# End-to-End Doctor-Patient Transcription Pipeline (Azure)

Domain: Healthcare

Client : Kaiser Permanente

Platform: Azure

Tech Stack: Azure API Management, Azure Data Lake Gen2, Azure Functions, Azure Databricks, Azure Synapse Analytics, Power BI, Delta Lake, Python, Spark

## 🔶 Problem Statement

The client aimed to analyze historical and ongoing doctor-patient conversations for insights such as treatment patterns, diagnosis frequency, and overall consultation quality improvement.

## 🚀 Goal

To build a secure, scalable data pipeline that ingests conversation transcripts (from Whisper/OpenAI), processes them through Azure services, and enables downstream analysis using structured, queryable data.

## 🏗️ Solution & Architecture

We followed the Medallion Architecture with Bronze, Silver, and Gold layers, enabling traceability, data quality, and performance.

### 🔹 1. Raw Data Ingestion (Bronze Layer)

Transcriptions generated from Whisper API were sent to the pipeline via Azure API Management or a Function App. Data was stored in Azure Data Lake Gen2 (Raw Zone) or Blob Storage (Raw).

### 🔸 2. Initial Parsing (Azure Databricks – Bronze Layer)

Ingested text was parsed in Azure Databricks. Initial structuring, tagging with metadata (e.g., doctor ID, patient ID, timestamps), and formatting were done here.

### 🟠 3. Data Cleaning & Transformation (Silver Layer)

Cleaned, de-duplicated, and joined with metadata (doctor profiles, department mappings, etc.). Irregularities like filler words, nulls, and formatting issues were resolved.

### 🟡 4. Aggregation & KPIs (Gold Layer)

Aggregated data to compute:  
- Consultation counts  
- Frequent diagnosis patterns  
- Avg. consultation duration  
- Doctor-specific consultation metrics  
Prepared data marts for reporting.

### 📊 5. Analytics & Reporting

Final structured data was loaded into:  
- Azure Synapse Analytics for querying  
- Power BI for dashboards  
- Azure SQL DW as the reporting warehouse

## 👩‍💻 My Role

- Built the Azure Function & API integration to receive transcript data.  
- Developed the Bronze/Silver/Gold transformation pipeline in Databricks using Spark.  
- Designed data quality logic and ensured PHI/PII was anonymized as per compliance.  
- Created Power BI dashboards for tracking KPIs and treatment trends.

## 📈 Business Impact

- Enabled real-time and historical tracking of patient care trends.  
- Improved decision-making using diagnosis frequency and consultation quality data.  
- Delivered a fully automated, scalable pipeline with easy integration for new clinics/doctors.

## 🧱 Challenges Faced

- Handling unstructured audio transcripts with speaker overlaps.  
- Ensuring HIPAA-compliant storage and masking of PHI.  
- Managing semi-structured and inconsistent metadata across clinics.

## ✅ Outcome

The client now has a reliable, query-optimized, and scalable platform for clinical data analysis using voice-to-text transcripts. The system allows better medical research, improves service quality, and supports compliance.