Data Catalog - Michaël Bettan

Definition: Data Catalog is a fully managed and scalable data discovery and metadata management service that empowers organizations to quickly discover, understand, and manage all their data in Google Cloud.

SLA: 99.9%

Use Cases:

- **Data Discovery:** Quickly find and understand data assets across your organization.
- **Data Governance:** Enforce data policies and ensure data quality.
- **Data Catalog Management:** Centralized management of metadata across different teams and projects.

Billing:

- **Metadata storage**: Business metadata as well as any on-premises metadata ingested. Metadata storage is measured in binary gigabytes (GiB).
- **Data Catalog API calls**: Data Catalog read, write, and search API calls

Why Data Catalog ?

- Searching for insightful data
- Understanding data and Making data useful

Core Functionality

- **Scalability**: Automatically scales to handle large volumes of metadata.
- **Auto-Cataloging:** Automatically discovers and catalogs metadata from various services (BigQuery, Pub/Sub, Bigtable, Cloud Storage) during ingest operations. It also allows for manual and API-driven cataloging.
- **Metadata Types:** Supports both **technical** (schema, location, etc.) and **business metadata** (tags, descriptions, ownership, etc.).
- **Search & Discovery:** Provides a user-friendly search interface for quickly finding relevant data assets.
- **Security & Compliance:** Integrates with IAM for granular access control and with Cloud DLP for data sensitivity detection and classification.

Key Concepts

- Data Catalog handles two types of metadata:
  - **Technical metadata:** project information, asset name, schema name, description for BigQuery, etc.
  - **Business metadata:** tags, data stewards, rich text overview
  - Always linked to a technical metadata entry.
- **Business metadata**
  - PII
  - Delete-by dates
  - Business logic
- **Attributes types**
  - String
  - Boolean
  - Double
  - Enum
- **Two main functions:**
  - Searching for data entries for which you have access
  - Tagging data entries with metadata
- **Tag sensitive data automatically**, through Data Loss Prevention (DLP) integration
- **Structured Tags:** Allows adding custom business metadata as key-value pairs to tables and columns within BigQuery. This is crucial for business context and data discovery. Example: {"department": "Sales", "data_owner": "John Doe", "sensitivity": "Confidential"}
- **Entry Types:** Supports various entry types, including tables, datasets, columns, files, and more.
- **API Access:** Provides REST APIs for programmatic interaction and integration with other tools and workflows. This allows for automation of metadata management tasks.
- **Data Governance:** Facilitates data governance initiatives by providing a centralized view of metadata and enabling the enforcement of data policies.
- **Metadata Lineage:** While not explicitly a core feature of Data Catalog itself, it can be augmented with other GCP services to provide lineage information (e.g., using workflow management tools to track data transformations).

Tags and tag templates

- **Tags**: enable organizations to create, search and manage metadata for all their data assets in a unified service.
  - Tags = Annotations = “business metadata”
- Tags contain one or more fields where information can be stored.
- The fields in a tag are defined by a tag template, and each field can be used to store one or more values.
- Every tag is an instance of a tag template, which can be applied to an entire data asset, or to particular tables or columns.
- A tag on a column could tell you, for example, if that column contains PII, whether it's been deprecated, or what formula was used to calculate a certain value.
- Tag templates:
  - Define a new (custom) template
  - Reuse an existing public template

Data Ingestion & Cataloging Methods

- **Automatic Ingestion:** During data loading into supported services (BQ, Pub/Sub, Bigtable, GCS).
- **Manual Ingestion:** Using the Data Catalog UI or the API.
- **Third-Party Tool Integration:** Integrate with third-party tools and ETL processes to ingest metadata.

Entries and entry groups

- **Entries** represent data resources:
  - GCP resources, such as a BigQuery dataset or table, Pub/Sub topic, etc.
  - Custom resources with custom data types.
- Entries are contained in an **entry group**.
- **An entry group** is a set of logically related entries together with Identity and Access Management policies that specify the users who can create, edit, and view entries within an entry group.

Miscellaneous

- Search all your datasets with faceted-search
- Entry groups
- Policy tags
- Sync technical metadata automatically and create schematized tags for business metadata
- Auto-ingest tactical
- Highly sensitive data to the limited access