Quick notes

Sunday, April 20, 2025 10:36 PM

Create data sources :

1. hospital-a-mysql-source
   - patients (incr)
   - providers (full)
   - encounters (incr)
   - transactions (incr)
   - departments (full)
2. hospital-b-mysql-source
   - patients (incr)
   - providers (full)
   - encounters (incr)
   - transactions (incr)
   - departments (full)
3. claims
   - monthly basis for each hospital
   - in one month
     - 1 claim.csv for hospital-a
     - 1 claim.csv for hospital-b
4. cpt_codes
   - constant or standard codes for medical procedures
   - csv file

Data Ingestion :

- create dataproc cluster
- ingestion-1 => configs => hospital-a-mysql (tables) ==> gcs_bucket/landing/hospital-a/
- ingestion-2 => configs => hospital-b-mysql (tables) ==> gcs_bucket/landing/hospital-b/
- claims on monthly basis for two hospitals => csv files ==> gcs_bucket/landing/claims/*
- cpt_codes are constant => csv files ==> gcs_bucket/landing/cptcodes/*

Bronze:

- gcs_bucket/landing/hospital-a/* ==> external tables ==> bigquery/bronze_dataset/*
- gcs_bucket/landing/hospital-b/* ==> external tables ==> bigquery/bronze_dataset/*
- gcs_bucket/landing/claims/*.csv ==> dataproc ==> bigquery/bronze_dataset/claims
- gcs_bucket/landing/cptcodes/*.csv ==> dataproc ==> bigquery/bronze_dataset/cpt_codes

Silver:
- bronze_dataset/*{full/incr} ==> cdm, scd2, nulls, distinct, merge ==> silver_dataset/*

Gold:
- as per business usecases (silver tables ===> query ===> gold tables) ===> reports & dashboards
- ex : patients_summary (4 tables ==> joins, aggr, operations ===> patients_summary_gold)

Orchestration:
- create composer environment (workflow Orchestration tool)
- pyspark_dag ==> total 4 pyspark jobs
- bq_dag ==> total 3 bq jobs

CICD:
- sett up github
- set up composer env
- cloud build trigger
- test

gcs_bucket
- landing
  - hospital-a

- archive
  - 2025
    - 03
      - 24
        - patients_24032025.json
        - transactions_24032025.json
      - 25
        - patients_25032025.json
        - transactions_25032025.json
- patients
  - patients_26032025.json
- transactions
  - transactions_26032025.json

hospital-a-mysql > patients ----> full data ----> gcs_bucket/landing/hospital-a/patients

audit table -- in bigquery --