Oracle Warehouse Builder 10*g* Release 2

Transforming Data into Quality Information

An Oracle White Paper
January 2007

ORACLE

Note:

This document is for informational purposes. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described in this document remains at the sole discretion of Oracle.

Oracle Warehouse Builder 10gR2
Transforming Data into Quality Information

INTRODUCTION

Enterprises have always relied on data to be successful. Customers, products, suppliers, and sales transactions all need to be described and tracked in one way or another. Even before computers became commercially available, the data in the form of paper records has been vital to both commercial and non-commercial organizations. With the advent of computing technology, the sophistication of data usage by businesses and governments grew exponentially. The technology industry serving these needs has generated many buzz words that come and go: decision support systems, data warehousing, customer relationship management, business intelligence, etc., but the fact remains the same—organizations need to make the best use of the data they have to increase their efficiency today and improve their planning for tomorrow.

If the data is so fundamental to the business, it is not surprising that a lot of effort goes into acquiring and handling data and making it available to those who need it. In the process, the data is moved around, manipulated, and consolidated. The quality of the data is rarely a high priority as the pressure is on to deliver the project and "get the data out to the users." The justifications for not making data quality a priority are often just thoughts such as "our data is good enough" and "we can always clean it later."

Yet it is proven that data quality is one of the biggest obstacles blocking the success of any data integration project. The often-quoted estimate by The Data Warehouse Institute (TDWI) is that data quality problems cost U.S. businesses more than $600 billion a year— a very impressive number, but hard to relate to without first understanding the concepts of data quality and the technologies that address data quality issues.

This paper answers the questions: What is data quality? Why put an effort into data quality? Why is this effort most efficient inside the extract, transform, and load (ETL) process? How will Warehouse Builder make this effort successful?

You will discover how Warehouse Builder combines its core data integration functionalities with advanced data quality functionality.

WHAT IS DATA QUALITY?

Data quality is an all-encompassing term describing both the state of data that is complete, accurate, and relevant, as well as the set of processes to achieve such a state. The goal is to have data free of duplicates, misspellings, omissions, and unnecessary variations, and to have the data conform to the defined structure. Simply put, the data quality addresses the problem cynically but precisely summed up as "garbage in-garbage out."

A significant part of data quality deals with customer data—names and addresses, due to both their key roles in business processes and their highly dynamic nature. Names and addresses are ubiquitous—they tend to exist in almost every source and are often the only identifying data. Most matching applications rely heavily on names and addresses, because a common unique identifier is typically not available across systems. Consequently, whatever data is available must be used to determine if different individuals, businesses, or other types of records are actually the same. But names and addresses usually contain dirty data, since they often include nicknames, abbreviations, misspellings, bad fielding, etc. Furthermore, the name and address data consistently deteriorates over time as people move and/or change last names. Sophisticated name and address processing can solve these problems and have a significant positive effect on matching.

The focus of data quality on names and addresses sometimes causes a misconception that data quality is just about ensuring that names and addresses are correct and thus postal mailings will be deliverable to customers. Therefore, the thinking goes, if your business does not send bills or orders to customers by mail, the data quality is not that important. This is wrong for two reasons: 1. The correct and standardized name and address data is not the end goal, or not the only goal. The end goal is to identify and match customers reliably based on name and address data. 2. The data quality is certainly not limited to names and addresses. Any data, such as product data, will benefit from being standardized and complete.

DATA QUALITY IS CRITICAL

There are many challenges in building a data warehouse or performing any data integration project. Consequently there are many areas of importance for the tools that assist with building a data warehouse. Is it easy to build and maintain the data warehouse (productivity and maintainability)? Can all required sources of data be brought into the data warehouse (source integration)? Can the data warehouse continue to add data within allowed time periods and can it accommodate the growth (performance and scalability)?

However, to the end users of the data warehouse making the business decisions based on the data in the data warehouse, the most important question is "can I trust this data?" It doesn't matter if the data warehouse was built in record time, it practically maintains itself, every imaginable source system is in, and the loads are blazingly fast and are only getting faster. If the business users find the data in the data warehouse not to be trustworthy, they will not use it to make their business

decisions. All the impressive achievements in those other areas will then be meaningless.

Having complete, accurate, and relevant data free of all defects sounds great on paper but is never entirely achieved in reality. Instead, each organization defines, formally or informally, the acceptable level of data quality; i.e., a certain threshold that can be measured. Just as impressive are the figures that show the losses to businesses due to poor data quality and the estimates on how much is spent on trying to address the data quality issues. Therefore, it is not surprising that the true goal for businesses is not the absolute data quality in its academic sense but the acceptable level of data quality achieved efficiently in terms of time, effort, and money.

INTEGRATING DATA QUALITY INTO ETL

For anyone who follows the data integration industry through the press, analyst reports, or vendor events, there should be nothing new so far. The definition of data quality and the acknowledgement of its importance are universal in the industry.

Compared to other solution, the key difference of data quality in Warehouse Builder is the level of integration into the ETL process, making data quality transformations an inseparable function of data warehouse development.

Consider an analogy with a spell checker in a word processor. It would hardly be acceptable to anyone to close the document, start a spell check application, process and correct the document, then re-open it in the word processor—even if the spell checker has a great user interface of its own, is fast, and is otherwise pleasant to work with. Why not just invoke the spell checker from within a word processor? Better yet, why not correct the mistakes as the document is created? Of course there is no such problem in modern word processors. Spell checkers have long been integrated and work seamlessly. Another interesting parallel to note is that it's fairly unimportant to the user of a word processor that the actual spell checking technology is provided by a third-party vendor.

So within Warehouse Builder, data quality is built-in, using the same interface for design and management of data quality processes as for the ETL processes, just like your word processor and its spell checker.

Further, Warehouse Builder offers product-wide services that data quality transformations benefit from in the enterprise environments. Just like any other transformation in Warehouse Builder, data quality transformations are accessible in scripting language and through public application programming interfaces (APIs), in addition to the graphical user interface. The audit data pertaining to the execution of data quality processes is available in the same place and format as the audit data about other ETL processes.

The integration of data quality functionality in Warehouse Builder is not just for the sake of convenience. The higher goal is to promote the disciplined approach to data

quality, encouraging the ETL developers to think of data quality as they design the data integration processes and enabling them to incorporate it into these processes, as opposed to treating it as an afterthought.

DATA QUALITY IN ORACLE WAREHOUSE BUILDER 10GR2

What is it exactly that Warehouse Builder offers in the area of data quality? Warehouse Builder enables the following major data quality tasks:

- Data profiling
- Data rules derived from data profiling or entered manually
- Data auditors based on data rules
- Name and address cleansing
- Matching and merging of data

These functionalities are interdependent and intended to cover the complete cycle of data integration development, as illustrated in Figure 1 below.

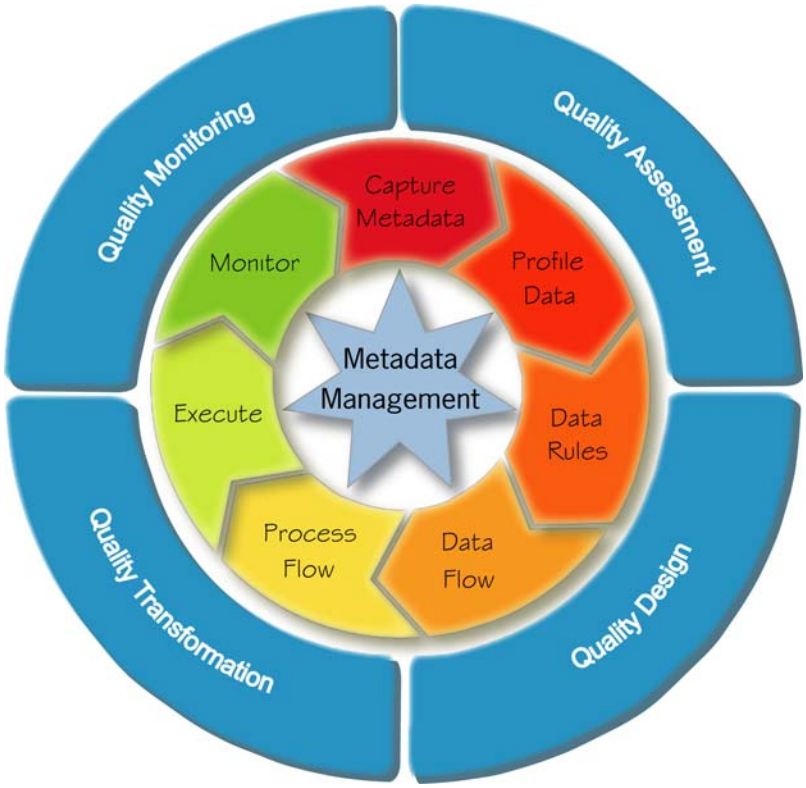

Figure 1. Data Quality Cycle

A typical use of data quality in Warehouse Builder will follow the cycle of Quality Assessment → Quality Design → Quality Transformation → Quality Monitoring.

Specifically, the steps performed in Warehouse Builder to insure data quality are:

- Starting at the 12 o'clock position, the metadata about data sources is captured.
- Next, the data sources are profiled.
- The data rules are then derived from data profiling or existing data rules are imported or entered manually.
- Data flows (mappings) are designed utilizing name and address and match-merge operators
- Data flows are combined in process flows, adding data auditors that measure data quality at any given point in the process
- The processes are deployed and executed, transforming raw data into quality information
- Information quality is, as a final step, continuously monitored in the operational environment by Warehouse Builder's data auditor programs

Data profiling, data rules and data auditors are brand new features of Oracle Warehouse Builder 10*g* Release 2. Name and address cleansing and match-merge have been introduced in prior releases of Warehouse Builder and enhanced for this release.

Data Profiling

The premise of data profiling is simple: look at the data content itself to incur the meaning of the data. The existing metadata is considered as well but it is not automatically trusted. If you've ever issued a query to check if there are null values in a given column, you have performed (very trivial) data profiling. If you've ever "eyeballed" the data to discover that a column of character data types actually contains only numbers or dates, you have performed (very unreliable) data profiling.

Data profiling in Warehouse Builder is a systematic analysis of data sources chosen by the user for the purpose of gaining an understanding of and confidence in the data. It is a critical first step in data integration process development that validates the assumptions about data or discovers new characteristics of the data. In the end, data profiling helps avoid the proverbial "code-load-explode" scenario.

Users' interaction with the profiling functionality in Warehouse Builder is through a wizard that lets them select which tables, views, dimensions, and cubes need to be profiled. Warehouse Builder then executes an asynchronous job that performs the necessary analysis. Once the job is complete the results are presented in the data profile manager as shown in Figure 2. The data profile manager organizes the data profiling findings by type as aggregation, data type, pattern, etc., arranged in tabs on the upper-right part of the screen. For each data object in a data profile, the tabbed interface shows each column and its data profiling metrics. The results can also be represented in graphs and diagrams. Most data profiling findings allow drill-down

into the underlying data that contributed to a given finding, signified by the URL-like notation. The drill-down data is in the bottom-left (distinct values) and bottom-right (complete records) parts of the screen.

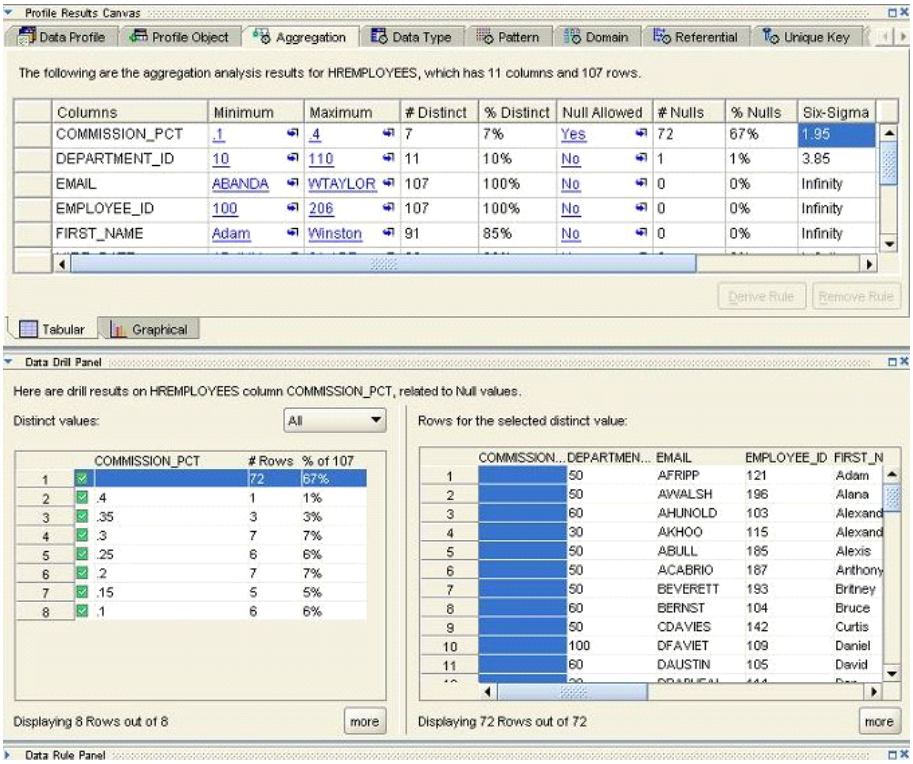

Figure 2. The Data Profile Manager

An example of the value of data profiling integration with ETL is viewing data profiling results directly from within a mapping editor for a given table, view, etc.

If such a table has been previously profiled, the data profile manager can be invoked through right-click action on a mapping table operator, with context automatically set. Profiling results are also available from within the mapping editor to verify what ETL logic could be required on this source.

Data profiling does not just work on Oracle database objects, you can profile other systems as well. Any application that is accessible via an Oracle Transparent Gateway is accessible to profiling, opening up profiling capabilities to such sources as Excel sheets (via ODBC). Flat files can be transparently profiled as well and using its SAP R/3 integrator even your SAP data is available for Warehouse Builder to profile. Imagine the wealth of information you have at your fingertips after profiling all these sources.

Data Rules and Automatic Data Correction

A logical question from looking at the impressive collection of data profiling results is: "What do I do with this?" Traditionally, it has been up to an ETL developer,

data analyst, or architect to interpret the data profiling results and make notes that influence the design of the data integration process. Warehouse Builder certainly allows such use, but it does not stop there. Instead, Warehouse Builder attempts to automate the actions, based on data profiling findings, through data rules and data correction automation.

A data rule is a user-entered or process-derived expression that determines the legal data within a data object or legal relationships between data objects. For example, a data rule could be "Gender is 'M' or 'F' or 'U' ." While reviewing the results in data profiling, it is possible to derive a rule that will populate legal values automatically, without typing anything. The data rule is then stored as a regular metadata object, accessible through the data rules node on the project explorer tree. A data rule can be attached to any data object, not just the one it was derived from.

Taking this automation even further, Warehouse Builder can generate the correction mappings that will evaluate the source data against chosen data rules and either remove or attempt to correct the invalid values, based on user choices. Figure 3 shows the data correction wizard, which specifies the strategy of dealing with invalid values. The result of this wizard is a mapping generated automatically that implements the chosen logic. The mapping can be reviewed and changed by the user. This is another example of the power of data profiling integration with the ETL tool.

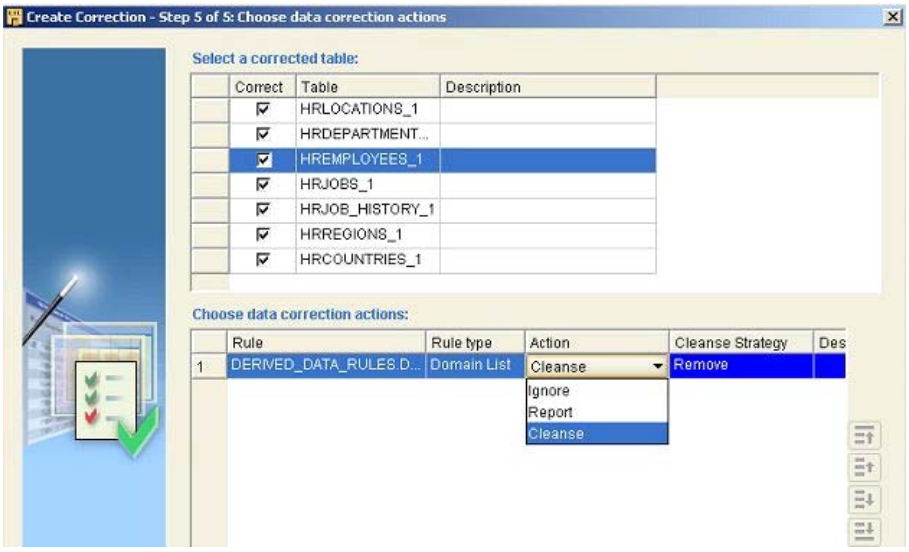

Figure 3, Data Correction Wizard

Data Auditors

To ensure information quality is monitored while the ETL processes are running on a scheduled basis, Warehouse Builder introduces data auditors.

Data auditors are process flow activities that evaluate one or more data rules against a given target (such as a table). Data auditors are not simply pass/fail checks; they have a tolerance threshold setting, expressed as a percentage of errors or as a six-

sigma value. Both the pass/fail as well as the actual audit assessment value can be used in the process flow.

Once the number of defective rows in the target exceeds the threshold, any action can be modeled in the subsequent process flow (e-mail, notification for action, abort of run etc.) and you can store the quality measure of this run.

Because data auditors are executed each time as part of the ETL process, they ensure continuous monitoring of the data quality, according to the data rules and ensure the circle of quality is complete.

Name and Address Cleansing

Name and address cleansing is a group of transformations performed on data containing individual and business names, as well as domestic and foreign addresses, for the purpose of improving the quality of data. Such transformations are usually referred to as parsing, standardization, correction, and augmentation.

- Name/address parsing is the breakdown of non-discrete input into discrete name or address components.
- Name/address standardization is the modification of components to a standard version acceptable to a postal service or suitable for record matching. For example 'Street' and 'Str' are standardized to 'St'.
- Postal correction involves matching an input address with postal database entries to verify and/or correct an address.
- Augmentation adds derived information to the data, such as gender based on name, or collection of census, and geo-location data.

The way Warehouse Builder accomplishes such transformations is as follows. At design time, the wizard-driven name and address operator is used to model name and address rules inside Warehouse Builder. This modeling is quite simple: tell the name and address operator what you know about your input data, through assignments of input roles, as illustrated in Figure 4. Then tell the name and address operator what output components you want to get back. The output components will contain parsed standardized and verified values, including the values that were not available in the input data at all. The selection of output components is shown in Figure 5.

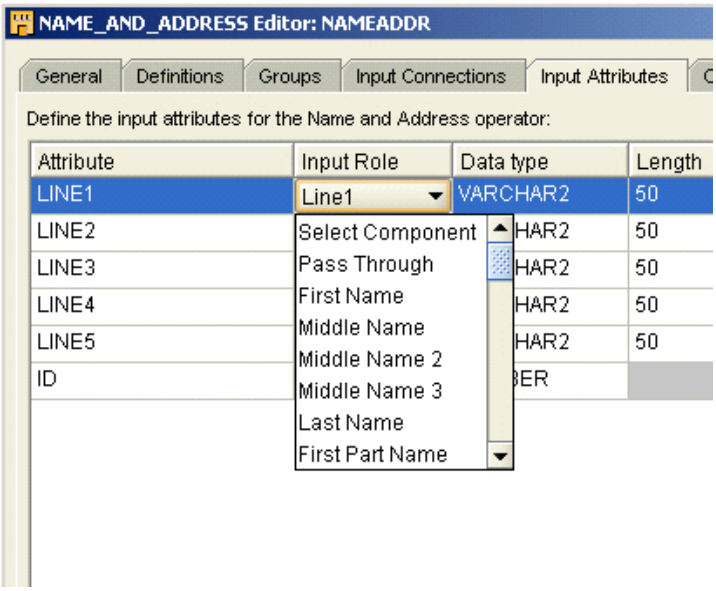

Figure 4. Assigning Input Roles in Name and Address Operator

At runtime, the deployed mapping containing name and address operator transparently accesses parsers and data provided by third-party vendors. These providers are companies that sell their software directly to customers, independent of Oracle. The full list of name and address providers certified by Oracle and their contact information is available on the Oracle Technology Network[1]. Note that the process of designing and running name and address transformations is always the same, regardless of which third-party provider (or even multiple providers) is used.

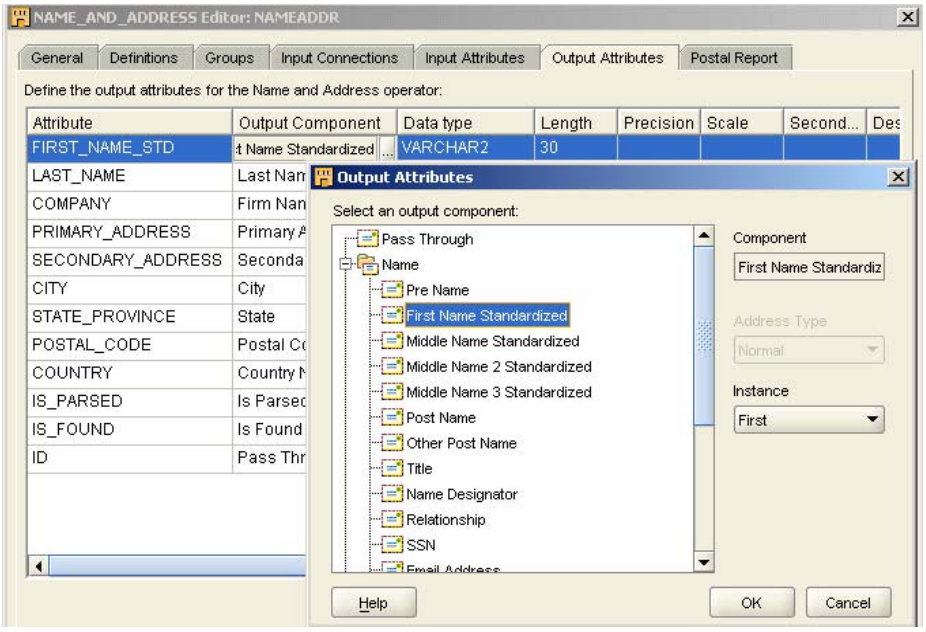

Figure 5. Selecting Output Components in Name and Address Operator

[1] http://www.oracle.com/technology/products/warehouse/htdocs/otn_partners.html

So after designing and running the name and address transformation, what do we get? The example comparing the input data before name and address cleansing to the output after is shown in Listing 1.

```
Input Roles      Input Line 1:  Mr. Bill Johnson Sr.
                 Input Line 2: Oracle Corporation
                 Input Line 3: 8500 Normandale Lake
                 Input Line 4: Suite 300
                 Input Line 5: Bloomington, Minnesota 55437

Output Components

Pre Name       : MR                          Primary Address: 8500 NORMANDALE
First Name     : BILL                        LAKE BLVD
First Name Std: WILLIAM                      Secondary Address: STE 300
Last Name:     : JOHNSON                     Postal Code       : 554373812
Post Name      : SR                          Address Type      : HD
Gender         : M                           Carrier Route     : C026
Is Parsed      : T                           Delivery Pt       : 99
Is Found       : T                           DPBC Check Digit  : 3
City Changed   : F                           Latitude          :44.856929
Postal Changed: T                            Longitude         :-093.356861
Firm           : ORACLE CORPORATION          MSA               :5120
Primary Range  : 8500                        MCD               :005
Primary Name   : NORMANDALE LAKE             Census ID         :0259035
Street type    : BLVD                        FIPS              :27053
```

Listing 1. Name and Address Cleansing

In name and address functionality, the following is new in the 10*g* Release 2 release:

- Additional output components supported by vendors are exposed
- A "pass-through" input role is added to allow source attributes to go through the name and address operator without being modified
- The code generated by the operator now uses a table function feature of the database, improving performance through parallelism

Matching and Merging of Data

Matching is the process of determining, through business rules, which records refer to the same logical data. Merging is the business rules-driven consolidation of the data from the matched set into a single record.

There are multiple uses of match-merge for different purposes that have their own terms:

- Deduplication is the process of matching and merging for the purpose of removing the duplicate records, especially customer-related records. This contributes to achieving the single view of the customer.
- House holding is the process of matching customers belonging to the same household, usually identified by the same address. Customer names are not merged; however, they are linked to the address that is stored once.

The benefit of house holding is the improved ability to understand and target customers.

- Record linking is the more generic instance of house holding. The records may need to be linked for purposes other than determining households, for example, linking business branches and subsidiaries to one parent entity.

It is also helpful to put the various terms pertaining to matching into perspective with Warehouse Builder. Warehouse Builder employs the elements of *fuzzy logic* and provides both *deterministic* and *probabilistic* matching algorithms:

- In general, fuzzy logic resembles human reasoning in its use of approximate information and uncertainty to generate decisions. In relation to matching, the term is used loosely to describe the approach that relies on rules that are imprecise rather than precise and operates on data with boundaries that are not sharply defined.
- Deterministic matching gives equal weight to the different types of information a record may contain. For example, a deterministic approach might place equal reliance on a match between the names on two records or a match between two birth dates.
- Probabilistic matching exploits the statistical probability that a match on particular items is more or less likely to indicate that the records are the same. For example, birth date information is subject to errors made by a mistake on a single digit, and the number of possible birth dates is relatively small. Names, in contrast, are more likely to be recognizable even if a single error is made. Probabilistic matching thus allows assigning appropriate weights to different attributes and then compares the total score to the threshold that defines a successful match.

Warehouse Builder accomplishes these tasks by using the match-merge operator with a wizard-driven interface. The match-merge operator comes with a powerful set of UI-controlled matching and merging algorithms. The user simply chooses the rule appropriate for the data being matched and the business requirement, followed by selection of parameters that control exact behavior of the rule. The whole operation is done entirely graphically with checkboxes, drop-down boxes, and scroll bars. It is possible to create very sophisticated match rules with only a mouse, i.e., not typing in any text. An example of a match rule creation is shown in Figure 6.

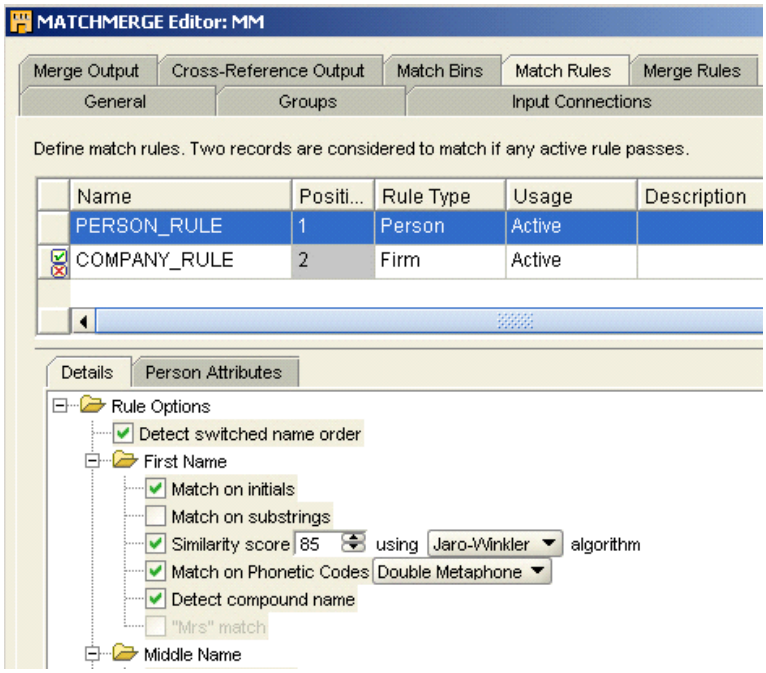

Figure 6. Creating a Match Rule in Match-Merge Operator

Similar to match rules, selecting the rule types and then supplying the parameters applicable to each rule create the merge rules. Figure 7 shows an example of how a merge rule is created.

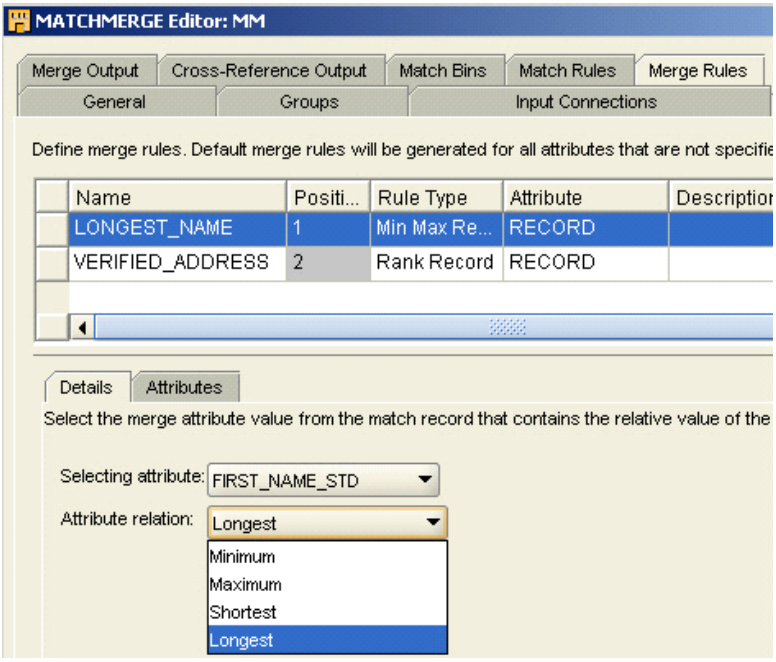

Figure 7. Creating a Merge Rule in Match-Merge Operator

With all the ease of use a graphical interface provides, the flexibility is not sacrificed. Match and merge rules both have custom rule types, allowing for a

PL/SQL-style procedural logic to be put in. Multiple rules can be called from other rules because they are represented as functions with input parameters.

At runtime Warehouse Builder executes the match-merge rules and transforms the input data, potentially containing duplicates and variations, into the consolidated output. Note that the implementation of match-merge operator in Warehouse Builder is entirely Oracle technology, with no reliance on a third-party.

In match-merge functionality, the following is new in 10*g* Release 2:

- More accurate and better performing phonetical algorithm—"Double Metaphone"
- More accurate and better performing similarity algorithm—"Jaro-Winkler"
- Optimization of matching "incoming" records against "existing" records through a "Match New Records Only" option, resulting in fewer matching comparisons performed

CONCLUSION

We have discussed the concepts of data quality, their importance in an enterprise as well as the distinct advantages of performing data quality processes inside a robust data integration product—Oracle Warehouse Builder.

Oracle Warehouse Builder addresses the challenges of data quality by offering data assessment, data cleansing, data integration, and data monitoring in one tool. It promotes the disciplined approach to data quality by making data quality processes easily available in the same development environment as the regular data transformation processes.

Warehouse Builder offers the flexibility and accuracy that data integration projects demand, yet at a low cost. By implementing data integration with Warehouse Builder, you will build a solid foundation for data quality in your enterprise.

Oracle Warehouse Builder 10*g* Release 2 – Transforming data into quality information
January 2007
Author: Jean-Pierre Dijcks

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
www.oracle.com