A Web-Based Approach for Combining Metadata, Search, and Data Profiling

Jeff Butler, Internal Revenue Service (Research, Analysis, and Statistics),

With nearly one petabyte of data serving hundreds of IRS research analysts, the Compliance Data Warehouse (CDW) brings together data and analytics for the world's largest tax agency on a massive scale. Metadata are available for more than 25,000 unique database columns and over 500,000 separate attributes through the CDW website and represents the largest database-driven repository of metadata in the IRS. Data profiling capabilities are also available through the CDW website, letting users quickly explore patterns through frequency tables, statistical distributions, trends, and geographic maps—often on billions of rows of data.

This paper discusses a web-based approach used by CDW for combining metadata, search, and data profiling into a single experience. Rules for standardizing metadata are outlined—including column definition templates, lookup reference tables, data types, and other artifacts. A database-oriented format is proposed for organizing and displaying search results for data stored in relational databases. Finally, a solution is presented for adding data profiling capabilities to both metadata and search results to let users dynamically explore and visualize patterns in data using a set of flexible ad-hoc database queries. Since it was implemented in 2009, this dynamic data profiling capability now accounts for roughly one-third of the 3,000 average daily database queries executed in CDW.