Democratizing Access to Claims Data: A First Step

Context & Problem Statement

Context:

- Synthetic healthcare claims data from a private Mexican insurer with complex data (patients, providers, claims, claim_providers).
- Multiple providers per claim causing data duplication challenges in analysis.

Problem:

- Business users need quick, accurate insights without deep SQL knowledge.
- Complex schema with many tables and relationships.

Solution & Technical Overview

Solution:

- Developed a star schema model with fact (claims) and dimension (patients, providers) tables to avoid duplication issues.
- For this first MVP version, only the primary provider per claim is considered for simplicity.
- Implemented a natural language interface using LangChain + OpenAI for non-technical users to ask questions in plain English.
- Connected to the synthetic claims database enabling real-time SQL generation and query execution.

Results / Benefits:

- Business users could now get quick, accurate answers to complex queries without needing SQL skills.
- This could enable exploration of trends, patient demographics, provider info, and claim details efficiently.
- It could significantly reduces dependency on technical teams and speeds up decision-making

Technical Overview & Next Steps

Technical Overview:

- Data stored in **PostgreSQL** with raw and staging schemas, modeled with **dbt**.
- **LangChain** toolkit connects **OpenAI GPT-4** with the database for natural language to SQL translation.
- Query results presented in **Jupyter notebooks** with user-friendly formatting.

Next Steps:

- Expand query capabilities and integrate user interface for stakeholders.
- Address the issue of ***multiple providers per claim*** by adjusting the data model—possibly adopting a snowflake schema or refining the current model.