MedStream - Real-Time Healthcare Analytics

Client Requirements

From: Dr. Sarah Mitchell - Chief Operations Officer, Northeast Health Collaborative
To: ABC Company - Data Engineering Solutions Partner
Subject: Development of a Real-Time Data Platform for Patient Encounters, Conditions & Claims Analytics

1. Business Background

Northeast Health Collaborative (NHC) operates a network of hospitals and outpatient centers across the New England region. The organization struggles with fragmented patient encounter data spread across different systems, making it difficult to analyze encounter trends, department utilization, financial performance, and patient outcomes.

Currently, leadership receives delayed reports and cannot monitor operations in near-real time, limiting decision-making during peak demand periods.

2. Business Objectives

- Monitor patient encounters in near-real time.
- Analyze trends by department, age group, and gender.
- Track financial indicators associated with patient visits including claims and payments.
- Provide leadership with self-service dashboards.
- Ensure the solution scales across all NHC facilities.

3. Functional Requirements

1. Data Sources

- Real-time encounter events are simulated using a Python-based producer publishing events to Azure Event Hubs (Kafka-compatible)
- Batch clinical datasets including patients, conditions, and organizations.
- Financial extracts from claims and claims transactions.

2. Data Processing & Storage

- Use Medallion Architecture (Bronze → Silver → Gold).
- Support schema evolution without pipeline failures.
- Maintain historical snapshots of patient and department data through batch refreshes in curated Silver and Gold layers.
- Model curated data in a Star Schema optimized for analytics.

3. Analytics & Reporting

- Leverage Azure Synapse SQL for analytical querying and validation on curated Gold datasets stored in ADLS Gen2.
- Power BI dashboards consume analytics-ready Parquet datasets directly from the Gold layer in ADLS Gen2.

4. Orchestration & Automation

- Automate ingestion pipelines using Azure Data Factory.
- Enable trigger-based and scheduled refresh processes.
- Refresh Silver and Gold layers automatically.
- Support monitoring and failure visibility through pipeline execution logs.

5. Data Quality

- Handle missing timestamps, duplicate records, and inconsistent values.
- Validate incoming data and handle quality issues during Silver-layer transformations.
- Clean and standardize records in the Silver layer.

6. Security & Governance

- Apply role-based access control across systems.
- Mask or anonymize sensitive fields when appropriate.
- Security controls are applied at the platform and storage level using Azure-managed access policies.
- Follow least-privilege access principles.

4. Deliverables

- Operational Azure data pipeline across ingestion, processing, and reporting.
- Synapse SQL layer enabling analytics queries.
- Analytics-ready Gold datasets exported as Parquet files for Power BI consumption.
- Interactive Power BI dashboards.
- Data quality checks embedded within transformation pipelines.
- Comprehensive project documentation (architecture, models, repository).

5. Success Criteria

- Dashboards reflect near-real-time operational insights.
- Pipelines operate fully automated without manual intervention.
- Schema changes do not cause system downtime.
- Stakeholders actively rely on dashboards for decision-making.