



Companies around the world have spent billions of dollars in recent years implementing enterprise applications to better integrate their corporate data and increase the overall value of existing information. With high-quality data, they hope to realize the significant savings that come from consolidated data, such as increased operational efficiencies, enhanced compliance initiatives and improved customer relationships.

These organizations have invested heavily in data-driven initiatives such as customer relationship management (CRM) and enterprise resource planning (ERP) applications, hoping to build more profitable relationships with customers and hold product-related expenses in check. Despite good intentions, however, most of these applications have faltered due to inefficient, outdated data. In fact, industry estimates show these projects fail or go over budget up to 75 percent of the time.

The very foundation of CRM and ERP systems is the data that drives these implementations. Beginning a data-driven initiative without first understanding the existing, underlying data is like repairing an automobile without first understanding the problems inside the engine. To repair the engine, the mechanic first has to understand the breadth and depth of the problem.

Successful data quality begins with a clear understanding of the integrity of your current data. Data profiling, also called data discovery, gives you the diagnosis of your existing data to begin building a successful data improvement and integration effort through consistent, accurate and reliable data throughout your organization.

## The Problems with Data

Data problems abound in most organizations. Some of the more common problems today include: outdated, inconsistent, missing, orphaned or duplicated data; data anomalies or outliers; and data that does not meet specified business rules. Before you begin any data quality improvement initiative, you should ask and address some key questions:

- Do you trust the quality of the data you are using in this initiative?
- Does the data for this initiative conform to the business rules monitoring process you expect to set up later?
- Will the existing data support the needed functionality?
- Is the data you are using complete enough to populate the needed data repository?

Engaging in any data initiative without a clear understanding of these issues will lead to large development and cost overruns or potential project failures. The effect can be incredibly costly. For example, one company spent more than \$100,000 in labor costs identifying and correcting 111 different spellings of the company AT&T. Data problems within your organization can lead to sub-standard customer relations, wasted expenses, poor decisions, lost sales, and ultimately, failed businesses.

*Do you trust the accuracy of your current enterprise data?*

## Data Profiling Defined

Many businesses and IT managers find that their data is often confusing and disorienting. Often, organizations do not – and worse yet, cannot – make the best decision because they can't get access to the right data. Just as often, decisions are made based on data that is faulty or untrustworthy.

In *Alice's Adventures in Wonderland*, when Alice asks the King where to begin, the King replied gravely, "Begin at the beginning." Regardless of the state of the information within your enterprise, the King had the right idea: Begin at the beginning.

*Data profiling gives a thorough diagnosis of your data.*

Data profiling is a fundamental step that should begin every data-driven initiative, yet it is often taken for granted. In fact, every ERP, CRM or data warehouse application project should start with data profiling. By identifying data quality issues at the front-end of a data-driven project, you can drastically reduce the risk of project failure later.

As mentioned earlier, 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 first goes through a series of diagnostic steps to determine the problem: he checks the battery, the starter, the fluids and the spark plugs. After a thorough diagnostic review, the mechanic validates the reliability of each relevant part and is ready to make the needed changes.

When proper data profiling methodologies are applied, you can also "look under the hood" to gain valuable insight into your business processes, refine these procedures over time and recommend new ways to refine and enhance the data-entry process.

## Discovering Your Underlying Data

Data profiling is the first step in the data quality process to help you diagnose – and repair – the problem. To help you "begin at the beginning," let's take a closer look at some of the data profiling techniques and processes being used today. These techniques 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?
- **Content discovery** – Is the data complete? Is it accurate? Does it contain information that is easily understood 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?

### *Structure discovery: Understanding metadata and data patterns*

By examining complete columns or tables of data, structure discovery – also known as structure analysis – helps you determine whether or not the data in that column or table is consistent and meets your expectations for this 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. Traditionally, a good place to start is by examining the actual state of data against the metadata on that data source.

*Metadata analysis determines whether original data expectations have been met.*

## Validation with Metadata

Most data has associated metadata – a 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 contains 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 designed to describe the data found 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.

Figure 1 shows information that a typical metadata report would display. In this case, the sales chart shows metric value counts compared with null counts for each individual field name (address, city, company, etc.).



**Figure 1: Metadata report for structure analysis of metric counts vs. null counts for individual fields.**

Metadata analysis determines 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 more informed decisions about how to use this data going forward.

## Pattern Matching

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

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 and 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. 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)	Percentiles
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.

### *Content discovery: Validating rules and assessing data completeness*

After you analyze entire tables or columns of data using these structure discovery techniques, you would then 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. Content discovery digs deeper and helps you determine which data values are inaccurate, incomplete or ambiguous.

Content 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 based on specific business rules to verify data that may be unique to your organization. Let's look in more detail at these techniques.

## Standardization

Unfortunately, data can be ambiguous. Organizational data often comes from a variety of sources: different departments, data entry clerks, partners, etc. 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 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:

- "GM," "General Motors," "G.M.," "g.m.," and "Genrl Mtrs" all represent the same company.
- "GM" in a database represents "General Motors" and "General Mills."
- "Brass Screw," "Screw: Brass," "Br. Screw," and "SCRW BRSS" all represent the same product.
- "100 E 4th Str," "100 East Fourth Str.," "100 East Fourth," and "100 4th Street" all represent the same address.

*Non-standard data foils analytical and operational efforts.*

Each of these values has 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.

These types of situations are endemic to databases worldwide. Fortunately, data profiling tools can discover these inconsistencies and provide a blueprint for a data quality technology to address and fix the problems at hand.

## 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 appear to be incorrect. But as the data grows, this quickly becomes an overwhelming 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 susceptible 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.

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.

## 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:

$$\text{Reorder\_Quantity} = (\text{QuantPerUnit} * \text{EstUnit}) \\ [\text{Unit\_type}] - \text{Inventory\_onHand}$$

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 major type 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 spread across many individual data stores make it difficult to develop a complete understanding of the data.

*Data  
profiling  
highlights  
potential key  
relationships  
across  
tables.*

Organizations today maintain a massive amount of data on customers, products, suppliers, personnel, finances and employees. Additionally, organizations get data from partners, purchase data from list providers and acquire industry-specific data from other sources.

As a result of the different sources of data, companies typically don't fully understand much of their data –and cannot effectively manage this data. To remedy this prevalent problem, you must understand all of these sources and the relationships of data across different applications.

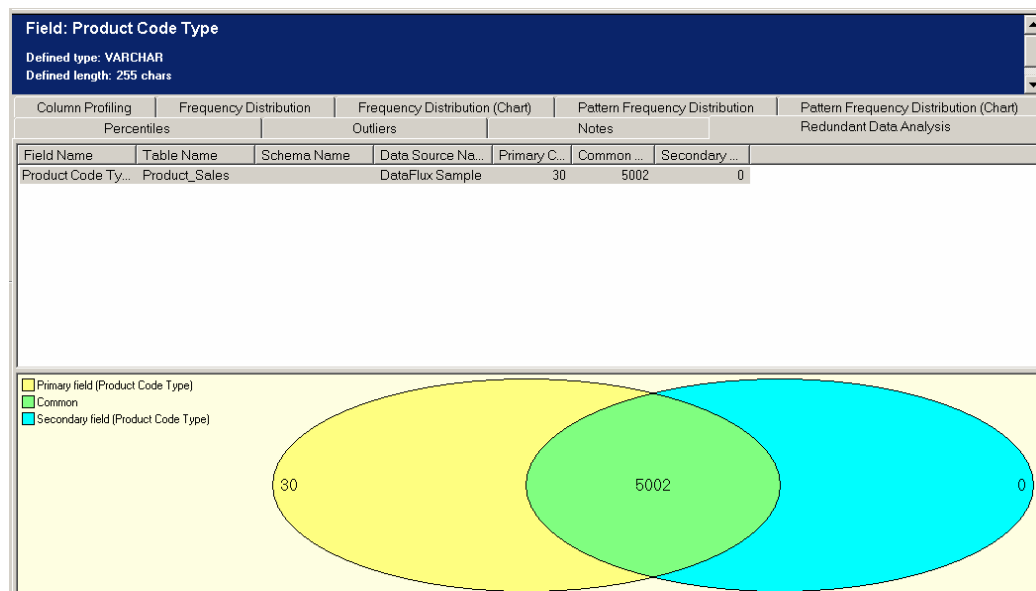
Relationship discovery helps you understand how data sources interact with other data sources. Consider some of these problems that can occur when data sources are not properly aligned:

- A 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.
- Your customer database has multiple customer records with the same ID.
- 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.
- 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.

Relationship discovery provides 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, your business analysts can profile this data to answer questions about the relationships between various records.

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

Figure 4 demonstrates an analysis of two product code types. In the first, or primary, field, 30 products are unique. An analyst would double click on these 30 to see just these unique records. The secondary field has no unique values. But relationship analysis shows that there are 5,002 overlapping records between the two products.



**Figure 4: Relationship analysis shows results of a comparison between two product code types.**

## Data Profiling in Practice

*Every successful data quality project starts with data profiling.*

Although excellent technology tools and methodologies exist today, many organizations continue to conduct data profiling tasks manually – or they ignore profiling altogether. Manual profiling may be practical when there are very few columns and minimal rows to profile. 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:

1. Initial profiling and data assessment
2. Integration of profiling into automated processes
3. Handoff profiling results to data quality and data integration processes.

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 sample list of these audits follows, along with an example of each.

<b>Type of audit</b>	<b>Example</b>
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.
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 see if the SKU groupings are correct and in line with expected frequencies.
Duplicate identification	If an inactive flag is used to identify customers that are no longer covered by health benefits, make sure duplicate files are marked inactive.
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.

One other concept to keep in mind when practicing these data profiling techniques is methodology. Data profiling is the first step in the data quality integration process that helps you diagnose your enterprise problems. But it is not a “one and done” project, something that you can apply once and be done with it. Proper data profiling must be part of a larger data quality methodology that ties these processes together in a cohesive fashion through an integrated and phased approach.

These phases, or “building blocks” of data quality, begin with data profiling and continue with data quality (standardizing, validating and verifying your data), data integration (accessing, linking and connecting data from multiple sources), data enrichment (augmenting and enhancing your data) and data monitoring (continual examining and auditing your data based on pre-built business rules).

## Conclusion

Like other astute business and IT managers, you know that your data is often outdated, redundant and inconsistent. And you’ve decided that you want to make data a strategic asset at your organization, understanding that this data needs to be consistent, accurate and reliable so your organization can make bold, intelligent decisions.

As the King stated the obvious in *Alice’s Adventures in Wonderland* – “Begin at the beginning” – the most effective approach to having clean data is to begin with data profiling. Data profiling is the important first step in the beginning of any effective data quality integration strategy. Proper data profiling must be part of a larger data quality methodology that encapsulates profiling as the first step in the process, alongside data quality, data integration, data enrichment and data monitoring.

A truly successful data quality initiative starts with a thorough and honest examination of your existing data and sources. Data profiling gives you the thorough diagnosis you need to begin building a credible foundation of high-quality data from throughout your organization.