



HEXAWARE

AI-Powered Digital Standards Library for Accelerated Regulatory Submissions

February 2026

CHAPTER 1

The Clinical Research Challenge

Current State Problems

The clinical research ecosystem is constrained by fragmented processes, manual review, and distributed tooling across SAS, R, and Python. These constraints cause unnecessary site burden, inflated costs, and delayed Database Lock (DBL).

In oncology, each day of DBL delay represents an opportunity cost of approximately \$840,000, underscoring the urgent need for modernization.

Traditional pipelines emphasize deterministic edit checks and manual reviews, creating bottlenecks that slow down the entire research process and increase the risk of errors.

The Multi-Language Complexity

Modern clinical research operates in a **multi-language ecosystem** that must integrate:

- **SAS:** Reliability, standardized procedures, proven regulatory alignment
- **R:** Extensive statistical libraries, advanced visualization, community-driven standards
- **Python:** Machine learning, big data processing, cloud-native integration

Organizations often lack a cohesive CI/CD foundation ensuring interoperability, reproducibility, compliance, and observability across all three languages at scale.

Our Solution: Standards-Led Agentic AI Platform

We introduce a standards-led, agentic AI platform and CI/CD pipeline that transforms data management from reactive oversight to proactive, autonomous governance.



Multi-Language CI/CD Architecture

End-to-end pipeline enabling SAS/R/Python interoperability atop standardized data (CDASH/SDTM/AdAM) and exchange formats (FHIR, Dataset-JSON)



Cloud-Native Deployment

Containerized deployment pattern providing reproducibility, scalability, and modularity across agents using Docker/Kubernetes



Advanced Validation Framework

Combines predictive ML/NLP risk detection, deterministic compliance checks, and external regulatory context for higher-fidelity oversight



Guardrails Generative Reporting

Automated CSR drafting and TLF generation using standardized AdAM with sentence-level provenance

Demonstrated Impact: Transformative Results

6x

Throughput Increase

More than six-fold increase in data cleaning throughput compared to traditional manual processes

54.67%

Error Reduction From

Baseline cleaning error rate before implementation of the autonomous platform

8.48%

Error Reduction To

Achieved cleaning error rate after platform deployment, representing an 85% improvement

13.6%

DBL Acceleration

Database Lock accelerated by at least 5 days per study

Cost Savings Per Phase III Trial

The platform translates into **millions of dollars in cost avoidance per Phase III trial**, with direct savings exceeding **\$361,500 per study**.

In oncology contexts where each day of DBL delay represents approximately \$840,000 in opportunity cost, a 5-day acceleration delivers substantial financial and competitive advantages.

The pipeline elevates auditability (21 CFR Part 11, GxP), traceability (sentence-level provenance), and operational scalability while enabling continuous, autonomous data quality management.



Regulatory Foundations: 21 CFR Part 11 & GxP Compliance

In the pharmaceutical sector, deployment of any software used to analyze or report clinical trial data must comply with **21 CFR Part 11** and **GxP (Good Practice) standards**. These regulations ensure that electronic records and signatures are as trustworthy and reliable as their paper counterparts.

ALCOA+ Principles for Data Integrity

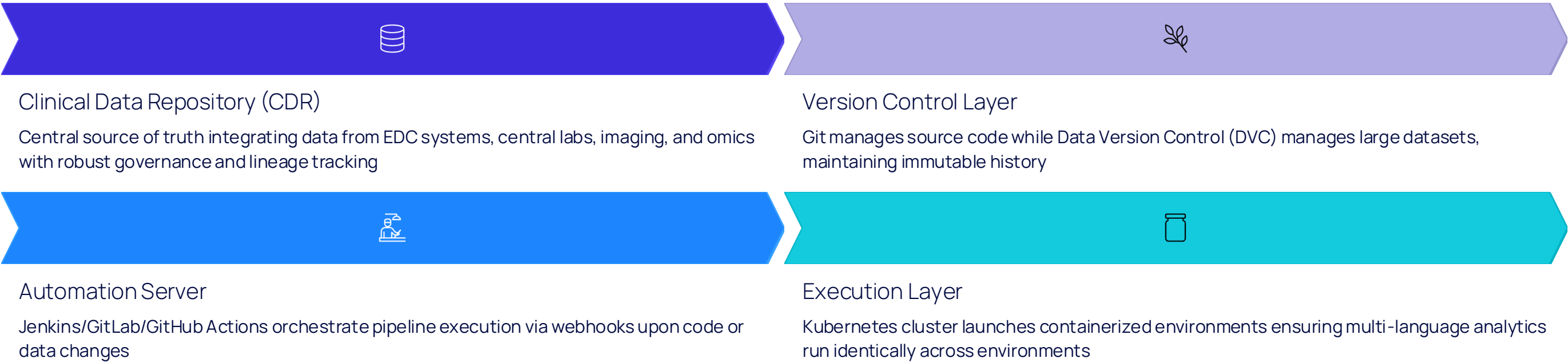
Data integrity is maintained throughout the automated pipeline by adhering to the ALCOA+ principles, which require data to be Attributable, Legible, Contemporaneous, Original, and Accurate.

	Attributable Requirement: Every data point or code change must be connected to a specific individual Implementation: Use of unique user accounts and digitally signed Git commits
	Legible Requirement: Data and logs must remain readable throughout the retention period Implementation: Automated archival of logs and outputs in human-readable formats
	Contemporaneous Requirement: Actions must be recorded at the time they occur Implementation: Computer-generated, time-stamped audit trails captured during pipeline execution
	Original Requirement: The primary record or a certified copy must be preserved Implementation: Immutable container images and version-controlled script repositories
	Accurate Requirement: Data must be correct, complete, and consistent Implementation: Automated unit testing, log parsing, and double-programming validation

 **Critical Context:** FDA findings show that approximately 40% of recent citations involve data integrity violations, such as missing audit trails or unauthorized data deletion. By automating the deployment process, organizations eliminate the manual "hand-offs" that are prone to such errors.

System Architecture: Digital Factory Assembly Line

The architecture functions as a digital factory assembly line, where multiple integrated layers harmonize to ensure that heterogeneous code and data are delivered with **verifiable reproducibility and GxP compliance**.



Key Architectural Components

Statistical Computing Environment (SCE)

- High-performance computing for concurrent usage
- Validated, compliance-first environment
- Supports SAS, R, and Python execution
- Unified workbenches with GUI access

Data & Model Lifecycle

- Data Version Control (DVC) for large datasets
- Model Registry (SAS Model Manager)
- Hashing to track data versions
- Links report versions to exact data versions

Compliance & Security

- Role-based access controls
- Time-stamped audit trails
- Electronic signature module
- Automated security scanning (SAST)

The Four Intelligent Agents

The Autonomous Multi-Agent Clinical Data Operations Platform (CDOP) comprises four specialized agents coordinated by the Agent Orchestration Layer (AOL), each automating critical aspects of the clinical data lifecycle.



Data Harmonization Agent (DHA)

Autonomously cleans and normalizes data within the Clinical Data Repository using AI models for entity resolution. Maps normalized data into standardized CDISC formats (CDASH → SDTM → ADAM) using R packages like {sdm.oak} and {admiral}, automating approximately 80% of standard domains. Generates auditable Python or R code artifacts ensuring traceability back to source data.



Risk & Validation Agent (RVA)

Functions as the Automated Validation Gateway, moving beyond simple log parsing to predictive quality assurance. Continuously monitors standardized datasets using predictive ML models (TensorFlow/Scikit-learn) and cross-references with external safety context via openFDA APIs. Detects protocol deviations and data drift in real-time, integrating CDISC CORE-aligned checks to ensure regulatory compliance.



Reporting & Submission Agent (RSA)

Utilizes finalized ADAM datasets to draft regulatory narratives and Clinical Study Report (CSR) sections using Retrieval-Augmented Generation (RAG). Ensures every numeric claim is linked back to specific data points in CDISC datasets, providing sentence-level provenance. Automates Tables, Listings, and Figures (TLFs) generation, shortening document development cycles by up to 68%.



Agent Orchestration Layer (AOL)

Coordinates closed-loop feedback between agents, scheduling tasks and managing escalations to Human-in-the-Loop (HITL) reviewers for critical decisions. Records every transformation, model version, and prompt in an Immutable Audit Store. Leverages Kubernetes to provision ephemeral containerized environments for multi-language execution, ensuring 21 CFR Part 11 and GxP compliance.

Implementation Workflow: From Ingestion to Submission

The end-to-end implementation follows a systematic workflow that automates approximately 80% of the clinical data lifecycle while maintaining rigorous compliance standards.



Continuous Integration & Deployment Best Practices

Continuous Integration (CI)

CI is the practice of frequently merging code changes into a shared repository, where each change triggers automated build and test processes.

Version Control with Git

Git creates a robust audit trail by documenting every change with clear author and timestamp. Organizations adopt "Conventional Commits" to standardize message formats:

- **fix:** Resolves bugs (e.g., "fix(adam): corrected age group derivation")
- **feat:** Adds new functionality
- **docs:** Updates documentation
- **test:** Adds or updates automated tests

Automated Testing Frameworks

- **Unit Testing:** {testthat} for R, SASUnit for SAS
- **Code Coverage:** {covr} measures test coverage
- **Double Programming:** Independent development with automated output comparison
- **Log Parsing:** Regex-based scanners flag warnings and errors

Continuous Deployment (CD)

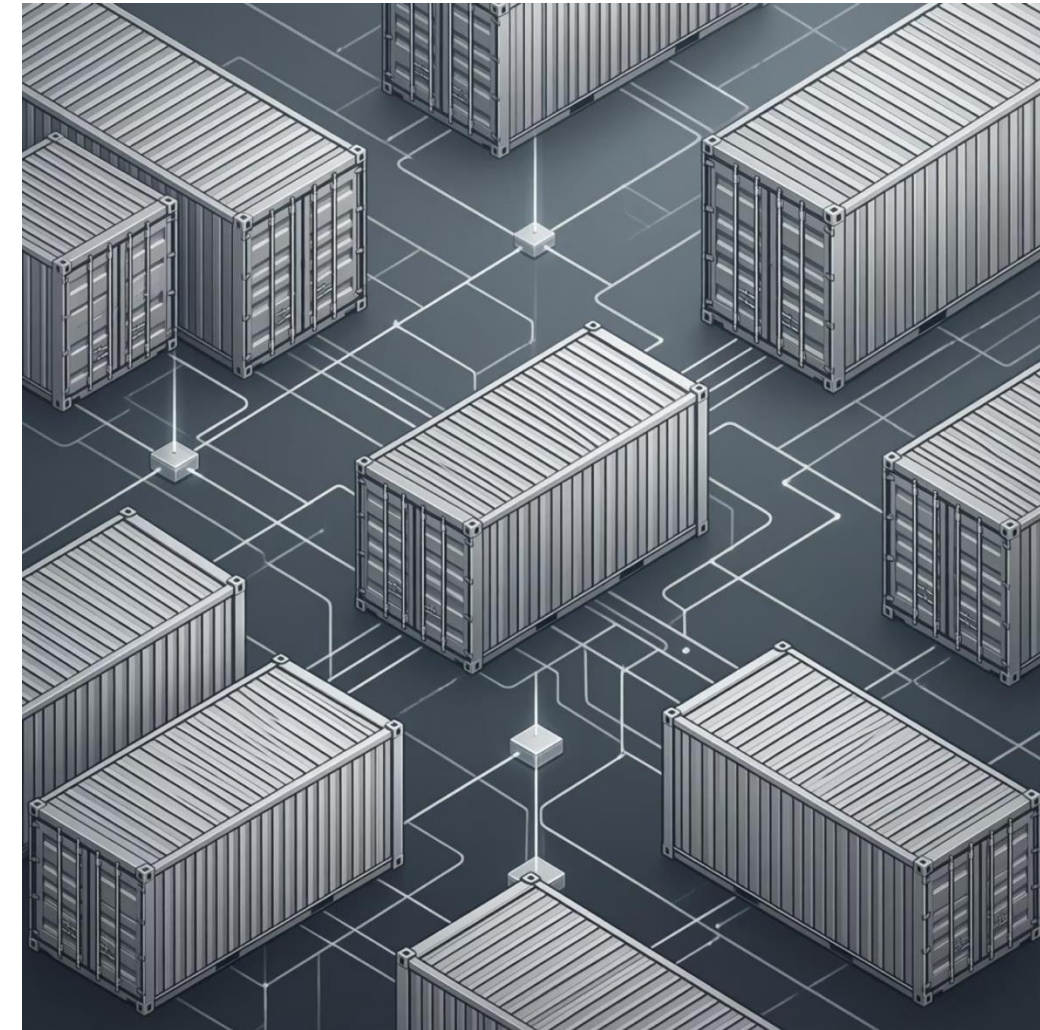
CD automates the release of validated code into production or staging environments, gated by quality checks and environment protections.

Containerization for Consistency

Containerization ensures software runs identically across development, testing, and production by bundling applications with their entire runtime environment.

- **Docker:** Mature configuration language, broad platform support, requires root permissions
- **Singularity (Apptainer):** Designed for HPC, runs as non-privileged user, automatic host filesystem integration

Containerization ensures **verifiable reproducibility**, allowing studies conducted years ago to be reconstructed with precision—a key requirement of EMA and FDA.



Conclusion: A Paradigm Shift in Clinical Research

The implementation of the **Autonomous Multi-Agent Clinical Data Operations Platform (CDOP)** signals a paradigm shift in how clinical trials manage the intersection of high-speed analytics and stringent regulatory requirements.

Key Achievements

The platform successfully harmonizes the reliability of SAS with the innovative statistical and machine learning power of R and Python, automating approximately **80% of the clinical data lifecycle**.

Through containerization (Docker/Kubernetes) and Data Version Control (DVC), the solution ensures verifiable reproducibility and absolute data integrity, adhering to ALCOA+ principles.

The empirical results demonstrate a **six-fold increase in data cleaning throughput** and error rate reduction from 54.67% to 8.48%, accelerating Database Lock by 13.6%.

Strategic Recommendations

Organizations should adopt these **automated, "compliance-first" environments** to eliminate manual hand-offs and ensure sentence-level provenance for all regulatory claims.

While SAS remains the authoritative system for core submissions, the integration of R and Python through a cohesive CI/CD foundation creates a **digital factory assembly line** that delivers high-quality clinical research with unprecedented speed and accuracy.

Contact Information

Authors: Avinash Patnaik and Varun Sethuraman

Company: Hexaware Technologies

Location: Pune, Maharashtra, India

Phone: 8280234647

Email: avinashp7@hexaware.com

Website: <https://www.hexaware.com>

Thank You

HEXWARE