phuse·share
connect·share·advance

Single Day Event

AI-Powered Clinical Data Sciences: Strategies for Speed, Accuracy, and Compliance

Indrasena Reddy Purini
Director, Elluminate services
eClinical solutions

phuse.global

The Challenge: Traditional Data Mapping

Labor-Intensive & Slow

Manual interpretation of diverse data dictionaries, study protocols, and regulatory standards like CDISC is slow and requires deep expertise.

High-Risk & Inconsistent

Misinterpretations from ambiguous variable names or undocumented derivations lead to costly compliance risks and data inconsistencies.

Heterogeneous & Complex

Data is collected from numerous, disconnected sources (eCRFs, lab systems, devices, ePROs) that are all structured differently.

What Are Agentic LLM Workflows?

An orchestration of autonomous Large Language Models (LLMs) that function as agents within a structured sequence of tasks. They are dynamic, adaptive, and capable of handling ambiguity.

- **Dynamic & Adaptive:** Unlike static scripts, they can reason, plan, and adapt to unforeseen challenges in the data.
- **Access to Specialized Tools:** Agents are given tools to use, such as code interpreters (Python, SQL) and API access to perform tasks.
- **Advanced Reasoning:** Capable of complex decision-making, such as inferring relationships between disparate datasets.
- **Verifiable Audit Trail:** Generates a complete, human-readable audit trail for GxP, documenting every decision and transformation.

The Agentic Process: Reflection

How Reflection Works

Agents are dynamic and iterative. They don't just provide the first answer. This "reflection" loop allows the agent to handle ambiguity and dramatically improve accuracy.

- An initial response is generated.
- The agent critiques its own work based on a set of rules.
- It incorporates this "self-feedback" to create an improved response.
- The final, validated response is delivered.

The 3-Phase Transformation Process

Phase 1: Autonomous Understanding

Schema Inference

Agents deduce field meanings from headers, data types, and sample values (e.g., 'VSDT' and 'visit_date' both mean "Visit Date").

Terminology Mapping

Matches local shorthand (e.g., "N/V") to standard, controlled vocabularies like CDISC SDTM or MedDRA.

Data Normalization

Standardizes inconsistent formats for dates (DD/MM/YYYY → YYYY-MM-DD) and units (mg → µg) across all sources.

Phase 2: Intelligent Query Generation

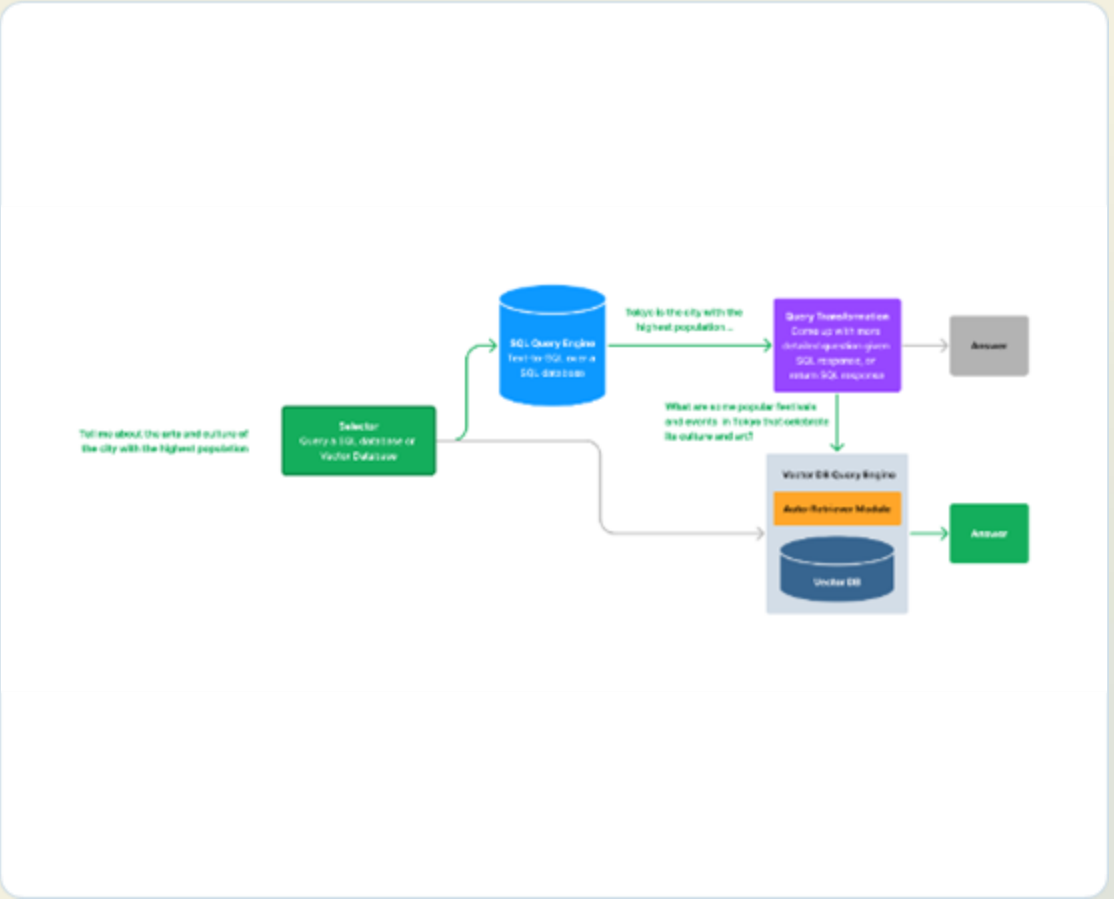
From Natural Language to SQL

Agents translate high-level clinical requests into executable, auditable queries.

Request: "Calculate average SBP at baseline and week 12 for all active-arm subjects."

Agent Actions:

- Automatically identifies keys (Subject ID, Visit) to join disparate datasets (e.g., Demographics, Vitals).
- Precisely handles complex date logic and visit windows.
- Selects the correct SQL functions (AVG, WHERE, GROUP BY).



Phase 3: Auditable Data Transformation

Automates SDTM/ADaM mappings using validated functions and concurrent documentation, ensuring every step is traceable for FDA/EMA audits.

- **Executable Code:** Generates modular Python or SQL scripts using validated, reusable functions (e.g., `decode()`, `datetime()`) for reliable conversion.
- **Structured Documentation:** Concurrently produces human-readable files (e.g., Markdown) detailing the logic, source/target fields, and all references.
- **Regulatory Readiness:** The code and its documentation are generated together, creating a transparent, auditable package ready for regulatory review.

The Impact: Transforming Clinical Data Workflows

Speed & Efficiency

Drastically reduces mapping and transformation timelines from weeks or months down to hours or days.

Accuracy & Consistency

Eliminates manual errors from misinterpretation and ensures standardized pipelines are used across all studies.

Compliance & Transparency

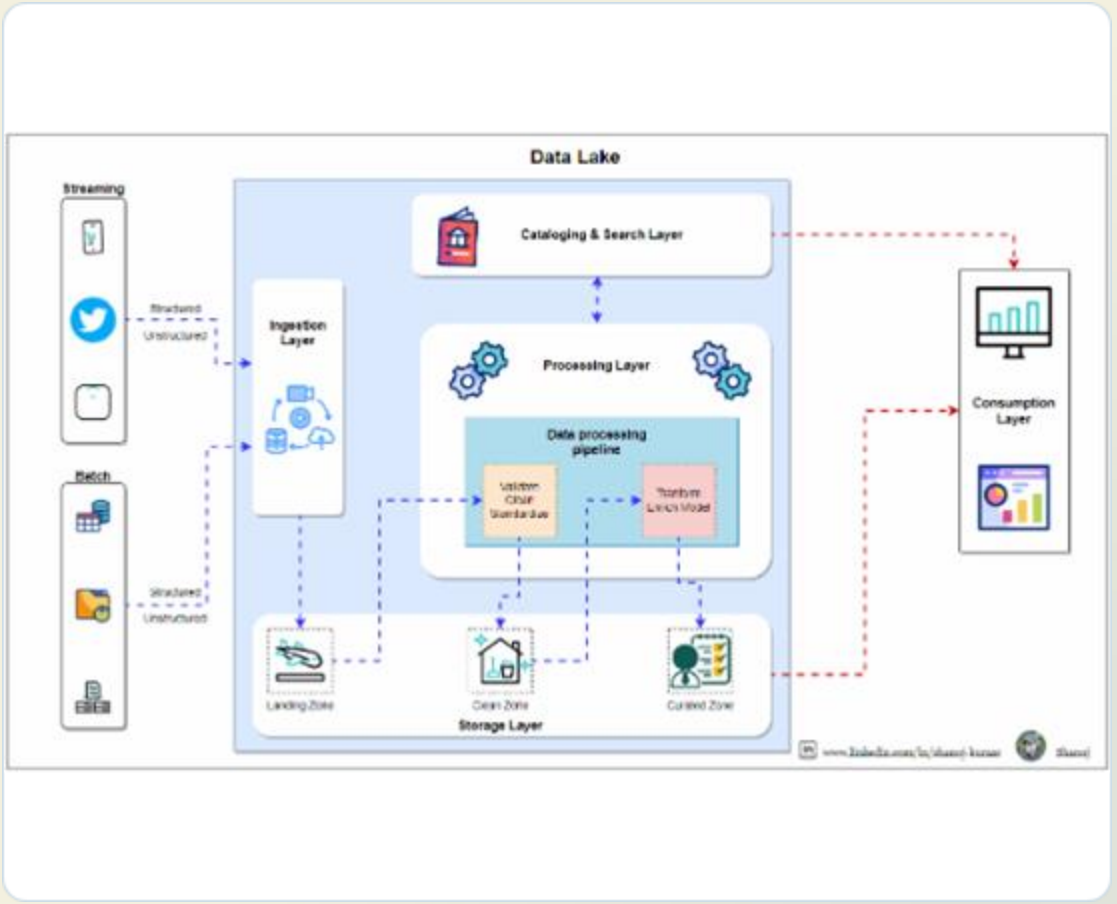
Generates complete, verifiable, and transparent audit trails for GxP, making the process "audit-ready" by design.

Proposed GxP-Compliant Architecture

A Governed, HITL Platform

This is not a single, uncontrolled model. It is a governed, human-in-the-loop (HITL) platform designed for compliance.

The architecture separates raw data, a semantic knowledge base, and a sandboxed execution environment. All components are governed by an "Orchestration Engine" that logs every action and manages the workflow.



Key Architectural Components

- **Data Foundation:** An immutable raw data store (Data Lake) where source data is preserved in its original state.
- **Knowledge Base (RAG):** Grounds agents in your specific context—study protocols, data dictionaries, and CDISC standards.
- **Tools Tier:** A controlled, sandboxed environment for safe code execution, preventing agents from affecting production systems.
- **Orchestration Engine:** The "Maestro" that assigns tasks to agents, governs the workflow, applies QC, and logs all activities.
- **User Applications:** Dashboards for human review, modification, and final approval of agent-generated work.

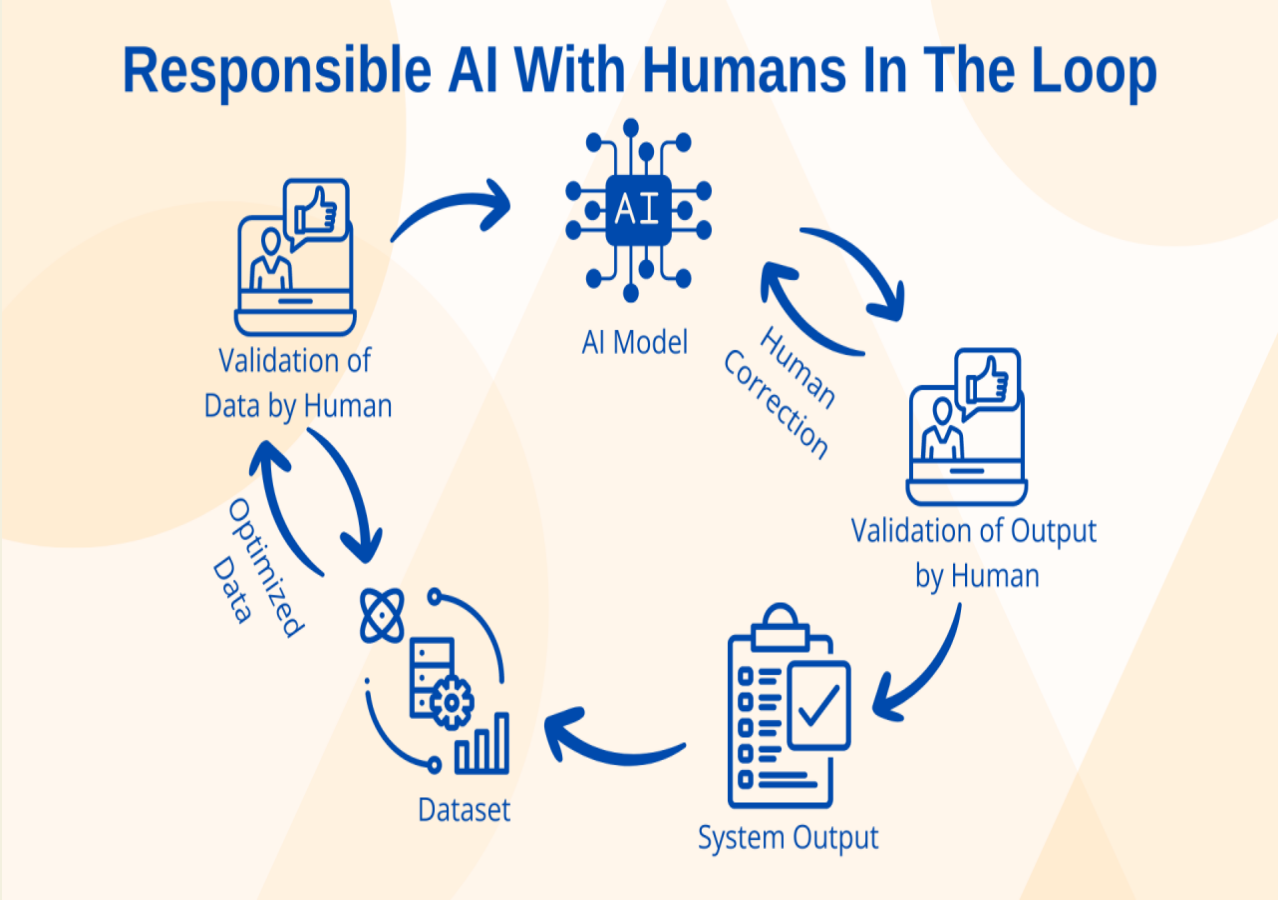
Human-in-the-Loop (HITL) Interface

Oversight, Not Automation

This is not a fully autonomous "black box." The platform is built to empower, not replace, clinical data managers and programmers.

Interactive dashboards ensure human experts have final oversight, control, and approval.

- AI-Assisted Mapping Suggestions
- Auto-Transformation Previews
- Validation & QC Dashboards
- One-Click Approval/Rejection



Compliance, Governance & Security

The entire architecture is designed to meet GxP and FDA 21 CFR Part 11 compliance requirements from the ground up.

- **Role-Based Access Control (RBAC):** Ensures users can only access data and perform actions appropriate for their role.
- **Tamper-Evident Logs:** All agent actions, user approvals, and data transformations are logged in a verifiable, immutable record.
- **PHI Redaction & Security:** Automated processes to redact or mask Protected Health Information (PHI) from logs and models.
- **Continuous Validation:** The platform and its components are subject to continuous validation and testing protocols.

Summary & Next Steps

Summary

Agentic LLMs redefine clinical data transformation. They move the industry from a manual, high-risk, and slow process to an automated, auditable, and intelligent workflow that delivers speed, accuracy, and compliance.

Next Steps

We are actively developing this future state and invite collaboration.

- Pilot Implementation on historical studies
- Integrate expert feedback from Data Managers
- Expand and refine model training
- Build scalable, multi-study ecosystems

Questions?

The Global Healthcare Data Science Community

Contact Channels
UK +44 1843 609600
office@phuse.global
phuse.global

Social Media
@phusetwitta
/phusebook
/phusetube
/company/phuse