

Profiling Time-Dependent Data

Data profiling is the process of analyzing actual data in order to understand its true structure and meaning. It is of critical importance because in most organizations existing metadata is incomplete, incorrect, and obsolete. Thus, data profiling must be the first step in any data-driven project.

With proliferation of efficient tools, data profiling has become one of the most common activities in data management. Unfortunately, many data profiling initiatives do not go beyond basic column profiling, that is gathering summary counts and statistics along with frequency and distribution charts for individual data fields. The main reason is that, historically, data profiling tools were built for column profiling even though they can be used to do a lot more.

While column profiling produces a wealth of valuable metadata, it falls far short of the mark when dealing with time-dependent data. Consider a simple example. Imagine you are dealing with a payroll table, which contains history of compensation data by employee-paycheck. The table will likely have the following fields:

- Employee ID
- Paycheck Effective Date
- Payroll Code
- Pay Amount
- A few more fields with additional details

Column profiling will easily produce the distribution of Paycheck Effective Dates and the frequency chart for Payroll Codes. We may then learn that the table has 10+ years of compensation history and 133 distinct payroll codes. We may identify and count the records with missing payroll codes and with strange negative pay amounts. We may identify payroll code “ZZ”, which is sometimes used in place of a missing value (often referred to as a default value).

All this information is extremely valuable. However, it leaves many open questions. Here are but a few:

1. *How much history do we have for each Payroll Code?* Indeed, the fact that the table has 10 years of history overall does not tell anything about the history for individual payroll codes. Some may only have been tracked for the last 6 months, while others may be historic and no longer used. Without this knowledge, we can make wrong assumptions about data availability.
2. *Do the Payroll Codes change meaning over time?* Indeed, it is quite typical for the same code to be used to track different types of compensation during different periods of time. Such situations are especially common as a result of database consolidations after corporate mergers and acquisitions. Without this knowledge, the data will be likely misused in any aggregate reports.
3. *Do the timestamps (Paycheck Effective Dates) follow certain patterns?* Indeed, it is quite common for some types of payments (such as year-end bonuses) to be only made during certain months. Other payments only fall on the first or last day

of the month, or only on business days. Knowledge of such patterns is key to the design of data quality rules that will find incorrect or missing data.

Unfortunately none of these questions can be answered from the analysis of individual column profiles. A different set of techniques is required – time-dependent data profiling.

The objective of time-dependent data profiling is to learn how much history exists for different data categories, does it follow any predictable patterns, and does the data meaning and patterns change over time. There are a great variety of techniques for time-dependent data profiling, ranging from simple timeline and timestamp pattern profiling procedures to complex approaches for analysis of event histories and state-transition models. Some are rather basic, while others are quite advanced and involve multi-dimensional analysis.

While I am not aware of any existing data profiling tools that explicitly target time-dependent data, in many cases the desired information can be gathered using column profiling tools and some simple SQL queries. But, do not expect a magic button. Time-dependent data profiling requires skill, experience, and “creative thinking”. The real challenge is to understand what to profile, how to organize the results, and what to look for. As usual, the devil is in the details.