phuse·share
connect·share·advance

Single Day Event

Elevating the Role of Data Scientists Through Digital Metadata and AI Integration

Avinash Patnaik,
Hexaware Technologies

Are you gambling with Submissions?

Feed for Thought

Programmers stuck in the loop of repetitive scrubbing

Relying on heavy human review and brittle rule upkeep, limiting scalability and consistency

What if a single mapping error could stop your submission?

Inconsistent mappings, and prolonged programming and validation cycles due to manual curation of documents

Are you willing to trade speed for hallucinations and lose reproducibility in regulated filings?

LLM-based tools bring risks like unpredictable outputs, hallucinations, and gaps in reproducibility and traceability.

How much hidden time and regulatory risk can you tolerate in manual curation

Slow decision-making and study timelines, increased regulatory risks, makes them engage in low-value repetitive work

Navigating the Roadblocks

MANUAL BURDEN

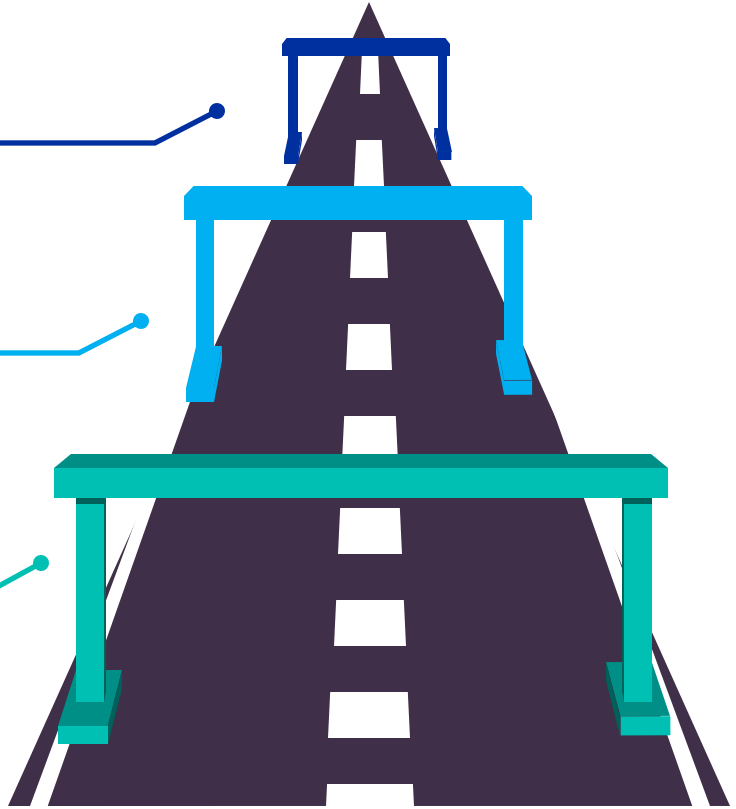
Labor-intensive, repetitive tasks, prolonged cycles, and manual curation lead to slow decision-making and extended timelines.

LACK OF STANDARDIZATION

Inconsistent data mappings, brittle rules, limited scalability, and hidden regulatory risk threaten submission integrity.

ADVANCED AI DILEMMA

Unpredictable outputs, hallucinations, and loss of reproducibility & traceability.



Solution Overview (1/2)

1

Intelligent Ingestion (Top Layer)

- Converts unstructured documents (protocols, CRFs, SAPs) into structured, actionable metadata **automatically**
- Natural language processing extracts key information without manual data entry
- Reduces protocol review time from days to hours

2

Trusted Metadata Core (Center)

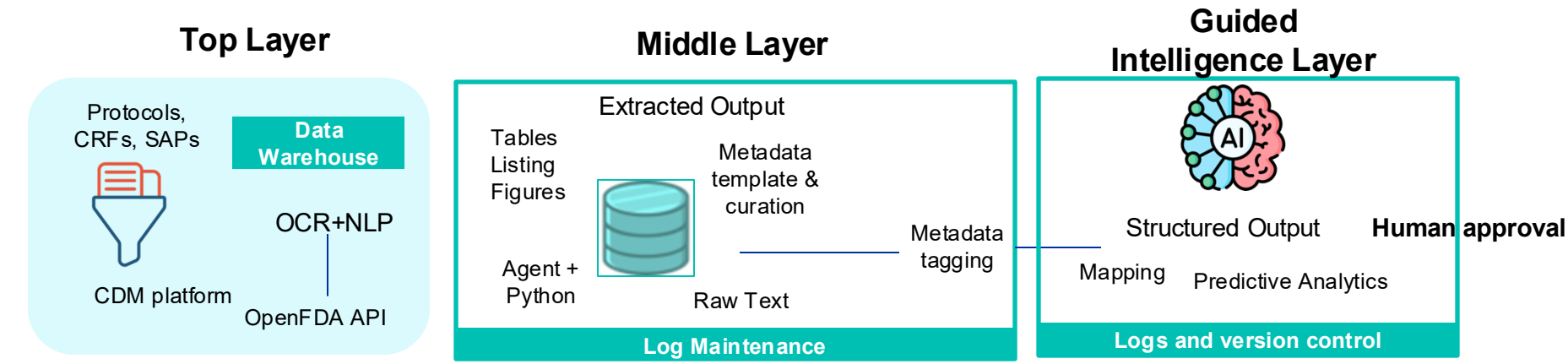
- Single source of truth with complete version control and lineage tracking
- All data transformations trace back to source with timestamps and approvals
- Instant audit readiness and regulatory confidence

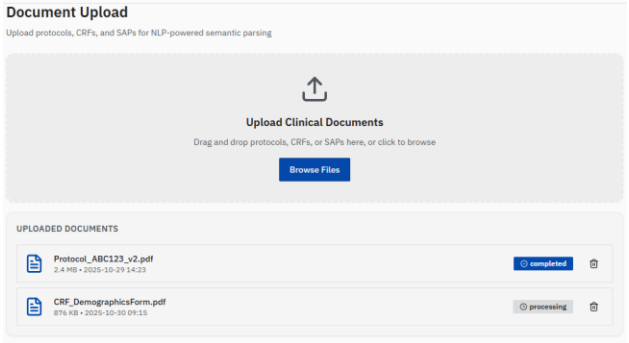
3

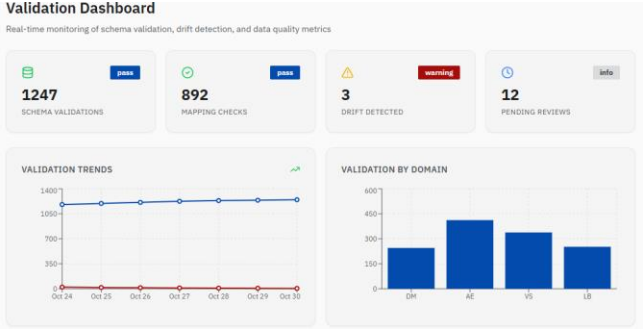
Guided Intelligence Layer

- AI proposes mappings and generates code, but **humans approve critical decisions**
- LLM agents suggest transformations → Experts review → System executes with full logging
- 60% faster than manual while maintaining 100% compliance control

Solution Overview (2/2)







Functionality Overview

PDF parsing and metadata anchoring

- Reviewer approve/edit updates metadata and triggers code/regeneration.
- Edits are captured as training artifacts to improve parsing and mapping accuracy. Support group BCP Plans

Continuous validation checks

- Rule-based schema checks plus statistical tests (ranges, distributions, missingness).
- Auto-create prioritized remediation tickets on rule or drift failures.

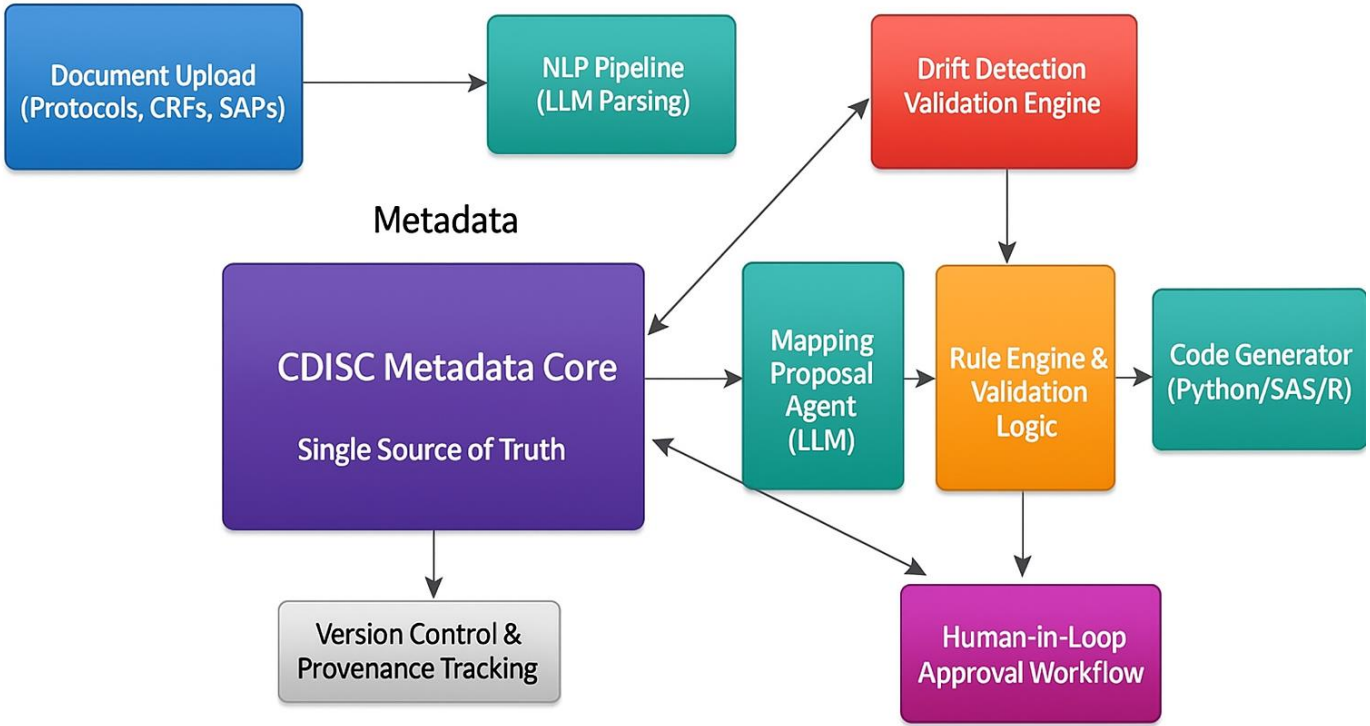
Confidence and provenance tagging

- Attach confidence score, model version and prompt ID to each extraction.
- Persist direct links to original PDF snippet for audit.

Dashboard UX for triage and audit export

- High-level KPIs with drill-down to variable-level diffs and extraction heatmaps.
- One-click export of audit package (prompts, model versions, code diffs, PDF snippets).

Functional Diagram



Regulatory Compliance

FDA/EMA Readiness

- Complete 21 CFR Part 11 audit trail
- Full data lineage and provenance
- Human oversight at critical decision points
- Model version tracking for reproducibility

Risk Mitigation

- AI proposes, humans approve
- Confidence thresholds trigger review
- Rule-based validators catch edge cases
- Rollback capability with version control

Mitigating AI risks

Risk	Our Safeguard	Business Outcome
AI Hallucinations	Human-in-the-loop approval for all mappings	Only validated transformations enter production
Unpredictable Output	Confidence scores + rule validation	Low-confidence proposals flagged for expert review
No Audit Trail	Complete provenance: Source → SDTM → ADaM → TFL	FDA/EMA inspection ready on day one
Lack of Reproducibility	Model version + prompt logging for every decision	Exact recreation of analysis possible years later

Benefits

Time Reduction

- Validation cut from days to hours (70–80% reduction).
- Protocol-to-mapping reduced to **30-40%.**
- Code generation automated via template-driven pipelines

Quality Enhancements

- Real-time schema-drift detection and continuous error checks.
- Unified mapping standards and automated reconciliation for consistent outputs.
- Continuous integration enabled validations reduce rework.

Productivity

- Manual programming reduced up to 60%, reallocating teams to higher-value work.
- Processes free up to **35–50%** capacity for strategic analytics.
- Data scientists focus on regulatory storytelling and advanced insights.

Regulatory Confidence

- Complete end-to-end data lineage is maintained.
- Model version tracking is implemented to ensure reproducibility.
- Human approval checkpoints are enforced for all critical decisions.

Scalability

- Reusable assets and modular templates scale trial capacity without proportional headcount.
- Faster amendments and consistent, validated outputs.

THANK YOU

The Global Healthcare Data Science Community

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

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