Analytics Engineer | Take Home Assignment

You've been tasked with building a career analytics data model and report. The business wants to understand *career progression patterns* across companies and industries to provide better recommendations to job seekers.

This challenge consists of three parts, and is intended to take no longer than **three hours**.

- Spending more time is unfair both to other candidates and to your own schedule! We'd love to see where you get with the challenge, even if you don't finish it.

Part 1: Data Engineering

You are provided with \``professionals_nested.json`\`, which contains career data for professionals in a nested document structure (similar to how it's stored in our MongoDB production environment):

Your tasks:

1. Design an ETL process to transform this nested data into a structured data model that supports the requirements in Part 2

2. Implement data validation and quality checks (watch for date inconsistencies, outliers, etc.)

3. Create a dimensional model with appropriate fact and dimension tables

4. Document your data pipeline approach and any assumptions made

Using pandas dataframes, duckdb, or other Python tools are all fair game.

Part 2: Requirements Gathering & Problem Space Analysis

Imagine you're meeting with the Customer Support team. They've told you: "We need to understand what factors lead to successful career growth across different industries."

1. List ~3 specific questions you would ask the team to clarify requirements

2. Based on your assumptions about their needs, identify 2-3 key metrics that would be valuable. Why are they valuable?

3. Explain how your data model supports answering these career development questions

4. Describe any additional data sources you would recommend integrating

Part 3: Airflow Integration & Production Considerations

1. Design an Airflow DAG that would automate the career analytics data pipeline.

Consider incremental data loading, data quality checks, and how you would refresh derived metrics. Please design your DAG according to following specs:

- Airflow version: 2.10.1
- Python version: 3.11
- Components:
  - Web servers: 1 vCPU and 2GB RAM
  - Workers: 1 vCPU and 2GB RAM
  - Schedulers: 1 vCPU and 2GB RAM
  - Database: 2 vCPU and 4GB RAM

2. Please write out any considerations about productionizing your work.

Prose and code example responses are accepted for this section.

Submission Format

- Code (public colab notebook or github link) with your data engineering work
- Document with your requirements analysis