What is an R&D Data Cloud?

R&D-focused, data-centric, cloud-native, and open

A Tetrascience Whitepaper

tetrascience

Table of Contents

R&D-focused, data-centric, cloud-native, and open

Biopharma R&D professionals are on a mission to accelerate discovery and improve human life — shortening time-to-market for new therapeutics. Data drives this mission — coming from diverse sources (instruments, software, collaborators) in a multiplicity of formats: rich with surface utility for analytics, automation, optimization; and full of undiscovered value for data science, big data analytics, and AI/ML enabled inquiry.

R&D data quantity, quality, and validity are now recognized as major competitive assets. A growing crowd of stakeholders — data scientists, tech transfer, procurement, operations, and other departments, plus external collaborators like CROs and CDMOs — clamor to turn this data into new discoveries, increased operational velocity (e.g., through automation), and other forms of value.

Life sciences are complicated

Biopharma R&D IT is finding that enabling this value-creation is much harder than it looks.

Huge roadblocks exist to making data findable, accessible, interoperable, and reusable (FAIR). For all enterprises, data is proliferating: growing geometrically in volume and velocity, and taking evermore-diverse forms, both structured and unstructured. Beyond this, biopharma R&D instruments and infrastructure, use cases, workflows, and requirements pose additional challenges, reflecting theirreducible complexity of doing science at scale. For example:

- Biopharma R&D has the largest number of instruments and software systems per scientist of any field of inquiry (Gartner), networked and non-networked, producing and consuming complex, diverse, and often proprietary file and data types
- A single small-molecule or biologics workflow (see illustration) can comprise dozens of sequential phases, each with many iterative steps that consume and produce data. As workflows proceed, they fork, reduplicate, and may transition among multiple organizations — different researchers, instruments, protocols
- Distributed research (e.g., collaboration with CROs and CDMOs) adds new data sources, formats, workflows, oversight and validation requirements

In the face of this, scientists and organizations adapt to minimize additional complexity. Data are manually extracted from instruments; saved in spreadsheets, .pdfs, and other file formats; dropped onto shared drives; transformed into reports and emailed; laboriously transcribed to Electronic Lab Notebooks (ELNs) and other tools of record.

Where domains of connectivity exist — for example, a fleet of instruments and their control software — they tend to be closed ecosystems — silos where data is aggregated for specific purposes rather than general utility.

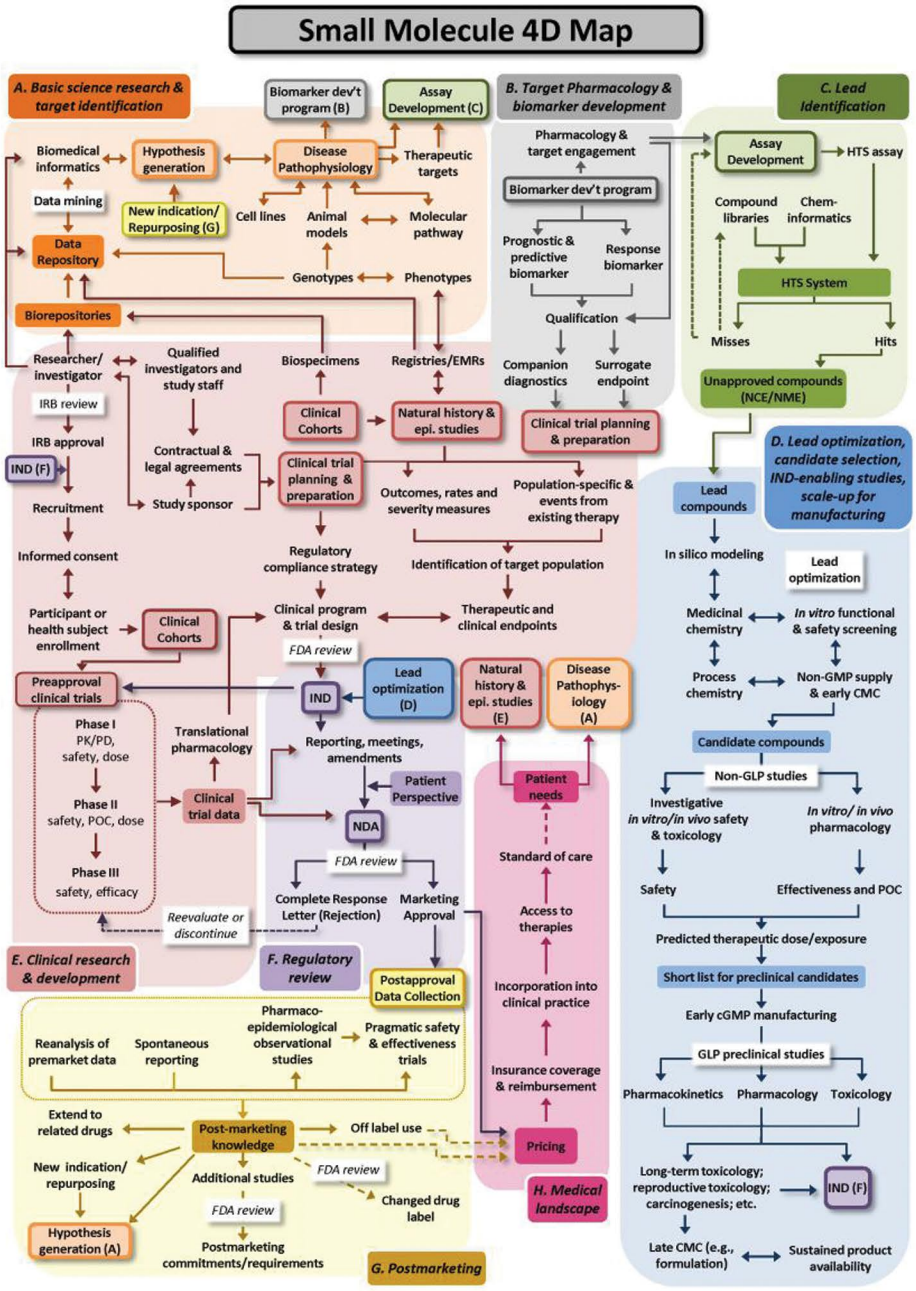

Figure 1: *Sample small molecule workflow, expressed as a 4D map. Wagner JA, Dahlem AM, Hudson LD, Terry SF, Altman RB, Gilliland CT, DeFeo C, and Austin CP. Drug Discovery, Development and Deployment Map (4DM): Small Molecules. Available at https://ncats.nih.gov/translation/maps. Last updated November 2017.* ***Download this map here.***

The result is wasted time and resources, and growing risk. Scientists spend significant percentages of working time transcribing, validating, and curating data, just to make the necessary connections and create the file artifacts needed to enable the work in front of them to proceed. Data scientists and data engineers spend even more time (double-digit percentages) finding and wrangling datasets for big data analytics, visualization, AI/ML training, and other applications. Manual transcription and ad-hoc, software enabled file transformations can introduce errors, make repeatability elusive, disrupt procedural transfer to collaborators, and damage regulatory compliance. As the cycle of data production and ad-hoc consumption continues, older data becomes less findable and accessible, sometimes forcing repetition of costly experiments.

Visualizing a real solution

Pushing back against chaos, biopharma R&D IT looks to ensure data's utility and long-term life cycle beyond the bench. But science-savvy developers quickly realize this is a tall order. Making data FAIR requires many steps:

- Accessing data where it resides — in networked and non-networked instruments, software systems, file-shares and elsewhere
- Parsing data robustly out of diverse file formats and representations
- Understanding, acquiring, and adding critical context (environmental, process-related, etc.) not saved with raw data
- Transforming and harmonizing data (plus metadata, tags, labels, and other decoration) into a unified hierarchy of (open) schemas that elucidate and preserve meaning, enabling search, comparison, analysis, and discovery
- Indexing and saving the data in an easily-consumed form (e.g., JSON), within a single, secure, highly resilient store that can work as "one source of truth"
- Distributed research (e.g., collaboration with CROs and CDMOs) adds new data sources, formats, workflows, oversight and validation requirements

On top of this, powerful functionality (and good user experience) are required to:

- Enable scientists, data scientists, data application builders, and other consumers to search, extract, and work with data easily
- Transform and synchronize data back, as needed, to LIMS, ELNs, and other systems of record, enabling scientists to continue working with their preferred/required/standard tools (but now with better-quality, richer data and many more ways to analyze it)

- Exchange data reliably with external collaborators like CROs and CDMOs
- Push data to automation, analytics, and other entities, tools, and applications

All while providing and enforcing security and access controls; making transactions auditable; enabling oversight, validation, and sign-off; and supporting other requirements of GxP, Title 21 CFR Part 11, and other relevant regulation.

The traditional approach: extended ad-hockery

Struggling to accelerate research by realizing at least some parts of this vision, R&D IT is faced with a range of (mostly bad) choices, each with significant costs and trade-offs.

Simplistic efforts to unblock and accelerate discovery — e.g., building point-to-point integrations between instruments, sensors, software, and other entities that produce data ('data sources') and applications that consume it ('data targets') — tends to be costly in both the short- and longer term, and may make for only superficial gains. Given the number and variety of sources, targets, connectivity and data-acquisition methods, parses, and required data transformations, any sizable R&D operation will require many such integrations. Of these, many will be complicated, fragile, and hard to maintain.

Under pressure to deliver tangible benefits quickly, integrations may be built to meet only simplistic goals: automating away manual parts of a process and/or solving for an immediate scientific use case. By omitting the necessary steps (adding metadata and other decoration, parsing, harmonizing, indexing, converting to consumable formats) required to make data FAIR, they may leave data in silos, or actually create new silos.

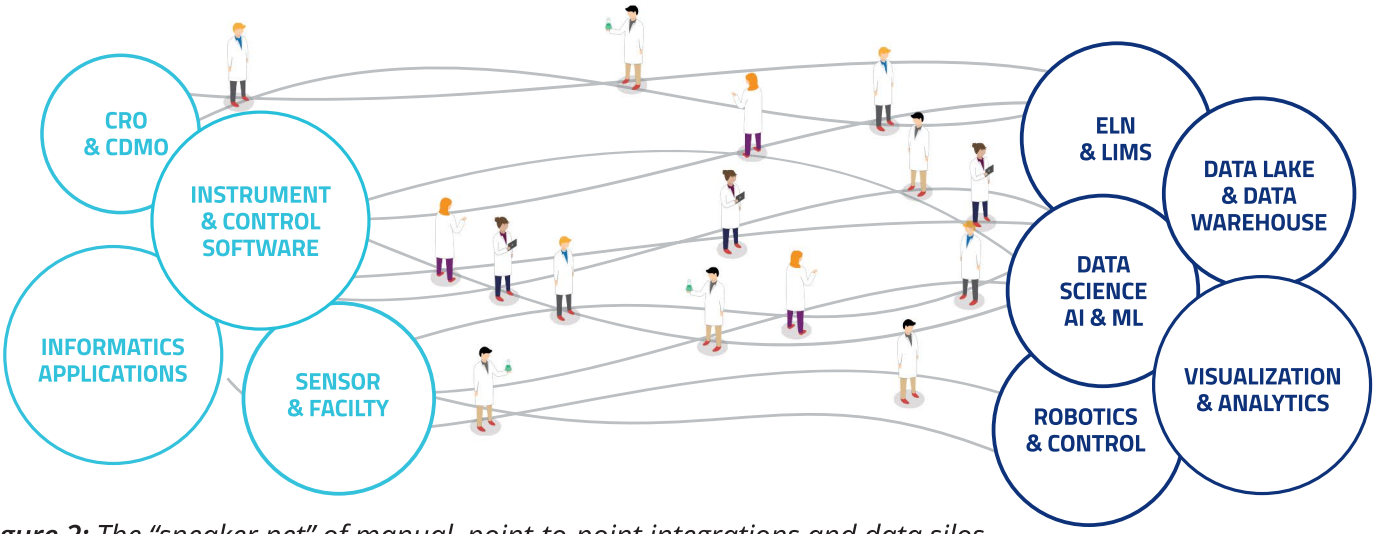

Figure 2: The "sneaker net" of manual, point-to-point integrations and data silos

More sophisticated efforts — for example, attempts to build data lakes or data warehouses from general-purpose building blocks — are risky, expensive, and prone to failure. Creating a "single source of data truth" is a best practice. But assembling a solution from components compels making heavy bets on each part of a solution stack, and committing to manage component interdependencies long-term. This tends to create non-strategic organizational spread — teams specialized around each component and around the challenges of running them at scale, in production. In practice, this may suck resources and attention away from the more important task of growing a data-centric IT organization, focused on using data to accelerate discovery and drive business value.

Building the core solution, too, is just the first step. Much more work is then required to create and maintain integrations with the huge and growing list of data sources and targets. And the team that builds the platform still owns this herculean task. Under pressure to deliver tangible benefits quickly, projects may veer away from ideal architectural goals (making data FAIR and building a single source of data truth) and towards delivering point solutions of limited general utility — effectively reproducing the drawbacks of point-to-point integrations on a new platform.

Efforts to press LIMS, ELN, SDMS, or other data-aggregating platforms into "single sources of data truth" may also fail to provide a complete or future-proofed solution. Designed to serve the needs of bench scientists and to support certain kinds of instruments and experimental workflows in compliant ways, these solutions are important sources and targets for data, but normally inadequate to serve broader needs of enriching and making data FAIR for the larger R&D enterprise. Larger R&D organizations may implement many of them, none owning more than a fraction of the global data pool; making many kinds of analysis difficult or impossible. They may not be cloud-native (see below), and may fail to satisfy resiliency, security, scalability, performance, and other technical requirements of a "single source of data truth." Their life cycles can be short — militating against use of this kind of solution as a long-term, canonical data repository. And the way they ingest and process data may not take all (or any) of the steps required to make data FAIR.

Cloud raises the stakes

All the above solutions, meanwhile, need to be aligned with biopharma IT's overarching goal of replatforming to the cloud. A future-proofed, mission-critical R&D data processing substrate — engineered to support big data and other science- and profit-accelerating programs — seems like a classic marquee project for delivering cloud's promised benefits:

- Redirecting spending from CapEx to OpEx
- Liberating growth, optimizing utilization, and controlling costs by leveraging elastic compute and storage

- Providing access to data from anywhere
- Empowering rapid development of new applications (including self-service data application development, perhaps by scientists themselves) using powerful public cloud services (e.g., ML-as-aservice, Hadoop-as-a-service, etc.)

But this is hard to do well. Replatforming to the cloud requires new skills and tools — for automation, monitoring, and other tasks — leading to non-strategic organizational spread for R&D-focused IT teams. Cost management in cloud environments is complicated, heavily conditioned on traffic and other hard-to-predict variables. Benefits can be elusive. Cloud bills are often larger than expected.

The biggest problem, however: merely replatforming to cloud doesn't get you closer to the optimal vision of an R&D data cloud, described above, because:

- A real R&D Data Cloud can't be efficiently realized as a conventional monolithic or tiered application architecture
- Cloud storage is only one component of the solution, and does not solve the big, combinatorial problem of connecting data sources and targets, breaking data out of silos, and making it FAIR

More on this, below.

Introducing the R&D Data Cloud

An R&D Data Cloud is purpose-engineered to solve the whole biopharmaceutical research and development data management puzzle, end-to-end. It can only be the product of *an organization with deep expertise in technology and life sciences; also committed to building and leveraging a broad, inclusive, and open network of partnerships across the R&D technology landscape.*

This combination of characteristics is required to engineer integrations that:

- Solve the whole problem of breaking data out of silos, enriching it with context, parsing it out of raw forms and into schemas that make it FAIR and enable value-extraction
- Leverage a full-stack product architecture and robust common functionality to ensure software quality / reliability, simplified integration lifecycle management, resource efficiency, and high performance
- That can be maintained and updated to ensure continuous compatibility with fast-evolving instruments, software, and other data sources and application targets
- Plug-and-play, and just work

We call such integrations "productized." They're much more complex than what most solution providers talk about as "integrations." Ingesting R&D data from innumerable sources isn't trivial, but it's also not the whole story. Conventional, point-to-point integrations often do the minimum required to make a data source/target connection work (and tend to be fragile despite doing less). But the data they process (and store) may remain inaccessible and/or unfindable, and may be stripped of context, limiting its long-term utility. General-purpose, industry-agnostic cloud storage solutions, likewise, can't generally supply the scientific and biopharma R&D insight, or leverage life sciences partnerships, required to create a library of productized integrations, or efficiently build and maintain new ones.

Doing the whole job: making data FAIR and storing it accessibly on the way to transforming and serving it to target applications — requires a modular, event-driven, distributed pipeline architecture that parses raw data, acquires metadata and other decoration, and transforms enriched datasets into fully-open, consistent intermediate data schemas (see below). This requires an organization that's mastered FAIR data processes, and can:

- Assemble the technical and scientific expertise to architect and extend a system of Intermediate Data Schemas (IDS) for use in data harmonization
- Devise an optimal modular, event-driven pipeline architecture for data ingestion, harmonization, and schematization
- Create and staff a mature process for developing, testing, and productizing complete integrations, including reverse-engineering and forensics
- Build and maintain partner relationships that can be drawn on to accelerate development of the integrations library: for knowledge, collaboration, validation, and for keeping integrations in synch with instrument, software, and application releases over time
- Leverage broad scientific knowledge and customer community interaction to master scientific and R&D process needs, determine and acquire critical metadata and decoration required to enrich raw data and place it in context, and solve around specific customer requirements

Combining all these superpowers lets a true R&D Data Cloud vendor create integrations that can be installed, configured, and just work — shortening time-to-value from months or years to weeks, or even days.

Compliance engineered into the solution. An R&D-focused Data Cloud must be engineered to enable and facilitate compliance with relevant regulatory requirements, doing so across the whole data life cycle. That means the solution provider needs to fully understand implications of 21 CFR part 11 and GXP and know how to deliver the assurances required under these regulations. The solution itself — in all its parts — needs to establish context

and chain of trust for data and maintain this end-to-end, including when data are processed, stored on, and retrieved from abstract cloud services. Doing this correctly can reduce the burden of data validation, speed audits, and reduce or eliminate risk and bottlenecks as data is drawn upon in the course of advancing candidate therapeutics from early research to development, preclinical, and manufacturing stages.

Data-centric

As discussed above, most strategies and systems proposed for solving the R&D data challenge tend to draw resources and attention towards non-strategic labor: just connecting things together, and/or just storing data, in or out of clouds.

By contrast, a true R&D Data Cloud treats data as the central concern, takes ownership, and manages data through its entire life cycle (which turns out to be a better, faster, more reliable way both of connecting things together and managing cloud data storage).

A true R&D Data Cloud pivots around a comprehensive Intermediate Data Schema (IDS). The IDS provides an extensible reference model, enabling harmonization and alignment of all exemplars of each kind of data entering the system, plus a framework for relating them together. Once raw data is parsed, metadata and other decoration is added, and datasets are schematized, data becomes self-documenting, searchable in new ways, and comparable. This is essential for reducing the enormous burden of manual data wrangling and curation on scientists, data scientists, and other data consumers. Equally essential to smoothly using data for analytics, AI/ML, automation, and innovation.

Cloud-native

An R&D Data Cloud can't easily be built, maintained, made resilient, or scaled up as a conventional monolithic or tiered application. Each source-to-storage or storage-to-target pipeline can require a different sequence of components, making different demands on compute, network, local storage, and other resources. Pipelines must scale collectively, and potentially at the component level, to handle loads and (where practical) continue processing data in close-to-realtime.

Conventional application architectures won't work. This is hard to do efficiently with a conventional computing architecture based on processes running on virtual machines, or even on container workloads. The number of entities, and their dynamic requirements, are variable enough to make utilization inefficient (read: you pay to run lots of servers and workloads that end up sitting around, waiting for work), orchestration challenging (read: it's

very hard to scale pipelines up and down, configure and deploy new pipelines efficiently), and housekeeping operations painful and risky (read: you may not be able to update an individual module without taking down its pipeline).

A true R&D Data Cloud requires more modularity and more deeply-automated orchestration and dependency management. It requires an event-driven microservices architecture that's substantially self-managing, self-scaling, and that simplifies the job of creating new pipeline modules (i.e., delivering value around data) by abstracting away other concerns (e.g., dependencies) that make conventional software development and DevOps complicated.

A full-stack solution is required to make data FAIR while minimizing operations toil and controlling cloud costs. Beyond this core architecture, a real R&D Data Cloud needs to fully and efficiently leverage advanced cloud compute, storage, network, observability, and other services to deliver more of the promised benefits of cloud technology, out-of-the-box.

Correct integration of these services minimizes day-to-day operational demands on R&D IT, minimizing non-strategic organizational spread, and freeing technologists to focus time, attention, skill, and budget on using and innovating around the data; gleaning insights from the data, accelerating discovery, and building business value.

Open

An R&D Data Cloud needs to work as the canonical substrate and single source of data truth for interconnecting and enabling (in principle) all the tools, platforms, applications, and workflows of a life sciences R&D operation. This means the organization providing such a solution needs to be:

Dedicated to an open vision of data and its value, and scrupulously vendor-agnostic — enabling open collaboration across a broad ecosystem of partners and community of users to share knowledge and produce value. An R&D Data Cloud needs to help break down "walled gardens," making data accessible. It needs to provide simple, well-documented, open APIs permitting easy access to all the data in its purview — it can't lock data away or leverage a business model that depends, to any degree, on making access to your data technically difficult or expensive or otherwise promoting "lock in." It also needs to stay in its lane, providing a data-centric substrate enhancing use of powerful, specialized tools (ELN, LIMS, others) by providing gold-standard data to all potential consumers.

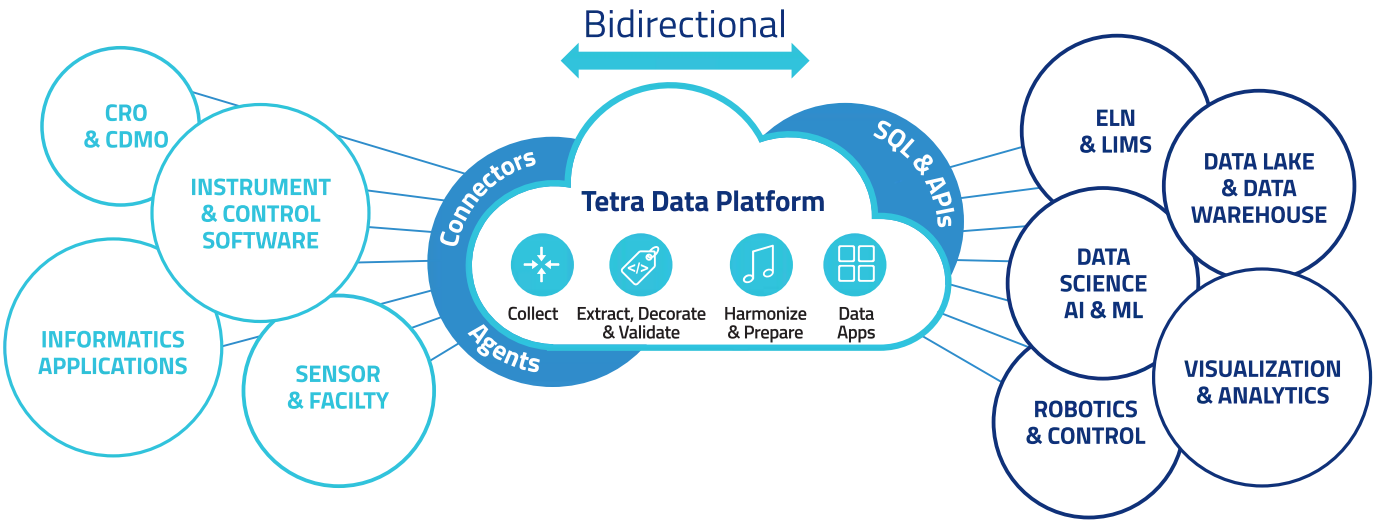

Figure 3: *The Tetra Data Platform (TDP), the technology foundation of an R&D Data Cloud*

Conclusion

Organizations seeking to derive maximum long-term value from R&D data should assess whether their proposed data management strategies are indeed data-centric, R&D-focused, cloud-native, and open. They should think clearly about the value of building out teams to architect, code, integrate, maintain, and operate such a solution — likely to dominate the organization's attention and inflect its collective risk profile for years. Alternatively, they should consider whether these same resources are better repurposed: initially to implement an off-the-shelf solution and quickly provisioning high-priority integrations, then extending the solution's functionality and support for scientific and business innovation.

For more information on how the Tetra R&D Data Cloud accelerates discovery in life sciences R&D, visit tetrascience.com.