# Data Classification and Storage Optimization

## By Alex Adamopoulos

**The cornerstone to developing a storage strategy that makes the most of your existing storage environment**

## Introduction

The surge of storage-related investments in recent years has caught the eye of cost-conscious executives—and with good reason! With the rapid fall of storage hardware and the focus on price-per-megabyte, many IT organizations have been meeting their rising data volumes by adding new disks and drives whenever and wherever it seemed they need them. They have purchased a panoply of storage hardware and software to support ERP, SCM, CRM, and other critical applications—and installed them in disparate locations within the organization. They've implemented advanced storage networking solutions—storage area networks, network attacked storage, and content addressed (should it be "addressed" or "addressable"?) storage—with no regard to the increased management and administrative costs that often accompany these complex storage infrastructures. It's no wonder, then, that CIOs and other executives are demanding a measurable return on their storage solution investments.

Much can be done to extract maximum value from your current storage environment. But first—and to succeed—you need to do your homework. You need to identify *what* information exists in your environment, *who* needs it and *when,* and you need to understand *how valuable* that information is to each of the individuals, groups, and business processes that will need it. This process is called *data classification*—and it is prerequisite to a complete Information Lifecycle Management strategy and any storage improvement strategy you will develop as well as your ability to manage how you store your data throughout its lifecycle.

The reason? N*ot all data is equal.* Therefore, when developing strategies for storing and managing information from inception to archive, the understanding and separating of business-critical information from information that needs to be archived—and everything in between, is essential. Also, because the usefulness of data to the business and to the application created it varies during its life, this means the level of immediacy needed for accessing that data also changes over time—and this has a direct impact as to where you decide to place that data in your storage infrastructure at different points in time.

# Data Classification and Storage Optimization

The entire concept of ILM and data related strategies begin with the understanding that the criticality of data classification as a fundamental component of developing a storage strategy will maximize current storage investments and will support Information Lifecycle Management efforts.

The data classification process includes such things as:

- How to set storage management goals
- What information needs to be identified, how to identify this information, and the resources and tools required to interview key stakeholders
- How to organize and classify information gathered
- How to apply this information to your goals
- And the implications of data classification on your storage management activities—for example, business continuity, archiving, or information retrieval.

## What Is Data Classification?

*First, a brief overview of data classification.*

DC is integral to a wide scope of information management processes and strategies, but what's most pertinent is how it is used to help companies identify the inefficiencies in their data storage and how they can modify how and where they use their existing resources to leverage their investment, improve efficiency, and reduce costs.

Data classification is a *process* that groups data in categories possessing similar characteristics. The classification process involves refining each group by defining its shared characteristics, for example, similar service goals. The purpose of this grouping is to facilitate a key storage objective.

For example, in order to create a comprehensive storage strategy, you may choose to classify your data by *business priority*—let's say, being first-to-market. This classification will ensure that your data is maintained on the appropriate storage infrastructure commensurate with its priority to the business and your business priorities.

Perhaps you want to design an information storage consolidation plan. In this case, you need to classify your data by *physical state* and *location.* This will allow you to identify and eliminate islands of data and bring the most important data items closer to their end users through a consolidation strategy.

Hopefully, you're seeing the pattern. All data is not equal. Active and online data that has a particularly high value needs to be available at all times for rapid access by multiple organizations and applications.

## Data Classification and Storage Optimization

Some data requires 100 percent instant accessibility around the clock—with no tolerance for downtime. Some data is more valuable to certain organizations than others and some changes in value over the time. And of course, some simply requires archiving for occasional access or long-term storage.

Understanding the value of your data to the day-to-day operations of your business—and thus how quickly it needs to be accessed and by whom—is therefore fundamental to designing an effective storage infrastructure strategy and to consolidating your information storage to leverage current investments into a more efficient, more manageable information infrastructure. Data classification is the essential first step to these processes.

What's driving the need—and the urgency—for storage optimization?

Economics, of course, is top of mind. You've likely made a considerable investment in recent years—and most likely are not leveraging them to their fullest due to an inefficient, distributed infrastructure, lack of interoperability, and underutilization or poor utilization of disk capacity. Budgets are tight—so the dictate now is, "do more with what you have."

Also, IT organizations must respond to an increasing demand for efficiency in IT operations. With the fast pace of the last decade, many companies implemented multiple, large-scale line of business applications with very little thought to creating an integrated solution. With the information that is required and acquired from these applications, the challenge now is to streamline the overall process and create a more enterprise-wide approach. Companies that recognize the power of this integration framework will be positioned for long-term growth and be able to weather various economic cycles.

Security is another key issue. Decentralized IT infrastructures come with security and continuity exposures that can threaten a company's future in the event of theft, viruses, hackers, and, as has been made abundantly clear, terrorism. Companies must be prepared to protect their data, as well as ensure business continuity in a variety of situations.

A proliferation of new regulatory demands is defining new parameters and requirements for storage, in terms of storage duration, format, and accessibility. Data classification will allow companies to comply with these new regulations using their existing storage system until they are prepared invest in new solutions specifically designed for compliance, and that will support long-term efforts more efficiently and cost-effectively.

Also as your organization grows, the amount of data that you must manage—and its value to the organization—continually changes. This means that your storage management system must remain dynamic.

And of course, there's the competitive factor. A more efficient storage system supports a more efficient business—and if your competition has better accessibility to data—and a lower TCO—their customers will benefit from better, faster service, at lower cost.

*So, how do you begin?*

The end objective of a data classification is ultimately to provide an actionable grouping of data within an enterprise. The resulting data classification model that you create can then be used to facilitate the implementation of storage-related program recommendations. Getting there involves a clearly defined, four-step process.

*First, you must define a data classification goal.*

Data classifications begins with your establishing a clear, storage-related *goal*. This is your tactical interpretation of your overarching objective. Although this may seem like a trivial exercise, defining an unambiguous, clearly defined data classification goal is an important first step that ensures that the output from the data classification exercise will produce an actionable framework and model that can then be used to drive strategies and implementation plans for the program.

*Second, you must select a perspective – the best view – of the data necessary to achieve your goal.*

"Perspective" is a horizontal view of your enterprise's data. It provides the focus that helps drive your remaining data classification activities in a direction that will produce the most actionable result. The most common perspective for a storage-related data classification is the *application perspective*.

This view classifies the data components of applications within the enterprise. Since most of the data in the enterprise is tied to applications, this perspective will likely help you classify the majority of the data within your enterprise. It will also produce the most actionable data classification model for your program objective since most storage-related recommendations are implemented at the application level within an enterprise. This perspective is also usually the easiest to execute as most IT practices require that applications be well-documented within the enterprise, and this documentation can easily be made available to support the classification effort.

Other perspectives include *business process, business object, logical state, and physical state*.

- *The Business process perspective* is mostly strategic in nature. It is particularly useful for tying business goals and impacts to storage initiatives. For storage management optimization, for example, this perspective could be used to classify the processes based on how critical the business function they support is to the enterprise. However, because the necessary classification criteria for this perspective are rarely documented and readily available within the enterprise, you may need to go through other classification steps for which more robust documentation and understanding is readily available and more easily extracted from the enterprise.
- *A Business object perspective* provides a guide for classifying data based on the logical entities or business objects within the enterprise—business objects being logical data constructs that represent a key concept or transaction within the enterprise. This perspective is most often useful when combining separate data infrastructures from different organizations, where it is helpful to classify the shared business objects by infrastructure requirements or business priorities so as to drive the consolidation effort toward an economical union of IT infrastructures and business goals.

- *A logical state perspective* provides a guide for classifying data based on *how* it is managed or used within the enterprise. This perspective is most useful when a more granular classification of data is required to drive your program objective. Since the logical perspective deals with the more tangible attributes of the enterprise data—data content, type, organization, and access method, for example—this perspective can be instrumental in providing focus and direction to drive a more technical recommendation or implementation.

- *Finally, a physical state perspective* will guide you in classifying data based on certain attributes of the data itself. For example, it can be used when the objective is to classify data based on what type of media is used, what type of server is used to house the data, where the data is located, and so forth.

To select a perspective, start with the “end in mind” and confirm *what* horizontal slice of data is meaningful to you in terms of helping you meet your goal. Understanding how various stakeholders view their data—meaning, what data they consider meaningful and data needs to be readily available to them—is essential to this step. And even though this is a relatively simple step, its importance cannot be overstated. Perspective selection ultimately drives the structure of the data collection effort in subsequent steps. If a perspective is chosen that does not align with your program objective and data collection goal, schedules and budgets may be affected.

## Data Classification and Storage Optimization

*Once you have selected the appropriate perspective, you are ready to assign schemes, classes, and rules.*

"*Schemes*" define the vertical segments of the view of data in your enterprise. They are what define the *value* of your data that will help you meet your program objective.

"*Classes*" are the names for the "containers" which will be used to group the chosen data. Classes and Rules work together to the group the data within the context of the scheme. Your program objectives will define the number of classes required in your data classification effort; the more classes you use, the more granular your data segmentation will be.

Finally, there are "*Rules*." "Rules" break up the chosen scheme into more narrowly defined categories that will provide direction and drive the implementation of your program objective. They determine how the selected data is sorted into the defined classes. Rules can be numeric measurements—for example, "business impact is greater than $2 million," or they can be non-numeric qualifications, such as "finance business function." The only criterion for a rule is that it provides a clear guide to sorting the selected data into the defined classes.

To define the scheme, you must be able to establish what makes data valuable to you—for example, "frequency of access." Selecting an appropriate scheme will enable you to define the rules for the data that meet your goal. How many classes you select, and what you name them, will depend on your program objective. They should be granular enough to provide actionable, meaningful segments of data that allow you to assign tiers of services. Give your classes names that are meaningful within the context of your program objective—for example, "mission critical," "business critical," "business important," or "business unimportant."

Once you've developed classes, you can define rules to quantify and quality the data that will be grouped within each class. Tied to the scheme, the rule further clarifies value. Define a single, narrowly defined, unambiguous rule for each class. This will make it easier to categories your data into different classes. Examples of rules are "accessed hourly," "accessed daily," "accessed weekly," and so forth. Consider including multiple layers of criteria—for example, "and" and/or "or" statements, that may mix quantitative and qualitative measures where appropriate.

*With these steps completed, you are now ready to collect and classify your data.*

There are several effective approaches to data collection. These include conducting one-on-one or and/or group interviews with targeted stakeholders and users throughout the organization, conducting working sessions and workshops with target audiences, collecting and reviewing existing documentation, and using scripts, queries, log files and/or packaged tools. The depth and breadth of your data collection effort should take into consideration your data classification goal and program objective. For example, when trying to classify data by criticality, all you would need to support the data classification is a list of data and the criticality rating that they carry within the enterprise. However, you may also need to collect more granular information—such as server names and locations, storage arrays connected to servers, and dependencies.

Once the required data has been collected, you can populate your data classification model framework. Depending on the tools you use, data collection and classification is possible in a single step. For example, at EMC, we have created a proprietary, MS Access-based system application mapping tool that allows us to perform simultaneous, real-time collection and population. Although manual collection and classification is possible, utilizing an automated tool to facilitate this step is strongly recommended, and will help speed project resolution significantly.

*So let's say that in order as part of your goal of fully leveraging your current investments and optimizing management across the environment, you want to reduce your storage costs by eliminating storage islands as well as unnecessary servers and support.*

To accomplish this goal, you'll need a list of business functions, classified by location. And to facilitate this, you would need to do three things: utilize the Application Perspective to classify applications by business function; classify applications again by location; and classify business functions by location. This classification will enable you to identify redundant data pools in multiple locations, and as a result, be able to drive tighter efficiencies in the data utilization metric—in other words, the percent utilization of a data item in the enterprise. This classification will also allow you to determine how to leverage older equipment, keeping the cost of storage metric down as well.

As I mentioned earlier, the output of a data classification effort is a populated data classification model that is *actionable.* It should be used as an input into subsequent program phases that are focused on developing strategic recommendations as well as designing and implementing storage-related improvements.

*Now, data classification may appear intuitive.*

And certainly, some amount is intuitive and site-specific, driven out of your business practices and environmental requirements. However, achieving the best results requires an *in-depth* understanding of how your organization works, the availability technology infrastructure, and the specific information needs for each business practice—both for today and for the future, in order to adequately plan for growth and change.

You will achieve maximum benefit, therefore, from taking an organized approach to data classification, especially by utilizing tools and methodologies to drive the process and elicit the optimum information. I cannot overemphasize the importance of making careful, well-researched and highly informed decisions on seemingly "obvious" points such as what your storage-related goal should be or how different groups perceive the value of data to that goal. Each decision drives the next, and a misalignment at any step in the process can have a significant impact on schedules, budgets, and the effectiveness of your data classification effort.

You can create your own tools and methodologies or utilize those provided by third-party data classification professionals, if you choose utilize their services. At EMC, we have created a toolset that has been field tested and proven to generate consistent and effective results from one data classification effort to another. Here's what we've found:

- A *data classification planning worksheet* can help pictorially explain how multiple perspective-scheme combinations may be necessary to create the relationship that will enable the project team to meet the data classification goal. Column headings may include program objective, data classification goal, perspective, scheme, classes, and rules.

- A *perspective, scheme, class and rule reference guide* can help you understand representative relationships. For example, consider the *physical state* perspective. You might define this as "a guide for classifying data based on certain attributes of the data itself and how it is being used within the enterprise." In this context, then a *location scheme* would be used to classify data based on the media type used to store the data—for example, mission-critical, business-critical, or non critical. *Rules* would be used as qualifiers to determine the location of the data—in this case, "files that reside on the SAN," "files that reside on direct attached," and "files that reside on tape."

- *Interview questionnaires to guide the gathering of detailed information gathering from key stakeholders and users…*a *system application mapping tool such the one as I mentioned earlier…*and *diagramming software* such as netViz—which allows you to create 3-D pictures of your enterprise network complete with data values—are also essential to creating a viable data classification model.

- And finally, there's the *data classification model* itself. This should be designed to conceptually depict the results of the classification in a manner that is meaningful and compelling. It need not be complex.

*Let's look at a practical, real-life application of what we've just described.*

The company in this example is a major financial services company. This company made the strategic decision to evaluate its current storage infrastructure in order to define a data classification strategy and storage policy designed to drive cost awareness and responsibility within its storage technology and storage management environments.

In kicking off the data classification effort, the project team established two objectives for the project. The first was to optimize storage-related spending through a tiered, service level approach to storage architecture. The second was to support continued application and data growth more effectively and predictably by establishing policies for more efficient data management and retention, and improving capacity planning and provisioning capabilities.

The data classification goal to support these objectives was then established. This goal was to institute a framework that: one, groups products based on business function criticality and defines corresponding storage services; two, simplifies capacity planning, provisioning and cost analysis by grouping products with like requirements; three, provides a mechanism to simplify communication of diverse storage services and costs for such services to business leaders; and four, is actionable and maintainable by the customer.

Once the objectives were clearly understood and the data classification goal was created, the team addressed perspective. They selected two: a *product perspective* because it was the most actionable perspective for the company, and an *application perspective* because it would allow them to create a tiered storage environment and migrate data associated with a product to the appropriate tier according to the data classification scheme.

For scheme, the team chose *business function criticality*. In other words, product data would be grouped into classes according to its associated business function criticality. The resulting classes would describe the criticality of the product and the rules would determine the class that the product falls within.

The resulting data classification model outlined two classes of criticality, Critical Business Data and Non-Critical Data. Critical Data is serviced by three tiers of criticality defined by the availability target, recovery time objective (RTO) and recovery point objective (RPO). The Non-Critical Class of data has a single tier of criticality with a range of 95-98% for the availability target, an RTO of less than 48 hours and a RPO of less than 24 hours.

In addition, an Archive class was created to accommodate data that can be placed off-line or near line once its immediate usefulness has been exceeded. This class would contain archive data that would be purged once it is no longer needed by the business unit that owns the data.

Availability of the tiers described the percentage of time product data must be available to business units and provided an indication of the amount of redundancy that must be placed within the primary storage environment in order to be able to satisfy the target metric. For example, in order to meet a 99%+ availability metric, RAID protection, redundant SAN connectivity, fiber route diversity, etc. must be in place within the architecture to protect from single points of failure and subsequent failure to meet the availability objective.

RTO and RPO are metrics that give way to backup and restore requirements. When combined with the amount of data that must be backed up and restored, the appropriate backup and restore technology can be selected to meet the demands of the data. For example, if a tier 1 application had 500GB of data to restore within 1 hour following an event, restoration within this timeframe would require advanced recovery from a business continuance device. Conversely, if a tier 2 application only had 10GB of data to restore in less than 24 hours, a tape restore would suffice.

This model set the foundation for defining the infrastructure requirements, services, and technology that will be would be used to design a tier of storage to meet the data class objectives.

Detailed product data was subsequently collected from the Product Owner and Systems Administrators. This included information regarding the physical infrastructure that supported the products as well as the criticality indicators—for example, availability, RTO and RPO—of the products themselves. Both Current and future Class field values were determined.

The team also examined the company's storage environment in two locations to determine how to classify their data in order to provide cost savings through a tiered storage environment. In addition to the data classification model, other recommendations relating to the classification and subsequent realignment within the model included Storage Virtualization, Storage Provisioning, Storage Configuration, Storage Capacity Planning, Data Classification Owner and Performance Management.

## Conclusion

Over time, the rapid pace of technical innovation will deliver new choices in storage devices and services that may render your existing environment obsolete or inefficient. But for now, the bottom line is that there is tremendous opportunity to more effectively utilize your existing storage assets—and at the same time, satisfy ever-changing business and regulatory requirements—by redistributing data across the infrastructure according its criticality. For example;

Hardware consolidation can yield reduced hardware and software support costs for your existing storage assets, coupled with increased efficiencies in managing storage, and reduced maintenance, license, and general support costs.

Also, cost savings can be achieved by archiving large volumes of extraneous or old data to release additional capacity from premium storage disks, avoid backups of extraneous data, and storing archived data on more cost-effective media, according to users' requirements.

Proper classification of your existing data is the necessary first step to accomplishing this. Careful consideration of different data classifications helps facilitate proper placement and management of data across its lifecycle. It also leads to other storage efficiencies including service level agreements, the development of data policies, and improved chargeback mechanisms.