Paper ET22

What Pharma Can Learn from Data Science Outside the Industry

Kris A. Wenzel, MMS Holdings, Canton Township MI, United States of America

Introduction

Pharmaceutical data systems are designed for integrity, traceability, and compliance, not necessarily for speed of insight or early decision making. That tradeoff is increasingly limiting timely decision making in fast moving studies, as data volumes grow steadily and data is generated more frequently. Waiting for finalized Standard Data Tabulation Model (SDTM) and Analysis Data Model (ADaM) introduces avoidable decision latency, particularly for operational oversight and safety review.

Other regulated industries have addressed similar challenges by treating analytics as a shared operating system, optimizing not only *which* decisions are made, but *how quickly* reliable signals can be acted upon. Decision speed and data quality are not opposing goals when data is engineered through repeatable, transferable processes with explicit guardrails.

This paper proposes a layered engineering approach that enables earlier, controlled access to analysis-ready data while maintaining pharma grade controls and traceability. By organizing data progression through the bronze, silver, and gold layers, organizations can make data visible sooner, clearly label its level of readiness, and still support downstream regulatory workflows. The focus is on standardizing ingestion, transformation, and semantic meaning of the data, rather than attempting to standardize vendor tools or raw data formats that are outside organizational control.

This approach allows teams to reduce decision latency without compromising compliance by making preliminary insights available earlier and refining them as data progresses toward submission quality artifacts. The result is a pragmatic phased model that delivers faster operational and safety insights today, while remaining fully aligned with SDTM and ADaM where required.

It is important to emphasize that this is *not* an argument for relaxing data quality standards, bypassing biometrics, or using preliminary data for regulatory submission. Rather, it is a proposal to engineer early analytics as a first class, governed capability, rather than treating them as ad hoc or disposable outputs.

Decision Speed and Analytics as a Shared Operating System

Decision speed matters in modern businesses, particularly in environments where safety, quality, and operational risk must be continuously monitored In industries such as automotive, companies actively measure how quickly they can move from a signal to an action, from the moment data is received to the point where a decision can be made. The goal is to shorten this time as much as possible. This is often referred to as **decision latency,** or time from signal to action.

In other safety-critical industries, early operational signals are made available through validated, repeatable processes, while full audit-grade validation occurs later in the lifecycle.

Pharma has traditionally been more focused on accuracy than speed, and rightly so. Clinical teams prioritize correct and complete data, and decision speed is often not explicitly measured or optimized. Reviews commonly wait for ADaM datasets or custom tables before action is taken. However, as study pace increases and trials generate more data more frequently, this model becomes increasingly strained. Reviewers and clinicians cannot afford to wait weeks for finalized outputs. Early signals need to be visible sooner, even if the data is not yet final.

Because of this, decision latency should be treated as a core metric in pharma. How long does it take to move from data availability to action? How can we improve decision speed while maintaining high data quality? And how do we treat slow data access as a business risk, rather than just an IT issue?

In other industries, analytics often functions as a shared operating system. Finance, manufacturing, and supply chain teams use the same metrics, see the same data, and tell the same story. This shared view creates alignment and enables faster, more confident decision making. In pharma, analytics is still largely siloed. Electronic Data Capture (EDC) system data is reviewed in one set of reports, lab data arrives separately from vendors, SDTM is viewed in different tools, and ADaM results are delivered later. There is no single, connected experience.

Pharma needs a consistent data foundation where information flows in a controlled way from EDC into SDTM. Users should be able to view source data and standardized data side by side. This enables faster safety monitoring and operational insight. Each stage serves a purpose: EDC supports rapid operational and safety signals, SDTM provides standardization and traceability, and ADaM supports regulatory submissions. These stages should be visible, not hidden.

The actionable takeaway is this: treat analytics as a shared system that supports daily decision making. Align teams around common data and metrics, and optimize not only *how right* decisions are, but also *how fast* they are made.

Fundamental Differences in Data Culture

One of the biggest differences I noticed when moving from automotive to pharma was the difference in data culture. The automotive Information Technology (IT) landscape is complex; there are many applications and external sources from which to obtain data. Not only are you contending with your own systems, but you are also constantly receiving and sending data to others.

As such, there are two key issues to contend with. The first is the need to funnel all this data into a cohesive system that can be used for reporting. The second is that many of these systems have their own rules and data quality issues, making it challenging to get a consistent, clear picture of your data. A good example of this is working with part numbers. Many systems have their own way of storing a part number, and in some cases the numbers are completely different. I have seen situations where the same part had up to five different numbering schemes.

In clinical studies, data is handled very differently than in industries such as manufacturing or automotive. The primary focus is scientific integrity and patient safety. Data from one study must always remain separate from another, since mixing data could lead to incorrect conclusions. For this reason, strong controls and checks are required throughout the data process. At the same time, clinical data standards define how data moves from EDC through SDTM and ADaM, which allows data to be handled in a structured and consistent way, even when data is processed early.

Clinical data usually moves in a sequence from EDC to SDTM, then to ADaM, and finally to tables, listings, and figures. Many reporting outputs are created late in this process, often after SDTM and ADaM are complete. In a Safety Review Committee (SRC) setting, reviews happen quickly and teams must detect safety issues as early as possible. When decisions depend on the end of the pipeline, action can be delayed.

By processing EDC data toward SDTM on a frequent basis, safety teams can generate clear summaries earlier in the study. These summaries are validated to be correct and repeatable based on the data available at the time, but they are not yet validated for regulatory submission. Even with these limits, the information can support timely safety decisions and help teams decide where to look more closely. This approach is like manufacturing, where teams rely on same day or next day data that is process validated, while full audit validation happens later.

Data Engineering’s role in Data Science

When I started working in pharma, I noticed there was a strong emphasis on data science. This makes sense, as most organizations aim to take clinical research data, analyze it, and publish or submit the results. However, there is a major step in this process that is often overlooked: how the data is engineered.

I consider data engineering to be the set of steps used to ingest and transform data so that it is ready for analysis. This includes activities such as curation and harmonization.

Data engineering is not simply writing a program to read a SAS file and make it usable by a dashboarding tool, it is much more . Data engineering is concerned with the entire process of moving data from receipt to analysis. It focuses on repeatable results, consistent data quality, and adequate reporting so that, as processes run, you know the data remains within established guardrails.

Manufacturing Grade Data Quality

Why Integration is Hard

The typical goal of a data warehouse is to pull in data from many sources and consolidate it into usable facts and dimensions. This can be challenging when you consider that each source system may use a different scheme for part numbers and other metadata.

When building a data warehouse for business reporting, you often have no control over the source system's data format or even its meaning. It is up to the warehouse builder to understand how to transform each system's data to form a consistent and holistic picture.

Example:
One system reports inventory on an hourly basis, starting with a beginning inventory amount and recording hourly increases or decreases. Another system reports inventory daily. These two datasets cannot be freely mixed. The solution is to take the hourly inventory data and aggregate it into a daily view. Once this is done, the data can be safely combined.

In pharma, I have seen an approach where data handling is retrospective, siloed, and compliance focused. Data is intentionally kept separate. The goal is to use a consistent, well defined, and well documented process to transform data, with a strong emphasis on provenance and repeatability.

Why Pharma Integration is Still Hard Across Studies

One key difference between automotive and pharma is that, in pharma, there is a tendency to keep data separate rather than aggregated. Projects remain isolated, and study data does not cross boundaries. This simplifies many of the issues seen in automotive, where combining data from many disparate sources is unavoidable. However, this simplification applies primarily within a single study or project.

What happens when we want to build a consistent process to handle study data across projects? Once you step outside a single study, new variables are introduced. You must contend with different EDC vendors, a variety of vendor specific data formats, and differences in how data is originally received. Even within the same EDC platform, data can be coded differently.

In this context, we cannot fully control the source systems. What we *can* control is the process used to ingest and transform the data.

The key takeaway is not to focus too heavily on application standardization. EDC vendors will change, coding rules will vary, and raw data formats will differ. Instead of spending effort on standardizing what cannot be controlled, focus on standardizing ingestion and processing pipelines. Embrace a layered architecture that moves data from receipt to a form that can be reliably used for decision making.

Data Harmonization at Scale

One challenge we consistently face is harmonizing data, and this becomes especially apparent when stacking studies. It would be ideal if everyone could simply agree on the same standards and move forward. Unfortunately, that is not how the world works, and instead we must deal with the data we receive.

In pharma, many of the harmonization challenges stem from attempts to force identical schemas, treatments, or representations of information. In some cases, this is simply impossible. Studies may have been conducted at different points in time, and it is not feasible to go back and change the standards or tools that were originally used. Trying to enforce a single tool or technology across all situations often leads to brittle systems. Tools can break when applied to problems they were never designed to solve, and this approach frequently creates resistance within organizations, as teams may feel that a mandated tool is not optimal for their needs. The result is increased friction, slower onboarding, and delayed project starts.

On the other hand, having too few standards can lead to an uncontrolled environment with significant local divergence. The same concept may be represented in many ways, turning harmonization into a highly manual, error-prone process. At that point, integration becomes slow, fragile, and difficult to scale.

So, the real question becomes: how do we make harmonization of something that can be automated and applied consistently across many studies? The answer is to focus on harmonizing the meaning of the data and the outcomes, rather than the source systems themselves.

For example, adverse events or concomitant medications may be collected using different schemas, terminologies, or levels of structure across studies. While it is rarely feasible to standardize how those data were originally captured, it is possible to align them to common semantic targets, such as MedDRA for events or WHO Drug for medications, in a controlled and repeatable way. What we need is a robust process that allows us to take the data as it is received and reliably transform it into a desired target format.

The goal is to standardize the semantic intent of the data. We need to be confident that we understand what the data means, that meanings are not mixed, and that outcomes are correct. The result should be data that is analysis ready and on a clear path toward producing regulatory facing artifacts.

Achieving this requires many pieces to come together in a coordinated way. In automotive, we addressed similar challenges by running processes nightly to prepare data and move it along a defined path. One of the most effective tools for enabling this was having a framework that clearly described how data progresses from raw intake to production ready outputs, such as dashboards or summary tables.

That framework is the medallion architecture, which we will discuss in the next section.

What is a Medallion Architecture?

The medallion architecture is a data engineering concept that describes how data progresses from initial receipt to being used for analysis. In this architecture, data mapping and transformation happens progressively across layers within the environment. In our case, those layers are **raw**, **bronze**, **silver**, and **gold**. Each layer has a specific purpose, reflecting both technical transformation and level of data readiness, which we will walk through in the following sections.

The medallion architecture provides a way to think about and organize data. We use four layers: raw, bronze, silver, and gold.

Raw Data Layer

As data is received from vendors or other systems, it is placed into the raw layer. The primary purpose of this layer is to collect and preserve data exactly as it is received. It serves as a place to organize data by source, project, and study, while also aligning it by the date received. It is an invaluable resource to fall back on if data ever needs to be reprocessed or audited.

The core intention of the raw layer is preservation. If we ever need to rebuild our system, the raw layer provides a reliable source of truth from which everything else can be reconstructed. Key characteristics of the raw layer include:

- Data is organized and curated, typically by sponsor, project, and study.
- Data is stored *as received*. If a ZIP file is delivered, it is stored as a ZIP file. If data arrives as SAS datasets, spreadsheets, or Comma Separated Value (CSV) files, those original formats are preserved.
- Data is stored by receipt date so historical versions can be referenced if inconsistencies arise later.

- No alterations, transformations, timestamps, or modifications are applied. The data is kept in its natural, original state.

Bronze Data Layer

The goal of the bronze layer is to store uniformly formatted, single file datasets organized by date received. At this stage, data is still organized by project and study, but it is converted into a consistent, tabular structure.

Data is unpacked and normalized into rows and columns. For example:

- An Excel file with multiple sheets is converted so that each sheet becomes its own table.
- JavaScript Object Notation (JSON) files are flattened into tabular structures.
- Data from various formats, delimited text files, SAS datasets, Excel files, is converted into a common format. In our case, that format is **Parquet**.

During this conversion, data encoding issues are resolved, and all data is stored using a consistent system encoding. In addition, tracking fields are added to each row, including:

- Load timestamp
- Original file name
- A provenance identifier that allows each record to be tracked through subsequent layers

By the time data reaches the end of the bronze layer, we can confidently say that:

- All data can be read consistently.
- All data has been treated uniformly.
- Structural and encoding inconsistencies have been resolved.

Silver Data Layer

The goal of the silver layer is to align data into CDISC familiar tables and variable names, and to begin stacking study data together. This is the point where we move from handling individual files, as in bronze, to building a domain-oriented data store.

For example, adverse event data across multiple studies within a project is stored in a single adverse event table with a consistent schema. At this stage, we are no longer managing files, we are building a database of domain data.

Key activities in the silver layer include:

- Organizing and stacking data across studies
- Adding a system generated study identifier. This identifier is independent of vendor provided values, which may be unreliable or inconsistently populated. This approach allows us to confidently associate data with the correct study.
- Aligning tables and variables to SDTM naming conventions where applicable
- Performing most of the data mapping and transformation

This is also where data from multiple sources is harmonized at a structural level. For example, laboratory data originating from EDC systems, local or regional labs are transformed and merged into a unified lab domain for the project. Additional steps include aligning data types, preserving original values, and stacking datasets across studies.

At the conclusion of the silver layer, data can be viewed consistently across studies. However, differences in coding and treatment may still exist between studies. To address this and prepare the data for analysis, we move to the gold layer.

Gold Data Layer

The gold layer is where data is prepared for reporting and analysis. This is where harmonization occurs to ensure consistency across studies, for example, standardizing adverse event terms or other key variables.

In the gold layer, we also build dimensions to support reporting. A common example is the **subject dimension**, which is typically derived from Demographic entries. In some cases, additional sources such as labs or Electrocardiograms (ECG) are used to account for subjects who may have screen failed before demographic data was collected.

Additional activities in the gold layer include:

- Constructing patient profiles by combining multiple domains such as subject, Adverse Events, Demographics, Concomitant Medications, Exposure, ECG, and Lab
- Deriving variables such as Adverse Event Start Day (AESTDY), based on the number of days between event start and study start.
- Creating study or project specific derived variables
- Building combined disposition tables from multiple silver layer sources

By harmonizing and preparing the data in the gold layer, it becomes suitable for higher level analysis. This is an ideal place to build dashboards, summary tables, and patient profiles.

It is important to note that gold layer artifacts are often derived from sources such as EDC extracts. While this data is extremely useful for early clinical review, it is not necessarily submission ready. This is like manufacturing, where operational data is used to monitor production but is not what would be sent directly to tax authorities. In pharmaceuticals, biometrics continues this process to produce submission quality SDTM and ADaM datasets.

gold layer objectives include:

- Preparing data for final use in dashboards
- Creating subject dimensions is usable across all domains.
- Building combined disposition and reporting tables
- Deriving analytical variables
- Producing the “public” view of the data for decision making

By this point, you should be able to see how the medallion architecture enables a repeatable, structured process. Data is continuously evaluated from the raw layer forward, allowing new data to be processed through bronze, silver, and gold in a consistent manner.

One key advantage of this approach is recoverability. If issues arise in the silver or gold layers, those layers can be confidently deleted and rebuilt entirely from raw. Achieving this requires a robust, repeatable processing framework, which we will discuss in detail in the next section.

Embracing Repeatable Processes

In automotive, data transformation is not just an ad hoc activity, it is a relied upon production process. Each run is fully automated and produces consistent, repeatable results. Processes are built to reliably read source data, ingest it, and apply transformations in a consistent way. The medallion architecture we just discussed provides a convenient

framework for organizing and reasoning about these transformations. The key goal is that, at any point, the system can be purged and rebuilt entirely from the raw layer and still produce the same results.

This approach moves us away from reliance on intermediate data and manual steps. While manual intervention can sometimes speed up delivery in the short term, it introduces risk and inconsistency. Automated, repeatable processing enables faster data delivery *and* significantly reduces the risk of error.

Another important goal is to put processing in place that treats data consistently every time. For example, when loading EDC data, the process should identify the data, recognize CDASH standards, and act accordingly. Data should move through the same mapping and transformation steps, with clear visibility into where and why transformations occur.

A good example is how vital signs and ECG data are received and processed. Although these domains represent distinct types of data, they often arrive in a similar structure, with values stored in column order that must be transformed into row-based representations. This transformation, commonly referred to as unpivoting columns, is required in both cases. Rather than writing domain specific programs for each dataset, a better approach is to implement a single, standard process that can be applied across all domains. The process itself remains unchanged and does not contain domain specific logic; instead, configuration supplies the domain context and defines how the unpivot should occur.

An ideal place to apply this approach is during the transformation of bronze layer data into silver. A load program that operates independently of domain and relies on configuration to drive behavior promotes consistent handling of data while maintaining clear provenance. This aligns with the medallion architecture, in which incoming data is landed in a centralized data lake and transformed through standardized processes as it moves from one layer to the next.

All incoming data should be cataloged by receipt date. Load processes should rely on watermarks that capture the most recent successful ingestion, ensuring that only newly received data is processed. To reduce risk and operational complexity, the pipeline should avoid subject level or record level incremental update logic. Instead, each load should treat the study as a single, complete unit of data and fully replace any previously loaded version. This wipe and replace approach favors correctness, transparency, and maintainability over fine grained incremental processing.

What can Pharma do now vs Later?

So, what can pharma do now, versus later? Pharma companies want high quality data, but they also need faster decision making. Some improvements can be made immediately through technology and process changes, while others take more time because they require people and teams to work differently. Understanding this distinction helps build a realistic and achievable roadmap.

What can pharma adopt now in the near term?

Here are some practical steps that can be taken immediately.

1. Organize Data with a Clear Structure

Pharma can adopt simple, proven ways to organize data, such as the medallion architecture. Store data exactly as received in the raw layer. Use bronze to convert data into reliable, readable tables. Use silver to align data to common standards such as SDTM. Finally, use gold to prepare harmonized data for dashboards and review.

This approach gives teams a shared understanding of where data lives, how it progresses, and what level of readiness each layer represents.

2. Improve Data Engineering Practices

Strong data engineering ensures that data is reliable and repeatable. This includes:

- Clear data contracts and processing rules

- Logging and monitoring pipelines
- Automated testing
- Tracking data provenance and transformations

These practices reduce manual effort, lower the risk of errors, and increase trust in data.

3. Deliver Usable Data Earlier

Teams should not always need final, submission-ready data to make decisions. Pharma can share data earlier, validated enough for review, while continuing to refine it for SDTM and ADaM downstream. This supports faster safety and operational decisions. Speed and accuracy can coexist if data is handled correctly and transparently.

4. Clarify Roles Between Teams

Data science, biometrics, and data engineering each play distinct roles. Clearly defining responsibilities helps teams work together effectively and avoids confusion or duplication of effort.

What Pharma Should Plan for Later (Long Term)

Some changes take more time because they involve people, culture, and habits.

1. Shared Decision Making

Teams need to agree on what data means, who owns decisions, and who is responsible for data quality. This shared understanding builds trust and enables faster, more confident decision making.

2. Semantic Governance

Rather than forcing everyone to use the same tools, pharma should focus on shared definitions, common metrics, and consistent outputs. This approach harmonizes data across studies without creating brittle systems.

From a data engineering perspective, this means embracing variation rather than trying to eliminate it. Tools and processes should be designed to handle differences in source systems, formats, and vendors while still producing consistent outcomes.

3. Cross Team Standardization

Groups such as data management, data science, and biometrics should align on common concepts and terminology. Where possible, they should agree on shared standards and favor simple, repeatable processes. This reduces friction, speeds onboarding, and improves communication across the organization.

Key Take Away

Not everything needs to change at once.
Do now: improve data organization, optimize data engineering practices, and provide earlier access to usable data.
Plan for later: cultural alignment, shared ownership, and governance.

The best approach is to build a phased roadmap:

- What can we do today?
- What can we do next month?
- What can we do in three months?
- What will take six months or more?

By separating quick wins from longer term change, pharma can move faster without sacrificing data quality or compliance.

Glossary

ADaM (Analysis Data Model): CDISC standard for analysis datasets used in statistical workflows and regulatory submissions.

Bronze / Silver / Gold (Medallion Architecture): Common data engineering pattern for raw, standardized, and analysis ready data layers.

CDASH (Clinical Data Acquisition Standards Harmonization): CDISC standard guiding how data are collected in EDC systems.

CDISC (Clinical Data Interchange Standards Consortium): Standards body responsible for CDASH, SDTM, and ADaM.

Clinical Domains (e.g., DM, AE, CM, EX, ECG): Standardized SDTM groupings representing related clinical data concepts.

Data Catalog: Metadata system describing available datasets and their intended use.

Data Lake: Central repository for large scale clinical and operational data.

Dimensions: Descriptive entities (e.g., subject, study) used to contextualize measured data.

Domain Oriented Data Store: Data organization structured by clinical domain rather than application or pipeline stage.

EDC (Electronic Data Capture): Systems used to collect clinical trial data at investigation sites.

EDC Extracts: Operational data exports from EDC systems.

Facts: Tables capturing measured events or observations.

Lineage: Tracking of data movement and transformation across processing stages.

Parquet: Columnar storage format optimized for analytical workloads.

Provenance: Metadata describing data origin and applied transformations.

Review-ready: Data suitable for internal review and decision making but not finalized for submission.

SDTM (Study Data Tabulation Model): CDISC standard for regulatory tabulation datasets.

Submission-ready: Data fully compliant with regulatory submission requirements.

Unpivot: Transformation converting wide data into normalized, row-based form.

Watermarks: Markers used to manage incremental data processing.

Contact Information

Kris A. Wenzel

MMS Holdings, Inc.

880 Commerce Blvd, Canton Township, MI 48187

kwenzel@mmsholdings.com

www.mmsholding.com

linked in @ https://www.linkedin.com/in/kriswenzel/