The success of any data warehouse rests on the quality of the data it contains. Yet, some business analysts are reluctant to accept the results their data warehouse provides. Why? Because they view the data as incomplete, redundant, or inconsistent. Other companies, meanwhile, are reluctant to centralize their data because they feel they can’t overcome data quality issues. In both cases, poor data quality is preventing businesses from realizing the full potential of their data.

You can take the first step toward protecting the integrity of your database by assessing your data quality with Teradata® Profiler, a unique combination of data assessment and exploration functions. Data quality profiling provides an array of advantages, including:

- Increased comfort level for analysts – add confidence in the quality of the data that produce your reports.
- Understand your data better – understand data characteristics and relationships better to validate business assumptions.
- Make better business decisions – avoid making costly mistakes based on erroneous data.

Teradata Profiler also brings you:

- A robust set of data exploration functions including descriptive statistical and visualization techniques.
- The power to analyze large volumes of complex data in multiple tables.
- In-database technology – the most efficient method of processing large data volumes by moving the functions to the data.
- The ability to capture data quality metadata, allowing you to audit data quality on an on-going basis to maintain the optimal level of data quality that your business requires.
- The ability to drill down on data for detailed analysis.

**Analyze All Your Data**

In addition to its inherent parallelism, each Teradata Profiler analysis has been optimized to minimize table scans. This in-database technique allows Teradata to deliver a highly scalable data quality assessment tool to analyze all your data, not just samples. For example, a customer tried to analyze two very large tables that required a join. The join was aborted after running for three hours because the data quality tool extracted the data and joined and analyzed each row outside of the database. Teradata Profiler joined the tables into a view and did the analysis in only five minutes providing a visual display of each variable and analysis.

Here’s how Teradata Profiler works:

**Data Explorer Analysis**

This feature allows analysts to explore multiple variables across multiple tables within a particular database in the same pass. It’s actually an intelligent data explorer because it determines the best analysis to run on data elements based on type and results from preliminary analysis.

The results are presented in a series of graphs in which each variable selected is represented with bar and pie charts to view the type, count, null, zero, and number of unique values depending on the variable type. Box and whisker graphs and histograms are also used to illustrate descriptive statistics, such as mean, max, min, standard deviation, and variable distribution. Data Explorer provides a browser for graphical objects to allow the analyst to view all selected variables to quickly identify missing values and other data quality issues. The analyst simply selects the variables, and the Data Explorer function does the rest.
**Drill Down Capability for Record-level Analysis**

Data Explorer and other descriptive analyses provide details about the statistical characteristics of each variable, such as the number of unique values, distribution of variables, along with other attributes helping analysts identify anomalies in their data. The drill down feature provides the next level of analysis allowing users to drill down and analyze data at the record level. By simply clicking on a graph or section of a graph, Teradata Profiler will automatically generate SQL to retrieve those records from the data warehouse allowing further analysis. Not only can users quickly identify questionable variables, but they can also retrieve the detailed data contributing to anomalies. Drill down capability is available for Data Explorer analysis, including individual values, statistics, histogram, and frequency analysis. See Figure 1.

**Descriptive Statistics**

All descriptive and exploratory functions used in the Data Explorer analysis are also available individually.

**Values Analysis** – provides a basic assessment of the accuracy and completeness of the data and identifies obvious cases where the data's values fall out of their expected domain. The Values Analysis can also be used to identify suspect values. When the analyst finds an unusual situation, such as negative values for the variable age, he or she can easily analyze values that fall out of the normal range.

**Frequency Analysis** – counts the occurrence of individual data values in columns that contain categorical data (non-numeric values). It can be useful in understanding the meaning of a particular data element, and it may point out the need to recode some of the data values found in the analytic data set. Frequency Analysis can also help analyze combinations of values occurring in two or more columns.

The Frequency Analysis can be used to detect a range of data quality issues, such as identifying duplicate values, structural analysis, and referential integrity of keys between two tables.

**Overlap Analysis** – provides counts of overlapping key fields among pairs of tables to identify referential integrity issues. For example, if you're building an analytic data set to describe customers, this analysis can help you determine if the customer, account, and transaction tables actually refer to the same customers.

**Statistical Analysis** – helps you understand the characteristics and properties of columns to identify their distribution, assess their quality, and look for outliers and other possible anomalies. Statistical Analysis provides several statistical measures for numeric data columns, using either population measures or sample
Teradata Profiler

Histogram Analysis – studies the distribution of continuous numeric values in a column by providing the data necessary to create a histogram graph. This type of analysis is sometimes referred to as binning because it counts the occurrence of values in a series of numeric ranges called bins. The Histogram Analysis provided in Teradata Profiler is particularly rich in functionality. It provides a number of ways to define bins, including multi-dimensional binning, overlaying of categorical data, calculation of numeric statistics within bins, and adaptive binning.

Scatter Plots – help identify relationships across two or three different variable combinations. These types of plots are used to investigate the possible relationship among two or three variables that both relate to the same event. You can often make inferences, depending upon the cluster of points within the scatter plot. See Figure 2.

The scatter plot is used as a visual method for detecting consistency violations. With the scatter plot, the assessor can visualize up to three variables at a time in a graphical representation of the data items. If you were to visualize the variable representing age, for example, you could clearly see any negative values or values beyond reasonable boundaries.

The Teradata Difference
Teradata Corporation has 25 years of experience building data warehouses where data quality is a critical core competency. Because of our experience in data warehousing and our in-depth understanding of key industries, we can provide the solution that meets your business requirements. In addition, Teradata technology provides the most efficient and scalable solutions for analyzing data directly in your Teradata Database.

For More Information
To find out more about how the Teradata Profiler can help you improve your data quality and grow a stronger, more productive business, contact your Teradata representative or visit Teradata.com.