Datum

Datum: Scientific Data Catalog

September, 2024

Overview

The data catalog market is currently flooded with a myriad of different products, but none serve the scientific community well. There are cloud-native tools like Databricks, Snowflake,to on-premise solutions like Collibra and Datahub. The common failing of all these tools however, is their inability to serve the scientific data community directly. Most catalogs are targeted towards financial, health, or user data - not sensor or scientific domain data. They also prioritize integrations that often don't exist or are just starting to be used in the scientific realm - all while ignoring common scientific tools and file types.

Datum is a catalog which targets the scientific data directly, including the tools and networks in which those tools are used. We work with the producers and consumers of the data where they are, targeting cloud and on-premise with a focus on classified networks.

Goals

1. **Seamless integration with scientific data:** Datum's automated scanning is built to extract metadata from scientific data formats without forcing users to convert to formats with less fidelity.
2. **Available where the data lives:** We target both classified and non-classified systems, focusing on enabling the use and access of the catalog wherever the data might live.
3. **High assurance and high availability:** You should never worry about the catalog being available to users, or about the data becoming lost or corrupted. We focus on

correctness from the ground up - mimicking the needs of the scientific reporting community.

4. **Semantic and AI-enabled search:** Data should be easily found and reasoned about, with minimal infrastructure. Datum is designed to enable semantic and AI augmented search with a minimum of infrastructure requirements.

Technical Features

Note: The features listed below are still under development and may change, slightly, upon final delivery of the product.

File Formats - Datum has the ability to read additional metadata and provides processing pipelines for the following file formats: Plain Text, PDF, LaTeX, HTML, Open Document Format (.odt), XML, CSV/TSV (and other standard delimiters), OpenDocument Database and Spreadsheets, Geo-Referenced TIFF, Common Data Format, HDF/HDF5, LabView TDMS, Excel, DeltaTables, Parquet, Apache Iceberg, Apache Hudi and many others.

Metadata Collection - Scanners for the local and networked file systems and cloud storage providers. Network integration with common databases such as MSSQL and MySQL.

User Plugin System - Users are able to provide either file processing, metadata extraction, or sampling plugins in the programming language of their choice.

Authentication/Authorization -: OIDC integration, SCIM provisioning and EntraID integration out of the box. Full user and group management system with a “least privilege” operating mode.

Governance - Customizable data governance platform; dictate and enforce required metadata, enforce data embargos, and enforce user agreements and NDAs before data access. Ability to create health checks on data, rejecting abandoned or poorly curated data and automatically removing it from the search index. Ability for users to submit corrections.

Search - Semantic search is a first class citizen. No licenses to expensive, external software required. Integrated use of vectors and vector-based search allows for AI agent integration at all levels of operation.

Metadata Model - Display and control data's lineage and connections to other data and data directories. Data is modeled after a filesystem - an organization instantly recognizable and navigable by most any user.

CLI and SDK - Ships with a Command Line Interface (CLI) tool and with a fully-featured Python SDK. This allows for rapid and programmatic use of Datum by every level of user.

Minimal Infrastructure - Datum ships as a single executable file and can be run on any operating system and most CPU architectures. Datum has no reliance on external databases, search indexing tools, or other outside services - and it runs equally well on edge computing devices, cloud services, or in a clustered HPC environment.