Technical Challenge: Data Engineer

1. Context

You are a Data Engineer on the data team at a HealthTech startup that manages a network of clinics. The Operations team needs to better understand doctor productivity and appointment trends.

Your task is to design and build a local ETL pipeline. This pipeline will ingest raw data (simulated in .xlsx files) containing information about doctors and appointments. You will process this data and load it into a local PostgreSQL database in a clean, analytical format.

---

2. Technical Requirements

1. **Environment:** The entire pipeline must run locally.
2. **ETL Script:** The pipeline must be written in **Python**.
3. **Extract:** The script must read the data from the two .xlsx files.
4. **Transform:** The script must perform necessary transformations.
5. **Load:** The script must load the cleaned data into two separate tables (e.g., doctors, appointments) within a dedicated PostgreSQL schema (e.g., healthtech).
    - The pipeline must be **idempotent**, meaning it can be re-run multiple times without failing or duplicating data (e.g., using a DROP/CREATE or TRUNCATE/LOAD strategy).
6. **Logging:** Your Python script **must** implement execution logging to report the pipeline's progress and any errors to both the console and a log file.

---

3. Deliverables

Your submission should be a link to a Git repository (e.g., GitHub, GitLab) containing the following:

1. **Python Pipeline Code:** All .py files required to run the ETL pipeline.
2. **Final dataset:**
   - Final version of the processed dataset before being uploaded to PostgreSQL.
3. **SQL Queries (queries.sql):** A single .sql file containing the queries required to answer the following business questions:
   1. Which doctor has the most confirmed appointments?
   2. How many confirmed appointments does the patient with patient_id '34' have?
   3. How many cancelled appointments are there between October 21, 2025, and October 24, 2025 (inclusive)?
   4. What is the total number of confirmed appointments for each doctor?
4. **Documentation (README.md):** A comprehensive README.md file that includes:
   - **Setup Instructions:** Clear, step-by-step instructions on how to set up the environment and run your pipeline
   - **Pipeline Explanation:** A detailed explanation of each stage of your ETL process (Extract, Transform, Load).
   - **AWS Architecture Proposal:** A section detailing which AWS tool you would use for each part of this process (Extract, Transform, Load) in a production environment, and a justification for *why* you selected each tool.