An Oracle White Paper
January 2013

Comprehensive Data Quality with Oracle Data Integrator and Oracle Enterprise Data Quality

Executive Overview

Poor data quality impacts almost every company. In fact, according to a research from Gartner "Companies routinely make decisions based on remarkably inaccurate or incomplete data, a leading cause of the failure of high-profile and high-cost IT projects such as business-intelligence and customer-relationship management deployments". Inconsistent, inaccurate, incomplete, and out-of-date data are often the root cause of expensive business problems such as operational inefficiencies, faulty analysis for business optimization, unrealized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.

These data quality issues and the business-level problems associated with it can be solved by committing to a comprehensive data quality effort across the enterprise. Oracle Data Integrator offers a complete data integration solution to meet any data quality challenge for any type of data domains with a single, well- integrated technology package.

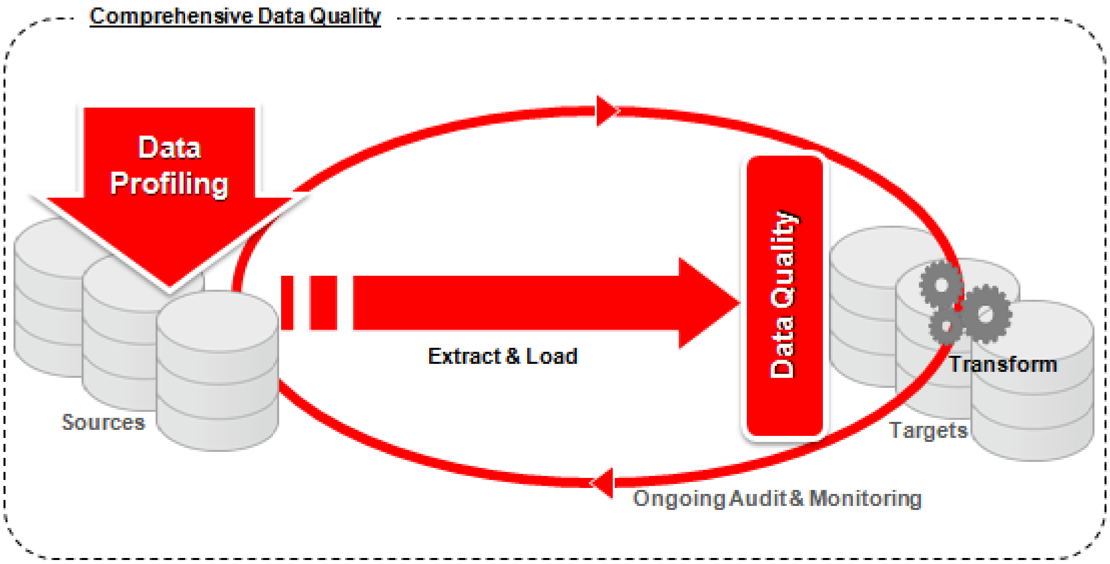


Figure 1 – Comprehensive Data Quality Process

Oracle's solution for comprehensive data quality includes two products: Oracle Data Integrator and Oracle Enterprise Data Quality. These best-of-breed technologies work seamlessly together to solve the most challenging enterprise data quality 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 analyzing 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 with full drill down capabilities. Profiling provides an initial baseline for understanding the ways in which actual data values in your systems fail to conform to expectations. When used prior to designing integration processes, data profiling helps reduce the implementation time and lowers the associated risks. In addition, advanced profiling capabilities ensure data assessment is not a one-time activity, but an ongoing practice that ensures trusted data over time.

Once data problems are well understood, the rules to repair those problems can be created and executed by data quality engines. 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, validation and de-duplication.

After data quality rules have been generated, fine-tuned, and tested against data samples, those rules must be added to data integration processes to ensure a pervasive data quality framework is in place across the enterprise. 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.

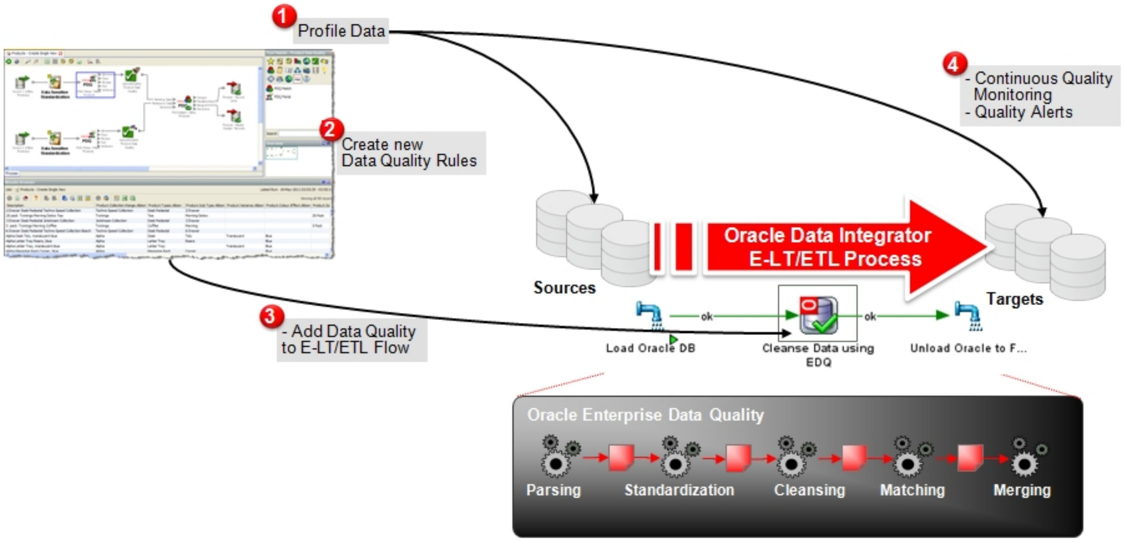

1 Profile Data
2 Create new Data Quality Rules
3 - Add Data Quality to E-LT/ETL Flow
4 - Continuous Quality Monitoring - Quality Alerts
Sources
Oracle Data Integrator E-LT/ETL Process
Targets
Load Oracle DB
ok
Cleanse Data using EDQ
ok
Unload Oracle to F...
Oracle Enterprise Data Quality
Parsing
Standardization
Cleansing
Matching
Merging

Figure 2 - Profile data, generate data quality rules, add to ETL flow, and execute overall data integration jobs

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. Data profiling creates a closed loop of continuous data quality monitoring and increasingly refined data repair.

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 Enterprise Data Quality technology which is integrated with Oracle Data Integrator.

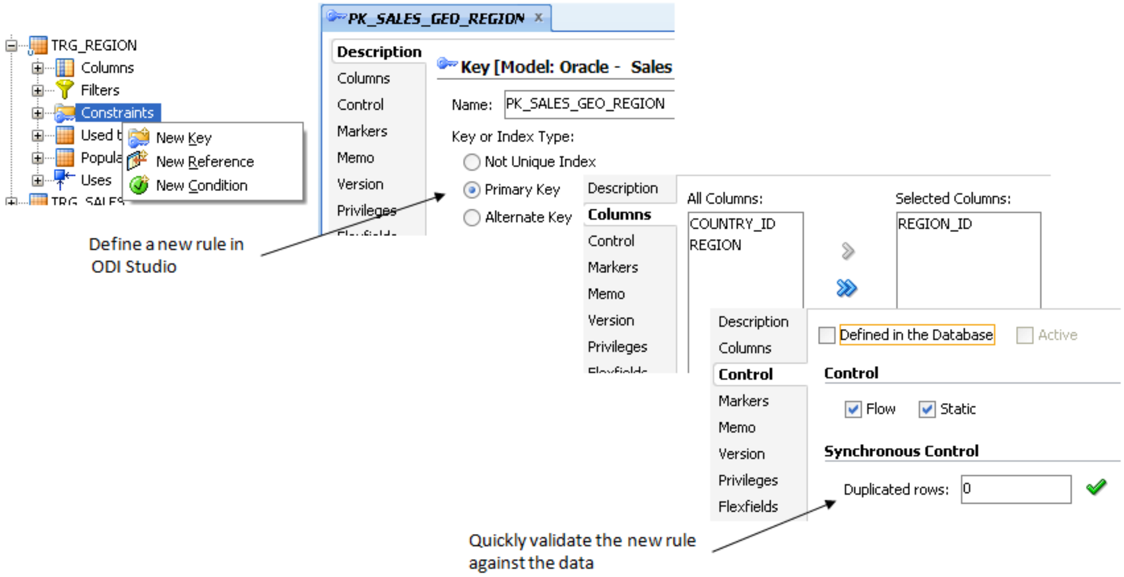
The following sections explain the profiling and quality functions available with the core Oracle Data Integrator technology, along with the more advanced features available with Oracle Enterprise Data Quality.

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 Studio. These rules can be inferred by looking at the data within Oracle Data Integrator. Developers can immediately test the new declarative rules against the data by performing a synchronous check.



Name: PK_SALES_GEO_REGION

Key or Index Type:

○ Not Unique Index

● Primary Key

○ Alternate Key

All Columns: COUNTRY_ID, REGION

Selected Columns: REGION_ID

☐ Defined in the Database ☐ Active

☑ Flow ☑ Static

Duplicated rows: 0 ✓

Define a new rule in ODI Studio

Quickly validate the new rule against the data

Figure 3 - Data quality rules can be checked against any ETL data inline with Oracle Data Integrator

Type of Rules

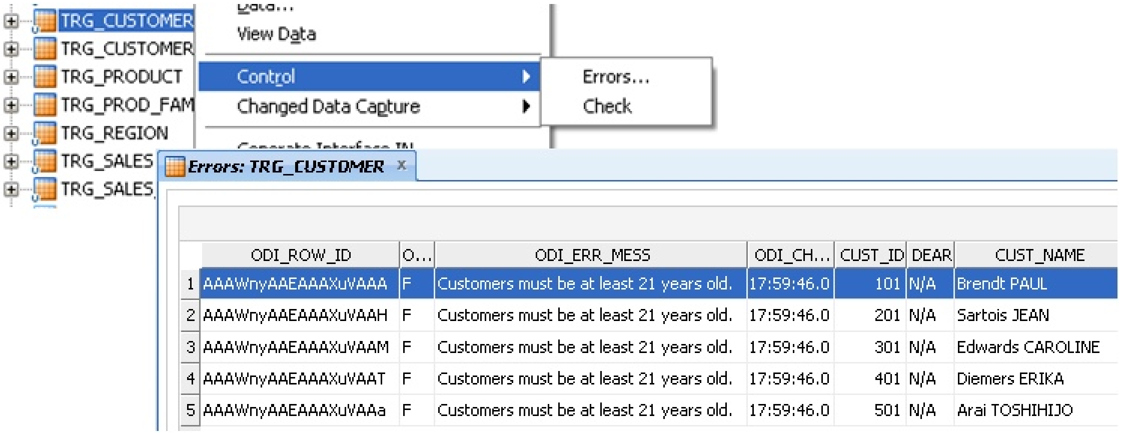
Rules for data integrity can include the following:

- Uniqueness rules
    - "Different customers must not have the same e-mail address"
    - "Different products must have different product and family codes"
- Referential integrity 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 Studio, or from any other front-end application.



	ODI_ROW_ID	O...	ODI_ERR_MESS	ODI_CH...	CUST_ID	DEAR	CUST_NAME
1	AAAWnyAAEAAAXuVAAA	F	Customers must be at least 21 years old.	17:59:46.0	101	N/A	Brendt PAUL
2	AAAWnyAAEAAAXuVAAH	F	Customers must be at least 21 years old.	17:59:46.0	201	N/A	Sartois JEAN
3	AAAWnyAAEAAAXuVAAM	F	Customers must be at least 21 years old.	17:59:46.0	301	N/A	Edwards CAROLINE
4	AAAWnyAAEAAAXuVAAT	F	Customers must be at least 21 years old.	17:59:46.0	401	N/A	Diemers ERIKA
5	AAAWnyAAEAAAXuVAAa	F	Customers must be at least 21 years old.	17:59:46.0	501	N/A	Arai TOSHIHIJO

Figure 4 - Erroneous data can be easily reviewed using Oracle Data Integrator Studio

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 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.

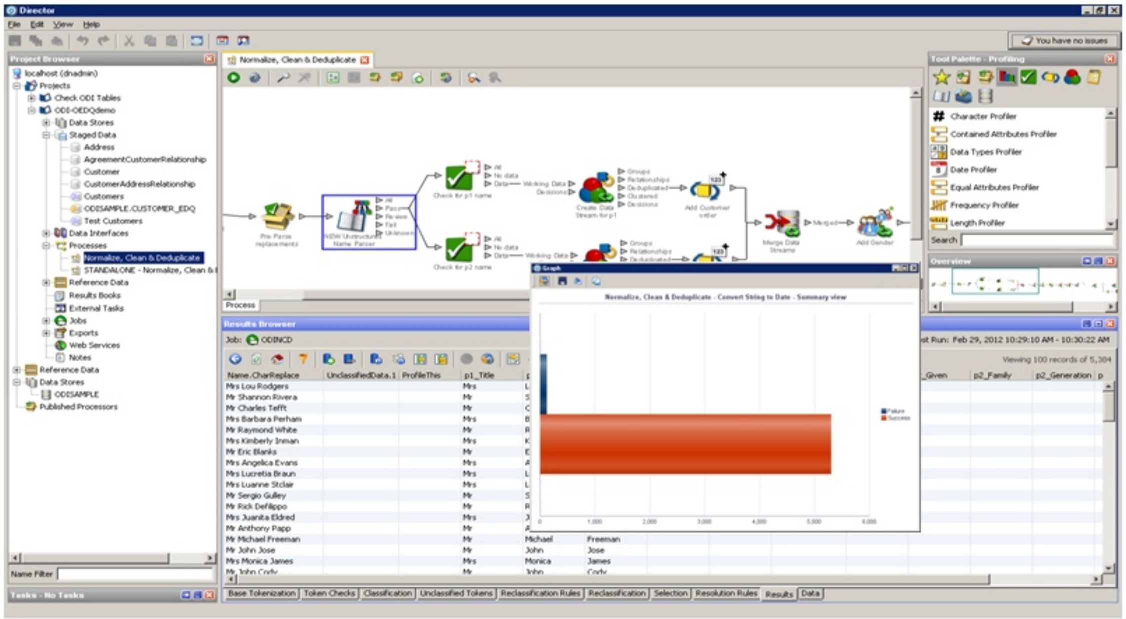
Advanced Data Quality and Data Profiling with Oracle Enterprise Data Quality

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 Oracle Enterprise Data Quality. Oracle Enterprise Data Quality provides an end-to-end solution to measure, improve, and manage the quality of data from any domain, including customer and product data. The combination of Oracle Data Integrator's best-of-breed E-LT capabilities with Oracle Enterprise Data Quality platform makes for an unbeatable solution to enterprise-scale data quality issues.

Oracle Enterprise Data Quality provides advanced features such as

- **Profiling**—Data Profiling capabilities help users to analyze and understand their data, highlighting key areas of data discrepancy and the business impact of these problems. Users can learn from historical analysis and define business rules directly from the data thanks to an integrated data quality user interface.
- **Parsing and Standardization**—A rich palette of functions to parse, cleanse and standardize any type of data is provided. Using easily managed reference data and a simple graphical configuration, users can quickly configure, package, share and deploy rules specific to their data and industry without any coding.
- **Matching and Merging**—Powerful matching capabilities allow users to identify matching records and optionally link or merge matched records together based on survivorship rules. Flexible and intuitive configuration capabilities allow matching and merging rules to be easily tuned to suit your needs.
- **Address Validation**—Oracle Enterprise Data Quality offers an optional address standardization and enhancement module which can be used in a data quality process. This component supports more than 240 countries worldwide and has built-in geocoding capabilities.

Oracle Enterprise 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 non-customer party data.



Name.CharReplace	UnclassifiedData.1	ProfileThis	p1_Title	p		_Given	p2_Family	p2_Generation	p
Mrs Lou Rodgers			Mrs	L					
Mr Shannon Rivera			Mr	S					
Mr Charles Tefft			Mr	C					
Mrs Barbara Perham			Mrs	B					
Mr Raymond White			Mr	R					
Mrs Kimberly Inman			Mrs	K					
Mr Eric Blanks			Mr	E					
Mrs Angelica Evans			Mrs	A					
Mrs Lucretia Braun			Mrs	L					
Mrs Luanne Stclair			Mrs	L					
Mr Sergio Gulley			Mr	S					
Mr Rick Defilippo			Mr	R					
Mrs Juanita Eldred			Mrs	J					
Mr Anthony Papp			Mr	A					
Mr Michael Freeman			Mr	Michael	Freeman				
Mr John Jose			Mr	John	Jose				
Mrs Monica James			Mrs	Monica	James				
Mr John Cody			Mr	John	Cody				

Figure 5 - Profiling and cleansing data using Oracle Enterprise Data Quality

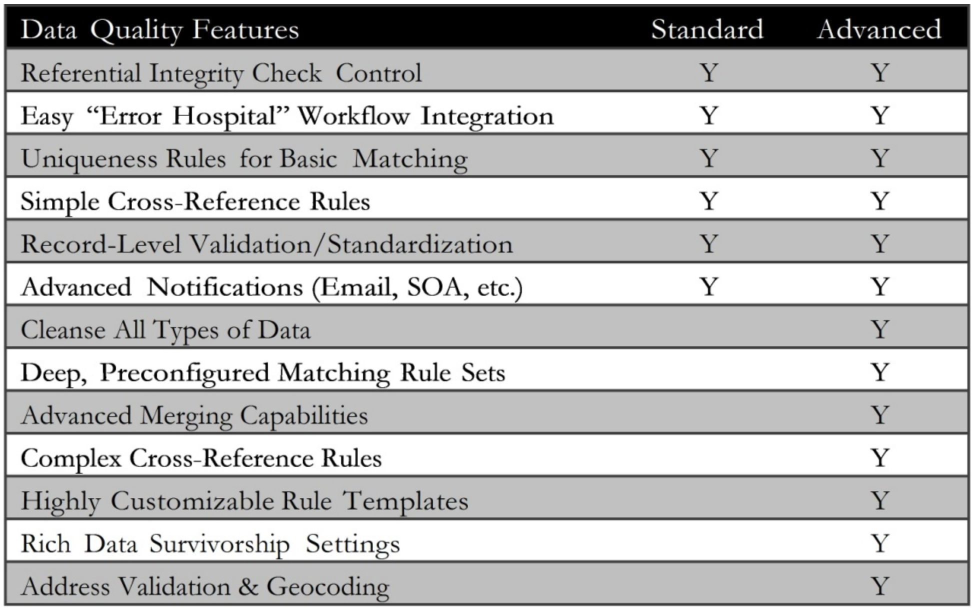
Oracle Data Integrator and Oracle Enterprise Data Quality are well integrated and users can leverage the data parsing, standardization, enrichment, and matching features of Oracle Enterprise Data Quality in their ETL processes.

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:

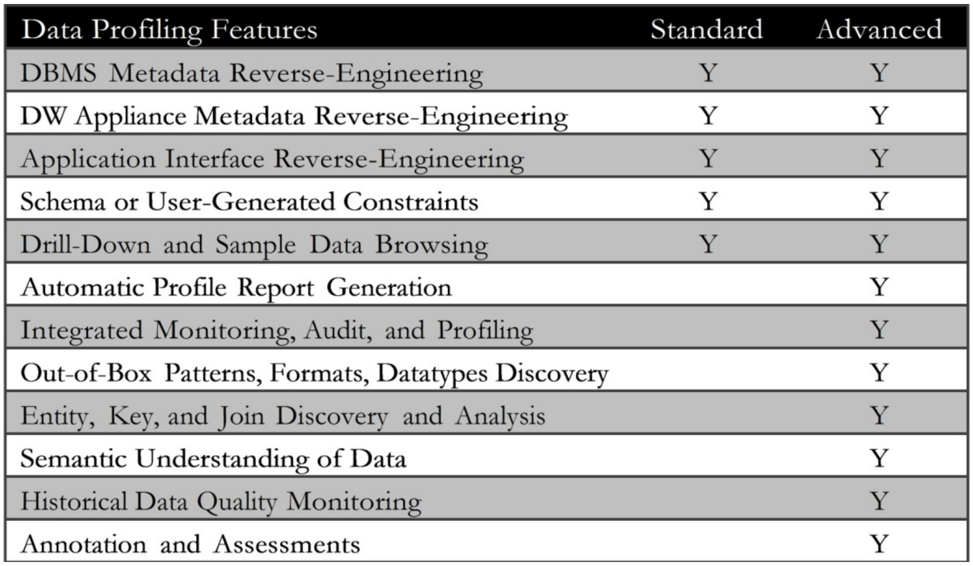
- What is the acceptable balance between high quality and low latency? (Typically, the higher the quality your data must be, the more time is required to introspect that data, apply cleansing algorithms, compare to trusted sources, and finally insert to a warehouse or operational system.)
- Can data 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 down the end-user application or entails a major front-office upgrade.
- Is standard data quality good enough, or do I need advanced abilities?

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 Enterprise Data Quality.



Data Quality Features	Standard	Advanced
Referential Integrity Check Control	Y	Y
Easy "Error Hospital" Workflow Integration	Y	Y
Uniqueness Rules for Basic Matching	Y	Y
Simple Cross-Reference Rules	Y	Y
Record-Level Validation/Standardization	Y	Y
Advanced Notifications (Email, SOA, etc.)	Y	Y
Cleanse All Types of Data		Y
Deep, Preconfigured Matching Rule Sets		Y
Advanced Merging Capabilities		Y
Complex Cross-Reference Rules		Y
Highly Customizable Rule Templates		Y
Rich Data Survivorship Settings		Y
Address Validation & Geocoding		Y

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 Enterprise Data Quality.



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
Integrated Monitoring, Audit, and Profiling		Y
Out-of-Box Patterns, Formats, Datatypes Discovery		Y
Entity, Key, and Join Discovery and Analysis		Y
Semantic Understanding of Data		Y
Historical Data Quality Monitoring		Y
Annotation and Assessments		Y

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

Oracle is committed to developing practices and products that help protect the environment

Comprehensive Data Quality
with Oracle Data Integrator and
Oracle Enterprise Data Quality
August 2012
Author: Julien Testut

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

Copyright © 2013, Oracle and/or its affiliates. All rights reserved. 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.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. UNIX is a registered trademark licensed through X/Open Company, Ltd. 0410

SOFTWARE. HARDWARE. COMPLETE.