A SUN STORAGE TECHNICAL BRIEF:
DATA PROTECTION

The 8 Steps to Cutting IT Operating Costs Without Sacrificing Data Protection

March 2007

Abstract

For the IT organization struggling to find ways to reduce costs, here's an essential starting point: data classification. By rethinking your company's approach to saving data over the long-term, you can make considerable cuts to your IT storage expenditures — without lowering the amount of data protection you have in place.

This paper highlights eight steps your organization must tackle when classifying data on the way to implementing a more cost-effective information lifecycle management (ILM) solution. The "ILM cube" is an innovative approach to lifecycle management, providing tiered storage solutions that align the cost of protection to the value and age of the data.

Table of Contents

1. Executive Summary

Data classification, in tandem with an information lifecycle management strategy, is a critical step in the struggle to drive down IT operating costs. Key to information lifecycle management is the identification of the value of data and the assignment of the appropriate level of storage based on that value. Prioritizing data according to its critical value allows an IT organization to optimize data storage and realize potentially sizable cost savings.

The eight-step data classification methodology outlined in this paper can be applied to any data generated in your organization. The result allows you to apply a tool called the "ILM cube", which can help your organization make appropriate storage decisions — premium disk versus Advanced Technology Attachment (ATA) tape, for example — based on data criticality.

2. Data Classification Benefits

The eight steps to the data classification range from determining what data is and isn't important to capturing that information in a software-based solution.

By using the value of data to guide placement in the storage farm, your organization can realize several competitive advantages, including:

- Improved return on investment (ROI) resulting from more intelligent spending on storage expenditures
- Improved data performance, integrity, and availability because appropriate controls are used across the enterprise
- Improved information protection because mechanisms are designed and implemented based on data value
- Improved hardware fulfillment processes because data classification at information "birth" delivers better lifecycle management
- Reduced management effort due to consolidated and standardized storage support and processes

3. Practical Implementation Steps

The following data classification process can help your organization obtain the analytics required to fully and systematically apply the ILM cube framework.

3.1 Determine Information Classification Inputs

You begin information classification by looking at a range of requirement inputs. The first six inputs relate to initial information requirements (those at data birth). There will be a discussion of the ILM cube itself later in this paper. But for now, the essence of the approach is that — based on your requirements — you assign a value to each input. You then combine and weight those inputs to determine the storage tier for the data using precious-metals descriptors (bronze for less critical data through platinum for the most critical). Step 4 discusses techniques for assigning values to these inputs. You may or may not use all of the inputs described, or you may have special needs to augment with other inputs. The inputs that follow are submitted merely as a starter set to consider.

The ILM cube relies on several inputs to help you arrive at appropriate storage solutions for mission-critical and other data.

3.1.1 Security Protection — Based on security organization requirements

What degree of access to the data does your organization need? How isolated does the data need to be? For example, real-time, mission-critical applications, such as those found in financial institutions and online gaming, use internal firewalls to partition information. Other issues to consider here are needs for cryptography, integrity (for example, write once, ready many [WORM] characteristics for archival information), and overall confidentiality.

3.1.2 Performance — Based on input from application developers

Those who actually create the applications that generate, store, and retrieve the data must specify input/output and nominal response times for that data. In other words, how quickly does the data need to be accessed? Issues associated with access include contention and blocking. If high performance is important, make sure the system will allow for sufficient concurrent data requests.

3.1.3 Availability — Based on application user requirements

For those who need to access the data to perform daily duties, continuity of delivery is a primary concern. This delivery is normally outlined in service level agreements (SLAs) that describe the guaranteed availability of data. Those SLAs will help guide you as to the data's criticality.

3.1.4 Recoverability — Based on operations and disaster recovery requirements

You rate — (assign values to) — each appropriate input to determine how critical the data is to your organization.

Recoverability describes just what data your organization needs to be able to get back in the event of a problem. There are four common components to recoverability:

- Recovery time objectives (RTO); in other words, the maximum allowable downtime resulting from data loss
- Recovery point objective (RPO); the point in time to which you can return in your backups (quarterly, weekly, daily, hourly)
- Recovery capacity objective (RCO); this relatively new term describes the percentage of initial post-disaster recovered capacity compared with the normal production environment capacity
- Granularity of recovery (volume level, file level, row level, and so on)

3.1.5 Scalability — Based on infrastructure and budgeting organization's requirements

Scalability addresses the anticipated growth of the file or database being stored. Will the amount of data be very dynamic or largely static? Overall, organizations plan for about 30% size increase over time, but this number can vary widely by industry and application. Key considerations here are providing adequate capacity management for both hardware and software and planning for adequate expansion of storage networks to accommodate growth.

3.1.6 Application Profile — Based on business owner's requirements

The application profile takes into account the use to which your organization puts the data being stored. For example, identify whether it is internal, customer-facing, or part of a business-to-business relationship. Depending on the data's use, it may be subject to internal information policies and standards or even to regulatory compliance. If data is mission-critical, consider assigning it a higher rating. Data required simply for compliance that may never be accessed can be given a much lower rating.

3.1.7 Lifecycle — Based on aging patterns established by developers

This last element addresses the value of the information as it decays over time. Whether you plan to keep data for 60 days or 60 years is a decision you should make at data birth. However, it is also one that you can reassess over the data's lifetime. In making storage decisions, take into consideration:

- The age of the data you are currently storing
- Patterns of data access
- Archive duration

Step 4 is where you use business impact analysis to determine exactly how important certain data is and how to measure that value.

As the data ages, consider deep archive requirements and deletion schedules (what the federal government calls a "disposition schedule"). You also need to look at various data archive services appropriate to the aging data, including SLAs and pricing.

3.2 Identify Business Owners

In identifying business owners, apply a line-of-business focus, and always keep in mind that data ownership is not an IT responsibility. However, gathering data about business functions, applications, and how the application uses IT is key. A business impact analysis (BIA) process can be helpful in identifying business owners.

To ensure an actionable plan, stakeholders at several management levels must be part of the process:

- Senior line of business management support is vital to success
- Chief information officer (CIO) and chief technology officer (CTO) sponsorship helps ensure adherence to new information lifecycle management policies.
- IT management must enact policies that give data owners authority commensurate with their responsibilities and business requirements.

3.3 Inventory Applications by Business Unit and Type

Inventory is a two-step process:

1. Inventory and categorize applications by department.
2. Organize the categories in terms of the inputs chosen from Step 1.

Poll the application owners and — if it hasn't already been done — create an application requirements analysis. The analysis should cover IT facilities and application systems (e.g., call center operations and major IT applications: CRM, ERP, supply chain, internal and Internet email systems, and so on).

3.4 Use BIA Methodology to Assign Data Value

A BIA is an excellent way to define and measure the information value of data to your organization, as well as to establish business continuity requirements. You can use a formal BIA or informal BIA-like classification process to:

- Identify the critical applications and quantify, in monetary terms, the impact of loss
- Identify critical business functions and the computer systems required to support them
- Identify the impact on other systems and lines of business if the computer systems required to support these critical business functions are not available
- Quantify the monetary impact if critical business functions cannot be performed due to a loss of required systems (cost per hour of downtime)
- Identify intangible impacts that would result if critical business functions cannot be performed due to a loss of required systems (e.g., loss of public confidence, violations of regulations and tariff agreements, damage to future business)
- Determine the recovery time frames (RTO, RPO) required of the systems supporting their critical business functions in the event of a loss of these systems
- Identify the adequacy of existing business continuity plans, systems and processes for each critical business function and application

Applications should be mined for risk and business dimensions such as:

- Operational impacts
- Financial impacts
- Current state of preparedness
- Available resources
- Critical files
- Critical time frames

Once you have this information, use the BIA to assign levels of criticality to the inputs in Step 1. You can then use these levels to help assign the initial storage tier in the ILM cube. The following illustration shows an example of how you might categorize severity levels for data loss.

3.5 Establish Data Management Goals

Begin by setting up a cross-functional information lifecycle management steering group (described in more detail later in this paper). The team should clearly define an unambiguous and simple set of data classification goals. These goals should produce a framework and model that can then be used to drive information lifecycle management strategies.

To help with goal setting, first create a high-level classification matrix — a horizontal view of the application of enterprise data, business processes, business objects, and logical state and physical state classifications. Questions to ask include:

- How many different types of business data does the organization support?
- What are the possible logical groupings of data?
- What is the longest duration we must keep information?
- What storage assets do we possess?

Close strategy sharing between the storage administrators and the database administrators across functional groups (i.e., production, development, and quality control) will facilitate creating an effective information lifecycle management framework.

3.6 Define Storage Tiers

Based on the results of Steps 1 through 5, your organization should have enough background to formally create a set of storage products by tier (typically kept to four to five for simplicity). Experience has shown that the cube structure is a practical way to convey information lifecycle management methodology. Certainly, similar frameworks have been employed before, but the ILM cube is unique in that it ascribes initial storage attribute levels to each tier as well as changing attributes as data value diminishes with age (a time component).

Note that based on requirements, some services or applications may fall into more than one tier. For example, certain data may require subsecond response time but does not need a mirrored solution. You will then need to make a judgment call about the best storage solution.

Tier 1 Platinum (real-time mission-critical) storage is configured to include mirrored enterprise-class disk, as well as remote, replicated, synchronous disk arrays. Many, but not all, financial or online applications fall into this category (e.g., you would probably move audit logs and reports to Tier 2 or Tier 3).

Tier 2 Gold (mission-critical) storage, which also includes enterprise-class disk, is configured to include local mirrors but not synchronous remote replication. Some online and many corporate applications (depending on their availability and RTO requirements) fall into this category.

Tier 3 Silver (mission-supporting) storage could be fulfilled with RAID 5 SAN arrays. These should be enterprise-class disk, but not requiring RAID 1 mirrors. Most corporate applications fall into this category, again based on availability and RPO requirements. Also, note that an opportunity may exist for migrating development and quality control applications from Tier 3 storage down to Tier 4.

Tier 4 Bronze (standard) storage may be fulfilled with lower-cost serial ATA fiber SAN arrays, iSCSI SAN arrays, or network-attached storage arrays (NAS) on gigabit

Ethernet or newer network transport technologies like iSCSI (non-SAN). Applications that fit into this category include reports, query extracts, and most development, email, and certain quality control applications.

Employ archive tape — such as access-centric, virtual, or performance-automated tape — for all application data deemed worthy of archiving (and thus requiring quick, but not instantaneous, restore time). For example, many organization members spend a great deal of manual effort to individually archive emails. This process could be automated at the archive tape (AT) tier. Very large databases used for queries can often leverage archive tape for cost savings. Other applications that can fit nicely onto AT include fixed content and certain data mandated by compliance regulations.

3.7 Define Lifecycle Classification and Data Movement Process

As information ages, data access patterns change. Most accesses occur within the first day or two of data creation. After one week the data is rarely accessed, and after 90 days the data is almost never accessed.[1] Therefore, different solutions are required for data and applications at different stages of the lifecycle.

The ILM cube provides storage tiers that address specific SLAs for inputs such as performance and recoverability. Simply rate your data based on the inputs to see what solution is most appropriate.

The ILM Cube

Initial storage attribute assignment

COST

Time

Storage tier	Protection	Performance	Availability	Recoverability	Scalability	Application profile	
Platinum Tier 1 (high-availability [HA] premium disk)	Remote synchronous, dual mirrors Recovery point objective (RPO) <5 min	Fast I/O response >800 I/O per second >2 sec response time	Enterprise class >99.999%	Cross-site replication RTO <5 min	High dynamic	Real-time mission-critical systems	Archive performance tape
Gold Tier 2 (Enterprise disk)	Mirrored RPO <15 min	Fast I/O response >800 I/O per second 2 sec response time	Enterprise class >99.99%	RTO <15 min	High dynamic	Mission-critical systems Corporate	Backup capacity tape
Silver Tier 3 (midrange disk)	RAID5 RPO <24 hours	Moderate I/O per second 100<> 800 query	High >99.95%	RTO <24 hours	Moderate change	Mission-supporting systems Corporate, QC	Deep archive
Bronze Tier 4 (serial ATA, NAS)	Alternate site recovery RPO <24 hours	No I/O rate SLA	Medium >99.99%	RTO <36 hours	Low Relatively fixed	Corporate, development	Densely stacked tape

Postst-ILM Storage Tiers — Note that the tiers and solutions shown are samples that would need to be adapted to fit your particular storage and SLA requirements.

1. The likelihood of data reuse is directly related to age. Storage expert Fred Moore of Horison Information Strategies has studied data retrieval patterns, finding that access activity declines sharply over the first week following creation of a file; after one month, the information is rarely accessed. According to Moore, "The probability of reuse of data has historically been one of the most meaningful metrics for understanding optimal data placement." (Storage Navigator, 2005)

Consider changing from a policy of treating all primary disk data the same — regardless of age — to a policy of leveraging the financial benefits of data migration as aging patterns dictate. The third axis on the ILM cube proposes three solutions for migrating and storing data as it decays:

- Access-centric or virtual tape — for archiving frequently requested information
- Capacity-centric tape — for backups that are not often referenced
- Densely stacked capacity-centric tape — for long-term and deep archive data, possibly using automated virtualization techniques

Time is a major component of the ILM cube. As data ages, its value may or may not decrease, and the storage solution must adapt accordingly.

Densely stacked data should be refreshed (placed on new media) at reasonable intervals to ensure that the media and hardware are available for data retrieval. Seven years is the current standard. Be sensitive to the life expectancy of media and your organization's ability to read it as technologies change.

The migration process will vary depending on the data type — file-based or database. Regardless of the data type, any migration data management tools you use should be as efficient, automated, and seamless as possible. The tools must include contingency plans to ensure retrieval of long-term archived data. Also, as was mentioned earlier, plan for the ultimate destruction of data based on business requirements.

3.8 Set Initial Classification and Lifecycle Using ILM Cube

Much of the information about data classification discussed previously has been couched in theoretical terms. But how is this classification actually captured in a way that is useful to your organization? The answer is with some kind of information lifecycle management data classification software. Such tools are available commercially and range from sophisticated Microsoft Excel spreadsheets to proprietary software solutions.

These tools should be customizable to address your particular data storage needs, but the ultimate goal is always to tier your organization's information into the array shown in the ILM cube. Get into the habit of using the ILM cube as a starting point for each new application coming online.

4. Implementation Approach

Applying a methodology as extensive as information lifecycle management is not a trivial undertaking, nor is it an overnight process. To be effective, it takes the cooperation of the entire IT organization.

Some recommended changes could be implemented with little effort, such as reassigning log files to a lower tier of storage. Others involve a more technical approach,

Specialized program groups are a critical organizational tool for managing and implementing information lifecycle management.

like implementing a virtual tape appliance or assigning a flash copy volume from the enterprise storage array to a modular storage array. Still others require a more politically sensitive coordinated effort, such as when moving development from Tier 1 to Tier 2 storage. In any case, executive management support and direction are critical to reaping the benefits of information lifecycle management.

4.1 Using ILM Program Groups

All organizations have a unique set of applications as well as unique business requirements. These requirements comprise the considerations your organization will need to take into account when evolving its storage infrastructure to align with changes in the storage solution landscape.

Instituting dedicated program groups to ensure that data classification is an organizationwide effort is key to any information lifecycle management implementation. Application standards across organizations are frequently not enforced, or they may be affected by leadership changes, mergers, and/or takeovers. Similarly, SLAs required for each tier may be interpreted differently from group to group within an organization.

To centrally manage the data classification effort and overcome these problems, consider implementing a dedicated information lifecycle management program group under the sponsorship of the CIO and business division heads.

Such an information lifecycle management group could be further broken down into subgroups to handle both architectural issues and issues specific to the application or business:

- Data classification — Focuses on applications and leads in implementing the eight-step data classification process. This group will also synthesize the outcomes of other groups to deliver the holistic information lifecycle management outcomes to the organization.
- Service level management — Focuses on the definition and promotion of practical storage SLAs and the associated management disciplines and tools required to measure and track performance.
- Data backup and recovery — Many organizations can find major optimization opportunities in the area of storage utilization. This group focuses on implementation of data backup and recovery optimization practices. The outcome from this group would form an important ingredient in service level management.
- Data archival — Focuses on the time-oriented dimension of the organization's storage architecture. The group would uncover opportunities in storage optimization by looking at the value of data from a time-from-birth perspective.
- Business recovery — Begins conducting a BIA and addresses the associated storage implications.
- Storage technology — Focuses on assessing existing and future technology for application within the organization.

5. Summary

This paper highlights the tasks your organization must tackle when classifying data prior to implementing an information lifecycle management solution. Additionally, it introduces a slightly different way to conceptualize lifecycle management — the ILM cube. This cube provides tiers of storage solutions appropriate to the value of the data. Additionally, it provides guidance for managing data as it ages in accordance with established service level agreements.

To learn how Sun's leading-edge data storage solutions may substantially reduce your IT operating costs, contact your local Sun representative. To find your local Sun representative please visit www.sun.com/contact.

Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 USA Phone 1-650-960-1300 or 1-800-555-9SUN (9786) Web sun.com

Printed in USA SunWin #503284 Lit. #SVWP12700-0 05/07