**🔹 What is a Dataset in GCP (BigQuery)?**

A **Dataset** in **BigQuery** is a top-level container that is used to organize and manage **tables**, **views**, **materialized views**, and **routines** (like stored procedures and functions).

**🔸 Key Characteristics of a Dataset:**

| **Feature** | **Description** |
| --- | --- |
| **Namespace** | Datasets act like namespaces to group related tables together. |
| **Access Control** | You can apply IAM roles to datasets to control who can access what. |
| **Location** | Each dataset has a geographic location (e.g., us, europe-west1) which determines where data is stored. |
| **Billing** | Usage and storage within a dataset contribute to your BigQuery billing. |
| **Encryption** | Datasets can use default or customer-managed encryption keys (CMEK). |

**🔹 Dataset Hierarchy in GCP**

GCP Project

└── BigQuery

└── Dataset (e.g., sales\_data)

├── Table (e.g., transactions)

├── View (e.g., recent\_sales\_view)

└── Routine (e.g., calculate\_discount())

**🔸 Creating a Dataset**

You can create a dataset via:

**✅ GCP Console (UI)**

1. Go to BigQuery.
2. Click on your project name.
3. Click “Create Dataset”.
4. Set Dataset ID, Data Location, and other options.

**✅ bq CLI**

bq mk --dataset --location=US project\_id:dataset\_name

**✅ Terraform**

resource "google\_bigquery\_dataset" "example" {

dataset\_id = "sales\_data"

location = "US"

friendly\_name = "Sales Data"

description = "Dataset for storing sales data"

default\_table\_expiration\_ms = 3600000 # 1 hour

}

**🔹 Common Use Cases**

* Data warehousing
* Storing machine learning training data
* Organizing analytical reporting data
* Centralizing logs, audit, or event streams

**🔸 Best Practices**

* Use meaningful dataset names (e.g., user\_metrics, finance\_2024)
* Apply fine-grained IAM roles at the dataset level
* Set dataset expiration policies to manage cost
* Always select the correct location for compliance & performance