Question 8: GCP ETL Design:

1. Extraction (Data Ingestion and Archival):

Google Cloud Storage (GCS) :

GCS will serve as the **Data Lake** for both the ingestion and the archival requirements. We will create 2 buckets for raw and transformed output each. We can create date partitions in the GCS bucket to keep the files date wise.

Cloud Composer (Apache Airflow) :

Cloud composer will trigger a DAG which will extract the data from legacy database and will be scheduled to run daily. This DAG can in turn trigger a Data fusion pipeline to fetch data from database and create raw csv files and place them in the raw GCS bucket.

2. Transformation:

Cloud Data Fusion or (Pyspark Job running on dataproc cluster):

We can create a CDF pipeline or pyspark job to read the csv files from raw bucket perform the filter, join etc transformations and create 2 outputs , one as CSV to be stored in transformed bucket and another to load BQ.

3. Load :

Write the final aggregated results into BigQuery tables (likely partitioned by ingestion time or a specific business date column).

##

#Question 10 : Pharmacy DB Design:

We will use Star schema to design this in a data warehousing system so that it will be easy for reporting as well.

1. Fact Table : FACT_PHARMA_SALES : This will have keys from all the dimensions and measurable facts:

Column Name	Data Type	Constraints/Description
sales_key	INT	PK, surrogate key for unique line item
date_key	INT	FK to Dim_Date, for transaction date
store_key	INT	FK to Dim_Store, for sale location
customer_key	INT	FK to Dim_Customer, for patient
medication_key	INT	FK to Dim_Medication, for medicine information
employee_key	INT	FK to Dim_Employee, for staff information
supplier_key	INT	FK to Dim_Supplier, for vendor management
quantity_sold	INT	Measure: units sold by store
unit_price	DECIMAL	Measure: per-unit sale price
total_amount	DECIMAL	Measure: line item revenue (quantity * unit_price)
cost_price	DECIMAL	Measure: acquisition cost
discount_amount	DECIMAL	Measure: discount applied
load_date	TIMESTAMP	Metadata: load timestamp for audit
source_system_id	VARCHAR	Metadata: to track source information

2. Dimension Tables :

DIM_DATE (date_key,full_date,year,month**)**

DIM_STORE (store_key,store_name,city,state,address,active_flag**)**

DIM_CUSTOMER (customer_key,customer name, age, gender, address**)**

DIM_SUPPLIER (supplier_key,supplier_name,contact)

DIM_MEDICINE (medication_key,medicine_name,price,category)

DIM_EMPLOYEE (employee_key, emp_name, role, manager_id)