Dharma Devi 2nd May Hw

Role: Data Ingestion Engineer:

In Data ingestion role, I will be responsible for building and maintaining pipeline that brings data from multiple sources

Data Ingestion pipelines from APIs,

1. Identify Data Sources, **Receive/Detect New Data**

Step 1:

Detect a new file in backend-team-bucket **backend team's Cloud Storage bucket** is **often the first step** in a data ingestion pipeline when:

**Validate & Standardize**

* + Check file headers
  + Confirm expected row count
  + Validate format (e.g., UTF-8 encoded CSV)

**Ingest to Raw Storage**  
 Move file into the project’s raw zone:

Source Bucket: gs://backend-team-bucket/

Target Bucket: gs://your-project-bucket/raw/

### **2. Upload Data to Raw (Bronze) Cloud Storage Bucket**

* Use gsutil, Python scripts, or scheduled jobs to upload:

bash

CopyEdit

gsutil cp gs://backend-team-bucket/\*.csv gs://devi-project-bucket/raw/

* Ensure naming conventions, folder structure, and versioning are consistent.

### **3. Metadata Capture**

* Log metadata like:
  + File name
  + Ingestion Time
  + Timestamp
  + Source system
  + Checksum/hash for data integrity
* Store in audit logs, BigQuery, or a metadata catalog.

4.**Trigger Downstream Processing**  
 Send a signal to Data Transformation team or trigger Cloud Composer/Dataproc job.