TDWI Monograph Series

Data Profiling:
Minimizing Risk in Data Management Projects

By Wayne W. Eckerson
Director of Research
The Data Warehousing Institute

November 2003

Table of Contents

Sponsored by DataFlux, a SAS Company

DataFlux offers complete, end-to-end data management solutions for organizations that want to realize more tangible, immediate value from their information assets. A wholly owned subsidiary of SAS, DataFlux helps companies improve the consistency, accuracy, and reliability of critical customer and business data. DataFlux solutions allow companies to build a solid information foundation that can enhance the effectiveness of data-driven applications, including customer relationship management (CRM), enterprise resource planning (ERP), data warehousing and database marketing.

To guide companies in the quest for better, more usable data, DataFlux relies on a proven data management methodology that encompasses data profiling, data quality, data integration and data augmentation. This comprehensive data management strategy allows DataFlux customers to analyze, cleanse, connect and enhance data. As the first step in the DataFlux methodology, data profiling is a crucial piece to any data-based initiative, allowing companies to understand the structure and relationships of existing information before attempting to improve the data. To learn more about DataFlux and its data management solutions and methodology, visit www.dataflux.com.

About the Author

WAYNE W. ECKERSON is the director of research for The Data Warehousing Institute (TDWI), the leading provider of high-quality, in-depth education and research services to data warehousing and business intelligence professionals worldwide. Eckerson oversees TDWI's Member publications, research reports, and consulting services.

Eckerson has written and spoken extensively on data warehousing and business intelligence since 1994. During the past two years, he has overseen the TDWI Report Series, which publishes in-depth reports on various topics in data warehousing and business intelligence. He also writes a regular column for Application Development Trends and the Business Intelligence Journal and is a regular contributor to DM Review. Eckerson frequently presents at industry conferences and is quoted in industry publications on data warehousing and business intelligence issues.

Besides his research duties, Eckerson coordinates TDWI's BI Strategies program, a quarterly educational event that brings together thought leaders and practitioners in the business intelligence field to discuss and share ideas about emerging trends and technologies. Prior to joining TDWI, Eckerson was a senior consultant at the Patricia Seybold Group and director of its Business Intelligence & Data Warehouse Service, which he launched in 1996.

About The Data Warehousing Institute

The Data Warehousing Institute (TDWI), a division of 101communications LLC, is the premier provider of in-depth, high-quality education and training in the business intelligence and data warehousing industry. TDWI offers quarterly educational conferences, regional seminars, onsite training, a professional membership, leadership awards, print and online publications, and a public and private (Members-only) Web site.

Since its founding in 1995, TDWI has trained thousands of data warehousing and business intelligence professionals in companies around the world. Its distinguished faculty consists of industry practitioners who share their knowledge about how to design, build, and manage data warehousing and business intelligence systems. TDWI currently conducts four training conferences each year in the U.S. as well as a dozen regional seminars. Its Membership includes business and technical professionals in medium-sized and large organizations around the world.

The Case for Data Profiling Tools

"Be prepared." It's more than just the Boy Scouts' motto; it's a rule we use to guide our personal and professional lives.

For example, most of us check the weather forecast in the morning to know how to dress for the day. Doctors and pharmacists analyze patient profiles before prescribing new medicines. Skilled entrepreneurs carefully evaluate the market before launching new products or services.

Code, Load, and Explode

Enter the World of Data. Oddly, it seems we leave this universal wisdom at the door when it comes to understanding data prior to integrating it with other sources of information. Often, we get halfway through a data integration project only to discover that our source data isn't what we thought it was! Character fields contain numbers; the gender field has five distinct values; invoices reference non-existent customers; sales orders have negative values; and so on.

We scratch our heads, pull in a subject matter expert, and apply new rules to fix this "bad" data. We run the process again and the same thing happens—we discover new data defects and again have to stop and rewrite our integration code. As schedules balloon, costs escalate, and tensions rise, we realize that we're stuck in an endless loop caused by undiscovered errors that pervade our data sources like mealy bugs in a grain tower. One data warehousing manager has described this pernicious cycle as "code, load, and explode."

Understanding Source Data Is a Major Challenge in Most Projects

Data Quality Obstacles. According to data warehousing professionals, the top two challenges in implementing ETL tools are "ensuring adequate data quality" and "understanding source data."[1] Most data warehousing professionals learn the hard way that fixing unanticipated defects and anomalies in source data is the primary reason for project delays, overruns, and failures.

Yet, the "code, load, and explode" phenomenon is not relegated to data warehousing; it afflicts any project that tries to integrate data from multiple systems. This includes customer relationship management systems, supply chain applications, and data migration initiatives. Most of the high-profile "failures" in these areas are due to organizations that underestimate problems associated with fixing poor data quality.

The Limits of Current Practices

Discovering Errors Too Late. Despite the growing publicity about the impact of poor quality data on data integration initiatives, most organizations take only minimal steps to understand—or "profile"—the data they want to integrate. In fact, almost two-thirds (62 percent) identify data quality problems *after the fact*—when users complain about errors in the new system, according to a TDWI survey. (See

[1] See "Evaluating ETL and Data Integration Platforms," TDWI Report Series, 2003, page 20. www.dw-institute.com/research.

Illustration 1.) When customers discover data defects in a system, they lose trust in it. If the quality problems are severe and persistent enough, the system can fail from lack of use.

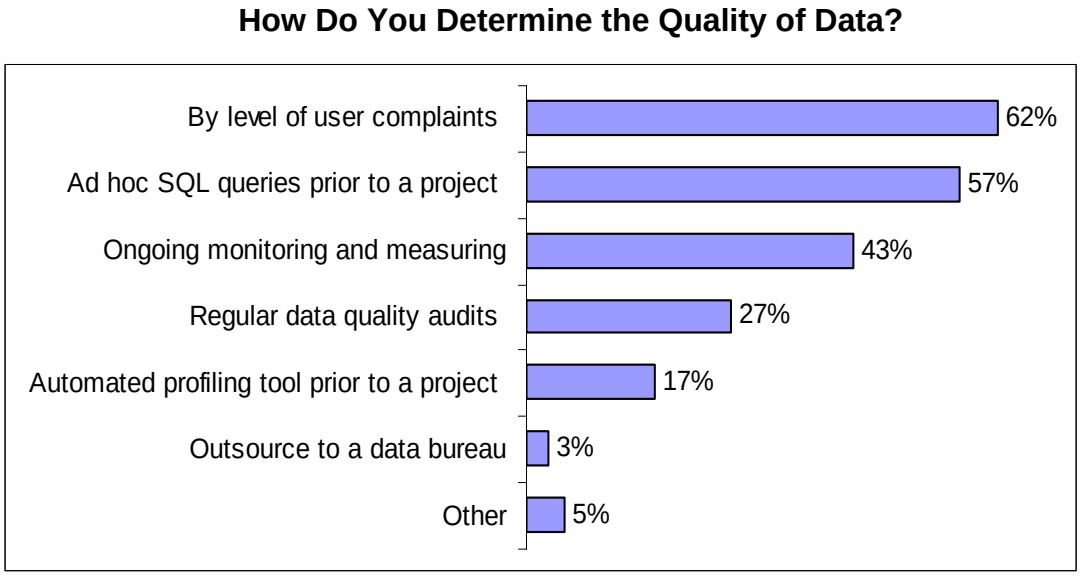


Illustration 1. Most organizations rely on user complaints and SQL queries to identify data defects. Less than one-fifth (17 percent) of organizations use data profiling tools. Based on 640 respondents from "Data Quality and the Bottom Line," TDWI Report Series, p. 21, www.dw-institute.com/research.

Unreliable Metadata. Part of the problem is that organizations rely too heavily on system catalogs (i.e., metadata) to provide an accurate description of the contents of source systems. These catalogs are notoriously unreliable because users often co-opt existing fields to enter new types of information.

For instance, telemarketers may insert cross-sell response codes into a birthday field because the marketing department wants to capture this data and there is no other place for them to enter it. In other cases, programmers who designed the system never documented its contents, or they left the company, taking their knowledge with them. And where catalogs exist, administrators rarely keep them up to date.

Also, data entry errors and poor validation routines permit errors to enter the system. Since many errors don't affect day-to-day operational processes, managers assume the data is accurate and consistent when it isn't. They only discover these "latent" errors when they try to integrate supposedly identical fields in two or more systems and nothing matches up.

Manual Methods Miss Errors

Manual Profiling. Another reason that analysts fail to discover data quality problems in advance is because they use manual methods to profile the contents of source systems. According to the TDWI survey above, a majority of organizations (57 percent) issue SQL queries to analyze source data. (See Illustration 1 above.) This usually means that they simply sample data in a few key fields to get a sense of what the data is like in those columns.

The hypothesis-driven approach takes too long and is error prone. It depends on the skill of the analyst to write accurate SQL and understand the business and its systems well enough to know what to look for in advance. This approach is like drilling for oil without a geologic map—there's little chance you'll find a deposit even if it's right underneath your feet!

The Answer: Automated Profiling Tools

Data Profiling Tools Scan Everything

To better understand source data and minimize risks in data integration projects, smart organizations are abandoning manual methods in favor of automated data profiling tools that take much of the guesswork out of finding and identifying problem data. These tools are now available from a variety of data quality vendors, including DataFlux Corporation, which is sponsoring this report.

Unlike user-generated SQL queries, data profiling tools scan every single record in every single column and table in a source system. And instead of just spitting out a list of data values, data profiling tools generate reports chock full of statistics and charts that make it easy to understand everything you ever wanted to know about your data. Thus, a data profiling tool is more likely to expose new or unanticipated structures and values in the data compared to manual profiling methods.

In general, data profiling tools enable users to better understand: (1) the structure and metadata of target systems, (2) the range and distribution of values in each column, and (3) relationships between columns in one or more tables, including primary/foreign key relationships. Also, many data profiling tools let analysts drill down from summary views of the data to actual data values to get a better sense of what records may or may not be skewing the summary results.

Benefits. The benefits of data profiling tools are enormous.

One user from a high-tech firm said a data profiling tool let them evaluate "100 percent of the data"—60 million records, 22 tables and 500 columns—in a matter of days compared to less than "half the data" in "three to four weeks" using manual methods. Moreover, the data profiling tool generated substantially more information about the data, accelerating the process of analyzing the data and creating appropriate data cleansing rules, he said.

100 Percent Accuracy and Completeness

Another company, the Automobile Club of Southern California (ACSC) attributes a data profiling tool with saving a data warehousing project that was plagued with data quality problems and mired in the "code, load, and explode" phenomenon. The data profiling tool let ACSC analyze two mainframe systems in six weeks with 100 percent accuracy, something it had not successfully done in 10 months using manual methods, according to its data warehousing manager.

Data Quality Audits. Besides analyzing data prior to integration projects, companies also use profiling tools to audit the cleanliness of existing databases. For example, a marketing manager at one company periodically "profiles" a marketing database to ensure it has no missing values or duplicate records. This regular audit gives her confidence to rely on the data to analyze and execute marketing campaigns.

Provide Evidence of Bad Processes

Exposes Inconsistent Business Processes. Another organization uses reports generated by a data profiling tool to convince executives to standardize key business processes. The reports depict data inconsistency across business units, where each unit uses different organizational structures and product codes. The reports provide top executives with concrete evidence that non-standard business processes will make it difficult to deliver a single version of the truth, a key strategic objective.

Delivers Trustworthy Data

In short, data profiling tools enable organizations to significantly reduce the time it takes to analyze data in a comprehensive and accurate fashion. It also can be used to regularly audit existing databases and demonstrate inconsistent business processes.

In many ways, a data profiling tool provides insurance for data integration projects: it mitigates the risks posed by poor quality data and gives you confidence that you can meet budgets and schedules as planned. More importantly, it enables you to deliver "trustworthy" data to end users so they can make accurate and effective decisions.

How to Use a Data Profiling Tool

Now that you know what data profiling is and what benefits it provides, you may wonder how it works. For example: who uses the tool? How do they use and analyze source data? How do they fix data problems that need correction?

Target Users

Profiling Tools Eliminate IT Intermediaries

Most data profiling tools are designed for business analysts—subject matter experts who understand the business and its data. The tools eliminate the need for IT administrators to serve as intermediaries between the data and subject matter experts. "I used to rely heavily on technical people to identify and fix problems in my database," says one marketing manager. "Now, I tell them the problems and likely fixes. It's empowering."

This contrasts with manual profiling methods in which IT people generate SQL queries and then turn the results over to business analysts for evaluation. However, if an organization needs to analyze dozens of sources with millions of records each, then a division of labor might make sense. In this case, an IT person may spend several days generating reports and identifying the most critical areas that a subject matter expert should focus on.

Analyzing Data: Three Steps

There are three basic steps that a subject matter expert might follow when analyzing data using a data profiling tool. These steps don't have to occur in this order and some steps may be skipped based on the nature of the data and its intended use.

1. Evaluate the data structure. Metadata identifies the number of records and null (i.e., blank) values in each field, data types (e.g., integers, characters), field length, unique values, patterns in the data, relationships to other fields, and so on. By first evaluating metadata, an analyst gains a general sense of the cleanliness and structure of the data and pinpoints potential problem areas.

2. Examine data values. Next, an analyst will want to analyze actual data values. This will help them determine whether the data adheres to business rules or not. If the data is deemed "invalid," the business analyst knows that he/she will need to devise rules to transform the data so that it meets corporate standards.

3. Analyze relationships. Finally, an analyst may want to understand the relationships of records within or between tables. For example, an analyst might validate a primary/foreign key relationship and discover that there are numerous customers who have purchased products (sales table) but haven't been billed for them because their customer numbers don't exist in the billing system.

Step 1: Evaluate Data Structure

Wendy, the Marketing Manager

To illustrate the use of a data profiling tool, we will use the example of Wendy, a marketing manager at a large U.S. publishing firm who wants to incorporate new customer names from a sales database into her marketing database, which she uses for telemarketing and direct mail campaigns, among other things.

Explore Metadata Anomalies. To begin, Wendy might point her profiling tool at the sales table within the sales system to view structural characteristics of every field in that table, including Address, City, Company, Phone Number, and so on. (See Illustration 2.)

Analyzing Field Structures

Profile Job 6 - dfPower Profile (Viewer)

File View Tools Help

Table: Sales
Total records: 5002

Field Name	Data Type	Primar...	Unique Count	Uniqueness	Pattern Count	Minimum Value	Maximum Value	Minimu...	Maxi...	Null Count	Blank Count	Actual Type	Count	Data Length	Mean
ADDRESS	VARCHAR	no	2389	75.79	1048	!0041 Nevad...	box 25-a-54	3	34	1850	0	string	5002	255 chars	(not applicable)
ADDRESS2	VARCHAR	no	4048	80.96	942	"12/25/76, T...	ZEPHYR COV...	5	37	2	0	string	5002	255 chars	(not applicable)
CITY	VARCHAR	no	391	59.15	52	ADAIRSVILLE	ZANESVILLE	4	19	4341	0	string	5002	255 chars	(not applicable)
COMPANY	VARCHAR	no	1293	39.47	840	1st Merit Bk	Yellow Stone ...	3	38	1726	0	string	5002	255 chars	(not applicable)
CONTACT	VARCHAR	no	3243	98.99	176	&rob Beckett	john doe	5	23	1726	0	string	5002	255 chars	(not applicable)
DATE	VARCHAR	no	323	9.86	4	1/1/1998	9/8/1997	8	10	1726	0	date	5002	255 chars	(not applicable)
GENDER	VARCHAR	no	0	0.00	0			(no data)	(no d...	5002	0	(no data)	5002	50 chars	(not applicable)
ID	INTEGER	yes	5002	100.00	(not applicable)	1	5002	(not a...	(not ...	0	(not appli...	integer	5002	10 chars	2501.5
MNFCTR NAME	VARCHAR	no	250	5.03	38	0ORTAL	ZOO DOO	1	14	27	0	string	5002	255 chars	(not applicable)
PHONE	VARCHAR	no	2890	66.88	7	(203)100-0263	999-7243	1	14	681	0	string	5002	255 chars	(not applicable)
Product Number	INTEGER	no	148	2.96	(not applicable)	0	499999	(not a...	(not ...	0	(not appli...	integer	5002	10 chars	83324.769092
Product Code	VARCHAR	no	4992	99.80	9	1002460360...	9988561501...	9	18	0	0	real number	5002	255 chars	(not applicable)
Product Code Type	VARCHAR	no	4	0.08	1	A	Y	1	1	0	0	string	5002	255 chars	(not applicable)
STATE	VARCHAR	no	64	4.55	1	AB	WY	2	2	3596	0	string	5002	255 chars	(not applicable)
TOTAL SALES	DOUBLE	no	164	3.28	(not applicable)	123	912341	(not a...	(not ...	0	(not appli...	double	5002	53 bit	6612.183395

Illustration 2. A data profiling tool captures accurate metadata by scanning actual data values in every column in a table.

Here, Wendy notices that the "Phone Number" field has seven different data patterns (sequences of alphanumeric characters), 681 null values, and a minimum field length of 1 and a maximum field length of 14. This analysis shows that besides the large number of missing and incomplete values, many phone numbers in the field don't adhere to a standard 10-digit U.S. telephone number. This may indicate that the data

may be bad, inaccurate, or filled with international numbers, thereby making it largely unusable in a telemarketing campaign.

Next, Wendy decides to run a pattern matching test to determine whether the seven data patterns represent valid telephone number sequences. (Her data profiling software uses a key in which "9" equals an integer, "A" equals an uppercase character, and "a" equals a lowercase character.) She discovers that one pattern sequence—"(999-999-9999"—has one record associated with it. Wendy surmises that a data entry operator accidentally typed a left parentheses before the phone number. She notes that she will have to create a rule to correct this pattern and record.

Step 2: Examine Data Values

Next, Wendy notices a pattern that ends in four capital letters, "9-999-999-AAAA". When Wendy drills into the actual data values, she realizes that the pattern refers to a toll-free vanity phone number (1-800-321-CARS). Since the number and pattern are legitimate, she knows she won't have to create a rule to "clean" or change this data.

Detection and Drilling to Detail

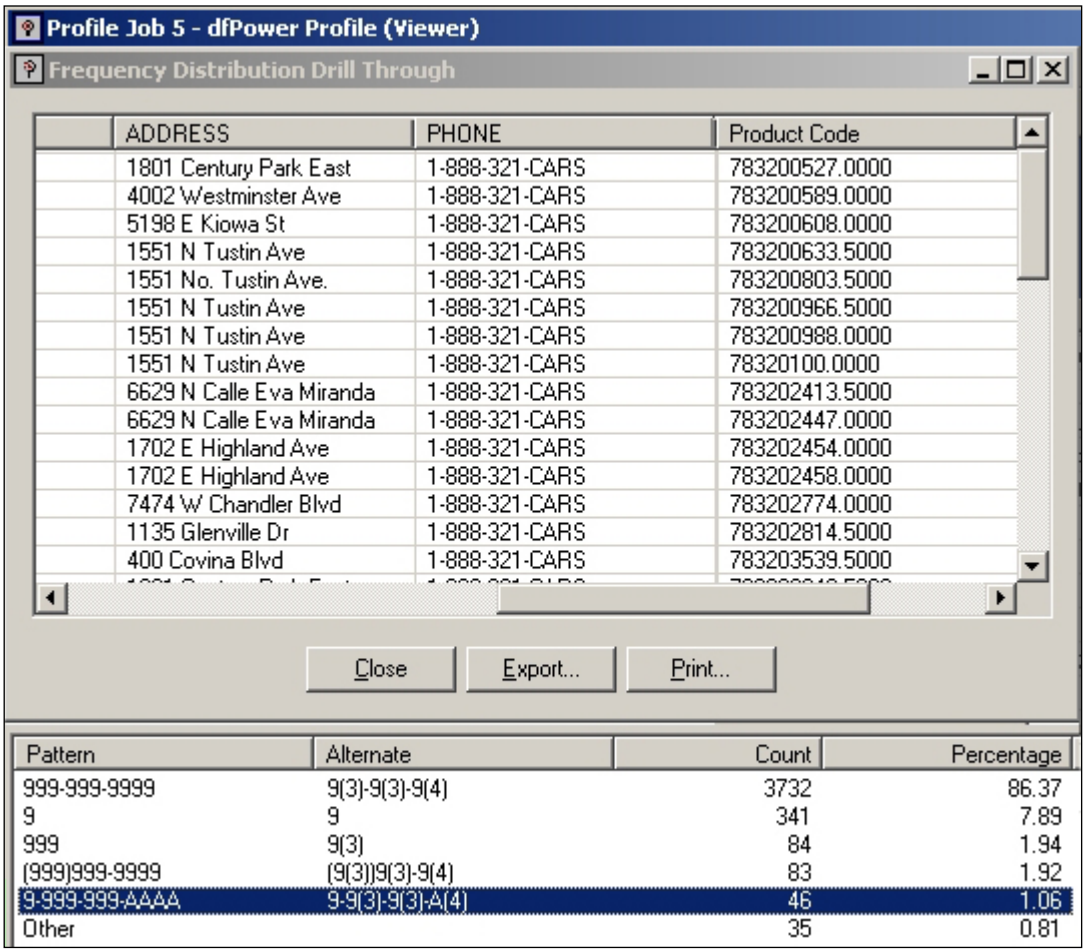
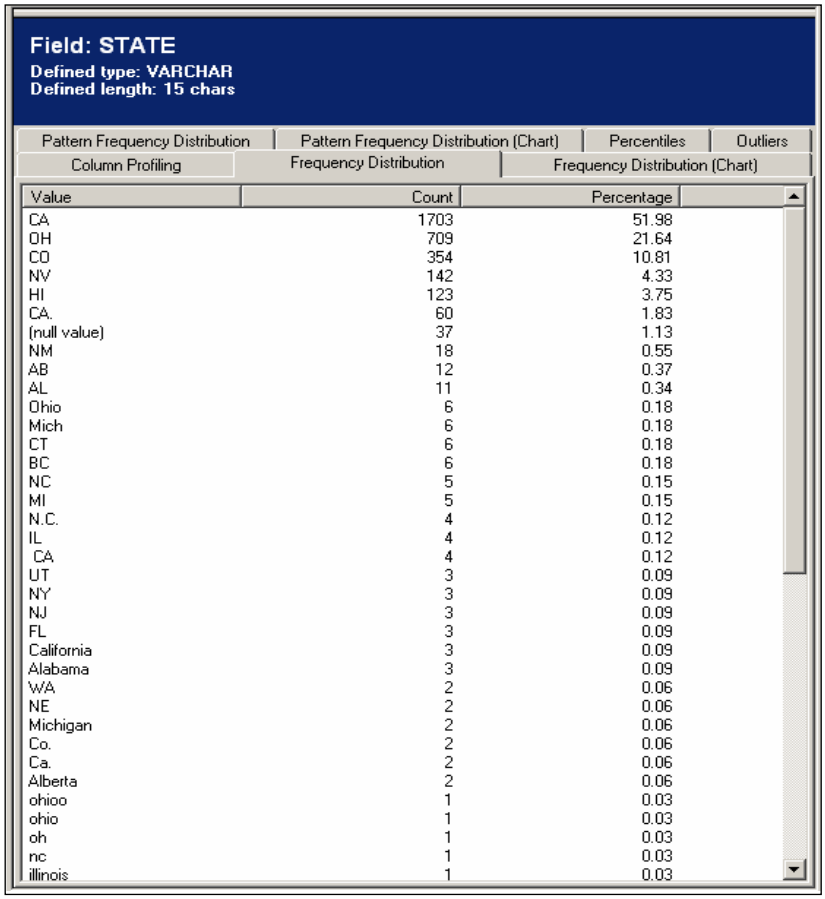


Illustration 3. Data profiling tools identify alphanumeric patterns in data and let you drill into a specific pattern to see the actual data values that conform to it.

Next, Wendy determines the cleanliness of the customers' addresses for use in direct mail campaigns. She drills into the "state" field to determine why there are 64 unique values for "state" when there are only 50 U.S. states! (See Illustration 2.) After running a frequency distribution analysis, she discovers that there are five different values for Ohio, four of which fail to conform to the standard postal code ("OH"): "Ohio" (occurs .18 percent of the time), "ohioo" (occurs .03 percent); "ohio" (.03 percent), and "oh" (.03 percent). (See Illustration 4.) Wendy will have to standardize all the non-conformant codes for Ohio and other states referenced in the database.

Step 2: Examine Data Values

Non-Standard State Codes



Value	Count	Percentage
CA	1703	51.98
OH	709	21.64
CO	354	10.81
NV	142	4.33
HI	123	3.75
CA.	60	1.83
(null value)	37	1.13
NM	18	0.55
AB	12	0.37
AL	11	0.34
Ohio	6	0.18
Mich	6	0.18
CT	6	0.18
BC	6	0.18
NC	5	0.15
MI	5	0.15
N.C.	4	0.12
IL	4	0.12
CA	4	0.12
UT	3	0.09
NY	3	0.09
NJ	3	0.09
FL	3	0.09
California	3	0.09
Alabama	3	0.09
WA	2	0.06
NE	2	0.06
Michigan	2	0.06
Co.	2	0.06
Ca.	2	0.06
Alberta	2	0.06
ohioo	1	0.03
ohio	1	0.03
oh	1	0.03
nc	1	0.03
illinois	1	0.03

Illustration 4. Wendy discovers that salespeople have failed to apply standard coding in the state field.

Step 3: Identify Relationships

Check Integrity. Finally, Wendy wants to discover how many customers in the sales database already exist in her customer file. To analyze the degree of overlap, she performs a redundancy analysis check, which compares records in two tables and depicts unique and overlapping records. (See Illustration 5.)

She discovers that there are more than a million customer records in the sales table that aren't in her database and 12,990 common or "redundant" records. She will then analyze the actual values to see whether the customer names are new or simply spelled differently than the ones in her database.

Redundancy Checking

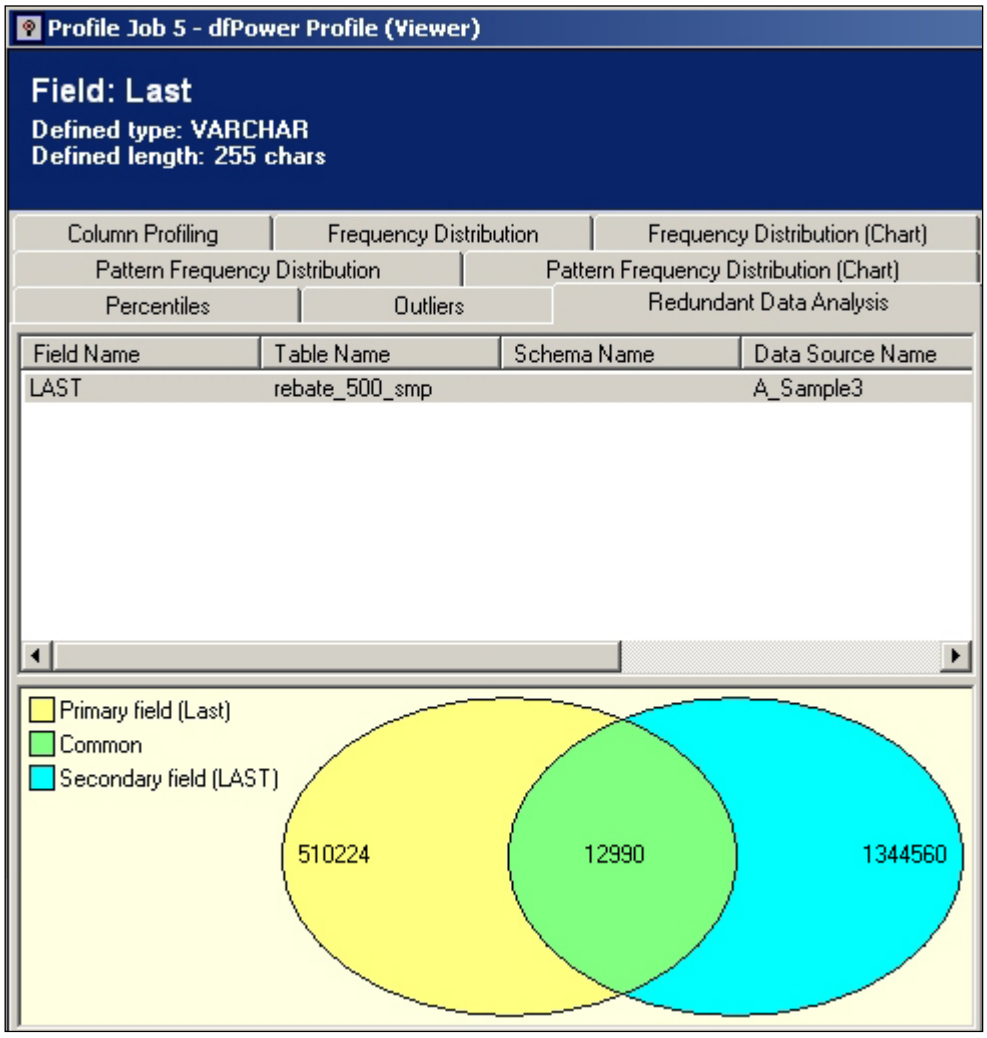


Illustration 5. Data profiling tools can identify the number of redundant records in different fields and drill into data values to view redundant or unique records. (Drill down not depicted here.)

Wendy's example demonstrates the power of data profiling tools to generate useful information quickly without having to know SQL or generate code. Business analysts simply point and click to discover interesting patterns and values in the data. Of course, data profiling tools are only as good as the analysts that use them. The tools won't help analysts take the right corrective actions unless they already have a deep knowledge of the business and its systems.

The Data Quality Process

Although data profiling is an important step in managing data quality, it is not the only one. Organizations that want to deliver high-quality data must not only analyze

the data, but clean and monitor it on an ongoing basis. In other words, data profiling tools must be implemented within the context of a total data quality program.

Profiling Is One of Nine Steps

Data Quality Methodology. The 2002 TDWI report titled *Data Quality and the Bottom Line* describes a nine-step methodology for ensuring high-quality data. The first three steps focus on organizational issues, such as launching a data quality program, developing a project plan, and building a data quality team. The second three steps involve assessments—analyzing business practices, data architecture, and source data (i.e., data profiling). The final steps entail taking action based on knowledge gained in prior steps: clean the data, change business processes, and monitor data quality.

Types of Data Quality Software. Many data quality tools today support several steps in this process, including data profiling as well as data validation and cleansing, data standardization and matching, and data augmentation. Some even bolt data quality tools into data warehousing tools that extract, transform, and load (ETL) data into a target system.

Creating Rules from Profiling Tasks

Defining and Creating Data Cleansing Rules

To return to Wendy's example in the previous section, Wendy's analysis might lead her to define two standard patterns or rules that all U.S. phone numbers: 999-999-9999 and 999-999-AAAA must adhere to. Then, she defines how all non-conforming patterns should be transformed to conform to the rules. She will do the same for the non-conforming values she discovered in the "state" and "company/organization" fields.

Today, Wendy and other business analysts typically write the rules in plain English and pass them to a system analyst who codes them in SQL. If the company owns data cleansing software, the analyst might define the rules in the tool's standardization or cleansing module. (See Illustration 6.)

Once the rules are defined, Wendy or the system analyst periodically applies them to her marketing database to catch any errors that have been introduced since the file was last "cleaned."

Data Quality Suites Should Have Integrated Metadata

Tool Integration. Since managing data quality is an integrated, multi-step process, the best data quality tools provide a high-degree of integration among different toolsets or modules and with external applications.

For example, the best data profiling tools pass metadata to data cleansing tools so rules can be quickly applied without duplicating effort. (Ideally, users should be able to write rules within a data profiling tool as they go along. However, few data profiling tools support rules creation today although some are working on it.) This level of integration is best achieved if data profiling and cleansing tools are offered by the same vendor.

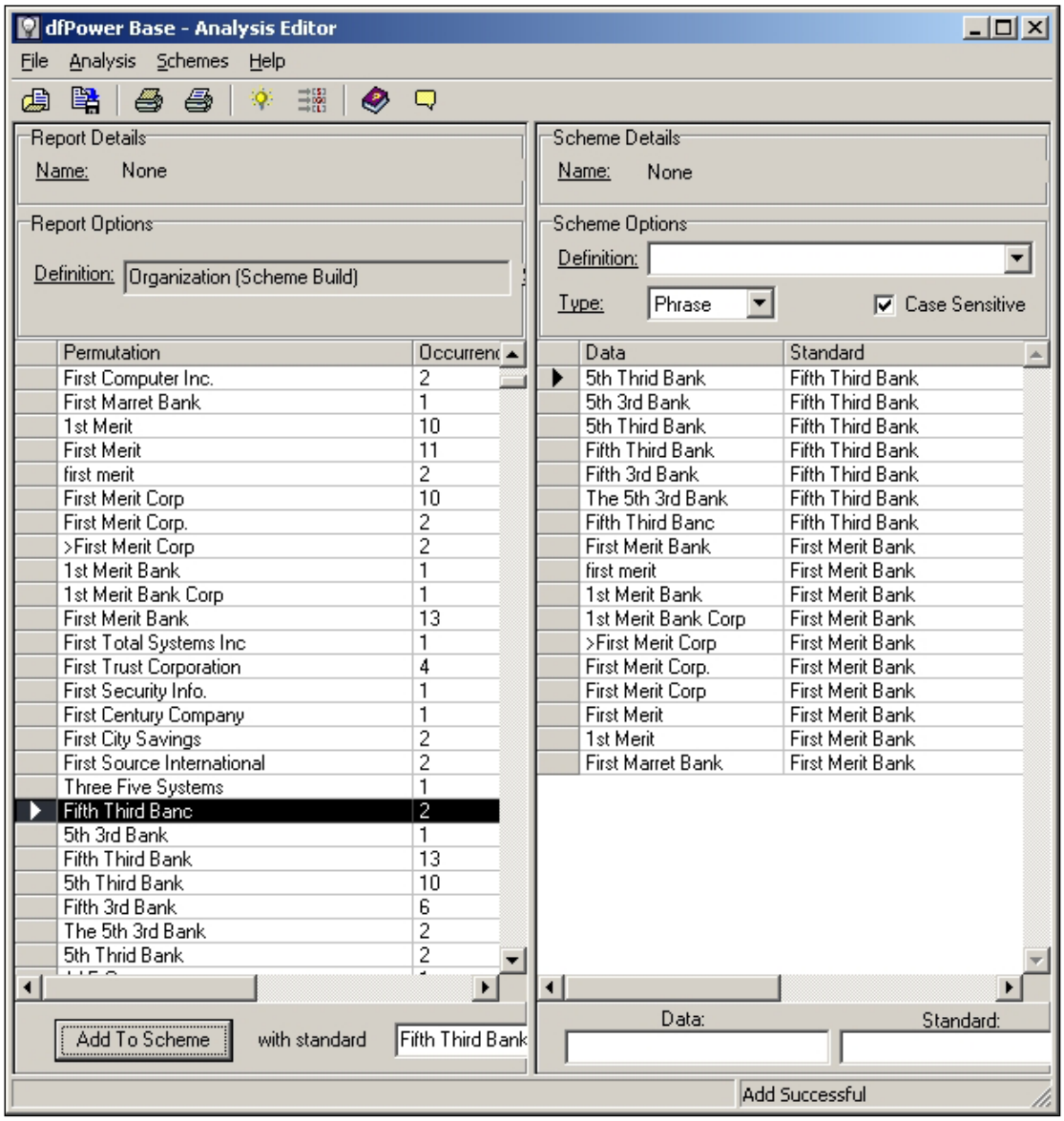


Illustration 6. This diagram depicts how a tool would automatically generate data standardization rules that an analyst would use to transform various spellings in a "company/organization" field to a standard code.

Integrating with ETL. In a data warehousing project, data profiling and cleansing processes should be integrated with ETL tools. The profiling tools pass metadata to cleansing tools which fix errors and spit out a clean file to the ETL tool for final processing. Again, this level of integration requires close coordination between data quality and ETL tools and is best achieved if a vendor owns both sets of tools.

Summary. Although data profiling is one tool within a data quality toolbox and one step within a data quality management program, it plays a critical role since it

provides the knowledge required to support ongoing data cleaning and monitoring efforts.

Evaluating Data Profiling Tools

Now that you better understand how to analyze, clean, and monitor data, you may wonder which data profiling tool you should purchase. The good news is that today there are more data profiling tools on the market than several years ago. However, not all data profiling tools work the same way, so you will need to carefully evaluate their capabilities.

Ideally, you should prototype a tool by running it against a subset of your data that you know well. This will help you understand whether the tool captures key characteristics in the data and how easy it is to use.

To help you in your quest, here are a few guidelines to help you find the data profiling tool that is right for your organization.

1. Ease of use. Can a business analyst use the tool without having to rely on someone from IT to run the software and generate the reports? The use of wizards, side-by-side panels that display summary and detail data, and visual workflows may indicate that the tool is appropriate for business users.

2. Collaboration. Good profiling tools educate the wider community about the nature of data that comprises a new or existing application. That means the tools should be able to generate easy-to-read reports in various formats (e.g., HTML, Word, Excel) that quickly communicate data quality issues to relevant constituents.

3. Direct Connectivity to Sources. To ensure adequate scalability, data profiling tools should connect directly to source systems rather than create copies of the data, which may take hours or more and chew up valuable processing power. Tools that connect directly to data sources also ensure that analysts are accessing the most up-to-date data values.

4. Generate Rules for Cleansing. It is not enough to identify data defects; analysts must be able to create rules to fix the problems. Ideally, analysts should be able to create rules within a data profiling tool, which can then pass the rules to a data cleansing tool for processing. In other words, data profiling tools should be tightly integrated with data cleansing tools to make the most efficient use of analysts' time.

5. Integration with Third-Party Applications. To automate routine data integration or migration processes, such as data warehousing loads, data profiling tools (and data cleansing tools) should be integrated with third-party applications, such as ETL and data integration tools. The profiling and cleansing jobs should be initiated and managed by these third-party applications.

6. Offer Broad Functionality. Data profiling tools should offer a comprehensive set of functionality for analyzing data structures, generating statistics about column values, and mapping dependencies within or among tables.

7. Price. Organizations are reluctant to purchase multiple tools to perform data integration projects. One reason many organizations rely on manual profiling is that they can't justify making a capital expenditure when they can use internal resources (i.e., programmers or systems analysts) to profile data manually. Given this context, data profiling tools need to provide excellent value for the investment. Ideally, they come bundled for free within a suite of data quality or data integration tools, so perceived costs are minimal.

Summary. There are likely many more evaluation criteria that you will want to apply when searching for a data profiling tool. The above criteria, however, provide a useful starting point.

Conclusion

As we have seen, data profiling tools promise to minimize the risks that undermine many data integration projects: unanticipated data defects in source systems. Data profiling tools provide insurance against project delays and cost overruns due to errors or inconsistencies in the data. Armed with a data profiling tool, you can have greater confidence that you can meet project deadlines and deliver trustworthy data to your customers.

Data profiling tools are designed to be used by business analysts, not IT administrators. The graphical tools make it easy for business analysts to scan 100 percent of the data with 100 percent accuracy. The tools generate reports that make it easy to identify structural and integrity anomalies in the data and drill into data values to verify the validity of those structures and relationships.

Since their inception, data profiling tools have been used in a stand-alone mode in which the process of analyzing data was divorced from the process of creating and applying rules to fix the data. During the next several years, we will see data profiling processes integrated more tightly within data cleansing and ETL tools. This will consolidate and streamline the process of detecting and fixing data quality problems in target databases and enable organizations to focus more on business issues than data issues.