Data Lake Value Proposition

Medical Data Processing Company

Christopher O'Hara, PhD, EngD

Agenda

- What is a Data Lake
- Components of a Data Lake
- Data Lake vs Data Warehouse
- Business Value of Data Lake Solution
- Proposed Data Lake Architecture for Medical Data Processing system

What is a Data Lake

Executive summary

A Data Lake is a **centralized repository** that stores all enterprise data at any scale. Data Lakes **eliminates capacity constraints;** no more forced deletions or expensive upgrades. They offer **real-time processing,** which replaces nightly batch windows. Some Data Lakes are **open-source technology,** thus **no vendor lock-in.**

Components of Data Lake

Ingestion Layer

Collects data from 8,000 facilities in real-time. Handles 77,000 files daily with automatic validation and routing.

Storage Layer

Unlimited capacity distributed storage with 3x replication. Eliminates the 8TB SQL Server bottleneck.

Processing Layer

Unified platform for batch, streaming, and SQL analytics. Replaces 70+ custom ETL scripts with metadata-driven automation.

Serving Layer

Optimized access for dashboards, reports, ML, and applications. Sub-second query performance at scale.

Data Lake vs Data Warehouse

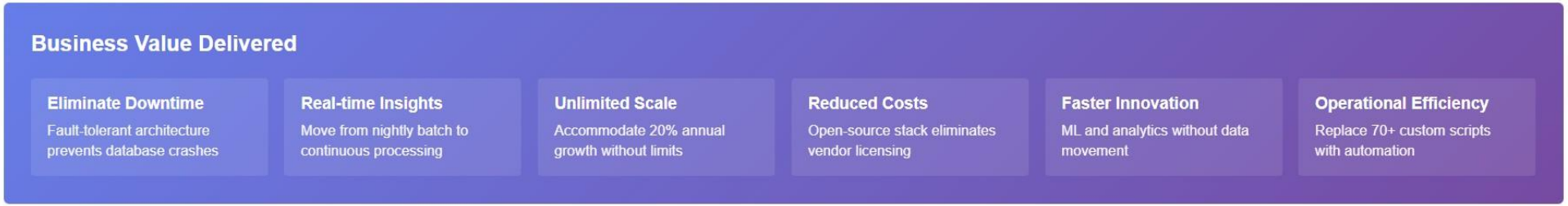
- Data warehouses require extensive planning before storing information
- Data lakes accept information in original form with flexibility for multiple uses

Our current SQL Server operates as a data warehouse - explaining our performance problems.

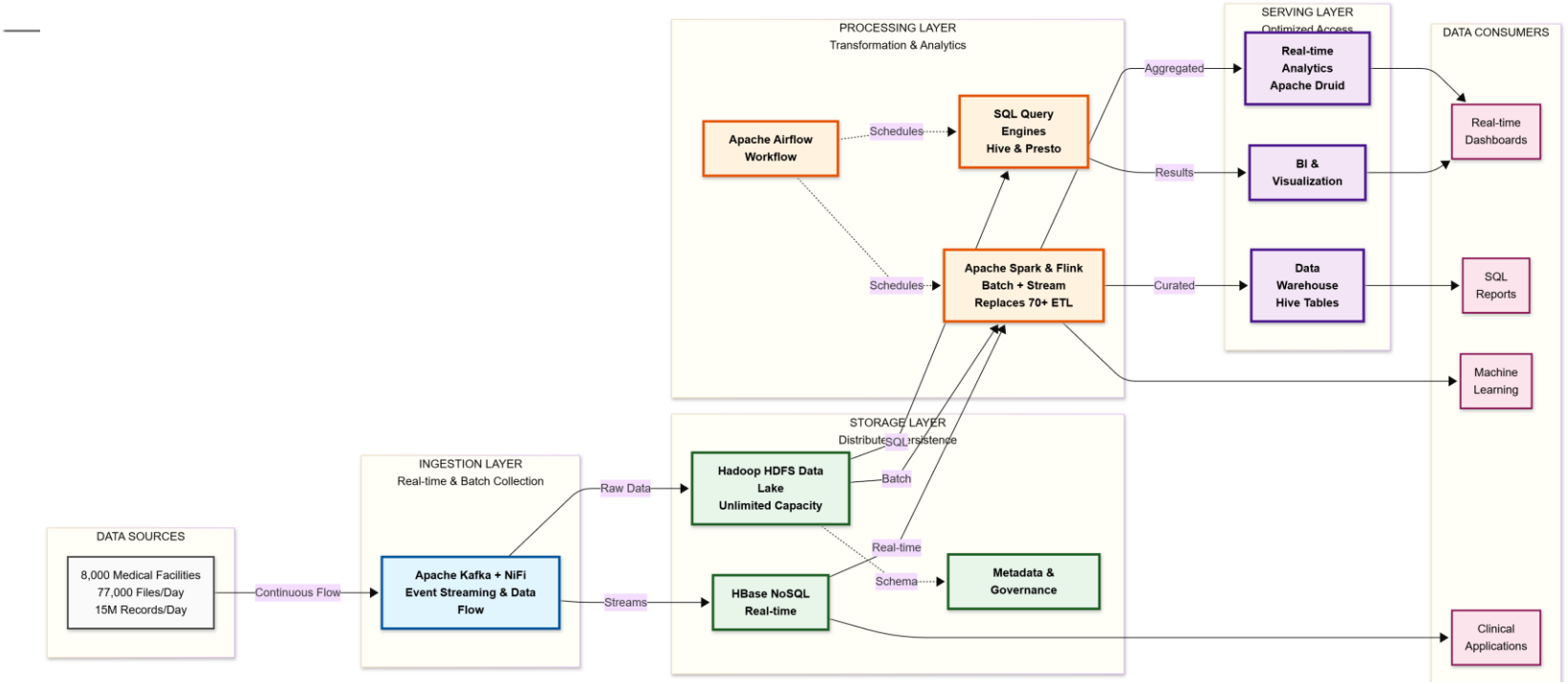
Aspect	Data Warehouse (Current)	Data Lake (Proposed)
Capacity	Fixed storage, requires hardware replacement	Unlimited, grows by adding standard servers
Flexibility	Rigid structure, months to add sources	Accepts any format immediately
Processing	Nightly batch only	Real-time + batch + interactive
Cost	Expensive proprietary licensing	Open-source, no vendor lock-in
Reliability	Single point of failure	Fault-tolerant, no downtime
Use Cases	Reports only	Reports + dashboards + ML + apps

Business Value of Data Lake

The **data lake** solution directly addresses our **three critical business challenges**. First, it **eliminates system downtime** that currently threatens customer relationships - the recent database crash that **took us offline** for hours cannot occur with fault-tolerant distributed architecture. Second, it enables **unlimited growth** to support our 8,000 facilities and 15-20% annual expansion without the capacity constraints forcing us to delete historical data today. Third, it **transforms our competitive position** by enabling **real-time dashboards** and **machine learning capabilities** that our current nightly-batch-only system cannot support, allowing us to deliver the advanced analytics our customers increasingly demand. Additionally, replacing 70+ custom scripts with **automated workflows** and **eliminating expensive proprietary licensing** reduces operational costs by an estimated **30-40%** while **improving reliability** and **accelerating innovation** timelines **from months to weeks**.



Data Lake Architecture



Architecture Layers

- **Ingestion:** Data collection & streaming
- **Storage:** Distributed persistence
- **Processing:** Data transformation
- **Serving:** Optimized access

THANK YOU