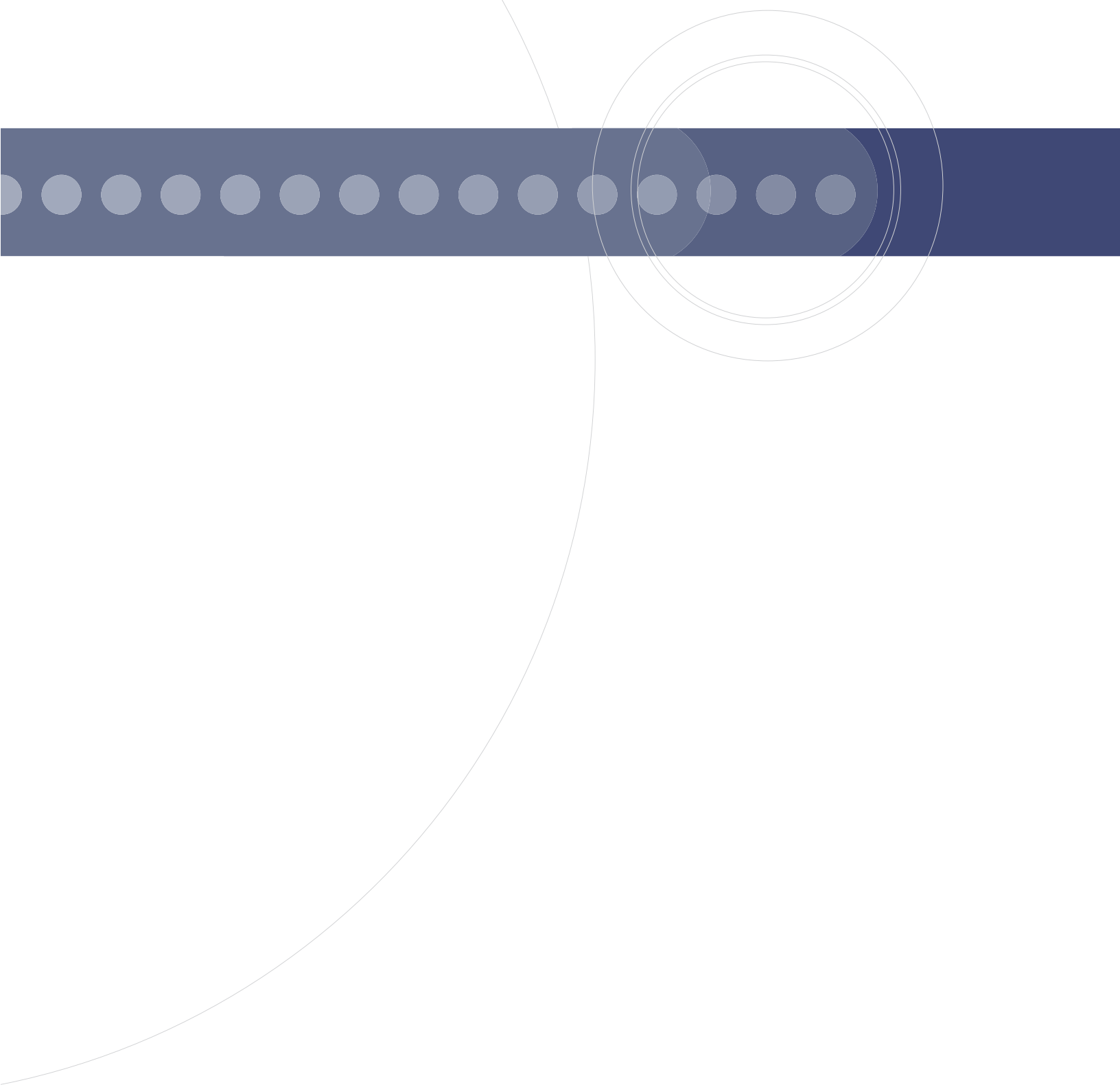


# Data Quality

*Driving Single View of Customer*





This document contains Confidential, Proprietary and Trade Secret Information ("Confidential Information") of Informatica Corporation and may not be copied, distributed, duplicated, or otherwise reproduced in any manner without the prior written consent of Informatica.

While every attempt has been made to ensure that the information in this document is accurate and complete, some typographical errors or technical inaccuracies may exist. Informatica does not accept responsibility for any kind of loss resulting from the use of information contained in this document. The information contained in this document is subject to change without notice.

The incorporation of the product attributes discussed in these materials into any release or upgrade of any Informatica software product—as well as the timing of any such release or upgrade—is at the sole discretion of Informatica.

Protected by one or more of the following U.S. Patents: 6,032,158; 5,794,246; 6,014,670; 6,339,775; 6,044,374; 6,208,990; 6,208,990; 6,850,947; 6,895,471; or by the following pending U.S. Patents: 09/644,280; 10/966,046; 10/727,700.

This edition published April 2006

## Table of Contents

<b>Purpose Of This White Paper</b> .....	<b>.2</b>
<b>Introduction</b> .....	<b>.2</b>
<b>Tools For The Job</b> .....	<b>.4</b>
<b>The Data Quality Process</b> .....	<b>.4</b>
Data Quality Profiling .....	.4
Standardization .....	.5
Matching .....	.5
Consolidation .....	.8
<b>Conclusion</b> .....	<b>.9</b>
<b>Appendix - Informatica Data Quality Product Suite</b> .....	<b>.10</b>



## Purpose Of This White Paper

The purpose of this white paper is to outline the importance of data quality with reference to single view of customer (SVC). In any organization SVC is the foundation of successful customer relationship management (CRM) across financial institutions, utility companies, telecommunications service providers and retail outfits. The paper defines the breath of the issue and how poor data quality can hamper, delay and even defeat organizations in their attempts to implement CRM and SVC. It describes a typical solution to the problem of poor quality customer data based on Informatica Data Quality enterprise data quality management software. The white paper goes through the steps that should be taken to ensure data quality issues are eradicated. It is the opinion of the author that SVC can only be achieved with reference to the state of the underlying data quality. The best way to achieve the necessary levels of data integrity, conformity and consistency is through the implementation of an end-to-end data quality process.

## Introduction

Single View of Customer (SVC) is the lynchpin of effective customer relationship management (CRM). With reliable SVC a CRM system can really deliver on its promise enabling your business to profit from understanding and anticipating the needs of current and potential customers. Yet even in organizations that have invested significantly in CRM initiatives, high quality and reliable SVC often remains an elusive goal. The root cause of this is data quality. Low quality customer identification and description data is pervasive in most organizations and is the major stumbling block in the way of achieving SVC.

Because CRM and SVC are so central to an organization's ability to do business, everyone who uses the system, from sales force and call center personnel to executive-level professionals and marketing teams, must be able rely on the quality of the data it contains. Without confidence in the currency and quality of CRM information these systems fall into disuse and ultimately fail. Successful organizations understand the direct link between data quality and business performance, and that information-intensive applications such as CRM can only achieve results if the data they depend on is reliable, complete and accurate. Analyst firm Gartner says data quality has emerged as the number one cause of CRM project failure.

The longer the problem is left to fester the worse it will get. According to PricewaterhouseCoopers customer data degrades at a rate of two percent a month, or nearly 30 percent per annum. Proactive efforts to identify and overcome data quality issues are vital for SVC and CRM applications to succeed. SVC implementation should start by identifying data quality problems, such as missing, non-standard or inconsistent data, and correcting such problems.

But achieving the high levels of data quality needed for SVC and CRM to succeed is more than a one-off exercise. The only way to ensure that accurate, consistent and timely SVC data is delivered into the future is through an ongoing data quality management program. Data quality levels need to be tracked and assessed against targets on a regular basis. New data should be cleansed and checked against business rules as it goes into the system to ensure the highest possible levels of data quality are maintained.

An end-to-end data quality management process such as that outlined in this white paper will enable any organization to rapidly identify and correct data problems and to prevent the erosion of data quality levels over time.

The four stages necessary to achieve this include Profiling, Standardization, Matching and Consolidation.

- **Profiling** Investigates the content of key data fields to Identify and quantify data quality problems
- **Standardization** Extracts, cleanses and standardizes key customer identification and description data
- **Matching** Matches selected customer identification and description data from each customer record to identify similar or related entities
- **Consolidation** Generates a final master table of unique customer records and key tables linking master data with source systems and transaction tables



## Tools For The Job

On large datasets it would be impossible to carry out all stages of a data quality management process manually. A powerful and flexible tool capable of implementing customizable business rules across any type of data is essential to carry out the task. Your data quality tool should make it possible to perform ongoing data cleansing in real time as new records are entered, evaluating whether the data is new or a modification of an existing record. Even with real time intervention companies will still need to run data quality checks in batches to ensure duplicates aren't introduced when a new mailing list is incorporated into an existing database for example. Checks should also be performed at regular intervals to minimize data degradation.

In the recent past data quality solutions tended to rely on separate products for the different stages of data quality analysis (profiling & matching) and enhancement (standardization & consolidation). Today offerings such as Informatica Data Quality provide a complete solution incorporating a wide range of data quality functionality, integration and reporting capabilities. Improved user interface and graphical capabilities has empowered the business user to control the entire data quality process, which leads to speedier and more effective resolution for all data quality issues. The goal is to deliver best of breed technology to implement a data quality strategy, regardless of the complexity of the underlying data.

## The Data Quality Process

To be really effective a data quality process needs to be easy to understand and should be repeated in a way that leads to a virtuous circle of ever increasing data quality improvement. The result of each round of profiling, standardization, matching and consolidation should include comprehensive reporting to monitor progress and provide a framework for ongoing data quality enhancement.

### Data Quality Profiling

As the starting point in the data quality process profiling delivers a complete investigation of the content of each of an organization's critical customer data fields. The goal is to identify problems that could prevent effective matching of the data. Data quality profiling effectively enables an organization to answer the following questions about its key customer data:

- What data fields are suitable for use in the matching processes?
  - Which fields have sufficient levels of completeness? e.g. if the GENDER field is only populated in 10 percent of cases then it will not be an effective matching field
  - How many fields contain valid and consistent values? e.g. the DATEOFBIRTH field may have 20 percent of values set to default 01/01/1901
- What standardization/cleansing is required for each data field prior to the matching process?
- What matching rules are likely to be effective? e.g. partially incomplete or invalid fields can be used in the matching process, but rules must be formulated to ensure that they are only used when a valid output is present.

Profiling involves applying specialized algorithms to investigate the content of different field types:

- Text profiling for name, address, email and other free text fields
- Character profiling for date, telephone and other code fields
- Number ranging for numeric fields
- Frequency counts for defined value fields, such as gender, title etc.

During the profiling phase validity checks are carried out on actual values or data patterns using lookup tables and rules wizards. The output of the profiling phase of any comprehensive data quality process will be a series of reports identifying data quality issues for each selected attribute. Informatica Data Quality's powerful and easy to read reports offer drill-down functionality allowing users to view the underlying data associated with each data quality problem. At this stage Informatica Data Quality also generates metadata that enables easy parsing and standardization of data prior to matching.

## Standardization

The goal of the standardization phase of a data quality management process is to remove or flag problems prior to moving on to the matching phase. In the standardization phase each of the key data fields is passed through a series of user defined rules to remove inconsistencies and unconformities identified during the data profiling stage. The output is a range of new data fields containing standardized and enhanced data.

During standardization the following tasks are carried out:

- Noise removal e.g. comments in free text fields such as "Incorrect Address" will inhibit the matching process if not recognized but can be extremely useful if recognized and extracted
- Parsing of data e.g. name, address, descriptions
- Standardization of terms e.g. use of dictionaries to correct common misspellings etc., or use of transforms to standardize date or telephone number formats
- Derivation of missing data values e.g. derivation of gender from firstname or title
- Generation of data quality flags for use in matching rules e.g. records with multiple incomplete or invalid values can be flagged.

The new data fields created during the standardization phase may be used solely for the matching process. They may also be written back to the original source file to replace the original low quality data fields.

Developing standardization rules using Informatica Data Quality is fast and easy to understand using point and click components that leverage the data quality metadata generated during the profiling phase. It is highly flexible users can define any number of output data fields and can build multiple layers of data cleansing rules for each field. For example an input "customer name" field containing non-standard business and customer names may be passed through the standardization module to generate the following new data fields - Customer type (Business or Person), Company Name, Title, Firstname, Middlename, Surname, Gender, Postnomial. This is achieved by a combination of dictionary lookups, positional parsing and business rules. The dictionary manager within Informatica Data Quality enables easy lookup of user created dictionaries within the product or third party tables stored in other systems.

## Matching

Effective record matching within a dataset or across multiple datasets is a highly complex task requiring sophisticated algorithms and iterative tuning of the process for each individual dataset. But cracking the matching problem is at the core of achieving and maintaining SVC. In choosing the tools to use at this phase of the data quality management process flexibility and usability are essential attributes.

Every dataset is different and every organization requires different business rules for classification of duplicate customers. Data characteristics and business rules change continuously and so must matching rules. The key to success is to empower users who know and understand the underlying data to control matching rules from definition to configuration, testing and implementation.

Informatica Data Quality is designed for use by in-house teams of data and business analysts, enabling users to quickly and easily develop customized rule sets for automated data quality reporting, validation and correction.

Flexibility - the product's powerful match scoring algorithms can be applied to any range or type of data fields. Matching rules can be tailored and tuned to overcome the specific data quality problems within your dataset and to utilize all the available customer identification and description fields.

Usability - the product's user interface allows business analysts/data analysts who know the data and understand the business needs to configure and test matching rule sets. No knowledge of coding or scripting is required. This means that matching rule sets can be configured quickly and easily and modified over time to deliver optimum results.

To deliver matching that is both accurate and high performance, Informatica has devised a two-step matching process that involves pre-grouping of records by generation of complex grouping keys, followed by matching of grouped records to identify duplicate or related customers.

### Matching Stage 1: Pre-Grouping

The most accurate way to identify duplicate or related records is to compare a large number of attributes from each customer record with all other records in the dataset. However the processing time to carry out this process grows exponentially as the dataset grows. As such this process becomes extremely inefficient for large datasets.

Informatica Data Quality's pre-grouping process significantly reduces the number of computationally expensive record-pair comparisons. It does this by labeling records so that only those that have some basic similarities will be compared. As such it massively improves matching performance while maintaining matching accuracy.

The pre-grouping process works by generating a set of keys based on features within each customer record. Only record pairs with one or more similar or identical keys will be grouped for Stage 2 of the matching process.

For example: take the following four records:

REC	Last	First	Address 1	Address 2	Address 3	Address 4	Date of Birth	Phone
1	Smith	John Paul	14 Trebovir Road		London	Sw5 9LY	20/04/69	020 7-8643567
2	Paul	Smithe	10 Trebovir Road		Sw5		01/01/69	864 3567
3	Jim	Smeath	25 Tra tham Park	London			02/04/1969	020 7-6754497
4	Murphy	James	10 Huntley Park	Fulham	London	Sw5 8RY		

The pre-grouping process will generate keys that will group the first three records for matching. All of these records have significantly similar features (Smith, Smithe and Small have phonetic similarities as do Trebovir and Tratham) while record four will not be grouped with the others because its features are not similar.

The pre-grouping process works by stripping away unnecessary information and applying phonetic and character extraction transforms to generate multiple keys describing the most significant features in the data. The use of multiple keys means that even records with one or more different significant features (e.g. name/address) can still be grouped for matching provided some significant features are the same. The use of phonetic and other transforms means that significant differences (e.g. misspellings, abbreviations, non-standard formats) within the significant features will be overcome in the key generation.

## **Matching Stage 2: Matching**

Stage 2 of the matching process compares pre-grouped records using a range of powerful algorithms and user defined matching rules. It identifies pairs or clusters of records that relate to similar or associated entities.

The Informatica Data Quality matching module provides: A selection of powerful, best practice matching algorithms each one optimized for a particular field type, such as:

- Customer and company names
- Mixed and non-standard address fields
- Post codes/telephone numbers
- Long free text fields
- Numbers/dates
- A rules wizard enabling users to define unlimited matching conditions and match over-ride rules
- Match scoring systems allowing all matches to be graded based on confidence and different consolidation processes to be applied depending on the grade
- Multiple output options including HTML reports, match pair listings and match key tables
- Full configurability: Within minutes users can build and run a matching process using data attributes and matching rules selected or defined by the user. By iteratively running the matching process against different samples of data, the optimum matching rules and weightings for that dataset are defined.



## Consolidation

Informatica Data Quality's consolidation functionality provides the capability to integrate the matching outputs with transactional and other operational systems to deliver an effective, ongoing single customer view. An effective consolidation process must provide the following elements:

- A high quality customer reference dataset containing only unique customer records created by taking the highest quality information from each of the clusters identified during the matching process
- A set of key tables (relationship tables) to maintain the relationship between:
  - Old customer identifiers in source systems and the new customer key in the reference dataset
  - Related records (households, subsidiaries, employees) etc within the reference dataset and other systems
- Processes for updating the reference files and all of the key tables when new customer records are added/deleted to/from the system
- Dynamic linkages between the reference table and key tables and the transaction systems/analytical systems which utilize the single customer view.
- Manual over-ride process for rejecting/deleting relationships created which have been identified to be incorrect

The setup of each of the above elements is typically highly specific to each individual organization as they depend on the source systems, target processes, internal business rules etc. As a result significant custom development is typically required to put these elements in place.

Informatica Data Quality is designed to be easily configured. It enables users to build highly customized consolidation processes to meet their specific technical and business requirements without having to resort to coding or scripting.

The software includes a range of pre-built components that contain the basic functionality required to deliver each of the above elements. At the same time all components are sufficiently configurable to enable the structures and processes to be customized to fit the needs of each organization. Informatica Data Quality Designer allows users to configure these components and embed the required business rules and connections by dragging and dropping components onto a desktop data quality plan.

In Designer the configured components are stored as graphical "plans" for deployment as batch or real time processes.

Some of the capabilities of these components include:

- Ability to build, populate and update customized key tables within the internal Informatica Data Quality database repository
- Ability to build, populate and update customized key tables in external databases (e.g. Oracle, SQL Server, DB2)
- Lookup and replace component to remove or replace old customer keys in multiple source tables based on dynamic data feeds from matching outputs
- Ability to maintain dynamic linkages across multiple systems to ensure any updates to customer keys are immediately fed through to all related systems
- Rules-wizard to specify rules for merging of records
- Consolidation Viewer tool for manually reviewing and updating merged records and match clusters

## Conclusion

Data quality is emerging as one of the most significant challenges for companies implementing SVC or CRM. Getting the data right from the start can save the headache of having to explain why a CRM project has failed. To be successful, any SVC project must have a data quality and enhancement program at its core. Informatica provides a revolutionary approach to low quality data discovery and cleansing that enables business users to maintain control of the process from planning to implementation.

Informatica Data Quality software is a powerful product suite providing the complete range of functionality required to resolve and overcome the data quality problems that inhibit SVC. It provides a complete solution both in the initial phases of implementing a single customer view and for the ongoing management of new data entering the system.

Informatica Data Quality is designed to combine the power and flexibility to build highly customized data quality rule sets, with the ease of use that means these rule sets can be built without the need for complex scripting or low level coding. The software works across all master data types such as materials data, customer data, asset codes and vendor data, to solve a wide range of data quality problems. With powerful reporting tools, the product enables users to measure and improve data quality on an ongoing basis. The Informatica data quality management process and its Designer, Runtime and RealtimeSDK components ensure continued cleansing and transformation of data shared by multiple applications, enabling organizations to gain real competitive advantage from their data.



## Appendix - Informatica Data Quality Product

### Informatica Data Quality

The latest version of Informatica Data Quality is a significant advance. It introduces a new architecture that supports enterprise-wide data quality. The product consists of four integrated components: Designer, Server, Runtime and RealtimeSDK; and a comprehensive library of reference data and content. Each component provides all the key functions necessary for implementing business-focused data quality rules and processes – data analysis, cleansing/standardization, matching, consolidation/survivorship and reporting & monitoring. All Informatica Data Quality components share an integrated repository that enables management, control and reuse of data quality rules and reference data.

### Key Features of Informatica Data Quality

#### Designer

Designer is the central component of Informatica Data Quality. Based around a highly intuitive drag and drop user interface, Designer enables non-technical users to easily build complex, customized data quality rules and plans. The data quality plans drive automated data quality analysis, reporting, cleansing and transformation through Informatica Data Quality's Server, Runtime or RealtimeSDK. Designer is also available as a standalone desktop product that can be used for one-off data-cleansing projects.

#### Server

Server is a scalable server for developing, testing, deploying and managing Informatica's data quality analysis and enhancement processes across the enterprise. It enables remote execution of data quality plans and processes leveraging hardware resources across the network to deal with large data sets and complex data quality requirements. The server enables organizations to re-use data quality business rules, storing and deploying the same rules in all environments. With it, data quality solutions can be deployed across a number of platforms including Windows and Unix, and run interactively, as scheduled batch jobs or in real-time.

#### Runtime

Runtime is a high performance data processing engine designed for easy integration at one or multiple points within a company's information systems. Runtime enables companies to deploy customized data quality analysis and enhancement processes as a scheduled or automated batch process.

## **RealtimeSDK**

RealtimeSDK is a set of APIs used to deploy data quality plans, built in Designer for real-time validation and correction of data. RealtimeSDK is designed to prevent low quality data entering a database by identifying and correcting data quality problems in data capture or data acquisition processes.

### **Technical Requirements:**

Informatica Data Quality can accept data from multiple sources including, delimited or fixed width files or directly from Oracle, DB2, SQL Server, SAP or any ODBC compatible database.

### **Operating Environments:**

Informatica Data Quality Designer runs on Windows NT, 2000 and XP.

Informatica Data Quality Server, Runtime and RealtimeSDK run on Solaris, AIX, HP-UX, Linux, Windows NT, 2000 and XP.







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](http://www.informatica.com)

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

© 2005 Informatica Corporation. All rights reserved. Printed in the U.S.A. Informatica, the Informatica logo, and, PowerCenter are trademarks or registered trademarks of Informatica Corporation in the United States and in jurisdictions throughout the world. All other company and product names may be tradenames or trademarks of their respective owners.

J50850 (04/11/2006)