Data Catalog - A Brief Introduction

- Data Catalog is a fully managed and scalable metadata management service that empowers organizations to quickly discover, understand, and manage all of their data in Google Cloud.
- It offers a simple and easy-to-use search interface for data discovery, a flexible and powerful cataloging system for capturing both technical and business metadata, and a strong security and compliance foundation with Cloud Data Loss Prevention (DLP) and Cloud Identity and Access Management (IAM) integrations.

Why do you need Data Catalog?

- Most organizations today are dealing with a large and growing number of data assets.
- Data stakeholders (consumers, producers, and administrators) within an organization face multiple challenges:
  - **Searching for insightful data**
    - Data consumers don't know the location and origin of data. They have to navigate data "swamps".
    - Data consumers don't know what data to use to get insights because most data is not well documented and, even if documented, is not well maintained.
    - Data can't be found and is often lost when it resides only in people's minds.
  - **Understanding data**
    - Is the data fresh, clean, validated, approved for use in production?
    - Which dataset out of several duplicate sets is relevant and up-to-date?

    - How does one dataset relate to another?
    - Who is using the data and who is the owner?
    - Who and what processes are transforming the data?

- **Making data useful:**
    - Data producers don't have an efficient way to put forward their data for consumers. If there's no self-service, consumers may overwhelm producers. Several data engineers can't manually provide data to thousands of data analysts.
    - Valuable time is lost if data consumers have to find out how to request data access, request it, wait without a defined response time, escalate, and wait again.

Without the right tools, the challenges become a major obstacle to the efficient use of data. Data Catalog provides a centralized place that lets organizations achieve the following:

- Gain a **unified view** to reduce the pain of searching for the right data.
- Support data-driven decision making and accelerate the insight time by enriching data with **technical and business metadata**.
- Improve **data management** to increase operational efficiency and productivity.
- Take **ownership** over the data to improve trust and confidence in it.

Data Catalog functions

Data Catalog provides three main functions:

- Searching for data entries for which you have access
- Tagging data entries with metadata
- Providing column-level security for BigQuery tables

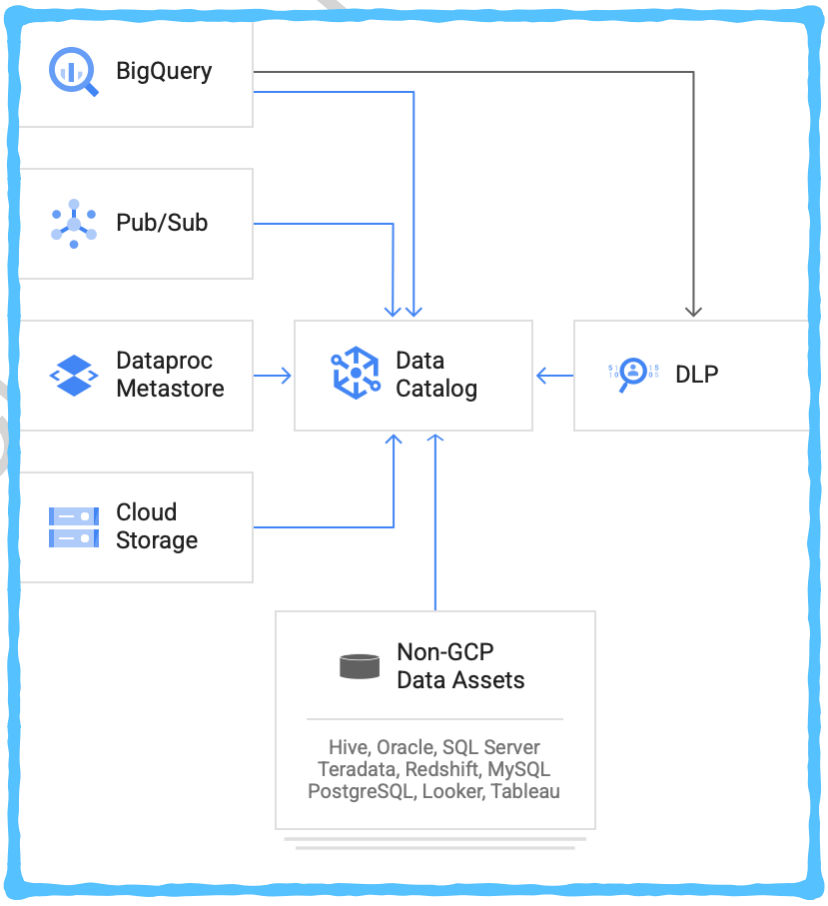
In addition, Data Catalog can leverage the results of a Cloud Data Loss Prevention (DLP) scan to identify sensitive data directly within Data Catalog in the form of tag templates.

Data Catalog metadata

- Data Catalog handles two types of metadata: **technical metadata** and **business metadata**.

How Data Catalog works

- Data Catalog can catalog asset metadata from different Google Cloud systems.
    - You can also use Data Catalog APIs to integrate with custom data sources.
    - After your data is cataloged, you can add your own metadata to these assets using tags.



Automatic catalog of assets

For a given project, Data Catalog automatically catalogs the following Google Cloud assets:

- BigQuery datasets, tables, views.
- Pub/Sub topics.
- Dataplex lakes, zones, tables, and filesets.
- (Public preview): Dataproc Metastore services, databases, and tables.
- (Public preview): Analytics Hub linked datasets.

Access Data Catalog

You can access Data Catalog functionalities using:

- Dataplex UI in the Google Cloud console
- gcloud command-line interface (CLI)
- Data Catalog APIs
- Cloud Client Libraries

Why use Dataplex?

- Dataplex is an intelligent data fabric that helps you unify distributed data and automate data management and governance across that data to power analytics at scale.
- Enterprises have data that's distributed across data lakes, data warehouses, and data marts.
- Dataplex lets you discover, curate, and unify this data based on your business needs, and centrally manage, monitor, and govern this data.
- Dataplex helps you to standardize and unify metadata, security policies, governance, classification, and data lifecycle management across this distributed data.

How Dataplex works

- Dataplex manages data in a way that doesn't require data movement or duplication.
- As you identify new data sources, Dataplex harvests the metadata for both structured and unstructured data, using built-in data quality checks to enhance integrity.
- Dataplex automatically registers all metadata in a unified metastore.
- You can also access data and metadata through a variety of Google Cloud services, such as BigQuery, Dataproc Metastore, Data Catalog, and open source tools, such as Apache Spark and Presto.