AI-Ready Enterprise Data Pipeline & Governance Blueprint

FAIR, ALCOA+, & CDISC Compliant Architecture for R&D & Data Science

"Bridging Regulatory Integrity with Modern MLOps & Data Analytics"

Executive Summary

Modern Data Science, AI, and R&D initiatives require data that is both **highly utilitarian** and **strictly defensible**. By unifying **FAIR** principles (Findable, Accessible, Interoperable, Reusable) with **ALCOA+** integrity standards (Attributable, Legible, Contemporaneous, Original, Accurate, Complete, Consistent, Enduring, Available), this blueprint establishes a gold-standard data architecture for regulated environments.

The proposed framework leverages Data Vault modelling, Databricks Delta Lake (Medallion architecture), and dbt-driven DataOps to deliver auditable, historized, and AI-ready data products. Automated QA gates, cryptographic validation, and dynamic access controls ensure compliance with GDPR, HIPAA, FDA, and CDISC standards, while real-time telemetry and cross-functional stewardship align data engineering directly with ML/RAG workloads, feature stores, and advanced analytics. This proposal outlines a phased, production-ready roadmap to transform fragmented data pipelines into a trusted, scalable, and inspection-ready data ecosystem.

Conceptual Framework: FAIR × ALCOA+ for AI-Ready Data

Data Science teams thrive on reusable, semantically rich datasets, while regulatory bodies demand immutable provenance and strict access controls. This architecture resolves that tension by treating compliance as a *feature* of the pipeline, not a bottleneck:

- **FAIR = Data Utility for AI/ML:** Standardized metadata, ontology mapping, and catalog-driven discovery accelerate feature engineering, RAG context retrieval, and cross-study analytics.
- **ALCOA+ = Data Integrity for Compliance:** WORM storage, cryptographic hashing, system-level audit trails, and deterministic lineage guarantee legal defensibility and reproducibility.
- **Synergy:** Every dataset flows through automated validation gates that score both *utility* (FAIR) and *integrity* (ALCOA+), ensuring only trusted, model-ready data reaches the semantic layer or vector databases.

Phase 1: Architecture & Governance Foundation

Robust compliance and AI readiness require an architecture designed for historization, scalability, and semantic clarity across clinical, pre-clinical, and R&D domains.

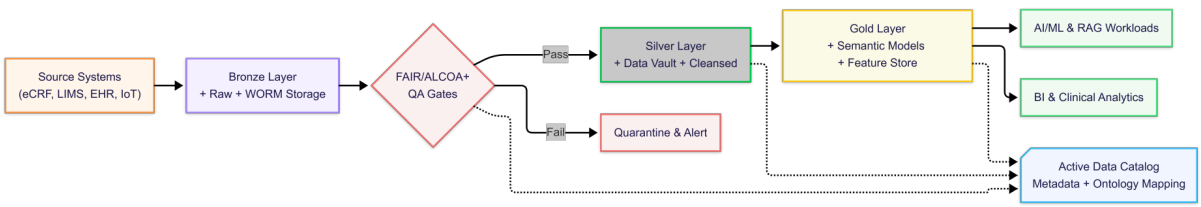

Figure 1: **End-to-End Lakehouse & Governance Flow:** Data moves from source systems through Bronze (raw/WORM) layers, passes automated FAIR/ALCOA+ QA gates, transforms into Silver (Data Vault), and finally serves Gold (semantic/feature store) layers for AI/BI consumption.

- **Enterprise Data Modelling (Data Vault 2.0):** Deploy hubs, links, and satellites to decouple raw ingestion from business rules. This natively supports ALCOA+ by preserving source-original data while enabling agile, auditable transformations for DS feature stores.

- **Lakehouse Infrastructure (Databricks & Delta Lake):** Implement a Medallion architecture (Bronze → Silver → Gold) with Delta Lake's ACID transactions, time travel, and schema enforcement. Optimized for both operational BI and distributed ML training workloads.
- **Active Data Catalogs & Ontology Mapping (FAIR Engine):** Integrate enterprise catalogs (e.g., Alation, Collibra, Unity Catalog) with mandatory metadata tagging, lineage tracking, and mapping to unified ontologies (SNOMED CT, HL7 FHIR, CDISC). Unregistered or semantically misaligned datasets are rejected at ingestion.
- **Immutable Audit Trails & Privacy-by-Design Access:** Enforce WORM storage (e.g., AWS S3 Object Lock) for raw clinical/R&D data. Transition to Attribute-Based Access Control (ABAC) that evaluates user role, data sensitivity (PHI/PII tags), and active compliance training status, satisfying *Accessible* without compromising privacy.

Phase 2: Automated QA & DataOps CI/CD

Quality assurance shifts from retrospective manual reviews to proactive, code-driven validation embedded directly into ETL/ELT pipelines.

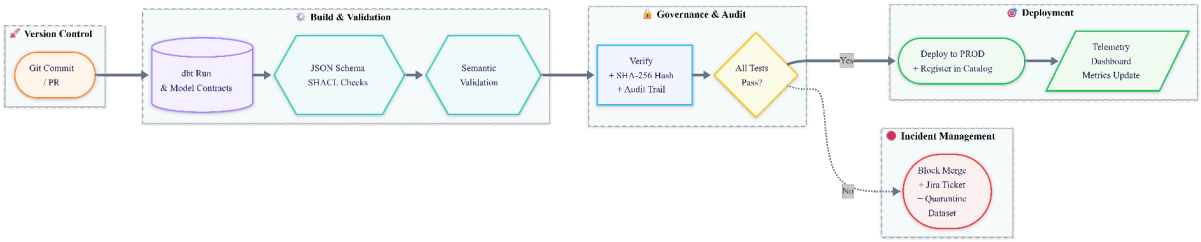

Figure 2: **Automated CI/CD & Validation Pipeline:** Code commits trigger dbt runs, followed by sequential Schema/Ontology checks and Hash/Audit verification. Only validated data is registered in the Catalog and deployed to production; failures are quarantined.

- **dbt-Driven Pipeline Observability:** Standardize transformations using dbt with incremental loading, strict model contracts, and auto-generated documentation. Integrate with CI/CD runners to enforce peer review, testing, and version-controlled deployments.
- **FAIR Validation Testing:**
  - *Metadata Completeness:* Automated JSON Schema/SHACL checks quarantine datasets missing mandatory tags (e.g., study ID, consent status, data owner).
  - *Semantic Integrity:* Ontology mapping scripts flag unapproved vocabularies (e.g., "heart_attack" vs. "myocardial_infarction" SNOMED ID) before Silver-layer promotion.
  - *Access Protocol Simulation:* Automated penetration tests deploy "dummy" AI agents to verify row-level security and ABAC policies block unauthorized PHI/RAG queries.
- **ALCOA+ Validation Testing:**
  - *Audit Trail Verification:* Synthetic data injections validate that CRUD operations immutably record user IDs, server timestamps, and change reasons.
  - *Cryptographic Hashing:* SHA-256 checksums enforced across all pipeline hops guarantee *Accurate* and *Complete* data fidelity from eCRF/source to Gold layer.
  - *Endurance & Retention Drills:* Scheduled disaster recovery simulations and automated data destruction tests validate regulatory retention/deletion policies.

Phase 3: Continuous Monitoring, AI Lineage, Human Oversight

Automation handles pipeline execution; human governance prevents semantic drift, aligns data products with business outcomes, and ensures inspection readiness.

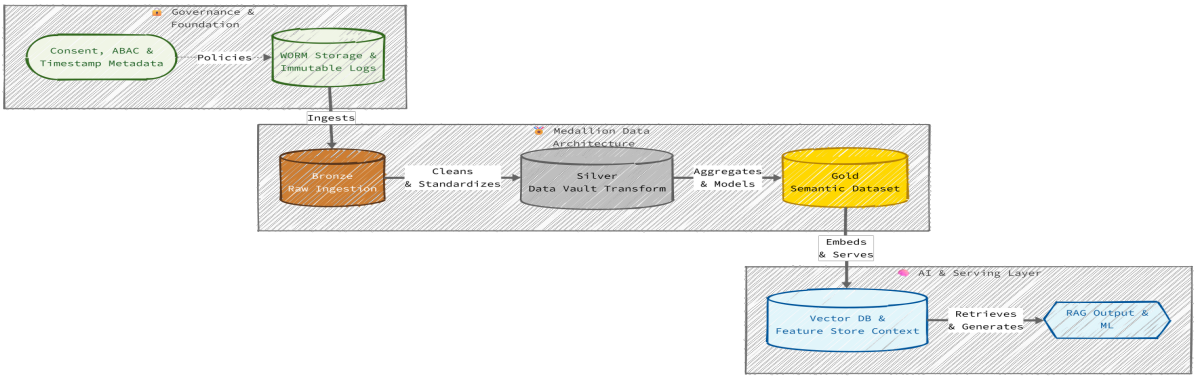

Figure 3: **AI/RAG Lineage & Audit Traceability:** Tracing model outputs or RAG insights backward through Vector DB contexts, Gold datasets, and Silver transformations to the original Bronze raw data and immutable audit logs, proving end-to-end compliance.

- **Dual-Scoring Compliance Dashboards:** Real-time telemetry scores every dataset on FAIR (metadata richness, interoperability, reuse potential) and ALCOA+ (provenance, accuracy, auditability). Alerts route directly to Slack/Teams and Jira for rapid remediation.
- **Domain Data Stewards & Cross-Functional Loops:** Appoint scientific, technical, and business stewards per data domain. Agile feedback cycles ensure ETL logic, feature definitions, and RAG context windows directly support active DS/ML use cases.
- **Internal Mock Inspections & AI Traceability:** Routine simulated FDA/GDPR/CDISC audits trace complex outputs (e.g., RAG-generated clinical insights, ML model predictions) backward through Delta Lake time travel, dbt lineage, and catalog metadata to raw source data, proving end-to-end compliance and reproducibility.

Implementation Roadmap & Success Metrics

Phase	Timeline	Key Deliverables & KPIs
Phase 1	Weeks 1–4	Data Vault + Medallion scaffold, ABAC rollout, Catalog integration, WORM storage. *KPI: 100% raw data immutability, catalog coverage >90%*
Phase 2	Weeks 5–8	dbt CI/CD pipelines, FAIR/ALCOA+ automated tests, SHA-256 hashing, access simulations. *KPI: <2% pipeline failure rate, QA automation >85%*
Phase 3	Weeks 9–12	Telemetry dashboards, steward onboarding, mock audits, RAG/ML lineage tracing. *KPI: Audit traceability <15 mins, FAIR/ALCOA+ scores >4.5/5*

Conclusion

This blueprint transforms data governance from a compliance overhead into a strategic enabler for Data Science, AI, and R&D. By embedding FAIR utility and ALCOA+ integrity directly into the pipeline fabric, organizations gain faster model iteration, reproducible research, audit-ready AI outputs, and scalable DataOps. The proposed architecture is cloud-agnostic, tool-flexible, and designed to evolve alongside emerging ML/RAG workloads while maintaining strict regulatory defensibility.