

Data Profiling: Myth and Reality

Data profiling has become one of the most common activities in IT, and for very good reasons. Every data quality management guru will tell you that data profiling is the first step towards better information quality. Every data warehousing professional knows that you must profile the source data before implementing a new BI application. A data migration consultant will place data profiling on the first page of the project plan. Master data management starts with data profiling and it is a cornerstone of any metadata repository. So, we all know that data profiling is very important.

With proliferation of data profiling tools we also all know what data profiling is. Even those of us who have never done profiling! To test this theory I actually like to ask attendees of my classes at various conferences: “What is data profiling?” I get a forest of raised hands. The most common answer is:

Data profiling is what those tools do, you know. Sift through the tables and give you all kinds of statistics about each column.

Unfortunately this is a very simplistic and erroneous view.

Let me draw an analogy. Here is the definition of cooking from Wikipedia.

Cooking is the act of preparing food for eating by application of heat.

So, cooking is what those stoves do, you know. Of course, real cooking is not that simple. Wikipedia itself quickly qualifies the definition a couple sentences later:

Cooking is the process of selecting, measuring and combining of ingredients in an ordered procedure in an effort to achieve the desired result. Factors affecting the final outcome include the variability of ingredients, ambient conditions, tools, and the skill of the individual doing the actual cooking.

That explains why I cannot achieve “the desired result” when my wife is absent, despite having the variability of ingredients, ambient conditions, and a good stove.

The same is true with data profiling. It is not what the tools do. It is what **you do** with them. Those of us who have done profiling know that having a tool sift through the tables, collect statistics, and build nicely-looking frequency charts is just a first step. Analyzing these raw data and looking for meaningful useful information is the key to success and that part proves to be far more difficult and time consuming than clicking on a few buttons within a magnificent data profiling tool.

Let’s consider a simple example – a table in the human resources database with an attribute named “DateOfHire”. According to the data dictionary it stores original hire date for all company employees. Now, maybe it does and maybe it does not. We cannot know for sure without analyzing actual data. One way to find out is to look at every record and check what it says, but this is enormously time consuming.

So a better idea is to sift through all the records automatically and build a value frequency chart. Now we must analyze it. Here are some possible observations.

- We may see 534 employee records with DateOfHire value 7/4/2006. Has the company hired that many new employees on the same day? Or, maybe, the company acquired a subsidiary on that day and it was the date of acquisition that was entered into the “DateOfHire” field for all employees of that subsidiary.
- We may find 923 employee records with DateOfHire value 1/1/1980. Sounds strange if we know that the company was not even founded until 1995. Maybe these are not real values, but rather substitutes for missing values used during data entry because the data entry screen does not allow blanks.
- We may further discover that 1,575 employee records have NULL value. Do these values represent data errors? Or, maybe, the company tracks period of employment for its contractors in a separate table, so these missing values are legitimate.
- Finally, we may observe that the value distribution for DateOfHire shows 99.9% of all values after year 2003. This is strange for a company that was in business since 1995. Maybe, for some obscure reason dating back to prior system conversion, all hire dates for older employees are stored in a different table.

The frequency chart itself does not offer any answers! Instead it gives us an opportunity to find the right questions to ask about your data. This is really the key to data profiling. It is the process of analyzing the data and identifying right questions to ask. This is the work for human data experts, rather than tools. Of course, this work is time-consuming when you are looking at the frequency charts and value distribution for 1,000 attributes in 100 tables.

This realization unfortunately comes too late for companies that purchase new data profiling tools and expect to immediately save time and money. The opposite happens: better data profiling tools create more work for data analysts. The problem is especially disastrous when companies lay off data analysts expecting the profiling tools to replace them.

Conclusion

The rumors of importance of data profiling are absolutely true: it pays off more than any other investment in data quality management. However, significant effort by the data analysts is required to make it work. As is common in many areas of IT, good software by itself does not solve problems; only people who know how to use it do!