Meta Data Management in the Data Warehouse Environment

A Logical Approach to Meaningful Data Analysis

Abstract

“Without an effective way to manage data warehousing meta data, companies cannot automate the processes involved in running a data warehouse, or give end users a convenient way to understand the origins and nature of the data in the warehouse.”

—Wayne Eckerson, The Rationale for Data Warehousing Meta Data

Table of Contents

Executive Summary

Information Asset Management is intrinsically valuable—it is the lifeblood of any complex business. Companies depend on all types of information—financial reports, text documents, audio and video files, graphics, maps, messages and transactions—to make business decisions.

Ascential™ Software gives companies the infrastructure solutions they need to make the most of the information they have—focusing on five essential aspects of Information Asset Management:

- collection,
- validation,
- organization,
- administration and
- delivery.

Today's dynamic and fast changing business environment creates volumes of data that quickly become useless without effective management. To make sound decisions that affect the bottom line, users need information that helps them interpret the data. They need clear, precise, accessible and effectively managed meta data.

Meta data is generated whenever data is defined for, added to, deleted from, modified in, or moved to or from, a warehouse. Through the management of meta data, data definitions can be standardized, data relationships clarified, and data made available when and where it is needed.

Meta data sharing is realized by tool collaboration and sharing, made possible through the selection and seamless integration of best-of-breed tools. Today, vendors, systems integrators, and businesses generally employ one of two methods for achieving meta data integration. The more common involves constructing bridges from one data warehouse tool to another. This approach becomes unworkable whenever one of the participating tools is changed by its vendor. Alternatively, the common meta data model method offers a “least common denominator” approach in which only an agreed-upon set of meta data is contained or one vendor establishes a “standard” that others must follow. The Ascential approach avoids the short-comings of these methods. It features MetaStage™, a patented component of the Ascential DataStage® XE Suite. MetaStage provides visibility into the entire data warehouse environment by integrating islands of information into the warehouse architecture. It combines robust meta data sharing, analysis, and management across a spectrum of warehousing tools.

The Promise ... and the Problems

The Role of Meta Data in the Data Warehouse

It is frequently said of high technology that "the only constant is change." Enterprise data warehouses, by their very nature, are highly dynamic—perpetually changing to keep pace with evolving business (especially e-business) needs. Over time, data stored in a warehouse inevitably becomes outdated or changed in various ways. As it does, reporting and analysis of warehouse data becomes less and less accurate. Users find themselves groping for meaningful answers to such questions as: What does the data mean? What is its structure and format? Where did it come from? How was it calculated? When was it loaded? Who owns it? Where is it used?

Properly managed, meta data enables users to arrive at the answers. Meta Data—loosely defined as "data about data"—is generated whenever data is added to, deleted from, modified in, or moved to or from, a data warehouse. It can be found throughout the enterprise in a patchwork of repositories and proprietary meta data stores.

It's difficult to overstate the advantages of managing meta data:

- *Consistency of definitions*—One department refers to "revenues," another to "sales." Are they talking about the same activity? One subsidiary unit talks about "customers," another about "users" or "clients." Are these different classifications or different terms for the same classification? Effective meta data management can ensure that the same data "language" applies throughout the organization.
- *Clarity of relationships*—Meta data management illuminates the associations and interactions among all components of the warehouse environment: business rules, tables, columns, transformations, and user views of the data, to name a few. By clarifying relationships throughout the data warehouse environment, managed meta data enables warehouse managers and knowledge workers to see the bigger picture—to fully understand the meanings of the data assets, and to accurately predict and manage the impact of changes to the environment.
- *Availability of information*—Meta data exists "behind the scenes," revealing the origin of data, who defined it, when it was modified, and much more. Traditionally hidden, meta data must now be made visible to company knowledge workers on demand. New standards and technologies like XML and the Web create a perfect vehicle for delivery.

Technical and Business Meta Data

Two broad categories of meta data are found in the warehouse environment:

- Technical meta data provides a detailed blueprint that IT can use to build and maintain the warehouse. Technical meta data typically includes database implementation names, table and column sizes, data types, and structural information such as database key attributes and indices. Providing data warehouse designers and developers broad access to technical meta data allows more rapid implementation of changes and rollout of future projects.

- Business meta data includes those descriptions of data that are not related to software implementations—for example, the business name, business rules in relation to other data, and the owner of the definition. Business meta data gives end users a roadmap for navigating all of the data in the enterprise by documenting what information is available in the warehouse and, when accessed, provides a context for interpreting the data. Business descriptions of a data element, information on when the data was loaded, and how it was calculated or transformed, all prove invaluable to users in order to understand and trust the data and use it to make sound business decisions.

Why Has Using Meta Data Been So Difficult?

When building or extending a data warehouse, developers employ a variety of tools for system modeling, database design, data quality assessment, data movement, scheduling, analysis, and reporting. The need for close collaboration among these tools is critical as data is defined, transformed, moved, and accessed for business intelligence (BI) purposes. The key to achieving tool collaboration and sharing lies in the selection and seamless integration of best-of-breed tools—but so far, in the large majority of enterprises, this easy-sounding remedy has been more hoped-for than achieved.

The consequences of inability to manage meta data are many—and severe:

- Changes in the data warehouse are difficult to manage, and thus frequently can't be made in time to match the pace of change in the business.
- Data can't be compared and analyzed across departments and processes. Stove-piped data marts have to be reengineered into an enterprise warehouse to support sound, consistent business decisions that transcend divisional boundaries.
- Redundant, ad hoc systems—only temporarily useful—are extremely costly to maintain, integrate, and retire.
- BI and modeling tools can't be integrated into the warehouse environment, hence multiple versions of "truth" are scattered throughout private stores of business data across the enterprise.
- Meta data can't be shared among products without rekeying, which is time-consuming and introduces errors into the environment.
- Documentation is out of date or incomplete, affecting the ability to manage changes in the environment and ultimately undermining the knowledge workers' confidence in the data they are using.

Not There, Not Accessible, or Not Intelligible

Most businesses struggling to manage their meta data are finding that it is either nonexistent or inaccessible, or accessible but unintelligible. This situation can be due to any of several reasons:

- The business lacks the tools or organizational structure necessary for developers and end users to create valid meta data;

- Information exists in unconnected, frequently redundant or temporary gathering places that function as "islands," "silos," or "stove-pipes," making it impossible to implement changes that will ripple through the entire enterprise; or

- The meta data can't be accessed because the tool that created it is no longer available; other tools can't be substituted because they don't speak the same language as the originating tool. Example: An extraction and transformation tool is typically unable to use the information generated by a warehouse modeling tool.

Whatever the cause, the negative effect is that comparing and analyzing data across departments or processes becomes virtually impossible. Meta data is the logical way to create a uniform corporate definition of data, yet no industrywide meta data standard has been adopted. (Although the Common Warehouse Model promises to become the industry meta data standard, its effectiveness depends on wholesale, widespread and consistent adoption by all relevant tool vendors – which historically, at least, has not happened with other proposed 'standards.')

Looking for the Right Tool

As data warehouses and data marts continue to proliferate, businesses need meta data management more than ever to function effectively. These organizations can point to a substantial wish-list of capabilities they would like to see in an ideal meta data management tool. Of primary importance, developers and end users alike need a single authoritative source in order to use captured meta data. Furthermore, they also need to be able to maintain meta data throughout warehouse development and deployment, and make it accessible by any number of tools used to create and maintain the business intelligence infrastructure. To accomplish these objectives, the infrastructure needs a translation service that would permit data to be shared among the warehouse tools suite. This service would provide fine-grained semantic integration, enabling maximum sharing without gaps or information loss; and it would also provide the ability to add new tools to the warehouse environment without creating additional maintenance problems.

The "ultimate" meta data management tool would enhance data availability by offering the capability to publish Meta data out to the Web. It would also be tightly integrated with the BI reporting and analytical tools. These enhanced capabilities would yield the consistency of definitions, clarity of relationships, and availability of information that are unavailable through current approaches.

Before introducing the Ascential solution, we'll examine the prevailing approaches to tool integration and see why they are proving unsatisfactory.

Conventional Approaches to Tool Integration

Tool Bridges

Today, vendors, systems integrators, and businesses employ one of two methods for achieving meta data integration. The more common involves constructing bridges from one data warehouse tool to another. In this approach, the developer first becomes familiar with the underlying schemas of the source and target tools, then extracts the meta data from the source tool and changes it into the format of the target tool. The same process in reverse is used for bi-directional exchange.

To a limited extent, tool bridges furnish a quick and easy solution to meta data exchange. Over time, however, the quick fix begins to break down. Problems emerge when vendors deliver revised or updated versions of their tool, and existing bridges will not work with the new versions. Consequently, additional precious developer time must be spent studying new schemas and building new bridges.

Furthermore, the bridge approach will not scale as the warehouse grows and the number of integrated tools increases. Assuming that two bridges are required to integrate two tools, six bridges will be needed to integrate three tools, twelve to integrate four tools, and so on. Like their real-world counterparts, these bridges will eventually "fall down" if not properly maintained. When one of those four integrated tools changes its schema, then six bridges must be changed. Clearly, the endless writing and maintenance of bridges is likely to devour an unacceptable amount of scarce, expensive development resources.

For these reasons, tool bridges tend to be most appropriate, and function most effectively, as a tactical, unidirectional meta data exchange solution. They are an inadequate answer to the comprehensive management of warehouse meta data.

Common Meta Data Model

An alternative integration method is the common meta data model, which typically takes one of two forms:

- The first type of model contains only that set of meta data on which a group of vendors can agree to use for data exchange, which then becomes limited to those items described in the model. In this "least common denominator" method, the more vendors involved in the agreement, the smaller and more restrictive the set tends to become. Examples include ANSI IRDS, IBM's AD/Cycle, PCTE, CDIF, and the Meta Data Coalition's OIM, which was recently integrated with the Common Warehouse Model from the Object Management Group.

- The second model type is put forth by a vendor as "the standard"; other vendors must support this model to achieve integration. This approach is not only somewhat arbitrary, it fails to embrace all of the meta data available for sharing within tool suites. Informatica's MX2 is an example of a vendor-supplied extraction, transformation and load (ETL) standard.

The Ascential Approach to Meta Data Management

It follows that a truly workable solution must avoid the shortcomings of tool bridges and common models. In the Ascential approach, meta data is integrated by means of a comprehensive model that represents the union of shareable meta data from all data warehousing tools. Unlike the other approaches, this integration model is already built in, and reuse is transparent to users. Most important, it enables identification of the rich relationships among the various tools.

We've spoken of the three major goals of meta data management: consistency of definitions, clarity of relationships, and availability of information. The Ascential approach achieves these objectives by:

- Allowing each tool to manipulate meta data via its own meta data model;
- Enabling most tools' meta data objects to be shared;
- Incorporating schema evolution strategies; and
- Eliminating construction and proliferation of custom bridge code.

A Solution That Caters to Change

In setting out to deliver a better way of managing meta data, Ascential sought a solution that would represent the native perspectives of the tools that participate in the warehouse environment, and also incorporate a unique atomic, neutral warehouse model—the MetaHub. We wanted to maximize semantic sharing among an integrated tool set, while minimizing or eliminating the need for the participating tools to change their structure. At the same time—keeping in mind that, indeed, change is the only constant—we needed to ensure that our solution would allow for, and manage, the inevitable changes in participating tools and the integration hub over time.

MetaStage: Bringing Coherence and Understanding

The Ascential approach to meta data management features MetaStage, a patented component of the Ascential DataStage XE suite. MetaStage is a single tool for meta data control that brings coherence and understanding of data to people who need to use it. By integrating islands of information into the warehouse architecture, MetaStage provides visibility into the entire data warehouse environment. It offers robust meta data sharing, analysis, and management across a spectrum of warehousing tools, including business and modeling tools, DataStage, online analytical processing (OLAP) servers, and BI reporting tools.

These are some typical MetaStage use scenarios:

- A knowledge worker wants to know the definition and source of a term such as "corporate price" without leaving the analytical tool he or she is currently using.
- When a data manager changes a corporate pricing algorithm, he or she needs to know how many BI tool reports are affected by this single change to the corporate data model.
- The warehouse administrator has a visual representation of the previous 24 hours of warehouse production runs; he or she needs to find out if the correct version of the pricing algorithm was used to populate the Sales data mart.

MetaBrokers™

MetaBrokers™ are MetaStage exchange utilities that contain knowledge of a tool's data model and its relationship to the MetaStage integration hub. They decompose and recompose tool model components into simple semantic units for movement in and out of the MetaStage MetaHub™ Directory—the storage schema that represents the essential semantic units for a shared environment. A separate MetaBroker exists for each data warehousing tool.

Conventional meta data integration may deal only with the form of the meta data, ignoring the semantic meaning of what is transferred. As a result, users are likely to make inaccurate assumptions about the meta data or have to rekey the missing information. By providing not just form but also contextual meaning, MetaBrokers provide a deeper level of meta data interchange. Now the full value of warehouse data can be exploited—quickly and without guesswork or tedious rekeying.

Currently, MetaBrokers are available for DataStage and a number of widely used design and BI tools, and support for others is being added. As new MetaBrokers are incorporated into the shared environment, they can immediately share meta data with all other tools previously integrated into that environment. Further, MetaBrokers can be developed which provide compatibility between whatever prevailing meta data standards finally win the standards war and the tools which support those standards—making this logical approach effectively "future proof."

MetaStage Architectural Components

- The *MetaHub Directory* is the database that contains the data warehouse integration model—the storage schema that represents the essential semantic units for a shared environment.

- The *Explorer* is the primary client interface, letting users navigate through, analyze, publish, and subscribe to, meta data stored in the MetaHub Directory.

- *MetaBrokers* enable the sharing of meta data among all of the tools in the warehouse environment. With MetaBrokers, tools can share meta data without having to change their internal meta schema to conform to a common model.

MetaStage Functional Components

How the tools are mapped to the integration model, and thus indirectly with each other, provides a rich source of meta data for advanced management and reporting—for example, cross-tool impact analysis and data lineage. Through its Explorer, MetaStage provides a number of sophisticated exploring, analysis and use functions.

Impact Analysis
Through Impact Analysis, developers can examine all of the relationships associated with an object. In this way, they are able to assess the potential impact of changes across the integrated environment before they occur. For example, the developer can predict the effects that a change to the table definition or business logic will have on the entirety of ETL jobs and business reports.

Data Lineage
Basically, the function of Data Lineage is to answer the question, “Where did this data come from?” Data Lineage reports events that occurred to create and update warehouse data. When event meta data is connected to design meta data, users get a complete view of the data’s history and can judge how it affects business analysis. Data Lineage reporting opens up an overall view of the environment that, in turn, optimizes data warehouse and data mart creation. A special Process MetaBroker automatically populates the MetaStage MetaHub Directory with ETL event meta data from DataStage, and can work with any other tools that emit event-type data.

Meta Data Sharing and Documentation
The Publish and Subscribe capability lets users publish standard meta data from the MetaHub Directory to a variety of targets. Other users can subscribe to meta data publications on a one-time or a recurring basis. When meta data in the MetaHub Directory changes, subscribers can be automatically notified. Developers can use Publish and Subscribe to share and reuse DataStage design components across multiple projects, shortening the warehouse development cycle. By the same mechanism, analysts developing BI reports and OLAP tool users can reuse technical and business meta data generated by any of the other warehouse tools.

For documentation purposes, any meta data can also be output in HTML or XML format and further customized via XSL style sheets for presentation on any platform supporting a Web browser. Further, MetaStage is tightly integrated with Axielle, Ascential Software’s enterprise portal product, allowing automatic publishing of MetaHub Directory contents to a portal for viewing by business users.

Looking Ahead

Meta data is nothing less than the foundation on which business knowledge and decision-making are built. By providing essential content to the enterprise portal, it serves as a critical piece of the enterprise information infrastructure.

Our ultimate goal is to fuel the integration of knowledge throughout the enterprise by offering robust, flexible architectures that embrace a variety of model standards. MetaStage adds significant meta data management services to the entire data warehouse, including source databases, data models, ETL tools, and analysis tools. We intend to offer the capability for heterogeneous cross-tool meta data sharing and analysis by continuing to enhance our cross-tool analysis and query capabilities. We will exploit XML integration to help e-businesses communicate more effectively.

MetaStage can potentially support any type of meta data from any data warehousing tool. We plan to further exploit this potential by delivering key MetaBrokers and providing MetaBroker development capabilities for our partners and customers.

“The only constant is change.” At Ascential, we’re not only ready for change, we thrive on it. This philosophy will continue to guide our direction as we move ahead.

About Ascential Software, Inc.

Ascential Software is the market leader in helping companies turn unrefined data and content into reliable, reusable information assets so they can make better business decisions and operate more efficiently — maximizing their return on information. Ascential defines a framework for Information Asset Management based on experiences with over 1,500 customers worldwide. Ascential offers infrastructure solutions and services that are open, scalable and able to manage the wide variety and volume of information assets that complex organizations rely on. Ascential Software is based in Westboro, Massachusetts with affiliate offices worldwide.

Ascential™ Software

50 Washington Street
Westboro, MA 01581
Tel. 508.366.3888
www.ascentialsoftware.com

ASCENTIAL SOFTWARE REGIONAL SALES OFFICES

North America	800 486 9636	Japan	81 3 5562 4500
	508 366 3888	Asia	852 2824 0981
Northern Europe	44 20 8818 0700	South Africa	27 11 807 0313
Southern Europe	33 (0) 1 4696 37 37	Australia/New Zealand	612 9900 688
Central & Eastern Europe	49 89 20707 0	Latin America	55 11 5188 1000

WP-601A-0301

Printed in USA 01/01.