KYTHERA
DECIPHERING HEALTHCARE

Wayfinder DataSync

A Scalable Architecture for Integrating Provider and Multi-Source Healthcare Data

Executive Summary

Healthcare organizations generate vast amounts of data from patient care, billing, and payer interactions, yet much of that data remains underutilized. Fragmented systems, inconsistent identifiers, and privacy constraints make it difficult for data and technology leaders to see a unified picture of patient journeys or market performance.

Wayfinder DataSync represents a significant advancement in healthcare data engineering. It unifies a provider's internal billing and operational data with Kythera's multi-source third-party healthcare data, creating a comprehensive, analysis-ready foundation.

By combining privacy-preserving de-identification, machine learning-based record linkage, and automated quality validation, DataSync transforms disconnected data streams into a governed, scalable, and auditable architecture that enables confident decision-making across clinical, financial, and strategic domains.

The Technical Challenge in Context

Health systems face not only a "data volume" problem but also a "data orchestration" problem. Data originates from dozens of internal and external systems, including EHRs, billing, payer feeds, scheduling, and laboratory systems, each using different schemas, coding standards, and identifiers.

The result is fragmentation that prevents data teams from maintaining referential integrity or confidence in analytic outputs. Key challenges include:

- **Disparate Schemas and Standards:** Variations across HL7® FHIR®, and proprietary formats complicate integration.
- **Inconsistent Identifiers:** Patients may appear under multiple IDs or incomplete demographic details.

- **Privacy and Compliance Constraints:** HIPAA requirements demand de-identification, but naïve approaches can break record linkage.
- **Data Quality and Lineage Gaps:** Duplicates, missing values, and lack of version control erode trust in downstream analytics.

Traditional ETL pipelines are not equipped to handle these challenges at the scale or complexity of healthcare. DataSync was designed to fill that gap.

Technology Overview: The DataSync Architecture

Wayfinder DataSync is a modular, high-performance data orchestration system that automates complex healthcare data management at scale. DataSync provides a managed service model in which client environments are provisioned under Kythera's Databricks account, ensuring both control and governance.

Each instance is securely isolated, allowing Kythera to handle orchestration, monitoring, and optimization while the healthcare organization retains full ownership of its data.

1. Ingestion & Standardization

DataSync ingests both provider and third-party data sources, including billing data, payer claims, and reference datasets, into a governed Wayfinder environment. Schema mapping and normalization occur automatically through configurable templates aligned with healthcare ontologies such as CPT, HCPCS, and ICD-10.

2. De-Identification Layer

Patient privacy is maintained through a trusted, industry-standard tokenization process that replaces personally identifiable information (PII) with non-reversible tokens before data enters the Wayfinder environment. Kythera's Wayfinder enables this process through its governed architecture, ensuring that de-identified records can be securely linked and analyzed across multiple datasets while maintaining HIPAA compliance.

Because Wayfinder DataSync operates under a Bring Your Own Data (BYOD) model, organizations can integrate their internally de-identified data or apply existing tokenization workflows within the same privacy-preserving framework.

3. Patient Mastering Engine

The Patient Mastering Engine applies a blend of deterministic and machine learning-based probabilistic matching to align patient identifiers across millions of records. It uses feature creation (e.g., ZIP + DOB + NPI overlaps), blocking logic, and fuzzy-matching algorithms to improve linkage accuracy and reduce duplication.

4. Harmonization & Enrichment

Once patient identities are mastered, DataSync harmonizes data by aligning code sets and structures across multiple vendor formats. It applies address standardization at big-data scale, converting every address to USPS format and appending latitude and longitude coordinates for geospatial analysis.

Additional enrichment steps include:

- Column value standardization and correction of nulls, invalid dates, or malformed entries.
- Imputation of missing facility addresses and correction of provider misattribution using NPI logic.
- Integration of Kythera's curated directories to append practitioner, organization, and taxonomy attributes.

5. Governance, Orchestration & Monitoring

DataSync employs Apache Airflow to orchestrate pipeline execution across multiple job types, including incremental updates, full "base" cycles, and directory refresh flows. These processes maintain data consistency and reduce latency without reprocessing entire datasets.

Data validation and quality checks are built into each stage of the workflow, ensuring that data completeness, duplication, and match confidence are continuously monitored. Each record maintains full lineage through Databricks Unity Catalog, enabling traceability and auditability for every transformation.

Key Technical Differentiators

- **Privacy-Preserving Linkage:** Secure tokenization and pseudonymization maintain compliance while enabling cross-dataset record linkage.
- **Healthcare-Specific Processing:** The Wayfinder features JAR extends Databricks with domain-specific functions such as NPI validation, gender normalization, NDC code translation, and address imputation.
- **ML-Driven Record Matching:** Probabilistic algorithms enhance accuracy even with incomplete identifiers.
- **Automated Validation & Benchmarking:** Continuous rule-based checks ensure analytic reliability across billions of rows.
- **Scalability & Governance:** Built on Databricks' distributed compute infrastructure with full lineage tracking and access control.
- **Flexible Integration:** Supports both Kythera-supplied and customer-provided datasets within a single governed instance.

CASE EXAMPLE:

Integrating Provider and Multi-Source Data for Measurable Results

A regional healthcare system deployed Wayfinder DataSync to integrate its internal billing data with Kythera's multi-source claims data. Prior to implementation, the organization had limited visibility into patient movement beyond its own facilities and could not quantify referral leakage.

Process

1. Data was ingested from internal billing systems and external payer claims into a secure Wayfinder environment.
2. Tokenization de-identified patient identifiers and enabled compliant record linkage. Token linkage: ~98% average match rate.
3. Patient Mastering applied probabilistic matching to more than two million records, improving linkage success rates by over 20 percent.
4. Address and code harmonization standardized inputs across systems, creating a single, reliable dataset.
5. Enrichment routines imputed missing referral and facility data, enabling complete patient journey reconstruction.

Outcomes

- **Data Completeness:** Improved patient visibility from approximately 75 percent to more than 90 percent.
- **Leakage Detection:** Uncovered a 23% uplift in leaked encounters identified.*
- **Leakage Reduction:** Reduced leakage data noise 11-36% across service lines.**
- **Revenue Recovery:** Quantified nearly $3 million in recoverable revenue tied to out-of-network total knee procedures.
- **Operational Efficiency:** Reduced data preparation cycles from weeks to hours.

*ER Visit to specialty

**Primary care visit to specialty

Integration and Deployment Model

Wayfinder DataSync is delivered as a managed Databricks environment under Kythera's governance model. This allows healthcare organizations to maintain control of their data while Kythera manages orchestration, optimization, and monitoring.

The Wayfinder Asset Distribution framework synchronizes internal and external data via a secure S3-based DataMart. Automated replication jobs transfer and register assets into the client's environment, with full authentication and audit logs for traceability.

Deployment timelines are minimal. Once customer data access is established, average deployment time is approximately one week. DataSync supports multi-cloud operation (AWS, Azure, or GCP) and can run either in the client's Databricks workspace or as a Kythera-managed instance.

Conclusion

Wayfinder DataSync establishes a governed, automated, and scalable framework for healthcare data integration. By unifying internal provider data with Kythera's multi-source healthcare data within the Databricks environment, it enables organizations to maintain data quality, ensure compliance, and generate analysis-ready datasets that support strategic and operational objectives.

DataSync provides a durable technical foundation for ongoing data engineering and analytics innovation. Its flexible architecture can be deployed within a customer's Databricks environment or operated as a Kythera-managed instance, ensuring performance, security, and governance requirements are met in any configuration. This adaptability allows healthcare organizations to evolve from fragmented, siloed information to a continuously improving, high-confidence data environment.

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.

KYTHERA™
DECIPHERING HEALTHCARE

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