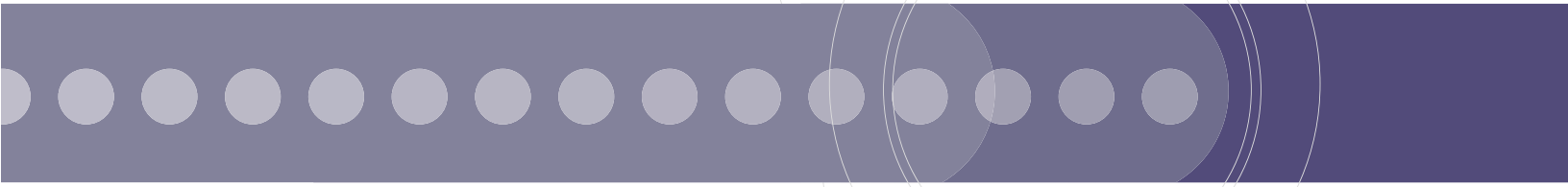
INFORMATICA®

WHITE PAPER

Addressing Data Quality at the Enterprise Level

Six questions your organization should ask to ensure high-quality data



This edition published October 2005

Table of Contents

Executive Summary

Data quality encompasses more than finding and fixing missing or inaccurate data. It means delivering comprehensive, consistent, relevant, and timely data to the business regardless of its application, use, or origin.

Ensuring data quality is a challenge for most organizations—partly because they may not be fully aware of their own data quality levels. Without this information, they cannot know the full business impact of poor or unknown data quality—or how to begin addressing it.

Poor data quality is a costly issue. Findings from a PricewaterhouseCoopers Global Data Management Survey found that 75 percent of those surveyed reported significant problems as a result of defective data, more than 50 percent had incurred extra costs due to the need for internal reconciliations, 33 percent had been forced to delay or scrap new systems, 33 percent had failed to bill or collect receivables, and 20 percent had failed to meet a contractual or service-level agreement.[1] Through 2007, more than 50 percent of data warehouse projects will have limited acceptance, if not outright failure, as a result of lack of attention to data quality issues (0.8 probability).[2]

New compliance drivers are also pushing organizations to scrutinize the quality of their data. Accounting reform legislation such as the Sarbanes-Oxley Act in the United States and Basel II in Europe demands that information be accurate and transparent. High-quality data is essential for companies to meet these new requirements.

The stakes are high. Put simply, to realize the full benefits of investments in enterprise computing systems, your organization must have a detailed understanding of the quality of its data—where quality is poor, how to clean it, and how to keep it clean. By making data quality a strategic priority, your organization better positions itself to streamline operations, grow revenue, keep costs in check, and achieve long-term competitive advantage.

This white paper outlines the questions your organization should be asking to gain confidence in the quality of your enterprise-wide data. After reading this paper, you'll understand:

- What is meant by the term "data quality."
- Why data quality is critical to your business operations.
- The different factors that compromise data quality.
- The various strategies your organization can use for achieving data quality.
- How the Informatica platform can help your organization achieve data quality.

[1]Friedman, Ted, Scott D. Nelson, John Radcliffe. "CRM Demands Data Cleansing." *Gartner Research*, 3 December 2004.
[2]Hostmann, Bill, Frank Buytendijk, Ted Friedman. "Avoid the 'Fatal Flaws' of Business Intelligence and Corporate Performance Management." *Gartner Research*, 9 June 2005.

What is data quality?

Data quality is a broad umbrella term for the accuracy of a particular piece or set of data, and for the way in which data enters and flows through the enterprise. Organizations may not be aware of the full business impact of poor or unknown data quality if they define the term too narrowly.

To evaluate your organization's data quality, consider the following:

- **Existence.** Does your organization own the data?
- **Validity.** Do the data values fall within an acceptable range or domain?
- **Consistency.** Does the same piece of data have the same values even when stored in multiple locations?
- **Timeliness.** Is the data available to the business processes that use it, at the optimal time?
- **Accuracy.** Does the data correctly describe the properties of the object it is meant to model?
- **Relevance.** Does the data support your business objectives?

Ensuring data quality requires more than finding and fixing missing or inaccurate data within the enterprise. It means delivering comprehensive, consistent, relevant, and timely data to the business regardless of its application, use, or origin.

Why is data quality important?

Achieving and maintaining high-quality data is critical to efficient IT and business operations, as well as to the success of strategic business initiatives and to your company's long-term competitive advantage.

Minimizing IT project risk

Data quality is vital for IT organizations, especially when tackling data integration projects. Clean, reliable, current source data minimizes project risks. Addressing data quality as a key requirement from the outset of data integration projects prevents project delays and overruns. It saves costs, too; even something as simple as de-duplicating data can yield significant savings. High-quality data streamlines IT operations and makes the best use of limited resources. Instead of dedicating staff and project cycles to detecting and resolving unanticipated defects and anomalies in source data, re-working code, or maintaining data, your IT organization can focus its energy on delivering projects that drive revenue to your enterprise's bottom line.

Making timely business decisions

Your executives depend on the IT organization to deliver high-quality data they can immediately use to make strategic business decisions. Companies may suffer productivity losses, incur significant costs, or lose hard-earned competitive advantage when the quality of their data is poor or questionable. For example, if executives are working with data that's inaccurate, incomplete, inconsistent, or outdated, their ability to make quick and informed business decisions is compromised. As the old adage goes, time is money. Slow responses to market changes—or miscalculated responses based on poor-quality data—open windows of opportunity to your competitors.

Ensuring regulatory compliance

High-quality data supports operations by helping organizations meet the requirements of the Sarbanes-Oxley Act in the United States and Basel II in Europe, which demand control and accurate reporting of business performance. Reliable, auditable, high-quality data provides your organization the transparency it needs to comply with these regulations. High-quality data also supports your organization behind the scenes, with data matching and compliance applications and justice, intelligence, and anti-fraud systems.

Who says data quality is an issue?

"The Data Warehousing Institute (TDWI) estimates that poor quality customer data costs U.S. businesses a staggering $611 billion a year in postage, printing, and staff overhead."

—Wayne Eckerson
Director of Research
TDWI

Expanding the customer base

The importance of gathering, presenting, and maintaining high-quality data in all your customer interactions—everything from correctly spelling a customer's name to ensuring up-to-the-minute product and pricing data are listed on your Web site—cannot be underestimated. Accurate, current data in customer-facing systems, such as customer relationship management (CRM), call center, and marketing applications helps your organization deliver superior levels of customer service, increase the effectiveness of cross- and up-sell campaigns, and convert more prospects to customers.

"...Projects don't fail for technical reasons, they fail on the issues that humans ought to be able to resolve and don't. Issues like data quality..."

—Mike Whitehorne
Business Intelligence Editor,
ServerManagement
(formerly *Enterprise Server Magazine*)

How is data quality compromised?

The quality of data can be compromised depending on how the data is:

- **Entered.** Data quality suffers when data fields are left blank or filled in incorrectly. Examples of data quality problems arising from data entry errors include:
  - Staff may enter cross-sell response codes in a purchase date field because there is no other place for this particular type of data.
  - Customers may mistype their names or addresses when placing an order using a Web site.
  - Companies may use forms-based applications that provide insufficient field validation at data entry time. Alpha and numeric characters may be mixed, or there may be problems with upper and lower case. For example, if a customer ID number has an upper-case "I" instead of a numeral "1," a company may not be able to reliably track or contact this customer, or may have duplicate entries for the same customer.
  - Customers and staff may enter calendar dates in different formats (i.e., mm/dd/year versus dd/mm/yr) or use incompatible currency units.
- **Maintained.** Each act of data maintenance creates a potential for changes that may have unpredictable consequences.
- **Processed.** When incorrect data enters a system, it may be propagated across multiple systems, thus compromising data quality throughout an organization. Even relatively straightforward data quality errors can mushroom into a complex tangle that undermines organization-wide data quality.
- **Received.** As organizations increasingly outsource business processes to third parties or work with partners and suppliers where data is out of their immediate control, external data of questionable origin or quality might enter an organization and proliferate.
- **Stored**. Storing data in multiple systems often puts data consistency at risk.

How can my organization achieve high quality data?

Your organization can achieve high-quality data. But with systems and applications frequently receiving new data and undergoing incremental changes, ensuring data quality can't be a one-time event. All organizations need to manage data quality in a phased, iterative, ongoing process that includes data quality assessment, planning, and strategy selection and implementation.

Step 1: Data quality assessment

The first step is to assess the current state of data quality to build the business case for a data quality initiative. Profiling the data provides an accurate assessment of the content, quality, and structure of your organization's data, regardless of whether that data is stored in enterprise resource planning (ERP) or CRM systems, legacy and mainframe systems, partner and supplier systems, or in individual spreadsheets, documents, .pdf files, or desktop databases. It's important to involve executive leadership in the data quality assessment and to present what you've found through profiling. The most effective data quality assessments list all the issues, prioritized by maximum impact on the business, to help IT secure sponsorship.

Step 2: Data quality planning

With a clear understanding of the content and quality of your organization's data, the next step is to develop an incremental project plan to correct current errors and prevent future ones. Effective plans address the quality of data in an organization's existing application base and specify ways to ensure that new applications incorporate data quality principles from the start.

Step 3: Data quality strategy selection and implementation

Selecting a strategy to address data quality throughout the enterprise requires balancing the cost of each data quality initiative against its impact. There are two types of strategies: addressing data quality at the point of entry in the enterprise is called an "upstream" approach, while implementing it at an application that consumes data from operational systems (e.g., a data warehouse) is a "downstream" approach.

Upstream strategies

An upstream strategy examines the logic, data, and processes of an existing application for opportunities to address the anomalies uncovered through profiling. This may include changing application logic, adding better form validation, improving the processes associated with data entry, and almost always requires an effort to cleanse the data already in place. It may also necessitate incorporating data quality functions in the application itself, for example, validating data against known examples as it enters the enterprise.

Because this approach solves data quality issues at the source, it benefits every application that draws from that source and every application that will draw from it in future. Selecting and implementing an upstream approach is appropriate if your strategy dictates the highest levels of customization, control, and integration into the operational environment. However, upstream strategies are expensive. They require changes to the application, its logic, and the data. Organizations should be aware that resistance may be encountered when changing existing solutions. For upstream strategies to succeed, projects require clear ownership, consensus, and the support of staff from other teams.

How much does poor data quality cost?

To analyze the cost of bad data, William Weil of Innovative Systems, Inc., completed the following analysis. A given customer list for a company is probably 90 percent accurate. Assume 10 percent of customer details are inaccurate, and of those, 5 percent (i.e., 0.5 percent of the total number of records) have unusable addresses that could have been corrected. The cost of retaining each customer is estimated to range from $100 to $1,000 per customer. In a huge business organization, with around a million customers, 0.5 percent of its customer base (5,000 customers) may be lost unless they are identified correctly in the organization's database. Therefore, that company is likely to lose 5,000 customers at a direct cost of $500,000 to $5,000,000.

"Data quality must be viewed as a business issue and approached in a structured manner. Data quality initiatives are multifaceted, and there is no single key to success. However, a methodology for executing a data quality initiative provides a plan that can increase the chance of achieving benefits. Organizations must seek out data quality methodologies that are best-suited to their culture, style, type of data quality problems and goals. Complete methodologies will address not only technology issues, but people and process considerations as well."

—Ted Friedman
"Data Quality Methodologies:
Blueprints for Data Quality Success"
Gartner Inc., July 2005

Downstream strategies

Downstream strategies resolve data quality issues at the target application or warehouse, not at the data source. Since data can be modified in flight, it can be improved in the process, typically through data transformation, name and address cleansing, and lookup validation. A downstream strategy benefits only consumers of the data who are working with the target applications or data warehouses, but it is considerably easier and less expensive to implement than an upstream strategy.

Choosing the right strategy

Regardless of whether your organization chooses to adopt upstream or downstream strategies—or a few of each type—you should make sure:

- The scope of your organization's data quality program is carefully controlled.
- Executive buy-in has been secured.
- Processes for measuring current and future ROI have been established.

In deciding to adopt an upstream or downstream strategy, your organization should measure the cost of changing upstream applications and their data and compare that to the ad-hoc costs of managing data quality downstream as data is moved for other purposes.

In either case, continual monitoring and progress reporting are essential to ensuring data quality and building trust in the data. If users decide that data is defective, your organization will have to work long and hard to change that perception. Publishing progress reports and data scores from a monitoring system helps build (or re-build) users' confidence.

Because data quality is vital to achieving maximum results from your enterprise data integration implementations, consider investing in proven technology to help analyze and address data quality, both upstream and downstream.

How can the Informatica enterprise data integration platform help me address data quality?

Informatica offers an enterprise data integration platform with data quality capabilities that allow your organization to implement a comprehensive, full lifecycle approach to addressing data quality, both upstream and downstream. This enterprise data integration platform is called Informatica PowerCenter®.

Companies of all sizes rely on PowerCenter to access, discover, and integrate data from virtually any business system, in any format, and deliver that data throughout the enterprise at any speed. PowerCenter features robust data quality capabilities in a single integrated environment.

PowerCenter's data quality capabilities allow organizations to standardize, validate, and correct any data source. Doing so maximizes the integrity and value of your organization's most important information assets and provides users with accurate, business-relevant information.

Unlike standalone data quality tools, PowerCenter features robust connectivity, metadata management, parallel performance, and linear scalability—capabilities that deliver the infrastructure and processing power necessary to meet enterprise data quality requirements.

The platform also safeguards process quality by ensuring the use of standard definitions, providing development standards, and enforcing access security. It supports operational quality through capabilities for audit and control, system management, and error handling.

Through comprehensive and powerful data quality capabilities, PowerCenter is able to support a large number of data quality projects, speed the overall project lifecycle, and help organizations have confidence that the data they use to make strategic business decisions is accurate, current, and consistent. PowerCenter is the ideal platform for addressing data quality at the enterprise level.

Informatica PowerCenter ensures data quality through five core capabilities:

1. **Data profiling.** Assesses the current and ongoing quality of data
2. **Data transformation and correction.** Parses, transforms, and corrects existing data
3. **Data cleansing and enhancement.** Cleanses and/or enhances data with additional fields and information such as postal codes and geographic data
4. **Data matching and consolidation.** Matches similar records and performs de-duplication and consolidation based on set criteria
5. **Data quality process management.** Repeats the data quality process in an iterative fashion to measure ongoing effort

Data profiling

Informatica PowerCenter's data profiling capabilities support effective data integration by providing comprehensive, accurate information about the content, quality, and structure of data in virtually any operational system. Leveraging the platform's unified data integration architecture and codeless development environment, your development teams can discover and understand source data and easily create and leverage historical profiling performance metrics. PowerCenter presents data profiling information in easy-to-understand reports and charts, with metrics that quantify the impact over time of quality initiatives. These make it easy for your business analysts to understand the data quality issues.

The data profiling capabilities in PowerCenter help your IT group deliver projects to specification and on time, because it enables them to scope out data quality issues faster and with much greater accuracy than traditional methods of writing SQL queries. Those other methods of profiling data simply sample data in a few key fields to get a sense of what the data is like in those columns, but the results can be—and usually are—wildly inaccurate. Unlike hit-or-miss user-generated SQL queries, PowerCenter's data profiling capabilities scan every single record in every single column and table in a source system. They generate reports that make it easy to understand the state of the data, clearly exposing all new or unanticipated structures and values.

Because data profiling provides better information from the outset, it significantly streamlines subsequent development. PowerCenter is unique because it provides a single environment for both profiling and addressing quality issues, reducing the number of exploratory development iterations and accelerating development of transforms and mappings. IT departments can be more productive since they use the same user interface for the entire process.

As an integrated capability, data profiling leverages the existing PowerCenter infrastructure. No additional infrastructure set-up is required. Your IT department can profile any source that PowerCenter can reach. It ensures maximum performance regardless of sampling algorithm. Its proactive error identification and handling provides a detailed assessment of the severity of any quality issues.

The data profiling capabilities of PowerCenter promote reusability. PowerCenter automatically generates data profile mappings, and IT can reuse these to continually measure quality improvement and to produce components that can be reused for development of the data correction process.

The Role of Data Quality in Data Integration Projects

Traditionally, data quality has been addressed as part of data warehousing projects, because moving data out of operational systems into the warehouse offers an opportunity to refine the data as it is moved. As the data integration technology that powers this process has increasingly been adopted to support other initiatives-such data migration, data consolidation, data synchronization, master data management, and outsourcing-these initiatives now offer opportunities to improve data quality.

Additionally, as the concept of an integration competency center (ICC) has evolved into a prevalent, successful IT model, data quality has become a core service provided by an ICC team. An ICC leverages integration technology, expertise, processes, and best practices to utilize IT resources for business benefit.

A good example of the extension of data integration technologies to data quality initiatives is a data migration and consolidation project for the Israeli Defense Force (IDF), which used the Informatica PowerCenter® data integration platform instead of a hand-coded solution. When the IDF embarked on its ambitious initiative to migrate data from mainframe and legacy applications to SAP R/3, it wanted to address data quality issues throughout the migration process and ensure the new system was populated with high-quality data. The IDF relied on PowerCenter to profile, access, transform, cleanse, and deliver data in one smooth operation. The initial assessment exposed an alarming quantity of issues with source-level quality that would otherwise have gone undetected.

Thanks to PowerCenter, in the whole process of cleansing and transforming millions of records there were only 350 data record errors that went undetected in the final product data load. IDF also used PowerCenter to de-duplicate its data, reducing the number of parts tracked in SAP by 25 percent and cutting overall data volume by 20 percent, saving on storage requirements.

"Data standardization, cleansing, matching, profiling, and other functions that comprise data quality are vital to customers' ability maximum results from their data integration implementations. The Informatica ... PowerCenter platform [gives] their customers proven technology."

—Eric Rogge
Vice President and Research Director
Ventana Research

PowerCenter offers a broad array of data profiling functionality, including

- Column level
  - Range validation, min/max, average count
  - Domain validation: for example, pre-determined set of values
- Source level
  - Row count
  - Redundancy evaluation
- Intersource level
  - Outer join analysis
  - Cardinality analysis
- Integrated reporting and analysis
  - All results captured in a "profiling warehouse" or schema
  - Bundled reports for rapid analysis
  - Custom reports for "dashboard" views for business justification

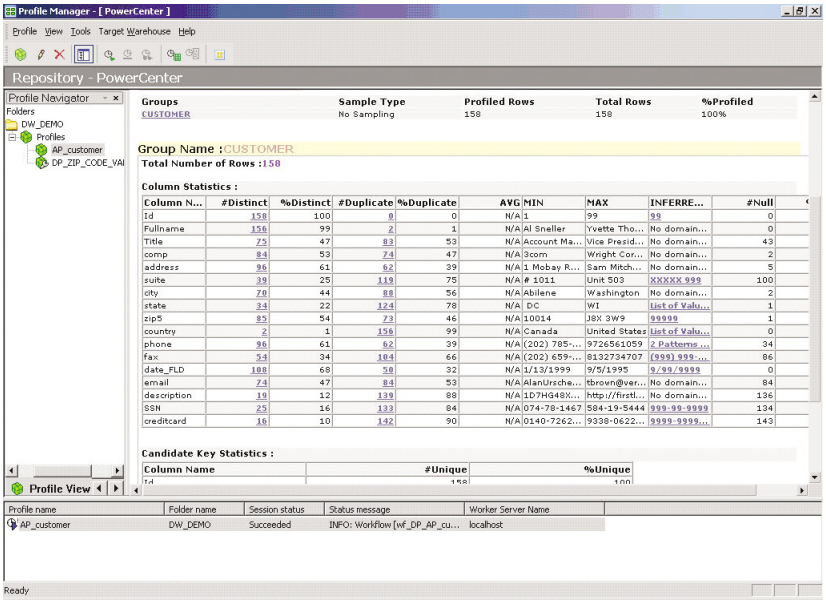


Figure 1: Example of PowerCenter data profiling assessment report

Data transformation and correction

Informatica PowerCenter was originally developed to address the data transformation and conversion process associated with data warehousing. That core capability resulted in the development of a robust function library capable of handling the majority of data-related manipulations required for data quality initiatives. This core capability has continually evolved to support other data integration initiatives including data migration, consolidation, master data management, and outsourcing.

PowerCenter intelligently parses and identifies individual components of data. Parsing isolates data and applies structure to each data element, speeding the data integration transformation processes and reducing the risk of error. PowerCenter identifies data elements—including unstructured data—and separates them into individual fields.

To transform data, PowerCenter converts it into a common internal format and structure to achieve a consistent output. It can rank and sort data, converting text strings based on phonetics into encoded values to use for comparison as necessary. Depending on the data source or target, PowerCenter may use any of the following to obtain the desired result:

- Pivoting arrays for normalization
- Filters for subset processing
- Custom and stored procedures
- Lookups to augment data
- Joiners for disparate tables and files
- Sequence generator for surrogate keys
- Aggregators for sums and averages
- Routers for branching and case logic
- Expressions to translate data (for example, concatenate, instring, rpad, ltrim)
- Source qualifiers for specific data selection

PowerCenter's wizard-driven business rules and robust transformation library reduce the development effort necessary to transform and correct data, thereby increasing developer productivity and reducing training requirements. Developers using PowerCenter can take advantage of all functions from a common toolbar to develop the business rules to perform this critical function.

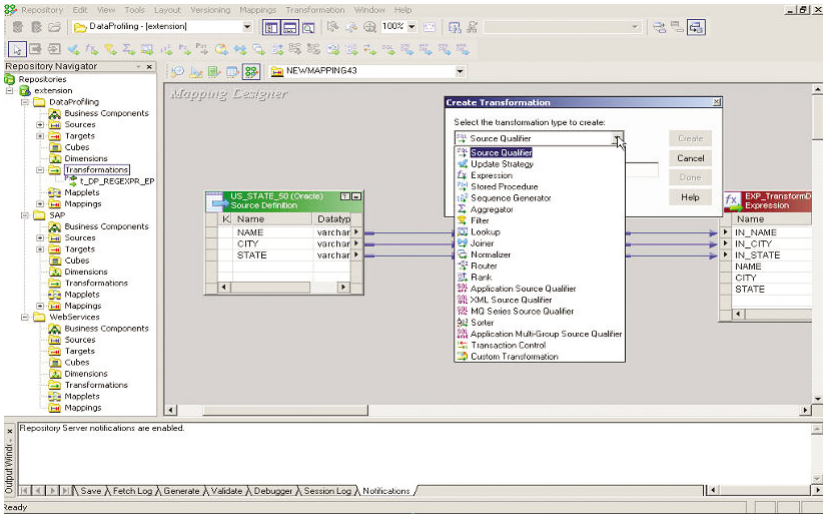


Figure 2: Example of data transformation library in PowerCenter

Data cleansing and enhancement

PowerCenter's data cleansing and enhancement capabilities enable you to cleanse, standardize, and enhance data through rich parsing and address standardization and geographic name and address libraries. These data cleansing capabilities provide users greater confidence in their analytical systems by creating an accurate "single version of truth." Embedded within the PowerCenter environment, they leverage the connectivity, metadata, parallel performance, and linear scalability of the PowerCenter platform, enabling you to reconcile data more quickly.

PowerCenter's data cleansing and enhancement capabilities allow users to:

- **Parse fields by patterns.** PowerCenter includes robust features for parsing data elements, including unstructured data, by identifying data elements in customer files and separating them into individual fields. For example, parsing converts a field containing "John Taylor, 123 Anywhere Street, Hometown, IL" into five separate fields with "John," "Taylor," "123 Anywhere Street," "Hometown," and "IL." Parsing isolates data and applies structure to each data element, helping to speed data integration transformation processes.
- **Standardize format and structure of output record.** PowerCenter allows users to convert data into different format, structure, and content to gain desired output.
- **Enhance/correct names and addresses through geo and postal code libraries.** Address standardization based on the most current postal authority regulations compares and corrects address information against postal service directories from more than 195 countries. PowerCenter leverages an unrivaled repository of postal knowledge embodying global logistical data structures and content, ensuring greater quality and accuracy.

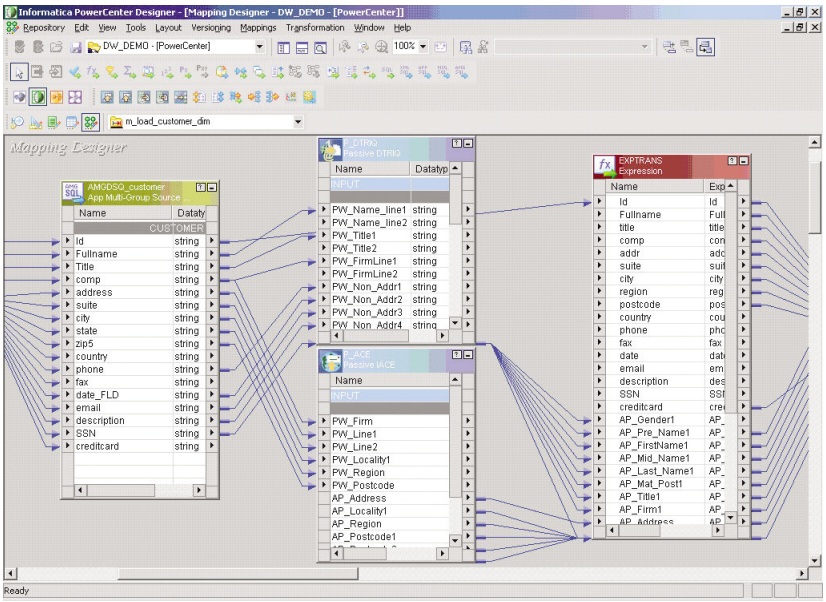


Figure 3: Example of integrated data parsing and cleansing in PowerCenter

Data matching and consolidation

Once data has been transformed, cleansed, and enhanced, it needs to undergo de-duplication and consolidation to produce the final clean representation. One of the major problems associated with poor data quality is the duplication of data across similar records. This complicates efforts made during data migration projects to get to a “single version of truth” for a new application.

PowerCenter provides matching and consolidation capabilities that help your team identify and remove data duplication. PowerCenter searches existing records for similar records using specific business-defined criteria. Match standards and business rules eliminate any doubt as to whether two records refer to the same individual customer or household.

Identifying all related records is not enough to create a single, consistent view. With its consolidation capabilities, PowerCenter uses the data found during matching to combine all similar data into a single consolidated view. Used by the Israeli Defense Fund to migrate data from mainframe and legacy systems to SAP R/3 instances in multiple locations across Israel, the cleansing and deduplication capabilities of PowerCenter reduced data volumes by roughly 20 percent, saving on storage requirements.

Key features of the data matching and consolidation capabilities of PowerCenter provides powerful “fuzzy logic” and other matching functions to identify similar or duplicate records. These functions simplify the need to write custom logic to perform this step in the process.

PowerCenter also provides flexible, tunable business rules, driven through its graphical user interface. These rules allow your developers to determine the best course of action when PowerCenter identifies duplicate records.

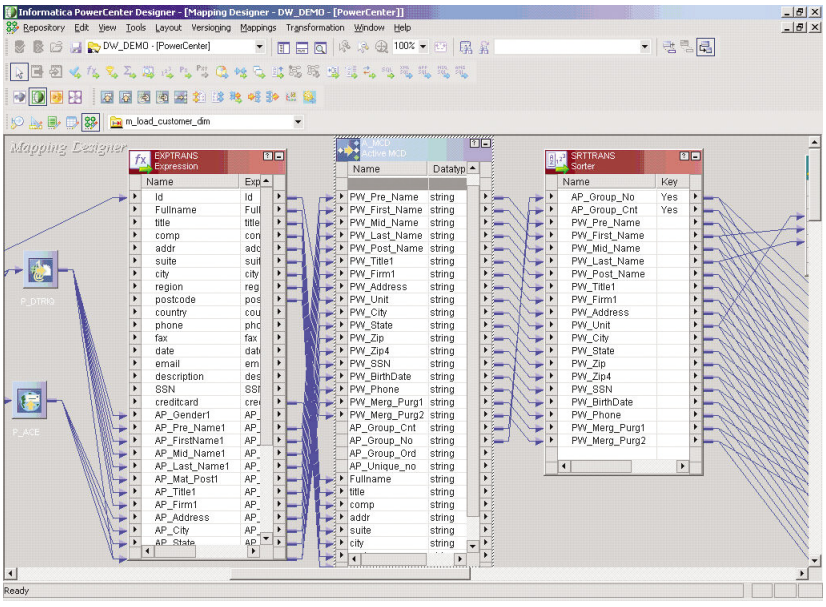


Figure 4: Example of integrated data matching and consolidation in PowerCenter

Data quality process management and control

Implementing a repeatable, consistent, secure process for addressing data quality is one of the biggest challenges of a new data quality initiative. Fortunately, PowerCenter provides a complete environment for creating, managing, and auditing the full data quality lifecycle from the data processing point of view. Using a single solution to address data quality as an end-to-end process saves your organization development and maintenance time and reduces initial and ongoing costs.

Key features of PowerCenter's data quality process management and control capabilities include:

- **Scheduling and workflow** features that execute and control the technical data quality process. Because data quality is an ongoing process, PowerCenter captures runtime statistics, provides dashboards and reports, and allows for automatic recovery/restart if anything goes wrong.
- **Security features** that ensure correct data access privileges. PowerCenter enables encryption of data in stream and use of authentication infrastructure such as LDAP directories.
- **Data lineage and audit features** to address compliance requirements. PowerCenter provides key data lineage and change analysis capabilities that allow IT to show where data comes from and the business rules applied during data quality processing. These capabilities ensure transparency and compliance with efforts to improve data quality on an ongoing basis.

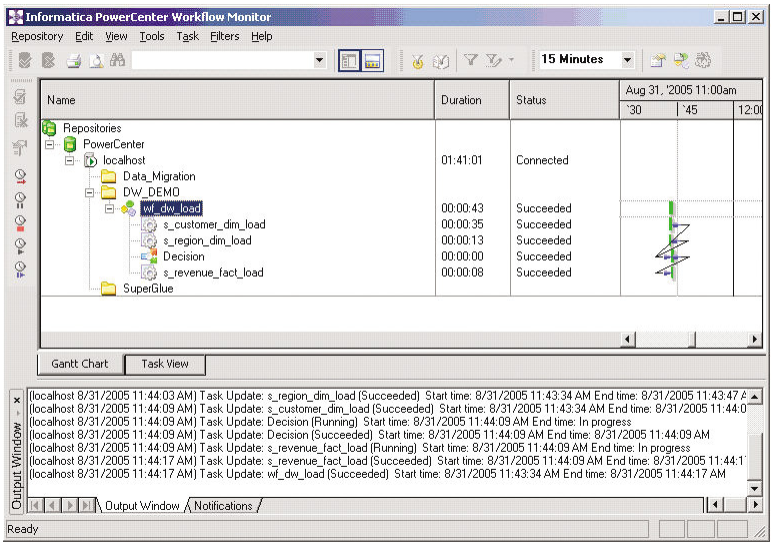


Figure 5: Example of PowerCenter workflow monitor

Where can my business go for more information?

Data quality is a complex set of issues that organizations must address in order to keep customers satisfied and compete effectively. Assessing your organization's data quality issues and then adopting a proactive, automated, and multi-phased data quality strategy built on Informatica PowerCenter and its capabilities is the fastest and most cost-effective way to resolve current and ongoing data quality issues. This approach will help ensure that your organization benefits from accurate, valid, consistent, relevant, and timely data to support your current and future business needs.

To find out more, please visit us at www.informatica.com or call us at (800) 653-3871.

INFORMATICA®

Worldwide Headquarters, 100 Cardinal Way, Redwood City, CA 94063, USA
phone: 650.385.5000 fax: 650.385.5500 toll-free in the US: 1.800.653.3871 www.informatica.com

Informatica Offices Around The Globe: Australia • Belgium • Canada • China • France • Germany • Japan • Korea • the Netherlands • Singapore • Switzerland • United Kingdom • USA

J50612 6631 (10/14/2005)