Paper DH09

The Olympics of Data Ingestions

Cathryn Herbst, Formation Bio, Cambridge, MA, United States

ABSTRACT

It's the shared goal of many in industry to collect and analyze clinical data faster. However, many data vendors require manual retrieval and reconciliation before analysis can begin. To both accelerate and de-risk clinical data analysis, our Data Management and Analytics (DMA) and Data Platform Engineering teams have worked together to modernize our clinical data pipeline, harnessing the power of data engineering to entirely automate retrieving and validating clinical data. We do so by selecting vendors with flexible delivery endpoints, executing data extraction on a scheduled basis, and embedding data quality checks. With near-real time access to disparate clinical data sources, medical monitors can more quickly identify and rectify patient safety issues and protocol deviations, while data analysts and stats programmers can sooner interpret their topline clinical results. This presentation will give you the details needed to revolutionize your own data pipelines and enhance your analytical workflows.

1. INTRODUCTION - WELCOME TO THE OLYMPICS OF DATA INGESTIONS

Bringing a new drug to market is a race against time. Every pharmaceutical company is competing to develop safe and effective treatments faster, but a major hurdle stands in the way - data. Drug development is like a decathlon, requiring excellence across multiple disciplines, from clinical trials and regulatory submissions to manufacturing and commercialization. Yet, drug development teams often find themselves bogged down by fragmented, inconsistent, and slow-moving data, creating obstacles at every turn.

At Formation Bio, we believe that overcoming these data challenges requires more than just better tools, it demands a fundamental shift in how clinical development leverages technology. By integrating AI, automation, and scalable data infrastructure, we aim to transform the way data is structured, processed, and applied, ensuring that insights drive faster and more efficient decision-making. Central to this approach is the ability to turn raw, messy data into a structured, decision-ready asset.

This paper outlines a data decathlon training regimen for transforming raw, messy data (Bronze) into standardized, structured data (Silver) and ultimately into fully governed, decision-ready data (Gold). By adopting modern data platform practices such as the medallion framework, scalable infrastructure, and a data-as-a-product mindset, organizations can make data ingestion a repeatable, automated, and scalable process. The goal is to spend less time wrestling with data and more time analyzing it, accelerating decision-making and enabling better patient outcomes.

2. THE PRESEASON STRUGGLES - WHY PHARMA TEAMS STRUGGLE TO MAKE THE MOST OF THEIR DATA

Just like an athlete needs the right training regimen to avoid burnout before the big race, pharma companies need strong data ingestion strategies to avoid data burnout in the data decathlon: siloed data, manual and slow data processing, slow decision making, and manual compliance reviews.

2.1 - TOO MANY EVENTS, NOT ENOUGH COORDINATION: DATA IS FRAGMENTED ACROSS SILOS

Drug development generates massive amounts of data, but too often, this data is trapped in silos, making coordination nearly impossible. Preclinical teams store toxicology results in unstructured PDFs. Clinical teams pull patient data from multiple electronic data capture (EDC) systems in different formats. Regulatory teams manually compile submission data from various sources. Business development teams create landscape maps. Without a centralized approach, each team operates in isolation, leading to inefficiencies, delays, and lost insights.

2.2 - NO REAL-TIME PERFORMANCE TRACKING: DATA PROCESSING IS MANUAL AND SLOW

When clinical and commercial teams lack real-time performance tracking, they are forced to operate reactively, leading to delayed decision-making and missed opportunities. Safety monitoring teams may not detect unexpected adverse event trends until weeks later because they manually pull data from disparate EDC and safety databases. Commercial teams struggle to assess launch readiness because market access data is not integrated with clinical and supply chain forecasts. Clinical trial teams experience costly delays as they wait days for patient enrollment updates, preventing timely recruitment adjustments.

2.3 - LACK OF HISTORICAL CONTEXT: REINVENTING THE WHEEL FOR EVERY STUDY

Clinical trials often repeat past mistakes because critical insights from previous studies are not easily accessible. Patient recruitment bottlenecks, unexpected safety signals, and protocol deviations remain buried in static documents or siloed databases. Chemistry, Manufacturing, and Controls teams may struggle to link prior formulation challenges to current development work, leading to redundant testing and delays. Without structured historical benchmarks, teams designing a Phase 3 trial may lack key data from similar programs, making it harder to optimize enrollment strategies.

2.4 - DATA GOVERNANCE? MORE LIKE DATA GYMNASTICS

Regulatory submissions require strict data traceability, yet many organizations still rely on manual processes to track changes across datasets. When submitting a New Drug Application, sponsors must manually reconcile versions, increasing the risk of compliance issues and regulatory delays. Without a structured system, ensuring that datasets align with SDTM and CDISC standards becomes a time-consuming, error-prone task.

3. TRAINING FOR GOLD - MODERNIZING DATA INFRASTRUCTURE FOR DRUG DEVELOPMENT

Instead of struggling without a training plan and facing the data burnout problems, pharma companies must create their data decathlon training plan. While many teams treat data ingestion as an IT problem rather than a strategic function, viewing it as a core business capability enables faster decision-making, reduces inefficiencies, and prevents costly mistakes. There are three key pieces to the data training plan

1. Thinking of data as a product
2. Structuring data using the medallion framework
3. Adopting modern infrastructure

3.1 - THINKING OF DATA AS A PRODUCT: THE KEY TO A WINNING STRATEGY

Before implementing any structural frameworks, organizations must shift their perspective on data. Data is not just a byproduct of research but a strategic asset that requires careful management, clear ownership, and thoughtful design to serve multiple stakeholders effectively. A data-as-a-product approach ensures that data is not only collected but structured, governed, and optimized for long-term use across clinical, regulatory, and commercial teams.

This mindset brings several key advantages. Clear ownership and accountability of datasets and data concepts prevent duplication and inefficiencies, reducing the burden on technical teams while empowering domain experts to manage and refine data. Well-structured datasets serve multiple use cases, eliminating redundant processing and ensuring smooth integration across different phases of drug development. A productized approach also promotes continuous improvement, where data is iteratively refined for AI-driven insights and future applications. Standardized, reusable datasets enhance scalability, making it easier to accelerate regulatory submissions, support new drug approvals, and leverage cross-program learnings. High-quality, trusted data also speeds up decision-making by reducing the time spent on cleaning and reconciling datasets. By treating data as a product, organizations can unlock its full potential, ensuring it drives faster, smarter decision-making across the entire drug development lifecycle.

Example 1: Protocol Deviation Data Used Across Multiple Teams

Without a data-as-a-product mindset, different teams handle protocol deviation data in silos. Clinical operations teams track deviations to ensure site compliance, regulatory teams manually extract insights for submission reports,

and quality assurance teams conduct separate audits, leading to redundant work and inconsistencies. Each team collects and processes the same data differently, causing delays and inefficiencies.

With a data-as-a-product approach, protocol deviation data is ingested into a unified data platform where it is structured, validated, and made available for multiple use cases. Clinical teams can monitor deviations in real time, regulatory teams can pull submission-ready reports instantly, and quality teams can perform audits without redundant data requests. This reduces inefficiencies, enhances compliance, and ensures alignment across teams while cutting data preparation time

Example 2: Iterative Updates to Real-World Evidence for Faster Decisions

Without a data-as-a-product mindset, real-world evidence (RWE) datasets (such as patient outcomes from prior clinical trials or post-market surveillance data) are only analyzed once fully complete, which can take months or years. Regulatory and commercial teams hesitate to act on partial data, delaying go/no-go decisions for new indications or market expansions.

With a data-as-a-product approach, RWE datasets are continuously updated and refined, allowing teams to make early, informed decisions. AI-driven data validation highlights reliable insights even before a dataset is fully complete, enabling regulatory teams to preemptively draft submission strategies and commercial teams to adjust market positioning based on emerging trends. Instead of waiting for a perfect dataset, iterative updates allow faster decisions, reducing the time needed for regulatory filings and market entry by months.

3.2 - THE MEDALLION FRAMEWORK: STRUCTURING DATA FOR SUCCESS

Just like athletes use their training plan to compete for Bronze, Silver, and ultimately Gold, data must go through its own transformation to reach peak performance. Raw, unstructured data starts at the Bronze level, where it is ingested and stored. It advances to Silver as it is cleaned, standardized, and structured for analysis. Only when it reaches Gold does it become a high-quality, AI-ready, decision-grade asset that can drive real impact across clinical, regulatory, and commercial teams.

Bronze Medal - Taming the Chaos of Raw Data

Bronze data is the starting point of any data pipeline, often messy, inconsistent, and scattered across multiple formats and storage locations. It includes everything from CSVs, Excel files, and JSON exports to PDFs, SFTP dumps, and manual extracts. Data at this stage lacks validation, structure, and standardization, making it difficult to use effectively for analysis or decision-making.

To transform raw data into something usable, organizations should first centralize it using API-based ingestion and data lakes, ensuring all data is collected in a single storage layer. Automating extraction from unstructured sources like preclinical lab notebooks and PDF reports can significantly reduce manual effort and improve efficiency. Tagging metadata such as timestamps, source systems, and data owners enhances traceability and governance, laying the groundwork for higher-quality data downstream.

✅ *Example Impact:* For regulatory documents from the FDA, AI tools can be used to extract data tables from a PDF turning the data into structured data. This allows other teams to make use of the data immediately upon ingestion instead of waiting to receive value from the data until the data reaches the gold layer. Teams can also identify any quality issues earlier in the process instead of waiting until the final data product to flag any data quality issues.

Silver Medal - Standardizing & Structuring Data for Analysis

Reaching the Silver stage means your data is starting to take shape, but it is not fully refined. At this level, data is structured and some transformations have been applied, but it may still contain inconsistencies or gaps. Business rules are not yet fully implemented, making the data useful for exploratory analysis but not yet ready for regulatory submission or high-stakes decision-making.

To progress further, organizations need to implement robust transformation and validation processes. Tools like dbt can standardize data formats, while automated quality control checks, such as duplicate detection and outlier flagging, help catch errors early. Enforcing data contracts ensures schema consistency across different datasets, preventing downstream mismatches and inconsistencies.

AI-driven data modeling can automate the initial structuring of datasets, using historical trends and domain-specific rules to generate an optimized schema. Instead of requiring data engineers to manually define tables and relationships, AI provides a first-pass model that human experts can refine. This approach accelerates data structuring, ensures standardization across datasets, and reduces the time spent on manual data preparation.

✅ *Example Impact:* In the Silver layer for preclinical data, toxicology results from various lab formats can be converted into standardized, queryable tables, allowing for easy cross-study comparisons. On the regulatory side, ensuring datasets comply with SDTM and CDISC standards helps streamline submission readiness. By leveraging AI to auto-validate and structure these datasets, companies can reduce regulatory review times, allowing for faster approvals and drug launches.

Gold Medal - The trusted AI-Ready Data Layer

Reaching the Gold level means your data is no longer just structured and standardized—it is fully validated, compliant, and ready for the most critical applications. At this stage, data is aggregated across multiple sources, seamlessly integrating clinical, commercial, and safety datasets. It is directly consumable for regulatory submissions, AI models, and business intelligence dashboards, ensuring decision-makers have access to real-time, high-quality insights.

Maintaining Gold status requires a strong data governance strategy. Automating end-to-end data lineage tracking with tools ensures transparency and traceability. Governance and access controls must be in place to protect sensitive datasets, ensuring that only authorized users can access regulatory and compliance-critical data. Enabling self-service analytics empowers non-technical users to explore and analyze data without bottlenecks, increasing efficiency across teams.

AI-powered data lineage tracking and automated governance frameworks can ensure full traceability of clinical, regulatory, and commercial data. AI-driven monitoring detects inconsistencies, unauthorized changes, and potential compliance risks in real time. By implementing AI-driven data governance tools, organizations can improve regulatory compliance readiness and reduce audit preparation times.

✅ *Example Impact:* Implementing AI-driven data governance would have automated data lineage tracking ensuring regulatory teams always work with validated datasets. AI-powered compliance monitoring could proactively detect inconsistencies in trial data, allowing teams to address data integrity issues before submission. This would reduce compliance risk, streamline audit preparation, and cut regulatory submission review times, ensuring a faster and more reliable approval process.

3.3 - ADOPTING MODERN DATA INFRASTRUCTURE: THE FOUNDATION FOR A WINNING DATA STRATEGY

Modern infrastructure is the backbone that makes the principles of data as a product and the Medallion Framework operational at scale. Without a unified, scalable, and automated data platform, even the best-designed data processes will struggle with inefficiencies, inconsistencies, and bottlenecks. By investing in cloud-native architectures, real-time data processing, AI-driven automation, and a single integrated data platform, pharma companies can ensure that their data strategy is not only effective today but remains adaptable for the future.

Enabling a Unified Data Platform for Cross-Functional Collaboration

One of the most significant barriers in traditional pharma data management is fragmentation because different teams use different systems, leading to duplicated work and inconsistent insights. A unified data platform consolidates all clinical, regulatory, safety, and commercial datasets into a single ecosystem, ensuring seamless interoperability across teams. Instead of relying on multiple, disconnected systems, a centralized data platform built on cloud technologies like AWS, GCP, or Snowflake enables real-time data sharing and cross-functional collaboration. When picking a data platform, make sure you have the right team in place and are building the data platform at the right level of technical detail.

✅ *Example Impact:* Instead of clinical operations teams manually reconciling patient enrollment data across multiple EDC systems, a unified platform enables automatic ingestion, transformation, and validation of trial data. This reduces data reconciliation efforts and allows teams to make informed recruitment decisions in real-time.

Scalability and Automation for Faster Data Processing

Legacy, batch-based data systems slow down decision-making by requiring manual intervention at multiple points and data processing delays. By adopting modern, event-driven architectures and scalable cloud computing, organizations can process data in real-time as it is generated. AI-driven automation further enhances efficiency by continuously monitoring, structuring, and validating incoming data streams.

✅ *Example Impact:* AI-powered adverse event detection systems can analyze patient safety data as it is ingested, flagging potential concerns in real-time. This allows pharmacovigilance teams to identify safety signals within hours instead of weeks, enabling faster regulatory interventions and improved patient safety.

Ensuring Data Quality and Governance with AI

With an increasing volume of data generated across trials and commercial activities, maintaining regulatory compliance and data integrity is critical. Modern infrastructure and data contracts support AI-driven data lineage tracking, anomaly detection, and automated compliance monitoring, ensuring data is always audit-ready and meets stringent regulatory standards.

✅ *Example Impact:* Instead of manually compiling audit trails for an FDA submission, a pharma company can leverage AI-powered data governance tools that automatically track every data transformation, ensuring full transparency and traceability. This reduces audit preparation time ensuring a smoother, faster approval process.

Empowering Self-Service Analytics for Faster Insights

A modern infrastructure enables domain experts to access and analyze data without needing help from the technical team. With self-service analytics tools and AI-driven insights, decision-makers can rapidly generate reports, detect trends, and optimize strategies without being dependent on engineering teams.

✅ *Example Impact:* A regulatory team preparing a submission can instantly access validated trial data, generate compliance reports, and simulate approval scenarios using AI-driven analytics. This reduces the time required to prepare a submission package and improves regulatory success rates.

By combining modern infrastructure, the Medallion Framework, and a data-as-a-product approach, pharma companies can transform how they manage and use data—eliminating inefficiencies, accelerating decision-making, and ensuring compliance, all while reducing operational overhead.

4. WINNING THE DRUG DEVELOPMENT DATA DECATHLON

Successfully implementing data as a product, the medallion framework, and modern infrastructure stack requires a phased approach that balances immediate impact with long-term scalability. This section outlines a structured, incremental adoption strategy designed to help pharma companies modernize their data infrastructure with minimal disruption while continuously proving the value of the platform. The steps are sequenced starting with the lowest-cost, highest impact actions.

Incremental Adoption Strategies for Implementation

Step	What to do	Biggest Risk to Achieving It	Impact Once Completed	Example Success Criteria
1. Define and Prioritize Core Data Models for Clinical, Regulatory, and Business Needs	Define the essential data models needed for clinical, regulatory, and commercial functions to ensure interoperability and consistency.	Lack of alignment between different teams on data definitions.	Creates a unified structure for data ingestion, transformation, and governance.	Core data models approved and adopted by at least 3 functional areas.
2. Centralize Raw Data in a Data Lake (Bronze Layer)	Establish a single location (e.g., S3, Snowflake, or GCP Storage) to collect raw clinical, preclinical, regulatory, commercial, and diligence-related data. Ensure metadata tagging and lineage tracking.	Resistance from IT/security teams due to compliance concerns.	Eliminates data silos and manual file exchanges, enabling long-term scalability.	All relevant source systems successfully integrated, with metadata tagging and lineage tracking enabled.
3. Automate All Data Ingestions with Standardized Pipelines	Move from manual data pulls across all sources (EDCs, safety, real-world evidence, commercial, diligence) to automated extractions via APIs or scheduled batch processing.	Vendor constraints or lack of API support for legacy systems.	Reduces delays in data availability from weeks to hours, accelerating decision-making.	80%+ reduction in manual data pulls, with automated ingestion running on schedule.
4. Implement and Enforce Standardized Data Contracts & Schema Validation (Silver Layer)	Define and enforce common schemas for clinical, regulatory, and diligence data (e.g., SDTM/CDISC) to ensure standardization.	Schema drift from vendor updates; resistance from data providers.	Ensures clean, analysis-ready data that aligns with regulatory and business expectations.	95%+ compliance with defined schema validation rules across ingested data.

5. Implement Comprehensive Data Quality Checks (Silver Layer)	Add automated anomaly detection (e.g., missing values, protocol deviations, duplicate records) to all ingestion pipelines.	Initial false positives may cause user frustration, requiring fine-tuning.	Reduces time spent on data cleaning by analysts and statisticians.	Reduction in data quality errors by at least 50% over the first 6 months.
6. Develop and Deploy Real-Time Clinical, Regulatory, and Commercial Dashboards (Gold Layer)	Enable real-time views of enrollment, adverse events, diligence analysis, and protocol deviations with self-service analytics.	User adoption; need for training to ensure effective use.	Provides near-instant insights into study execution, deal analysis, and compliance.	75%+ of functional teams regularly use the dashboards within the first 3 months.
7. Integrate AI-Driven Data Harmonization Across Clinical, Regulatory, and Business Data (Gold Layer)	Automate mapping of clinical, regulatory, diligence, and real-world data sources into a unified model for longitudinal analysis.	Variability in source data structure may require ongoing tuning.	Unlocks cross-study and cross-indication insights for accelerated drug development and diligence evaluations.	AI-driven harmonization used in at least 3 major cross-study or diligence analyses within the first year.
8. Establish Automated Governance, Data Lineage, and Audit Trails (Gold Layer)	Enable automated lineage tracking, access controls, and data versioning to support regulatory audits and M&A diligence.	Compliance teams may be skeptical without clear SOPs.	Ensures regulatory and business readiness and supports historical data reconstruction.	Successfully pass at least one internal or external audit with full data lineage tracking.
9. Enable Secure Data Sharing & Self-Service Analytics Across the Organization	Provide controlled access for biometrics, regulatory, commercial, and M&A teams to explore structured data securely.	Fear of exposing sensitive data; need for access control policies.	Accelerates decision-making across functions by breaking down silos.	At least 5 functional teams actively using self-service analytics within 6 months.

10. Expand Advanced Predictive Analytics and AI for Business & Clinical Decision Support	Leverage AI for trend detection in clinical, regulatory, and commercial data to improve forecasting and decision-making.	Resistance to adopting AI-driven insights; need for robust validation.	Improves probability of success for trials and acquisitions by identifying high-value assets and risks early.	AI-powered predictive models used in at least 3 strategic decisions within the first year.

5. CONCLUSION - ARE YOU READY TO WIN GOLD?

Most pharma companies are stuck at the Bronze level, dealing with fragmented, unstructured data. Some make it to Silver, where data is cleaner but still not fully optimized. Only a few reach Gold, where data is validated, AI-ready, and a true competitive advantage. To compete at the highest level, organizations must modernize their data ingestion strategies now.

Where does your data stand in the Olympics? Run a data ingestion fitness test and assess whether you are operating at Bronze, Silver, or Gold. The path to better, faster decision-making starts with building a strong data foundation because in drug development, winning with data means getting life-saving treatments to patients faster.

REFERENCES

OpenAI. *ChatGPT, Feb 10 version*. 2025, https://openai.com.

ACKNOWLEDGMENTS

Formation Bio PHUSE Team - Will Linto, Emily Yates, Andrew Burd, Melanie Hullings, Matt Luppino, Tasnima Tanzim, Will Linto, Laura Ng

RECOMMENDED READING

https://www.getdbt.com/blog/data-product-data-as-product
https://www.mckinsey.com/capabilities/quantumblack/our-insights/how-to-unlock-the-full-value-of-data-manage-it-like-a-product
https://learn.microsoft.com/en-us/azure/databricks/lakehouse/medallion
https://www.databricks.com/glossary/medallion-architecture

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Author Name: Cathryn Herbst
Company: Formation Bio
Email: cathryn@formation.bio
Website: formation.bio