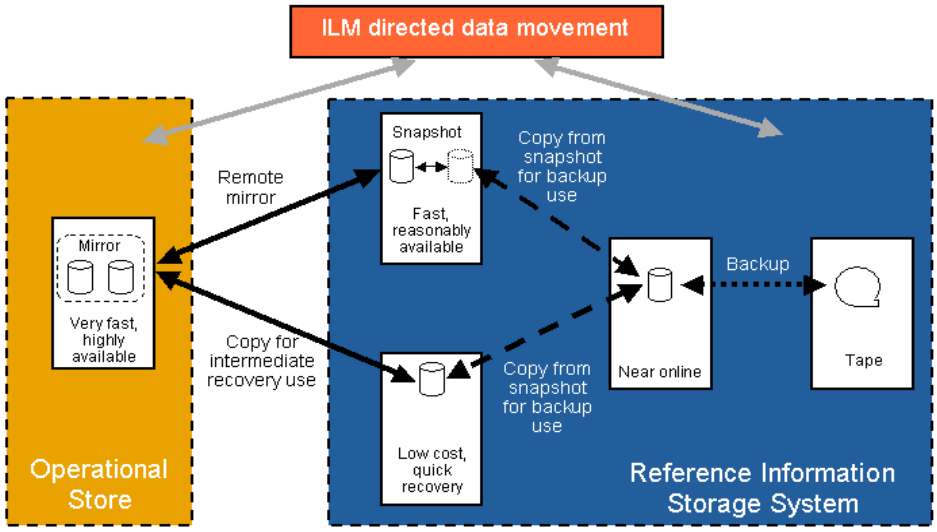
The near term focus of Data Movement Plug-ins is to improve archiving functionality and the management of disk-based replicas. This provides a foundation for adding true backup-to-disk capabilities in the near future.

In general, Data Movement Plug-ins are storage application–aware modules responsible for implementing the directives issued by the Policy Manager. This includes making copies or moving data to where it should be. These actions involve replication, backup, archive, deletion, and data migration (HSM) activities.

The Data Movement Plug-ins also interface with business applications to copy and move data between storage containers to meet customer-specified goals. In other words, the plug-ins provide the integration code needed to interface ILM with business applications. The plug-ins also enable data objects to be defined at application- and user-specific granularity, such as mailboxes or mail messages as opposed to SCSI blocks.

In an ILM solution, there are various types of data movement depending in the policies established through the Policy Manager. Examples of these data movements are shown in Figure 11. In addition to data movement policies shown, it must be pointed out that the data movement is not just between storage boxes, but the migration is among different QoS classes of storage determined by the Policy Manger. As previously discussed, this might even be a single storage box that contains a number of QoS classes of storage.

Figure 11.



One of these data movements is archiving, which moves data to a durable archival repository (with a searchable content index and retrieval capabilities) for the purpose of long-term data retention. However, the data prior to its long-term retention could involve intermediate data movements such as replication for recovery purposes. The data is moved from the operational store to the reference store by way of a replication application such as a remote mirror for recovery purposes.

This is operational data being actively operated upon.

In another scenario, if a replica of a volume is taken for re-purposing (for example, each night a snapshot of our OLTP database is taken to perform data mining), then this copy stays as operational or active. While if the replica is made for disaster tolerance reasons, then the copy will be inactive but still must reside in an operational data store (since it may become operational at any moment).

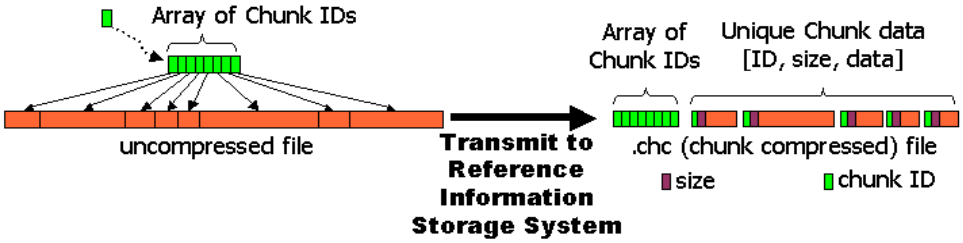
Archive and backup are operations that move or copy data into the reference information storage system. In addition to archive and backup operations, other forms of data movement can be invoked to move static data from the operational store to the reference store. These processes were described earlier.

That is, all ILM data movement operations have been categorized into four major categories: migration, replication management, archive, and backup. In addition to the need for a reference store, we will implement these data movement functions and put in place the management framework to control and automate them.

Chunking

Chunking is accomplished by applying special algorithms to data. These "hash" algorithms identify "milestones," or repeating patterns, in the data. Breaking the data at these milestones creates chunks. When the hash of a chunk is calculated, the Reference Store is queried to determine whether it has seen the hash before. If it has not, the chunk associated with the hash is transmitted to the store. What this means is that decisions about whether or not to place a chunk of data in the Reference Store are based on the movement and analysis of very small amounts of information (the hashes) rather than on the larger quantities of actual data. This is very efficient in terms of elapsed time and network bandwidth requirements. Figure 12 illustrates the chunking and data movement processes.

Figure 12.



If the file changes (even by only a few bytes) or some other process such as another backup job, or an archive job, tries to put that same file into the store, only the new or changed chunks will actually be written. That is, we have very fine grain duplicate detection on objects stored. This radically reduces the amount of data that must be stored. Also, this means that the Reference Store is responsible for:

- Storing chunks and the hash information that associates collection chunks with data (files, versions of files, and so on)
- Retrieving and reassembling appropriate chunks when an application requests specific data

Chunking enables a number of important benefits:

1. Backups take a fraction of the time because they must save only the changed chunks. For file systems, we can now store every version of every file that was ever created (because the difference between all the versions is tiny in terms of used storage capacity). For example, in test cases, backup in concert with chunking was measured at 40 seconds, compared with 7–54 minutes for same data using traditional backup methodologies.[7]
2. Reduces network bandwidth by as much as 99%, again because chunking greatly reduces the amount of backup data that must be moved across the network.[8]
3. Storage capacity needs are reduced by as much 90%, again based on HP testing.
4. Very fast restores for large files because we need restore only the changed blocks, which are then reassembled to restore the original data (assuming original is still online).
5. Lower total Total Cost of Ownership (TCO) for data protection because administrative costs are lower (no manual tape handling, easier access to recovery sources) and network costs are lower.
6. 100% online data recovery even during other operations.
7. No failed restores due to bad or missing tapes.
8. Embedded remote replication to protect against site failures.

Applying ILM to the email problem

This section illustrates a scenario that many customers will face, due in part to regulatory and compliance requirements for saving vast amounts of aged data for long periods of time—to meet some regulations this means indefinitely. Because of its business importance to customers, HP initial ILM solutions deal with archiving data for email. The same University of California–Berkeley study mentioned previously estimated that email generates about 400,000 terabytes of new information each year worldwide[9].

One of the first elements of ILM required by organizations is a mechanism for archiving emails and attached documents and makes them accessible for years to come. In the case of email, the raw data that is recovered in the form of a users Exchange account is not useful until that user utilizes it for a business purpose. It is at this point it becomes useful data or information.

Another consideration is that, since this information must be recoverable in a specified time, it might need to be migrated among storage devices as new storage systems and/or technologies are introduced. It might even need to be readable even if the application that originally created it is no longer available.

As a result, an ILM solution using archiving technology becomes a key ingredient. Archiving is extremely important because it provides:

- A key part of the solution, which achieves legally acceptable compliance for a variety of requirements
- Durable protection against loss or corruption of data (data is stored offline and offsite on durable media)
- A means to reduce operational information disk storage (that is, high cost) needs by providing a cost-effective resource to which static data or reference information is securely archived

The successful implementation of an archiving solution involves both the customer and the supplier working hand-in-hand to ensure that all technology and procedural aspects are covered. After the implementation has been completed, there will need to be an ongoing support to ensure the integrity and maintenance of the archive.

---

[7] HP laboratory test results
[8] HP laboratory test results
[9] How Much Information 2003, University of Californian–Berkeley Study, October 2003

Summary

HP has long recognized that customers use data for a variety of purposes, and that data usage and the business results derived from it are key factors that impact how that data will be stored, accessed, and managed. In addition, HP understands that a set of data may be used for more than one purpose (user, application, business process, and so on). Because of this, HP believes it is important to manage data and its access in the context of business usage and to extend data management to encompass the relevant life cycle for the data.

The impact of current and new compliance and regulatory mandates will have a dramatic impact on how customers manage their data. However, ILM in general and the HP ILM solution specifically will enable enterprises to deal with this requirement more effectively and efficiently.

Regulatory compliance focuses on the retention of specified records for prescribed (customer selectable) periods of time. Retention involves storing records on media whose durability is certified for the required periods of time—or even indefinitely—and whose data integrity is assured. In this context, data integrity refers to the recoverability of data and inalterability—protection against overwriting, erasing, or otherwise modifying or altering the data. Also, the retention process must be capable of verifying that the expected data was in fact written to the expected (target) storage medium.

To comply, archived data must be retrievable in its original form. Thus, the repository in which the data is stored must be indexed and searchable. These are native attributes of the Reference Store. And since the index and the database that drives it form the heart of the archive, the HP solution is designed to meet all of these criteria. It also fully indexes both onsite and offsite storage media. Furthermore, the HP solution is designed to facilitate almost instantaneous record retrieval.

Record retention compliance should also provide an audit trail that shows all queries, retrievals, and administrative operations performed on the retention system. The HP ILM solution provides this while at the same time allows audit records to be maintained for customer-specified durations.

Security is critical when dealing with compliance. For that reason, the HP solution applies a variety of security measures throughout the system. For example, HP supports digital signatures on SmartCell disks. Digital signatures also have the capability to validate stored data against tampering. The entire system requires password access, including access through a Web browser.

How HP will deliver ILM

The HP strategy intimately ties ILM to the business applications it supports. As such, HP plans to roll out ILM in three key directions:

1. A family of application-specific solutions (application-specific integration elements and services that can be used to implement them). The first of these will be focused on email and Microsoft Office document archiving and management. More application-specific solutions will be released over the next several years as new storage technologies, data movement plug-ins, and application modules are created.
2. Evolutionary storage systems will continue to enhance the networked storage infrastructure. New classes of storage systems, enabled by new device and subsystem technologies, will be introduced as indicated in the roadmap in Figure 13.
3. New storage technologies, such as Operational and Reference Store management applications, advanced data movement, and function-specific management modules will be introduced.

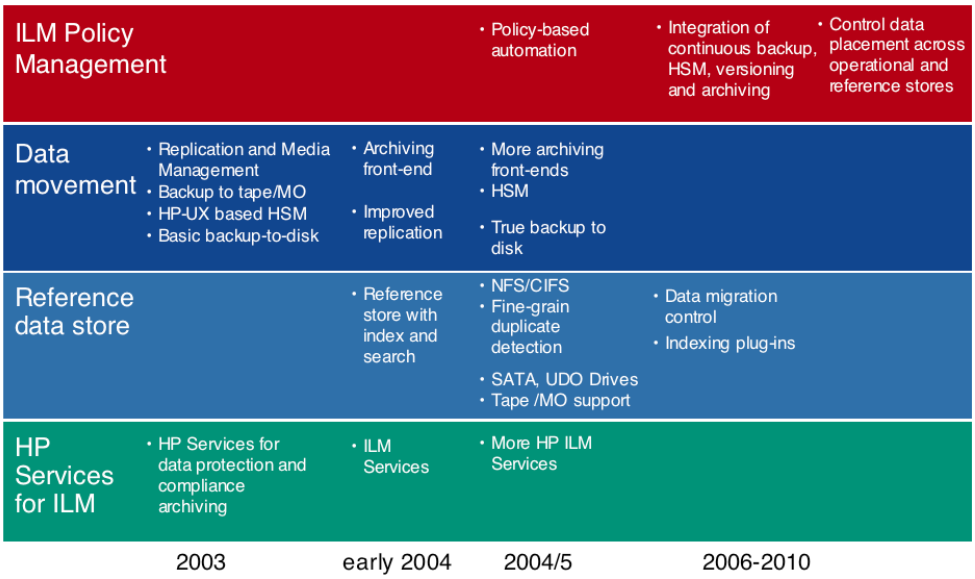
Storage Builder and Storage Data Protector are foundational management applications that will be leveraged as HP rolls out ILM. Storage Builder provides capacity views (in a business application context for applications like Oracle and Exchange). It will be used to provide application-specific host-volume-file-storage array-LUN mapping. These views will be used to track, analyze, and tune space

allocations based on trending information. Data Protector provides managed backup, archive, and recovery capabilities that can be coupled with point-in-time copy and other management functions.

The first solution is specific to email archiving. Subsequent solutions will be focused on other applications like Oracle and SAP, and may be specific to usage like medical imaging, legal record retention, CRM, and ERP.

ILM management capabilities are expected to roll out in approximately the following time frames.

Figure 13.



For more information

HP Information Lifecycle Management web page:

http://h18006.www1.hp.com/storage/highlights/09172003.html

Information Lifecycle Management business brief:
ftp://ftp.compaq.com/pub/products/storageworks/prodbriefhpilmnov.pdf

5982-3398EN, 01/2004