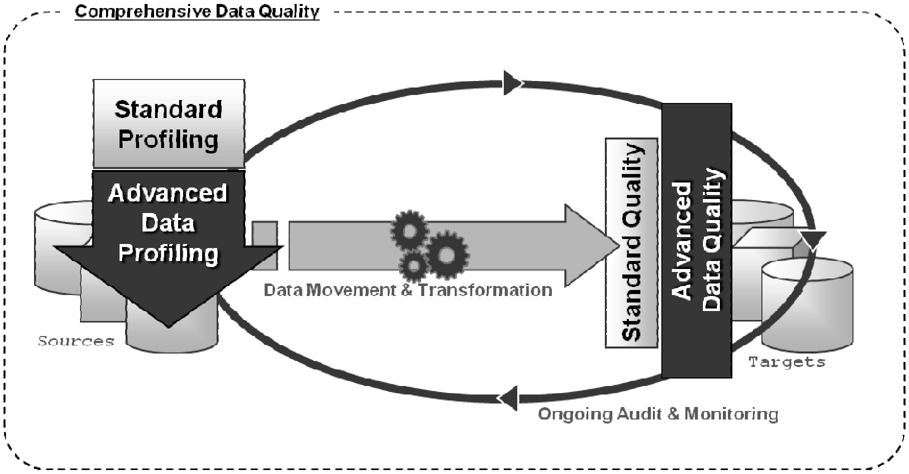
Comprehensive Data Quality with Oracle Data Integrator

An Oracle White Paper
Updated December 2007

Oracle Data Integrator ensures that bad data is automatically detected and recycled before it is inserted in the target applications. Oracle Data Integrator offers standard data quality features, while more advanced matching and deduplication capabilities are available with Oracle Data Quality for Oracle Data Integrator.

EXECUTIVE OVERVIEW

Poor-quality data afflicts almost every company of moderate size and operational complexity. In fact, inconsistent, inaccurate, incomplete, and out-of-date data are often the root cause of expensive business problems such as operational inefficiencies, faulty analysis for business optimization, unrealized economies of scale, and dissatisfied customers. Savvy IT managers can solve a host of these and other business-level problems by committing to a program of comprehensive data quality. Oracle Data Integrator offers a comprehensive data quality solution to meet any data quality challenge for any type of global data with a single, well-integrated technology package.



Oracle Data Integrator provides standard and advanced data quality features.

Oracle's solution for comprehensive data quality includes three products: Oracle Data Integrator, Oracle Data Profiling, and Oracle Data Quality for Oracle Data Integrator. These three best-of-breed technologies work seamlessly together to solve the most challenging enterprise data governance problems.

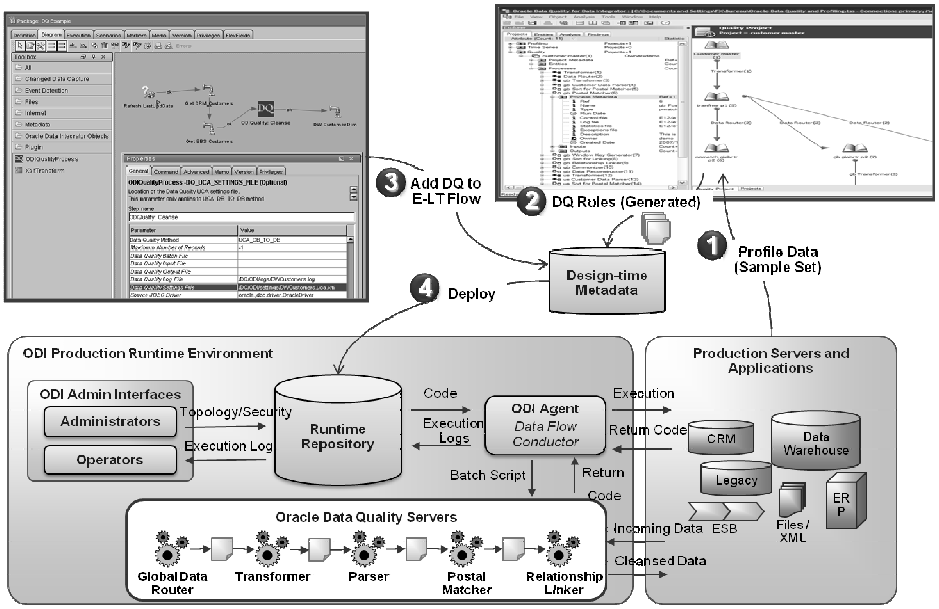
INTRODUCTION

The first step in a comprehensive data quality program is to assess the quality of your data through data profiling. Profiling data means reverse-engineering metadata from various data stores, detecting patterns in the data so that additional metadata can be inferred, and comparing the actual data values to expected data values. Profiling provides an initial baseline for understanding the ways in which actual data values in your systems fail to conform to expectations. Advanced profiling capabilities ensure data assessment is not a one-time activity, but an ongoing practice that ensures data quality over time.

Once data problems are well understood, the rules to repair those problems can be created and executed by data quality engines. For both standard data quality and advanced data quality, an initial set of rules can be generated based on the results of profiling, then users that understand the data can refine and extend those rules. Data quality rules range from ensuring data integrity to sophisticated parsing, cleansing, standardization, matching, and deduplication.

After data quality rules have been generated, fine-tuned, and tested against data samples from within a unified design environment, those rules must be added to data integration processes. Data can be repaired either statically in the original systems or as part of a data flow. Flow-based control minimizes disruption to existing systems and ensures that downstream analysis and processing works on reliable, trusted data.

Finally, the data integration processes—including data quality rules—are placed into production. The runtime performance and reliability of the data quality servers used to process these rules is of utmost importance, whether the rules are applied to batch data flows or to real-time data movement. Advanced profiling creates a closed loop of continual data quality monitoring and increasingly refined data repair.



Profile data, generate data quality rules, add to ETL flow, and execute jobs in real time or batch.

Any data quality problem should be solved using these basic steps. Some data quality challenges can be solved with the standard quality features included with Oracle Data Integrator. More troublesome problems will require the advanced capabilities available with the optional Oracle Data Profiling and Oracle Data Quality technologies as fully integrated components of Oracle Data Integrator. The following sections explain the profiling and quality functions available with the core Oracle Data Integrator, along with the more advanced features available with the optional components, which can enhance the capabilities of your comprehensive data quality solution to meet your specific needs.

STANDARD DATA QUALITY WITH ORACLE DATA INTEGRATOR

Oracle Data Integrator enables application designers and business analysts to define declarative rules for data integrity directly in the centralized Oracle Data Integrator metadata repository. These rules are applied to application data—inline with batch or real-time extract, transform, and load (ETL) jobs—to guarantee the overall integrity, consistency, and quality of enterprise information.

Defining Business Rules

Oracle Data Integrator can automatically retrieve existing rules that have been defined at the data level (such as database constraints) using a customizable reverse-engineering process. Developers can also create new declarative rules without coding by using the graphical user interface in Oracle Data Integrator Designer. These rules can be inferred from data discovery and profiling within Oracle Data Integrator. Developers can immediately test the new declarative rules against the data by performing a synchronous check.

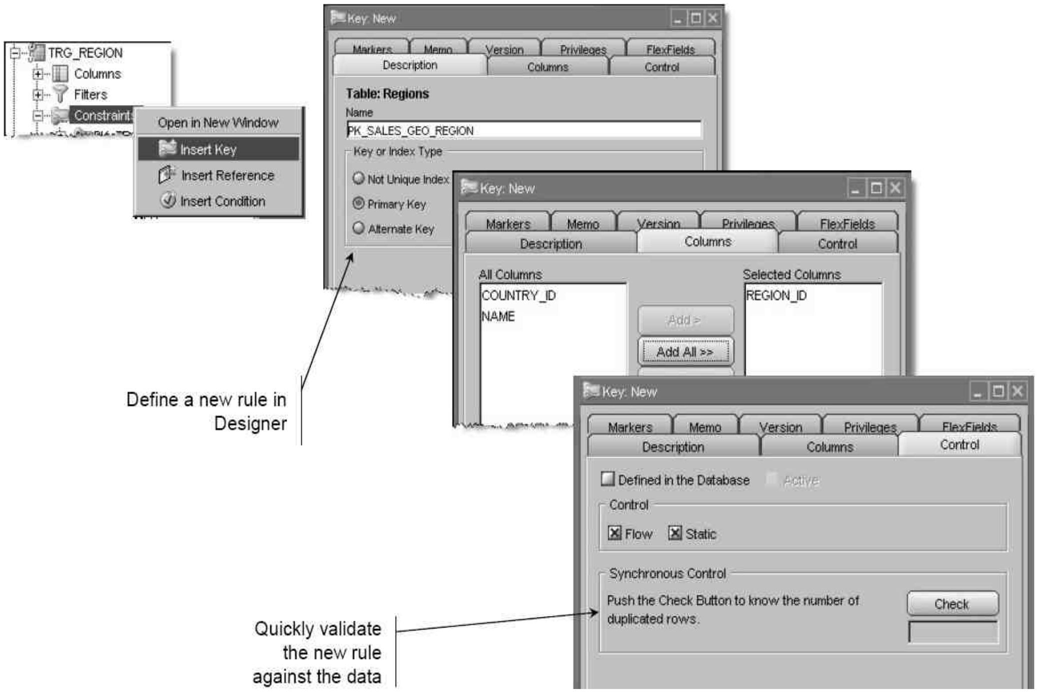


Table: Regions

Name: PK_SALES_GEO_REGION

Key or Index Type

○ Not Unique Index
● Primary Key
○ Alternate Key

All Columns: COUNTRY_ID, NAME

Selected Columns: REGION_ID

☐ Defined in the Database ☐ Active

Control

☒ Flow ☒ Static

Synchronous Control

Push the Check Button to know the number of duplicated rows.

Define a new rule in Designer

Quickly validate the new rule against the data

Data quality rules can be checked against any ETL data inline with Oracle Data Integrator Designer jobs.

Types of Rules

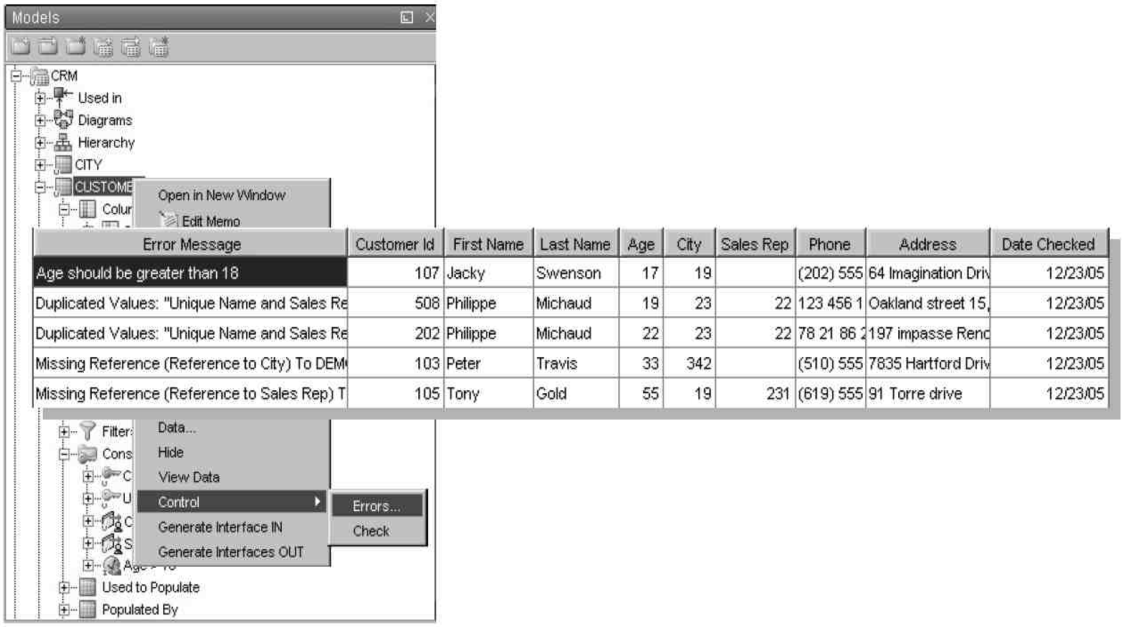
Rules for data integrity can include the following:

- Uniqueness rules
  - “Different customers must not have the same e-mail address”
  - “Different products must have different product and family codes”
- Simple and complex reference rules
  - “All customers must have a sales representative”
  - “Orders must not be linked to customers marked as ‘Invalid’”
- Validation rules that enforce consistency at the record level
  - “Customers must not have an empty zip code”
  - “Web contacts must have a valid e-mail address”

Enforcing Business Rules

Oracle Data Integrator’s customizable Check Knowledge Modules (CKMs) help developers automatically enforce the data integrity of their applications based on declarative rules that have been captured by Oracle Data Integrator. These CKMs generate the code necessary for static or dynamic data checks and also for any error recycling that is performed as part of the integration process.

Audits provide statistics on the integrity of application data. They also isolate data that is detected as erroneous by applying the business rules. Once erroneous records have been identified and isolated in error tables, they can be accessed from Oracle Data Integrator Designer, or from any other front-end application.



Error Message	Customer Id	First Name	Last Name	Age	City	Sales Rep	Phone	Address	Date Checked
Age should be greater than 18	107	Jacky	Swenson	17	19		(202) 555	64 Imagination Driv	12/23/05
Duplicated Values: "Unique Name and Sales Re	508	Philippe	Michaud	19	23	22	123 456 1	Oakland street 15,	12/23/05
Duplicated Values: "Unique Name and Sales Re	202	Philippe	Michaud	22	23	22	78 21 86 2	197 impasse Renc	12/23/05
Missing Reference (Reference to City) To DEM	103	Peter	Travis	33	342		(510) 555	7835 Hartford Driv	12/23/05
Missing Reference (Reference to Sales Rep) T	105	Tony	Gold	55	19	231	(619) 555	91 Torre drive	12/23/05

Erroneous data can be easily reviewed using Oracle Data Integrator Designer’s graphical user interface.

This extensive audit information on data integrity makes it possible to perform a detailed analysis, so that erroneous data can be handled according to information technology strategies and best practices. For example, the following are four ways erroneous data might be handled:

- **Automatically correct data—**Oracle Data Integrator offers a set of tools to simplify the creation of data cleansing interfaces that can be scheduled to run at predetermined intervals.
- **Accept erroneous data (for the current project)—**In this case, interface developers need precise rules for filtering out erroneous data later, using Oracle Data Integrator filters.
- **Correct the invalid records—**In this situation, the invalid data is sent to application end users via various text formats or distribution modes, such as human workflow, e-mail, HTML, XML, flat text files, and so on, using Oracle Data Integrator packages.
- **Recycle data—**Erroneous data from an audit can be recycled into the integration process.

All these strategies can be automated using Oracle Data Integrator's inline interfaces and packages—without any additional data quality components. Therefore, Oracle Data Integrator puts data quality at the very heart of integration processes with robust standard data quality capabilities.

ADVANCED DATA QUALITY AND DATA PROFILING

In situations where the business requirements demand the most advanced data quality capabilities, Oracle Data Integrator can meet those demands with optional functionality available in two new products. Oracle Data Profiling and Oracle Data Quality for Oracle Data Integrator are the result of a joint engineering project between Oracle and the award-winning Trillium Software, whose platform has been the market-leading industrial-strength data quality and profiling platform for many years. The combination of Oracle Data Integrator's best-of-breed E-LT capabilities with Trillium's award-winning data quality platform makes for an unbeatable solution to enterprise-scale data quality issues.

Oracle Data Quality for Oracle Data Integrator and Oracle Data Profiling provide advanced features to run name and address cleansing processes such as

Unless a systematic and reliable approach to data quality is implemented, poor-quality data can propagate across the information technology infrastructure and may "contaminate" other applications.

- **Standardization and cleansing—**Contact and address information is cleansed and standardized based on the country of origin, or according to corporate standards.
- **Address validation and enrichment—**Addresses are validated against national postal authority files and possibly corrected. Additional geographic or third-party information (latitude and longitude, or market targeting information, for example) can also be added to the address information.
- **Matching and deduplication—**Duplicate records are identified and merged into a unique "best" record, and all duplicates are linked to it. Consolidation can then be performed with Oracle Data Integrator to create a unified view with no loss of original data.

Oracle Data Quality not only provides these features for global data—with built-in rules sets for different countries and support for Unicode and double-byte data—

but its cleansing features can also be used against product data, brand data, financial data, and other types of noncustomer party data.

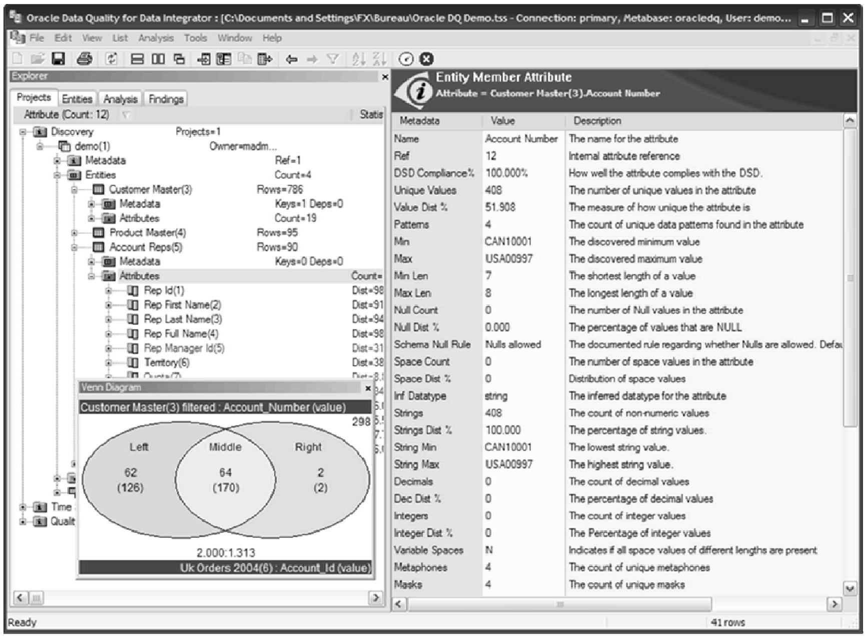
Oracle Data Integrator can use the data standardization, enrichment, and deduplicating features of Oracle Data Quality in real-time—for event-driven changed data capture—or in batch mode.

Further, with Oracle Data Profiling, data stewards and data quality officers can create "bottom-up"—or data-driven—data quality rules by actually working with sample data. In this way, the designers can look at statistically significant portions of their actual data, find outlier populations that they might not have been aware of, and then dynamically build the rules to cleanse and scrub that data. Finally, the data steward or data quality officer can automatically generate the Data Quality Project for Oracle Data Quality to execute at runtime.

These new "bottom-up" capabilities further improve productivity and accuracy by directly linking the profiling activities with the data quality activities—and distinguish the toolset from others that still enforce a "top-down" approach to data quality rule building.

Some key capabilities of Oracle Data Profiling include

- **Entity discovery and analysis—**Oracle Data Profiling collects metadata and data from sources and analyzes it to consolidate information and statistics such as attribute lengths, maximum and minimum values, value distributions, patterns, data types, and so on. It automatically applies advanced profiling techniques to identify potential problems with data fields such as nonconforming zip codes or customer/product codes, misspellings, duplications, or punctuation issues.
- **Natural drill-down—**The user interface and graphical views enable users to browse back and forth through the analysis results using a natural drill-down approach.
- **Keys, functional dependencies, and joins discovery and analysis—** Oracle Data Profiling detects and presents potential keys, with their degree of uniqueness, while identifying duplicates and other inconsistencies. It can also detect functional dependencies between attributes within a given entity (a shipped order should have an invoice number), as well as relationships between entities (joins). Quality analysts can also create all these types of quality rules within the user interface.
- **Quality monitoring over time—**The Time Series feature allows users to evaluate data quality over time by performing regular assessments of the data quality. E-mail notifications warn business users when certain service-level requirements are not met.


Oracle Data Quality for Data Integrator : [C:\Documents and Settings\FX\Bureau\Oracle DQ Demo.tss - Connection: primary, Metabase: oracledq, User: demo...
File Edit View List Analysis Tools Window Help
Explorer
Projects
Entities
Analysis
Findings
Attribute (Count: 12)
Statis
Venn Diagram
Left
Middle
Right
Ready
41 rows

Discovery Projects=1
demo(1) Owner=madm...
Metadata Ref=1
Entities Count=4
Customer Master(3) Rows=786
Metadata Keys=1 Deps=0
Attributes Count=19
Product Master(4) Rows=95
Account Reps(5) Rows=90
Metadata Keys=0 Deps=0
Attributes Count=
Rep Id(1) Dist=98
Rep First Name(2) Dist=91
Rep Last Name(3) Dist=94
Rep Full Name(4) Dist=98
Rep Manager Id(5) Dist=31
Territory(6) Dist=38
Time
Qualit

Customer Master(3) filtered : Account_Number (value)

298

62 (126) · 64 (170) · 2 (2)

2.000:1.313

Uk Orders 2004(6) : Account_Id (value)

Entity Member Attribute

Attribute = Customer Master(3).Account Number

Metadata	Value	Description
Name	Account Number	The name for the attribute
Ref	12	Internal attribute reference
DSD Compliance%	100.000%	How well the attribute complies with the DSD.
Unique Values	408	The number of unique values in the attribute
Value Dist %	51.908	The measure of how unique the attribute is
Patterns	4	The count of unique data patterns found in the attribute
Min	CAN10001	The discovered minimum value
Max	USA00997	The discovered maximum value
Min Len	7	The shortest length of a value
Max Len	8	The longest length of a value
Null Count	0	The number of Null values in the attribute
Null Dist %	0.000	The percentage of values that are NULL
Schema Null Rule	Nulls allowed	The documented rule regarding whether Nulls are allowed. Defau
Space Count	0	The number of space values in the attribute
Space Dist %	0	Distribution of space values
Inf Datatype	string	The inferred datatype for the attribute
Strings	408	The count of non-numeric values
Strings Dist %	100.000	The percentage of string values.
String Min	CAN10001	The lowest string value.
String Max	USA00997	The highest string value.
Decimals	0	The count of decimal values
Dec Dist %	0	The percentage of decimal values
Integers	0	The count of integer values
Integer Dist %	0	The Percentage of integer values
Variable Spaces	N	Indicates if all space values of different lengths are present
Metaphones	4	The count of unique metaphones
Masks	4	The count of unique masks

Oracle Data Profiling's intuitive user interface speeds project delivery times.

Oracle Data Profiling takes the input from the user and then converts that to an automatically generated set of data quality rules.

CHOOSING THE RIGHT TOOLS

Not every data integration project requires advanced data quality and data profiling abilities, but how do you choose the right tool for each project? Certain trade-offs should be considered along functional abilities, while others should be along performance and architectural implications on overall quality of service (QoS) and service-level agreements (SLA). Here are some questions to consider when selecting a comprehensive data quality solution:

- What is the acceptable balance between high quality and low latency? (Typically, the higher the quality your data must be, the more time is required to introspect that data, apply cleansing algorithms, compare to trusted sources, and finally insert to a warehouse or operational system.)
- Can data quality be enforced at point-of-entry, or only in batch? (Often, the best way to improve data quality is to prevent bad data from the outset—but sometimes this can be impractical if it slows the end-user application down or entails a major front-office upgrade.)
- Is standard data quality good enough, or do I need advanced abilities? (Sometimes it is sufficient to be able to enforce formats and constraints and provide core pattern matching and find/replace functions, without upgrading to highly advanced data quality features.)

The following table details some of the differences between the standard data quality features of Oracle Data Integrator and the more advanced quality features of Oracle Data Quality for Oracle Data Integrator.

Data Quality Features	Standard	Advanced
E-LT Style Integrity Check Control	Y	Y
Easy “Error Hospital” Workflow Integration	Y	Y
Uniqueness Rules for Basic Matching	Y	Y
Complex Cross-Reference Rules	Y	Y
Record-Level Validation/Standardization	Y	Y
External Key Constraint Checking	Y	Y
Cleanse All Types of Data	Y	Y
Advanced Notifications (E-mail, SOA, etc.)	Y	Y
Out-of-Box International Rule Sets		Y
Street-Level Global Post Codes		Y
Highly Customizable Rule Templates		Y
Deep, Preconfigured Matching Rule Sets		Y
Rich Data Survivorship Settings		Y
Out-of-Box Data Quality Sample Projects		Y

Note that the standard data quality features can be delivered with an inline ETL process that usually operates with set-based efficiency—that is, data quality can be achieved with extreme performance benefits while batch data is being extracted, loaded, staged, or transformed.

The following table details some of the differences between the standard data profiling features of Oracle Data Integrator and the more advanced features of Oracle Data Profiling.

Data Profiling Features	Standard	Advanced
DBMS Metadata Reverse-Engineering	Y	Y
DW Appliance Metadata Reverse-Engineering	Y	Y
Application Interface Reverse-Engineering	Y	Y
Schema or User-Generated Constraints	Y	Y
Drill-Down and Sample Data Browsing	Y	Y
Automatic Profile Report Generation	Y	Y
Integrated Monitoring, Audit, and Profiling		Y
Out-of-Box Data Profiling Sample Projects		Y
Entity, Key, and Join Discovery and Analysis		Y
Automatic Runtime Project Generation		Y
Time Series–Based Quality Monitoring		Y
Annotation and Assessments		Y

Note that the standard data profiling features are delivered as part of the ETL design process, thereby providing productivity gains when reverse-engineering metadata from databases, large-scale data warehouse appliances, and data-centric applications. These benefits become readily apparent in team-based settings where many legacy systems must be profiled for the data integration design to be completed.

CONCLUSION

Comprehensive data quality should be a key enabling technology for any IT infrastructure, and it is critical to solving a range of expensive business problems. Comprehensive data quality is particularly important in the context of any data integration process to prevent data quality problems from proliferating. Oracle Data Integrator's inline, stepped approach to comprehensive data quality ensures that data is adequately verified, validated, and cleansed at every point of the integration process. Oracle Data Integrator has both standard and advanced data quality capabilities, which feature the same high performance and simplicity that are characteristic of the entire Oracle Fusion Middleware technology stack.

ORACLE

Comprehensive Data Quality with Oracle Data Integrator
Updated December 2007

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com

This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.