Data Profiling Basics

What is Data Profiling?

Data profiling is a process for analyzing large data sets. Standard data profiling automatically compiles statistics and other summary information about the data records. It includes analysis by field for minimum and maximum values and other basic statistics, frequency counts for fields, data type and patterns/formats, and conformity to expected values. Other advanced profiling techniques also perform analysis about the relationships between fields, such as dependencies between fields in a single set and between fields in separate data sets.

Why Do People Profile?

People may want to profile for several reasons, including:

Assessing risks—Can data support the new initiative?

Planning projects—What are realistic time lines and what data, systems, and resources will the project involve?

Scoping projects—Which data and systems will be included based on priority, quality, and level of effort required?

Assessing data quality—How accurate, consistent, and complete is the data within a single system?

Designing new systems—What should the target structures look like? What mappings or transformations need to occur?

Checking/monitoring data—Does the data continue to meet business requirements after systems have gone live and changes and additions occur?

Key Considerations for Selecting a Data Profiling Tool

1. Who is profiling: business users, IT, or both
2. Common enviroment to communicate, review, and interpret results
3. Complexity of analysis, number of sources
4. Security of data
5. Ongoing support and monitoring

TRILLIUM SOFTWARE®

HARTE HANKS

Who Should Be Profiling the Data?

Data profiling is primarily considered part of IT projects, but the most successful efforts involve a blend of IT resources and business users of the data. IT, business users, and data stewards each contribute valuable insights critical to the process:

IT system owners, developers, and project managers analyze and understand issues of data structure: how complete is the data, how consistent are the formats, are key fields unique, is referential integrity enforced?

Business users and subject matter experts understand the data content: what the data means, how it is applied in existing business processes, what data is required for new processes, what data is inaccurate or out of context?

Data stewards understand corporate standards and enterprise data requirements as a whole. They can contribute to both the requirements for specific projects and the corporation.

How Do People Profile Data?

The techniques for profiling are either manual or automated via a profiling tool:

Manual techniques involve people sifting through the data to assess its condition, query by query. Manual profiling is appropriate for small data sets from a single source, with fewer than 50 fields, where the data is relatively simple.

Automated techniques use software tools to collect summary statistics and analyses. These tools are the most appropriate for projects with hundreds of thousands of records, many fields, multiple sources, and questionable documentation and metadata. Sophisticated data profiling technology was built to handle complex problems, especially for high-profile and mission-critical projects.

How Do Data Profiling Tools Differ?

Data profiling tools vary both in the architecture they use to analyze data and in the working environment they provide for the data profiling team.

Architecture option: Query-based profiling Some profiling technologies involve crafting SQL queries that are run against source systems or against a snapshot copy of the source data. While this generates some good information about the data, it has several limitations:

Performance risks: Queries strain live systems, slowing down operations, sometimes significantly. When additional information is required, or if users want to see the actual data, a second query executes, creating even more strain on the system. Organizations reduce this risk by making a copy of the data, but this requires replicating the entire environment—both hardware and software systems—which can be costly and time-consuming.

Traceability risks: Data in production systems changes constantly. The statistics and metadata captured from query-based profiling risk being out of date immediately.

Completeness risks: It is difficult to gain comprehensive insights using query-based analysis. Queries are based on assumptions, and the purpose is to confirm and quantify expectations about what is wrong and right in the data. Given this, it is easy to overlook problems that you are not already aware of.

Profiling by query is valuable when you want to monitor production data for certain conditions. But it is not the best way to analyze large volumes of data in preparation for large-scale data integrations and migrations.

Architecture option: Data profiling repository

Other profiling technologies profile data as part of a scheduled process and store results in a profiling repository. Stored results can include content such as summary statistics, metadata, patterns, keys, relationships, and data values. Results can then be further analyzed by users or stored for later trending analysis.

Profiling repositories that allow users to drill down on information and see original data values in the context of source records provide the most versatility and stability for non-technical audiences. Independence from operational source systems coupled with the vast amount of metadata and information derived from a point in time profile provide a cross-functional team of business and IT resources a common, comprehensive view of source system data from which traceable decisions can be based.

Volume considerations: Should tables or files enter into the range of hundreds of millions of records, a profiling repository strategy should be considered. With volumes this large, the best strategy may be a blend between weekend-scheduled profiling processes and focused, non-contentious query-based profiling, closely monitored by IT.

Work environment: Multi-user workspace

Some profiling tools are designed as desktop solutions for resources to use as a team of one. How many resources will be involved in your data profiling efforts? For large projects, there is generally a cross-functional team involved.

Consider the environment a profiling tool provides since multiple users with different skills, different expertise, and varying level of technical skills all need to be able to access and clearly see the condition of the data. Even if some prospective data profilers are skilled in SQL and database technologies, profiling tools that foster collaboration between business users and IT offer greater value overall. With a common window on the data sets, people with diverse backgrounds can concretely and productively discuss the data, its current state, and what is required to move forward.

Work environment: Graphical Interface

Because users may not be familiar with database structures and technologies, it is important to find a tool that provides an intuitive, easy-to-learn graphical user interface (GUI). Appropriate security features should also be a part of the work environment, to ensure that access to restricted fields or records can be allowed or denied, for sensitive information.

What Follows Data Profiling?

Once the task of data profiling is complete, there is more to do. Keep in mind both the short- and long-term goals driving the need to profile your data. Leverage your investments by understanding what follows and see if there are logical extensions to your profiling efforts that can be executed within the same tool.

ETL projects for data integration or migration use profiling results to design target systems, define how to accurately integrate multiple data sets, and efficiently move data to a new system, taking all data conditions into consideration.

Data quality processes that improve the accuracy, consistency, and completeness of data use results to identify problems or anomalies and then develop rules for automated cleansing and standardization.

Data monitoring initiatives use profiling results to establish automated processes for ongoing assessment of key data elements and acute data conditions in production systems. The profiling repository captures results, sends alerts, and centrally manages data standards.

TS Discovery for Data Profiling

TS Discovery is unlike many other data profiling tools on the market. It performs as a best-of-breed data profiling tool, and also has several key differentiators that advance its overall value:

User Interface— the interface is designed specifically for a business user. It is intuitive, easy to use, and allows for immediate drill down for further analysis, without hitting production systems.

Collaborative Environment— team members can log into a common repository, view the same data, and contribute to prioritizing and determining appropriate actions to take for addressing anomalies, improvements, integration rules, and monitoring thresholds.

Profiling Repository—the repository stores metadata created by reverse-engineering the data. This metadata can be summarized, synthesized, drilled down into, and used to recreate original source record replicas. Business rules and data standards can be developed within the repository to run against the metadata or deployed to run systematically against production systems, complete with alert notifications.

Robust profiling functions—in addition to basic profiling, TS Discovery provides advanced profiling options such as: pattern analysis, soundex routines, metaphones, natural key analysis, join analysis, dependency analysis, comparisons against defined data standards, and regulation against established business rules.

Improve Data Immediately—data can be cleansed and standardized directly using TS Discovery. Name and address cleansing, address validation, and recoding processes can be run using TS Discovery. Cleansed data is placed in new fields, never overwriting source data. The cleansed files can be used immediately in other systems and business processes.

Helpful Modeling Functions—data architects and modelers rely on results from key integrity, natural key, join, and dependency analyses. Physical data models can be produced through the effects of reverse engineering the data, to validate models and identify problem areas. Venn diagrams can be used to identify outlier records and orphans.

Monitoring capabilities—monitor data sources for specific events and/or errors. Notify users when data values do not meet predefined requirements, such as unacceptable data values or incomplete fields.These powerful features give users the environment necessary to understand the true nature of their current data landscape and how data relates across systems.

TS Discovery: Investing in the Future

A data profiling solution cannot exist in a vacuum because it is also a part of a larger process. While data profiling is a way to understand the condition of data, TS Discovery provides bridges to larger initiatives for data integrations, data quality, and data monitoring. Trillium Software is committed to expanding those bridges. It continues to innovate and provide new ways to track, control, and access data across the enterprise. It also continues to integrate TS Quality functionality into TS Discovery to establish and promote high quality data in complex and dynamic business environments.

Harte-Hanks Trillium Software
www.trilliumsoftware.com

Corporate Headquarters
+ 1 (978) 436-8900
trlinfo@trilliumsoftware.com

European Headquarters
+44 (0)118 940 7600
trillium.uk@trilliumsoftware.com

HARTE HANKS TRILLIUM SOFTWARE