Dataplex - Michaël Bettan

Definition: Intelligent data fabric that unifies your distributed data to help automate data management and power analytics at scale.

Use Cases:

- *Rapidly curate, secure, integrate, and analyze any type of data, at any scale*
- *Organize data without data movement, automatic data discovery, metadata harvesting, lifecycle management, and data quality with built-in AI-driven data intelligence.*
- *Central policy management, monitoring and auditing for data authorization, retention, and classification.*

Billing: based on pay-as-you-go usage:

- Dataplex processing (standard and premium)
- Dataplex shuffle storage
- Metadata storage
- Data Catalog API calls

Data Mesh

Data Mesh promotes **distributed data ownership** while **centralizing governance** and **data discoverability**. It shifts the responsibility for managing, producing, and consuming data to **domain-specific teams** that have the most context for that data.

- **Dataplex** provides centralized data governance while enabling distributed ownership through "virtual lakes."

Dataplex

- A **unified data management platform** that unifies data lakes and warehouses to **simplify data governance** and **accelerate analytics**.
- **Organizes disparate data sources** across multiple clouds and formats into logical lakes and zones for unified management.
- Offers **built-in security, governance**, and **integrations** with other GCP services for seamless data analysis and insights.

Key Capabilities

Centralized Data Management:

- Unifies data lakes and warehouses across multiple storage systems (Cloud Storage, BigQuery, etc.) for centralized visibility and control.
- Organizes data into lakes and zones based on business logic for easier management.
- Integrates with data governance and security tools for policy enforcement.

Enhanced Data Governance:

- Enforces data access and usage policies across all unified sources for compliance and privacy.
- Provides data lineage tracking to understand data origins and transformations.
- Enables data quality management to ensure accuracy and reliability.

Simplified Analytics:

- Provides a unified interface for data discovery and exploration.
- Simplifies data access for analysts and data scientists using preferred tools (BigQuery, Spark, etc.).
- Accelerates analytics and machine learning workflows with streamlined data access.

Security and Compliance:

- Enforces granular access controls and data masking for sensitive data protection.
- Integrates with Google Cloud's security infrastructure for threat detection and prevention.
- Supports compliance with industry regulations (GDPR, CCPA, etc.).

Data Profile

Data profiles give you insights into your data:

- **Data types:** What kind of information is in each column (e.g., numbers, dates, text)?
- **Completeness**: How much data is missing?
- **Uniqueness**: Are there duplicate values?
- **Value distributions**: What are the common values in a column?

Automated Tagging: You can automatically tag tables in Dataplex based on the insights from your data profiles. For example, you could tag a table as "PII" if a profile detects sensitive information like names or addresses.
Improved Data Discovery: Tags make it easier to find and understand data within Dataplex.
Data Governance: Tagging helps you manage and control access to sensitive data.

Security Access Controls

Project Level:

- **Dataplex Admin:** Full control over all Dataplex resources in the project.
- **Dataplex Editor:** Can manage Dataplex resources but cannot grant access to others.
- **Dataplex Viewer:** Read-only access to all Dataplex resources in the project.

Lake Level: grant permissions within a specific lake.

- **Lake Admin**: Full control over a lake and its contents.
- **Lake Contributor**: Can create, update, and delete resources within a lake.
- **Lake Reader:** Read-only access to a lake and its contents.

Zone Level: grant permissions within a specific zone.

- **Zone Admin:** Full control over a zone and its contents.
- **Zone Contributor**: Can create, update, and delete resources within a zone.
- **Zone Reader:** Read-only access to a zone and its contents.

Data Roles control access to the data within Dataplex assets.

- **Data Reader:** Read-only access to data.
- **Data Writer:** Permission to write data.
- **Data Owner:** Full control over data, including granting access to others.

Lakes Zones

Purpose: Logical groupings within a Dataplex lake that allow you to organize and manage your data assets based on criteria like data sensitivity, department, or business domain. Benefits:

- **Simplified data management**: Break down large data lakes into manageable units.
- **Improved data governance**: Apply different policies and controls to different zones.
- **Enhanced security**: Isolate sensitive data in dedicated zones.

Virtual Lakes

- A **lake** is the highest-level abstraction. It represents a **logical container** for organizing and managing your data across multiple data storage systems (e.g., Cloud Storage, BigQuery).
- **Purpose**: It allows you to apply governance, metadata management, and monitoring across your entire data ecosystem.
- **Use case**: A lake can represent an overarching domain in your data mesh architecture, such as a specific business function (e.g., marketing, finance) or a data product.

Zones

- A **zone** is a sub-component of a lake that organizes data into logical groupings, typically based on the lifecycle or stage of the data (e.g., raw, curated, or analytics data).
- **Purpose**: Zones allow finer segmentation of data within a lake. This can reflect different stages in the data processing pipeline (like landing raw data, transforming it, and storing curated results).
- **Types of zones**:
  - **Raw zone**: Stores unprocessed, incoming data (e.g., from logs, transactions).
  - **Curated zone**: Stores transformed and cleansed data, ready for consumption or analysis.
  - **Landing zone**: A temporary holding place for files before processing.
- **Use case**: You might create separate zones for raw, transformed, and curated data within a single lake, ensuring logical separation and governance at different stages of the data lifecycle.

Policy Tags

Purpose: Enable you to define and apply business-relevant metadata to your data assets. These tags can represent classifications (e.g., "Confidential", "PII"), compliance requirements (e.g., "GDPR", "HIPAA"), or data sensitivity levels. Following benefits:

- **Improved data discovery:** Make it easier to find data based on its classification.
- **Enhanced data governance**: Enforce policies and controls based on tags.
- **Automated data management**: Trigger actions (e.g., data masking, access control) based on tags.