Enterprise Data Architecture Book
Healthcare Multi-Product Platform
Analytics • ML • GenAI • Data Mesh

1. Executive Summary

This book defines a complete, enterprise-grade data architecture for healthcare organizations operating multiple product lines. It integrates operational data, analytics, ML, and GenAI with strict governance, privacy, and patient safety requirements. It is designed for CIOs, CDOs, Chief Architects, and regulators.

2. Architecture Principles

1. Data is a product
2. Patient safety first
3. Federated ownership, centralized governance
4. Event-first ingestion
5. Analytics isolated from operations
6. Explainable AI over black-box models
7. Privacy by design
8. Lineage everywhere

3. End-to-End Logical Architecture (Visual)

[Sources] → [Ingestion] → [ODS] → [Lakehouse] → [Semantic Layer] → [Analytics / ML / GenAI] Sources: EHR, Claims, Pharmacy, Labs, Apps, Devices, Partners Ingestion: Events, CDC, APIs, Batch ODS: Service DBs, Event Store, Cache Lakehouse: Raw / Refined / Curated / Feature / Secure Serving: FHIR models, Metrics, Features Consumption: BI, ML, GenAI, APIs

4. Data Mesh Operating Model

Each product line owns its data as a product while the platform team provides shared capabilities. Domain Data Products: - Care Delivery - Member Experience - Provider Operations - Claims & Finance - Digital Channels Platform Capabilities: - Ingestion framework - Lakehouse - Governance - Quality - Security - ML platform - GenAI gateway Governance is enforced through contracts, not meetings.

5. Databricks Reference Implementation

Ingestion: Auto Loader, Delta Live Tables Storage: Delta Lake Processing: Spark, Structured Streaming Governance: Unity Catalog ML: MLflow, Feature Store GenAI: Mosaic AI + RAG BI: Databricks SQL

6. Snowflake Reference Implementation

Ingestion: Snowpipe, Kafka Connector Storage: Snowflake tables Processing: Tasks, Streams Governance: Horizon, Tags, Masking ML: Snowpark ML GenAI: Cortex + External RAG BI: Native + external tools

7. AWS Reference Implementation

Ingestion: Kinesis, MSK, DMS Lake: S3 + Iceberg Processing: Glue, EMR Governance: Lake Formation ML: SageMaker GenAI: Bedrock + RAG BI: QuickSight

8. Azure Reference Implementation

Ingestion: Event Hubs, Data Factory Lake: ADLS + Delta Processing: Synapse, Databricks Governance: Purview ML: Azure ML GenAI: Azure OpenAI + RAG BI: Power BI

9. Analytics Architecture

Certified datasets, semantic models, governed metrics, and operational analytics ensure trust and reuse. Clinical analytics require explainability and auditability.

10. Machine Learning Architecture

Feature stores, training pipelines, registries, serving, drift detection, bias monitoring, and human-in-the-loop validation are mandatory for healthcare.

11. Generative AI Architecture (Safe-by-Design)

RAG-only, governed sources, prompt versioning, audit logs, and human approval for clinical outputs.

12. Governance, Security & Privacy

HIPAA, state laws, consent enforcement, masking, tokenization, lineage, and retention are embedded into the platform by default.

13. Observability & Reliability

Pipeline health, freshness, anomalies, cost, model drift, and hallucination detection are monitored continuously.

14. Common Failure Modes & Mitigations

PHI leakage, schema drift, broken pipelines, untrusted analytics, model bias, vendor lock-in are mitigated through architecture and automation.

15. Implementation Roadmap

Phase 1: Foundation (ingestion, lakehouse, governance) Phase 2: Analytics & ML enablement Phase 3: GenAI enablement Phase 4: Data mesh scale-out

16. Conclusion

This architecture enables safe scaling of data, analytics, ML, and GenAI across healthcare product lines while maintaining trust, compliance, and operational excellence.