



Data Profiling

The Foundation for Data Management

Prepared by:
DataFlux Corporation

Data Profiling: The Foundation for Data Management

Contents

Executive summary.....	1
Through the looking glass: What is data profiling?	2
The problems with data.....	3
Structure discovery: Understanding data patterns and metadata	4
Data discovery: Business rule validation and data completeness discovery	7
Relationship discovery: Data redundancy and similarity discovery	10
Data profiling in practice.....	11
The Four Building Blocks of Data Management.....	13
Data profiling: Understand your data	13
Data quality: Standardize, validate and verify your data	13
Data integration: Combine data from multiple sources.....	13
Data augmentation: Enhance and enrich data	14
Conclusion	14
Getting started.....	15

Figures

Figure 1: Metadata report on a character field.	4
Figure 2: Pattern frequency report for telephone numbers.	5
Figure 3: Statistics on a column of loan data.	6
Figure 4: Frequency distribution on state data.	8
Figure 5: Outlier report on product weight.	9
Figure 6: Results of primary key/foreign key analysis.	11

Executive summary

Current data quality problems cost U.S. businesses more than \$600 billion per year.¹

Not so long ago, the way to become a market leader was to have the right product at the right time. But the industrial and technological revolutions of the last century created a market with many companies offering the same products. The path to market leadership required companies to design and manufacture products cheaper, better and faster. And as more businesses entered the market with lower barriers of entry, products showed fewer distinguishing characteristics that define a market leader—resulting in a more commodity-based marketplace.

With narrow margins and constant competition, organizations realized that a better product no longer guaranteed success. In the last 10 years, organizations have concentrated on the optimization of processes to bolster success. Profits are as much the result of controlling expenses as from generating additional revenue.

To realize significant savings from expenses, companies throughout the world are implementing two primary enterprise applications: enterprise resource planning (ERP) and customer relationship management (CRM). Each of these applications focus on driving increased efficiencies from core business processes, with ERP systems focused on holding expenses “in check,” and CRM systems working to build more profitable relationships with customers.

Successfully implemented, ERP systems help companies optimize their operational processes and help reduce processing costs. On the opportunistic, customer-facing side of profit-seeking, companies realize that customers are expensive to acquire and maintain, leading to the deployment of CRM systems. At the same time, organizations have developed data warehouses in an effort to make more strategic decisions across the enterprise—spending less and saving more whenever possible.

But a new age in enterprise management is here. The very foundation of ERP and CRM systems is the data that drives these implementations. Without valid corporate information, enterprise-wide applications can only function at a “garbage in, garbage out” level. To be successful, companies need high-quality data on inventory, supplies, customers, vendors and other vital enterprise information. Or their ERP or CRM implementations are doomed to fail.

In their most recent Global Data Management Survey, PricewaterhouseCoopers writes that, “The new economy is the data economy.” The survey adds, “Companies are entering a crucial phase of the data age without full control or knowledge of the one asset most fundamental to their success—data.”² The successful organizations of tomorrow are the ones that recognize that data (or, more accurately, the successful management of corporate data assets) will determine the market leaders in the future.

If your data is going to make you a market leader, it must be consistent, accurate and reliable. Achieving this level of prosperity requires solid data management practices including data profiling, data quality, data integration and data augmentation. And any data management initiative begins with profiling, where you analyze the current state of your data—and begin to build a plan to improve your information. This paper discusses data profiling in detail, what it is and how it can be deployed at your organization. The paper will also look at how data profiling fits into the broader data management process of your organization.

¹ “Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data.” The Data Warehousing Institute. Report Series 2002.

² Global Data Management Survey. PricewaterhouseCoopers. 2001.

Through the looking glass: What is data profiling?

**Where shall I begin, please your Majesty?
“Begin at the beginning,” the King said gravely.**

Lewis Carroll
Alice’s Adventures in Wonderland

It might seem odd to introduce a section on data profiling with a quote from *Alice’s Adventures in Wonderland*. Yet, many organizations find that their data is as confusing and disorienting as the whimsical world of Wonderland. Consider the following quotes from the book. Do they describe Alice’s level of confusion as she faced an onslaught of unintelligible answers and nonsensical conclusions? Or do they summarize how you feel about your organization’s data?

- “I don’t believe there’s an atom of meaning in it.”
- “This is very important... Unimportant, of course, I meant – important, unimportant, unimportant, important.”
- “I think you might do something better with your time than waste it asking riddles that have no answers.”
- “Curiouser and curiouser”

Many business and IT managers face the same problems when sifting through corporate data. Often, organizations do not—and worse yet, cannot—make the best decision because they can’t get access to the right data. And just as often, a decision is made based on data that is faulty or untrustworthy. But regardless of the state of the information within your enterprise, the King in *Alice’s Adventures in Wonderland* had the right idea: “Begin at the beginning.”

Data profiling is a fundamental, yet often overlooked, step that should begin every data-driven initiative. Every ERP implementation, every CRM deployment, every data warehouse development and every application rewrite should start with data profiling.

Industry estimates for ERP and data warehouse implementations show these projects fail or go over-budget 65-75% of the time. In almost every instance, project failures, cost overruns and long implementation cycles are due to the same problem—a fundamental misunderstanding about the quality, meaning or completeness of the data that is essential to the initiative. These are problems that should be identified and corrected prior to beginning the project. And by identifying data quality issues at the front-end of a data driven project, you can drastically reduce the risk of project failure.

To address information challenges at the outset, data profiling provides a proactive approach to understanding your data. Data profiling, also called data discovery or data auditing, is specifically about discovering the data available in your organization and the characteristics of that data. Data profiling is a critical diagnostic phase that arms you with information about the quality of your data. This information is essential in helping you determine not only what data is available in your organization, but how valid and usable that data is.

Profiling your data is based on the same principle that your mechanic uses when you take your car to the shop. If you take your car in and tell the mechanic that the car has trouble starting, the mechanic doesn’t say, “Well, we’d better change the timing belt.” The mechanic goes through a series of diagnostic steps to determine the problem: he checks the battery, checks the fluids, tests the spark plugs and checks the timing. After a thorough diagnostic review, the mechanic has validated the reliability of different parts of the engine, and he is ready to move forward with the needed changes.

Starting a data-driven initiative (ERP system, CRM system, data warehouse, database consolidation, etc.) without first understanding the data is like fixing a car without understanding the problems. You may get lucky, but chances are you will waste time and money doing work that is neither complete nor productive. And you are likely to become just another failure statistic for ERP, CRM or data warehousing implementations.

With proper data profiling methodologies, you can also gain valuable insight into your business processes and refine these procedures over time. For instance, a data analyst conducts a profiling routine on a CRM database and finds that over 50% of the product information is inaccurate, incorrect or outside of the standard parameters. The data analyst can then go to other departments, such as sales and business development, to find out how product data is entered into the system—and find ways to refine and enhance this process.

To help you “begin at the beginning,” data profiling encompasses many techniques and processes that can be grouped into three major categories:

- Structure discovery – Does your data match the corresponding metadata? Do the patterns of the data match expected patterns? Does the data adhere to appropriate uniqueness and null value rules?
- Data discovery – Are the data values complete, accurate and unambiguous?
- Relationship discovery – Does the data adhere to specified required key relationships across columns and tables? Are there inferred relationships across columns, tables or databases? Is there redundant data?

The next section will look in detail at the structure discovery, data discovery and relationship discovery routines—and how you can use these profiling techniques to better understand your data.

The problems with data

Data problems abound in most organizations. These problems can stymie your data initiatives—data inconsistencies, anomalies, missing data, duplicated data, data that does not meet business rules, orphaned data and many more problems. Before you begin any project initiative, you need to know basic information in support of that initiative:

- Do you trust the quality of the data you are using in this initiative?
- Will the existing data support the needed functionality?
- Is the data you are using complete enough to populate the needed data repository?
- Does the data for this initiative conform to the expected business rules and structure rules?
- Can you access the needed data sources for your initiative?

Engaging in any data initiative without a clear understanding of these issues will lead to large development and cost overruns or potential project failures. GartnerGroup estimates that through 2005, “more than 50 percent of business intelligence and customer relationship management deployments will suffer limited acceptance, if not outright failure, due to lack of attention to data quality issues.”³ From a real-world perspective, the effect can be incredibly costly; one company spent over \$100,000 in labor costs identifying and correcting 111 different spellings of the company AT&T.

Because companies rely on data that is inconsistent, inaccurate and unreliable, large-scale implementations are ripe for failure or cost overruns. More disturbing, the organizations usually

³ “A Strategic Approach to Improving Data Quality,” by Ted Friedman. GartnerGroup. June 19, 2002.

do not understand the magnitude of the problem or the impact that the problems have on their bottom line. Data problems within your organization can lead to lost sales and wasted expenses. Poor decisions. Sub-standard customer relations. And ultimately, failed businesses.

Data profiling is the first step to help you diagnose—and fix—the problem. Now, let’s look in more detail at the types of discovery techniques you should consider during the data profiling process.

Structure discovery: Understanding data patterns and metadata

By examining complete columns or tables of data, structure analysis helps you determine whether or not the data in that column or table is consistent and meets the expectations that you have for the data. There are many techniques that can validate the adherence of data to expected formats. Any one of these techniques provides insight about the validity of the data.

Let’s look at some of the common structure analysis techniques and the potential problems that these techniques can uncover. We compare metadata to the actual data, consider basic patterns and formats of the data and examine how basic statistics can uncover potential data anomalies.

Validation with metadata: Most data has some associated metadata or description of the characteristics of the data. It may be in the form of a COBOL copy book, a relational database repository, a data model or a text file. The metadata will contain information that indicates data type, field length, whether the data should be unique and if a field can be missing or null.

This metadata is supposed to describe the data that is present in the table or column. Data profiling tools scan the data to infer this same type of information. Often, the data and the metadata do not agree, causing far-reaching implications for your data management efforts. For example, consider a 10 million row field with a field length of 255 characters. If the longest data element in the data is 200 characters, the field length is longer than required, and you are wasting 550MB of disk space. Missing values in a field that should not have missing values can cause joins to fail and reports to yield erroneous results. Figure 1 shows the types of information that a typical metadata report on a character field (here, product description) should contain.

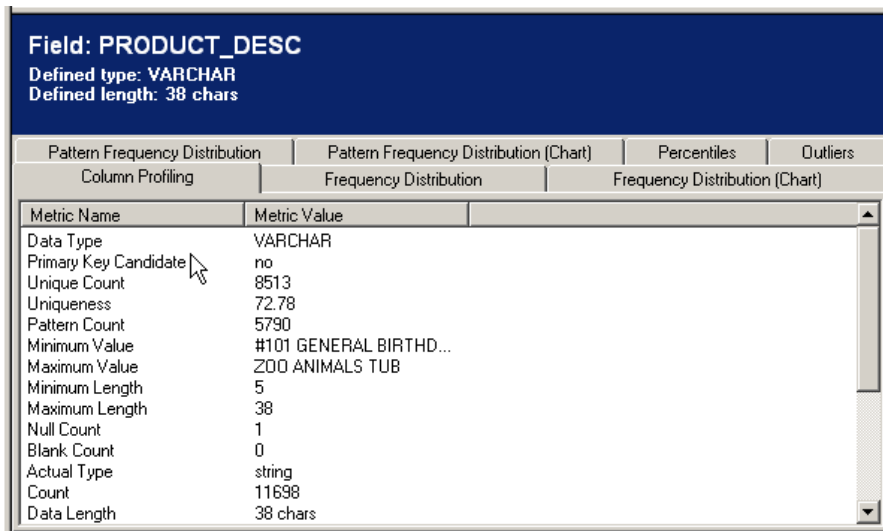


Figure 1: Metadata report on a character field.

Metadata analysis helps determine if the data matches the expectations of the developer when the data files were created. Has the data migrated from its initial intention over time? Has the purpose, meaning and content of the data been intentionally altered since it was first created? By answering these questions, you can make decisions about how to use the data moving forward.

Pattern matching: Typically, pattern matching is used to determine if the data values in a field are in the expected format. This technique can quickly validate that the data in a field is consistent across the data source—and that information is consistent with your expectations. For example, pattern matching would analyze if a phone number field contains all phone numbers. Or if a social security field contains all social security numbers. Pattern matching will also tell you if a field is all numeric, if a field has consistent lengths and other format-specific information about the data.

As an example, consider a pattern report for North American phone numbers. There are many valid phone number formats, but all valid formats consist of three sets of numbers (three numbers for area code, three numbers for exchange, four numbers for station). These sets of numbers may or may not be separated by a space or special character. Valid patterns might include:

- 9999999999
- (999) 999-9999
- 999-999-9999
- 999-999-AAAA
- 999-999-Aaaa

In these examples, “9” represents any digit, “A” represents any upper case alpha (letter) character and “a” represents any lower case alpha character. Now, consider the following pattern report on a phone number field.

Field: Phone			
Defined type: VARCHAR			
Defined length: 15 chars			
Column Profiling	Frequency Distribution	Frequency Distribution (Chart)	
Pattern Frequency Distribution	Pattern Frequency Distribution (Chart)	Percentiles	Outliers
Pattern	Count	Percentage	
999-999-9999	3166	96.73	
(999)999-9999	42	1.28	
(999) 999-9999	34	1.04	
999 99 9999 999	20	0.61	
999 999 9999	5	0.15	
999-999-AAAA	2	0.06	
9-999-999-9999	2	0.06	
a	1	0.03	
99 99 9999 999	1	0.03	

Figure 2: Pattern frequency report for telephone numbers.

The majority of the phone data in this field contains valid phone numbers for North America. There are, however, some data entries that do not match a valid phone pattern. A data profiling tool will let you drill through a report like this to view the underlying data or generate a report containing the drill-down subset of data to help you correct those records.

Basic Statistics: You can learn a lot about your data just by reviewing some basic statistics about the data. This is true for all types of data, especially numeric data. Reviewing statistics such as minimum/maximum values, mean, median, mode and standard deviation can give you insight into the validity of the data. Figure 3 shows statistical data about personal home loan values from a financial organization. Personal home loans normally range from \$20,000 to \$1,000,000. A loan database with incorrect loan amounts can lead to many problems, from poor analysis results to incorrect billing of the loan customer. Let’s take a look at some basic statistics from a loan amount column in the loan database.

Field: LoanAmount	
Defined type: double	
Defined length: 53 bit	
Pattern Frequency Distribution	Pattern Frequency Distribution (Chart)
Column Profiling	Frequency Distribution
	Frequency Distribution (Chart)
Metric Name	Metric Value
Data Type	double
Primary Key Candidate	no
Unique Count	1140
Uniqueness	70.11
Pattern Count	(not applicable)
Minimum Value	-223000
Maximum Value	9999999
Minimum Length	(not applicable)
Maximum Length	(not applicable)
Null Count	2
Blank Count	(not applicable)
Actual Type	double
Count	1628
Data Length	53 bit
Mean	114348.170972
Median	4888499.5
Mode	0
Non-null Count	1626
Nullable	YES
Ordinal Position	7
Decimal Places	0
Standard Deviation	429438.361236
Standard Error	10649.778281

Figure 3: Statistics on a column of loan data.

This report uncovers many potential problems with the loan amounts (see arrows above). The minimum value of a loan is a negative value. The maximum value for a loan is \$9,999,999. There are two loans with missing values (Null Count). The median and standard deviations are unexpectedly large numbers. All of these indicate potential problems for a personal home loan data file.

Basic statistics give you a snapshot of an entire data field. As new data is entered, tracking basic statistics over time will give you insight into the characteristics of new data that enters your systems. Checking basic statistics of new data prior to entering it into the system can alert you to inconsistent information and help prevent adding problematic data to a data source.

Metadata analysis, pattern analysis and basic statistics are a few of the techniques that profiling tools use to discover potential structure problems in a data file. There are a variety of reasons that these problems appear in files. Many problems are caused by incorrectly entering data into a field (which is most likely the source of the negative value in the home loan data). Some problems occur because a correct value was unknown and a default or fabricated value is used (potentially the origin of the \$9,999,999 home loan).

Other structure problems are the result of legacy data sources that are still in use or have been migrated to a new application. Often during the data creation process for older mainframe systems, programmers and database administrators designed shortcuts and encodings that are no longer used or understood. IT staff would overload a particular field for different purposes. Structure analysis can help uncover many of these issues.

Data discovery: Business rule validation and data completeness discovery

After you analyze entire tables or columns of data with the structure discovery steps, you need to look more closely at each of the individual elements. Structure discovery provides a broad sweep across your data and often points to problem areas that need further investigation. Data discovery digs deeper and helps you determine which data values are inaccurate, incomplete or ambiguous.

Data discovery techniques use matching technology to uncover non-standard data, frequency counts and outlier detection to find data elements that don't make sense, and data verification according to specific business rules that may be unique to your organization. Let's look in more detail at each of these techniques.

Standardization: Unfortunately, data can be represented ambiguously. Data in an organization often comes from a variety of sources: different departments, different data entry clerks and different partners. This is often the root of an organization's data quality issues. If multiple permutations of a piece of data exist, then every query or summation report generated by that data must account for each and every instance of these multiple permutations. Otherwise, you can miss important data points, which can impact the output of future processes. For example:

- "IBM," "Int. Business Machines," "I.B.M.," "ibm," and "Intl Bus Machines" all represent the same company.
- Does the company "GM" in a database represent "General Motors" or "General Mills?"
- "Brass Screw," "Screw: Brass," "Bras Screw," and "SCRW BRSS" all represent the same product.
- "100 E Main Str," "100 East Main Str.," "100 East Main," and "100 Main Street" all represent the same address.

Each of these values all have the same meaning, but they are represented differently. The analytical and operational problems of this non-standard data can be very costly, as you cannot get a true picture of the customers, businesses or items in your data sources. For instance, a life insurance company may want to determine the top ten companies that employ their policyholders in a given geographic region. With this information, the company can tailor policies to those specific companies. If the employer field in the data source has the same company entered in several different ways, inaccurate aggregation results are likely.

In addition, consider a marketing campaign that personalizes its communication based on a household profile. If there are a number of profiles for customers at the same address, the addresses are represented inconsistently. Variations in addresses can have a nightmare effect on highly-targeted campaigns, causing improper personalization or the creation of too many generic communication pieces. These inefficiencies waste time and money both on material production and creative efforts of the group, while alienating customers by ineffectively marketing to their preferences.

While these are simple data inconsistency examples, these and other similar situations are endemic to databases worldwide. Fortunately, data profiling tools can discover these inconsistencies, providing a blueprint for data quality technology to address and fix these problems.

Frequency Counts and Outliers: When there are hundreds or even thousands of records that need to be profiled, it may be possible for a business analyst to scan the file and look for values that don't look right. But, as the data grows, this quickly becomes an immense task. Many organizations spend hundreds of thousands of dollars to pay for manual validation of data. This is not only expensive and time-consuming, but manual data profiling is inaccurate and prone to human error.

Frequency counts and outlier detection give you techniques that can limit the amount of business analyst fault detection required. In essence, these techniques highlight the data values that need further investigation. You can gain insight into the data values themselves, identify data values that may be considered incorrect and drilldown to the data to make a more in-depth determination about the data. Consider the following frequency distribution of a field containing state and province information.

Field: STATE		
Defined type: VARCHAR		
Defined length: 15 chars		
Pattern Frequency Distribution	Pattern Frequency Distribution (Chart)	Percentiles
Column Profiling	Frequency Distribution	Frequency Distribution (Chart)
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

Figure 4: Frequency distribution on state data.

The frequency distribution shows a number of correct state entries. But, the report also shows data that needs to be corrected. Incorrect state spellings, invalid state abbreviations and multiple representations of states can all cause problems. California is represented as “CA,” “CA.,” “Ca.,” and “California.” Non-standard representations will have an impact any time you are trying to do state-level analysis. The invalid state entries may prevent you from contacting certain individuals; the missing state values make communication even more problematic.

Outlier detection also helps you pinpoint problem data. Whereas frequency count looks at how values are related according to data occurrences, outlier detection examines the (hopefully) few data values that are remarkably different from other values. Outliers show you the highest and lowest values for a set of data. This technique is useful for both numeric and character data.

Consider the following outlier report (showing the 10 minimum and 10 maximum values for the field). In Figure 5, the field is product weight, measured in ounces, for individual-serving microwaveable meals. A business analyst would understand that the valid weights are between 16 and 80 ounces.

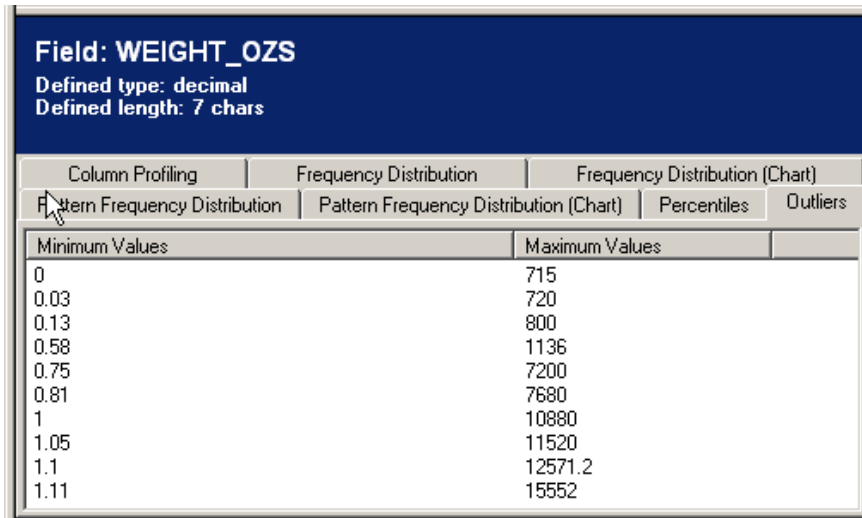


Figure 5: Outlier report on product weight.

However, as you can see, there are many outliers on both the low end and the high end. On the low end, the values were probably entered in pounds instead of ounces. On the high end, potentially these are case or pallet weights instead of individual serving weights. Outlier detection allows you to quickly and easily determine if there are gross inconsistencies in certain data elements. Data profiling tools can let you drill through to the actual records and determine the best mechanism for correction.

Business Rule Validation: Every organization has basic business rules. These business rules cover everything from basic lookup rules:

Salary Grade	Salary Range Low	Salary Range High
20	\$25,000	\$52,000
21	\$32,000	\$60,000
22	\$40,000	\$80,000

To complex, very specific formulas:

$$\text{Reorder_Quantity} = (\text{QuantPerUnit} * \text{EstUnit}) - \text{Inventory_onHand}$$

You can check many basic business rules at the point of data entry and, potentially, recheck these rules on an ad-hoc basis. Problems that arise from lack of validation can be extensive, including over-paying expenses, running out of inventory and undercounting revenue.

Since business rules are often specific to an organization, you will seldom find data profiling technology that will provide these types of checks “out-of-the-box.” These pre-built business rules may provide domain checking, range checking, look-up validation or specific formulas. In addition to the canned data profiling validation techniques, a robust data profiling process must be able to build, store and validate against an organization’s unique business rules.

Applications today need the ability to store, access and implement these basic business rules for data validation. Data profiling should use these same data validation rules to monitor and identify violations of these business rules.

Relationship discovery: Data redundancy and similarity discovery

The third and final major phase of data profiling is relationship discovery. This aspect of profiling discovers what data is in use and links data in disparate applications based on their relationships to each other or to a new application being developed. Different pieces of relevant data that are spread across many separate data stores make it difficult to develop a complete understanding of the data.

Organizations today maintain an enormous amount of data, such as customer data, supplier data, product data, operational and business intelligence data, financial and compliance data and so on. In addition, organizations get data from partners, purchase data from list providers and acquire industry-specific data from other sources. Companies typically don't fully understand all of their data—and cannot effectively manage their data—until they understand all of these sources and the relationships of data across different applications.

Relationship discovery helps you understand how data sources interact with other data sources. Consider some of these problems that can occur when data sources are not properly aligned:

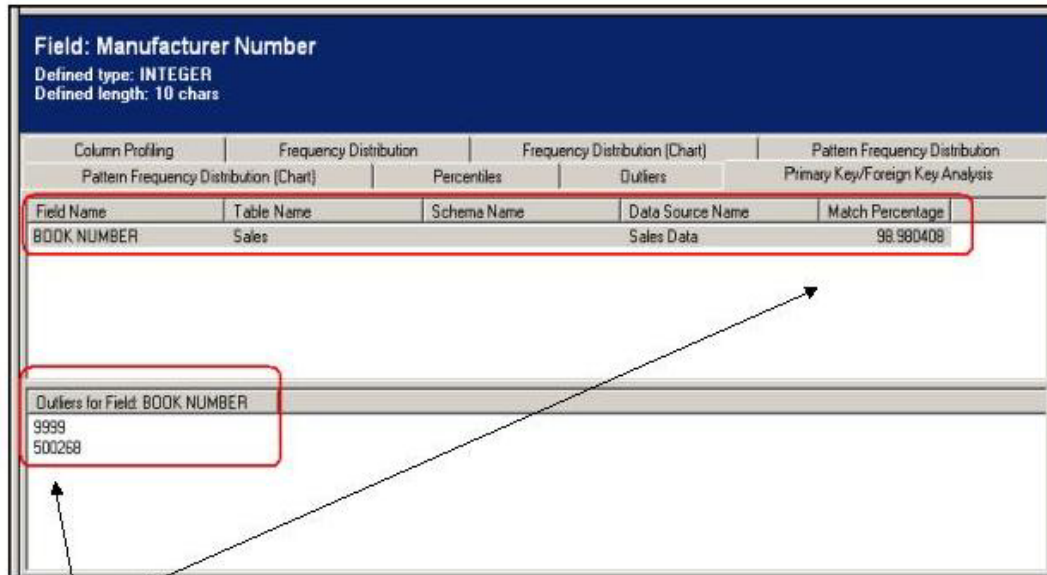
- A product ID exists in your invoice register, but no corresponding product is available in your product database. According to your systems, you have sold a product that does not exist.
- A customer ID exists on a sales order, but no corresponding customer is in your customer database. In effect, you have sold something to a customer with no possibility of delivering the product or billing the customer.
- You run out of a product in your warehouse with a particular UPC number. Your purchasing database has no corresponding UPC number. You have no way of restocking the product.
- Your customer database has multiple customer records with the same ID.

Relationship discovery provides you with information about the ways that data records relate. These records can be multiple records in the same data file, records across data files or records across databases. With relationship discovery, you can profile your data to answer the following questions:

- Are there potential key relationships across tables?
- If there is a primary/foreign key relationship, is it enforced?
- If there is an explicit or inferred key relationship, is there any orphaned data (data that does not have a primary key associated with it)?
- Are there duplicate records?

Relationship discovery starts with metadata to determine relationships, using any metadata available about key relationships. The documented metadata relationships need to be verified. Relationship discovery should also determine, in the absence of metadata, what fields (and therefore, what records) have relationships.

Once potential relationships are determined, further investigation is needed. Does my relationship provide a primary/foreign key? If so, is my primary key unique? If not, which records prevent it from being unique? With my key relationships, are there any outstanding records that do not adhere to the relationship? Figure 6 shows the results of a primary key/secondary key analysis, where two products listed in the sales data did not exist in the products table.



Two products have been sold that do not exist in the products table – the IDs have a 98.98% match rate

Figure 6: Results of primary key/foreign key analysis.

Data profiling has many different aspects. This section has covered some of the more basic types of profiling techniques. Any solid profiling initiative should cover the structure, data and relationship aspects and generate the reports and business rules you need to fully understand (and repair) your data.

Data profiling in practice

Data profiling is not a glamorous task. It is also not something that you can do once and forget about it. Proper data profiling methodology must become a standard part of both your business and IT infrastructure to allow you to continuously diagnose the health of your systems.

Today, many organizations attempt to conduct data profiling tasks manually. With very few columns and minimal rows to profile, this may be practical. But, organizations today have thousands of columns and millions (or billions) of records. Profiling this data manually would require an inordinate amount of human intervention that would still be error-prone and subjective.

In practice, your organization needs a data profiling tool that can automatically process data from any data source and process hundreds or thousands of columns across many data sources. Data profiling in practice consists of three distinct phases:

- Initial profiling and data assessment
- Integration of profiling into automated processes
- Handoff of profiling results to data quality and data integration processes

The most effective data management tools can address all of these initiatives. Data analysis reporting alone is just a small part of your overall data initiative. The results from data profiling serve as the foundation for data quality and data integration initiatives—allowing you to automatically transfer this information to other data management efforts without losing the context or valuable details of the data profiling.

The first part of the process to achieve a high degree of quality control is to perform routine audits of your data as discussed in this paper. A list of these audits follows, along with an example of each.

Type of audit	Example
Domain checking	In a gender field, the value should be M or F.
Range checking	For age, the value should be less than 125 and greater than 0.
Cross-field verification	If a customer orders an upgrade, make sure that customer already owns the product to be upgraded
Address format verification	If "Street" is the designation for street, then make sure no other designations are used.
Name standardization	If "Robert" is the standard name for Robert, then make sure that Bob, Robt. and Rob are not used.
Reference field consolidation	If "GM" stands for "General Motors", make sure it does not stand for "General Mills" elsewhere.
Format consolidation	Make sure date information is stored yyyyymmdd in each applicable field.
Referential integrity	If an order shows that a customer bought product XYZ, then make sure that there actually is a product XYZ.
Basic statistics, frequencies, ranges and outliers	If a company has products that cost between \$1,000 and \$10,000, you can run a report for product prices that occur outside of this range. You can also view product information, such as SKU codes, to view if the SKU groupings are correct and in line with the expected frequencies.
Duplicate identification	If an inactive flag is used to identify customers that are no longer covered by health benefits, make sure all duplicate records are also marked inactive.
Uniqueness and missing value validation	If UPC or SKU codes are supposed to be unique, make sure they are not being reused.
Key identification	If there is a defined primary key/foreign key relationship across tables, validate it by looking for records that do not have a parent.
Data rule compliance	If closed credit accounts must have a balance of zero, make sure there are no records where the closed account flag is true and the account balance total is greater than zero.

Auditing the data as it stands in the systems is not enough. Data profiling needs to be a continuous activity. Your organization is dynamic and evolving. New business initiatives and new business rules continuously generate and incorporate new data into your systems. Each of these new elements brings the potential for more data problems and additional integration headaches.

The rules that you create as part of your initial data profiling activities should be available throughout the data management processes at your organization. As you monitor the consistency, accuracy and reliability of your data over time, you need to apply these same rules

to these ad-hoc data checks. As you investigate data profiling tools, look for tools that can integrate rules and technology into scheduled data profiling processes to track the changes in data quality over time.

Finally, you must also maximize the relationships between data elements, data tables and databases. After you get an overall view of the data within your enterprise, data management solutions must provide the ability to:

- Fix business rule violations.
- Standardize and normalize data sources.
- Consolidate data across data sources.
- Remove duplicate data and choose the best surviving information.

As part of your initial profiling activities, you can develop and implement all required business and integration rules. A robust data management tool will provide the ability to integrate the data validation algorithms as part of standard applications at your organization.

The Four Building Blocks of Data Management

Data profiling is the beginning of an effective data management strategy. Although profiling techniques provide an essential first step, there is much more to a complete data management strategy. The foundation of data management consists of four technology building blocks: data profiling, data quality, data integration and data augmentation.

Your data initiatives cannot succeed unless you have a technology and methodology that addresses all four of these areas. These building blocks each require unique processes, but the building blocks form the necessary parts of a complete data management strategy.

Data profiling: Understand your data

As this paper has discussed, data profiling encompasses such activities as frequency and basic statistic reports, table relationships, phrase and element analysis and business rule discovery. It is primarily done before any data-oriented initiative and often can be used to pinpoint where further efforts need to be focused.

Data quality: Standardize, validate and verify your data

Data is often invalid, out of range or incompatible with current business rules. Data can be misspelled. And data becomes outdated. By checking domain, range and missing values, you can create correction algorithms to identify and correct data problems. Information on customers, suppliers and products all have unique validation rules.

Standardization is also an essential part of data quality. Standardization generally refers to defining consistent and conforming rules for representing similar concepts. Before you can successfully integrate different data sources, the data sources must be normalized so that the same concepts are represented in the same fashion. The best way to correct, standardize and verify your data is to use a reference database or a defined set of business rules and corporate standards. The data quality building block includes technologies that encompass parsing, transformation, verification and validation.

Data integration: Combine data from multiple sources

Whether identifying similar data within and across data sources or removing and consolidating duplicate data records, data integration is necessary to obtain a true understanding of your organization. Data integration can occur at the individual level, at the household level (for example, all customers at the same address), at the business or corporate level, at the product level, at the supplier level or some other combination of attributes.

Data integration requires powerful matching technology that can locate less obvious members of a related group. Data integration technologies will recognize that “Jim Smith at 100 Main Street” and “Michelle Smith at 100 Main St” are members of the same household. A good solution will also recognize that two people with different last names living at the same address could be spouses or members of the same household. In addition, data integration technology can determine that two items are the same. A good data integration tool can determine that “1/4 x 3 wood screw zinc” and “screw, wood (zinc) 1/4 x 3 inches” are the same product. Data integration gives you the ability to join data based on similar concepts as well as exact data matches.

Data augmentation: Enhance and enrich data

This final building block appends new data and provides missing information. Common data augmentation for business or consumer data includes demographic, geographic and credit information. Data augmentation can also encompass data management algorithms and methodologies that combat unique market or industry data problems. For example, data augmentation can provide additional compliance support by supporting matching technology for things like OFAC (Office of Foreign Assets Control) or the PATRIOT Act. Data augmentation also can provide industry- or process-specific enrichment; for example, you can conduct commodity coding to allow your organization to understand the spending techniques company-wide for goods and services.

Conclusion

So, you want to make data a strategic asset at your organization. You understand that your data must be consistent, accurate and reliable if you want your organization to be a leader. As the King said in *Alice’s Adventures in Wonderland*, “Begin at the beginning.” The most effective approach to consistent, accurate and reliable data is to begin with data profiling. And the most effective approach to data profiling is to use a tool that will automate the discovery process.

But data profiling, while a critical piece of your efforts to strengthen your data, is only the first step. In addition, you’ll need a methodology that ties these process steps together in a cohesive fashion. A comprehensive strategy requires technology in the four building blocks of data management—data profiling, data quality, data integration and data augmentation—to achieve success.

Getting started

A pioneer in data management since 1997, DataFlux is a market leader in providing comprehensive, end-to-end data management solutions. DataFlux products are designed to significantly improve the consistency, accuracy and reliability of an organization's business-critical data, enhancing the effectiveness of both customer- and product-specific information.

The process of data management begins with a discovery or data profiling phase that asks one critical question: What points of data collection might have relevant, useful information for your data-based applications and initiatives? Once you begin to understand your data, you can correct errors and use this information to build more productive CRM, ERP, data warehousing or other applications.

DataFlux provides a total solution for your data management needs, which encompasses four building blocks:

- **Data Profiling** – Discover and analyze data discrepancies
- **Data Quality** – Reconcile and correct data
- **Data Integration** – Integrate and link data across disparate sources
- **Data Augmentation** – Enhance information using internal or external data sources

DataFlux's end-user product, dfPower® Studio, brings industrial-strength data management capabilities to both business analysts and IT staff. dfPower Studio is completely customizable, easy to implement, intuitive and usable by any department in your organization. With dfPower Studio, you can identify and fix data inconsistencies, match and integrate items within and across data sources and identify and correct duplicate data. dfPower Studio also provides data augmentation functionalities that allow you to append existing data with information from other data sources, including geographic or demographic information.

Blue Fusion® SDK, a software developer kit, is a packaged set of callable libraries that easily integrates the core DataFlux data management technology into every aspect of your systems, including operational applications and analytical applications. dfIntelliServer™ is a software developer kit built on Blue Fusion SDK and provides a client/server architecture that allows for real-time data validation and correction as data is entered into Web pages or other applications. Working independently or together, dfPower Studio, Blue Fusion SDK and dfIntelliServer ensure a comprehensive data management environment, allowing for better business decisions and improved data-driven initiatives.

DataFlux is a wholly-owned subsidiary of SAS, the market leader in providing business intelligence software and services that create true enterprise intelligence.