

Managing with metadata

Need help mitigating data warehouse errors? Try improving your metadata first.

by Ashok K. Nag

Data warehousing professionals keenly understand the importance of metadata. This “data about data” is used to answer questions and solve problems, and it provides the thread that connects and stabilizes the components of a data warehouse architecture. Most vendor solutions incorporate metadata exchange capabilities, as this functionality has become an essential requirement of industry standard tools.

However, while metadata’s criticality in a data warehouse environment is universally accepted, end users’ metadata requirements often go underappreciated and unmet. In fact, a technocratic view of metadata has prevailed, something that contributes to many data warehouse problems, particularly in terms of usage.

A sound and robust metadata management strategy must be implemented to meet the needs of end users. In most data warehouse implementations, though, metadata and the reporting and query environment aren’t even integrated. So what do business users need, and how does the data warehouse architect provide it?

To help answer this question, it’s first necessary to spell out the importance of metadata in managing and maintaining a data warehouse over a longer period. The main point? It may be possible to build a data warehouse and make it available to end users without having



comprehensive metadata. However, over a longer period, such a strategy is fraught with grave risk.

Metadata and data warehouse risk management

Any large data warehouse runs the risk of becoming obsolete and disused over time due to a number of reasons. The four most important reasons are:

> **Source system changes.** It is very common to find that source system architecture has undergone substantial modifications, even complete restructuring, after the enterprise data warehouse (EDW) has been built and made operational. Changes could result through many ways, like the replacement of the old operating system (OS) by a new one, a redesign of old table structures and changes in data types and data formats. All of these changes could materially affect not only the extract, transform and load (ETL) process but also the existing reporting and query environment. The only way to mitigate the risks associated with such changes is to have a comprehensive metadata database encompassing all aspects of the source system and ETL workflow details. The best strategy would be to build a metadata-driven data warehouse so all processes in the data warehouse-related workflow are built in a generic fashion, and process parameters are obtained from the metadata database.

> **Changes in the business environment and business model.** The business environment is inherently dynamic and competitive. Organizations respond to environmental changes by restructuring their business models, mergers, acquisitions and consolidations. In the process, the enterprise data model at the foundation of an EDW would require changes ranging from substantial to minor additions and deletions. Handling these changes could be very smooth if the data warehouse was built with a robust metadata strategy.

> **Changes in accounting standards and regulatory prescriptions.** Whenever accounting standards change, it has a direct impact on the definitions and, to a varying extent, on the underlying concepts for many existing data elements in the data warehouse. Similarly, regulators may redefine the existing terms and concepts. For example, banking regulators may redraw the definition of non-performing loans in a particular context. In such situations, it would be extremely difficult to maintain integrity and temporal compatibility of data in a data warehouse without relevant metadata.

> **Changes in the data warehouse management team.** Like all software applications, a data warehouse is also subject to people-risk during the maintenance and enhancement phase in the absence of adequate documentation of the application developed. A good metadata strategy is the only insurance against occurrence of this risk.

Types of metadata

In a broad sense there are only two types of metadata: ETL-related metadata (also called technical metadata) and business metadata. As ETL-related metadata draws its lineage from the data dictionary of industry standard databases, it has matured considerably in the last decade, both in terms of its coverage and comprehensiveness.

The same cannot be said about the business metadata, for which industry standards are yet to emerge. Though the Open Information Model of Metadata Coalition had proposed business engineering and a knowledge management sub-model, the main motivation driving this effort was to provide a standardized framework for tool integration by vendors. Moreover, business metadata, or semantic metadata, is looked upon in this framework as a translation layer enabling business users to get a description of the database content in a language they understand. These models do not address the

issues of required content of business metadata in a generic sense.

Components of business metadata

Business metadata can be structured into nine constituent components, some of which are dynamic, the rest static. The dynamic components are data- and/or time-dependent in the sense that as new elements are added to a given data series, these components of metadata also undergo changes. These components are detailed below, using specific references to the banking and financial sectors in the area of data risk management as an example.

1 Concepts and definitions

An organization's business practice, standards and conventions generally define the contextual meaning and definitional boundary of data items available in the organization's database. The organization's data warehouse has to store and provide access to the end users to this essential component of metadata. A good metadata practice is to impose a classification scheme for all these business data items and represent the resulting conceptual categories through a navigable hierarchical structure. For example, let us consider the data item "credit rating" of a corporation in the database of a banking firm. Since "credit rating" is a measure of quality of credit—an attribute of loan assets of the bank—an end user should be able to drill down to this data item from a conceptual category like "asset quality." For all practical purposes, constructing a meaningful and easily understood classification scheme in regard to a specific domain like banking shouldn't pose insurmountable problems.

2 Sources of data

For a business user, a data source identifies the business unit or organization that is the owner of the data. For example, if "credit

Taking the next step in metadata management will make risk management of the data warehouse easier and more efficient, giving a new edge over the competition.

rating” is provided by a credit agency through a specified report, the same should be identified as the data source. If data on an economy’s total credit is from the central bank in the country and is published in its monthly bulletin, then the same should be reported as the data source. On the other hand, the technical metadata would describe the source database and the specific table in it as the data source.

3 Units of measurement

It is always advisable to determine one standard unit of measurement, say, millions of domestic currency for credit amount, in an organization’s data warehouse. The end user must be informed about the unit of measurement of any data item in a report or query result.

4 Coverage and granularity of data

By “coverage,” we mean geography or business units to which a given data item refers. For example, a report on credit growth must explain whether it refers to domestic credit or global credit, whether some business units’ data is missing and so forth. By “granularity,” we mean the lowest level at which data is available. End users will be interested to know at what level of detail some data items are available.

5 Compilation methodology

For a derived variable, it is necessary to know the method of compilation. When the data is sample estimates, the estimation methodology must also be explained to end users. For example, data on profitability must be accompanied with complete details about its calculation procedure.

6 Frequency

As any data warehouse contains historical data, metadata should inform the end users about the lowest frequency at which data is available within the warehouse.

7 Availability

This is a dynamic component of metadata, giving the time horizon for which data is available for a given item. As new data is added to a given series, it should be reflected in this component of metadata.

8 Temporal compatibility

A data item may undergo significant changes in terms of its definitional boundary, geographical and business unit-wise coverage and compilation methodology. An end user needs to be informed about all these changes so that he or she can make an informed decision about the real import of the observed temporal pattern in the data series. This is one of the most important components of metadata that would determine the usability of data from a data warehouse. For example, in many countries the definition of non-performing assets has undergone changes in the wake of Basel II recommendations. This would impact any trend analysis of the quality of a bank’s credit portfolio. Without this necessary background information, such trend analysis would be rendered meaningless.

9 Metadata navigation requirement by end users

The metadata browsing environment and the reporting and query environment are generally not integrated in the majority of data warehouse implementations. While browsing business metadata, it isn’t possible to launch a report on a particular data item, nor is it possible to browse

related metadata while viewing a report. The business users need this facility most in an EDW. End users should be able to view or create a report with all associated metadata, as they would get it in a paper report in the form of footnotes to a table. It should be possible to query metadata databases to know what data from which period and up to which period is available in the data warehouse.

What’s next?

Of the two kinds of metadata, business and technical, development on the former has remained underdeveloped in recent years while technical metadata has grown in leaps and bounds. Most work on business metadata has concerned itself with specific components without improving it as a whole.

However, development of business metadata is becoming equally crucial to businesses’ success. Together with technical metadata, it helps end users mitigate risk, solve problems and assist data warehousing. Further enhancements to metadata capabilities—especially to the nine dynamic components listed earlier—will enable end users to better manage metadata in these capacities.

Companies need to be willing to give end users the tools to better navigate business metadata. Taking this next step in their metadata management will make risk management of their data warehouses easier and more efficient, and will give them a new edge over the competition. **T**

Dr. Ashok K. Nag is senior vice president of Riskraft Consulting and a former senior executive of the Reserve Bank of India. He is a well-known expert in the area of data warehousing and data mining, and lectures on the topics. Dr. Nag has also published more than 25 articles in national and international magazines.