

Kimball Design Tip #59: Surprising Value of Data Profiling

By Ralph Kimball

Data profiling is something of a quiet little corner of data warehousing. I suspect that most of us think of data profiling as something you do after most of the ETL system has been built. In this view, data profiling checks for small anomalies in the data that may require cleanup before the real production data is delivered. Finding these anomalies would seem to save the data warehouse team from little surprises after going into production.

During the past year I have dug deeply into the back room ETL processes required to build a data warehouse while working on a new ETL book with Joe Caserta. Perhaps the biggest revelation of the whole project has been discovering how undervalued data profiling is in the average data warehouse project.

What is data profiling?

Data profiling is the systematic up front analysis of the content of a data source, all the way from counting the bytes and checking cardinalities up to the most thoughtful diagnosis of whether the data can meet the high level goals of the data warehouse.

Data profiling practitioners divide this analysis into a series of tests, starting with individual fields and ending with whole suites of tables comprising extended databases. Individual fields are checked to see that their contents agree with their basic data definitions and domain declarations. It is especially valuable to see how many rows have null values, or have contents that violate the domain definition. For example, if the domain definition is “telephone number” then alphanumeric entries clearly represents a problem. The best data profiling tools count, sort, and display the entries that violate data definitions and domain declarations.

Moving beyond single fields, data profiling then describes the relationships discovered between fields in the same table. Fields that implement a key to the data table can be displayed, together with higher level many-to-1 relationships that implement hierarchies. Checking what should be the key of a table is especially helpful because the violations (duplicate instances of the key field) are either serious errors, or reflect a business rule that has not been incorporated into the ETL design.

Relationships between tables are also checked in the data profiling step, including assumed foreign key to primary key relationships and the presence of parents without children.

Finally, data profiling can be custom programmed to check complex business rules unique to a business such as verifying that all the preconditions have been met for granting approval of a major funding initiative.

Hopefully as I've been describing the “features” of data profiling, you have thinking that data profiling really belongs at the very beginning of a project, where it could have a big effect on design and timing. In fact, I have come to the conclusion that data profiling should be the mandatory “next step” in every data warehouse project after the business requirements gathering. Here are the deliverables of data profiling that I have come to appreciate during my recent ETL research project:

- A basic “Go – No Go” decision on the project as a whole! Data profiling may reveal that the data on which the project depends simply does not contain the information from which the hoped for decisions can be made. Although this is disappointing, it is an enormously valuable outcome.
- Data quality issues that come from the source system that must be corrected before the project can proceed. Although slightly less dramatic than canceling the whole project, these corrections are a huge external dependency that must be well managed for the data warehouse to succeed.
- Data quality issues that can be corrected in the ETL processing flow after the data has been extracted from the source system. Understanding these issues drives the design of the ETL transformation logic and exception handling mechanisms. These issues also hint at the manual processing time that will be needed to resolve data problems each day.
- Unanticipated business rules, hierarchical structures, and FK-PK key relationships. Understanding the data at a detailed level flushes out issues that will permeate the design of the ETL system.

Finally a big benefit of data profiling that perhaps should be left unstated (at least while justifying the data warehouse to executives) is that data profiling makes the implementation team look like they know what they are doing. By correctly anticipating the difficult data quality issues of a project up front, the team avoids the embarrassing and career-shortening surprises of discovering BIG problems near the end of a project.