



Data Profiling and Mapping

The Essential First Step in Data Migration and Integration Projects

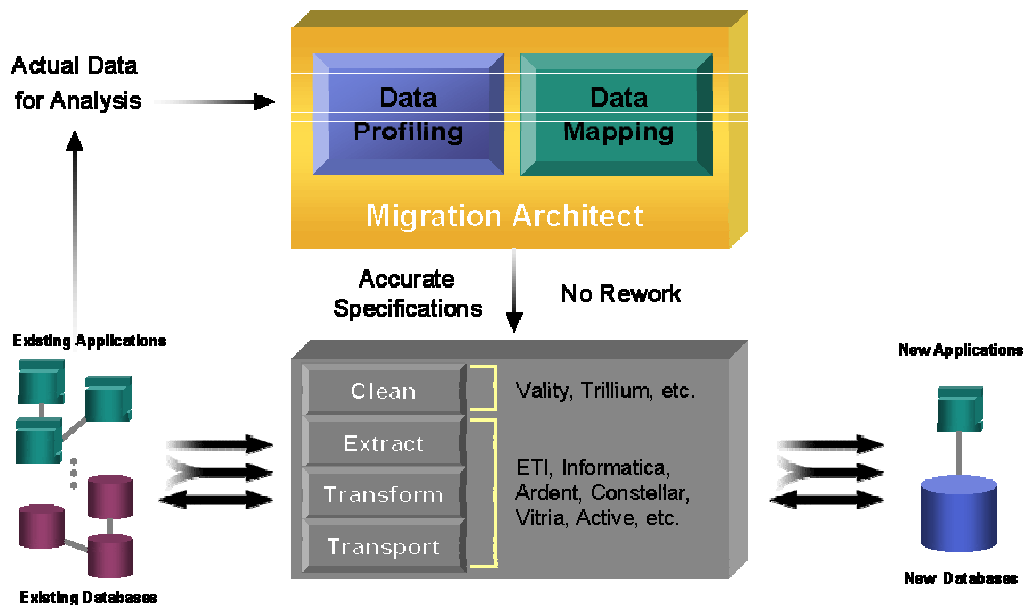
An Evoke Software White Paper

Summary

At any given time, according to industry analyst estimates, roughly two-thirds of the Fortune 1000/Global 2000 are engaged in some form of data migration or integration project—including implementation of new ERP, CRM and e-commerce applications, data consolidations, data quality improvements, and creation of data warehouses and data marts. These projects are driven by increasing worldwide competition, industry consolidation, and constant pressure to increase revenues and profits.

To achieve success, each of these projects must follow a path that begins with studying the source data to thoroughly understand its content, structure, and quality—a process called Data Profiling. Once the data has been profiled, an accurate set of mapping specifications must be developed based on this profile—a process called Data Mapping. These two processes of Data Profiling and Mapping comprise the essential first steps in any successful data migration or integration project and should be completed *prior to* attempting to extract, scrub, transform, and transport the data.

Migration Architect The Data Profiling and Mapping Solution



A new category of software that automates many of the complex processes involved in Data Profiling and Mapping has emerged to simplify and accelerate these projects. The remainder of this document discusses the market drivers and opportunity for this type of software, presents an overview of Data Profiling and Mapping, and concludes with a look at associated benefits.

Market Drivers and Opportunities

The only true constant for today's corporations is change. Technology standards have evolved dramatically over the years. As a result, corporate computing environments are overpopulated with disparate, poorly integrated applications and databases. For a variety of reasons—financial pressures, increased competition, ongoing deregulation, mergers and acquisitions, and European currency concerns—corporations are consolidating their data sources and integrating their applications. This involves moving from legacy environments to packaged applications based on modern relational technologies, as well as using data from existing applications to support second-generation e-business applications, corporate-wide CRM implementations, and enterprise information portals.

According to recent studies conducted by The Standish Group, a research advisory firm based in Dennis, Mass., 15,000 data migration projects with budgets of \$3 million or greater began in 1999 at a total cost of \$95 billion. Data from other market research firms confirms the significant market opportunity. AMR Research estimates that ERP conversion projects will generate \$52 billion by 2002. Gartner Group/Dataquest forecasts that \$2.6 billion will be spent on data warehousing software in 1999, growing to \$6.9 billion by 2002. International Data Corporation forecasts that e-commerce revenues, including business-to-consumer and business-to-business, will top \$1 trillion by 2003, and will require extensive investments in supporting technology.

Unfortunately, most of these data migration, integration, and consolidation projects don't go as smoothly as anticipated. The Standish Group studies indicate that 88 percent of the 15,000 data migration projects starting in 1999 will either overrun or fail. One of the primary reasons for this extraordinary failure rate is the lack of a thorough understanding of the source data early on in these projects. Conventional approaches to data profiling and mapping can create nearly as many problems as they resolve—data not loading properly, poor quality data and compounded inaccuracies, time and cost overruns and, in extreme cases, late-stage project cancellations. The old adage 'garbage in, garbage out' is very applicable here. The market opportunity is enormous for an innovative solution that can help companies solve these ongoing problems.

Fortunately, such a solution exists in the form of Data Profiling and Mapping software from Evoke Software Corporation.

What Is Data Profiling and Mapping?

Data Profiling and Mapping is the process whereby the content and structure of legacy data sources are examined and understood in detail, and mapping specifications are produced for the successful movement and transformation of the data from source to target. This process consists of six sequential steps, three for Data Profiling and three for Data Mapping, with each step building on the information produced in the previous steps. The resulting transformation maps are used as specifications to drive data extraction, scrubbing, transformation, and transport processes. These data movement processes can be implemented by writing custom code or using third-party migration or integration products.



Conventional Approach to Data Profiling and Mapping: Problems and Pitfalls

The conventional approach to Data Profiling starts with a large team of people (data and business analysts, data administrators, database administrators, system designers, subject matter experts, etc.). These people meet in a series of joint application development (JAD) sessions and attempt to extract useful information about the content and structure of the legacy data sources by examining outdated documentation, COBOL copy books, inaccurate metadata and, in some cases, the physical data itself. Typically, this is a very manually intensive process supplemented, in some cases, by semi-automated query techniques. Profiling legacy data in this way is extremely complex, time-consuming, and error-prone. And, once completed, the team has achieved only a limited understanding of the source data.

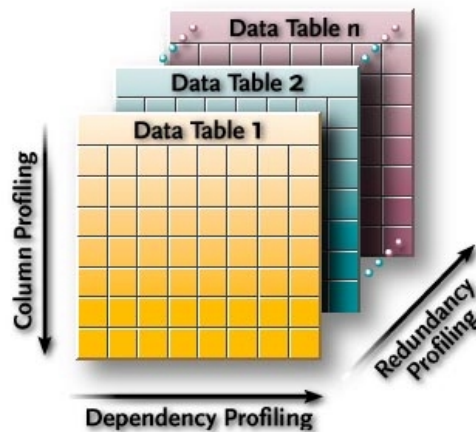
At that point, according to the project flow chart, it is time to move on to the mapping phase. But since the source data is so poorly understood and inferences about it are largely based on assumptions rather than facts, this phase typically results in an inaccurate data model and set of mapping specifications. Based on this information, the data is extracted, scrubbed, transformed, and moved or integrated.

Not surprisingly, in almost all cases, the new system doesn't work correctly the first time. Then the rework process begins: redesigning, recoding, and retesting. At best, the project incurs significant time and cost overruns. At worst, faced with runaway costs and no clear end in sight, senior management cancels the project, preferring to give up the promised but apparently unattainable benefits in favor of the status quo.

Data Profiling with Migration Architect

Unlike conventional approaches to Data Profiling, Migration Architect utilizes a combination of automated discovery and interactive analysis to provide data and business analysts—for the first time—with a thorough understanding of their source data, including content, structure, quality, and integrity. With Migration Architect, data sources are profiled in three dimensions: down columns (Column Profiling); across rows (Dependency Profiling); and across tables (Redundancy Profiling).

MIGRATION ARCHITECT™ PROFILES DATA IN THREE DIMENSIONS



Column Profiling

Column Profiling analyzes the values in each column or field of source data, inferring detailed characteristics for each column, including data type and size, range of values, frequency and distribution of values, cardinality, and null and uniqueness characteristics. Interactive drill-down capabilities allow analysts to detect and analyze data content quality problems and to evaluate discrepancies between the inferred, true metadata and the documented metadata.

Dependency Profiling.

Dependency Profiling analyzes data across rows—comparing values in every column with values in every other column—and infers all dependency relationships that exist between attributes within each table. This process cannot be done manually. Dependency Profiling identifies primary keys and whether or not expected dependencies (e.g., those imposed by a new application) are supported by the data. It also identifies ‘gray-area dependencies’—dependencies that are true most of the time, but not all the time—usually an indication of a data quality problem. Dependency Profiling is critical to the subsequent elimination of duplicate information and the production of a true third normal form model of the data sources.

Redundancy Profiling.

Redundancy Profiling compares data between tables of the same or different data sources, determining which columns contain overlapping or identical sets of values. It looks for repeating patterns among an organization’s ‘islands of information’—billing systems, sales force automation systems, post-sales support systems, etc. Redundancy Profiling identifies attributes containing the same information but with different names (synonyms), and attributes that have the same name but different business meaning (homonyms). It also helps determine which columns are redundant and can be eliminated, and which are necessary to connect information between tables (foreign keys needed for referential integrity). Redundancy Profiling eliminates processing overhead and reduces the probability of error. As with Dependency Profiling, this process cannot be done manually.

Data Mapping: The Remaining Three Steps

Once the Data Profiling process is finished, the profile results can be used to complete the remaining three Data Mapping steps: Normalization; Model Enhancement; and Transformation Mapping.

Normalization

Using the information gathered from the first three Data Profiling steps, Migration Architect builds a fully normalized relational model based on the consolidation of all the data. Because the model is fully supported by the data—rather than by assumptions and inaccurate metadata—it will not fail.

Model Enhancement

Users can modify the normalized model by adding structures to support new requirements, or by adding indexes and denormalizing the structures to enhance performance. Then Migration Architect produces data definition language (DDL) for the resulting model, complete with referential integrity instructions. The DDL is tailored to Oracle, Informix, Sybase, DB2 for AIX, or ANSI SQL-92 environments as required. It may also be imported into graphical design tools such as Computer Associates' *ERwin*.

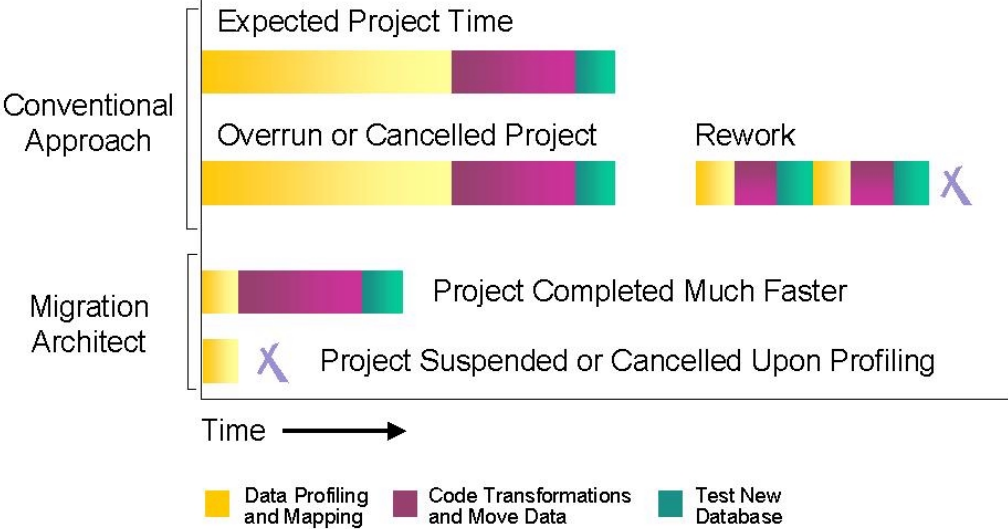
Transformation Mapping

When data model modifications are complete, Migration Architect creates a set of transformation maps that show the relationships between columns in the source files and tables in the enhanced model, including attribute-to-attribute flows and scrubbing and transformation requirements. These maps provide essential information to the programmers creating conversion routines to move or integrate data.

Industry studies suggest that conventional approaches to Data Profiling and Mapping take between three and five hours per attribute (or data element). And this does not include Dependency and Redundancy Profiling, complex processes that typically involve millions of comparisons between attributes within the same table and across tables of disparate sources so as to uncover functional dependencies, primary and foreign keys, duplicate data, etc.

In sharp contrast to conventional methods, Data Profiling and Mapping with Migration Architect can be done in a fraction of the time—minutes per attribute instead of hours—while providing users with a much more thorough understanding of the source data. Based on this understanding, projects are completed successfully the first time, accelerating time-to-benefit and lowering project cost and risk. In addition, an accurate data profile improves communication between IT professionals and business end users—groups that rarely speak the same language—by providing the data facts necessary to make objective business decisions. This results in higher data and application quality and greater end user and customer satisfaction.

Accelerate Project Completion with Migration Architect



The Bottom Line

Developing an accurate profile of existing data sources is the essential first step in any successful data migration or integration project. Data Profiling software enables a small, focused team of technical and business users to quickly perform the highly complex tasks necessary to achieve a thorough understanding of source data—a level of understanding that simply cannot be achieved through conventional manual processes and semi-automated query techniques.

Data Profiling software enables data migration, integration, and consolidation projects to be completed successfully the first time, eliminating extensive design rework and late-stage project cancellations. It can even warn IT management if the business objectives of the project are not supported by the data, dramatically lowering project risk and enabling valuable resources to be re-directed to other, more fruitful projects. Finally, Data Profiling will deliver higher data and application quality, resulting in more informed business decisions and greater revenues and profits.