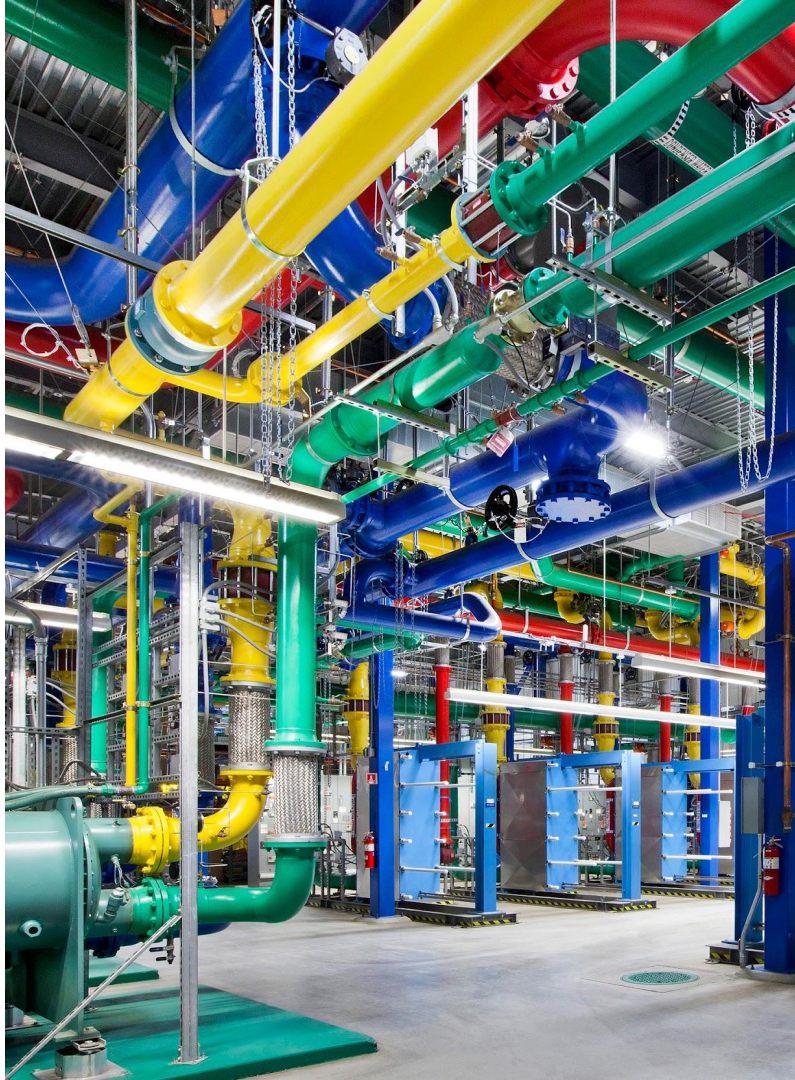
datalabs

Data Engineering

Technical Test

Google Cloud



Requirement Data Engineering Test

This test is designed to evaluate your ability to:

- Build and a ETL pipelines.
- Design Data Warehouse models and architecture.
- Explain end-to-end data pipelines.
- (Optional) Work with streaming data pipelines.
- User the technology that you are comfortable, for example : airflow, Dataflow, bigquery, Redshift, Greenplum, etc)

Please ensure you upload your project to GitHub with a clear README

Google Cloud datalabs

Task 1 (ETL Process)

- User python and airflow to create the pipeline
- Create 5 tables with the minimum record 1000. The following table you should create :
  - customers (customer_id, name, email, city, signup_date)
  - products (product_id, product_name, category, price)
  - transactions (transaction_id, customer_id, transaction_date, total_amount)
  - transaction_items (transaction_item_id, transaction_id, product_id, quantity, price)
  - marketing_campaigns (campaign_id, campaign_name, start_date, end_date, channel)

Google Cloud datalabs

Task 2 (Data Modeling)

- From the staging data, design a Data Warehouse schema.
  - Fact Table: fact_sales (from transactions + transaction_items).
  - Dimension Tables: dim_customer, dim_product, dim_date, dim_campaign.
- Create SQL DDL for all tables.
- Provide schema diagram (ERD).
- Deliverables:
  - SQL scripts for table creation.
  - ERD diagram.
  - Explanation of modeling choices.

Task 3 (Describe the Architecture)

- Describe the architecture of your data pipeline using draw io
- Put the architecture to be readable into a github

Task 4 Streaming Data (Optional Task : Would be additional value)

- Build a simple real-time pipeline:
  - Source: Dummy transaction generator (Python loop or Kafka producer).
  - Process: Aggregate transaction count per minute.
  - Sink: Console output or database insert.
  - Provide streaming architecture diagram.
- Deliverables:
  - Streaming code.
  - Pipeline diagram.
  - Explanation of the pipelines

Deliverables to completed the task

- GitHub repository including:
  - Sample data.
  - ETL script (Python/Airflow).
  - SQL DDL and schema diagram.
  - Data Warehouse architecture diagram (draw.io)
  - (Optional) Streaming pipeline.
  - README.md with instructions, assumptions, and setup.
- Assignment would be shared max (7 Jan 2026) via email

Google Cloud datalabs

Thank you

Google Cloud