TECHNOLOGY BRIEF

KYTHERA
DECIPHERING HEALTHCARE

Engineering Healthcare Intelligence at Scale with the Databricks Lakehouse

How Kythera Labs Uses Lakeflow, Spark, and Unity Catalog to Build Governed Healthcare Data Systems

Healthcare is one of the most data-rich industries in the world and one of the most structurally complex. Claims, EHR, pharmacy, and provider data generate billions of records each year, yet these datasets arrive fragmented across clearinghouses, payers, and operational systems. For organizations trying to understand care patterns, manage value-based contracts, or identify emerging treatment trends, the challenge is not simply accessing data, it is transforming fragmented healthcare transactions into reliable, governed intelligence. Achieving that transformation requires more than analytics. It requires architecture.

Kythera Labs operates at the center of this challenge, managing a massive data pipeline that processes approximately 2 billion medical transactions and 3 billion prescription transactions annually. To turn this high-volume, structurally inconsistent data into usable intelligence, Kythera built its healthcare decision platform, Wayfinder, on the Databricks Lakehouse. The platform is designed to transform fragmented healthcare records into identity-resolved, coverage-aware, enterprise-grade data products that can support large-scale analytics, AI workflows, and operational decision making.

As many organizations standardized on Databricks for large-scale transformation, analytics, and AI, an architectural pattern became clear: when data execution lives inside Databricks, managing orchestration outside the platform introduces unnecessary complexity. For regulated, high-complexity environments like healthcare and life sciences, that friction compounds quickly. As pipelines scale and analytical logic evolves, organizations increasingly consolidate orchestration directly into Databricks Lakeflow, standardizing around a Lakehouse-native architecture where ingestion, transformation, governance, and analytics operate within a single, unified system.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

The Data Reality: High Volume, Structural Inconsistency, Continuous Change

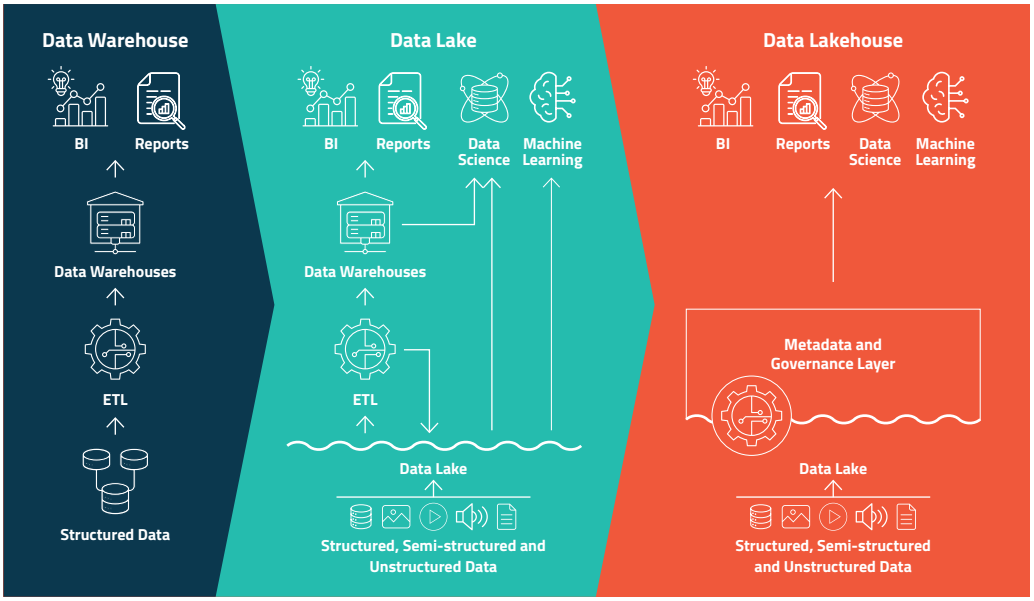
Healthcare claims data is not static. It is fragmented across multiple clearinghouses and payers, refreshed weekly or daily in open claims feeds, subject to 10-20% annual plan switching in some markets, and incomplete or inaccurate-to name just a few common factors impacting fidelity.

Health systems rely on this data to manage referral network stability, site-of-care migration, value-based care delivery, and M&A strategy. But ingesting claims data is only the first step. The real work happens in transformation. Identity resolution changes over time, coverage modeling assumptions evolve, and attribution definitions must be versioned. Every layer must be reproducible and that is where the Databricks Lakehouse is foundational.

Why the Lakehouse Architecture Matters in Healthcare

Kythera, with extensive experience in healthcare and life sciences, has built an enterprise data and decision intelligence platform on Databricks that combines Lakehouse-native data engineering with agentic AI. To support large-scale healthcare analytics, Kythera selected the Databricks Lakehouse to unify data ingestion, transformation, governance, and analytics within a single scalable architecture.

Spark provides massively parallel processing capable of efficiently handling billions of transaction-level claims records, while Delta Lake delivers ACID-compliant storage, versioning, and time travel—capabilities that are essential when recomputing healthcare metrics as assumptions and models evolve.



The Lakehouse architecture unifies data engineering, analytics, and AI within a single governed platform, eliminating fragmentation between systems. This integration enables scalable, end-to-end healthcare data workflows with consistent governance and performance.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

As organizations increasingly standardize on Databricks for large-scale data execution, orchestration outside the platform can introduce unnecessary complexity. By moving orchestration into Databricks Lakeflow, Kythera consolidates pipeline execution where transformation, compute, and governance already live.

Most importantly, Spark Declarative Pipelines allow transformation logic to be defined in a modular, version-controlled way. In healthcare, where analytical definitions and identity resolution logic continually evolve, pipelines must be recomputed deterministically without destabilizing downstream analytics. The Lakehouse architecture makes that possible.

The Lakehouse architecture provides the technical foundation, but the real challenge in healthcare data engineering lies in transforming fragmented healthcare transactions into coherent analytical structures. Achieving that requires a series of governed transformation layers that reconstruct patient journeys, stabilize coverage assumptions, and normalize provider relationships across massive datasets.

Kythera implements these transformations as modular pipelines within the Databricks Lakehouse, turning fragmented healthcare feeds into stable, versioned healthcare data products.

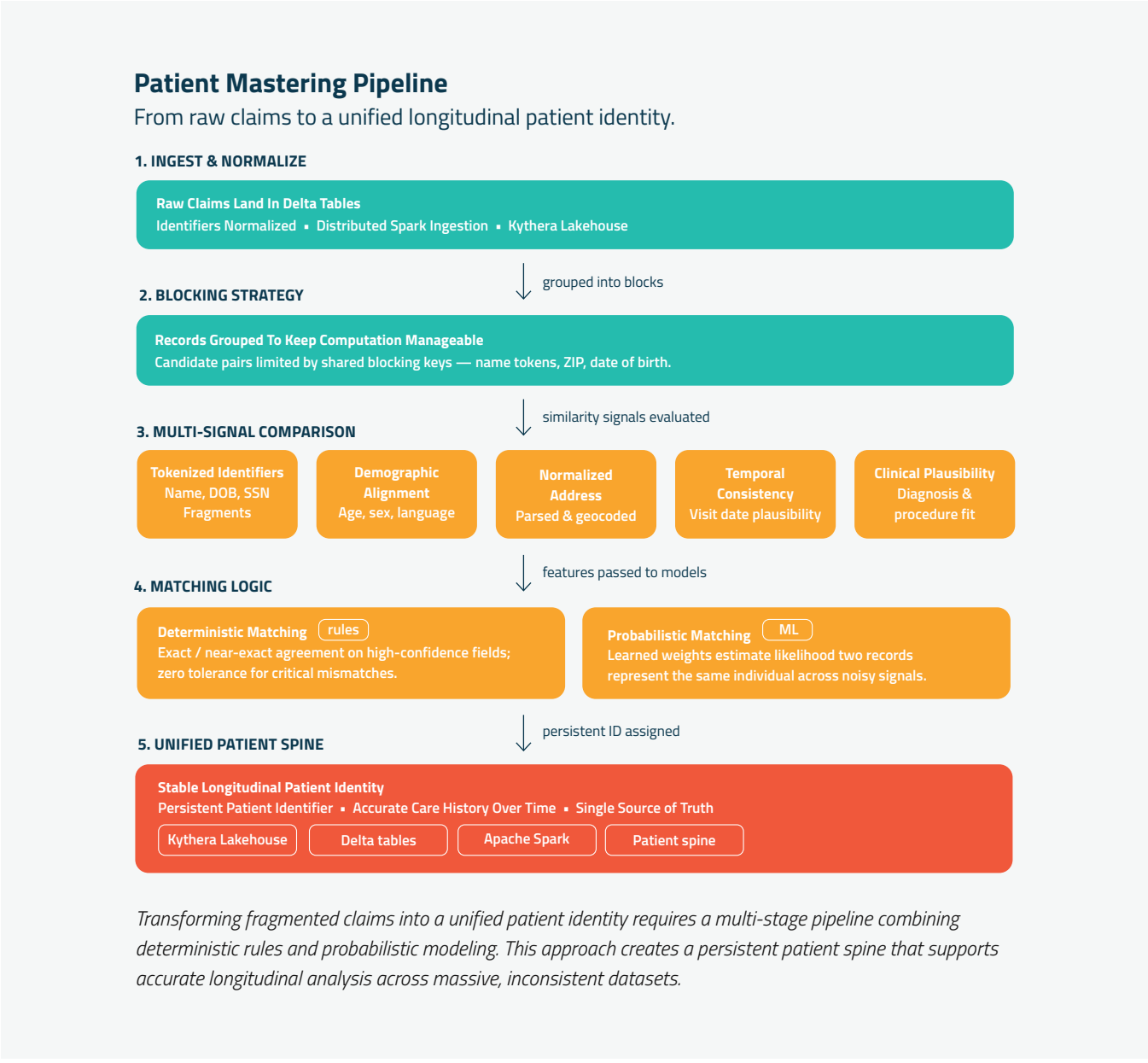
Identity Resolution at Billion-Row Scale

Effective patient mastering is not the result of a single matching step. It is the outcome of a carefully engineered pipeline. The process begins by organizing records into groups using strategies that keep computation manageable. Within these groups, multiple similarity signals are evaluated, including demographic alignment, tokenized identifiers, normalized address data, temporal consistency, and clinical plausibility. Machine learning models then evaluate these features to estimate the likelihood that records represent the same individual, after which persistent patient identifiers are assigned.

Even after patient identities are reconciled, duplication can still occur at the encounter level. Claims feeds often contain overlapping procedural records originating from multiple billing sources. Without additional normalization and consolidation logic, referral counts inflate, patient journeys fragment, and utilization metrics become unreliable. The objective is therefore not simply deduplication, but the creation of stable longitudinal patient histories that accurately reflect real care activity over time.

Kythera addresses this challenge through a patient mastering pipeline implemented inside the Lakehouse. Claims land in Delta tables where identifiers are normalized and distributed Spark processing applies deterministic and probabilistic matching logic to construct a unified patient spine.

Copyright© 2026 Kythera Labs Inc. All rights reserved.



Transforming fragmented claims into a unified patient identity requires a multi-stage pipeline combining deterministic rules and probabilistic modeling. This approach creates a persistent patient spine that supports accurate longitudinal analysis across massive, inconsistent datasets.

The result is a persistent enterprise patient identifier that enables reliable longitudinal analysis across fragmented healthcare records.

In healthcare analytics, identity resolution is not simply a technical challenge, it is a prerequisite for any reliable understanding of patient journeys, referral behavior, and treatment patterns.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

Solving the Denominator Problem with Coverage Modeling

Identity resolution reconstructs the patient spine, but understanding healthcare utilization requires a second critical layer: coverage modeling.

Closed claims datasets typically include enrollment files that define when a patient was covered under a specific health plan. Open claims feeds, however, often lack these enrollment records. This creates a fundamental analytical challenge known as the denominator problem.

Without reconstructed coverage windows, it becomes impossible to determine whether gaps in claims activity reflect true care fragmentation or simply periods when a patient was no longer observable within a payer network. As a result, utilization rates can be misinterpreted, referral leakage may be overstated, and risk models can produce unstable outputs.

Kythera addresses this challenge by implementing coverage modeling pipelines within the Lakehouse. These pipelines reconstruct longitudinal coverage timelines using a combination of:

- payer normalization and classification
- temporal continuity inference across fragmented claims feeds
- observability modeling to estimate active coverage periods
- validation against adjudicated claims and enrollment data where available

Because these transformations run inside the Data Lakehouse, coverage assumptions are versioned and fully reproducible. As payer classification rules evolve or new data sources become available, coverage timelines can be recomputed deterministically without destabilizing downstream analytics.

By stabilizing the denominator, coverage modeling provides the foundation for reliable healthcare analytics including referral network analysis, utilization modeling, and value-based care performance measurement.

Supporting these transformation layers at scale requires pipelines that remain reliable even as data sources evolve, analytical definitions change, and new claims arrive continuously.

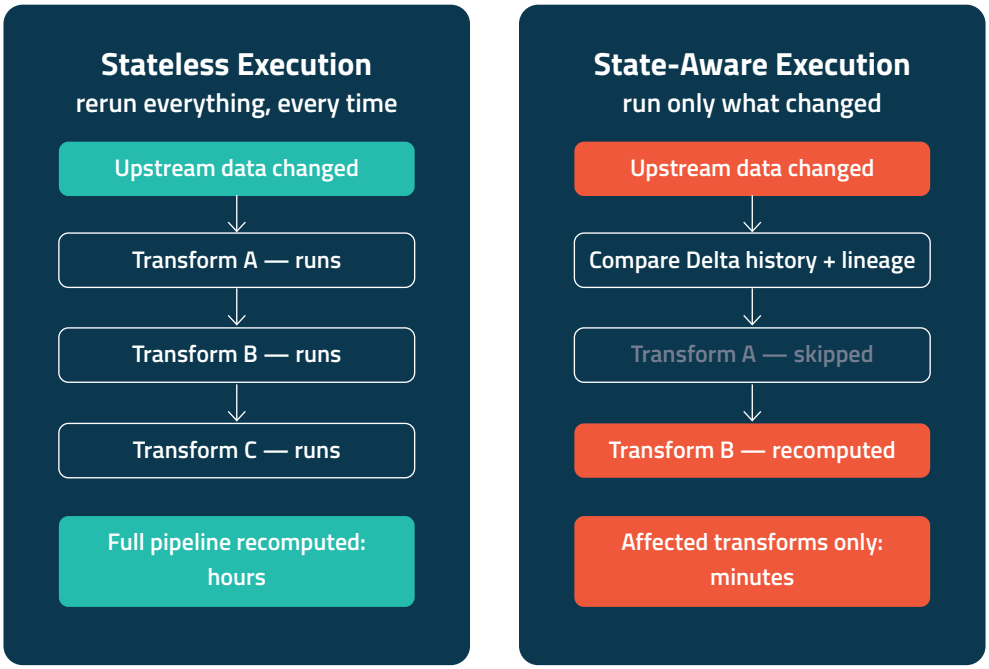
Copyright© 2026 Kythera Labs Inc. All rights reserved.

Architectural Foundations for Governed Healthcare Data

Healthcare analytics requires more than scalable compute. To support reliable reasoning across claims, EHR data, and provider networks, the underlying architecture must ensure that pipelines are reproducible, governed, and resilient to constantly changing data. Kythera's platform, built on the Databricks Lakehouse, achieves this through three core architectural capabilities: state-aware execution, execution-native orchestration, and unified governance through Unity Catalog.

State-Aware Execution

Healthcare data pipelines must operate in an environment where the underlying data is continuously evolving. Claims arrive late, provider affiliations change, payer classifications shift, and identity resolution logic improves over time. Pipelines must therefore understand not only what transformations exist, but also the state of the data those transformations operate on.

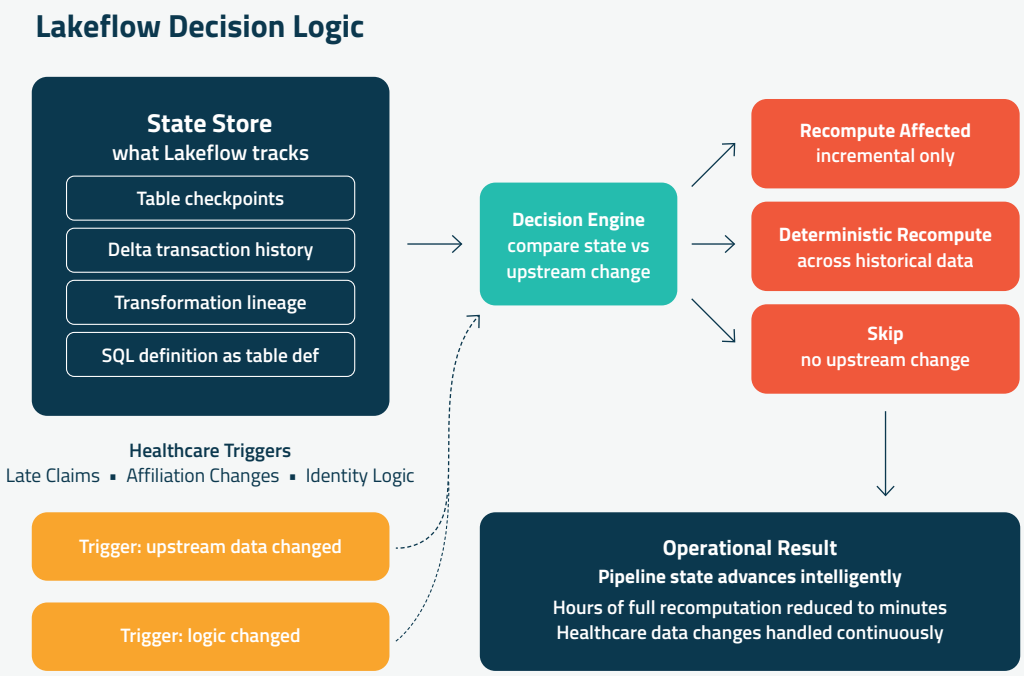


State-aware execution dramatically reduces compute overhead by recomputing only what has changed rather than rerunning entire pipelines. This enables faster, more efficient processing in environments where healthcare data is continuously evolving.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

Lakeflow pipelines provide state-aware execution, particularly when implemented with Delta Live Tables. Rather than treating transformations as stateless queries, Lakeflow tracks table state, checkpoints, lineage, and Delta transaction history. When a SQL transformation is defined within a Lakeflow pipeline, the platform materializes it as a managed table whose definition includes the transformation logic itself. Because the command becomes part of the table definition, Lakeflow can inspect the transformation, evaluate its lineage, and compare it against the Delta history of upstream datasets.

This allows the pipeline to determine what actually needs to run next. If upstream data changes, the system recomputes only the affected transformations. If logic changes, deterministic recomputation can occur across historical data. Instead of rerunning entire pipelines blindly, the system advances the pipeline state intelligently.



By tracking lineage, table state, and upstream changes, Lakeflow intelligently determines when to recompute, skip, or fully reprocess data. This ensures pipelines remain both efficient and reproducible as data and logic evolve.

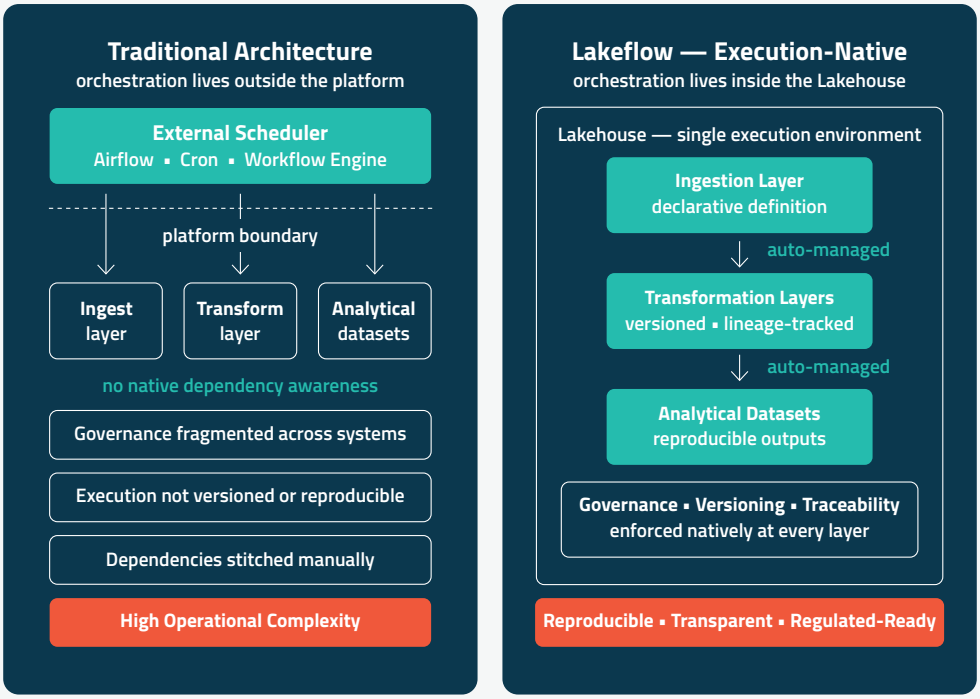
The operational impact is significant. Large-scale recomputation tasks that previously required hours of full pipeline execution can often be reduced to minutes because the platform understands both the transformation logic and the historical state of the data.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

Execution-Native Orchestration

Traditional data architectures often rely on external schedulers and workflow tools to coordinate pipelines. This creates operational complexity: orchestration logic lives outside the data platform, dependencies are fragmented across systems, and governance becomes harder to enforce consistently.

Orchestration Architecture



Embedding orchestration directly within the Lakehouse eliminates external dependencies and simplifies pipeline management. This approach improves traceability, governance, and reliability in complex, regulated healthcare environments.

Lakeflow introduces execution-native orchestration, where pipeline coordination lives inside the Lakehouse itself. Ingestion, transformation layers, and downstream analytical datasets are defined declaratively as part of the same pipeline definition. The platform automatically manages dependencies, execution order, and incremental updates.

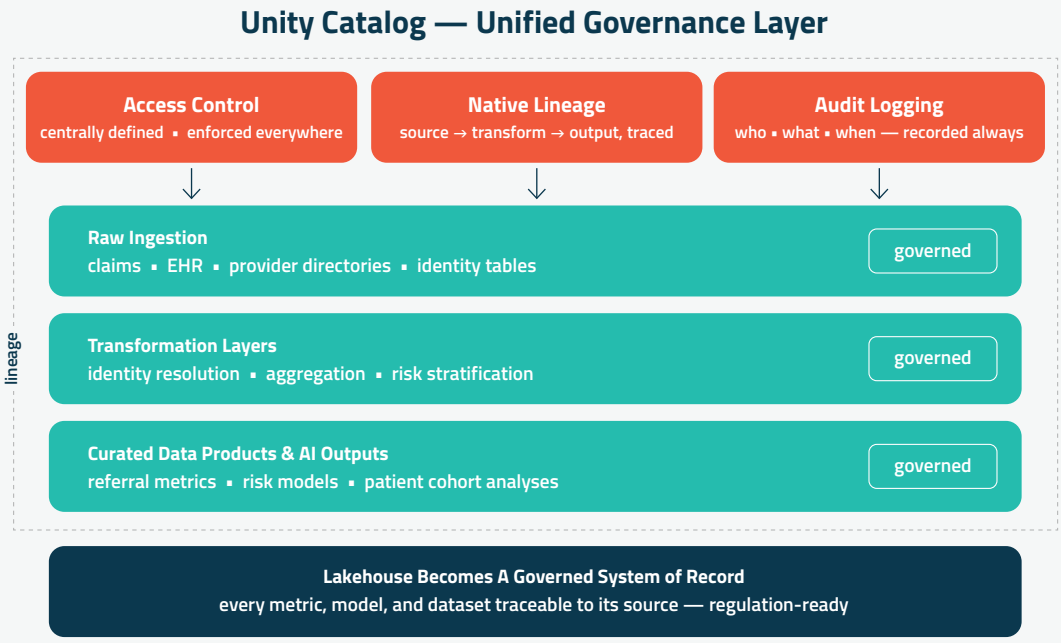
This approach eliminates the need to stitch together external triggers and workflow engines. Instead, orchestration becomes an intrinsic part of the data platform. Workflows are versioned, traceable, and executed within the same environment where data transformations occur.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

For regulated healthcare workloads, this architectural discipline is essential. Execution logic must be transparent and reproducible so that analytical outputs, whether referral metrics, risk models, or patient cohort analyses, can be reliably reconstructed.

Unified Governance with Unity Catalog

In regulated domains, governance cannot be layered onto analytics after the fact. It must be embedded directly into the architecture of the platform.



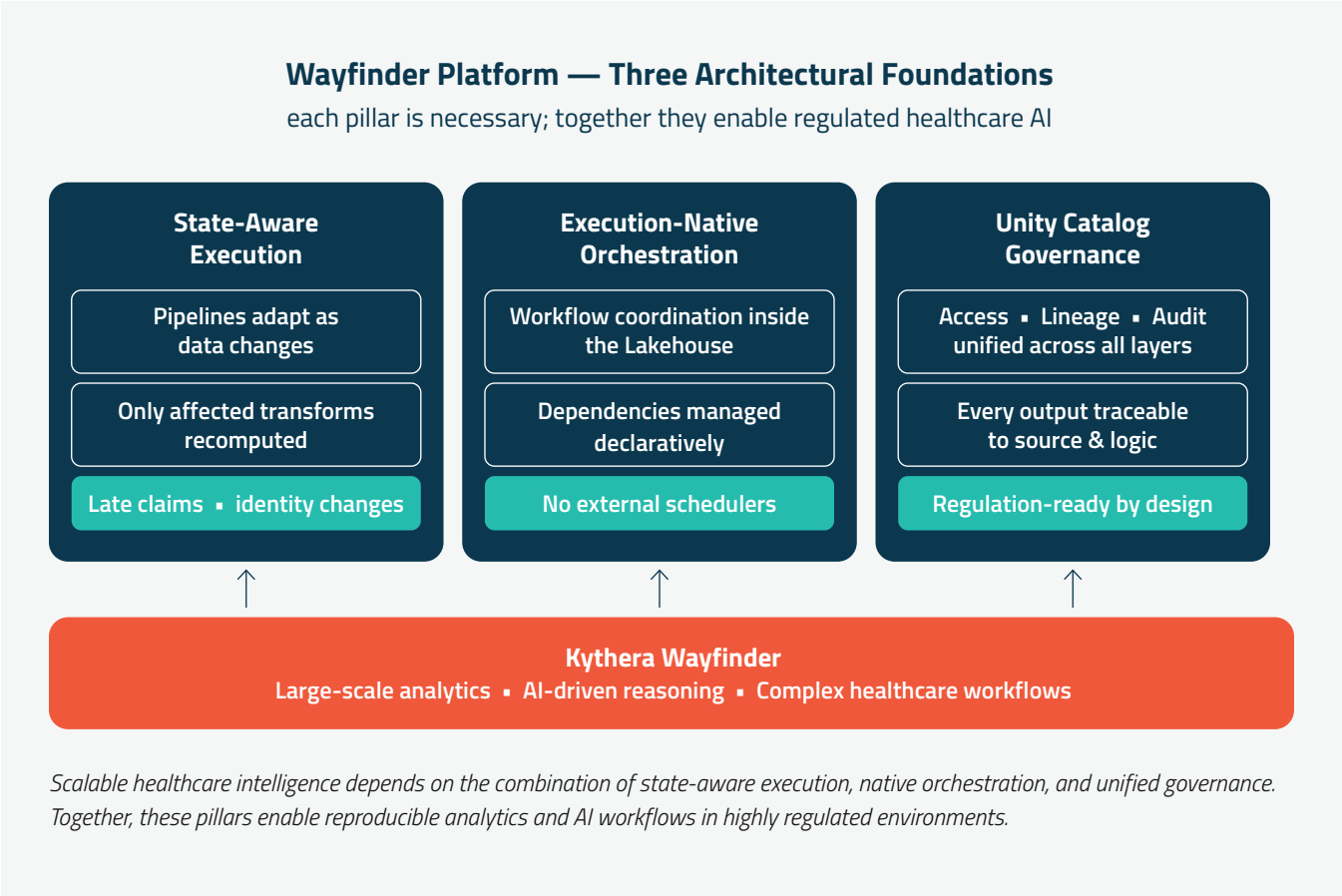
Centralized governance enforces consistent access control, lineage, and auditability across all data and AI assets. This creates a fully traceable system of record, ensuring compliance and trust in healthcare analytics outputs.

Unity Catalog provides a unified governance layer across data, pipelines, and AI workloads. Access policies are defined centrally and enforced consistently across tables, views, notebooks, and pipelines. Sensitive healthcare datasets such as claims records, provider directories, and patient identity tables can therefore be controlled through a single, governed framework.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

Equally important is native data lineage. Unity Catalog tracks how data moves from raw ingestion through transformation layers to curated data products and AI-driven outputs. Every metric, model input, or analytical dataset can be traced back to its source tables and transformation logic.

Audit logging further strengthens this governance model by recording who accessed data, what transformations were executed, and how datasets evolved over time. For healthcare organizations operating under regulatory and contractual scrutiny, this centralized visibility simplifies compliance while strengthening trust in analytical outputs.



Scalable healthcare intelligence depends on the combination of state-aware execution, native orchestration, and unified governance. Together, these pillars enable reproducible analytics and AI workflows in highly regulated environments.

Together, these capabilities transform the Lakehouse from a simple analytics environment into a governed system of record for healthcare data. State-aware execution ensures pipelines adapt intelligently as data changes. Execution-native orchestration embeds workflow coordination directly within the platform. Unity Catalog provides the governance and lineage required for regulated environments.

This architectural foundation is what allows Kythera's Wayfinder Platform to support large-scale analytics, AI-driven reasoning, and complex healthcare workflows with confidence.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

Referral Network Computation with Distributed Spark

Once patient identity and coverage timelines are stabilized, healthcare utilization patterns can be analyzed at scale. One of the most operationally important of these analyses is referral network behavior.

Referral analytics requires longitudinal linking of primary care encounters to downstream specialist utilization across multi-year claims datasets. These are not simple aggregations. Referral relationships form a distributed network structure where patients move through sequences of providers, facilities, and care settings over time.

At national scale, this network spans billions of encounters and millions of providers, making traditional relational analysis insufficient.

Apache Spark enables Kythera to compute referral network metrics across this distributed graph efficiently. Primary care visits, specialty encounters, and facility utilization are linked through identity-resolved patient journeys, forming referral edges that describe how patients move through the healthcare system.

From this structure, Kythera computes metrics such as:

- network stability across provider groups
- specialty concentration and fragmentation measures
- site-of-care migration patterns

Provider hierarchies and facility affiliations are incorporated as governed transformation layers so referral mapping reflects real-world alignment structures rather than billing artifacts.

Because these computations are defined declaratively within Lakeflow pipelines, attribution logic can evolve without destabilizing the underlying architecture. Updated definitions such as changes to referral attribution rules or provider hierarchy mappings can be versioned and recomputed deterministically across historical datasets.

Answering referral questions reliably requires reconstructing the operational reality of healthcare delivery:

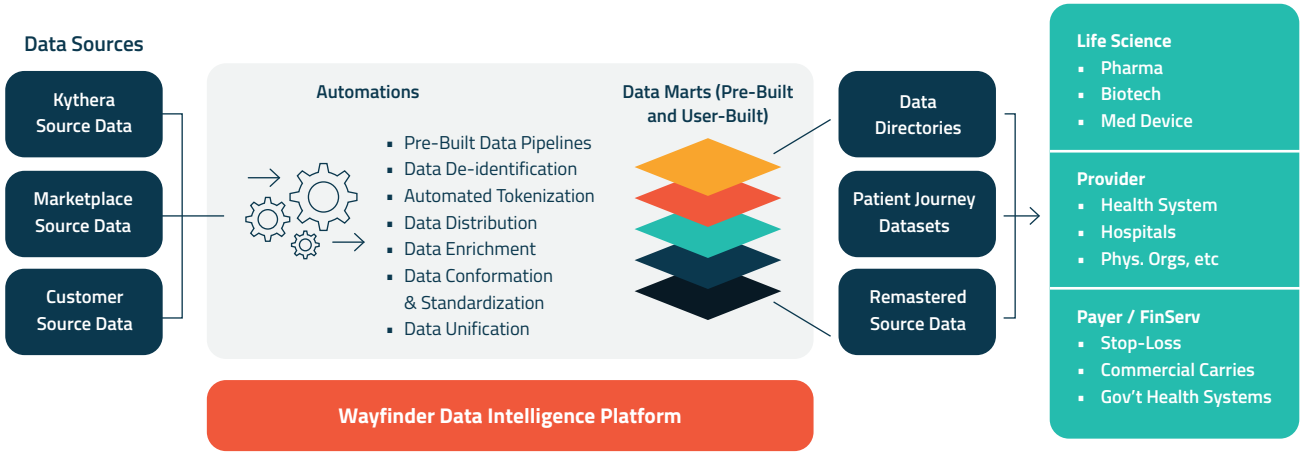
- rebuilding longitudinal patient journeys across fragmented claims feeds
- validating coverage continuity and payer exposure
- normalizing provider and facility hierarchies
- resolving identity across systems and transactions
- preserving full lineage for audit and regulatory review

While the Lakehouse architecture enables these computations at scale, organizations still need a way to operationalize this intelligence within decision workflows.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

Operationalizing Healthcare Intelligence with Wayfinder

The Lakehouse architecture enables large-scale healthcare computation, but organizations still need a way to translate that analytical capability into operational insight. This is where Wayfinder becomes critical.



Wayfinder integrates ingestion, transformation, and analytics into a unified workflow that delivers decision-ready healthcare data products. This architecture enables organizations to move from raw data to actionable insight without managing fragmented pipelines.

Wayfinder libraries connect the Databricks tools to Kythera's and customer's data to produce decision ready data instead of multiple pipelines and raw data. Wayfinder surfaces governed healthcare data products through decision-focused workflows designed for health system and life sciences teams.

Within this architecture:

- Lakeflow manages ingestion and pipeline orchestration
- Delta Live Tables define declarative transformation layers
- Identity mastering produces a persistent patient spine
- Coverage modeling stabilizes utilization denominators
- Referral graphs are computed from harmonized patient journeys
- Unity Catalog enforces governance, lineage, and access control

Because these components operate within a unified Lakehouse environment, Wayfinder can deliver analytical outputs that remain reproducible, auditable, and continuously recomputable as new healthcare data arrives.

Instead of relying on retrospective dashboards, Wayfinder enables organizations to investigate healthcare dynamics through structured analytical workflows. Users can explore questions such as:

Which patient cohorts have experienced shifts in care patterns over the past 60 days, and what is the downstream financial or therapeutic impact?

Copyright© 2026 Kythera Labs Inc. All rights reserved.

For a health system, this might mean identifying primary care physicians whose referral patterns are shifting outside the network and understanding the potential effect on downstream specialty revenue or risk contracts.

For a life sciences organization, the same architecture can detect changes in treatment patterns within a target patient population and evaluate how provider networks influence therapy adoption.

In this way, Wayfinder transforms governed Lakehouse data products into continuous healthcare observability, allowing organizations to monitor evolving care delivery.

Key Takeaways

- Healthcare intelligence requires architectural discipline. Fragmented claims, EHR, and provider data must be transformed into identity-resolved, coverage-aware data products before reliable analytics or AI can occur.
- Lakehouse architecture enables governed healthcare data systems. By unifying ingestion, transformation, analytics, and governance within Databricks, organizations can manage large-scale healthcare data pipelines more reliably.
- State-aware pipelines and execution-native orchestration simplify complex workflows. Lakeflow allows pipelines to recompute deterministically as healthcare data evolves, reducing operational complexity while improving reliability.
- Distributed Spark computation makes large-scale healthcare network analysis possible. Referral patterns, care journeys, and treatment dynamics can be modeled across billions of encounters using identity-resolved patient data.
- Operational intelligence emerges from governed data products. Platforms like Wayfinder translate Lakehouse data architecture into decision workflows that allow healthcare and life sciences organizations to monitor care patterns and respond to change in near real-time.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

Engineering the Foundation for Healthcare Intelligence

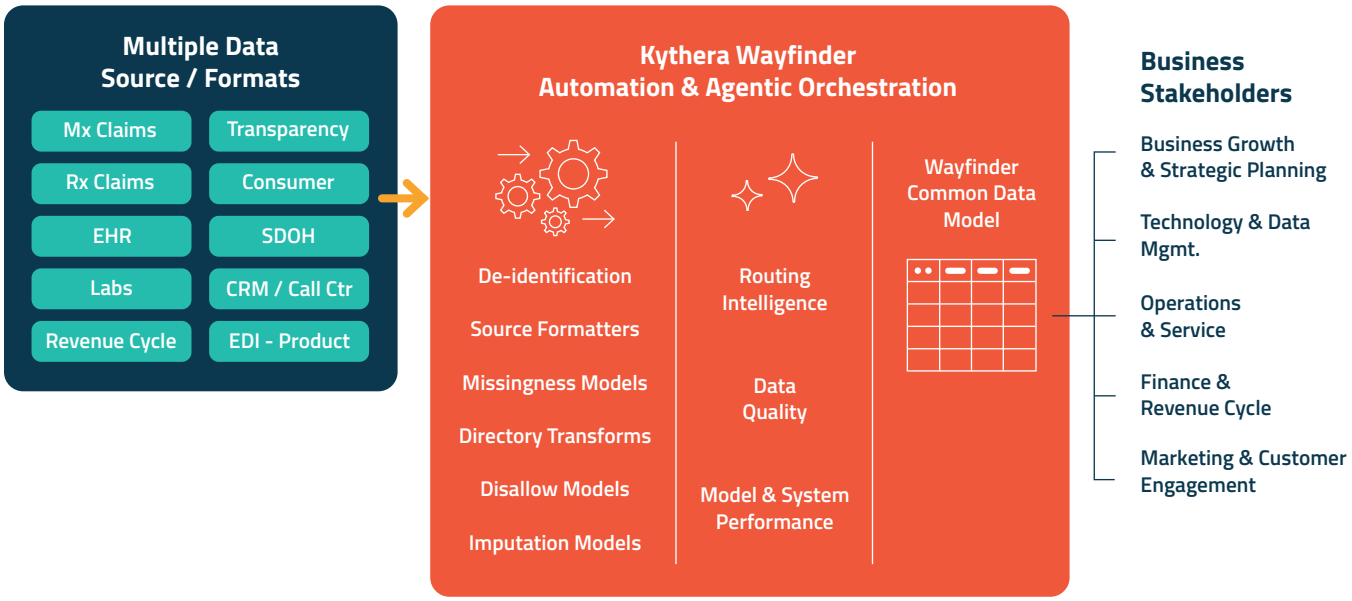
Healthcare data systems are often built incrementally with new dashboards layered onto existing pipelines and analytics tools added as new questions emerge. Over time, these environments accumulate complexity, making it increasingly difficult to ensure that analytical results remain consistent, explainable, and trustworthy.

Kythera's approach starts from a different premise: intelligence must be built on a disciplined data architecture. The Databricks Lakehouse provides the foundation for this approach by unifying large-scale data processing, versioned storage, and governance within a single platform. Lakeflow enables execution-native orchestration and state-aware pipelines that keep transformation logic deterministic even as healthcare data continuously evolves. Unity Catalog ensures that data lineage, access control, and auditability remain embedded across every layer of the platform.

On top of this foundation, Wayfinder transforms governed healthcare data products into operational intelligence workflows, allowing organizations to observe changes in care delivery, referral networks, and treatment patterns as they happen.

Agentic Orchestration Foundation

Proactively managing data workflows to serve multiple stakeholders across the enterprise.



Agentic orchestration coordinates complex data workflows across diverse sources, models, and business functions. This enables organizations to operationalize data intelligence across stakeholders while maintaining quality, performance, and governance.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

As healthcare and life sciences organizations increasingly rely on data to guide strategic decisions, the challenge is no longer simply generating analytics. It is building systems that can continuously interpret complex healthcare data while remaining explainable, reproducible, and governed.

In healthcare, intelligence does not emerge from data alone, it emerges from the architecture that makes that data understandable.

Connect with Kythera. Kythera is a data technology company that brings unprecedented clarity and structure to complex real-world healthcare data. Kythera's Wayfinder Technology Platform, supported by pre-configured pipelines, processing libraries, analysis tools and remastered datasets, helps Healthcare and Life Sciences organizations work with greater speed, scale and confidence. Learn more at www.kytheralabs.com.

Copyright© 2026 Kythera Labs Inc. All rights reserved.

KYTHERA™
DECIPHERING HEALTHCARE

346 Main Street
Franklin, TN 37064
www.kytheralabs.com