PharmaSUG 2025 - Paper SI-225

Stop Making More Uncontrolled Copies of Your CDISC Data Sets

Anthony Chow, CDISC;
Robert Garelick, Independent

ABSTRACT

The pharma industry faces significant challenges with data traceability and fidelity due to the pervasive creation of physical copies of data and derivatives that are uncontrolled. This practice diminishes data quality, increases regulatory risks, and creates inefficiencies in compliance and governance processes. Our approach proposes a paradigm shift: treating all datasets as traceable views derived from a single source of truth. By leveraging modern technologies such as data lakes, event streams, and unified analytics platforms, organizations can eliminate unnecessary redundancies, enhance traceability, and improve data fidelity. We highlight how tools like dbt and Apache Spark enable end-to-end logging of transformation logic, ensuring compliance and reproducibility.

This framework offers key benefits, including eliminating data silos, enabling scalable automation, and fostering collaboration through a unified data platform. Additionally, it provides futureproofing by integrating real-time and historical insights while reducing labor-intensive manual processes. Our presentation will showcase practical implementations of these strategies, demonstrating how they streamline data workflows, enhance regulatory compliance, and ultimately empower organizations to make data-driven decisions efficiently.

BACKGROUND

This industry faces a widespread and critical challenge: the creation of many uncontrolled copies of data. Much like a game of telephone in a digital context, this practice results in significant losses in traceability, fidelity, and efficiency, making it an unsustainable approach. It is essentially a breakdown in data provenance practices, which makes understanding the lineage of the data (where it came from, how it was transformed, and why) difficult. This problem is unfortunately widespread, especially in the biopharma industry where datasets pass through many layers of transformation and data owners.

TYPICAL WORKFLOWS

In the context of regulatory data submissions for marketing approval, it is essential to integrate raw data from various sources related to a research study—such as clinical trial data from eCRFs, electronic data transfers from central labs, and patient diaries from electronic patient-reported outcomes (ePRO). These raw data are used to create physical data tabulation datasets in the CDISC Study Data Tabulation Model (SDTM) format.

Many sponsor pharma organizations outsource this task to CROs, transferring the raw data to them via FTP. Once the CROs complete the task, they transfer the processed data sets back to the sponsor. This process may be repeated multiple times, especially when the CRO lacks an online data review platform, leading to numerous data set transfers.

This scenario involves frequent handoffs and often lacks proper data extraction and receipt documentation. Critical details, such as the origin of a data element, how it was created, who modified it, and any transformations it underwent, are often missing. Maintaining a comprehensive record of the data's history is vital to ensure its authenticity and reliability.

PAIN POINTS

1. Loss of original source information: People don't know where the data originated.
2. ETL processes risking data fidelity: Transformations introduce the risk of errors, inconsistencies, or loss of critical details.
3. Labor-intensive documentation: Manually maintained "mapping specification" or "data specification" documents are prone to error, hard to update, and unsustainable.

4. Compliance risks: Poor traceability can lead to difficulties in meeting regulatory requirements and data governance standards.

A DIFFERENT MINDSET

The industry needs to embrace a paradigm shift to address the root causes of data fidelity and traceability issues. By moving away from the traditional "data set from data set" model (which leads to disjointed, non-linear processes), we want to introduce a systematic and scalable framework that prioritizes traceability and data quality.

"VIEW" FROM SOURCE WITH TRACEABILITY

- This shifts the focus to treating all data sets as derived views of a single source of truth.
- It eliminates unnecessary intermediate data sets (`x -> x' -> x''`), reducing redundancy and the chances of errors introduced at each step.
- Built-in traceability (via transformation logic logging) ensures that every change can be traced back to the source.

LEVERAGING MODERN TECHNOLOGIES

- Data lakes: Store raw and processed data centrally, enabling you to always reference original data without unnecessary duplication.
- Event streams (e.g., Kafka): Allow for real-time tracking of data transformations, enabling traceability and provenance in streaming data workflows.
- Scalable data stores: Develop an efficient strategy for data storage and retrieval. Avro, for example, excels in data serialization and streaming use cases, while Parquet is optimized for analytics and storage efficiency in read-intensive workloads.
- Unified data analytics platforms (e.g., Databricks, Snowflake): Enable integrated workflows that connect ingestion, transformation, and analytics with traceability baked in.

LOGGING FULL TRANSFORMATION AND COMPUTATIONAL LOGIC

- The {sdtm.oak} package [1] is an open-source R tool designed to assist the pharmaceutical programming community in developing SDTM data sets. It provides a machine-executable framework for modular programming and aims to automate the conversion of raw clinical data into SDTM format, which is essential for regulatory submissions.
- Leveraging metadata to drive machine-executable data transformation enables the automatic rendering of mapping or data specification documents, significantly reducing manual intervention, improving efficiency, and eliminating error-prone processes
- Systematically logs every data transformation step, storing detailed records as metadata alongside the resulting data set to ensure full traceability.
- Tools like dbt (data build tool) and Apache Spark already allow this kind of lineage tracking, which could be expanded across all datasets.

KEY BENEFITS

This approach offers several key benefits that enhance data management and analysis. First, it eliminates data silos by centralizing processes around a single source of truth and utilizing "views," thereby preventing the proliferation of unnecessary data copies that often lead to misalignment. Additionally, it enhances traceability through end-to-end logging, simplifying regulatory audits, error investigations, and ensuring reproducibility. Improved data fidelity is another advantage, as always referring back to the original source minimizes the risk of propagating errors through derivative datasets.

This approach also promotes scalability and automation, leveraging modern tools and platforms designed for large-scale, distributed data processing, which significantly reduces manual effort and labor-intensive documentation. Finally, it provides future-proofing by enabling both real-time insights, such as live dashboards, and historical views through event streams and unified platforms. This

eliminates the need for repeatedly creating static datasets, ensuring a streamlined and efficient process.

A unified platform enables holistic data governance. Technologies such as Microsoft Fabric change how organizations manage and analyze data throughout its entire lifecycle. With end-to-end visibility, organizations gain full data transparency, reducing risks like siloed operations, lost context, and duplicated efforts.

LEVERAGING DATA ARCHITECTURES

Effective data management and governance hinge significantly on the choice of data architecture. A robust architecture serves as a blueprint that facilitates efficient data operations, enhancing traceability, fidelity, and compliance. Properly structured, it can minimize redundancies, streamline data flows, and ensure that data provenance and lineage remain transparent and intact.

In this context, two contemporary architectural blueprints merit attention due to their prevalence in the data management space and distinct approaches: the Medallion Architecture and the Data Product Architecture.

INTRODUCING ARCHITECTURES

When considering data architecture, there are many blueprints and methodologies to choose from, each with its own strengths and trade-offs. There is no single right or wrong approach. Instead, architects need to make choices that fit their organization's goals, technical needs, and level of data maturity. Among the many frameworks that data architects discuss and sometimes debate, two stand out as especially influential: Medallion Architecture and Data Product Architecture. Below, we take a closer look at these two approaches, exploring their core principles, how they are implemented in practice, and the different ways they manage and deliver data.

1. Medallion Architecture

The Medallion Architecture (Figure 1) organizes data into a structured three-tiered model:

- **Bronze (Raw)**: Stores raw, ingested data without any transformations.
- **Silver (Refined)**: Cleanses and conforms data through standard transformations and basic quality checks.
- **Gold (Curated)**: Hosts highly curated datasets ready for business analytics and reporting.

This structure is intuitive and has historically helped manage complexities by compartmentalizing data processing tasks into clear, logical stages.

In pharma contexts (Figure 2), clinical and imaging data (e.g., CRF, eDT, IxRS, DICOM, Nifti) enter through a landing zone and undergo a structured transformation process. The data progress through three layers: Bronze (raw data), Silver (clinical and trial management operational data store), and Gold (specialized for safety, pharmacovigilance, and regulatory operations). Ultimately, this refined data supports analytics platforms such as Spotfire, JMP Clinical, and Alteryx, facilitating data-driven pharmaceutical insights.

Figure 3 presents a reference implementation of Medallion Architecture. The process starts with different data sources like ODM, FHIR, Amazon S3, FTP, and streaming data securely sending information through a landing zone. This landing zone is equipped with virtual machines, firewalls, role-based access control, and identity and access management to ensure security. From there, the data moves into a SaaS layer that includes storage, ETL pipelines, data lakes, and either data warehouses or lakehouses. Once processed, the data fuels self-service analytics, enabling business intelligence, informatics, AI and machine learning, as well as GraphQL applications.

2. Data Product Architecture

Data Product Architecture (Figure 4) treats data entities as individual "products" aligned explicitly with consumer needs and business context. This architecture focuses on:

- **Context-Driven Data Management**: Aligns data structures around specific business use cases.
- **Decentralized Ownership**: Encourages distributed data management responsibilities.

- **Proactive Data Quality**: Embeds rigorous quality and traceability checks at the source, aligned closely with end-use objectives.

Comparing the Two Architectures

Architectural Element	Medallion Architecture	Data Product Architecture
Core Philosophy	Transformation-stage based (Bronze-Silver-Gold).	Business-context driven, use-case-specific data products.
Data Flow Direction	Pull-based (consumers must actively retrieve and transform).	Push-based (business context actively informs upstream data preparation).
Traceability	Traceability fragmented across layers; each transformation introduces complexity.	Built-in, continuous traceability from the data source itself; less complexity in lineage tracking.
Data Fidelity	Risk of compounding errors due to repeated transformations.	High fidelity preserved by minimizing unnecessary transformations.
Operational Efficiency	Multiple redundant transformations; higher storage/compute costs.	Lean, purpose-specific processing; reduces redundancy and associated costs.
Consumption Flexibility	Typically batch-oriented; limited consumption options without additional processing.	Offers flexible consumption modes (batch, streaming, APIs) tailored to business needs.

Pros and Cons of Each Architecture

Comparison of Pros: Medallion vs. Data Product Architecture

The following table outlines the advantages of the Medallion and Data Product architectures, highlighting their strengths in data management, flexibility, and governance:

Aspect	Medallion Architecture	Data Product Architecture
Structure & Intuitiveness	Straightforward layering model (Bronze, Silver, Gold)	Proactively integrates business context from earliest stages.
Responsibilities & Governance	Clear division of roles between data engineering teams	Strong traceability & transparency, ideal for regulated environments.
Industry Recognition	Established method, widely recognized across industries	Highly flexible, accommodating diverse consumption patterns (batch, streaming, APIs).
Data Quality & Fidelity	–	Enhanced data fidelity through minimal, targeted transformations.
Scalability & Flexibility	–	Scalable and adaptable; promotes decentralization and empowers domain experts.

Comparison of Cons: Medallion vs. Data Product Architecture

The following table summarizes the disadvantages and potential challenges of implementing each architecture:

Aspect	Medallion Architecture	Data Product Architecture
Data Errors & Transformations	Can propagate and compound errors due to repeated transformations	–
Operational Overhead	High operational overhead and resource usage due to redundant copies and transformations	Potentially higher upfront planning and coordination required
Business Context	Lacks early-stage business context, causing inefficiencies downstream	Requires mature data governance and robust infrastructure to manage decentralization
Flexibility & Adaptation	Limited flexibility in diverse use cases without added complexity	May involve cultural shifts and adjustments in team roles and responsibilities

By explicitly understanding and contrasting these two approaches, organizations can make informed decisions tailored to their specific needs. The chosen architecture becomes a strategic enabler—enhancing data traceability, ensuring data fidelity, and improving the overall quality of data-driven decisions.

ALIGNMENT WITH CDISC 360I INITIATIVE

The approach presented in this paper aligns directly with the objectives of the CDISC 360i initiative [2], particularly regarding enhancing end-to-end data traceability and governance through modern, metadata-driven clinical data standards. Like CDISC 360i, our framework advocates for a paradigm shift from redundant data copying to deriving data sets as views from a centralized source of truth, leveraging technologies that enable clear, auditable transformation logic.

Our use of contemporary data architectures such as the Medallion and Data Product architectures directly parallels CDISC 360i's goals of creating structured, transparent, and reusable data pipelines. Both methodologies emphasize metadata-centric automation, reducing manual interventions, improving fidelity, and simplifying regulatory compliance. This alignment underscores our shared vision of streamlined, robust, and scalable clinical data workflows.

By adopting approaches consistent with CDISC 360i, our proposed framework positions organizations to not only enhance compliance and operational efficiency but also to fully embrace futureproofed, interconnected data ecosystems essential for clinical research excellence.

REFERENCES

1. Ganapathy R, Forys A, Manukyan E, et al. sdtm.oak: SDTM Data Transformation Engine (Version 0.1.1) [R package]. CRAN. 2024. Available at https://cran.r-project.org/package=sdtm.oak.
2. CDISC 360i. *CDISC*. Acccessed March 28, 2025. Available at https://www.cdisc.org/cdisc-360i.

ACKNOWLEDGMENTS

We sincerely thank Sam Hume, DSc, for his thoughtful feedback and encouragement. His insights on data product architecture, data mesh, and the potential for broader industry collaboration were especially valuable and deeply appreciated.

RECOMMENDED READING

Pereyra H, Sloan J. “Challenges of today’s monolithic data architecture.” IBM. Accessed March 28, 2025. Available at https://www.ibm.com/think/insights/monolithic-data-architecture-challenges.

Strengholt P. 2023. Data Management at Scale. 2nd ed. ISBN 1098138864: O'Reilly Media, Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Anthoy Chow
CDISC
achow@cdisc.org

Robert Garelick
Independent
robert.garelick@gmail.com

APPENDIX A: DIAGRAMS

Figure 1: A general view of the medallion architecture

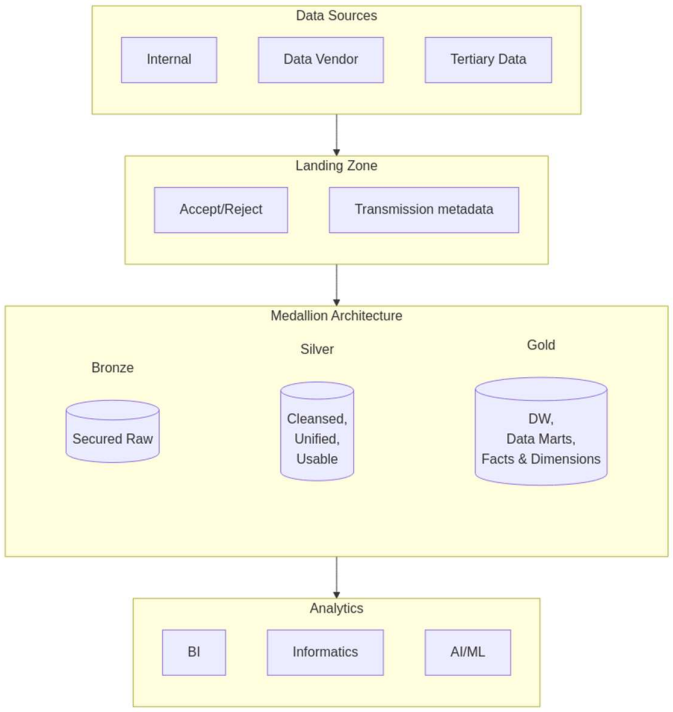

Figure 2: Same general view of medallion architecture contextualized with pharma characteristics

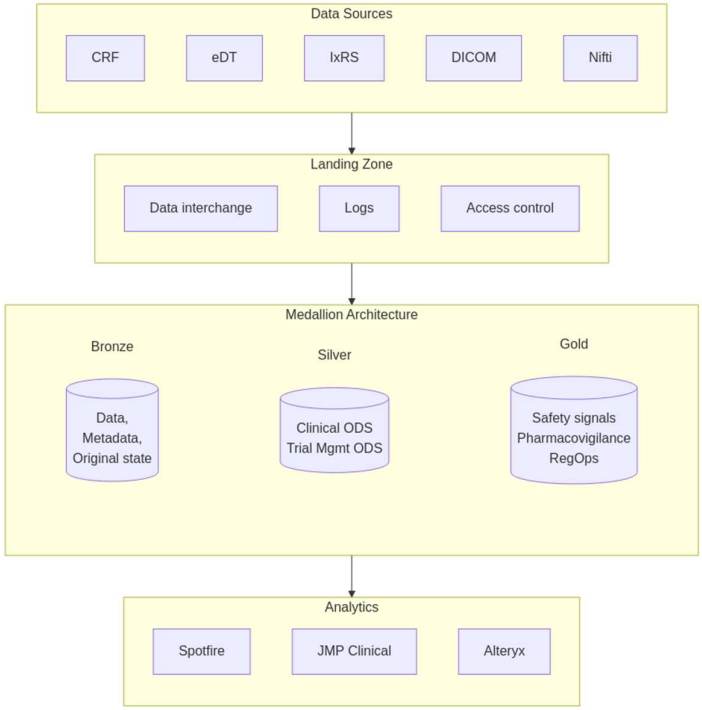

Figure 3: A reference architecture for medallion architecture

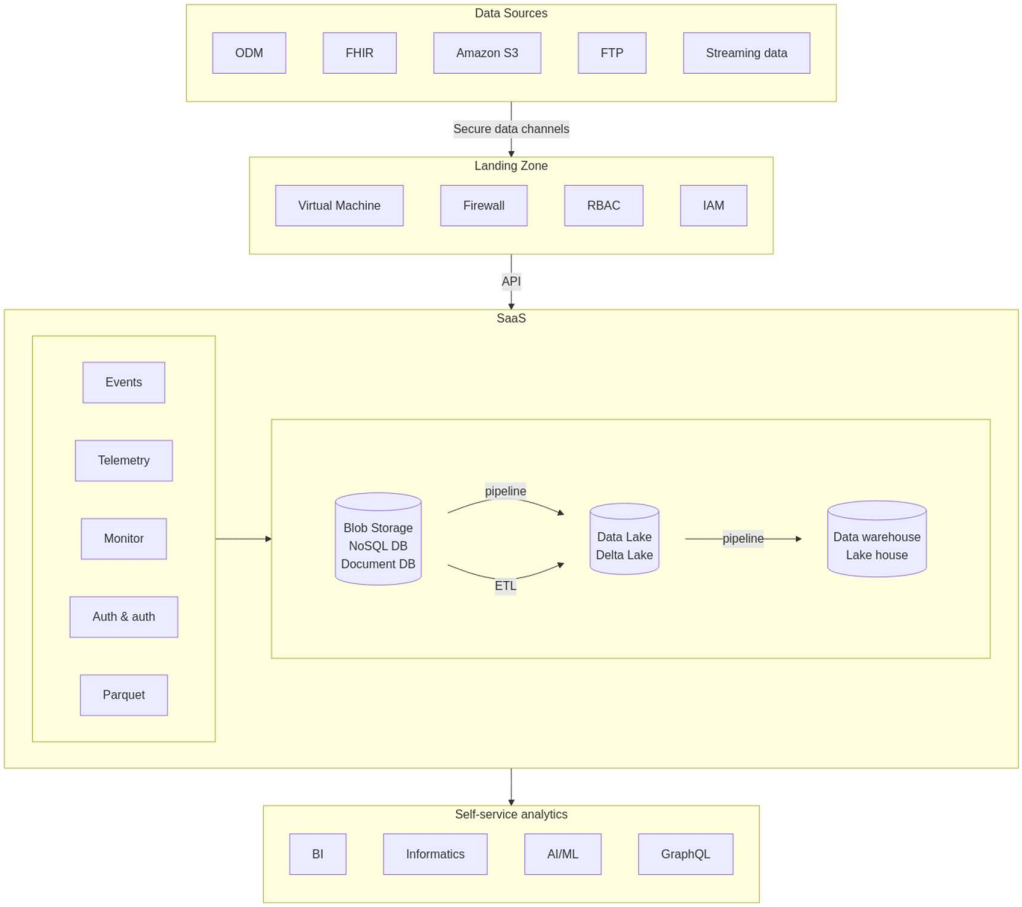

Figure 4: A general diagram of the data provider architecture