**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?** 2](#_Toc179458568)

[**1.1 Key Concepts of Dataplex** 2](#_Toc179458569)

[**2. How to Use Dataplex?** 2](#_Toc179458570)

[**3. Use Cases of Dataplex** 3](#_Toc179458571)

[**4. Get Started with Dataplex** 3](#_Toc179458572)

[**4.1 Prerequisites** 3](#_Toc179458573)

[**Step 1: Create a Dataplex Lake** 3](#_Toc179458574)

[**Step 2: Create Zones within the Lake** 4](#_Toc179458575)

[**Step 3: Ingest Data into the Zone** 5](#_Toc179458576)

[**Step 4: Catalog the Data** 5](#_Toc179458577)

[**Step 5: Analyze Data with BigQuery** 5](#_Toc179458578)

[**Step 6: Manage and Monitor** 5](#_Toc179458579)

# **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.

# **4. Get Started with Dataplex**

## **4.1 Prerequisites**

**1. Enable the Dataplex API:**

* In the Google Cloud Console, go to the **API & Services** section.
* Search for **Dataplex API** and enable it for your project.
* Ensure that you have the necessary permissions to create and manage Dataplex resources.

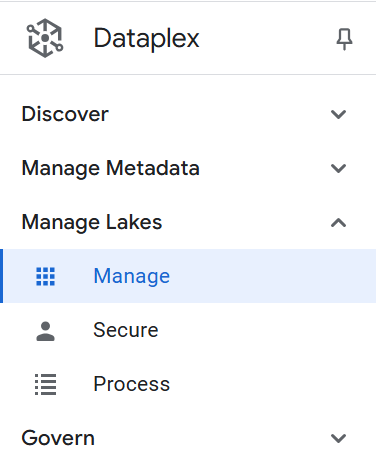
**2. Set up GCS and BigQuery:**

* You'll need some data stored in Google Cloud Storage (GCS) or a dataset in BigQuery.
* If you don’t have any data yet, you can create a sample CSV file and upload it to a GCS bucket.

## **Step 1: Create a Dataplex Lake**

A lake is a top-level resource in Dataplex. Follow these steps to create it.

1. In the **Google Cloud Console**, go to the **Dataplex** section.
   * Navigate to **Navigation Menu > Dataplex > Manage Lakes > Manage**.



1. Click on **Create Lake**.
2. Fill in the necessary details:
   * **Lake ID**: Choose a name for your lake, like “my-lake”.
   * **Region**: Select the appropriate region where you want your data lake to be created (e.g., asia-south1).
   * **Project**: Ensure the correct project is selected.
   * **Note**: Once created, the ID and region can’t be modified.
3. Click **Create**.

## **Step 2: Create Zones within the Lake**

After creating a lake, you can add zones to organize your data.

1. Once the lake is created, go to the **Zones** tab of your lake.
2. Click **Create Zone**.
3. Fill in the details for the zone:
   * **Zone ID**: Name your zone (e.g., raw-data-zone).
   * **Type**: Choose the zone type (e.g., Raw for raw, unprocessed data).
   * **Description**: Add a description for the zone.
4. For **Asset Type**, choose **Cloud Storage** if you want to link data from GCS, or **BigQuery** if you’re working with BigQuery datasets.
5. Choose the appropriate **storage path** (e.g., a GCS bucket or BigQuery dataset).
6. Define **Permissions** to control who has access to this zone.
7. Click **Create** to finalize the zone.

## **Step 3: Ingest Data into the Zone**

If you're using **Google Cloud Storage** as the data source, ensure your data is already uploaded to a GCS bucket.

1. Upload a sample file to a **GCS bucket**.
   * Example: If you have a CSV file data.csv, upload it using the GCS browser or gsutil cp command.
2. Once the data is in GCS, it will automatically be available in the zone you created in Dataplex if the GCS path matches the zone’s storage path.

## **Step 4: Catalog the Data**

After data ingestion, you can catalog it for easy discovery and access.

1. Go to the **Google Cloud Console** > **Dataplex** > **Zones**.
2. Select the zone containing your data.
3. You can use **Data Catalog** to create tags and metadata for your data, making it easy to search and discover datasets.
   * Click on a dataset or file and add relevant metadata tags for discovery.

## **Step 5: Analyze Data with BigQuery**

You can now analyze the ingested data using BigQuery directly from within Dataplex.

1. In Dataplex, go to **Assets** and locate the data you want to analyze.
2. Use **BigQuery** to query the data directly from GCS or BigQuery.
   * If your data is stored in GCS as CSV, you can create an external table in BigQuery to query it:

CREATE EXTERNAL TABLE my\_project.my\_dataset.my\_table

OPTIONS (

format = 'CSV',

uris = ['gs://my-bucket/path/to/data.csv']

);

* + You can then run queries on this external table without moving the data.

## **Step 6: Manage and Monitor**

You can now monitor your data lake and zone through the **Dataplex Console**:

* Check data usage, performance metrics, and governance policies.
* Ensure compliance with policies by auditing data access and usage patterns.

**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.)