Paper 102-30

Data Profiling: Designing the Blueprint for Improved Data Quality

Brett Dorr, DataFlux Corporation, Cary, NC

Pat Herbert, DataFlux Corporation, Cary, NC

ABSTRACT

Many business and IT managers face the same problem: the data that serves as the foundation for their business applications (including customer relationship management (CRM) programs, enterprise resource planning (ERP) tools, and data warehouses) is inconsistent, inaccurate, and unreliable. Data profiling is the solution to this problem and, as such, is a fundamental step that should begin every data-driven initiative. This paper explores how data profiling can help determine the structure and completeness of data and, ultimately, improve data quality. The paper also covers the types of analysis that data profiling can provide as well as how data profiling fits into an overall data management strategy.

INTRODUCTION

Organizations around the world are looking for ways to turn data into a strategic asset. However, before data can be used as the foundation for high-level business intelligence efforts, an organization must address the quality problems that are endemic to the data that's available on customers, products, inventory, assets, or finances. The most effective way to achieve consistent, accurate, and reliable data is to begin with data profiling. Data profiling involves using a tool that automates the discovery process. Ideally, this automation will help uncover the characteristics of the data and the relationships between data sources before any data-driven initiatives (such as data warehousing or enterprise application implementations) are executed.

THE CASE FOR DATA PROFILING

Not so long ago, the way to become a market leader was to have the right product at the right time. However, the industrial and technological revolutions of the last century created a market in which many organizations offered the same products. The path to market leadership required organizations to design and manufacture products cheaper, better, and faster. As more businesses entered the market due to lower barriers of entry, products showed fewer distinguishing characteristics and market leaders disappeared. This situation resulted 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. In today's economy, profits are as much the result of controlled expenses as they are from the generation of new revenue.

To realize significant savings in expenses, organizations throughout the world are implementing two primary enterprise applications: ERP and CRM. Each of these applications focus on driving increased efficiencies from core business processes, with ERP systems trained on keeping expenses "in check," and CRM systems working to build more profitable relationships with customers.

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

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, organizations need high-quality data on inventory, supplies, customers, vendors, and other vital enterprise information. Otherwise, their ERP or CRM implementations are doomed to fail.

The successful organizations of tomorrow will recognize that data (or, more accurately, the successful management of corporate data assets) determines the market leaders of the future. If data is responsible for making market leaders, then it must be consistent, accurate, and reliable. Achieving this level of precision requires solid data management practices that use data profiling as the starting point for any data management initiative. This involves analyzing the current state of your data and building a plan to improve your information assets.

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 paper 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 issues when sifting through corporate data. Often, organizations do not—and worse yet, cannot—make the best decision because they can't access the right data. Just as often, a decision is made based on data that is faulty or untrustworthy. Regardless of the state of the information in 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, including ERP implementations, CRM deployments, data warehouse development, and application re-writes.

Industry estimates for ERP and data warehouse implementations show that these projects fail or go over-budget 65 to 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 problems should be identified and corrected prior to beginning the project. By identifying data quality issues at the front-end of a data-driven project, the risk of project failure can be drastically reduced.

Data profiling, which is also referred to as data discovery, provides a structured approach to understanding your data. Specifically, it can help to discover the data that's available in your organization and the characteristics of that data. Data profiling is a critical diagnostic phase that gives you information about the quality of your data. This information is essential in helping you to determine not only what data is available, but how valid and usable that data is.

The process of data profiling relies on the same principles that your mechanic uses when examining your car. If you take your car in and tell the mechanic that the car has trouble starting, the mechanic doesn't immediately say, "Well, we'd better change the battery." The mechanic uses a series of diagnostic steps to determine the problem: he checks the battery, checks the fluids, tests the spark plugs, and then 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, and so on) without first understanding the data is like fixing a car without understanding the problems. You might get lucky, but chances are you will waste time and money doing work that is neither complete nor productive. With this approach, you are likely to become just another failure statistic in ERP, CRM, or data warehousing implementations.

PROBLEMS WITH DATA

Data problems abound in most organizations. Data problems can include data inconsistencies, anomalies, missing data, duplicated data, data that does not meet business rules, and orphaned data—just to name a few. These problems can limit or even ruin your data initiatives. Therefore, before you begin any data initiative, you need to know basic information about your data, such as:

- Do you trust the quality of the data that you're using in this initiative?
- Will the existing data support the needed functionality?
- Is the data you're 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 understanding these issues will lead to large development and cost overruns or potential project failures. From a real-world perspective, the effect can be incredibly costly. For example, one organization spent over 100,000 dollars (USD) in labor costs identifying and correcting 111 different spellings of the organization name AT&T.

Because organizations rely on data that is inconsistent, inaccurate, and unreliable, large-scale implementations are ripe for failure or cost overruns. More disturbing, the organizations usually do not understand the magnitude of the problem or the impact that the problems have on their bottom line. Data problems within an organization can lead to lost sales and wasted money, poor decisions, sub-standard customer relations, and ultimately, failed businesses. Data profiling is the first step toward diagnosing and fixing problematic data.. To help you "begin at the beginning," data profiling encompasses many techniques and processes that involve three major categories:

- Structure Discovery
- Data Discovery
- Relationship Discovery

STRUCTURE DISCOVERY: DOES DATA MATCH METADATA? DO PATTERNS MATCH AS EXPECTED?

Structure Discovery techniques enable you to take a broad view of the data that you plan to use. By examining complete columns or tables of data, structure analysis helps you to determine whether the data in that column or table is consistent and if it meets the expectations that you have for the data. Many techniques can validate the adherence of data to expected formats. Any one of these techniques provides insight into the validity of the data.

Let's review some of the common structure analysis techniques and the potential problems that these techniques can uncover. In particular, we will explore three techniques: validation with metadata, pattern matching, and the use of basic statistics.

VALIDATION WITH METADATA

Most data has some associated metadata or a description of the characteristics of the data. It might 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 the data type, the field length, whether the data should be unique, and if a field can be missing or null.

Metadata is supposed to describe the data that's in a 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, which causes far-reaching implications for any data management effort. For example, consider a 10 million-row field that has 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.

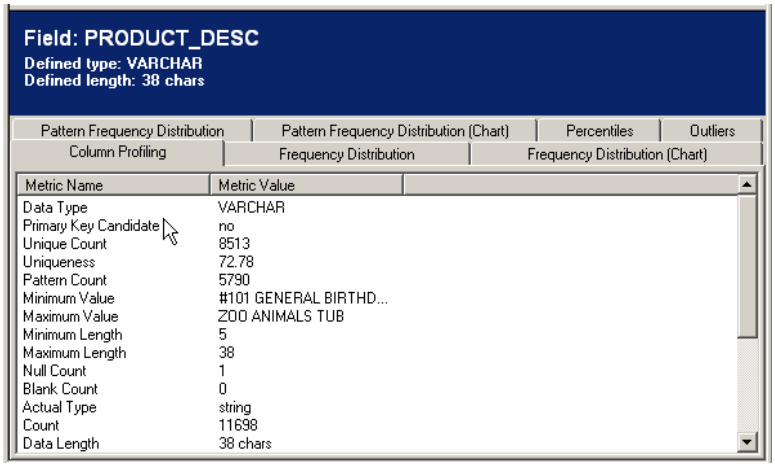

Figure 1. Metadata Report for a Character Field

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? The answers to these questions will help you make decisions about how to use the data in the future.

PATTERN MATCHING

Typically, pattern matching determines if the data values in a field are in the expected format. This technique can quickly validate whether the field data is consistent across the data source, and whether the information is consistent with your expectations. For example, pattern matching would analyze if a phone number field contains all valid phone numbers or if a social security field contains all valid social security numbers. Pattern matching can also tell you if a field is completely numeric, if a field has consistent lengths, and other format-specific information about the data.

For example, consider a pattern report for North American telephone numbers. There are many valid phone number formats, but all valid formats consist of three sets of numbers (three numbers for the area code, three numbers for the exchange, and four numbers for the station). These sets of numbers might be separated by a space or by a special character. Valid patterns include:

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

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

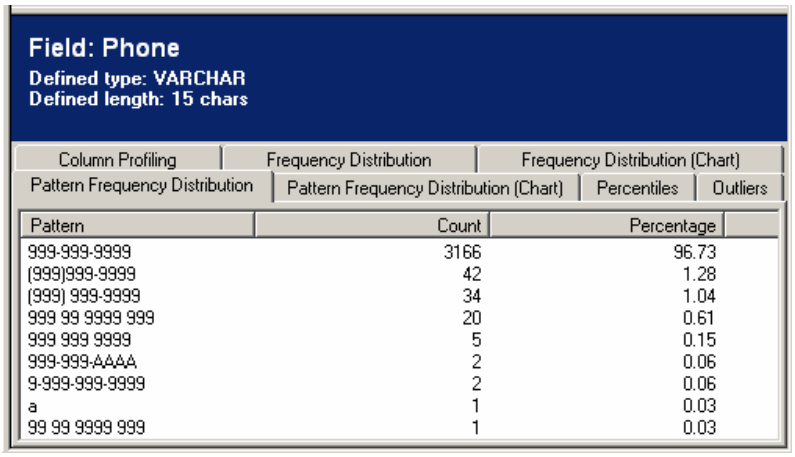

Figure 2. Pattern Frequency Report for Telephone Numbers

The majority of the data in the Phone field contains valid telephone numbers for North America. However, some data entries do not match a valid telephone-number pattern. A data profiling tool enables you to drill through a report like this to view the underlying data or to generate a report that contains the drill-down subset of data to help you correct those records.

BASIC STATISTICS

You can learn a lot about your data by reviewing some basic statistics. This is true for all types of data, especially numeric data. Reviewing statistics such as minimum and maximum values, mean, median, mode, and standard deviation can give you insight into the validity of the data. Figure 3 shows statistical data about home loan values from a financial organization. Personal home loans usually range from 20 thousand to 1 million dollars. A loan database that contains 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 a loan database.

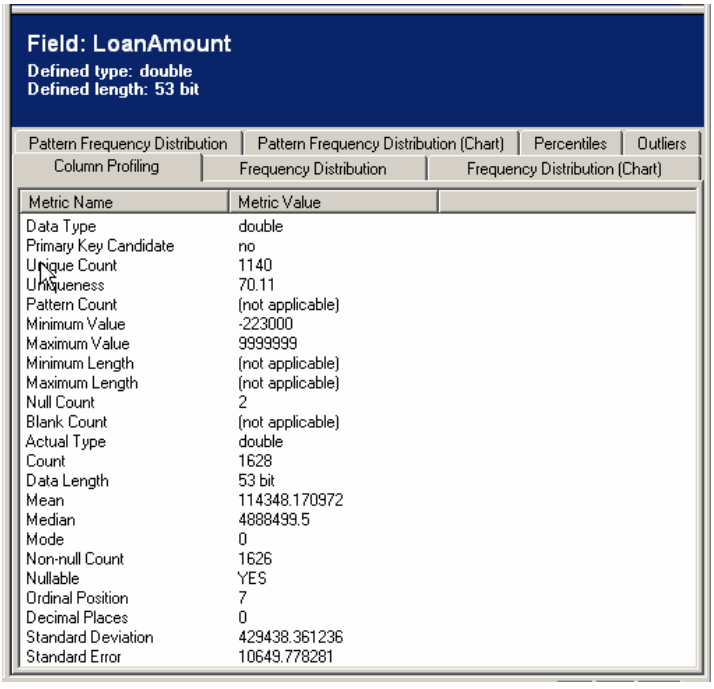

Figure 3. Statistics for a Column of Loan Amount Data

The report in Figure 3 uncovers many potential problems with the loan amounts. The minimum value of a loan is a negative number (-223000). The maximum value for a loan is 9,999,999 dollars. There are two loans with missing values (Null Count). The median and standard deviations are unexpectedly large numbers. All these values indicate potential problems for a home loan data file.

Basic statistics give you a snapshot of an entire data field. As new data arrives, tracking basic statistics over a period of time will give you insight into the characteristics of the new data that enters your systems. Checking the basic statistics for the new data prior to entering the data into the system can alert you to inconsistent information and help prevent introducing problematic data into 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. Many problems are caused by incorrectly entering data into a field; this is the most likely source of the negative value in the home loan data in Figure 3. Some problems occur because a correct value was unknown and a default or fabricated value is used (potentially the origin of the 9999999 maximum home loan value).

Other structure problems are the result of legacy data sources that are still in use or have been migrated to a new application. During the data creation process for older mainframe systems, programmers and database administrators often 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: DATA COMPLETENESS DISCOVERY AND BUSINESS RULE VALIDATION

After you have analyzed entire tables or columns of data using Structure Discovery steps, you need to look more closely at each of the individual data 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 and frequency counts and outlier detection to find data elements that don't make sense. Let's look at each of these techniques in more detail.

STANDARDIZATION

Unfortunately, data can be ambiguous. Data in an organization often comes from a variety of sources: different departments, various data entry clerks, and various 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 that is generated by that data must account for each instance of these multiple permutations. Otherwise, important data points can be missed, which can affect the output of future processes. For example:

- "IBM," "Int. Business Machines," "I.B.M," "ibm," and "Intl Bus Machines" all represent the same organization.
- Does the organization "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 the values for a specific item has the same meaning, but the representations are different. The analytical and operational problems of this non-standard data can be very costly because you cannot get a true picture of the customers, businesses, or items in your data sources. For example, a life insurance organization might want to determine the top ten organizations in a given geographic region that employ their policy holders. With this information, the organization can tailor policies to those specific organizations. If the employer field in the data source represents the same organization 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 often 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 the creative efforts of the group, while alienating customers by not marketing to their preferences, effectively.

These simple examples of data inconsistency and other similar situations are common to databases worldwide. Fortunately, data profiling tools can discover such inconsistencies, provide a blueprint for data quality technology to address, and fix the problems.

FREQUENCY COUNTS AND OUTLIERS

Frequency counts and outlier detection provide you with techniques that can limit the amount of fault detection that's required by a business analyst. 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 might be considered incorrect, and drill down to the data to make a more in-depth determination about the data.

Consider a frequency report that details the ways that North American states were represented in a data source. The frequency distribution shows a number of correct state entries. The report also shows data that needs to be corrected. Incorrect spellings, invalid abbreviations, and multiple representations of the same state can cause problems. California is represented as "CA," "CA.," "Ca.," and "California." Non-standard representations will have an impact anytime that you are trying to perform state-level analysis. The invalid state entries might prevent you from contacting certain individuals while the missing state values might make communication attempts problematic.

Outlier detection helps you to pinpoint problem data. Whereas frequency count looks at how values are related according to data occurrences, outlier detection examines what you hope are a 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. Figure 4 shows a sample outlier report (which contains the 10 minimum and 10 maximum values for the field). The field is product weight, measured in ounces, for individual-serving microwavable meals. A business analyst would understand that the valid weights are between 16 and 80 ounces.

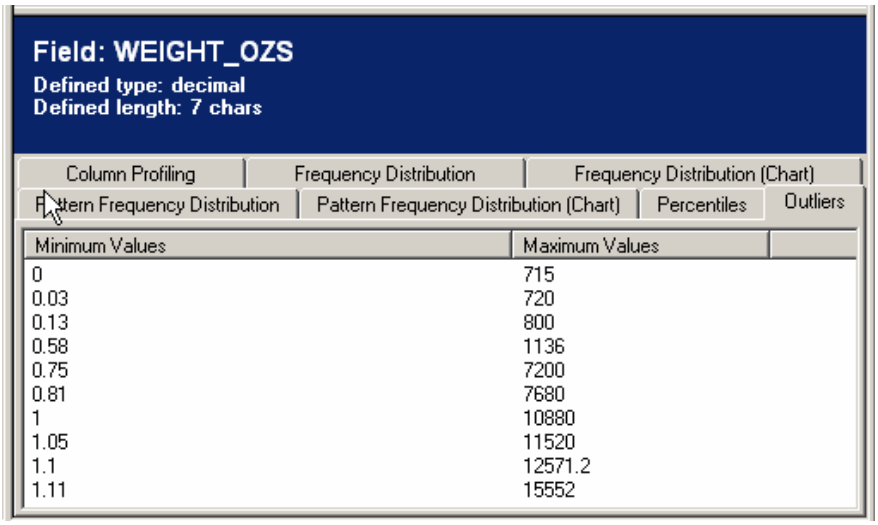

Figure 4. Outlier Report for Product Weight

However, 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, the values could potentially be case or pallet weights instead of individual serving weights. Outlier detection enables you to determine if there are gross inconsistencies in some data elements. To view the data in context, 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 (USD)	Salary Range High (USD)
20	25,000	52,000
21	32,000	60,000
22	40,000	80,000

to complex, very specific formulas:

```
Reorder_Quantity = (QuantPerUnit*EstUnit)
[Unit_type] - Inventory_onHand
```

You can check many basic business rules at the point of data entry and, potentially, re-check these rules on an ad hoc basis. Problems that arise from lack of validation can be extensive, including over-paying invoices, running out of inventory, and undercounting revenue. Business rules are often specific to an organization, and you will seldom find data profiling technology that will provide these types of checks "out-of-the-box." Instead, pre-built business rules typically provide domain checking, range checking, look-up validation, and specific formulas. In addition to the generic type of data profiling validation techniques, a robust data profiling process must be able to build, store, and validate against an organization's unique business rules.

RELATIONSHIP DISCOVERY: DATA REDUNDANCY AND SIMILARITY DISCOVERY

After Structure Discovery and Data Discovery techniques have been implemented, you are ready to put the Relationship Discovery data profiling technique into practice. This aspect of profiling is used to establish links among the data in disparate applications or databases. These interdependencies are based on the relationships that different data stores have to each other or to any new application that is under development. Relationship Discovery is essential. Without it, the different pieces of relevant data that are located across many separate data stores can make it difficult to develop a complete understanding of enterprise data.

Today, organizations 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 receive data from partners, purchase data from list providers, and acquire industry-specific data from other sources. Typically, organizations don't fully understand all their data, especially about the quality of incoming information. Worse yet, organizations cannot effectively manage their data until they understand all available data sources and the relationships that these sources have across different applications.

Relationship Discovery identifies 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 that has a specific UPC number. Your purchasing database has no corresponding UPC number. You have no way of re-stocking 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 by using any available metadata to determine key relationships. The documented metadata relationships must then be verified. Relationship Discovery should also determine, in the absence of metadata, what fields (and, therefore, what records) have relationships.

After potential relationships are determined, further investigation is necessary. Does a relationship provide a primary or a foreign key? If so, is the primary key unique? If not, which records prevent it from being unique? Within key relationships, are there any outstanding records that do not adhere to the relationship? Knowing this information can help you link and join data, and save you time and money on large-scale data integration projects.

DATA PROFILING IN PRACTICE

Today, many organizations attempt to conduct data profiling tasks manually. If very few columns and minimal rows exist to profile, then this might be practical. Unfortunately, most organizations 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. Putting data profiling into practice consists of three distinct phases:

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

The most effective data management tools can address all these phases. 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. It's best to look for a data profiling solution that enables you to construct data correction, validation, and verification routines directly from the profiling reports. This will help you to combine data inspection and correction phases, which will streamline your overall data management process.

To achieve a high degree of quality control, the first part of any comprehensive data quality improvement process involves performing 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, then make sure that the customer already owns the product.
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," then make sure it does not stand for "General Mills" elsewhere.
Format consolidation	Make sure that date information is stored as `yyyymmdd` 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 organization has products that cost between 1000 and 10000 dollars, you can run a report for product prices that are not in this range. You can also view product information, such as SKU codes, to find out 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, then make sure all duplicate records are also marked inactive.
Uniqueness and missing value validation	If UPC or SKU codes are supposed to be unique, then make sure they are not being reused.
Key identification	If there is a defined primary key/foreign key relationship across tables, then validate it by looking for records that do not have a parent.
Data rule compliance	If closed credit accounts must have a balance of 0, then make sure there are no records where the closed account flag is true and the account balance total is greater than 0.

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

Finally, you must maximize the relationships among data elements, data tables, and databases. After you obtain an overall view of the data in your organization, your 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.

CONCLUSION

Data quality is more than a technological fix. To create data that can drive accurate decision-making and provide solid analytics on your corporate information, you must begin by taking an accurate assessment of your current data landscape. As the king said in *Alice's Adventures in Wonderland*, "Begin at the beginning." The most effective approach to achieving consistent, accurate, and reliable data is to begin with data profiling. The most effective approach to data profiling is to use a tool that automates the discovery process.

Keep in mind that 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 that addresses the five building blocks of data management—data profiling, data quality, data integration, data augmentation, and data monitoring.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Brett Dorr
DataFlux Corporation
4001 Weston Parkway, Suite 300
Cary, NC 27513
Work Phone: 919-674-2153
Fax: 919-678-8330
E-mail: brett.dorr@dataflux.com
Web: www.dataflux.com

Pat Herbert
DataFlux Corporation
4001 Weston Parkway, Suite 300
Cary, NC 27513
Work Phone: 919-674-2153
Fax: 919-678-8330
E-mail: pat.herbert@dataflux.com
Web: www.dataflux.com