

Combining Data Profiling and Data Modeling for Better Data Quality

Table of Contents

Executive Summary

SECTION 1: CHALLENGE 2

Reducing the Cost and Risk of Data Integration

Improving Data Quality and Analyzing Cross-system Data

Matching Design With Reality

Understanding Legacy Data

SECTION 2: OPPORTUNITY 4

Ensuring the Data That Drives Your Business

Data Profiling

Cross-system Data Analysis

Matching Design With Reality

Uncovering Legacy System Structure and Relationships

SECTION 3: BENEFITS 7

Competitive Advantage Powered by Timely, Quality Data

SECTION 4: CONCLUSIONS 7

Executive Summary

Challenge

Maintaining a competitive edge in any market requires faster, better decisions based on accurate enterprise data. Today, that means companies must manage and maintain their data while facing increasingly complex challenges, such as mergers and acquisitions, budget reductions and evolving technologies. In order to build high-performance business applications, data marts and data warehouses (DWs) or Master Data Management (MDM) initiatives — and get the information you need out of those systems — you must:

- Ensure consistent understanding of your existing and legacy database environments
- Build a solid database foundation for the future
- Control the total cost of ownership (TCO) of the entire data lifecycle

Opportunity

A robust data infrastructure consists of both accurate data and a strong architectural foundation for storing and retrieving this information. Accurate data helps ensure that strategic decisions are made on valid information, while a strong architectural foundation helps ensure that the data is delivered quickly and efficiently.

Data profiling helps you assess the quality of your information and generate metrics and statistics to quantify the results. If an organization is making decisions based on certain information, it is important that this data is reliable, complete and free from error. Data modeling helps you understand both the structure and meaning of your information. Business definitions can be maintained in a data model and structural information can be defined to help ensure that data is stored efficiently. For legacy systems (where this structural information is not well-defined), data profiling can help assess the structure based on statistical analysis of data values. Thus, the combination of data profiling and data modeling is crucial for maintaining the quality of the information that drives your business decisions.

Benefits

CA ERwin® Data Profiler helps increase the quality of your critical data assets by performing cross-system analysis, generating robust data-quality metrics and statistics and validating instance data with database design and architecture. With a strong understanding of your existing data infrastructure, you can reduce the risk of rework and the cost and time involved in complex data integration, data migration, DW and MDM projects.

CA ERwin® Data Profiler analyzes the data in your existing data sources, whether they are databases, VSAM files, spreadsheets or others, and helps reveal the data structures of those sources based on their content. CA ERwin® Data Modeler enables you to visualize these complex data structures and establish enterprise-wide standards for managing data.

The combined solution simplifies data management projects and helps improve the accuracy of the data within them — the basis of your most strategic decisions — thereby helping to increase revenue. It also helps ensure a more efficient data architecture, which helps you to reduce costs.

Reducing the Cost and Risk of Data Integration

In the age of information, organizations depend on their data assets for mission-critical decision making of all kinds — from assessing the position of a competitor, to forecasting future sales, to determining inventory levels for future purchases. Business intelligence (BI) applications make it easier for business and IT alike to gain access to that strategic data in the form of customized reports.

To be effective, front-end BI applications must be supported by a robust data architecture that ensures that the information populating the back-end reports is accurate, timely and high-quality. And in addition to demanding information that is correct, most users want it at the speed of business. But, again, a fast response time for reports requires the right underlying architecture — one that ensures that data is structured and stored in a way that decreases retrieval time, reduces redundancy and minimizes storage volumes.

In other words, you need to be able to quickly understand and cross-reference your data sources and generate accurate, reusable data models. The combination of data profiling and data modeling can assist with creating valuable, well-managed information that helps drive revenue and decrease costs.

Improving Data Quality and Analyzing Cross-system Data

The costs associated with poorly documented, or completely undocumented, data sources often represent a large portion of a project's overall budget — putting the success of any DW, BI, MDM or integration-oriented data management project at risk. Data profiling helps increase the integrity of your critical data assets by performing cross-system analysis, generating robust data quality metrics and statistics and validating live instance data with database design and architecture.

The task of improving data quality through profiling, however, is compounded when there are multiple, disparate data sources to evaluate. As a result of mergers, acquisitions or geographically and organizationally disparate business units, information is often lacking in centralization and reliability. Thus, it is important to identify overlapping data and combine it with data quality statistics in order to create a single source of record.

For example, if two banking organizations merge, they are likely to be interested in ascertaining how many total customers they have, how many might have accounts with both institutions and whether the information stored for each customer is the same (account number, address, age, gender and so on.)

Matching Design With Reality

Most data systems make use of a data model to design the structure and meaning of information. Take, as an example of a simple data structure, the column headings of a spreadsheet. These headings define what type of data should be stored in each column, which, in other words, is, the actual and intended meaning and context of the information assets, or metadata. But in too many organizations, the best intentions of the data architecture team are not followed, and columns are used for an unfortunate and all too vague *something else*.

Rather than add a new column to the data source to store the information they need, rogue data entry personnel simply appropriate an “empty” column for an entirely different purpose. In the example below, for instance, the “Year Purchased” field was used for a promotion code instead. Imagine the difficulties that would result from running or using a report on this data with the goal of determining how long specific customers have been with the company.

FIGURE A

Data quality can only be achieved when data values are compared with design architectures.

COMPARING DATA VALUES WITH DESIGN ARCHITECTURE IS ESSENTIAL TO ENSURING DATA QUALITY

Customer Name	Company	Address	Year Purchased
Joe Smith	Komputers R US	11 Kleiner Ct	AHE92834
Mary Jones	Big Bank Co	10 Gulf Road	UIE238982
Proful Bishwal	Little Bank Inc	1081 Main St	MEO28082
Ming Lee	My Favorite Store	PO Box 987	IYE987234

The table above shows a mismatch between the design architecture (column headers) and the actual data values. The 'Year Purchased' column, which is circled in blue, contains values that are clearly not years, indicating a data quality issue. The diagram also highlights the 'Data Modeling (structure)' for the first three columns and 'Data Profiling (values)' for the entire table.

Understanding Legacy Data

While the problem of database design failing to align with data values can cause reporting errors, it is compounded for many legacy systems because they have no structure defined at all. Imagine trying to understand the meaning of a spreadsheet with no column headers. For the many legacy sources with millions of rows of data, this task becomes even more difficult without the help of an automated system. To be more specific, the primary and foreign keys that are defined in many relational database systems are often not defined at all in legacy data sources.

Ensuring the Data That Drives Your Business

CA ERwin Data Profiler can help you overcome your traditional data analysis and discovery challenges through powerful cross-system data analysis and profiling that integrates with data modeling to ensure both quality data and quality designs. The solution finds hidden inconsistencies in data, provides robust statistics to correct errors in your database or modeling environment and offers integration with CA ERwin Data Modeler, which you can leverage to compare live instance data with model design.

CA ERwin Data Profiler helps you reduce the costs and risk associated with data integration by providing reusable, automated, cross-data-source discovery, analysis and profiling combined with industry-leading data modeling.

Data Profiling

Using a series of rigorous calculations, statistics and metrics, CA ERwin Data Profiler provides an intuitive interface that highlights the critical errors requiring correction in your databases, spreadsheets and legacy applications. For example, before a BI report that lists all of the customers in an organization is run, it is critical to validate relevant customer records to help ensure that there are no issues, such as duplicates, blank fields or unexpected data in a field. Any of these errors could cause incorrect information to be reported, which, in turn, could lead to faulty business decisions.

Cross-system Data Analysis

Most data profiling tools allow you to analyze only a single data source at a time — making the process of comparing the data of more than one system a manual, expensive and time-consuming task. But, CA ERwin Data Profiler includes cross-system data profiling through a series of intuitive wizards — simplifying this potentially complex task — and ensuring that it is easier, faster and less prone to error. Even more compelling is the fact that you can analyze various multiple data sources (up to 20 simultaneously) to identify overlapping and unique attributes and discover attribute supersets and subsets to get a single, comprehensive view of important information.

A common example is consolidating a single view of the customer from several sources — as a result of mergers, acquisitions, internal consolidation of multiple applications and so on. CA ERwin Data Profiler helps you compare these sources, see where overlaps occur and identify a single source of record.

FIGURE B

CA ERwin Data Profiler delivers a single source of record from all compared data sources.

CA ERWIN DATA PROFILER PULLS TOGETHER INFORMATION FROM ACROSS THE BUSINESS



Matching Design With Reality

The architects responsible for building the designs for storing data don't usually have a simple way to test their physical data values in order to confirm that the rules they've defined in the model are being followed. The combination of CA ERwin Data Modeler and CA ERwin Data Profiler helps architects perform data-quality metrics on live data, allowing them to find discrepancies between design and implementation.

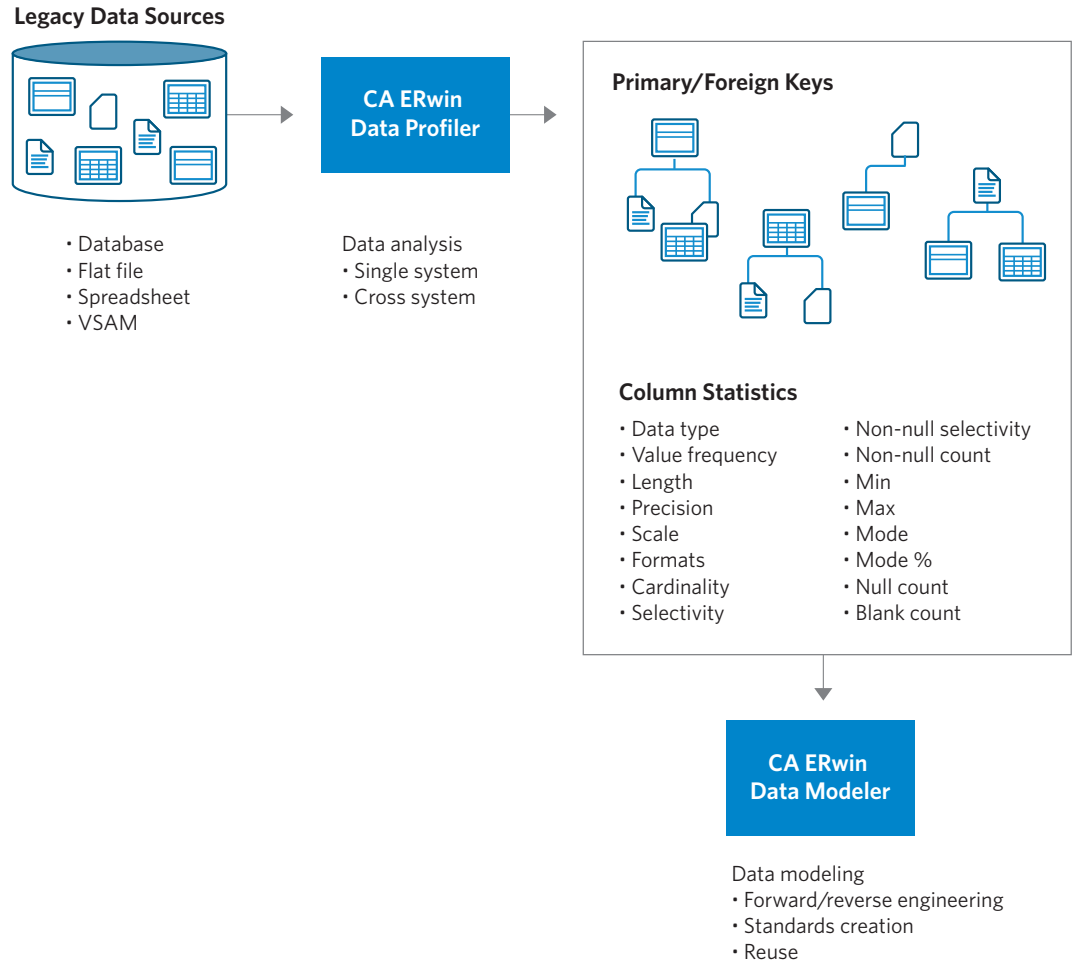
Uncovering Legacy System Structure and Relationships

CA ERwin Data Profiler can infer primary and foreign keys based on data values and helps you to choose the best fit based on your interpretation of the results. Once this structure has been determined, it can be exported to CA ERwin Data Modeler to be leveraged in future enterprise data modeling and architecture efforts.

FIGURE C

Metadata Data from CA ERwin Data Profiler can be leveraged in CA ERwin Data Modeler in related data modeling and architecture projects.

STRUCTURAL INFORMATION INFERRED FROM CA ERWIN DATA PROFILER CAN BE EXPORTED TO CA ERWIN DATA MODELER



SECTION 3: BENEFITS

Competitive Advantage Powered by Timely, Quality Data

A cross-system analysis and data profiling solution, CA ERwin Data Profiler is the newest addition to the CA ERwin Modeling family. This solution delivers robust data profiling management, cross-system overlap analysis and data model reconciliation that offers:

- Cross-system data analysis of up to 20 sources
- Interactive, side-by-side comparison of record values between and within data sources
- Identification of “source of record”
- Column-level profiling and statistics
- Primary and foreign key discovery
- Integration with CA ERwin Data Modeler for synchronization between design and data

With it, you can immediately improve the value of the information that constitutes the lifeline of your organization — *and increase your competitive advantage.*

Specifically, CA ERwin Data Profiler reduces the time required to document cross-reference data sources and increases the accuracy and depth of understanding of your data foundation, thereby:

- Decreasing integration, delivery and maintenance costs
- Improving data quality and accuracy
- Integrating data design with implementation to ensure standards and quality

SECTION 4: CONCLUSIONS

In today’s information-based economy, data-intensive efforts are likely to continue to grow as organizations attempt to gain a robust, consolidated view of their strategic data assets. And critical enterprise initiatives, such as those for MDM, business intelligence, data warehousing and information governance, can benefit from the combination of data profiling, which helps ensure that the data contained in related reports and applications are based, is of high quality, and data modeling, which helps ensure a robust architectural structure for defining, storing and retrieving data assets. CA provides this functionality in a single product family — with the industry-leading combination of CA ERwin Data Profiler and CA ERwin Data Modeler — powerful, best-of-breed tools that help ensure that data design matches reality and that data is of high quality.

For more information on the CA ERwin Modeling family of products in your region, visit: ca.com/contact/rmdm.

CA (NYSE: CA), one of the world's leading independent, enterprise management software companies, unifies and simplifies complex information technology (IT) management across the enterprise for greater business results. With our Enterprise IT Management vision, solutions and expertise, we help customers effectively govern, manage and secure IT.

334520109

Learn more about how CA can help you transform your business at [ca.com](https://www.ca.com)

