tetrascience

The 8 trends redefining scientific data management in biopharma

WHITE PAPER

Table of Contents

Executive Summary

The biopharmaceutical industry is experiencing a seismic shift in how scientific data is managed and analyzed. This transformation is being driven by emerging technologies like AI, which biopharma leaders are embracing to dramatically improve innovation, lower costs, and accelerate drug development. To achieve this goal, organizations must fully leverage their scientific data. These datasets are some of the largest, fastest growing, and most valuable in the world. However, they are typically trapped in silos and incompatible formats that severely limit their utility.

To overcome these challenges and unlock the full value of scientific data, the industry is undergoing several key shifts:

1. On-premises → cloud-native data management
2. File-based → engineered data
3. In-line → event-driven data processing
4. Disjointed analysis → co-located cloud-based data apps
5. Endpoint-specific SDMS → endpoint-agnostic data cloud
6. Single-company → multi-company collaboration
7. IT project → data product
8. File management as an end state → an onramp to analytics, ML, and AI

Biopharma organizations that fail to evolve their data strategies risk falling behind competitors. Those that adapt quickly will be positioned to fully capitalize on Scientific AI and thus maintain a competitive edge.

Introduction

Over the past few years, there has been a profound evolution in the ways scientific data is generated, transferred, analyzed, and stored. These advancements reflect broader macro-trends that transcend individual organizations and impact the entire biopharmaceutical industry. Despite the significant progress made so far, we stand at the cusp of a radical transformation in scientific data management and analysis.

This transformation is catalyzed by the need to harness the full potential of data through advanced technologies such as cloud computing, machine learning (ML), and artificial intelligence (AI). It could impact the entire value chain for therapeutics, unlocking billions of dollars through faster development times, reduced costs, and the discovery of more blockbuster drugs. The changes underway are poised to redefine how biopharma organizations use scientific data and derive value from it.

Challenges of scientific data in biopharma

To understand the reasons for this shift, we first need to examine the industry's core asset: scientific data. This data is dispersed throughout a biopharma organization, generated by numerous groups across diverse locations. Many scientific teams use a multitude of instruments in various labs, floors, buildings, and sites worldwide. Plus, biopharma companies increasingly depend on a network of external partners, such as contract research organizations (CROs) and contract development and manufacturing organizations (CDMOs), which generate a significant portion of their data.

When we account for all scientific groups within an organization and its collaborators—including those involved in discovery, development, and quality control—we find thousands of locations where data files are stored, locked away in a plethora of silos. This fragmentation adds to the inherent complexity of scientific datasets and presents significant challenges in managing, accessing, and utilizing them effectively.

Below, we explore several factors that contribute to the challenges of managing scientific data.

INHERENT COMPLEXITY OF SCIENTIFIC DATA

Scientific data is complex due to its diverse sources and formats. Numerous instruments, each producing data in proprietary formats with limited interoperability, contribute to this complexity. For instance, lab instruments like spectrometers, plate readers, liquid chromatography systems, and electron microscopes generate data in formats unique to their manufacturers. This results in a heterogeneous mix of data formats that are challenging to integrate and analyze cohesively.

Additionally, this issue is compounded by the various types of scientific data generated throughout experimental workflows. These include time-series data, multi-dimensional numerical data, long-term study data, unstructured text, audit trails, and proprietary file formats. Each type requires specific expertise to interpret and integrate effectively. Consequently, robust data management practices are essential to ensure that data remains usable and valuable.

COMPLEX SCIENTIFIC WORKFLOWS

Scientific workflows involve many steps that generate data. These include experimental design (recorded in electronic lab notebooks), sample preparation, sample testing, analysis of results, and report creation. Each of these steps adds to the overall complexity of scientific data.

FRAGMENTED SCIENTIFIC VENDOR ECOSYSTEM

The vendor landscape in the life sciences is highly fragmented, with limited integration, standardization, and collaboration. Instruments and software applications from different vendors often cannot exchange or read each other's data, leading to data silos.

VENDOR LOCK-IN

Vendor lock-in further aggravates the situation. Vendors often lack the motivation to be compatible with each other. Instead, they are usually incentivized to do the opposite—trap users within their own ecosystems using proprietary data formats and limited export capabilities. These practices restrict the freedom to use the best tools on the market and create significant barriers to data integration and analysis. The lack of standardized data formats and interoperability hampers the ability to aggregate, contextualize, and harmonize scientific data across lab instruments and applications.

EXTERNAL COLLABORATION

External collaboration and outsourcing add another layer of complexity. Biopharma companies typically work with contract organizations for various research and development activities. However, contract organizations often use general-purpose tools like email and file-sharing platforms to share results with their clients, usually in PDF files and spreadsheets. These technologies fall short of meeting the needs of biopharma scientists: they fail to capture the full context of scientific data, make it difficult to extract and reuse data, and provide little to no data traceability.

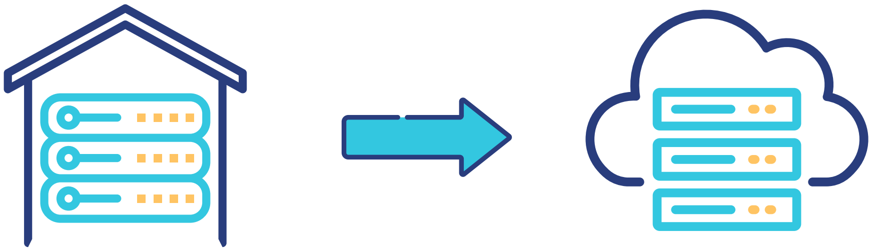
CULTURAL HABITS OF SCIENTISTS

Under tight deadlines, scientists may prioritize speed over rigor, leading to documentation delays and incomplete records. Informal or selective recordkeeping can jeopardize data integrity and result in the loss of important experimental context. This practice also leads to inconsistent or missing metadata among individuals and groups. Manual-intensive workflows in non-regulated environments are particularly susceptible to these problems.

Paradigm shifts in scientific data management

Companies around the world are waking up to a realization: they have been mishandling and underutilizing their scientific data assets. Data connectivity and automation are not enough. Increasingly, they're moving beyond outdated data management strategies to establish a foundation for analytics and Scientific AI. As a result, there is a paradigm shift underway in scientific data management. This shift is catalyzed by the need to effectively leverage scientific data with advanced analytics and AI. It's independent of any individual company. And it's only just beginning. Let's examine the many benefits of this new paradigm.

1. FROM ON-PREMISES TO CLOUD-NATIVE DATA MANAGEMENT

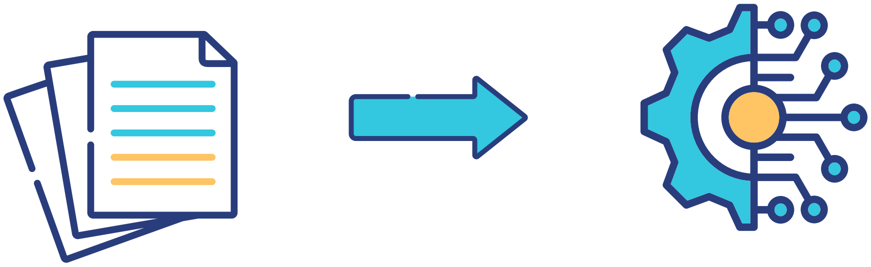


The transition from on-premises, siloed infrastructures to cloud-native services has transformed data management. A cloud-native solution offers a lower total cost of ownership, better scalability, and support for the high workloads that advanced analytics, ML, and AI demand. This shift is propelled by advancements in AI and ML by leading tech companies, further enhancing the capabilities of scientific data analysis.

Cloud-native data management provides significant flexibility. Traditional on-premises systems require substantial investment in hardware, software, and maintenance. In contrast, cloud solutions operate on a pay-as-you-go model, reducing upfront costs and allowing organizations to scale resources according to their needs. This scalability is crucial for handling the large volumes of data generated by modern scientific activities.

Additionally, cloud-native systems enable better data sharing across geographically dispersed teams. Scientists can access data and tools from anywhere in the world, facilitating global collaborations. Moreover, cloud providers offer robust security measures, ensuring that sensitive scientific data is protected against unauthorized access and breaches.

2. FROM FILE-BASED TO ENGINEERED DATA



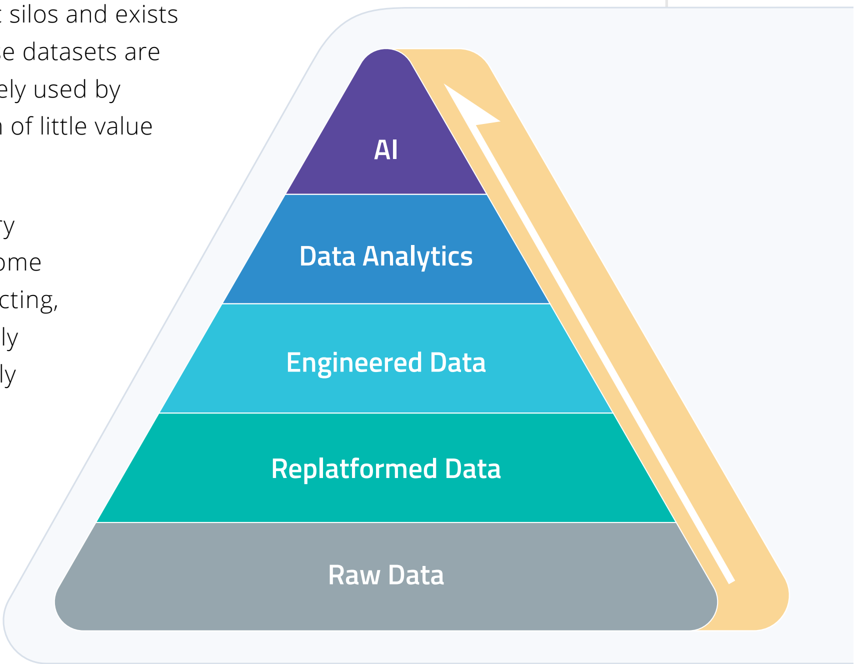
Files segment your data in containers that limit access. Scientists are not interested in the file per se but rather the information it contains. Data that is purpose-engineered for science includes critical scientific context, uses multimodal taxonomies and ontologies, and can be reused across scientific applications.

Organizations are moving away from a file-oriented mindset to a data-centric approach, especially when dealing with multiple data types—unstructured, semi-structured, and structured. Their focus is now on the valuable information that the file holds. This requires understanding table structures, schemas, taxonomies, and ontologies.

This paradigm shift can be challenging because many scientists are accustomed to viewing files as the primary means of data storage and transfer. For example, CROs and CDMOs usually convert data into files (e.g., PDFs or spreadsheets) and email them to scientists, who extract and interpret the information. While files remain important and will continue to be used, transitioning to a data-centric approach is essential.

Engineered data transcends the limitations of traditional file-based systems. By incorporating scientific context and metadata, it becomes easier to integrate and analyze data from diverse sources. This approach allows stakeholders to combine multimodal datasets and derive deeper insights. Large-scale, engineered data is a prerequisite for training and using AI models. Raw scientific data, no matter its volume, is unsuitable for AI applications. It must undergo an immutable series of operations to become AI ready (see box).

The scientific data journey



Raw data often resides in vendor-specific silos and exists in proprietary, unstructured formats. These datasets are typically too small and static to be effectively used by advanced analytics and AI, rendering them of little value to biopharma organizations.

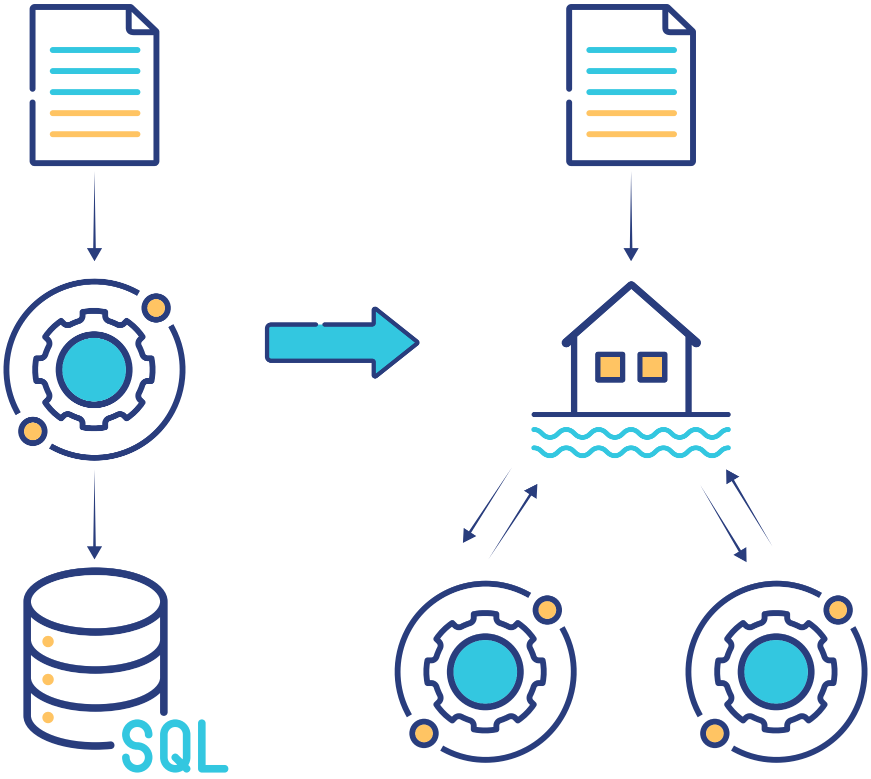
Data replatforming is the first necessary step to improve datasets. Data must become compliant, liquid, and accessible. By collecting, centralizing, and contextualizing previously siloed data in the cloud, scientists can fully leverage modern cloud infrastructure. They gain unprecedented elasticity for storing, processing, easily retrieving, and accessing their data.

Data engineering requires deep scientific, data, and technology expertise to fully convert scientific data into liquid, large-scale, harmonized, and compliant data with industrialized scientific taxonomies and ontologies. Only these open, vendor-agnostic, AI-enabled datasets can serve as the fundamental building blocks for Scientific AI.

Data analytics is a key avenue for extracting value from scientific data. Although numerous vendors and tools can help organizations leverage their data, fragmented data models will result in massively suboptimal outcomes. Organizations need to engineer their scientific data to optimize it for their preferred dashboards, visualization tools, and analytics applications.

AI can help answer the most pressing questions in science. But organizations often lack internal resources or skills to put the right data strategies into place to capitalize on cutting-edge AI applications and collaborative workflows. Only purpose-engineered, liquid data can enable groundbreaking Scientific AI–driven outcomes across the value chain.

3. FROM IN-LINE TO EVENT-DRIVEN DATA PROCESSING



There is a growing recognition that simple point-to-point integrations are inadequate. They follow linear, predefined workflows that extract only a subset of information from files and load it into SQL databases or data warehouses. This in-line approach to data processing does not easily adapt to new use cases. When changes are necessary, they often require extensive reengineering of the entire workflow, making data management more complex and inefficient.

In contrast, event-driven data processing offers greater flexibility and customization. In this model, data is initially centralized in a cloud solution like Amazon S3, which is cost-effective due to intelligent tiering and archival options. Data workflows are then executed in response to specific events or triggers. By using tools like AWS Lambda or event buses, data processing becomes dynamic and responsive. Data consumers can subscribe to published data streams, enabling them to iteratively explore and extract value from the data, even if they are initially unsure of their exact needs.

The lakehouse architecture complements the event-driven approach by providing a unified storage system. It merges the scalability and flexibility of data lakes with the performance and reliability of data warehouses. By combining a lakehouse architecture with event-driven data processing, organizations can achieve a more dynamic, efficient, and responsive data management system. Data consumers can uncover new insights and applications as they engage with the data, leading to more successful implementations without the need for constant reformatting or downsampling.

4. FROM DISJOINTED ANALYSIS TO CO-LOCATED CLOUD-BASED DATA APPS

Traditionally, scientists analyze data by transferring it from laboratory instruments to analysis software. This manual process involves multiple data stores, including thumb drives, leading to serious concerns about regulatory compliance. As a result, data workflows become slow, cumbersome, error-prone, and insecure.

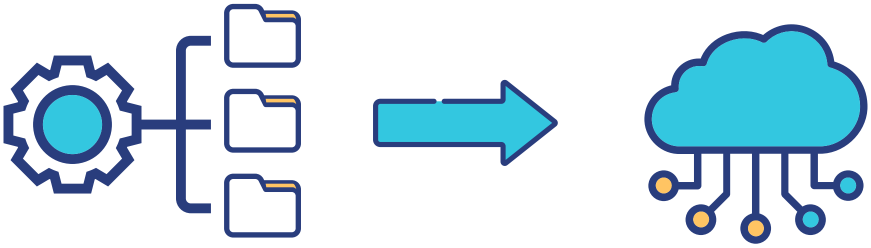
The cloud has greatly improved the situation. By assembling all data in one place, scientists can easily retrieve data, upload it into their preferred tools, and share their work with project stakeholders. Better yet is bringing the analytical software into the cloud where the data is co-located. This results in seamless access to data, reduced errors, better collaboration, and increased scientific productivity.

Consider a common scenario where project data is centralized in the cloud. Scientists often need to download files onto their laptops and then import them into various analysis tools, such as HALO for imaging, Skyline for mass spectrometry, and FlowJo for flow cytometry. They frequently express frustration about finding data in the cloud and downloading it instead of accessing it directly from familiar folder structures on their instruments. Centralizing the data in this manner doesn't immediately streamline their analysis workflow.

The industry requires a model where data and analysis applications are co-located in the cloud. This approach, similar to how Google Drive allows users to open and edit files directly within the platform, enables scientists to work more efficiently. Tools like Google Colab and Jupyter Notebook exemplify this trend, providing integrated environments for data access and analysis.

However, not all applications are designed for co-location, and licensing issues can pose additional challenges. Nonetheless, the co-location of data and applications is a significant trend that several successful companies are adopting.

5. FROM ENDPOINT-SPECIFIC SDMS TO ENDPOINT-AGNOSTIC DATA CLOUD



Over the past two decades, nearly every instrument or informatics application vendor has either acquired or developed its own data management solutions. This has become a competitive necessity in a fierce market. Unfortunately, each vendor's scientific data management system (SDMS) contributes to a fragmented landscape of data silos. Endpoint-specific SDMSs are a dead end for scientific data, severely limiting the data's value (see box).

In response, organizations are adopting cloud solutions that are independent of the endpoint. These solutions provide large-scale, liquid, and FAIR (findable, accessible, interoperable, reusable) datasets, forming the foundation for AI and advanced analytics applications. Vendor-agnostic architectures eliminate conflicts of interest and enable seamless data flows between the data cloud and various endpoints. Consequently, scientists can use the best instruments and software available, regardless of the vendor, to advance their work.

A potential concern with this approach is that the data cloud itself could become another silo. To address this, it must be part of a data mesh. A data mesh enables organizations to create a decentralized data architecture that treats data as a product managed by domain-specific teams. This ensures that data is not confined to specific endpoints but is accessible and usable across the entire organization. Companies like Databricks are leading this transformation, enabling data from different nodes to converge into a unified, horizontal consumption layer.

Why a traditional SDMS is not an option anymore

Traditional SDMSs were designed to store and archive data for regulatory compliance, not to prepare data for AI applications. They might be adequate for collecting instrument and application data, cataloging data by adding some metadata, and archiving data in a compliant manner. But most traditional SDMSs have serious limitations in supporting AI initiatives.

Inflexible data flow: Traditional SDMSs have few options for data flow and processing. For example, they might be unable to send data to multiple destinations. If they can't provide the flexible data liquidity required by biopharma teams, they become a data graveyard.

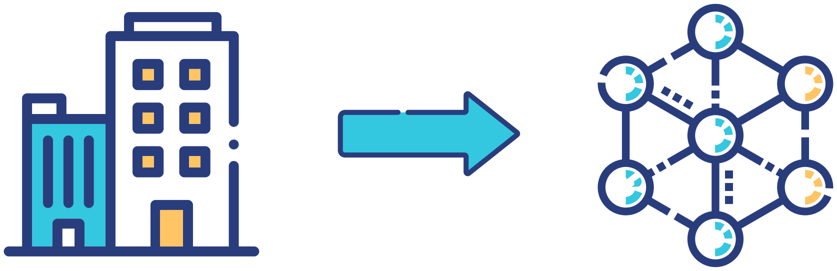
Little data engineering: SDMSs are designed to store data but not transform it. Traditional SDMSs don't attempt to engineer data for scientific use cases. They don't produce data in a standardized, harmonized, future-proof format that is engineered specifically for data science, analytics, or AI.

Poor discoverability: SDMSs might add metadata to files, but because they don't typically harmonize metadata taxonomies and ontologies, they can make it difficult for scientists to discover new or historical datasets. Data is searchable and consumable only if someone knows precisely what terms or labels to query. In many cases, lab scientists end up re-running an assay or an experiment because that's easier than finding historical data.

Lack of scalability: On-premises SDMSs cannot be scaled easily or cost effectively. Each upgrade requires multiple changes, including upgrades for the database, servers, and file storage. If SDMSs employ cloud services at all, they often use the cloud as another data center. Consequently, SDMSs are not the best environment for assembling the large-scale datasets required for AI.

SDMSs simply aren't designed to prepare data for AI. Some SDMS vendors might tack on capabilities to address deficiencies. But in general these legacy solutions cannot provide sufficient data liquidity, allow adequate searchability, enable data accessibility, or efficiently scale up to support the massive volumes of AI-native data needed for AI models.

6. FROM SINGLE-COMPANY TO MULTI-COMPANY COLLABORATION



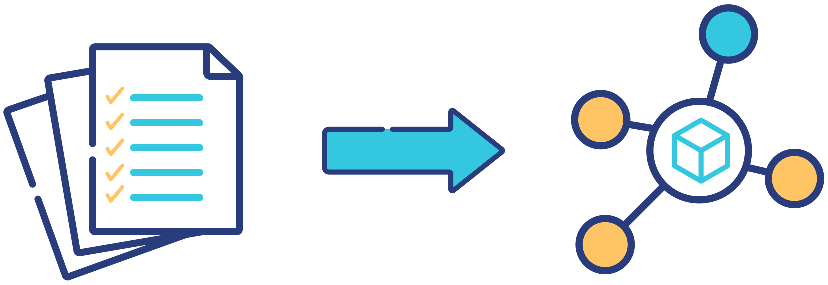
Five years ago, industry conversations centered on “my company’s data.” However, outsourcing has since advanced to the next level. With the rise of external collaboration, data management strategies must now support data sharing between companies. Despite this need, organizational boundaries—such as security protocols, data ontologies, and operational cadence—often restrict data flow, limiting the potential for deeper collaborative insights.

To this day, many contract organizations email PDF reports or spreadsheets to biopharma clients. This approach is no longer sufficient. The new paradigm requires a multi-company data architecture, enabling seamless collaboration across an ecosystem of partners.

Data companies like Databricks excel in this area, designing architectures that prioritize collaboration and data sharing. Their reference architectures, which include features like Delta Sharing, facilitate secure and efficient data exchange across different accounts.

However, implementing this collaborative data model is challenging due to organizational boundaries. Success often starts within divisions of a biopharma company or among acquired companies. Once refined internally, these data-sharing practices can be extended to external partners. This fosters an ecosystem where liquid data with common ontologies can be securely exchanged.

7. FROM IT PROJECT TO DATA PRODUCT



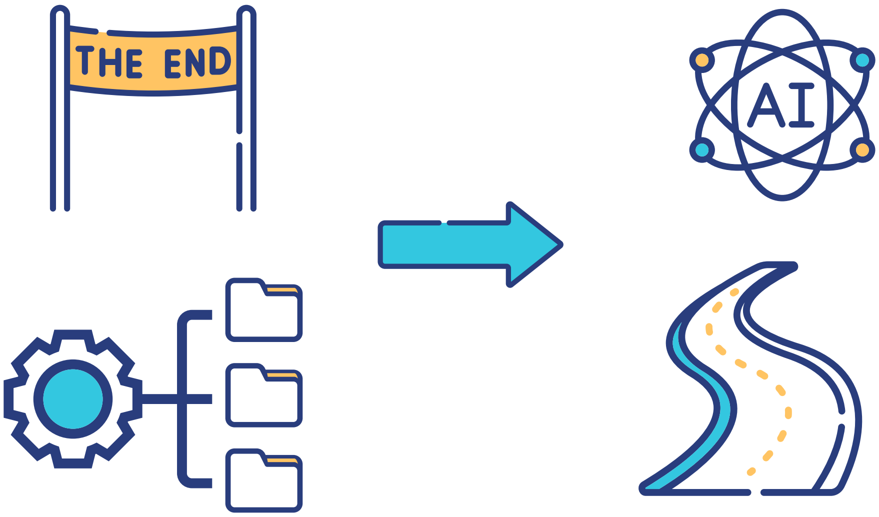
Data integration projects often prioritize technology over results. They focus more on the technological "plumbing" than on the scientific use cases that emerge. A better strategy is to prioritize the data product itself. By involving scientific stakeholders early in the planning and execution phases, product owners can better align with desired outcomes and unlock new scientific use cases.

Historically, data projects have been led by R&D IT teams, business IT teams, and informatics teams, following a traditional project management mindset with a clear beginning and end. These projects follow predefined charters and plans, aiming to achieve specific goals without deviation. However, this approach is becoming less effective as the paradigm shifts toward treating data as a product.

Adopting a data product mindset changes how organizations operate. It emphasizes continuous evolution rather than a fixed start and end. This shift necessitates identifying stakeholders—those producing, governing, and using the data—and involving them throughout the process. Like any product, a data product starts simple and becomes more complex over time, adapting to new needs and opportunities.

By thinking of data as a product, organizations become more attuned to both their goals and those of their collaborators. This approach promotes ongoing improvement and ensures that the focus remains on leveraging data to drive scientific and business outcomes.

8. FROM FILE MANAGEMENT AS AN END STATE TO AN ONRAMP TO ANALYTICS, ML, AND AI



Scientific data management should not be viewed as the final destination for your data. Instead of focusing solely on consolidating all your data, the emphasis should be on what you want to achieve with it. How will the data be used today? What other insights might you need tomorrow? Scientific data management is just a starting point.

In the past, data management was often treated as an end goal. Success was measured by the amount of data consolidated, the number of terabytes or petabytes stored, or the number of instruments connected. There was much less focus on how the data would be used.

The paradigm is shifting. Data aggregation and integration are now seen only as entry points toward achieving broader goals in AI and analytics. This shift requires a change in mindset, recognizing that data management is part of a larger journey—one that uses a scientific data and AI cloud as the foundation and onramp to analytics and AI.

One challenge in this shift is that different teams handle various stages of the data journey. At the lowest tier, lab IT personnel may be responsible for systems like an electronic lab notebook (ELN), while at the highest tier, data teams focus on analytics and AI. These teams often do not interact with each other. Shortsighted decisions made by the IT team, such as using middleware to connect lab instrument software to an ELN, can hamstring the data team (see box).

To ensure success in this new paradigm, every step of the data journey must involve communication and collaboration among all teams, including IT, science, and data science. Ideally, the organizational structure should include a product owner or an analyst who oversees the entire process, ensuring continuity and alignment of goals from data collection to analytics and AI implementation. This integrated approach will yield greater success and unlock the full potential of your data.

The middleware approach is a dead end

Middleware is purpose-built software that acts as an intermediary layer to transfer scientific data between instruments and applications. Take, for example, middleware used to input results from a chromatography data system (CDS) into a laboratory information management system (LIMS) or ELN. Typically, it extracts a only limited set of data from the CDS—such as peak area, height, and analyte name—which are needed for a LIMS report or ELN schema.

However, this middleware approach is fundamentally flawed. It extracts data solely based on the narrow requirements of a single endpoint like a LIMS or ELN at the time of creation. These applications are designed for scientific workflow tracking, experiment recording, and report generation—not for storing all the data.

Although middleware may seem convenient for point-to-point integrations, it does not align with a data-centric strategy. Instead, it serves the narrow needs of a specific application rather than treating data as a reusable asset. As a result, data teams must spend significant time and resources reworking data for analytics or AI.

By taking a middleware approach, organizations create a dead end for their scientific data. Only a fraction of the information and context is captured in the process, making these datasets far from future-proof and unsuitable in quantity or quality for advanced analytics and AI.

Conclusion

The landscape of scientific data management and analysis is undergoing a remarkable transformation. Biopharma organizations are recognizing the need to extract maximum value from their scientific data. AI holds enormous promise in this regard: it has the potential to bring safer, more effective drugs to market faster and at lower costs.

To realize the benefits of Scientific AI, organizations must overcome the inherent challenges of integrating, engineering, and analyzing one of the world's largest and most complex datasets. Scientific data remains widely dispersed and siloed within organizations and across partner ecosystems. To make matters worse, many instrument and software vendors use proprietary data formats and practices that trap customer data within their walled gardens.

Breaking free from these limitations requires biopharma organizations to move away from traditional data management practices and embrace modern, integrated, and collaborative frameworks. Key to this evolution are the eight paradigm shifts discussed above.

The shift in scientific data management and analysis is not merely a technological evolution. It represents a fundamental rethinking toward data centricity and how data can be used to drive better scientific and business outcomes. We're just at the beginning of this transformation, but it will likely accelerate quickly.

Biopharma organizations that fail to adopt these new paradigms risk falling behind their competitors, with the gap only widening with time. In contrast, those that embrace these changes will be able to harness the full power of their scientific data, maintaining or even increasing a competitive edge.

Learn how TetraScience can help you adopt the new paradigms in scientific data management and analysis.

Visit Tetrascience.com

tetrascience

Corporate Headquarters | 177 Huntington Avenue, Suite 1703, Boston, MA 02115