A Quick Guide to Data Catalog

What is a Data Catalog?

A data catalog is a collection of metadata, combined with data management and search tools that helps data consumers find the data that they need. The data catalog serves as an inventory of available data and provides information to evaluate the fitness of data for intended uses. -- Adapted from: Wells, Dave. (2020, January). Introduction to Data Catalogs. Eckerson Group.

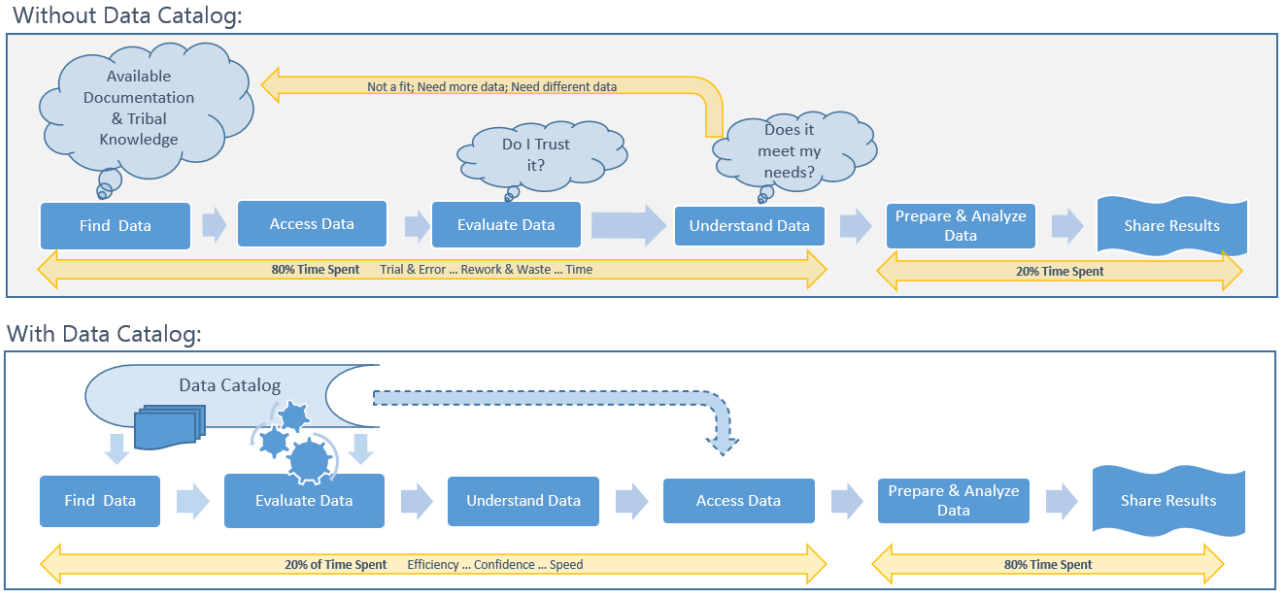
Some tools and use cases differentiate between the internal, technical metadata store (compute catalog) and the external-facing consumer/contributor metadata store and user interface (user data catalog).

Compute Catalog	**Stores all the structured information about the various databases, tables, including column and column type information.**
User Data Catalog	Stores profile information about the data assets [and/or datasets]; supports data asset discovery and access; makes structure metadata accessible to data consumers.

Challenges of Data Management With and Without Data Catalog

Organizations that do not have a catalog of data assets and their sources are impeded in their efforts to perform data analysis and data exploration across the organization. With a catalog, data consumers can spend less time trying to understand the data and instead spend the majority of their time analyzing or using the data to reach their intended goal.

Figure 1: Challenge Without Data Catalog & Solution With Data Catalog



A data catalog provides a central location for data consumers to explore an organization's available data and for data contributors to share and exchange knowledge and insight of their data assets. Without a catalog, the valuable insights around data and data assets remain fragmented; consumers are left to rely on

tribal knowledge and their own devices to gain an understanding of a data asset's contents and purpose. Data consumers can more easily discover new data assets and understand their purpose, and data contributors can benefit by reducing the burden of responding to questions and queries about their data that could easily be answered by the descriptive metadata in a catalog.

Who Uses a Data Catalog?

Data Stewards can see how their data fits in with the organization as a whole and can use that higher perspective to plan for proper Data Management and Data Quality Assurance.

Data/Business Analysts benefit from the descriptive metadata providing context to data assets for the purpose of extracting business value.

Data Engineers & Data Scientists will be able to discover, understand, and utilize existing data while avoiding the creation duplicate data.

Data-Focused Executive Leadership become better informed about their organization's data landscape and can make better informed strategic decisions for the organization.

Other Line of Business Data Consumers will use a data catalog for a wide range of data inquiries.

-- Ehtisham Zaidi, Guido De Simoni (2019).

Features of a Data Catalog

There are many modern Data Catalog tools available that offer various features, including varying degrees of automation. Some of the following features can be found among today's available catalog tools:

- **Native Integration:** Data may be coming from a database, data warehouse, data lake, or any number of additional sources. The Data Catalog helps normalize data from various sources to provide a single source of reference for all of the organization's data.
- **Self Service:** Data catalogs are self-service platforms, giving Users quick and convenient access to the organization's data.
- **Marked Relationships:** Data catalogs make the relationships between data assets visible by connecting databases, tables, or data assets that relate to a common entity.
- **Data Classification:** Data sources and fields are tagged, either manually or dynamically depending on the tool, to allow powerful search and categorization capabilities. Identifying where data sets contain PII (Personally Identifiable Information) or other sensitive data can assist organizations to stay in compliance with internal security policies.
- **Data Profiling:** Data catalogs can include an analysis on a database or data asset beyond simple metadata. Data profiles may include information such as row counts, top values in a column, null counts, distinct value counts, and much more. Users can gain greater insight into the data without performing their own analytics and without even needing to open the data.

- **Data Lineage:** Data catalogs can show the complete history of a piece of data, where it originated from, and what transformations it has taken, with some tools capable of backtracking through the ETL (extract, transform, load) processes that produce integrated data assets.
- **User Feedback and Improvement:** Modern data catalogs allow Users to rate data assets for accuracy and to assign different levels of trust. Better data perception results in better business decisions made from using the data. In some data catalogs, the Users can also add comments or initiate discussions within the catalog tool.

Steps to Creating a Comprehensive Data Catalog

1. Identify valuable data assets to include and identify which are considered authoritative data assets.
2. Adopt a data catalog Standard like the Data Catalog Vocabulary (DCAT) standard, a Resource Description Framework (RDF) vocabulary designed to facilitate interoperability between data catalogs published on the web.
3. Collect structural metadata from the data sources. Some data catalog tools can crawl through databases and retrieve metadata automatically.
4. Upload an existing data dictionary (or build one) of each data source.
5. Create data asset profiles, analyze database tables or data assets and provide that analysis in the data catalog.
6. Categorize data assets into Data Domains or other groups.
7. Mark relationships among data, making connections between database tables, data assets, domains, etc. related to the same entity. This can be done manually or through a tool's advanced algorithms.
8. Build data lineage by identifying the data assets that are integrated from two or more data sets, and by analyzing the ETL processes that are performed on each data asset.
9. Secure sensitive metadata that needs protecting. Sensitive metadata may need to have special permissions for who can access or view it.
10. Set up the catalog in an easy to access location, e.g. within a web app.

Use a Data Catalog Template

CMS provides an Excel template to capture all of the necessary metadata for a basic data catalog. The tables and attributes in the template are aligned with attributes used by the CMS Enterprise User Data Catalog (EUDC), and this template can be used as a stepping stone towards integration with the EUDC and/or the Enterprise Data Lake (EDL).

Data_Catalog_Template_v2.xlsx

Examples of Public-Facing User Data Catalogs

In addition to providing data catalog features, implementations can be extended to actively facilitate data access.

- **Data.CMS.Gov** https://data.cms.gov/ - provides data on Medicare, other programs, and the Marketplace.
- **US Data Catalog** https://catalog.data.gov/dataset - opens over 250,000 data sets to the public.
- **USGS Science Data Catalog (SDC)** https://data.usgs.gov/datacatalog - publishes geological and weather data and serves its metadata to data.doi.gov, data.gov, and other government agencies.