Data Classification:
Key to a Successful Tiered Storage Strategy
Data Classification: Key to a Successful Tiered Storage Strategy

A “Key to a Successful Tiered Storage Strategy,” a simple yet insightful guide, is easy to read and comprehend. With a wealth of information out there, it delivers straightforward information that highlights key topics in a quick read.

Introduction ................................................................. 3
Why Tiered Storage? ....................................................... 3
Structured Approach to Tiered Storage ............................ 5
Data Classification ............................................................. 6
Information Category Assignment ........................................ 8
Tiered Storage ..................................................................... 8
Introduction

With storage growing at 50% to 80% compounded annually the expanding volume of inactive data is a concern for companies of all sizes. Analysts and editors agree that 60 to 80% of all stored data is inactive. Data, once it’s created, is rarely accessed after one to two months. Consequently, the ever-increasing demand for storage capacity creates an opportunity for a tiered storage approach.

IT organizations need to find ways to manage storage and data to get the best value from their storage investments. Historically, storage managers concentrated on safeguarding their organization's data and ensuring its availability. A common approach was to view data as a monolithic entity which grew too fast and consumed ever increasing capacity.

While companies recognize that not all of their data is of equal value, storage managers generally viewed all data as equal. Consequently systems and behaviors reflected that view. All data on similar storage. One backup model. A single tier of storage with the same protection level is easier. Implementing a tiered storage approach requires a significant investment in time and resources. There was no delineation among critical, important, and unimportant data.

Helping storage managers recognize those differences and do something about them is at the heart of tiered storage architecture and its driver, data classification.

Data classification is not tiered storage. Data classification is the decision making process that identifies data and determines its value to the organization. Tiered storage is hardware, software, and processes that implement data classification plans.

Data classification is pointless unless you intend to go to a tiered storage architecture. It’s virtually impossible to place the right data in the appropriate storage tiers without first classifying your data. One way to look at it is that tiered storage is implementation of storage decisions made during the data classification.

Why Tiered Storage?

Tiered storage is basic reorganization of corporate data assets onto a variety of enterprise storage media. It involves selection and implementation of:

- storage systems
- software to manage and optimize storage
- policies and procedures needed to operate each tier
The drivers for implementing tiered storage are well known:

- rapid continued growth of data at 50% to 80%
- up to 80% of data is inactive
- arrival of lower cost, adequate performing SATA drives
- maximizing storage investment -- $/GB is decreasing yet still a significant budget factor
- changed data rules – compliance and legal requirements forcing increased volumes of data retained for longer periods

Achieving a functional tiered architecture requires a significant amount of work by IT staffs. Storage administrators must classify applications and associated data, determine which type of storage device to use for each, and manually configure servers and storage to implement this new scheme. Then they need to monitor and re-allocate storage to ensure optimum data placement.
Structured Approach to Tiered Storage

There are a variety of existing and emerging tools to help with this task. Whether manual, automated, or a mixture of both, the process consists of:

- Evaluation – using agreed upon criteria to establish a method of objective data assessment that leads to classification (in this paper we use SNIA criteria)
- Classify – define levels of information value – we use three (Primary, Secondary, Archival)
- Assign – match classes of information to appropriate storage tiers – we use three for simplicity
- Protect – implement appropriate data protection schemes that make sense for the data value

Figure 1 shows a graphic representation of this process.

Figure 1. A structured approach to tiered storage includes logical progression of data evaluation to classification to storage tier assignment. Adapted from SNIA Data Management Forum.
Data Classification

Before you can design a tiered storage environment you first must decide how to segment or classify your data. Data classification is primarily a manual, human process. Software does exist to help discover and evaluate information assets. However, no tool exists that can tell you what information is worth to your business.

Each company must answer the information value question for itself. Defining data categories according to their level of criticality is essential. This is a non-trivial exercise that should involve senior business managers in the discussion. From an initial two or three person effort (which may be perfect for some organizations), an enterprise team usually includes IT, information management, information security, finance, business and legal. Additional corporate departments may also become involved at some point during the classification process.

While there are a variety of ways to classify data, a good example comes from the Data Management Forum of the Storage Networking Industry Association (SNIA: www.snia.org). SNIA defined information classes in straightforward categories:

- Mission-critical
- Business vital
- Business important
- Important productive
- Not important
- Discard

These categories are useful in determining storage tiers. They are also useful in a broader Information Lifecycle Management (ILM) context and have relevancy in security, business continuity, and disaster recovery aspects.

Information can be evaluated on a number of factors to objectively place it in the above categories. To determine how information should be categorized, the following evaluations can be made:

Usage pattern of the information:

- Frequent
- Regular
- Periodic
- Occasional
- Rare
- On demand or request
- Never

Information availability requirement:

- Immediate
- Reasonable
- Defined time frame
- Extended time frames
• Limited
• Not defined or unnecessary

Financial impact of information unavailability:

• Significant and immediate
• Significant long- and/or short-term
• Potential long-term
• Possible
• Unlikely
• None

Operational impact of information unavailability:

• Significant and immediate
• Significant over time
• Probable over time
• Possible over time
• Doubtful
• None

Compliance impact of information unavailability:

• Definite and significant
• Eventual
• Probable
• Potential or possible
• None expected

Classifying existing data and maintaining policies on data movement is time-consuming and resource-intensive. Plus, it’s frequently difficult to get business leaders to give IT input in the data classification process. This often prevents companies from implementing a multi-tiered storage approach.

It’s also useful to remember that while information classification is a necessary step toward matching storage tiers to specific business goals, security and other facets of the information need to be integrated into the classification model as well.

The good news is that there is an emerging group of independent data classification software vendors. We include a sample here for reference purposes only.

Archivas
http://www.archivas.com/

AppIQ
http://www.appiq.com/

Abrevity
http://www.abrevity.com/

Arkivio
http://www.arkivio.com/
Most of these products create classifications by reading and capturing metadata on every file. Data includes information such as file type, name, size, date created and date last accessed. They also collect file owner user information from the network directory service.

**Information Category Assignment**

Using the SNIA evaluation criteria you can place your data in the following categories

<table>
<thead>
<tr>
<th>Primary</th>
<th>Mission Critical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Business Vital</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Secondary</th>
<th>Business Important</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Important Productive</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Archival</th>
<th>Not Important</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Discard</td>
</tr>
</tbody>
</table>

Obviously this is arbitrary. However it makes sense if you keep your objectives clearly in sight -- important, transactional information on high-end storage; less-important, less frequently accessed information on lower-performing, cheaper storage; archive that which is unimportant; discard the truly useless.

Companies have a range of data used for different purposes. It makes sense to store it on different disk types. Matching storage price/performance to application need has a very straightforward value proposition.

**Tiered Storage**

Shifting less valuable data to less-expensive storage media allows higher storage capacities at a lower cost per GB. However, when you shift to larger, lower-cost storage, frequently there’s a performance tradeoff at a technical level. And this is perceived as service tiers as well as storage tiers.

Yet many companies experience improved storage performance by tiering. When you store all company data on a single tier (e.g., FC), all network users compete for access. When you separate storage into tiers, I/O operations are spread out among multiple tiers. This reduced competition for I/O time may:

1.) allow for good performance at the SAS or SATA tier
2.) improve top tier performance by reducing I/O requests arriving FC drives
3.) enhance storage service for all users

As you consider the various tiers you want to deploy, it’s a good idea to keep in mind the various storage specific attributes and behaviors you require for each tier. In no particular order these are:

- **Performance**: Each tier will have specific requirements spelled out in SLA’s in terms of application performance and end user delivery. Each tier will also require certain levels of device inherent tools and configuration advantages (RAID, optimization, e.g..)
- **Scalability**: All tiers need to be flexible and easy to increase, decrease, physically move, logically partition, and configure – the lower tiers less so.
- **Backup**: If you decide that data must be maintained – it must be backed up regardless of tier. Your backup strategy must be inclusive and flexible. Tier 3 can be attended to less frequently than tier 1.
- **Recovery**: With multiple tiers recovery can occur from one to another. It’s fairly important for different tiers to interact with one another – particularly in a DR or business continuity emergency. Also important for each tier to be independently recoverable.
- **Replication**: Ideally you should be able to replicate information within tiers, in or out of a frame, to other tiers or platforms locally or remotely. You may need to acquire additional tools to facilitate.
- **Availability**: Establish availability metrics in relation to storage performance and SLAs.
- **Operations and Management**: Ensure you have the operational and management team in place to accomplish a tiered storage implementation. This includes personnel, tools, training, and skills necessary at all levels in the organization.

Implementing a tiered storage strategy is probably best done incrementally. After you define your data classes and associated storage tiers, implement in a controlled and limited manner. Take one class of data at a time. Do it in bite size stages where you control success. This limits risk, is relatively non-disruptive, lets you show immediate results, and gives you continuous learning.

It’s good to remember that there’s no one-size-fits-all tiered storage solution. Your implementation will be specific to your organization based on business needs, data classification requirements, and IT budgets.

Larger enterprises might deploy different storage arrays for each tier, while smaller companies may opt for a mix of drive types and create multiple tiers within a single array. Companies might choose combinations of FC and SAS drives, while others select FC, SATA and tape.

Bottom line: successful tiering requires an understanding of business needs – which points right back at data classification.
About Mosaic Technology

For over 20 years Mosaic Technology has provided IT Infrastructure solutions to companies around the world. We help companies evaluate their IT environments and develop solutions that meet IT and business needs. We deliver solutions that are easy to implement, easy to use, and deliver immediate and long term ROI.

Quality Partnerships and Vendor Independence

Mosaic partners with companies that deliver quality and value. Our product portfolio includes a variety of best-of-breed options - often from competing manufacturers. This lets us keep an independent approach to technology and ensures our clients receive the best solution for their needs and budget.

In order for us to deliver solutions that fit, we first study and understand our clients environment. Our sales and technical staff work with you through a collaborative assessment process. Together, we identify strengths and weaknesses within your IT Infrastructure. Only then do we recommend appropriate solutions that are a fit for - and will improve - your organization.

Simplifying Infrastructure -- Solutions and Support

Simple is better. Whether designing a tiered storage implementation or redesigning your backup strategy - simplification makes sense. With Mosaic you get a complete portfolio of proven technologies that will streamline your operations. Our areas of expertise include:

- Storage – Primary, Secondary and Archival
- Backup, Recovery and Archiving
- Servers
- Networking
- Software Solutions
- Professional Services
- Maintenance
- Flexible Leasing Options
- Legacy Systems and Products

**Mosaic Value**

We give you independent technical input to your IT planning and execution. We work with you to understand and identify solutions that work for you. Our product portfolio and technical resources let us be product neutral. So solutions we propose are always driven by your needs not by a specific manufacturer or technology.

Legacy Support -- Used Systems & Trade-ins

Mosaic maintains a dynamic refurbished business. We can meet your needs or support your legacy systems with quality used systems or components. And we can give you aggressive trade-in value on your older or ‘coming-off lease’ equipment.

On-Site Assessments

The best way to experience the Mosaic Value is to see for yourself. Every IT department is on a continual improvement path. We can give you an independent set of eyes on specific areas of your operations. It won’t cost you a dime and you will get fresh, independent, vendor neutral input into your backup, storage, network, or related area. Give us a call to make an appointment: East Coast (603) 898-5966 | West Coast (425) 462-5004.