Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P:508.872.8200 F:508.935.4015 www.idc.com

IDC EXECUTIVE BRIEF

Managing Data Strategically

December 2005

Adapted from *Capturing Meaning in Metadata Management Software: Semantic Prerequisite to the Coherent Information Environment*, by Carl W. Olofson

IDC #33057

Introduction

The demand for nimble and flexible business processes has driven interest in developing a dynamic enterprise, which in turn requires dynamic information technology. Dynamic IT includes the requirement of "on demand" or "real time" data access and ability to deliver just-in-time information that's useful and actionable, and can fuel event-driven IT actions. This means using all the information inside an organization that may be currently bottled up in relational tables, formatted files, discrete online content documents and systems, and email by putting it all together in service of the organization's mission. It means managing data as a strategic asset. This Brief examines strategies for the systematic use, reuse, and understanding of data, as well as the architecture for ensuring data availability and security.

Understanding Data

For many years organizations have needed and wanted something very basic from their computer systems — the ability to ask a question and get an answer, without having to log into several different systems, use multiple applications and query tools, and know query syntax. Organizations have also wanted their computers to perform tasks that are explainable in business terms, as well as the ability to start or adjust those tasks using business terms.

Many piecemeal attempts have been made at achieving this ability, including point-to-point integrations to ensure consistency of applications, at least regarding key information, such as customer information, and to coordinate the operations of existing applications according to the rules of the business using business process automation (BPA). Such efforts have usually involved some limited metadata used to capture data formatting and transformation rules.

Data yields information when its definition is understood or readily available and it is presented in a meaningful context. Yet even the information that may be gleaned from data is incomplete because data is created to drive applications, not to inform users. Only by presenting data together with its definitions in a meaningful context, along with the available online content, and supporting any random combination of queries and actions, can organizations begin to manage their data as a strategic asset.

Metadata is the data that holds application data definitions as well as their operational and business context, and so plays a critical role in data and application design and development, as well as in providing an intelligent operational environment that's driven by business meaning.

Metadata is sometimes called "data about data." This definition, however, is not only incomplete, but is inadequate in understanding the full role of metadata. Metadata isn't always just about data — it can be used to document or provide a structure for managing data, applications, or an IT environment.

Metadata can be used merely to manage the rules for data storage, retrieval, presentation, and validation. Such operational metadata is necessary but not sufficient to build an integrated data environment. The metadata must provide more if the data and, ultimately, the representation and use of data are to be managed and understood.

To achieve the systematic use, reuse, and understanding of data (this last activity accomplished by its conversion into information) necessary for a fully integrated data-management environment, the full definition, context, and usage parameters of the data must be captured as well. These elements, taken together, provide the semantics of the data, without which systematic use of data outside of the application for which it was designed is impractical, if not impossible.

The Role of Semantics

People tend to take for granted the processes by which they apply semantics to human language; they don't really think about them. Oftentimes people expect to be able to naturally infer meaning from structured data in the same way. But assuming that sufficient meaning can be naturally inferred from structured data usually proves to be a big mistake. Structured data generally offers few clues as to its meaning, and sometimes those clues turn out to be false.

Some people tend to naively assume, for instance, that relational databases are self-documenting, and that they can interpret the data from the names of tables and columns in them. Not only is this not true, but when the far more precise understanding necessary for data transformation is called for, users can be left quite ill-equipped to do more than guess as to how to go about it .

For instance, one might assume that a column in one table called CUST-NUM and a column in another table called USER-ID contain data that mean the same thing, but this could be completely wrong. Sometimes a meaningful unit of information cannot be derived from a row in a table, or even from a table altogether, but it can be derived by putting different column values in select rows from one table together with column values in select rows from one or more other tables. Some data in the database may provide input to a formula; the result of the formula is meaningful and is presented to the user, but if the user doesn't know the formula (often buried in program code), the data itself is useless.

Data semantics involve four elements:

- The definition of the data, both elementary (such as field or columns) and in structures or combinations (such as record types or tables, database schemas, and larger data systems)

- The context of the types of data in question within a larger structure of meaning (such as a business model)

- The context of the actual data itself in the database, files, and so forth in relation to other data

- The knowledge and experience of the user

The weaker the former two elements are, the stronger the latter two must be if the user is to understand the data at all. The former two have to do with the structure, definition, and rules of types of data; the latter two have to do with interpreting instance data as it occurs in the database or file and are dependent on the actual operation of the application, and on the user. The former two must be captured in metadata and are necessary to govern large-scale data integration.

The four taken together enable smarter query and reporting functionality across the entire IT environment, and they enable the delivery of information, derived from structured data, that is meaningful and actionable for the user.

Metadata-Driven Data Management

The front-and-center role of metadata is to manage the relationship between a master data system, such as a data hub, within an organization, and the organization's various application databases.

Databases contain basically two kinds of data — either reference data or transaction data. Reference data, of which master data is a subset, is data typically about people or things, such as customers, employees, or products. Transaction data is all the information related to particular transactions, such the sale of a product. Transaction data has a definite life cycle, one that's typically shorter than the life cycle of reference data.

Whether reference or transaction, however, the more critical the data is, the more critical the metadata is. Therefore it's important to optimize the management of metadata through synchronization. This in turn can help organizations manage and integrate data that's shared across systems, such as lists of customers, suppliers, accounts, or organizational units.

Availability and Security Considerations

Managing multiple data sources creates the need for broader and more straightforward administration. But this need in turn creates its own challenges and opportunities aplenty for database managers. Data integration projects are increasing the complexity and fragility of data environments, putting a premium on the ability to keep databases up and available and to ensure good performance in handling database requests. IT departments need to approach the question of data availability and up-time proactively, while continuing to control costs.

Approaches to ensuring data availability must be operationally efficient, and focus on preventative maintenance and automation. Ensuring that the right information arrives to the right user at the right time — all cost-effectively — requires automating maintenance tasks, making performance problems visible, and correcting potential problems. Tools that are integral to ensuring data availability include:

- Database performance and tuning
- Back-up and recovery
- Replication
- General administration
- Database development

The burdens placed on organizations by strict regulatory requirements and increasing exposure to data theft means that decentralized data systems without the proper processes and tools for data governance are literally a threat to business health. As IT organizations pull data from separate repositories, the ability to track and confirm changes to the data could impact, for example, financial reporting and therefore represent a Sarbanes-Oxley control risk.

Data security that's applied via applications according to specific databases and data stores, or via firewalls, isn't enough. Organizations need to govern data usage from an enterprise perspective, and address security issues directly. The processes and tools that enable this governance include:

- **Data auditing**, which includes the ability to detect suspicious data access by reviewing audit trails. Also, and especially for compliance and assurance of confidentiality, being able to demonstrate that all data access was legitimate and conformed to existing regulations and confidentiality agreements

- **Database access control**, which includes ensuring timely establishment of appropriate data access constraints, reconciling them with overall security definitions, and controlling data access at various levels of granularity; on the user side at the user, role, and group level, and on the data side at the database, table, column and possibly row level.

- **Change and configuration management**, which includes tracking changes for compliance reasons

- **Encryption**, consisting of client-level encryption, explicit encryption (i.e., via a stored procedure or client-side package), implicit encryption (i.e., encryption provided by the DBMS that's transparent to the client), and encryption “on the wire” (i.e., between client and server)

- **Real-time security monitoring** (related to data monitoring)

Ensuring that data access is regulated is a challenge that must be addressed on three levels — the conceptual, the reporting, and the physical. A unified approach that integrates all three domains will help organizations develop and sustain a meaningful data governance policy. Organizations evaluating database administration (DBA) tools and utilities as part of a data-asset management strategy should consider the following:

- Seek to move from DBA point products to database management suites

- Tie data quality and security software to enterprise compliance and data consolidation priorities

- Implement data management and data quality software in combinations as part of an enterprisewide data management platform

Although database management systems (DBMSs) are becoming more self-tuning and self-maintaining, higher-level management functions that span DBMSs from multiple vendors will be in demand as long as enterprises maintain databases driven by software from more than one vendor.

Conclusion

Data yields information when its definition is understood or readily available and it is presented in a meaningful context. Yet even the information that may be gleaned from data is incomplete because data is created to drive applications, not to inform users. Metadata is the data that holds application data definitions as well as their operational and business context, and so plays a critical role in data and application design and development, as well as in providing an intelligent operational environment that's driven by business meaning.

Only by presenting data together with its definitions in a meaningful context together with the available online content, and supporting any random combination of queries and actions, can organizations fully harness and exploit their data assets. These data semantics, effectively applied, enable smarter query and reporting functionality across the entire IT environment, and they enable the delivery of information, derived from structured data, that is meaningful and actionable for the user.

With a unified, global approach to enterprise data, organizations can ensure accuracy, manage regulatory compliance, and ensure that the data infrastructure constantly maps to business needs.

COPYRIGHT NOTICE

The analyst opinion, analysis, and research results presented in this IDC Executive Brief are drawn directly from the more detailed studies published in IDC Continuous Intelligence Services. Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from IDC. Contact IDC Go-to-Market Services at gms@idc.com or the GMS information line at 508-988-7610 to request permission to quote or source IDC or for more information on IDC Executive Briefs. Visit www.idc.com to learn more about IDC subscription and consulting services or www.idc.com/gms to learn more about IDC Go-to-Market Services.