**Dataplex**

Google Cloud Dataplex is a service designed to help organisations manage, govern, and analyse large-scale data across different storage systems and data lakes.

Contents

[**1. What is Dataplex?** 1](#_Toc179390478)

[**1.1 Key Concepts of Dataplex** 1](#_Toc179390479)

[**2. How to Use Dataplex?** 1](#_Toc179390480)

[**3. Use Cases of Dataplex** 2](#_Toc179390481)

# **1. What is Dataplex?**

At its core, **Dataplex** provides a unified platform for managing distributed data (like data stored in BigQuery, Cloud Storage, or other systems) by creating **data lakes** and **data zones**. It helps simplify data governance while providing the ability to curate and analyse data seamlessly.

## **1.1 Key Concepts of Dataplex**

1. **Data Lakes**: A data lake is a centralised repository that allows you to store all your structured and unstructured data at any scale. With Dataplex, you can organise your data into **zones** for better management and security.
2. **Data Zones**: Zones are logical partitions within a data lake. Typically, data lakes are broken into different zones for raw data, curated data, and sometimes a sandbox for experimental data. Each zone may have different access rules or processing requirements.
3. **Data Catalog**: Dataplex integrates with Google Cloud's **Data Catalog**, which helps in indexing and discovering datasets across your data lake. You can tag and annotate datasets for better metadata management.
4. **Governance**: Dataplex comes with built-in **data governance** capabilities. It allows you to define and enforce policies across your data, ensuring security, privacy, and regulatory compliance. You can manage who has access to which zones or datasets and audit the usage.
5. **Data Processing and Analytics**: Dataplex makes using data for analytics and machine learning easy. You can analyse data across different zones using tools like BigQuery and Dataproc and even integrate with Vertex AI for machine learning purposes.
6. **Serverless Data Management**: One of the main benefits is that Dataplex operates in a **serverless** manner, meaning you don't need to manage infrastructure. Google Cloud handles the scaling, availability, and performance behind the scenes.

# **2. How to Use Dataplex?**

Here’s a simplified step-by-step process of how you might start using Dataplex:

1. **Create a Lake**: First, define a data lake. This is essentially the top-level structure that will hold your data zones.
2. **Add Zones to the Lake**: You can then define zones, such as raw, curated, and sandbox zones, each with its own access control and processing rules.
3. **Ingest Data**: You can import data into your zones from various sources, such as Google Cloud Storage (GCS) or BigQuery.
4. **Organize and Curate**: Once the data is ingested, Dataplex provides tools to clean, curate, and classify the data for easier management and governance.
5. **Enforce Policies**: Define governance policies, such as access controls, auditing, and privacy policies, to ensure compliance with organisational and regulatory standards.
6. **Analyze Data**: Use integrated tools like BigQuery or Dataproc for running queries, analysing, and generating insights from the data.
7. **Monitor and Optimize**: Dataplex comes with monitoring capabilities to help you track data usage, performance, and cost. It helps you optimise how you use your data lakes.

# **3. Use Cases of Dataplex**

* **Data Governance**: Enforcing consistent governance rules across distributed datasets.
* **Analytics**: Querying data across storage systems (like Cloud Storage and BigQuery) without moving data.
* **Data Lakes and ML**: Building data lakes for ML pipelines that can be used with tools like Vertex AI.

**Entry Group**

An **Entry Group** in Dataplex is a powerful tool, serving as a logical container for organising metadata entries for data assets such as tables and files. It empowers you to efficiently manage, search, and govern metadata across your data lake or other data systems. It is closely integrated with **Google Cloud Data Catalog** for centralised metadata management.

* **Lakes:** Sales Data Lake
* **Zones:**
  + **Raw Zone**: Contains raw CSV files from different sales systems.
  + **Processed Zone**: Contains processed sales data in BigQuery.
  + **Curated Zone**: Contains analytics-ready data products (BigQuery views or derived tables).

You could create three **Entry Groups**:

* **Entry Group 1:** raw\_sales\_data (contains entries for GCS files in the Raw Zone)
* **Entry Group 2:** processed\_sales\_data (contains entries for BigQuery tables in the Processed Zone)
* **Entry Group 3:** curated\_sales\_data (contains entries for data products in the Curated Zone)

Data Mesh: [What Is a Data Mesh? | IBM](https://www.ibm.com/topics/data-mesh#:~:text=A%20data%20mesh%20is%20a%20decentralized%20data%20architecture,ownership%20to%20the%20producers%20of%20a%20given%20dataset.)