White Paper

We make sure

Information Lifecycle Management (ILM) Overview

Issue July 2006

Pages 9

Contents

1. Summary

Information Lifecycle Management (ILM) has the potential to significantly improve cost positions (Total Cost of Ownership, productivity, etc.) and Service Level Agreements. ILM also provides functions, which help companys to fulfill legal and regulatory requirements as far as archiving is concerned. The following paragraph depicts a concise summarization of ILM.

Information Lifecycle Management (ILM) is a storage management concept that actively manages information objects through their whole lifetime.

ILM is not a product but it is a combination of processes and technologies.

The ultimate goal of ILM is having

- **the right information at the right place**
- **in compliance to legal regulatory requirements and**
- **in accordance to Service Level Agreements (SLA)**

at lowest cost possible.

An optimization process within a rules engine determines the most suitable location for the managed information objects. The most important parameters for the optimization process are requirements of business process on the one hand (value of information, security requirements, Service Level Agreements, etc.) and the cost structure of the storage hierarchy on the other hand. The results of the optimization process are decisions as to where best to store information objects or how to control backup, replication, migration, relocation and archiving functions. For an efficiently functioning ILM, certain advance provisions are required. Virtualization for the online, nearline and NAS areas are just some examples. The separation of the logical view from the physical view puts ILM in the position to optimally place information objects on the basis of process decisions. The white paper therefore attempts to focus on the dynamic processes that make up the actual core of ILM.

2. Driving forces

The idea of implementing effective ILM is by no means new. A lot of functions have been available in the mainframe area for decades. In the past, the approach tended to be less systematic and less formal. The IT and storage environment has undergone enormous change in many areas, which means that a professional, comprehensive concept is now required. The major driving forces are unrestricted data growth, internal and external obligations in the form of legal and regulatory demands, intensifying cost pressure and the technological innovations of recent years.

2.1 Data growth

Businesses and institutions can bear witness to the extreme levels of data growth experienced in many sectors.

According to a study of the Radicati Group the number of received / sent e-mails per user and day within enterprises will rise from about 130 in 2006 to about 170 in 2010. In the same time the average size of e-mails with attachments (which account for 25% of all e-mails) will rise from an average of 420 KB in 2006 to 470 KB in 2010. That means, for each e-mail box 3 TB of data has to be handled under the assumption of 200 working days a year. On basis of the above mentioned growth this size will rise to about 4 TB per mail-box and year in 2010.

Another area experiencing extremely high growth is the health sector. In the Storage Magazine article “Optimize your storage for fixed content, issue may 2003”, it was shown that a Picture Archive and Communications System (PACS) in a large hospital can easily generate more than five Terabytes of digital X-ray or MRI information (Magnetic Resonance Imaging) annually. Even double digit TeraBytes growth rates have been reported.

Similar growth rates can also be found in other sectors. Examples include research and development, video editing, video streaming, Customer Relationship Management (CRM), quick decision (parallel analysis of the content of multiple sources for critical keywords), fraud detection, monitoring, etc. This trend can also be seen in the banking and insurance segment where mainframes have strong positions.

2.2 Legal and regulatory demands

Legal and regulatory demands define how and how long information is stored and in what form it has to be made available when requested. The basis for such demands is formed by legislation or regulation at international, national, communal, association and organizational level. Internal company controls are also important in this context. International presentations and documentation frequently cite American laws and controls including SEC Rule 17a-4, 21CFR Part 11, HIPPA, Sarbanes-Oxley, GoBs, DoD 5015.2. Comparable requirements can also be cited for other countries. In Germany, for example, the “Grundsätze der ordnungsgemäßen Buchführung (GoB)” (similar to the Generally Accepted Accounting Principles in the US), the “Grundsätze zum Datenzugriff und zur Prüfbarkeit digitale Unterlagen (GDPdU)” (principles governing data access and the verification of digital documentation), the “Handelsgesetzbuch (HGB)” (Commercial Code) and the “Abgabenordnung (AO)” (Tax Code) etc. apply. There are also, as is the case in the United States, a multitude of additional regulations. One such example are the controls of the “Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin)” (federal supervisory agency for financial services), which apply to banks, insurance companies and stock trading. There are signs of an emerging trend towards harmonization in Europe, although expectations should not be too high in the short term. It is the duty of those responsible to work together with internal or external auditors, chartered accountants or other authorized experts and carry out binding advise to define the type of storage and retention periods. The majority of demands appear to be technology neutral, which leaves plenty of scope in relation to the technical realization in ILM. One thing, however, is certain – the amount of information to be stored and the length of retention periods will continue to increase significantly.

2.3 The value of data during the lifecycle

The value of data changes over the course of time. This fact can best be illustrated by means of an example involving a simplified analysis of some of the processes in a bank transaction. The most important information is the account balance. This data will always have a high value, as it is from this information that debts and liabilities are determined. A specific business transaction (withdrawal, deposit, transfer, etc.) must be kept online until the bank statement is printed. After this time, the individual transaction no longer has to be held in the online primary storage. The value is, however, required for the creation of monthly business reports or for calculating the balance sheet total. Following publication of the balance sheet total, the individual values play a lesser role and thus have a lower value. However, the value of this data changes abruptly when, for example, monthly incoming payments have to be checked for loan applications, when follow up orders, queries etc. have to be processed, or when audits or tax audits are due. Access to the individual values is required in some circumstances. A permanent change in value therefore does not take place until query deadlines or prescribed archiving times have been passed. It is then at the discretion of the bank to either delete the data or continue to hold it for a longer period of time. In some cases there are also situations where data has to be deleted after a certain period of time. Cases of this nature are often to be found in sectors where monitoring data (e.g. video surveillance) has to be deleted after a given number of days. This may require certain technologies that prevent the data from ever being restored.

2.4 Cost pressure

Reducing costs is a permanent goal. Over the course of time there have only been short periods where cost pressure has been pushed into the background. E-business is a case in point. When e-business first arrived, some businesses made expansion their highest priority. In the end, however, it just didn’t work out for a great many businesses (for the sake of clarity, a clear distinction needs to be made between the unworkable e-business models of the past and the sound Internet services, e-commerce and Intranet/Extranet applications that we have today). Today, the issue of reducing costs is once again top priority. The main reasons may lie in globalization and the difficult economic situation in many regions. Cost reductions must be targeted at the right areas. Costs can be divided into HR (human resources), technology and process costs. HR costs tend to rise permanently if you assume a constant headcount. The percentage drop in technology costs in recent years has reached double digits, based on the same capacity and no change in technology. Process costs are for the most part individual, though they can be strongly influenced by automation. The one commonality is the constraint, whereby more has to be achieved with fewer resources. ILM provides automated processes to help better utilize existing technology. This in turn reduces HR costs, as the right information is automatically made available at the right time in the right place.

2.5 Technological innovation

The emergence of new storage technologies such as ATA disks (Advanced Technology Attachment), which have different price/performance positioning compared with Fibre Channel/SCSI disks, has expanded the classical storage model. Today we have at least three main hierarchies – leaving aside the fact that there are different technologies in the online area (high-end, mid-tier and entry RAID systems) and different technologies in the nearline area (tapes as well as cartridges, CDs, DVDs, etc.). The hierarchy begins with the high performance, high availability, online primary area. The main characteristic of the online primary area is the fact that data is available immediately. Thanks to block level access, access speeds are in the millisecond range. In the enterprise sector, this hierarchy usually consists of FC or SCSI disks with high performance and high availability. Unfortunately, high performance and high availability also come at a high price. Next in line to the online primary area are the online secondary areas. This hierarchy also has direct access feature but, unlike the online primary area, is constructed using high capacity disks with moderate throughput values but lower costs per Megabyte. Normally the secondary online area is based on ATA disk technologies. The notion ATA shall be treated as collective term even if some vendors make use of different technologies. Some of the most important application scenarios are to be found in archiving and in disk-to-disk backup. Storage systems can either be based exclusively on ATA disk systems or they can house FC / SCSI disk in parallel with ATA disks. Due to the ability to freely configure the systems the storage system can optimally adjusted e.g. for OLTP or backup / archiving tasks. At the other end of the hierarchy lies the nearline area with tape technology and much lower price levels per Megabyte. Access speeds in this hierarchy range from a few seconds to a few minutes, with the requirement that the tapes be managed in automatic libraries. Tape technology has undergone major development in recent years, with cartridge capacity currently lying at 200 Gigabytes and set to increase to 400 and 800 Gigabytes respectively in the near future. In addition to the nearline level, it is also possible to have a migration level with shelves and manual access, for example.
The challenge is to make optimum use of the hierarchies with the help of ILM.

3. The ILM concept

3.1 Basic methodology

ILM is not a product, but a combination of processes and technologies. It determines how information flows through a storage infrastructure made up of different storage hierarchies/storage classes. It does not examine individual input/output requests (I/Os), but information objects. Examples of information objects include volumes (LUNs), file systems, files, e-mail messages including attachments, digitized contracts/policies, patient records in hospitals, etc. The processes determine at what time and in which storage hierarchy/storage class an information object is to be available. The storage hierarchy/storage class is determined in an optimization process. Variables for the optimization process incorporate, on the one hand, the value of the information in the information object, the frequency of access, the age of the information object, legal/regulatory requirements, etc. and, on the other, the costs per storage unit, the required access speed and availability, the available capacity, etc. in the relevant storage hierarchy/storage class. The optimization process is a recurring process, which is performed for the first time when an information object is created. The information objects are then reevaluated at regular intervals. If, following evaluation, it is determined that it would be better to store an information object in a different storage hierarchy/storage class, the process initiates the necessary steps (e.g. migration or relocation to a different storage hierarchy/storage class, archiving, deletion of the information object, etc.). Optimization processes can also take place on an event-driven basis outside of the defined cycles.

3.2 External input

The optimization process described above requires additional decision criteria, which need to be incorporated as external requirements (not from the IT department). The most important of these are:

- The value of the information at the relevant time
- The agreed Service Level Agreements (SLA)
- Definition of access rights/data integrity

3.2.1 The value of information

Defining the value of information is a management task that is performed on the basis of the business processes with the mapping of the business logic. It is often the case that the business processes are not clearly enough defined. In such instances, external consultants should be brought in. Once the value of the information has been defined, the protection of the information can also be substantiated. The most important considerations are the demands to be met by RPO (Recovery Point Objectives) and RTO (Recovery Time Objectives), for example.
RPO defines the maximum loss of information that may occur when a fatal error occurs. To put it in fairly general terms, one could ask how many transactions would be lost if either the system crashed or, if the worst comes to the worst, the entire computer center including the information was no longer available.
RTO, on the other hand, defines the time between a damaging event and the point at which the IT system needs to be operational again. In certain sectors, this is not merely an internal decision, but is influenced by association requirements or even by legislation.

3.2.2 Service Level Agreements

Service Level Agreements (SLA) define the service that an end user, for example, can expect when accessing an information object. SLAs should, in principle, be defined for every provision of service. This aspect takes on a greater significance in ILM, as dynamic processes run on the basis of the changing value of the information over the course of its lifecycle. For example, information objects can be moved from the online primary area to slower, more cost-effective storage media if they have not been accessed for a long time. If information objects that have been moved need to be accessed again, it must be clear that the access will be for a longer time, but may not exceed a specified maximum time.

The SLA definitions specify the technology to which the information objects can be moved after the given times.

3.2.3 Access rights/data integrity

Like the value of the data, access rights must principally be defined from a management viewpoint. For example, it must be decided whether data should only be accessible internally or whether it should also be accessible externally. In reality there are, of course, different internal user groups with different access rights. ILM must take these requirements into account so that they cannot be violated by ILM actions. After data has been migrated, for example, the same access rights that were in place before the data was migrated must continue to exist. The same applies to backup and replication etc. Archiving follows different rules. After arching information object nobody should be able to delete or modify an information object before the expiration date has expired. If modifications are necessary a new version of the object has to be created. The old object must not be altered in any circumstance. Protection demands can change during the lifecycle of an information object.

3.3 Storage hierarchies/storage classes

ILM can work more effectively the better the entire storage area is structured. This is normally done through storage hierarchies and storage classes.
Storage hierarchies are produced as a result of the classification of storage technologies in ascending or descending order according to a technical or cost viewpoint. Technical aspects and cost viewpoints are often mutually dependent. For example, storage systems with fast access are in a higher price range than storage systems with slower access.

Storage classes are the result of the logical structuring of the storage area. Storage hierarchies can thus be structured in more detail. In addition, it is possible to make a distinction based on the number of copies of information objects to be created automatically, even when the storage objects are in the same storage hierarchy.

3.4 Structuring the data

In the current model we have evaluated the data globally and defined the containers (information objects). This framework needs further refinement, as otherwise the highest level of requirements would have to be applied to data that does not have such requirements. In structuring the data, we divide the entire dataset into logically connected subareas. The easiest way to structure data is by area of responsibility (or department). Examples include development, purchase, sales, marketing, HR, etc. Data can also be structured according to user group. An orthogonal division is created through the separation of “structured” and “unstructured” data (see section “Driving forces”). Structuring according to application appears more practical, as applications are often used at inter-departmental level. The model becomes more interesting if the vertical procedures can attribute individual data objects. It would be ideal if, for example, a process on the SMTP server could determine whether incoming e-mails are tax-related. It would therefore not be necessary to archive all e-mail processes as a precaution.

Especially in the file system area first approaches for automated data classification have arrived. In these solutions appliances are placed into the storage network which assess step by step all files according to defined rules and trigger necessary actions. For instances the appliance can scan all files if the string “confidential” is contained in the content. If such a file is stored on a open accessible filer mistakenly then the appliance could move this file from the open accessible filer to file with controlled access rithts.

By structuring the data, it is possible to define different processes and optimization points for the different data areas.

3.5 Benefits

The aim is to guarantee optimum performance at enterprise level by storing data in different storage classes according to their importance and in compliance with the legal / regulatory requirements and in accordance to SLAs at lowest cost possible, without unduly preventing users from accessing the data. Key variables for the optimization process are Service Level Agreements (be they of an internal or external nature) and the available budget/configurations.

The concept is based on the idea that the value of most information and, consequently, the access models for the information objects change over the course of time. Automated processes help ensure that the information objects are available at all times in the right place at the lowest possible cost. This therefore reduces investment costs.

Hardware costs, however, account for only a small portion of total costs. Different Total Cost of Ownership analyses estimate this portion to be only one third to one fifth of total costs. The remaining costs are infrastructure, administration and operating costs. The punctual archiving of information objects eases the burden on the online primary area, for example, for which costly backup processes are executed. Archived information objects need only be saved once in principle. Archiving can therefore reduce daily, weekly, or monthly backup volumes. The situation is even more serious when it comes to restoring data. If old information objects are not removed from the online primary area, all objects have to be reloaded before the application can go online again.

ILM provides the basis for guaranteeing Service Level Agreements, as the dynamic process always ensure optimum utilization of the individual storage classes. Service centers can forward these Service Level Agreements directly to their customers. When complied with, Service Level Agreements also have an effect on customer loyalty. Customer inquiries, whether they are made directly over the Internet or indirectly through employees, can be responded to in the shortest time possible. Customer satisfaction increases accordingly.

4. Advance provisions

4.1 Consolidation and SAN/NAS infrastructure

From a conceptual point of view, it doesn't matter with ILM whether the IT configuration is operated in a decentralized or centralized model as long as there are communication paths to the remote sites. But the communication overhead is much higher in a distributed model compared to a consolidated model. All consolidation potential should therefore be utilized before introducing ILM. Consolidation may be the key technology when it comes to, on the one hand, creating coherent storage hierarchies and, on the other, keeping the number of storage classes to a reasonable level.
Practical examples of effective consolidation can be found in the e-mail sector. With Exchange Server 2003, for example, several hundred e-mail servers can be concentrated on just a few servers. It is much easier to connect a migration or archiving level with a few servers than it is with several hundred servers that may be distributed over long distances.

Intelligent storage networks (SAN, NAS, IP Storage) make implementation of dynamic ILM decisions at operational level much easier. In intelligent storage networks, every server has access to every storage area in principle. This creates the requirement whereby, for example, the most suitable storage class has to be selected for the first assignment.

4.2 Virtualization

Effective ILM assumes problem-free access to resources. When deciding which technologies to use for storing information objects, technical attributes should play only a minor role. This is achieved today by means of virtualization. A distinction must be made here between virtualization concepts in the online and nearline and NAS areas. In the online area we have solutions based on servers, networks and storage systems.

A detailed description of the different virtualization concepts is available on the white paper “Virtualisation, basic building block for dynamic IT”

4.3 Storage Resource Management

In Storage Resource Management (SRM), the characteristics of available storage systems, physical disks, logical and virtual volumes, file systems, files, database containers, database contents, etc. are recorded and stored in a central, structured database for long-term administration. This forms the basis for successful cross-server and cross-platform data consolidation and the evaluation of storage utilization for trend analyses, threshold value monitoring or usage accounting. The storage classes and the sizes required can be derived from the data.

5. The ILM model

The ILM model consists, on the one hand, of process instances and, on the other, of the transition matrix. Standard process instances include initial allocation, data backup including recovery, replication, migration (HSM), archiving, relocation and deletion/destruction of information objects.

In **initial allocation**, ILM specifies which storage class an information object is to be stored in. Today, this is still mainly done by means of a fixed assignment of file system to virtual volumes or LUNs. However, there are also file systems that already automate this process.

Conventional data **backup** at basic and file level is normally based on a version schema. In other words, the information objects are copied to another physical medium at regular intervals without destroying the previous copy. This allows access not just to the last version of an information object, but also to older versions. Databases and applications provide restore functions up to a specific transaction status using the log files.

Replication is available in the form of “continuous mirroring” and “point-in-time” copies.
With continuous mirroring, every write operation is performed either synchronously or asynchronously in the mirrored dataset as well. Data that is modified in a dataset is changed in the other dataset as quickly as possible. The original dataset and mirrored dataset should always be linked.

Point-in-time copies create a logical copy of the dataset at a specific point in time. Even when changes are made in the original dataset, the copy remains in the same logical state that it was in when created at the specific point in time.

High flexible configurations with the highest requirements for RPO / RTO need additional technologies in the form of Continuous Data Protection which surmounts Point-in-Time Copies. Continuous Data Protections extends regular Point-in-Time Copies by enabling a reset process to a discretionary point in time in the past. Going back to a discretionary point in time is realized that for each write action the new data including a time stamp and the old data is stored. Continuous Data Protection can be implemented either in the file system or on the block level.

Migration moves information objects form one storage class to a different storage class without changing the access syntax. Information objects can, for example, be migrated to the online secondary area or to tape libraries. System integrated migration will first copy the information object to the new storage class and then convert the original object into a pointer that refers to the copied object in the other storage class. This has the advantage of the application or user being able to use the same access syntax as if the file were still in the online primary area. The only difference from the end user's point of view lies in the access time, as the system must first recall the information object in the original storage class in the event of it being accessed. Migration processes are transparent to the application.

Migration is to be disassociated from systems where a linear file system that behaves like an online file system is realized, but information objects are written quickly to offline storage classes. These systems do not recognize policies, but instead use the online storage as a cache. It is not possible to make a deterministic forecast as to whether an information object still exists in the cache. Probabilities can be specified on the basis of the cache size and access behavior if necessary.

Archiving is the targeted storage of an individual information object or a related set of information objects in a different storage class and the definitions specifying the length of time over which the information objects cannot be deleted. A key feature of archiving is the additional storage of logical metadata for each archiving request. The metadata can either be created manually or automatic indexing can be performed. Retention periods are specified on the basis of the business processes. For example, German tax law prescribes, in addition to many other deadlines, a typical retention period of six and ten years. Depending on the sector, there are also much longer retention periods (e.g. design drawings or approval documentation for a nuclear power plant). The ILM processes must guarantee compliance with these deadlines.

The **relocation** of an information object is, in principle, an irreversible process. Relocation is an important process instance when old hardware is to be replaced with new hardware. Information objects are principally LUNs or volumes. With relocation, the challenge lies in the transparent execution of the process, which usually takes a long time and following the transparent change of the addressing mechanisms. Following a relocation, all areas are released on the original.

At the end of the lifecycle of an information object there are the **deletion and destruction** process instances. With deletion, the space is only logically released. Destruction means that the data areas are overwritten many times with specific alternating bit patterns so that the possibility of data reconstruction (even with sophisticated physical analysis tools) is ruled out.

The dynamic process is described in the **transition matrix**. After initial allocation, all transitions into all other states are possible. A backup or a replicate can be created from an information object and the information object can be moved, relocated, archived or deleted. After deletion, especially after destruction, transition to another state is no longer possible, while an initial allocation can once again be derived from saved or replicated information objects. Migrated or archived information objects can once again be made available for initial allocation. Information objects often go from the backup state to the replicate state, whereby the data backup is automatically replicated after being written to a backup medium. There are also other possible transitions. For example, information objects can be moved from the backup area or archived. For information objects with fixed content, initial allocation in the archive area is also possible. Special systems are suitable for information objects with a fixed content. They provide the basis for compliance with legal and regulatory demands in the archiving area.

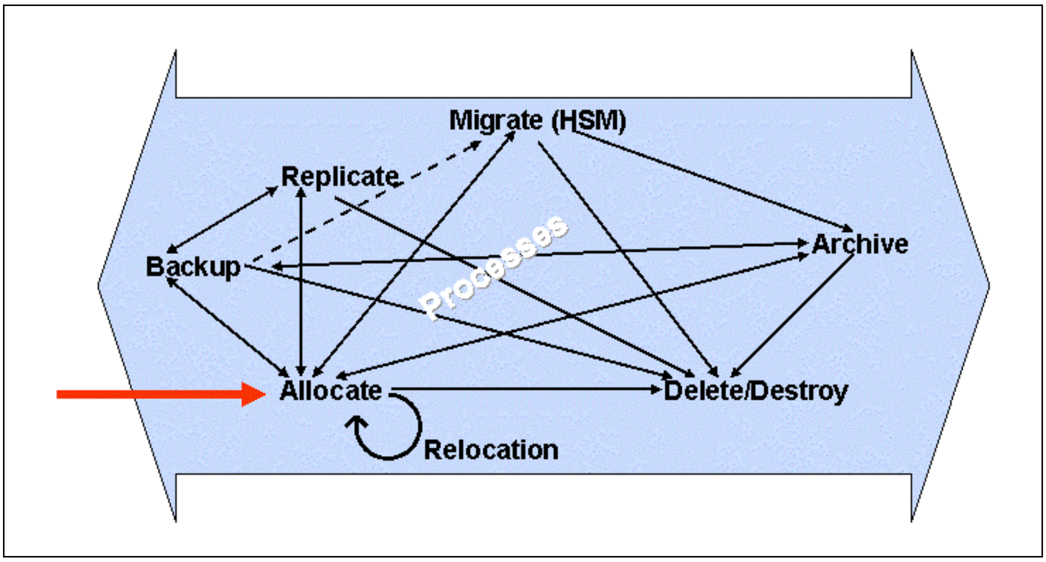

Figure 1: Transition matrix

6. Conclusion

The industry is implementing ILM solutions with emphasis. Although there are already partial solutions available, an integrated solution as described in the theory behind the concept is still a few years away. Partial solutions show that the industry is already on the way towards an overall solution. Partial solutions can more than likely utilize a high percentage of the potential that currently exists. **There are therefore few arguments against beginning ILM immediately.** The key is implementing business process requirements. A great deal more integration is still required here. It is often difficult to know all of the legal and regulatory demands on the one hand and correctly estimate the value of the information object on the other. This becomes all the more difficult when group or departmental interests differ within an organization. Neutral advice from an external expert can help in this case.

Published by department:

Simon Kastenmüller
Phone: ++49 (0)89 3222 1934
Simon.Kastenmueller@fujitsu-siemens.com
http://www.fujitsu-siemens.com/

Extranet:

http://extranet.fujitsu-siemens.com/