DataOS®: Metadata and the Data Catalog for the Modern Data Stack

Modern

Anyone who works with data knows that it's long past time for data catalogs to catch up with the rest of the modern data stack. Data is no longer consumed primarily by the IT team. Today, data teams include data analysts, data scientists, product managers, business analysts, citizen data scientists, and more. Each of these people has their own favorite data tools and even different languages for describing data.

Too often collaboration dissolves into chaos and confusion. Frustrated questions like, "What does this column name mean?" and "Why are the numbers on the dashboard wrong again?" slow data teams to a crawl. To help ensure this doesn't happen, data teams and other users can leverage metadata for answers instead.

Metadata facilitates collaboration

Metadata is a solution to enable collaboration across business units and to make data easier to find and use. Metadata (documentation, queries, history, glossaries, etc.) makes data understandable.

Anyone who has used a library catalog is familiar with metadata; tags like author, date of publication, subject, Dewey Decimal Number and more help you locate a book and determine whether it's useful for what you have in mind. However, the modern idea of metadata dates to the 1990s and the rise of the Internet.

As the Internet grew, data and metadata exploded. IT teams were given ownership of data in most companies and placed in charge of creating an "inventory of data," the way a grocery store might inventory apples and soap. Setting up these inventories and keeping them current were constant struggles for IT.

Data catalogs arose during the Hadoop era (2010s). They evolved as companies realized that they needed to improve the data inventories of the 1990s-2000s by adding new business metadata. The idea was to help the expanding class of business users understand their datasets and put the data in a business context.

Data stacks have evolved but metadata solutions haven't

These early data catalogs were clumsy, and specialized solutions were lacking. The earliest adopters of the modern data stack and most large tech companies resorted to building their own proprietary solutions, such as Airbnb's Dataportal, Facebook's Nemo, LinkedIn's DataHub, Lyft's Amundsen, Netflix's Metacat, and Uber's Databook. Small companies, without the resources for such in-house projects, had to wait for solutions to arrive.

And arrive they did, eventually, with tools such as Apache Atlas. Still, while the rest of the data stack has evolved in the past few years, and tools like Fivetran and Snowflake let users set up a data warehouse in hours once they are

installed, data catalogs have not kept up. Even trying out current metadata tools involves significant engineering time for setup, plus weeks of back and forth with a sales rep to get a demo.

It's time for a metadata solution that is just as fast, flexible, and scalable as the rest of the modern data stack. In January, 2021, Prukalpa Sankar wrote on towardsdatascience.com, "[I]n the next few years there will be the rise of a modern metadata management product that takes its rightful place in the modern data stack." These new data catalogs will be based on principles of data and data use that have developed alongside the data stack.

There's more to your data assets than tables

Today's BI dashboards, code snippets, SQL queries, models, recordings, presentations, and Jupyter notebooks are all data assets. All can be searched and analyzed for valuable information. All can be enriched and made more usable through appropriate metadata.

Metadata itself is a data asset — and "big data"

A modern data catalog should leverage metadata as a form of data that can be searched, analyzed, and maintained in the same way as all other types of data. The ability to process and understand metadata will help teams understand and trust their data better.

For example, query logs are just one kind of metadata available today. SQL query logs, properly analyzed, allow us to create column-level lineage, assign a popularity score to every data asset, and even deduce the potential owners and experts for each asset. Quality ratings from users, indexed by a data source, can identify source problems that can be addressed to improve data quality throughout the organization.

Visibility into all your data, rather than siloed point solutions

Today's data catalogs have greatly improved discoverability, but they still do not give organizations a "single source of truth" for their data. Information about data assets is usually spread across tools for data lineage, data quality, data preparation and cleanup, and more. Data silos still impede discovery and enrichment. And dark data remains dark, hidden, and unused (let alone catalogued).

DataOS meets the need

DataOS has these principles at its heart. Its metadata engine, Metis, allows DataOS to apply rich metadata covering all aspects of a dataset, from lineage to documentation. Sitting atop your data ecosystem, DataOS accesses every dataset in your organization, without moving the data, to eliminate silos and dark data.

For more information, or to arrange a demonstration (in days, not weeks), email us at **info@TMDC.io**.

BY E. WALLACE

TheModern DataCompany

The Modern Data Company
306 Cambridge Ave
Palo Alto, CA 94306
TheModernDataCompany.com
info@TMDC.io

DataOS®: Metadata and the Data Catalog for the Modern Data Stack