TheDataWeb

TheDataWeb

TheDataWeb is a platform for bringing together many different datasets and disseminating them to users. It can be used by one organization, or it can link multiple databases across different agencies into a single virtual data repository. The data may reside in different formats, database engines, and/or on different servers with unique operating systems and security models, but the user experience is the same regardless of access.

Historically, once a dataset was created each group who wanted to act on the data made a copy on their own machine with which to work. These groups often re-packaged the data and disseminated them as millions of individual tables or spreadsheets. Instead of one copy of the data, we have the original data, the multitudes of copies of the data, and millions of tabulations. This is a problem because data providers have to maintain multiple copies of data, thus increasing costs and causing inconsistencies when updates are not made across all data products.

TheDataWeb implementation at the U.S. Census Bureau provides access to over 828 datasets with new files added monthly. It provides statistical business intelligence (BI), allowing everyone to access public government databases, as well as providing secure access to authorized users of non-public data. Even though the data reside on many different servers, they can be combined into webpages that act as data-driven applications and dashboards, which turn numbers into meaningful information easily understood by the public, used by policy-makers in decision making, or reviewed by analysts concerned with data quality.

Designed over sixteen years ago, TheDataWeb architecture is continually optimized, updated, enhanced, and extended to ensure parity with state of the art technological innovation. One of its major strengths is the flexibility of the design. TheDataWeb is designed as a Service Oriented Architecture (SOA) and is generalized to be platform agnostic and to support an enterprise solution to data dissemination. SOA systems allow for agile development and can be quickly adapted to new hardware, software, and security models, thus leading to cost savings.

TheDataWeb's power and flexibility have their foundation in its metadata, which is a generalized and evolving repository whereby the metadata are stored separately from the data themselves, allowing for maximum flexibility. Metadata include information about where the data are stored and how the data are structured which allows the data to be retrieved across different servers. The metadata also include other information that are usually stored in technical documentation, like labels for the variables and values, and comparability information across time. The metadata are stored separately from the datasets, allowing a single source of data to be re-purposed in a variety of ways. This is a core tenet of the Digital Government Strategy and has been part of TheDataWeb philosophy from the beginning.

Features of TheDataWeb Architecture:

- Works with any type of data, including: microdata records, longitudinal linked records, time series records, and aggregate (pre-tabulated) records.
- Works with multiple data engines, including: MySQL, Oracle, Sybase, SQL Server, and DataWeb Transverse File System
- Uses industry best practices to provide fast response on very large databases.

- Minimizes space requirements by dynamically defining and creating tables as needed rom a single underlying dataset.
- Supports over 100 different Census-defined geographic summary levels.
- Supports the concept of geographic vintages, allowing the system to make use of information about changes in geographic boundaries.
- Includes a generalized mapping solution for all datasets by associating each dataset to a particular set of shape files for that geographic vintage.