

The Case for Data Classification as the Foundation for Intelligent Information Management

Data classification is the act of collecting “information about data” that allows users to separate data into “piles” for more effective use. Data classification is of particular interest to storage vendors and users, and will soon be of interest farther up the software stack. What is it about data classification that is potentially so useful? Why should data classification be important to IT strategists?

The immediate uses of data classification in the storage area are relatively obvious, and include Information lifecycle management (ILM), storage management, and business compliance processes. ILM is the policy-driven management of information as it changes value throughout its lifecycle from conception to disposition. Data classification allows users to apply effective policies based not just on the information at hand (when the data was created) but also on richer notions of its utility at various stages of the lifecycle, its “fit” with business policies, and its ability to speed key queries if placed in speedier storage.

In storage management, data classification helps IT management to use tiered storage more effectively (with the resulting cost savings) and, by getting rid of unnecessary data in the to-be-destroyed category, frees up existing storage — thus deferring the cost of additional storage investments. In the business compliance area, data classification allows users to more easily identify data to be retained for legal discovery and regulatory compliance, and to assign the correct mix of storage assets for this purpose.

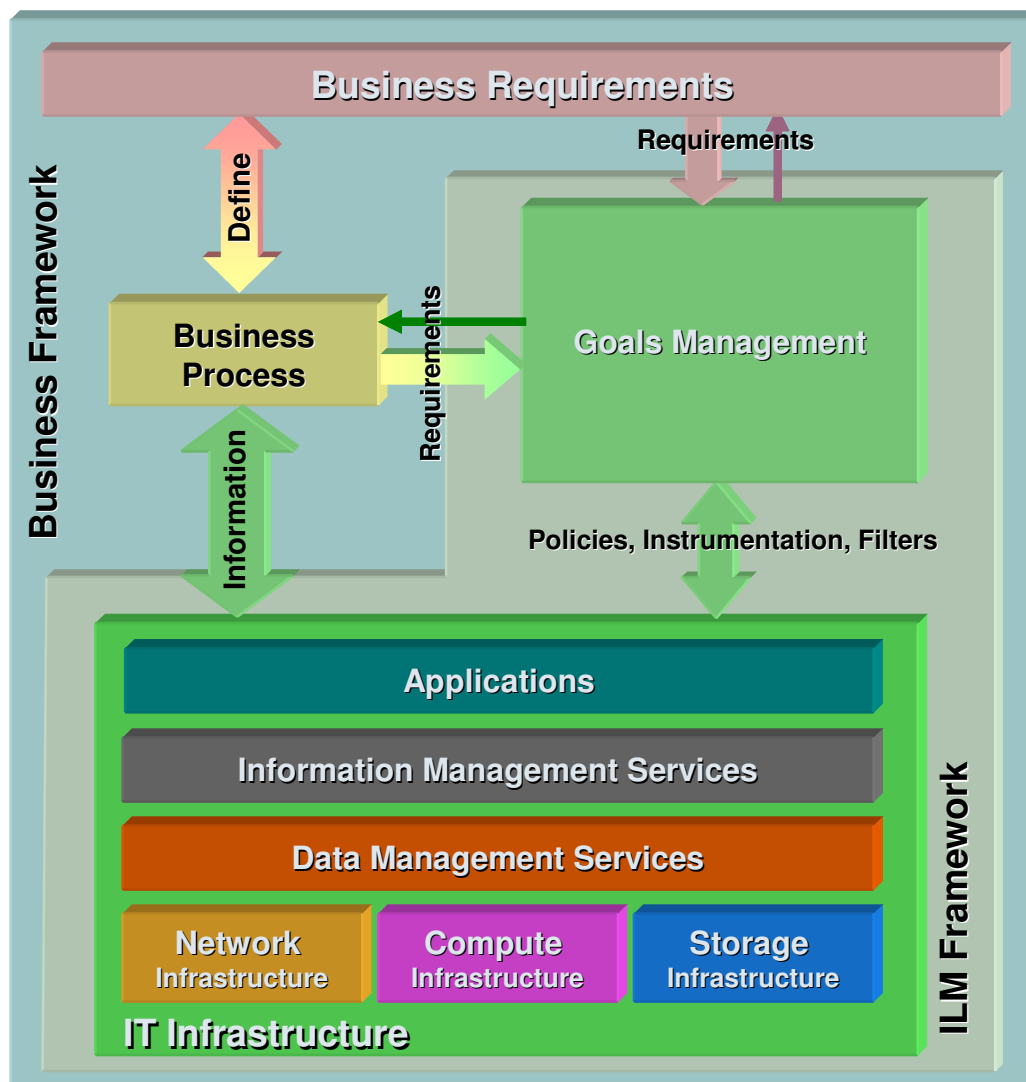
These uses of data classification, while compelling to the storage specialist, are too far “down in the details” to entertain the interest of the CEO (with the exception of compliance, which while essential, is still a targeted application of data classification). However, it is also true that buy-in at the business level is necessary to achieve good implementation and use of data classification. Executives and managers need to define the rules that enable the initial classification of data — business rules that are turned into policies that can be embedded in software as the basis for ongoing data classification.

As it turns out, data classification has justifications that reach beyond storage, all the way to the business level. Data classification is a necessary step towards improving the messy way that data is organized today, giving users a handle on the relationships between various “data archipelagoes.” Improving data accessibility, flexibility, and quality using data classification allows more rapid, more soundly based decisions that begin to implement the “real-time enterprise” that CEOs dearly desire. Finally, data classification can also serve as a foundation for “intelligent information management”, a newly arrived concept that aims to allow proactive tuning and movement of data to react rapidly to, and even anticipate, the enterprise’s need to use proprietary information for competitive advantage.

The Starting Vision

The process of taking data classification not only deep into storage software but above the storage level to the rest of an enterprise architecture starts with the vision of the Storage Networking Industry Association's Data Management Forum (SNIA DMF). The Forum's vision (Figure 1) of ILM is of a framework that links the ILM framework within the IT infrastructure to a business framework — and that business framework includes business requirements, business processes, and goals management. Most importantly, that link is enabled by a global metadata repository.

Figure 1: SNIA's Visionary Model for Information Lifecycle Management



Source: Storage Networking Industry Association, 2005

The Curse of the Data Archipelago

The next step in understanding the close connection between storage needs for data classification and the needs of the business is to realize the fragmented state of enterprise data, and the potential usefulness of technologies such as data classification that diminish or eliminate that fragmentation.

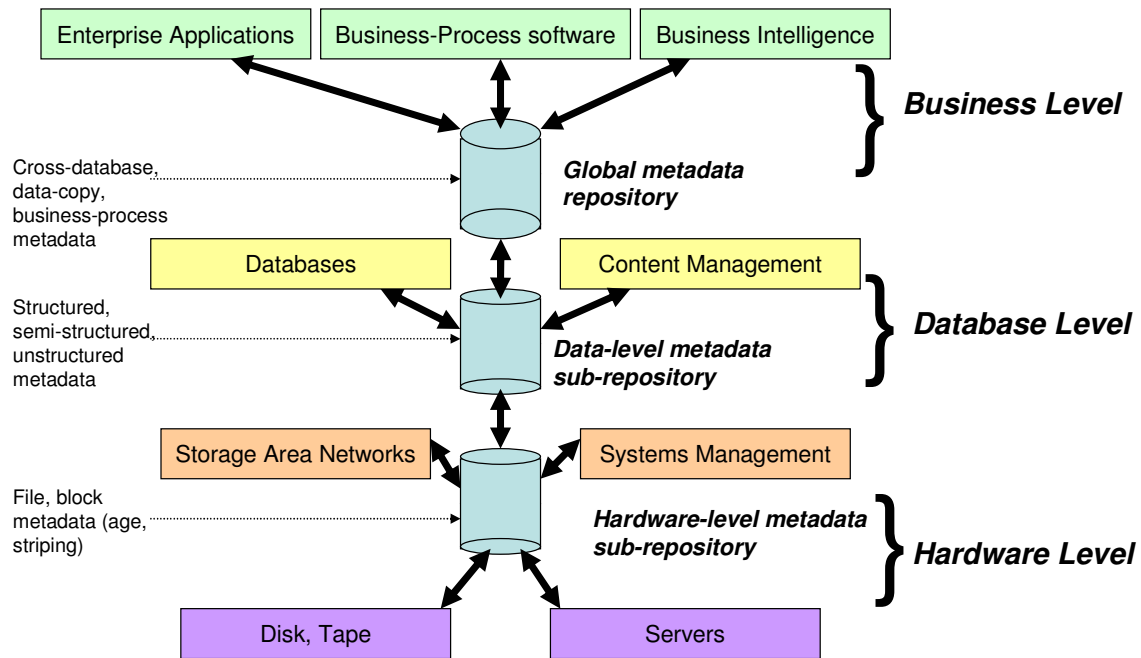
A real-time enterprise (RTE) supports on-demand access to key information, by making key business processes accessing this information faster, with the result of enhanced competitive advantage. “Real-time enterprise” IT infrastructures enable any member of the organization, at any time, to react to information about key new events (occurring inside or outside the enterprise), immediately, with the appropriate decision and action. However, because of the proliferation of isolated data archipelagoes, no one in the typical large enterprise knows where all of the data critical to this fast reaction is located, let alone how some critical data (such as customer records) relates to other critical data (such as product records). And the enterprise is continually generating new types of key data, such as RFID-process data and online customer relationships — making “real-time” reaction enabled by useful information ever more difficult to attain.

One obvious strategy to create an RTE is simply to identify processes that are business-critical, and speed them up. However, this strategy surprisingly often results in long-term failure. Why?

- Today, no database can deliver instantaneous access to the terabytes or even petabytes of mission-critical and business-critical information present in the average large enterprise — and the situation is getting worse, not better. Storage size at Fortune 2000 enterprises doubles every two years or less while major-database speed improves by only 40% at most every two years.
- New data sources are arriving at an increasing rate. Thus, while IT races to incorporate today’s data sources, new ones continue pile up, waiting to be incorporated -- so that the enterprise falls farther and farther behind the RTE ideal. Examples of the new data sources include RFID, the “deep Web,” and increasing amounts of data from more closely connected suppliers and customers.

The foundation for a practical answer to this “curse of the data archipelago” is to create a global metadata repository that would classify all of an enterprise’s data by its storage and information characteristics, and would allow storage administrators, database administrators, and business strategists to get a handle on the full scope of the enterprise’s proprietary information, as well as allowing rapid, comprehensive queries to detect and react to key business events “in real time” (see Figure 2). And so data classification at the storage level can form the foundation for the global metadata repository that enables the RTE.

Figure 2: Possible Global Metadata Repository Architecture



Source: Infostructure Associates and Mesabi Group, October 2006

Solving the Enterprise's Data Integration Problems

The usefulness of data classification and a global metadata repository go well beyond simply enabling a grand vision of a real-time enterprise. Consider the following statistics:

- 79% of all large enterprises have two or more data repositories; 25% have 15 or more.
- 40% of large enterprise IT budgets is spent on integration of data.
- 30% of the typical employee's time is spent searching for information.
- 30-50% of large-scale application-design time is spent dealing with multiple copies of the same data.
- 85% of information in the typical large enterprise is unstructured, and therefore typically not stored in "enterprise databases."
- There are 48 different financial systems in the average \$1 billion company, each typically requiring its own data store.
- 60% of CEOs believe that they need to do a better job of accessing information rapidly for swifter decision-making, while only 33% believe that the information they need is integrated or easy to get at.

These statistics indicate that because of data archipelagoes, every level of the enterprise (see Table 1) has data problems. By solving those data problems, data classification and the global metadata repository can ameliorate all the ills of business, IT and storage administration that these problems cause.

Table 1: The Large Enterprise's Data Problems

	Problems	Results
Business	<ul style="list-style-type: none">• Not globalized• Inaccessible (doesn't exist)• Incomprehensible	<ul style="list-style-type: none">• CEO doesn't know how he or she is doing• Enterprise reacts slowly• No one on same page
IT	<ul style="list-style-type: none">• Uncoordinated• Low quality (mistakes)• Low performance	<ul style="list-style-type: none">• High admin costs• Bad decisions• No decisions
Storage	<ul style="list-style-type: none">• Not available across vendors• Not classified by storage uses• Non-existent	<ul style="list-style-type: none">• Inefficient administration• No ILM

Source: Infostructure Associates and Mesabi Group, October 2006

By using metadata to classify information at all levels in a global repository, starting with storage, companies can tackle all of these problems effectively. Moreover, organizations can enjoy additional value in such key business areas as competitive advantage and data quality.

Data classification's impact on *competitive advantage* involves its ability to identify new relationships between data stored in a wide range of proprietary data sources. As the lag time before a competitor duplicates an application gets shorter and shorter, businesses are turning to proprietary information as a longer-term source of competitive advantage because competitors cannot duplicate it.

This often involves gaining deeper insights into customers and suppliers by identifying new relationships between data from different proprietary-data sources. Much of this information is scattered across multiple databases, spreadsheets, and rich-media Web sites, making it difficult to detect the connections between a customer's existing product preferences and new products arriving via the supply chain and R&D.

But data classification is up to the task. For example, Amazon.com might use a global metadata repository to "mine" its order-entry data about you, the customer, and note that you like a particular book author; it could then find out from its supply-chain management database that a new book by that same author will be forthcoming in three months, and solicit you by email to buy the book ahead of time. Result: a highly satisfied customer who sees no competitor that can provide the same service.

Moreover, data classification can help identify areas where relationships are undiscovered because the existing metadata, or the underlying data, is of poor quality. In Master Data Management, a new technology aimed at providing a

single view of key enterprise data such as customer information, user experiences show that customer records may have poor data-entry, or faulty defaults that cause poor data handling. The act of examining and classifying the customer records typically turns up these problems, and allows the user to correct them.

Intelligent Information Management

Information management is the management of data plus the content associated with the data that conveys the meaning of the data to the user, as well as of the relationships between bits of content as the information moves through the lifecycle of a business process. The “information about data” that conveys that meaning and those relationships is metadata. Intelligent information management is the enterprise-wide administration of metadata at the business level, for business needs, across all vendors and data types.

Typically, information management is not intelligent. That is, it is not carried out at the business level, does not operate across the enterprise or across vendors, and is in fact separated into silos that prevent IT strategists from using information effectively for the business’ purposes.

If an enterprise were to achieve intelligent information management, the IT strategists could for the first time ask questions such as:

- Should data be shared between my SAP and Siebel applications, or does it make more sense to provide a common interface at the composite-application level for better business processes?
- Should I migrate my business-critical data from the mainframe, to allow it to synch up with my online Web site’s customer data, or is it a better idea to leave it where it is and flow my Web site’s data to the mainframe?
- How should I replicate key data across geographies to maximize its availability at minimum cost?
- Where is the data that the CFO not only needs to know right now but will need at some point in the near future, and how can I make it available to him/her as swiftly as possible?

In other words, intelligent information management makes the data *plastic* (able to be moved or manipulated freely for better processing) and the information manager *proactive* (able not only to respond rapidly to business needs but also anticipate them).

Again, the foundation for intelligent information management is a global metadata repository, and the foundation for the global metadata repository is data classification. Storage vendors are already providing data classification and “indexes” (effectively, metadata repositories) that allow not only better ILM but also rudimentary querying of unprecedented performance and scalability.

Over time, these will combine with existing metadata above the storage level to allow enterprise-wide searches and data administration. Then, tools will come to analyze the data and carry out redistribution of data to improve performance,

robustness, and flexibility to handle transactional demand surges – effectively, a rudimentary intelligent-information-management utility. The point is that theoretically, an enterprise could do all of this today; but, practically, until a metadata repository ties all the information together, businesses cannot afford to try.

Mission Accomplished?

Enterprises are not necessarily standing at the starting line for data classification, ILM, and the global metadata repository, as some have made good progress in some areas; but a lot still has to be done. Moreover, the way ahead is not always clear. Starting with a targeted application, such as business compliance, may be tempting, but runs the risk that the work will be tailored solely to that focused business need and not serve as an extensible and expandable platform for future work. On the other hand, trying to achieve universal data classification may be biting off more than most enterprises are willing to chew.

Nevertheless, the game is worth the effort. Data classification's benefits are not narrow or minor. They range from straightforward operational benefits to strategic enhancements for the enterprise. They include:

- Basic cost savings through the use of tiered storage
- Better business compliance
- Serving as the foundation for solving enterprises' data integration problems
- Serving as the foundation for better data quality
- Providing the foundation for achieving competitive advantage through leveraging a key differentiator for any enterprise — its information.

Above all, data classification yields an ILM strategy that can serve as a fundamental foundation for intelligent information management. In doing so, data classification gives ILM a *raison d'être* that the business as a whole can understand. As a result, storage gets what it wants, IT as a whole gains some additional benefits, and the organization is able to more fully leverage its critical business information.