Paper ET23

The Clinical Data Lake: Empowering the Data Caterer Vision with Veeva, Databricks, SAS, and Open Source

Sascha Ahrweiler, Bayer AG, Wuppertal, Germany
Holger Dach, Bayer AG, Wuppertal, Germany

ABSTRACT

The Data Caterer vision transforms programmers into curators who prepare and deliver actionable insights. At Bayer, our clinical data lake makes this shift possible.

- Veeva acts as the structured pantry, ensuring organized, traceable data;
- SAS provides reliable analytics for regulatory-grade outputs and automated analyses;
- Open Source tools bring flexibility and innovation, supporting machine learning and modern visualization.

Early implementation highlights cover data flows, compliance safeguards, and readiness for AI. By blending trusted technologies with new approaches, we allow teams to focus on delivering high-quality insights rather than debating tools. This technical deep dive explores how these foundations work together, sharing practical lessons from Bayer's journey. The result: an environment where clinical data is not just stored but curated and served to power better decisions, embodying the Data Caterer vision with Veeva, SAS, and open source solutions.

INTRODUCTION – WHY RE-CONSIDER A NEW CLINCIAL DATA TECHNICAL LANDSCAPE

Clinical development technology landscapes were often built with a "decades" mindset. Systems were designed to last, customizations accumulated over time, and many processes evolved around functional boundaries. That approach can work—until the organization needs to move faster, reuse data across programs, and introduce automation and AI in a controlled way.

Our experience is that "digital transformation" is frequently interpreted as technology adoption. In practice, it is a change in how work is done. It changes how study intent is captured, how data is governed, how handovers are reduced, and how insights are produced and consumed.

OUR DIGITAL VISION

Our digital vision centers on improving work, not just installing platforms. We aim to rethink end-to-end processes, keep data accurate and traceable, and use digital routines to increase efficiency and effectiveness. The most important outcome is not a new tool landscape. The outcome is a more reliable path from study intent to decision-making.

THE "WHY" BEHIND THE VISION

We designed this architecture to improve operational efficiency without weakening scientific rigor. We want teams to collaborate more effectively and to deliver insights in a way that is customer-centric. At the same time, we need a stable and compliant foundation that supports automation and data-driven processes.

External drivers reinforce the need for change. Regulations evolve, legal obligations increase, and the technology market moves quickly. If we want to use modern capabilities responsibly, we must standardize core patterns and build the skills to operate them.

In practical terms, our strategic direction can be summarized as follows:

- Simplify the landscape and retire legacy systems where possible.
- Implement an end-to-end platform approach for study setup, conduct, and closure.
- Lead with a "Veeva-first" mindset to maximize synergies between vaults and align to industry standards.
- Build an analytics landscape that covers exploratory analytics, operational analytics, and statistical analytics, anchored on Databricks and SAS hosted in AWS Cloud.

- Enable and open source where it contributes value under governance

This landscape decision becomes much more meaningful when it is linked to a clear operating model and the users or certain persona, which we will discuss next.

MEET THE DATA CATERER: THE STATISTICAL PROGRAMMER OF THE FUTURE

In earlier presentations (including Sascha Ahrweiler's presentation at the 2025 PHUSE SDE Copenhagen and the 2026 PHUSE APAC Connect in Hyderabad), we introduced the role of a Data Caterer as a practical persona that describes how the statistical programming role might evolve in a modern clinical data supply chain.

We intentionally use a food analogy because it is easy to understand and principles are transferable. Data does not become valuable because it exists. Just like food, data becomes valuable when it is prepared, curated, and served in a way that supports the reliable decisions, we are all craving for. The role of the Data Caterer is to understand the cravings and preferences of their customers – for example clinical teams or regulatory agencies – and prepare the plates accordingly. Some might prefer fine dining, which could translate into a dashboard for quick decision making. Others might prefer a buffet style, which could translate into a data catalog of transformed data to enable a statistician to conduct modeling and simulations. And some might prefer a Swiss raclette style type of food, which could translate into some derived variables and conducted analysis you can put together to tell your own data story.

WHAT ENABLES THE DATA CATERER

The vision described above require a statistical programmer to apply a caterer mindset. The Data Caterer starts with the decision context based on good relationships with their customers all along the pharma value chain. Instead of optimizing single SAS or R programs, they focus on the analytical question and what stakeholders need to decide. They anticipate downstream requirements through strong stakeholder relationships. They also shape how insights are consumed by choosing the right format, the right timing, and the right level of explanation.

This shift requires both educational upskilling on the one hand and a stable technical environment on the other hand. We have covered the educational upskilling in the PHUSE 2026 APAC Connect presentation. In this paper, we will provide more insights into the data caterers kitchen – aka the technical landscape. The shift from a traditional statistical programmer to a Data Caterer is only possible when the end-to-end data flow is reliable. Standardized ingestion and a governed clinical data lake reduce the time spent reconciling sources and chasing provenance. A modern computing environment supports reproducibility and mixed tooling. Together, these building blocks allow the Data Caterer to focus less on manual wrangling and more on interpretation, trade-offs, and insight delivery.

HOW THIS DIFFERS FROM ADJACENT ROLES

The Data Caterer does not replace statisticians or clinical data scientists. Statisticians own methodological strategy and scientific inference. Clinical data scientists often drive exploratory analytics and advanced modeling. The Data Caterer owns decision readiness. They orchestrate the flow from data capture to analysis outputs and ensure the result is traceable, timely, and consumable.

This is a natural evolution of today's statistical programmer skill set. Statistical programmers have the unique combination to understand clinical data, have great technology insights, and know regulatory expectations by heart. The difference is the center of gravity: from code execution to service-oriented delivery.

With this Data Caterer persona established, the next step is building the "kitchen" for them.

DESIGN PRINCIPLES FOR BUILDING THE DATA CATERER'S KITCHEN

Around two years ago, we reviewed our full clinical data landscape. A key learning was that technology cycles are nowadays shorter than they used to be. Planning horizons are closer to three to five years. In that world, heavy customization becomes expensive quickly and limits agility.

We therefore adopted a pragmatic platform strategy. We use industrial-strength systems where stability and compliance are critical. We bring in open-source innovation where intent, design, and orchestration benefit from speed and openness. The goal is not to declare a single tool winner. The goal is to build an ecosystem that is compliant by design, flexible where it matters, and scalable for automation and AI.

The architecture is built around three pillars: Veeva, the Clinical Data Lake on Databricks (CDL), and the Statistical Computing Environment (SCE) on SAS Viya.

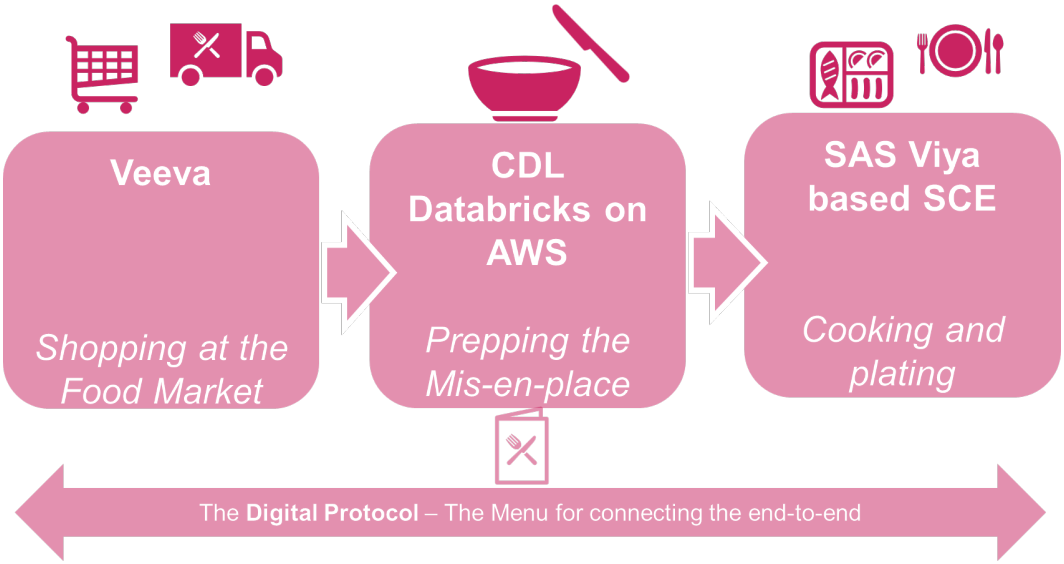

Figure 1: *High level illustration of our architecture for the end-to-end clinical data flow from document and data ingestions to transformations up to analysis and reporting*

SHOPPING AT THE FOOD MARKET – GETTING DATA IN THROUGH VEEVA

In the kitchen analogy, Veeva is the place where ingredients originate. This is where structure, provenance, and compliance are enforced closest to operational execution.

Veeva serves as an operations platform across protocol development, clinical operations, documentation, and submission. One reason for our "Veeva-first" approach is its broad footprint across functions. Vaults are designed as an integration backbone across the life science value chain. In practice, this covers workflows such as eTMF, CTMS, study startup, EDC/CDB, quality systems, training, and more.

Veeva Vault is designed for regulated workloads. The platform emphasizes security, validation, and consistency. It follows a metadata-driven approach that brings structured data, unstructured content, and workflow logic into a unified model. This matters because regulated work is not only about storing documents or rows in tables. It is about controlling lifecycles, ensuring audit trails, and managing change.

INTEGRATION PATTERNS THAT SUPPORT THE DOWNSTREAM ECOSYSTEM

To integrate with downstream environments, Vault provides interfaces for structured data access and workflow integration. Typical patterns include REST APIs and direct data interfaces for higher throughput. Where needed, server-side customization can be implemented with supported SDK approaches. Event-driven patterns enable near real-time reactions to content or data changes.

These interfaces are critical because a clinical data ecosystem is never a single system. It is a set of connected stations. Veeva provides a stable foundation for that connectivity.

PREPPING THE MISE-EN-PLACE – THE CLINICAL DATA LAKE (CDL) ON DATABRICKS

If Veeva is the food market, the CDL is the preparation station. This is where ingredients are washed, chopped, portioned, labeled, and made ready for efficient cooking.

The CDL acts as the central repository for valuable clinical trial data, both current and historical, excluding very old and non-relevant data. It is designed to support reuse, enrichment, and consistent exposure to downstream consumers. Technically, the CDL is built on Databricks hosted in AWS and uses Delta Lake as a storage layer.

A MEDALLION ARCHITECTURE THAT SEPARATES RAW FROM READY

We use a medallion architecture because it keeps responsibilities clear. Raw data lands first, then it is harmonized, and finally it becomes business-ready.

In the Bronze layer, the CDL stores raw study data from sources such as Veeva CDMS/EDC/CDB, external vendors, labs, PK/Bio, devices, and CRO deliveries. Data can arrive in different formats. The key is that it is cataloged immediately and governed from the moment it enters the platform.

In the Silver layer, the CDL reconciles and harmonizes data. Standard transformation logic and reference data are applied. This includes mappings driven by study metadata, unit conversions, medical coding reference tables, and transformation libraries.

In the Gold layer, the CDL produces curated data products that downstream teams can use reliably. Depending on the use case, this includes SDTM-like operational models, analysis input packages, and controlled release bundles.

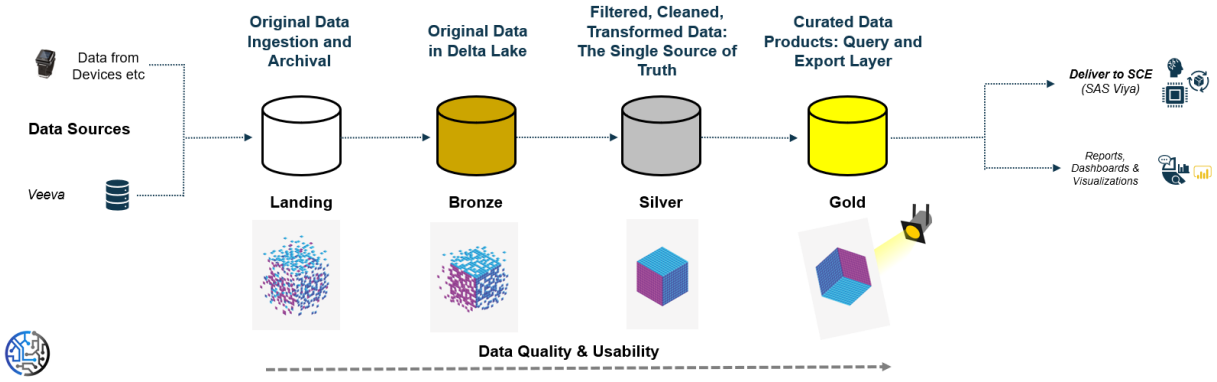

Figure 2: *Example Bronze → Silver → Gold flow with representative data products and downstream consumers.*

WHAT THE CDL PREPARES FOR THE ORGANIZATION

The CDL is designed to serve multiple consumers. It supports operational analytics such as RBQM and centralized monitoring. It supports medical review use cases such as anomaly and outlier detection. It also serves our statistical computing environment as a foundation for analysis dataset creation and other statistical data pipelines. In addition, it can feed dashboards for operational oversight.

To enable these consumers, the CDL exposes data through governed SQL endpoints, controlled sharing mechanisms, and APIs for automation.

CDL GOVERNANCE, LINEAGE, AND GXP READINESS

Governance is not optional in this environment. We use catalog and lineage capabilities to control schema evolution, document metadata, and provide traceability across transformations and releases. We also separate environments into Dev, Test, QA, and Prod, and we deploy through validated pipelines with change control.

From a GxP perspective, the CDL follows a qualified SDLC. Validation deliverables, release documentation, and audit readiness are built into the operating model.

A common misconception is that a data lake must behave like a traditional database lock environment. That is not the CDL's purpose. The CDL is optimized for ingestion, harmonization, enrichment, and reuse. It supports versioning and time travel so teams can reproduce results and understand what changed. However, the final regulatory lock belongs downstream, where regulated deliverables are produced and controlled promotion is applied.

COOK IT AND PLATE IT – THE STATISTICAL COMPUTING ENVIRONMENT (SCE) ON SAS VIYA

In the kitchen analogy, plating is where preparation becomes something a customer can consume. For clinical analytics, this is where analysis is executed, outputs are produced, and results are shaped for the decision context.

The SCE provides the controlled execution and delivery layer. It supports statistical analyses, regulated pipelines, and consistent output generation. It is not designed to be a raw ingestion platform. It is designed for reproducibility, access control, and controlled promotion.

WHY SAS VIYA AND WHY NOW?

Similar to other pharma companies, we have a home grown SCE, which has grown over a decade and always served its purpose to provide regulatory compliant deliveries. With the technology acceleration and especially with cloud adoption and AI readiness, we decided to modernize our environment significantly.

We are establishing our SCE on SAS Viya to ensure business continuity and to create a foundation for broader transformation objectives. This renewal program aims to modernize data and analytics capabilities while maintaining compliance for GxP and non-GxP use cases.

A central design goal is language-agnostic compute. The future is multilingual. We need the ability to run SAS, R, and Python under a validated umbrella, without fragmenting governance.

SCE TARGET STATE AND INTEGRATION

The target state is a Viya-based SCE with containerized compute and standardized environments. Data exchange with the CDL is based on governed sharing and APIs. An optional clinical repository component is evaluated through a structured A/B approach.

The practical operating principle is straightforward. We want scalable compute for transformations and analytics, but we also want controlled promotion, traceability, and stable delivery patterns for regulated outputs.

Role-based access and controlled unblinding are essential in clinical development. The SCE implements entitlements that distinguish standard access from blind-breaking access, and it enforces controlled workflows for unblinding and snapshotting. This ensures that interim and final analyses remain traceable.

Our SCE roadmap includes AI-assisted capabilities such as code copiloting, automated checks, and explainable derivation tracing across dataset pipelines. Over time, the goal is to enable “analysis applications” that product teams can consume under governance. In all cases, the operating constraint remains the same: automation and AI must be introduced with traceability, controlled change, and risk-appropriate validation.

THE MENU – THE DIGITAL PROTOCOL

Even the best kitchen stations do not work well if the recipe is unclear. In the following, we’ll dive into one of many examples, where the different technologies are connected. The example of a (digital) clinical protocol will illustrate the connections. In clinical development, the protocol defines study intent. When protocol intent is expressed only in documents, the organization spends time translating intent into operational configurations and downstream data definitions.

The digital protocol addresses this gap. It is a machine-readable representation of study intent that can connect design and execution. It supports standardized exchange across Veeva, the CDL, and the SCE. It can enable automation for downstream processes such as EDC screen generation, lab specifications, edit checks, and document creation.

In the analogy, the digital protocol is the recipe. It ensures every station works from the same intent and turns the data flow into a coherent, traceable decision pipeline.

WHERE OPEN SOURCE CAN ADD VALUE: OPEN STUDY BUILDER

Selected open-source solutions are particularly valuable upstream, where interoperability and flexibility matter most. Open Study Builder can help industrialize the digital protocol by making study intent structured and machine-readable. Importantly, this does not replace regulated execution platforms. It complements them by strengthening the “shared intent” layer.

THE DATA CATERER IN ACTION

With the above described kitchen in place, the Data Caterer becomes the customer-facing professional who turns protocol intent into decision-ready insight.

In practical terms, the Data Caterer ensures that data captured in Veeva is fit for purpose. They rely on the CDL to deliver harmonized, reusable data products with controlled versioning. They use the SCE to execute regulated pipelines and produce clear, traceable outputs.

An end-to-end scenario looks like this. Study intent is authored in a digital protocol approach and is versioned. Operational execution is configured in Veeva. Data is ingested into the CDL and is cataloged immediately. Harmonization and reconciliation occur through standardized transformation logic. Curated products are released with controlled versioning. The SCE consumes these products and produces regulated deliverables and decision packages. Throughout the flow, traceability is preserved from protocol intent to data product version to analysis output.

This is where the role shift becomes real. The Data Caterer spends less time chasing data provenance and more time ensuring that the result is decision-ready.

AUTOMATION: DATA FLOW AND AI

The architecture is intentionally designed to support automation and AI.

Automation helps industrialize routine steps across Veeva, the CDL, and the SCE. It improves consistency and traceability and reduces manual effort. AI adds a different capability. It can surface quality issues earlier, suggest derivations aligned with protocol intent, and support faster interpretation.

We use a simple operating principle. Automation is used for reliability. AI is used for foresight. Humans remain accountable for judgment, trade-offs, and customer-facing decisions.

In regulated environments, this only works with clear guardrails. We separate GxP and non-GxP contexts. We apply risk-appropriate validation to pipelines and models. We maintain audit trails and explainability. We enforce controlled access, including blinding and unblinding workflows.

PRACTICAL LESSONS LEARNED (EARLY OBSERVATIONS)

At the time of writing this paper, we are still in full implementation phase. Early implementation reinforced a set of pragmatic lessons:

1. Flow matters more than tools. If you cannot describe the end-to-end flow, you will optimize local steps and lose overall speed.
2. Governance must be designed in from day one. Lineage, versioning, and controlled sharing are product features, not afterthoughts.
3. Preparation and execution should be separated. The lake is for harmonization and reuse; regulated outputs require controlled execution.
4. Transformation libraries should be standardized early to avoid rebuilding complexity study by study.
5. The digital protocol is a multiplier. When study intent becomes machine-readable, the stations connect and automation becomes realistic. AI readiness then becomes less about model choice and more about controls: access, traceability, explainability, and change management.

CONCLUSION

Re-considering a clinical data technical landscape is not about choosing one tool. It is about building an ecosystem that aligns platforms, processes, and people around decision enablement.

In our approach, Veeva provides a governed operational backbone close to regulated execution. The CDL prepares and harmonizes data for reuse and analytics. The SCE provides controlled execution and delivery of regulatory-grade outputs. The digital protocol strengthens shared intent and connects the stations into a traceable pipeline. Open source adds flexibility where it is most valuable, especially upstream and in advanced analytics, under appropriate governance.

This is what makes the Data Caterer vision practical. When the kitchen is built well, clinical data is not only stored. It is curated and served in a way that helps stakeholders make faster and more confident decisions.

REFERENCES

- *Data is on the Menu! Are you still arguing about the oven? – From Coders to Data Caterers: Serving insight in an AI-Powered World,* Sascha Ahrweiler (Bayer AG), PHUSE SDE, Copenhagen, October 2025.

- *From Vision to Implementation: Building the Data Caterer's Kitchen at Bayer*, Sascha Ahrweiler (Bayer AG), Olivier Bouchard (SAS), PHUSE APAC Connect, Hyderabad, February 2026.

DISCLOSURE GENERATIVE ARTIFICIAL INTELLIGENCE USAGE

The authors wish to disclose that generative artificial intelligence (genAI) was used extensively in drafting this entire paper. Building upon initial outlines, detailed content for each section, and the overarching data caterer analogy provided by the authors, the comprehensive writing was conducted by generative AI tools. All genAI-generated output underwent careful review and thorough fact-checking by the authors, who assume full ownership and responsibility for the final content presented here.

The following AI tools were utilized:

- OpenAI (2026). ChatGPT-V5.2 (30 January version, Thinking mode) [Large Language Model].
- Microsoft (2026). MS365 Copilot GPT-5.2 (27 January version) [Large Language Model]

ACKNOWLEDGEMENTS

We would like to thank colleagues who contributed to the thinking and early implementation that informed this paper: Dirk Grosse, Konrad Nadobny, Robert Bergann, Bernd Heinrich, Astrid Scherer, Alexander Salzer, Holger Schimanski, René Wentzeck, and many more colleagues in the Bayer Data Sciences & Artifical Intelligence and other functions.

CONTACT INFORMATION (HEADER 1)

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

Author Name: Sascha Ahrweiler
Company: Bayer AG
Address: Aprather Weg 18a, 42113 Wuppertal, Germany
Email: sascha.ahrweiler@bayer.com

Co-Author Name: Holger Dach
Company: Bayer AG
Address: Aprather Weg 18a, 42113 Wuppertal, Germany
Email: holger.dach@bayer.com