Delivering information you can trust
December 2006

IBM Information Management software

Profiling: Take the first step toward assuring data quality

Contents

Organizations face an information challenge that begins with locating information, getting it when it is needed and providing it in the form needed. Once the data is found, the next steps are to discern further insights from it. Information validity and control are additional concerns.

The challenges only mount if businesses cannot ensure access to authoritative, consistent, timely and complete information. In fact, a wide range of corporate initiatives fail because of poor data quality. One way to address data quality is to profile source data when it enters the system.

Why profile data?

Analyst studies have shown that more than 75 percent of data integration projects either have cost overruns or fail. They often fail to deliver the required features, exceed their budget or end because of cancellation before they are completed. Why the high failure rate? The answer may lie in the traditional approach to data integration, which follows these basic steps:

Step 1. Analyze the users' needs and build a target database specification. Interviews with users result in devising a grand scheme for a database model to address all of the questions the user would like answered by the target application.

Step 2. Analyze the data sources available. Compiling and analyzing data sources from legacy systems, operational systems and other sources help determine their relevance to the target database. Documentation for these data sources may or may not be available, or it may be inaccurate. Samples of the source data are analyzed to detect the data properties.

Step 3. Build a set of source-data-to-target-database mappings. This step involves devising a plan to transform the various data sources to the target. Typically, it is performed by an extract, transform and load (ETL) tool or by hand-coded programs.

Step 4. Stage the data. Loading the source data into a staging area enables it to be massaged, cleansed and manipulated into the form needed for the target data store. Data quality software may be deployed during this stage to standardize and link records.

Step 5. Load the data. Moving the data from the staging area into the target application includes formatting the data for reporting. While this approach appears logical, it contains flaws that contribute to the high failure rate for data integration projects: It is highly dependent on manual effort, and it makes the fatal assumption that companies "know" their data.

Do not assume "we know our data"

The primary weakness in the traditional data integration approach is its assumption that data required for an application is actually available from the data sources. Major corporations have spent millions of dollars on data integration projects, only to discover that the source data will not support the target model. This can occur whether companies build the model internally or an enterprise application vendor defines it. Because the process consists of a series of disjointed steps, usually executed manually by independent teams of programmers, the discontinuity between the steps often leads to disaster.

Organizations typically spend 80 percent of their project budget on staging and loading data. Unfortunately, the actual mechanics of specifying a set of source-data-to-target mappings is a small part of the overall task of integrating multiple data sources. The real work lies in the *necessary* exercise of determining answers to several questions:

- *What exactly is in the source data?*
- *How is the data organized?*
- *What is the quality of the data?*
- *Is it fit for its intended purpose?*

Most data integration projects that overrun their budgets or fail entirely are often the result of not understanding the metadata. Without automated metadata reverse engineering tools, developers are left to investigate the source data by hand. Documentation for the metadata of legacy systems is usually incomplete at best, or at worst, non-existent. The personnel needed to interpret the data often have left the company, leading to haphazard guesses rather than a complete analysis of content. The result is a process where the integration of source data into the target data store is debugged far downstream in the development cycle. Problems in the metadata are reflected too late in the process—in production systems—rather than being addressed at design time.

A defect that is not detected upstream—during requirements or design—will cost from 10 to 100 times as much to fix later. With data integration, this translates into a significant financial loss for the enterprise that attempts to work with data without truly understanding the properties of the source data, and to manually build target databases. The lack of tools to detect problems in the ETL process upstream can cost businesses a significant portion of their data warehousing budgets.

Start profiling data for success in data integration

Poor data quality is the root cause of failure across a wide range of corporate initiatives. Profiling source data up-front generates significant benefits:

- *Helps reduces project risk*
- *Enhances ROI on a variety of enterprise projects, including business intelligence, enterprise application implementation, instance consolidation, single view of customer, master data management, and regulatory and compliance initiatives*
- *Validates business requirements as achievable or unachievable*
- *Helps ensure that disparate source data supports target requirements before the investment of time and resources in the data integration development effort*
- *Pinpoints data problems early during the project cycle, substantially reducing costly testing and correction efforts*
- *Enables more accurate project planning of resources (people, skill sets and time)*

IBM WebSphere Information Analyzer offers profiling tools

IBM® WebSphere® Information Analyzer, a module of IBM Information Server, brings automation to the critical and fundamental task of data source analysis, expediting comprehensive data analysis, reducing time to value and minimizing overall costs and resources for critical data integration projects. WebSphere Information Analyzer profiles heterogeneous source data—analyzing columns, tables, primary and foreign keys, relationships and redundancies.

WebSphere Information Analyzer helps users integrate multiple disparate systems by providing a complete understanding of the metadata and by discovering dependencies within and across tables and databases. Because the metadata is based upon the actual source data, its accuracy is typically 100 percent, reducing the project risk by uncovering integration issues before development begins. Leveraging this advanced data profiling capability can help companies achieve a robust and reliable implementation that avoids critical and costly data integration problems. WebSphere Information Analyzer can take the typical six- to eight-month project and deliver the same results in 30 to 60 days—a 70 percent average time savings.

Understand the source data

WebSphere Information Analyzer makes no assumptions about data content. It reads any source data and automatically analyzes and completely profiles the data so that the data properties—the metadata—are generated without error. The properties include the tables, columns, probable keys and interrelationships among the data. More than 30 out-of-the-box reports help users understand results quickly and efficiently.

Key processes lead to profiling success

Data profiling within WebSphere Information Analyzer includes several major processes.

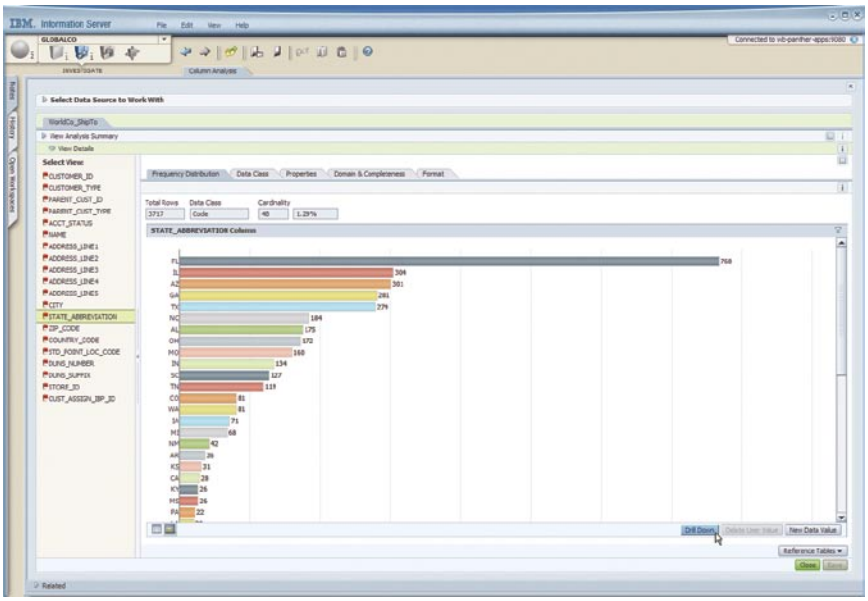
Column analysis

Column analysis examines all values for the same column to infer the column's definition and other properties such as domain values, statistical measures and minimum/maximum values. During column analysis, each available column of each table of source data is individually examined in depth. Many data properties are observed and recorded including:

- *Minimum, maximum and average length*
- *Precision and scale for numeric values*
- *Basic data types encountered, including different date/time formats*
- *Minimum, maximum and average numeric values*
- *Count of empty values, NULL values and non-NULL/empty values*
- *Count of distinct values or cardinality*
- *Full frequency distribution of values*
- *Full frequency distribution of data formats*
- *Count of data formats*

Figure 1 shows a frequency distribution.

Figure 1: Frequency distribution



Additionally, column analysis makes certain inferences about the data in the column; for example:

- *What data type, precision and scale apply to the column*
- *Whether NULLs are permitted*
- *Whether the column contains a constant value*
- *Whether the column values are unique*

Column analysis also supports the ability to assess completeness and validity of the values or formats in a column. Selecting by value, range or reference table, users can designate specified values as default or invalid or specified formats as non-conforming to standard. This information can be assessed and reported repeatedly over time to identify potential issues in data quality or it can be provided in reference tables to developers.

During column analysis, users create notes that can be shared with IBM Information Server data integration processes, contributing significantly to the project ROI.

Primary key analysis

Primary key analysis is the process of identifying all candidate keys for one or more tables. The goal is to detect a column or set of columns that might be appropriate as the primary key for each table.

WebSphere Information Analyzer immediately utilizes the results from column analysis to assess single-column primary keys without any additional processing. Further, the user can explore multicolumn primary keys, either through assessing a set of columns against a data sample with permutations of up to nine columns for uniqueness, or by directly assessing a chosen set of specified columns for uniqueness against the whole data source. The user can subsequently designate one candidate key as the primary key.

Foreign key analysis

Foreign key analysis is the process of comparing all columns in selected tables against the primary keys in those same tables. The goal is to detect the existence of a foreign key relationship between two tables based on the overlap of values between each specified column and the identified primary key.

Building off the primary key analysis and column analysis, WebSphere Information Analyzer first identifies the primary key for each table, and then finds columns across the specified tables or files with the same data classification and data type. Where these pairings are a match, the foreign key analysis process identifies overlapping data, from which the user can review and designate the primary key and corresponding columns as a foreign key relationship.

Cross-domain analysis

Cross-domain analysis is the process of comparing all columns in each selected table against all columns in the other selected tables. The goal is to detect columns that share a common domain. If a pair of columns is found to share a common domain, this might indicate consistency in the data stored between the two tables—such as consistent use of state or country codes—or it might simply indicate redundant data.

The common domain is noted from the perspective of both columns; that is, the user can review the association in either direction from either column. If the data is found to be redundant, users can mark it accordingly. This type of analysis can be performed repeatedly over time, both in the same sources or in new sources that are added to a project to continuously build out the knowledge of cross-domain relationships.

Avoid the pitfalls of the traditional manual process

Merging the traditional steps of data integration into an integrated process—with the addition of enlightened inferences from the metadata—can help avoid the pitfalls of the traditional manual process. Some advantages of using WebSphere Information Analyzer include:

- *The correct metadata is generated from the content that actually exists in the data, rather than from the wishful thinking of developers.*
- *Invalid data can be spotted and rectified early in the project.*
- *Accurate documentation for the source data is automatically created from reports in the system and verified by the user. The documentation is automatically generated and reflects the actual data in the source system.*
- *There is no dependence upon the programmers who developed the applications that produced the source data. Access to the data is the only resource needed.*
- *Keys are inferred from what is actually present in the data.*
- *Field types are inferred from what is actually present in the data.*
- *The true range of domain values for coded fields is generated and mapped as part of the specification.*
- *Dependency relations are inferred from what is actually present in the source data.*

The productivity achieved by utilizing WebSphere Information Analyzer reduces staffing requirements for a data integration project. This does not mean that using WebSphere Information Analyzer eliminates any possible problems in the process. Analysts and developers must still make informed decisions and apply their talents to the problems. But the elimination of the vast array of pitfalls that traditional multistep data integration projects encounter can dramatically reduce the time and effort needed for the project.

Customer implementations across multiple industries have shown that WebSphere Information Analyzer can take the typical six- to eight-month project and deliver the same results in 30 to 60 days. WebSphere Information Analyzer contributes to a favorable result, identifying serious problems with the source data early in the process when correction is far less costly in both time and budget.

IBM Information Server delivers information you can trust

By addressing key questions about source data at the beginning of any data integration project, WebSphere Information Analyzer is an integral part of IBM Information Server.

IBM Information Server is a revolutionary new software platform that helps you derive more value from the complex, heterogeneous information spread across your systems. It enables your organization to integrate disparate data and deliver trusted information wherever and whenever needed, in line and in context, to specific people, applications and processes. It helps business and IT personnel collaborate to understand the meaning, structure and content of any type of information across any source. It provides breakthrough productivity and performance for cleansing, transforming and moving this information consistently and securely throughout the enterprise, so it can be accessed and used in new ways to drive innovation, help increase operational efficiency and lower risk.

For more information

For more information about WebSphere Information Analyzer or IBM Information Server, contact your IBM marketing representative or IBM Business Partner, or visit **ibm.com**/software/data/integration

IBM Software Group
Route 100
Somers, NY 10589

Printed in the United States of America
December 2006

TAKE BACK CONTROL WITH **Information Management**

IMW11808-USEN-00