DATA PROFILING: WHAT, WHY AND HOW?

A TWO-PART INTRODUCTION AND OVERVIEW OF DATA QUALITY AND PROFILING

BY JASON HOVER

PART ONE: INTRO INTO DATA QUALITY

Like it or not, many of the assumptions you have about your data are probably not accurate. Despite our best efforts, gremlins inevitably find their way into our systems.

The end result – poor data quality – has a host of negative consequences. This brief article will provide an introduction to data quality concepts, and illustrate how data profiling can be used to improve data quality.

WHAT IS DATA QUALITY?

DATA QUALITY IS A MEASURE OF THE ACCURACY, VALIDITY AND COMPLETENESS OF DATA.

Is the data of sufficient quality to support the business purpose(s) for which it is being used?

Are any specific issues within the data decreasing its suitability for these business purposes?

DO MOST ORGANIZATIONS HAVE A DATA PROBLEM?

The short answer is "yes"; a study by Gartner estimated "more than 25 percent of critical data within fortune 1000 enterprises" to be flawed.

With the myriad of ways that data is captured (online transactions, automated device capture, manual screen entry, spreadsheet uploads, direct database changes), there are many opportunities for flawed data to enter source systems.

SO WHAT? DOES IT MATTER?

The costs of poor data quality are ongoing and substantial. A report from The Data Warehouse Institute concluded that data quality problems cost U.S. businesses more than $600 billion a year and that poor data quality leads to failure and delays of many high profile IT projects. Lack of "trust in the data" due to poor data quality leads to reduced or discontinued BI usage among consumers. Poor data quality also has legal/regulatory implications, especially in the wake of Sarbanes-Oxley, as accurate data is required in order to have accurate financial reporting.

"data quality problems cost U.S. businesses more than $600 billion a year"

PART TWO: DATA PROFILING OVERVIEW

WHAT IS DATA PROFILING, AND HOW CAN IT HELP WITH DATA QUALITY?

Data Profiling is a systematic analysis of the content of a data source (Ralph Kimball).

You must look at the data; you can't trust copybooks, data models, or source system experts. It is "systematic" in the sense that it's thorough and looks in all the "nooks and crannies" of the data. You have to know your data before you can fix it.

WHAT TYPES OF ANALYSIS ARE PERFORMED?

Completeness Analysis - How often is a given attribute populated, versus blank or null?

Uniqueness Analysis - How many unique (distinct) values are found for a given attribute across all records? Are there duplicates? Should there be?

Values Distribution Analysis - What is the distribution of records across different values for a given attribute?

Range Analysis - What are the minimum, maximum, average and median values found for a given attribute?

Pattern Analysis - What formats were found for a given attribute, and what is the distribution of records across these formats?

WHAT ARE SOME REAL-WORLD SCENARIOS?

Data profiling can add value in a wide variety of situations. The basic litmus test "is the quality of data important for this initiative?"

If the answer is "yes", then data profiling can help as it can quickly and thoroughly unveil the true content and structure of your data. Some example scenarios include:

Data Warehousing/Business Intelligence (DW/BI) Projects
These projects involve gathering data from disparate systems for the purpose of reporting and analysis. Data profiling can help ensure project success by:

- Identifying data quality issues that must be corrected in the source system
- Identifying issues that can be corrected in ETL processing
- Discovering unanticipated business rules
- Even potentially providing a "no-go" decision on the project as a whole

Data Conversion/Migration Projects

These involve moving data from a legacy system to a new system. Data profiling can help reduce project risk by:

- Identifying data quality issues that must be handled in the code that moves data from the legacy system to the new system
- Identifying data issues that may require a change to the target (new) system

Source System Data Quality Initiative

These projects endeavor to assess and improve the data quality of a given source system, seeking to fix existing issues as well as avoid those issues in the future. Data profiling can help maximize project ROI by:

- Highlighting the areas within the system suffering from the most serious and/or numerous data quality issues
- Identifying issues that may be the result of bad user input or errant system interfaces

DATA PROFILING THE OLD WAY

THE “MANUAL” APPROACH

Traditionally, data profiling required a skilled technical resource who could manually query the data source using Structured Query Language (SQL). There is often a disconnect between the business analyst who knows what the data should be, and the technical programmer who knows SQL.

DATA PROFILING THE NEW WAY

BENEFITS OF USING DATA PROFILING SOFTWARE

There are many benefits to be reaped by using software to automate the data profiling process, including:

Increased Speed (Resulting in Hard Dollar Savings)

Industry estimates for manual data profiling are approximately 3-5 hours per attribute; by using a data profiling tool, this can be reduced to 15-30 minutes per attribute.

- Sample ROI, assuming 1500 attributes: $281,250 minus the cost of data profiling software

More Thorough Analysis

With a manual approach, generally only a subset of the attributes and the rows are tested; with a data profiling tool, a thorough evaluation of the data can be performed.

Quote from DM Review: "Smart organizations are abandoning manual methods in favor of automated data profiling tools that take much of the guesswork out of finding and identifying problem data."

Common Repository

Data profiling tools provide a common repository for storing data profile results and other key metadata such as notes made during analysis.

- Data profile information is centralized
- Entire team can share and leverage the information

AVAILABLE TOOLS

A variety of options exist in the marketplace to help ease the challenge of data profiling ranging in capabilities and price. Tools like Datiris Profiler and Informatica Data Quality have been successfully deployed by a myriad of organizations. Implemented in the right way, such tools stand to sculpt the data profiling landscape, by reducing effort, broadening scope, and improving consistency across all data quality initiatives.

PARTNER WITH US

Do you have a current Data Quality initiative? How is your Data Quality project progressing? We'd like to hear from you. Take a few minutes to send us an email and let us know about your project. If you'd like further assistance with your current project, feel free to contact us directly. We'd be happy to talk with you and develop a proposal around your project.

About the Author:

Jason Hover has been engaged in the DW/BI and data quality/profiling space for over 13 years. He is a co-founder of Datiris, created by business intelligence & data warehousing professionals to meet the data profiling needs encountered on real projects.