

Using Dictionary Tables to Profile SAS® Datasets – Poster

Phillip Julian, Bank of America, Charlotte, NC

ABSTRACT

Data profiling is an essential task for data management, data warehousing, and exploring SAS® datasets. TDWI (<http://tdwi.org>) extends the usual definition of data profiling to include data exploration. This paper presents two SAS programs – Data_Explorer and Data_Profiler – that implement the TDWI definition.

These SAS programs are low-cost, free solutions for data exploration and data profiling. Data_Explorer searches for all SAS datasets, and gathers essential dataset and file attributes into a single report. Data_Profiler summarizes the values of any SAS dataset in a generic manner, and eliminates the need for custom SQL queries to learn what the data looks like. Because the profiler uses an efficient two-pass algorithm, a brute force approach, that includes everything plus the kitchen sink, can consume fewer resources than custom SQL queries. Profiler results are more complete because you get categorical details for all the columns of very big datasets.

These programs have been used in banking and state government, and should be useful in the pharmaceutical industry for validating SAS datasets and managing data content and changes in large data repositories.

NOTE

No paper is available for this poster. For details on this material, see paper BB-06, which explains the poster and includes the poster within the paper.