Innominds
Engineering Intelligence. Responsibly

Case Study

Building a Scalable EMR Data Pipeline for Advanced Healthcare Analytics

Modernized legacy EMR infrastructure

Secure data de-identification

Fully automated data pipelines

Scalable, analytics-ready platform

AI/ML-ready validated datasets

Client Overview

A leading global healthcare enterprise specializing in advanced wound care, ostomy care, continence care, critical care, and infusion devices operates numerous clinics worldwide. Its legacy Electronic Medical Record (EMR) system constrained data accessibility and limited the organization's ability to leverage clinical data for analytics, AI-driven insights, and product innovation.

Business Need

The client faced several critical limitations with its legacy EMR repository:

- No support for incremental data loads
- Data extraction limited to only 25 of 700 database tables
- Unpredictable output from legacy system interfaces
- Complex multilingual data support requirements
- A strong need for secure PHI/PII data handling

Technical Requirements

- Scalable cloud platform to ingest data from the legacy EMR
- Reliable incremental data load capabilities
- PHI/PII de-identification with automated data validation
- Secure connectivity and high availability
- Infrastructure provisioning through Infrastructure-as-Code
- Monitoring and event-driven workflow orchestration
- Centralized repository for analytics and BI reporting
- Multilingual data support and readiness for AI/ML use cases

Solution

A cloud-based ETL pipeline and data lakehouse architecture were designed and implemented to modernize data ingestion, processing, and analytics. Key components included:

Data Ingestion & Processing

- Established ETL pipelines with the legacy EMR as the source
- Implemented scalable cloud-based ingestion using AWS services
- Enabled reliable incremental data loading
- Automated data validation processes

Data Security & Compliance

- Applied PHI/PII de-identification using specialized healthcare NLP libraries
- Integrated a custom library from John Snow Labs
- Ensured compliant data handling throughout the pipeline

Analytics Platform

- Loaded curated data into the Snowflake data warehouse
- Deployed Power BI dashboards for business intelligence
- Enabled self-service analytics across stakeholder groups

Infrastructure Automation

- Developed Terraform scripts for AWS resource provisioning
- Implemented deployment pipelines for code management
- Built event-driven workflows with monitoring capabilities

This modernized platform established a unified foundation for enterprise analytics, reporting, and future AI initiatives.

Business Outcome

The transformation delivered measurable business value:

- Direct access to EMR data in Snowflake for enterprise analytics
- Fully automated infrastructure and data pipelines
- Reliable incremental data loading
- Reduced dependency on fragile legacy system interfaces
- Scalable architecture ready for integrating additional data sources
- Strong foundation for AI/ML-driven product strategy

Tools and Technologies

Data & Processing
PySpark | SQL

Cloud Services
Amazon S3 | Amazon Athena | AWS EMR Cluster
AWS Step Functions | Amazon EventBridge
Amazon CloudWatch | Amazon SNS | PrivateLink
VPC Endpoints

Data Warehouse & Analytics
Snowflake | Power BI

Infrastructure & DevOps
Terraform | Automated CI/CD pipelines

Specialized Libraries
Healthcare NLP tools for data de-identification