Data Architecture Reference
Healthcare Multi-Product Platform with Analytics, ML & GenAI

This document defines the enterprise data architecture for a multi-product healthcare platform, supporting clinical, member, provider, claims, and operational domains with advanced analytics, machine learning, and generative AI capabilities. The architecture enforces trust, safety, governance, and scalability while enabling innovation.

Data Architecture Principles

Data is a product; Patient safety first; Federated ownership with centralized governance; Event-first ingestion; Analytics isolated from operations; Explainable AI; Privacy by design; Lineage everywhere

End-to-End Logical Architecture

Sources → Ingestion → Operational Data → Lakehouse → Semantic Layer → Analytics / ML / GenAI

Data Ingestion Layer

Streaming (Kafka/Kinesis), CDC, APIs, Batch, Files. Events preferred for operational data.

Operational Data Layer (ODS)

Per-service databases, event stores, caches, and search indexes supporting real-time workloads.

Lakehouse Data Platform

Raw, Refined, Curated, Feature, and Secure zones with quality, lineage, and governance.

Semantic & Serving Layer

FHIR-aligned models, certified metrics, feature APIs, and analytics-ready views.

Analytics Architecture

Descriptive, diagnostic, predictive, prescriptive, and operational analytics.

Machine Learning Architecture

Feature store, training pipelines, model registry, serving layer, drift and bias monitoring.

Generative AI Architecture

RAG-based architecture with AI gateway, prompt management, validation, and audit.

Governance, Security & Privacy

Schema contracts, data quality rules, lineage, access control, masking, tokenization, consent.

Observability & Reliability

Pipeline health, freshness, anomalies, cost monitoring, model drift, hallucination detection.

Why This Architecture Works

Enables safe multi-product analytics, AI adoption, compliance, and scalable innovation.