What is Data Catalog?

Data Catalog is a fully managed, scalable metadata management service within Dataplex (/dataplex/docs/introduction).

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/docs) (DLP) scan to identify sensitive data directly within Data Catalog in the form of tag templates.

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 (/data-catalog/docs/how-to/custom-entries).

After your data is cataloged, you can add your own metadata to these assets using tags.

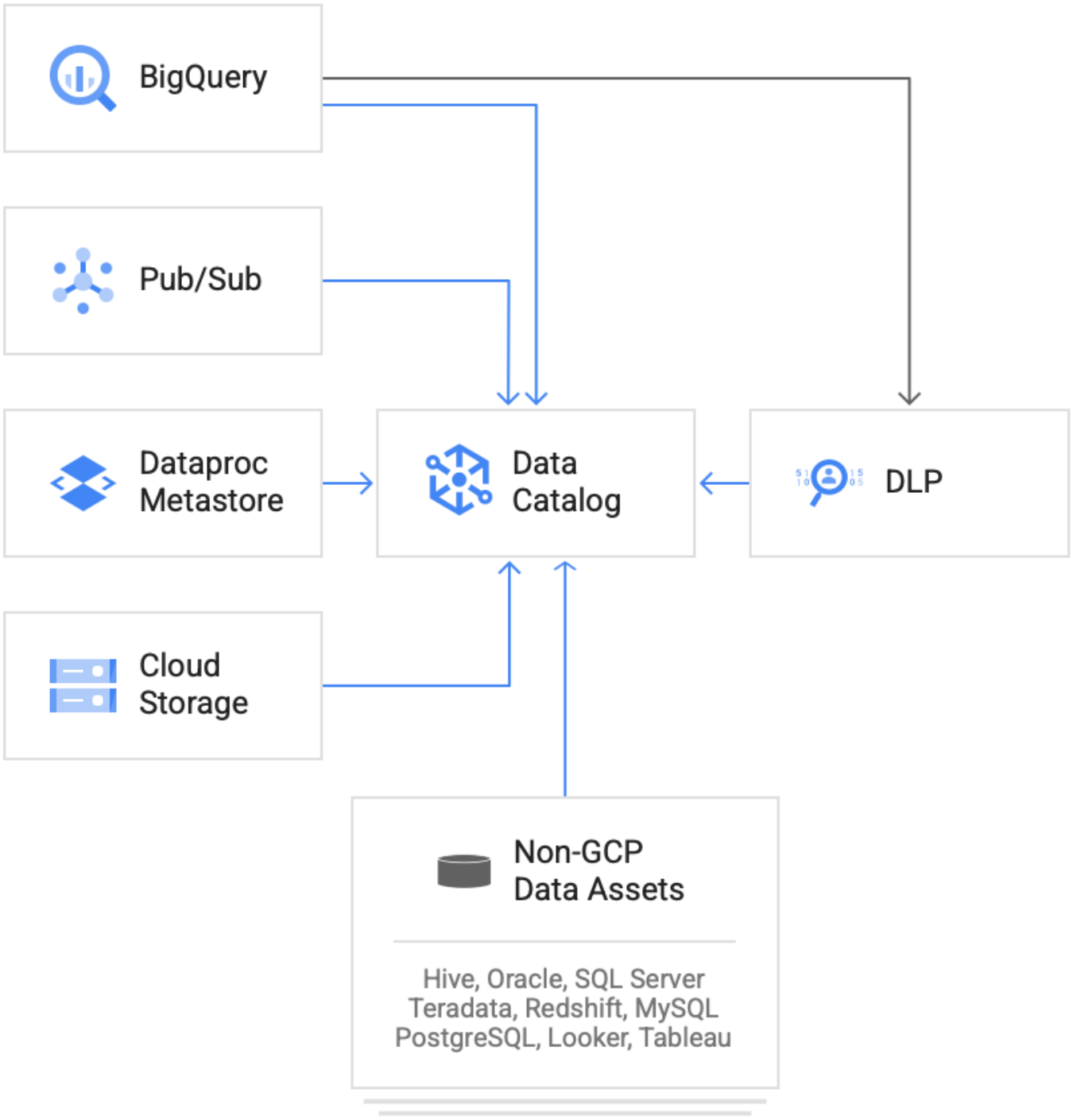

Figure 1. Architecture of Data Catalog

Data Catalog metadata

Data Catalog handles two types of metadata: **technical metadata** and **business metadata**. To know more about metadata, see Data Catalog metadata (/data-catalog/docs/concepts/metadata).

Search and discovery

Data Catalog offers a simple yet powerful predicate-based search experience for technical and business metadata associated with a data entry. You must have the permissions to read the metadata for a data entry so that you can apply search and discovery on the metadata. Data Catalog does not index the data within a data entry. Data Catalog only indexes the metadata that describes an asset.

Data Catalog controls some metadata such as user-generated tags. For all metadata sourced from the underlying storage system, Data Catalog is a read-only service that reflects the metadata and permissions provided by the underlying storage system. You can make edits in the underlying storage system to add, update, or delete the metadata of a data entry.

To know more about Data Catalog search, see Search for data assets with Data Catalog (/data-catalog/docs/how-to/search).

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.

In addition to cataloging assets within the project IDs for which you have metadata access, Data Catalog can catalog data stored in the BigQuery projects that contain public datasets.

Catalog non-GCP assets

To catalog metadata from non-GCP systems in your organization, you can use the following:

- Community-contributed connectors (/data-catalog/docs/integrate-data-sources#integrate_on-premises_data_sources) to multiple popular on-premises data sources
- Manually leverage the Data Catalog APIs for custom entries (/data-catalog/docs/integrate-data-sources#integrate_unsupported_data_sources)

Access Data Catalog

You can access Data Catalog functionalities using:

- Dataplex UI in the Google Cloud console (https://console.cloud.google.com/dataplex)
- `gcloud` (/sdk/gcloud/reference/data-catalog) command-line interface (CLI)
- Data Catalog APIs (/data-catalog/docs/reference#data-catalog-api-reference)
- Cloud Client Libraries (/data-catalog/docs/reference/libraries)

What's next

- To get started with Data Catalog tagging, see Create tag templates, tags, overviews, and data stewards (/data-catalog/docs/quickstarts/quickstart-search-tag).
- To get started with Data Catalog search, see Search and view data assets with Data Catalog (/data-catalog/docs/how-to/search).
- To integrate your data sources, follow the steps in Integrate Google Cloud and on-premises data sources (/data-catalog/docs/integrate-data-sources).

Last updated 2022-08-26 UTC.