International Journal of Engineering, Science and Mathematics
Vol.9 Issue 9, September 2020,
ISSN: 2320-0294 Impact Factor: 8.445
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell’s Directories of Publishing Opportunities, U.S.A

Enabling Data through Data Catalogs

Sowmya Tejha Kandregula | Data Governance Leader
Brambleton, Virginia | sowmyateja@gmail.com

Abstract

Organizations with successful data catalog implementations have experienced substantial improvement in enhancing the pace and quality of data analysis, allowing an effectual engagement of data analysts and data scientists for accuracy in decision making.In the age of big data and business intelligence, data catalogs are becoming the essence of metadata management, helping and guiding data users better understand their data and its importance.Before implementing a data catalog, one needs to understand its definition, need, features and mode of usage.

Keywords

Data Management, Data Governance Office, Critical Data, Data Custodian, Enterprise Metadata Collection, Metadata Management

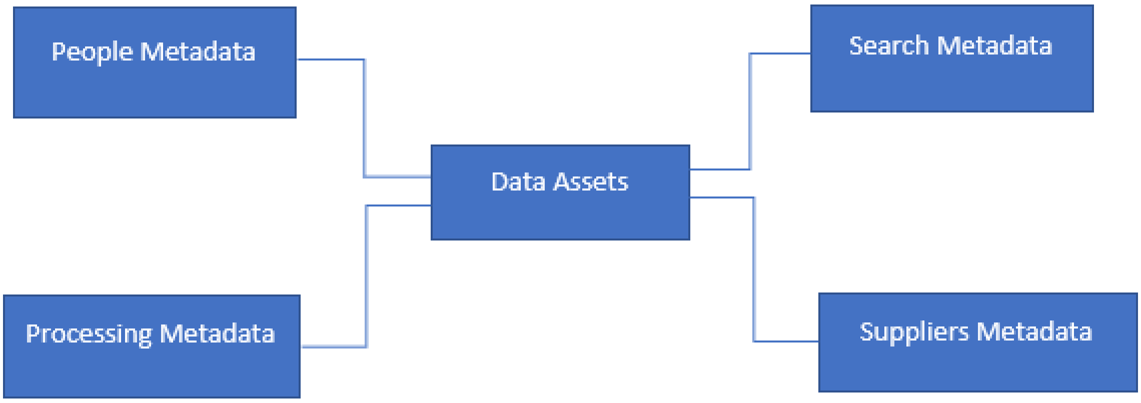
Introduction

A data catalog can be defined as a collection of metadata, typically used for data management with query access to help analysts and other data users find the data that they need. A data catalog contains lots of critical information about each piece of data, such as the data’s profile (statistics or informative summaries about the data) and lineage (how the data is generated). Not surprisingly, data catalogs serve as the go-to spot for data analysts, data stewards, data scientists and others to find and understand relevant datasets to build insights, discover trends, and identify new products for the company.

Data catalogs serve as an inventory of available data within the organization and provides access to evaluate the fitness of data for its intended use. With all its benefits, the effectiveness of the data catalog depends on the central capacity to provide a collection of metadata.

Why Data Catalog?

A data catalog focuses on data assets and connects the data sets within the assets with its related metadata (data about data) to help the users of the data understand it better.



- **Data Assets:** can be files, databases, or applications that data users need to find and access to generate insights for decision making. They could reside in data lakes, warehouses, or any other shared data resource.

- **People Metadata:** provides information on those who work on data assets. They could be consumers, curators, SMEs, or stewards.

- **Search Metadata**: supports tagging and searching for data within the assets using keywords to help people find the data.

- **Processing Metadata**: describes the transformations and derivations data goes through and how it is managed through its lifecycle.

- **Supplier Metadata**: describes data acquired from external sources providing insights on the sources and subscription and licensing constraints.

Building a Data Catalog

1. *Collecting metadata from organizational data systems*

The first step for building a data catalog is collecting the data's metadata. Data catalogs use metadata to identify tables, files and databases. The catalog crawls through various applications including but not limited to data management platforms, analytics and business intelligence platforms and custom applications to bring metadata (not the actual data) to the data catalog.

2. *Building a data dictionary*

The second step is to build a data dictionary or upload an existing one into the data catalog. A data dictionary contains the detailed description of every table or file and all their metadata entities.

3. *Data profiling*

The next step is to contour data to help the stakeholders view and understand the data quickly. These profiles are informative summaries that explain the data.

4. *Building a data lineage*

A visual representation of data lineage helps to track data from its origin to its destination. It explains the different processes involved in the data flow.

5. Organizing data for discovery

In this step, data from a table/file that is arranged in a technical format is organized in a way that would make sense to a business user. Appropriate tools and technologies are required on data assets so that data can be discovered, accessed, and trusted by business users.

How does a Data Catalog work?

A data catalog includes many features and functions related to the core capabilities of cataloging information – collecting data about data that identifies and describes the inventory of usable information. With a lot of usable and sharable data available, it becomes impractical to attempt cataloging as a manual effort.

Automated discovery of data assets on-the-go has become essential. The use of techniques such as artificial intelligence and machine learning for metadata collection and tagging becomes vital to get maximum value from data cataloging using minimal manual effort.

Apart from capturing metadata for the data, other essential features and functionalities captured by a data catalog include:

- **Data Search**: This capability includes a search of facets, keywords, and business terms. Business terms search capabilities are especially essential for non-technical users of data. The search capability can be organized by relevance and frequency of use, in turn providing the search results for the most relevant information.

- **Data Evaluation:** Choosing the right data asset depends on the ability to evaluate their suitability for the particular use case without needing to import the data first. An important feature of data evaluation includes the capability to preview data assets, see all associated metadata, access user ratings and reviews, and view data quality information.

- **Data Access:** The data access should be a seamless user experience with the data catalog implementing the access protocols directly or using access technologies. Data access functions include protection for security, privacy, and compliance of sensitive data.

Different capabilities can be explored based on the type of data catalog. The different types of data catalogs are utilized based on metadata and its importance to the organization.

Data Catalogs based on Metadata Categories

Different data catalog tools rely on the collection and use of a combination of metadata categories.

- **Technical Metadata Management Catalog:** A technical metadata management data catalog captures and provides structural information about the source and target data for integration development and ETL (extraction, transformation, and load). This data catalog mostly relies on:

    - Structural metadata (e.g., data element names, data types, and data element sizes)
    - Supplier metadata (e.g., data asset demographic information)
    - Processing metadata (e.g., data transformations and data derivations)
    - Query metadata (e.g., business glossary and data element definitions)

- **Data Lineage Tool:** A data lineage tool combines:

    - Supplier metadata such as the data owners
    - The details of the original sources from which the data asset is manufactured
    - Data production details with the data transformations, data derivations and the structure of the data processing pipelines from the processing metadata

- **Machine Learning Data Catalog:** A machine learning data asset inventory blends the practical aspects of the production of the data asset from:

    - Structure metadata (data element names, lengths, types)
    - Processing metadata (data transformation, derivations, and pipeline process maps)
    - Query metadata, including the semantic details and historical usage, to produce a searchable data catalog

- **Data Portal:** The objective of a data portal is transparency, and a data portal typically scans and then previews accessible data assets. To enable this, data portals combine:

    - Structure metadata
    - Supplier metadata
    - Query metadata

    The expected outcome is to provide a listing of available data assets, data element metadata, information about the different data sources, and data asset demographics such as several records or size in bytes. It also provides a means for browsing a subset of data instances within the data asset.

- **Data Governance Tool:** Data governance tools ensure data usability by monitoring data quality and alerting data stewards when issues emerge. These tools have evolved from metadata repository products to incorporate the definition of data quality policies and support operational data stewardship processes and procedures.

- **Data Security and Protection Catalog:** These types of data catalogs draw on:

    - User metadata to collect information about the different users, groups, and roles
    - Different classifications pulled from the query metadata
    - Data protection directives from governance metadata to enable the definition and implementation of runtime data protection and security policies

All solutions consist of multiple types of metadata consumed and utilized, and no single catalog tool has the capabilities to satisfy the extent of the need for a data catalog solution. Identifying the right data catalog solution requires attention to the organization's most critical user scenarios and requirements, such as:

- The scope of enterprise-wide business glossaries and data definitions
- Using metadata standards and defined procedures for collecting, documenting, and sharing the different classes of metadata
- Inferring data models and lineage through reverse-engineering
- Attentive data curation that establishes standardized processes for data asset configuration and preparation
- Simplifying intelligent query processing so data consumers can quickly find what they need and enabling data previewing for those seeking data assets to answer ongoing and emerging business questions
- Engineering, implementing, and monitoring data pipelines and the processing stages through which data streams for end-user reporting and analytics
- Operational data governance and assessing existing data governance and stewardship roles and their responsibilities

- Data validation and quality assurance for data trust
- Collaboration among data producers and different data consumers,
- Data content classification and how it relates to data organization and data protection.

Benefits of a Data Catalog

The benefits of data catalogs are reflected in the value and quality of metadata and the capabilities unlocked from it. The analysts observe the best benefits of data cataloging in their analysis. It provides business and data analysts complete visibility into the existing data, their content, and their quality and usefulness.
Quality and efficiency of analysis are substantially improved, and organizational analysis capacity increases without an increase in resources as analysts don't need to spend nearly as much time in finding, sorting, and cleaning data. Some of the most common benefits that can be readily observed by the implementation of the data catalog:

- Data efficiency
- Better clarity from a data context
- Reduced risk of error
- An improvement in data analysis

Conclusion

A data catalog benefits organizations in numerous ways. By using the right data cataloging tool, organizations can automate their metadata management process – including data mapping, data quality and code generation resulting in various downstream benefits including faster time to value (and) improved accuracy for data movement and/or deployment projects.

Data cataloging helps curate internal and external datasets for a range of content authors. It ensures effective management and monetization of data assets in the long-term if linked to broader data governance, data quality and metadata management initiatives. This becomes even more important given the rapidly changing privacy landscape.

References

https://www.eckerson.com/articles/choosing-a-data-catalog

https://www.gartner.com/en/documents/3838463

https://www.dbta.com/BigDataQuarterly/Articles/Driving-Actionable-Business-Intelligence-with-a-Data-Catalog-114562.aspx