



**TRILLIUM
SOFTWARE**
a division of Harte-Hanks

Improving Data Migration
with Automated Data Profiling

**Presented by:
Harte-Hanks Trillium Software**

Improving Data Migration with Automated Data Profiling

Data Profiling Keeps Data Migrations on the Fast Track

Companies are finding that they can streamline data migrations and save large amounts of time and expense by using an advanced data profiling tool during the preliminary stage of their projects.

“Large investments for enterprise application initiatives regularly fail to yield the expected results because of lack of attention to data quality issues and corresponding low adoption rates.”

Andreas Bitterer, META Group

The Data Migration Dilemma

Companies often view data migration as an unavoidable and rather unexciting step in building the right environment for a new enterprise application. But these projects are often far more complicated—and expensive—than companies anticipate, creating additional data problems and costing far more than was originally budgeted. Why does this happen? Industry researchers, like Gartner and META Group, agree that companies often make mistakes in initial assumptions, both about the integrity of the source data and what an ETL (Extract, Transform, Load) tool can do about it.

Costly assumptions

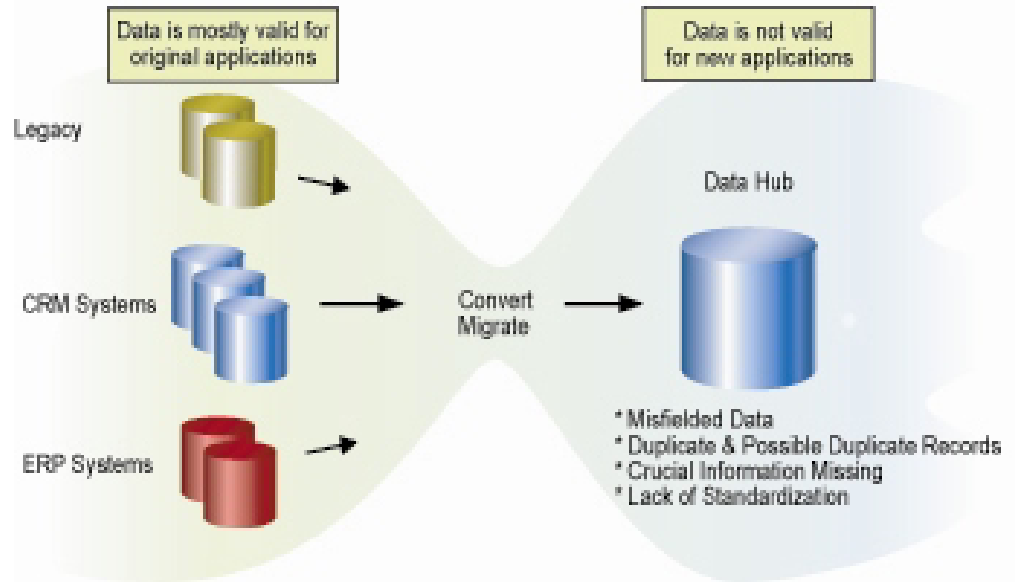
The reason companies assume the data is fine is simple and easy to understand: The data works pretty well in the original source systems. Therefore, the assumption goes, it should work in target systems. Data in source systems, however, may serve one specific purpose while the migrated data is supposed to serve another or multiple purposes, with different data quality requirements. Thus it is important to assess whether the data to be migrated is fit for the new purpose it will be used for.

Furthermore, source systems themselves often contain special routines for correcting their own data. These routines have been written over time to solve specific data anomalies in particular applications, but they are always lost during migration. As a result, data may not load properly, compounding inaccuracies, time and cost overruns, and, in extreme cases, late-stage project cancellations. Understanding the data within source systems, including all its inconsistencies and anomalies, is thus a critical step in any integration project.

Another source of additional data problems arises from the fact that data migration projects nearly always involve combining data from more than one system. Data records that may work relatively well within a single system, when combined with other sets of equally questionable quality, produce massively inconsistent, cryptic results, based on divergent purposes, and consequently different data structures, definitions, and sets of content.

ETL tools: What's possible, what's not

ETL tools offer great connectivity into disparate applications and they make it easy to map information from one system to another and handle metadata, but they fall short when it comes to many tasks, such as checking for continuity, profiling the integrity of the information, and identifying mis-fielded data, anomalies, and other inconsistencies. This type of analysis and intelligent transformation is crucial, researchers point out, but it is often overlooked, largely because organizations assume that the source data does not need to be fixed.



Data that works in specific legacy applications often does not fit precisely the new purposes it is intended to be used for after it is migrated and consolidated for a new enterprise application. To avoid costly rework after data is migrated, advanced data profiling tools uncover problems—inconsistencies, inaccuracies, and incomplete data fields, rows, or columns—that are better addressed prior to migration and in conjunction with ETL conversions.

Automated data profiling can reduce the time it takes to analyze data by 90% or more over manual methods.

Common Techniques for Analyzing Your Data

Companies can basically use one of three options when they want to analyze source data before migrating it to a new system:

- Review the data manually
- Write SQL scripts to profile the data
- Use a data profiling tool

An analysis of these options shows their effectiveness for large-scale migration projects that enterprises need today.

Manual review of data

The old-fashioned way of solving data quality problems with source data was to hire a team of database workers to manually analyze the data. However, especially with today's large databases, this can take years, several people, and a big budget. What's more, data changes constantly. Customers get married, move, buy new services or products, divorce, move again, and so on. Product information can be equally volatile, frequently making 30% or more of product information inconsistent or incorrect. A long-term manual process can never keep pace with such change and thus can never completely rectify the data.

Causes of poor data quality

- Lack of standards—Information represented in multiple ways.
- Staleness—Inconsistency that develops between data and the real-world entities it represents as those entities change over time.
- Missing/unrecognized data—The absence of vital information or the presence of information that is not easily understood—for example, because of language or character set variations, data in free-format text fields, and unknown abbreviations.
- Phantom data (“filler”)—Unusable information, often occurring when data entry personnel fill required fields with meaningless characters—for example, “999- 99-9999” in a social security number field.
- Misinterpretation - Misreading of data elements that look similar but have different meanings—for example, “CO” as an abbreviation for company, Colorado or care of.
- Misfielded data—Data entered into the wrong fields becomes meaningless.

Profiling in SQL

Another way to gain insight into source data is to write queries, typically SQL queries, against database information. There are a couple of problems with this approach: First, data from multiple sources—often a large number of sources, millions of rows of data, and diverse formats—must be consolidated. The data can be in flat files or DB2 files, in ASCII or UNICODE, on UNIX systems or mainframes. The mere act of consolidating these sources can present an IT challenge itself.

The second problem is the unsatisfactory results of traditional querying under these circumstances. Without a clear understanding of how the data is interrelated within a data set and how sets are related to each other, it is difficult to know what queries to run. Not knowing what to look for makes it hard to zero in on the inconsistencies and anomalies that need to be addressed. Often this method of querying creates more questions than it answers.

Manually cleaning the database with a team of data professionals or querying a database to identify problems produces the same results: inaccurate data models based on false assumptions and a new system that does not work. Then the rework process begins: redesigning, recoding, reloading and retesting. At best, the project incurs significant time and cost overruns. At worst, faced with runaway costs, senior management pulls the plug, preferring to make do with existing legacy systems rather than incur the costs of a seemingly endless data migration project.

Paving the Way for Data Migration Today

Today, many project managers and systems integrators have learned from past mistakes. They know the importance of understanding the data before starting a migration project. The data profiling process is the best way to accurately plan data integration and migration projects and eliminate the risks associated with data quality problems. It also reduces costs and increases productivity as much as 90% over manual methods.

Rather than viewing code and writing SQL scripts, multiple users are able to easily examine all the data with products like Trillium Software® Discovery, allowing even business users to be part of the discovery phase. Data profiling can be a process that automatically and completely identifies potential exceptions and anomalies, rather than leaving data migration managers with having to write scripts based on what they suspect the anomalies to be. Data profiling uncovers such things as missing and duplicate data, misspelled data, broken data rules, invalid data structures, incorrect content, and irreconcilable data.

Consolidating operational data for BI and reporting

In order to use operational data for business intelligence and reporting, Ford Financial Europe needed to consolidate multiple source systems into one, and they turned to Trillium Software Discovery for the solution. They needed a data analysis solution to identify data issues in three source data systems serving 12 countries. As Lee Edwards, data architect at Ford

Defining Data Quality

Data quality is measured by a data set's ability to meet the needs of business users. Beyond postal correction, enterprise data quality provides accurate, unified views of business entities to support sophisticated applications. In BI, better data quality means more reliable analyses. In CRM, it means a deeper understanding of the customer base. In ERP, it means data integrity across complex processes. For any data-dependent process, data quality determines the maximum business potential.

Here are some questions to ask when evaluating the quality of a data set:

Is it accurate? How free of error is the data?

Are the records complete? To what degree are the fields populated with data?

Is the data set consistent? Does it satisfy a set of internal constraints?

Is the data current? How up-to-date is the data?

How many of the records are unique? How redundant is the data?

Does the content fall within allowable values? How valid are the records?

Financial Europe explained: "It costs 100 times more to repair a problem with an actual integration interface than if the issue is identified before coding begins."

Using Discovery, Ford identified six major data issues in two source systems. They were able to drill down into them and rectify them in just a couple days. "Had we not identified these prior to integration, they would have taken around 180 days of effort to overcome," said Edwards.

By addressing the issues that the discovery process identified, Ford has improved its monthly reporting thus increasing support for the many decision making processes across the enterprise. In fact, recognizing the success of Discovery, other divisions of Ford in Europe as well as in the US are beginning to use the profiling process to fully analyze data before they take on large-scale enterprise-wide projects.

Synchronizing Supply Chain Data

The UK Ministry of Defence (MoD) also turned to Trillium Software® for help in consolidating hundreds of disparate systems. The extensive logistics network consisted of hundreds of systems, with each branch of the armed forces having its own unique set of systems. To describe inventory, the MoD used the NATO Codification System (NCS) to provide the basic code used by inventory, procurement, and supply systems. In theory, each item type should have had its own unique code, but the reality was very different. Across each of the services, three separate NATO Stock Numbers (NSNs) could be used to identify a single item. "Someone in the Navy could enter their stock number for a printed circuit board, but using the Army system, they'd get a plastic bag. To prevent the potentially catastrophic consequences of such a failure on operations, where lives could be at stake, we recognized that we needed a single master database for use by all," said Colonel Nigel Stafford, who was a team leader in the project.

"When we originally investigated the problem, we calculated that it would take 100 staff 178 years to manually validate all the Defense Logistics Organization (DLO) inventory management data. In just one instance, we determined that Discovery would take just over an hour to check 1.7 million records with an assumed seven attributes per record. It offered an enormous time reduction and therefore a direct cost saving," explained Stafford.

In the first two years of the project, it is estimated to have delivered supply chain cost savings of roughly \$36 million, simply by removing data inaccuracies and inconsistencies, leading to the elimination of duplicate and obsolete items.

Data Profiling Essentials for Data Migrations

Operational business units usually perceive data quality as an IT issue. But the reality is that, while the technology of data quality is often implemented and maintained by the IT staff, it is the business user that benefits most. Equally important, it is the business user who understands the meaning and usage of the data in the context of each application. For

these reasons, IT and business users need to agree on data standards and content and align their expectations about the key role data quality plays within the organization.

Automatic rapid data profiling

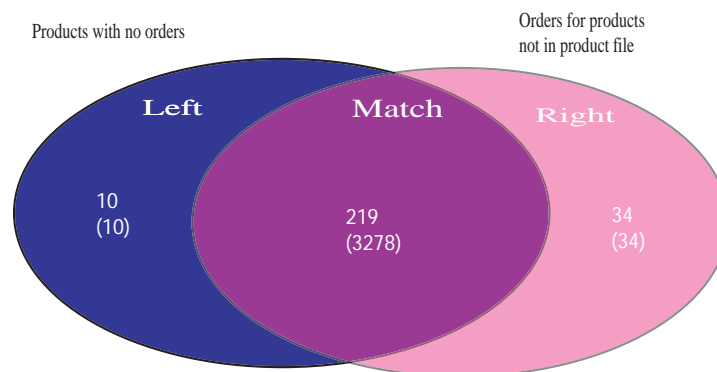
The best data profiling tools make no assumptions about the data. A data profiling tool should provide a comprehensive and automatic analysis of the full data sets to be migrated. It should give users the ability to explore the interrelationships between data fields within data sets and relationships between data fields across data sets as well.

Intelligent reporting and analysis

When focusing on inconsistencies and anomalies, no amount of sampling can deliver the precise and detailed kind of information that a full analysis of the data sets can do. Data profiling typically offers complete reports, including statistics, frequencies, ranges, data values, and content, including identifying redundant and missing data values, duplicates, and misspellings as well as incorrect data formats. Advanced data profiling tools deliver additional information about how well the data complies with data standards and business rules. They will give users options for customized reporting and provide the results in a way that directs users to potential problems.

Centralized repository for all data

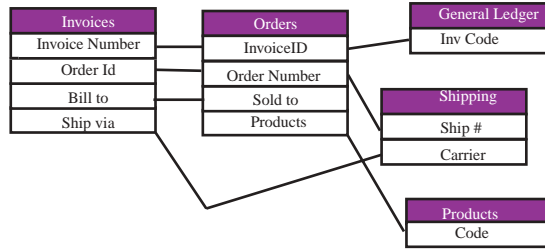
The best of data profiling software creates a centralized environment for comprehensive data analysis without impacting performance on operational systems. This environment should include all data, metadata, statistics, and documentation. Other types of profiling systems lower performance by querying directly against live systems or require separate copies of systems. With a centralized environment, users have a fast, scalable architecture that serves multiple users while isolating the data from any changes.



Venn diagrams help users visualize the relationships between data fields, columns, or rows, showing how consistent data is across data sets. The diagram pictured here shows how much overlap there is between a product database and invoices, revealing that some products have never been ordered and some invoices have been sent for nonexistent products.

“Data integration and building the underlying subject-area for the data mart / data warehouse consumes up to 70% of the effort in building an analytic solution.”

Henry Morris
VP of Research, IDG



Relationship diagrams can show how data content in one set relates directly to data in another set. In the diagram above, field names for invoices, orders, customers, and products vary from one data set to another but their correlations are revealed in the diagram.

User-defined business rules and data standards

Business users familiar with the data and the business meaning should be able to easily write business rules and check data for acceptable ranges and values. These invaluable tools allow users to automatically validate values and set key indicators for monitoring, compliance, and data governance. The ability to create standard definitions for data and business rules are critical functions during the data profiling and analysis process.

Drill down and graphical display

In order to fully understand what inconsistencies mean, users must be able to efficiently drill down into the data itself, examining, when necessary, specific content within its original context. The best data profiling tools offer an easy and intuitive graphical user interface that allows users to rapidly zero in on problems, explore multiple anomalies, define new data standards, and then quickly return to the summary report. Additional graphical tools for revealing the relationships between data fields across data sets should include Venn diagrams and relationship (dependency) mapping.

Enhanced communications

When analyzing data, users need flexible ways to communicate their data quality concerns, including features for annotating specific data directly within the working environment, for book marking issues for others to view, for copying data for further analysis, and for creating HTML reports for emailing to team members to alert them to data quality problems. Being able to save specific data, subsets of reports, and specialized data views allows team members to share their data discoveries with others and effectively work together to resolve problems.

End-to-End Data Quality Process

The data profiling phase is step one in creating consolidated and consistent views of all types of business data, including product data, sales data, supply chain data, and customer data. What's next? Solutions offered by Trillium Software form a complete end-to-end suite of data quality products and services. You have read about functionality available through Trillium Software Discovery. In addition, Trillium Software offers:

The risks inherent in using raw data are like the uncharted rocks hidden beneath the water's surface.

Trillium Software System®

Once the migration team understands the nature of the content to be migrated, it can enforce accurate business rules for cleansing, standardizing, matching, and consolidating data, using the Trillium Software System. This system offers connectors and APIs for integration with the ETL process for standard ETL tools, including Informatica PowerCenter, Oracle Warehouse Builder, Ascential DataStage, and Microsoft Data Transformation Services (DTS). Connectors are also available for Ab Initio, IBM Warehouse Manager for DB/2, and Business Objects Data Integrator, via those companies.

TS Enrichment

In conjunction with its parent company, Harte-Hanks, Trillium Software is able to offer a unique service to customers of the Trillium Software System. They can choose from over 5000 third-party sources to create full customer views, enriching their corporate data with demographics/firmographics, contact information, physical addresses, and preferred and optimal contact information. Harte-Hanks marketing database experts provide guidance in choosing the best third-party sources, and Trillium Software Professional Services ensures seamless appends to existing corporate data.

What is Data Quality?

Group	Quality	Issues Considered	Example of Data Quality Problem
Content within data values	Complete	Is all necessary data present?	17% of customer name data is blank
	Accurate	Does the data accurately represent reality or a verifiable source?	A Supplier is listed as 'Active' but went out of business six years ago
	Valid	Do data values fall within acceptable ranges defined by the business?	Transaction Data = 02/07/1902, but business started in 1969
	Fit For Purpose	Is the information valuable to the business? Does the data convey information that can intelligently be consumed by the business?	A person has a SIC code A business has a gender code
Structure of fields	Format	Do values follow consistent formatting standards?	Telephone number appears as xxxxxxxxxx, (xxx) xxx-xxxx, 1.xxx.xxx.xxxx, etc.
	Standard	Are data elements consistently defined and understood?	Gender code = M, F, U in one system and Gender code = 0, 1, 2 in another system
	Consistent	Do values represent the same meaning across systems and files?	Profit margin is calculated differently across units, using two different formulas
Relation to other data	Referential Integrity	Do records exist where expected? Do they contain unnecessary or inactive data? Are reference files/tables complete?	Transaction records carry codes for products that do not exist on the Product table/file
	Cardinality	Is the structure of relationships among entities and attributes maintained consistently?	A customer has more than one "Current Customer Profile" record

Avoiding Other Data Migration Hazards

By including data profiling as a step within a data migration initiative, companies can take their data migration projects to completion successfully the first time around eliminating extensive design rework and late-stage project cancellations. A good data profiling and data quality methodology will help assess the entire scope of a project, even warning IT management if the business objectives of the project are not supported by the data. Data profiling and data quality, if done correctly, can thus dramatically lower project risk, enabling valuable resources to be redirected to other, more fruitful projects.

About Trillium Software.

Trillium Software® a division of Harte-Hanks, has been selected by companies worldwide, both large and small, to improve their operational and analytic business decisions through accurate and timely information. Trillium Software offers an integrated suite of Total Data Quality software and services architected to turn raw, inaccurate data into usable, real-time information that supports business functions. The Trillium Software System® and Trillium Software Discovery are recognized as critical to the success of data warehouse, business intelligence, customer relationship management, enterprise resource planning, supply chain management, e-business, and other enterprise applications, and data integration, data migration, data stewardship, and data governance initiatives.

Trillium Software System® provides cleansing, standardization, matching, enrichment, and linking functions for global customer data and operational business data, in any and all systems and applications. Regardless of data source or structure, Trillium Software System ensures that data adheres to established standards, adaptable to fit each organization's specific needs. Both single and double byte data are processed in local languages to provide a unique and centralized view of customers, products and services.

Trillium Software Discovery provides data investigation, discovery, and profiling in support of quality, migration, integration, stewardship, and governance initiatives. Fully integrated with the Trillium Software System, Trillium Software Discovery includes a broad range of features that expand upon basic profiling, including automated monitoring, business rule validation, and trend analysis.

Trillium Software Enrichment provides additional data enhancement services to complement, supplement, and amplify data available in-house. Choose from over 5000 third-party data sources, and administrate enrichment through a single vendor.

Usage Notice

Permission to use this document is granted, provided that: (1) The copyright notice "©2005 by Trillium Software®, a division of Harte-Hanks" appears in all copies, along with this permission notice. (2) Use of this document is only for informational and noncommercial or personal use and does not include copying or posting the document on any network computer or broadcasting the document through any medium. (3) The document is not modified from the original version.

It is illegal to reproduce, distribute, or broadcast this document in any context without express written permission from Trillium Software™. Use for any other purpose is expressly prohibited by law, and may result in severe civil and criminal penalties. Violators will be prosecuted to the maximum extent possible.

This document and related graphics might include technical inaccuracies or typographical errors and is subject to change at any time by Trillium Software. Trillium Software does not guarantee the suitability of the information contained in this document, which is provided "as is" without warranty of any kind. Trillium Software hereby disclaims all warranties and conditions with regard to this information, including warranties and conditions of merchantability, whether express, implied, or statutory, fitness for a particular purpose, title and non infringement. In no event shall Trillium Software and/or its respective suppliers be liable for any special, indirect, or consequential damages or any damages whatsoever resulting from loss of use, data, or profits, whether in an action of contract, negligence, or other tortious action, arising out of or in connection with the use or performance of information available from this white paper.