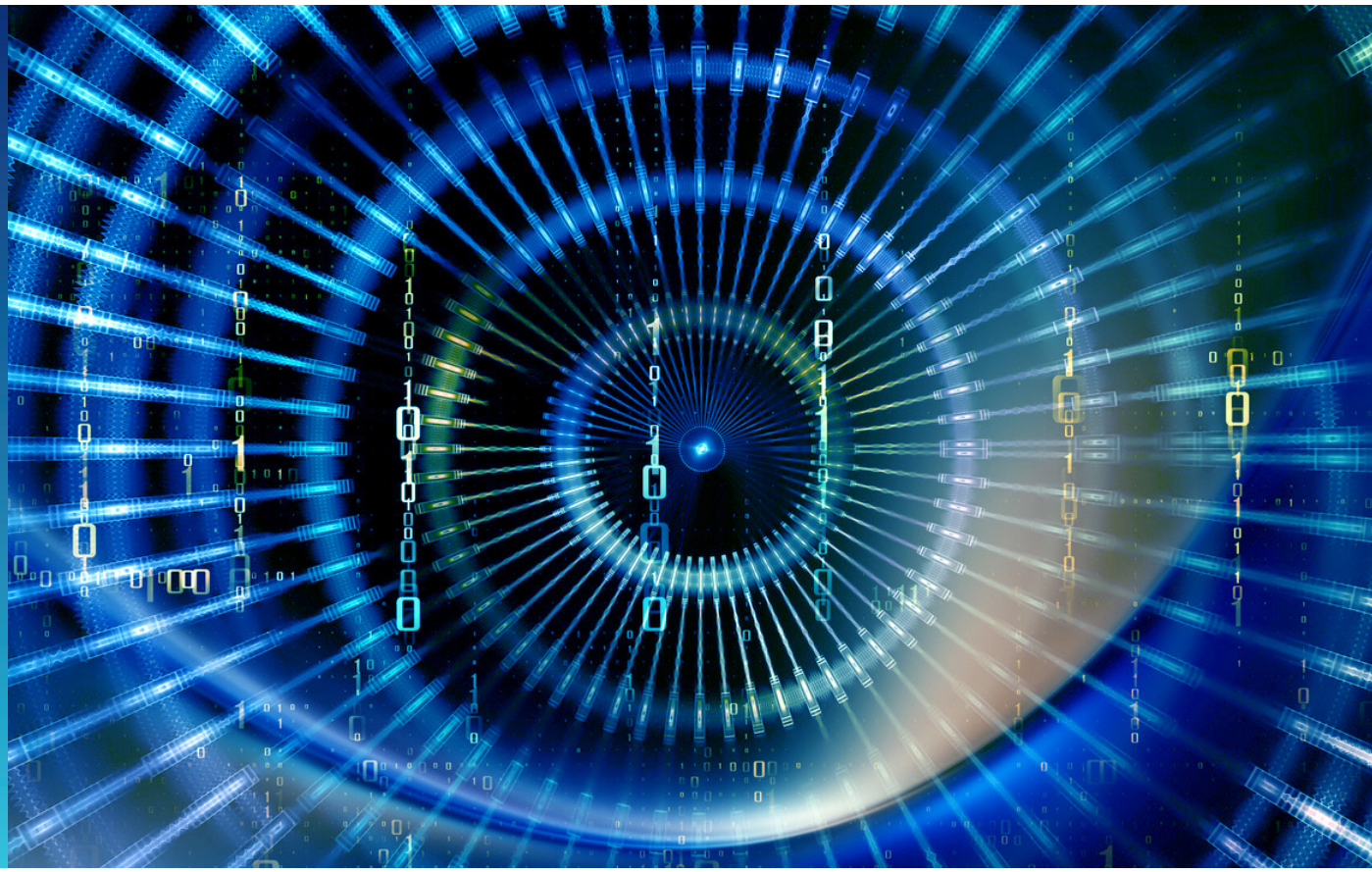
tetrascience

Manual No More: Automating the Scientific Data Lifecycle

WHITE PAPER



AUTHORS:

Spin Wang, CTO
Mike Tarselli, CSO

Manual No More: Automating the Scientific Data Lifecycle

TABLE OF CONTENTS

PREFACE

The world's scientific data today is a morass. 50-80% of research scientist and data scientist time is spent wrestling with data before they can focus on higher-value AI/ML and advanced analysis to help bring new life-saving therapeutics to market. This is due to siloed, fragmented data in hundreds of different proprietary formats, "sneaker-net" movement of data, and rigid, custom point-to-point integrations. A complete reimagining of discovery and development to drive scientific outcomes leads to a new world where science is accelerated with FAIR (Findable, Accessible, Interoperable, Reusable) harmonized data in the cloud, engineered by a cloud-native platform that delivers the benefits of continuous cloud innovation. Vendor products are connected via reusable multi-point productized integrations and vendors are connected via an open digital ecosystem. Harmonized scientific data are freely flowing from instruments through late stages in the biopharma pipeline to accelerate innovation. This paper describes how this new world has become a reality.

Introduction

We asked customers, colleagues, and advisers: "If you built a lab from scratch to orchestrate the free flow of information and data across your laboratory instruments and informatics applications to perform data analysis … what would you build?"

Here is their wishlist:

1. Seamless integration between ELN, LIMS, and instrument control software to automate export of experimental designs.
2. Connectivity with lab equipment, such as balances and pH meters, which places measurements in an easily locatable, centralized location.
3. Data prepared and formatted for modeling and advanced analytics.
4. Data analytics to gain new insights and automatically direct future experiments.

In essence, these customers want a connected and automated laboratory ecosystem that produces data prepared for secondary analyses, such as AI/ML, and adheres to **FAIR (Findable, Accessible, Interoperable, Reusable) data principles**.

How might you achieve this? A new category of technology provides a purpose-engineered, end-to-end solution to the biopharmaceutical research, development, and manufacturing (RD&M) data management puzzle. This technology couples the power and speed of a cloud-native scientific data platform — the Tetra Data Platform (TDP) — with the knowledge of life sciences, computer science and data engineering experts, and an inclusive and open network of partnerships across the RD&M technology landscape. Together, these solve the pervasive problem of information silos, underutilized data, and inefficient research processes that create significant drag throughout biopharma RD&M. Let's further explore how this is accomplished.

Automated and remote instrument control

Scientists design their experiment(s) in an ELN/LIMS such as PerkinElmer Signals. Next, they'll execute these experiment(s) on laboratory instruments.

To do so, you first need a way to easily interact with the laboratory instrument in a simple, vendor-agnostic manner. How can this be done? With an abstraction layer, which is an application programming interface (API) that conceals details of the underlying implementation, providing users with a known, common interface. Building an abstraction layer atop your execution systems — namely instrument control software or workflow orchestration software — allows any application to interface with instruments using a common web API without worrying about the heterogeneous and fragmented nature of lab data systems **(Figure 1)**.

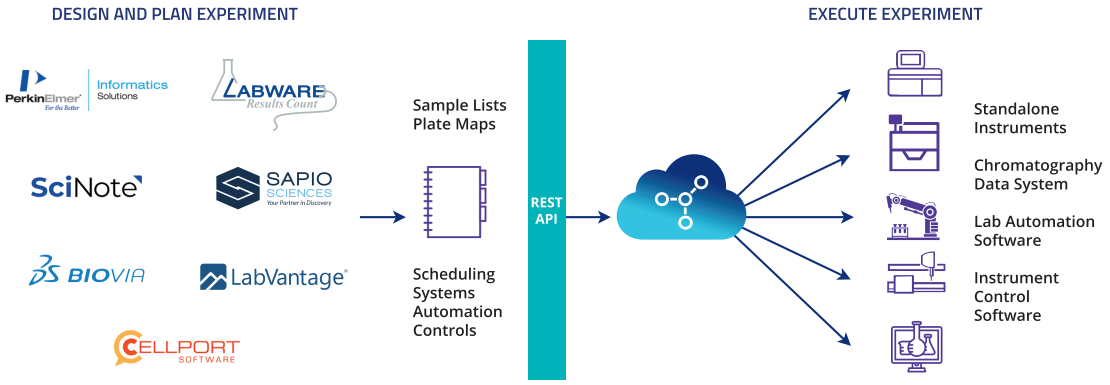


Figure 1 – A schematic depicting how one can interface with laboratory instruments and execution systems using web APIs to transfer experiment instruction or design to their target destinations

As a result, ELN, LIMS, or third-party informatics systems can easily interact with the instrument control software via TDP, using HTTP requests. For example, these requests may be used to create a list of samples to measure inside your chromatography data system (CDS) — e.g. Waters Empower or ThermoFisher Chromeleon. **Figure 2** shows how your scientific data engineers/R&D IT team can access Waters Empower using an **HTTP API request**:

```
curl --location --request POST 'https://api.tetrascience.com/v1/commands' \
--header 'Authorization: Bearer ,your-service-user-jwt-token>' \
--header 'Content-Type: application/json' \
--data-raw '{
  "target": {"name": "CDS 1"},
  "action": "TetraScience.Agent.empower.CreateSampleSetMethod",
  "metadata": "2020-11-03T18:28:03+00:00",
  "payload": {
      "samples": [{
          "name": "one",
          "type": "control"
      }, {
          "name": "two",
          "type": "sample"
      }]
  }
}'
```

Figure 2 – JSON example of an HTTP request to Waters Empower

Automated data acquisition and publication

After the experiment has been executed, the resulting data produced by the instrument must be collected from the instrument. Traditionally this step is another manual process where researchers must find locally stored data files, transfer them to a physical storage drive (or email the files to themselves), and clean the data — most often in Microsoft Excel — prior to analysis. This process introduces opportunities for improper naming conventions, loss of files, and introducing data errors, while taking time away from other high-value tasks.

TDP automates the collection, harmonization, and processing of raw experimental data and storage to the cloud as Tetra Data, a universally adoptable, vendor-neutral data model, built to support the data access and data management needs of biopharma organizations. Tetra Data is compliant, harmonized, liquid, and actionable to enrich scientific data, provide centralized access, flow across instruments and applications, provide provenance and context to prove compliance, and accelerate scientific outcomes.

Tetra Data is stored centrally with the original raw data and prepared to be queried. Importantly, crucial metadata, such as those related to data origin, data integrity verification, or instrument conditions, can be added to enrich Tetra Data through the use of configurable **data pipelines** **(Figure 3)**.

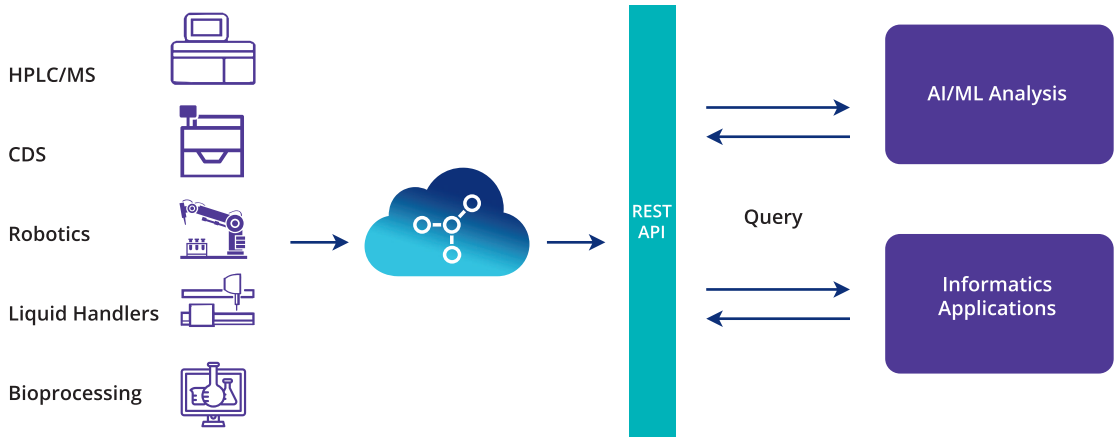


Figure 3 – A schematic depicting the extraction, centralization, enrichment, and harmonization of data from their sources

One common use case is to use pipelines to publish experimental data back to the **originating** application or applications to consume the data. For instance, if you developed an **HTS campaign** in Dotmatics Studies, you'd ideally expect your mass spectrometer, plate reader, and liquid handler metadata to populate the experiment once complete. If you designed a bioprocess template in Signals, you'd expect results from the analyzers and osmometers to populate the E-Workbook **(Figure 4)**.

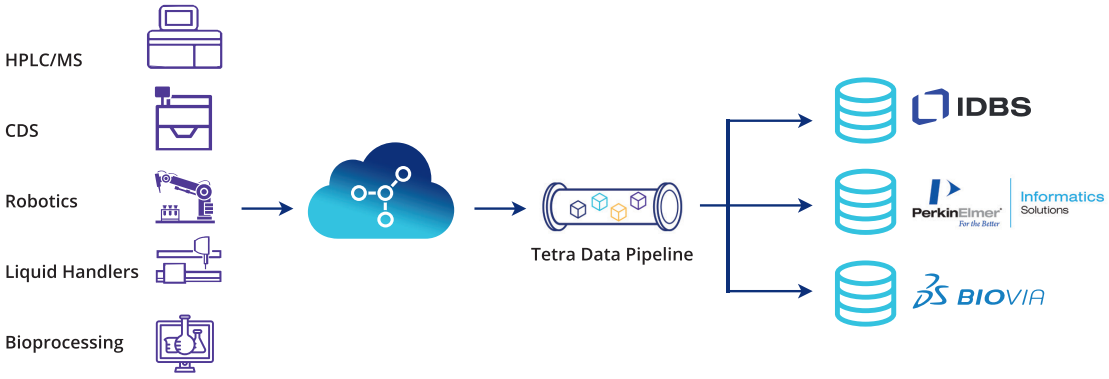


Figure 4 – A schematic representation of data pipeline-mediated transfer of experimental data back to the originating applications

Data pipelines for automated processing and decision making

Once the data are centralized, harmonized, and prepared for analysis, it's possible to leverage the configurable data pipelines to further automate workflows and decision making. To accomplish this goal, commands are sent that inform an instrument of next steps **(Figure 5)**.

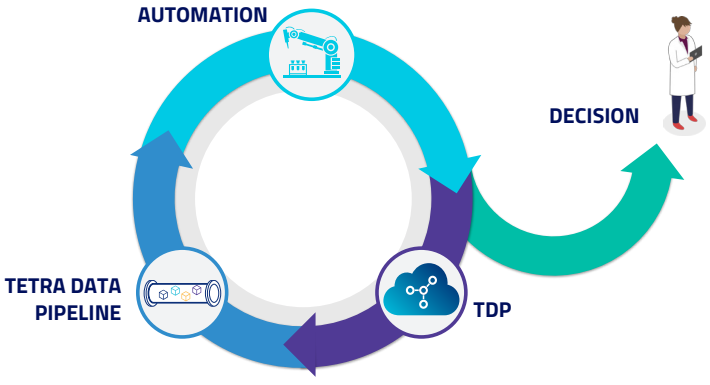


Figure 5 – A schematic representation of automated processing and iteration prior to research decision.

This communication can be achieved using common data engineering/data science languages, such as Python, to introduce customized machine learning and analytics logic. Perhaps most tantalizing, you could utilize this bidirectional path to connected instruments and software to build an automated **Design, Make, Test and Analyze (DMTA) loop** (**Figure 6**).

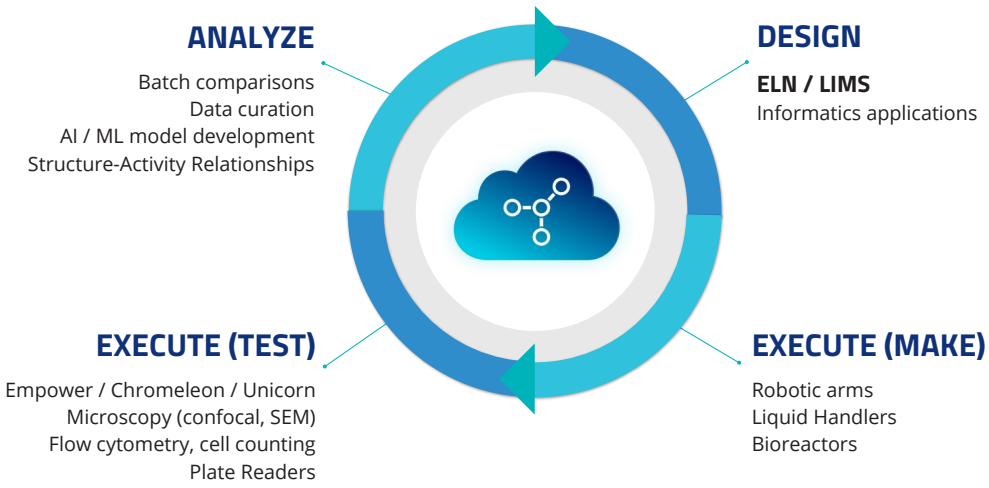


Figure 6 – A cycle chart showing the Design, Make, Test and Analyze (DMTA) Loop. TetraScience can help to automate each step.

Let's consider an example. Two experimental techniques, mass spectrometry (MS) and high-performance liquid chromatography (HPLC), rely on TDP to monitor the output of a high-throughput experimentation cascade. Once data are acquired for all wells of the microplate, TDP triggers pipelines that analyze the LC/MS data and determine the next steps for the experiment based on criteria established by the researcher. The data pipeline sends commands to a Titan Mosaic sample manager, a Beckman Coulter LabCyte Echo, Hamilton Vantage, or Tecan Freedom EVO (liquid handling systems), and a sealer/shaker to set up an initial reaction or assay.

Once the spectral and chromatographic data return to the platform, a data pipeline calculates the parameters for a new set of experiments, such as a concentration series, thermocycle, or additional mixing. Further loops permit a full picture of the experimental response landscape to emerge. **Figure 7** shows how this could be achieved using TDP.

```
import ts.devtools.context'
import pandas as pd

def decide_concentration(data):
  ########
  # add your customized logic
  ########
  return [{
    'A1': '10uMol',
    'A2': '1uMol'
    'A3': '0.1uMol'
  }]

def main(input, context)
    data_frame = pd.read_file(input)

    # handle the data
    # then run concentration series
    # on the next plate
    well_concentrations = decide_concentration(data_frame)

    # send the command to the liquid handler
    context.run_command({
      'target': {'connectorId': 'uuid-of-the-robotic-system-connector'},
      'action': 'TetraScience.Connector.freedomevo.dispense'
      'payload': {
        'well_concentrations': well_concentrations,
        'shake_time': '5min',
        'heat": '30degC for 10 hours',
        'sampling': '10uL'
              }
           })
```

Figure 7 – A code example of how previous results can be used to drive future experiments

In effect, with existing hardware and data collection, TDP would allow your screening facility to mimic the AI-powered discovery platforms of a **Cyclofluidic** or the **MIT-based platform** reported in late 2019.

Data applications for interactive analysis and guided ML

Think of this process like a sequentially iterative version of high-throughput experimentation: instead of running a massively parallel run under a specific set of conditions with a variable of interest, several smaller iterations could be done with smaller sample lists.

For instance, a dose-escalation study, a tolerance test, or a concentration check could all be done quickly with a suitable data model downstream to interpret results and send the next variable to test. Output can be checked by a human operator *ad hoc* during the run through an application like Jupyter Notebook, Qlik, Streamlit, Spotfire, or Tableau.

Automation does not only function to remove the need for scientist input. In fact, there are scenarios where automated data pipelines allow scientists to interact with a decision tree to guide the experiment based upon analysis in near real-time **(Figure 8)**.

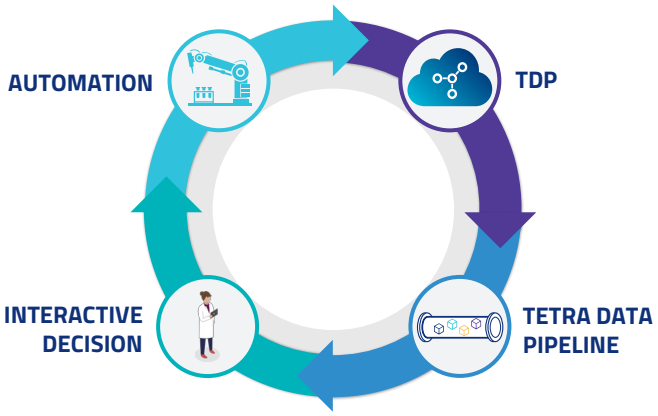


Figure 8 – A schematic representation of automated data processing with interactive decision making

In this example, automated data pipelines can narrow the total parameter space to a small set of options. Then data apps like Streamlit present results to a scientist who interactively makes decisions. The scientist may decide to select only the best responders to a given assay or condition set, or combine multiple "hits" or strong binders from multiple runs to reduce project time and conserve consumables. Whichever criteria they choose, automated data pipelines assist in the curation of Tetra Data that can then be automatically prepared, analyzed, and presented to the researcher to drive faster decision making.

Architectural considerations for seamless instrument control with data publication at scale

A major challenge organizations face involves direct control of an instrument from their ELN/LIMS, while ensuring a seamless return of the data with concomitant publication to the correct experiment or assay. Usually, this setup requires building a custom DIY tailoring of an on-prem ELN/LIMS or vendor customization. Both of these approaches lead to a tangled mess of point-to-point integrations and future maintainability challenges. Such integrations are brittle, costly to maintain, and (most importantly) create yet another scientific data silo. TDP uses metadata to track the origin of the instrument request and fulfills the data response to the ELN/LIMS without the need for complex, custom solutions.

Decoupling instrument control from data publishing enables scaled-up data science

Typically, scientific data automation is done in a fragile, point-to-point fashion. ELN/LIMS systems connect through instrument control software or lab automation software, such as HighRes Biosolutions Cellario, Biotek Gen5, Tecan Fluent, or Molecular Devices SoftmaxPro, which in turn send requests to laboratory instruments. The produced data only flows back to the requester (e.g. ELN/LIMS), preventing other informatics applications, assay warehouses, visualization tools, or AI/ML engines from receiving and using the data. In other words, instrument control and data publication are tied together in the same "handshake."

Disadvantages of point-to-point architecture include:

1. Difficulty having multiple requesters; organizations regularly need two or more systems interacting with the instruments, which is not possible with this architecture
2. ELN/LIMS typically don't focus on integration as their core product and business, rather integration layers are added on as an afterthought
3. Difficulty introducing AI/ML and decision making functionality in the data flow due to the nature of the handshake between the requester and responder
4. Limited FAIR impact, aggregated analytics, and **data liquidity**, since instrument data only gets published to the requester, thus preventing broader access to data by other informatics applications, data science, and analytics software.

Architecturally, TDP introduces the long-awaited abstraction layer and decoupling between instruments and ELN/LIMS. In this decoupled architecture:

1. All scientific data are harmonized and prepared for consumption by any application, along with the metadata coming from the ELN/LIMS, enabling data science and advanced analytics
2. All the requests/commands are logged in the TDP for inspection, trending, and analysis

Once TDP connects to a lab workflow, **all** data from that process are automatically stored within TDP, or deployed in an organization's own data lake. Post-acquisition data science pipelines can be set up to run and provide greater instrument control and flexibility in how the data are published to an ELN or other target.

Delivering better data for faster science

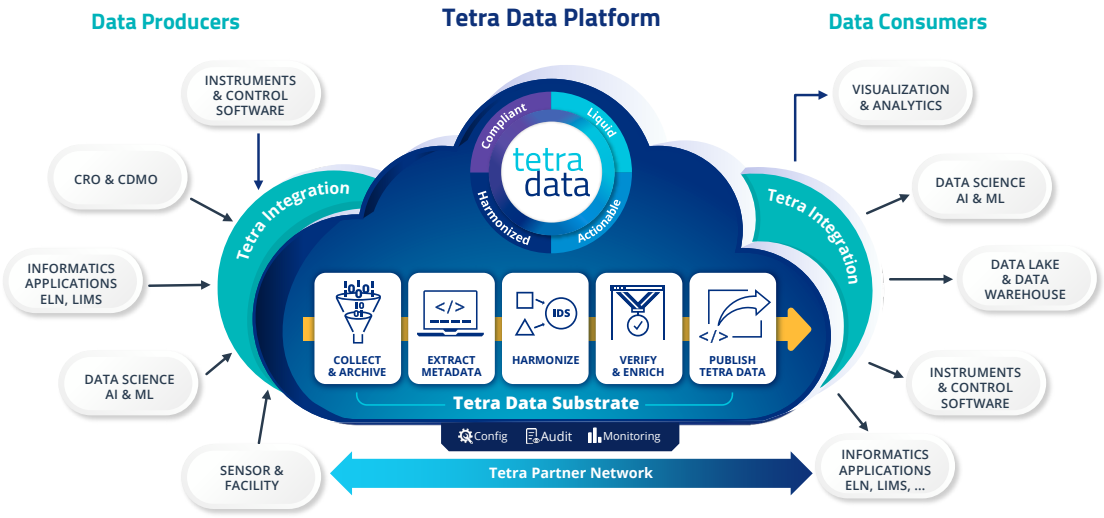


Figure 9 - A visual representation of the Tetra R&D Data Cloud, which comprises the enabling platform (TDP) and the connected network of vendors (TPN) whose products create and consume the data. Together, TDP and TPN create an open ecosystem with unfettered data movement across the drug discovery, development, and manufacturing ecosystem.

Conclusion

To answer the question we initially posed at the top: "If you could build a lab from scratch…", you would need a solution that ensures FAIR data, automates processes using API connectivity, harmonizes data and primes them for secondary analyses, abstracts the challenges of remote communication with instruments and applications, and enables machine-driven experimental optimization.

Modern-day science requires data that are:

- Accessible, searchable, and actionable
- Harmonized from disparate data models
- Fluid between instruments, systems, and software
- Compliant with GxP and other regulatory requirements

TetraScience delivers all of these data attributes to organizations so they can derive maximum long-term value from their RD&M data. For more information on how to automate data flow across its entire lifecycle and accelerate drug discovery and development, please **contact our team**.

tetrascience

TetraScience is the R&D Data Cloud company with a mission to transform life sciences R&D, accelerate discovery, and improve and extend human life. While the company was founded in 2014, we began our R&D Data Cloud journey in 2019 with the origin of the Tetra Data Platform (TDP).

Corporate Headquaters | 177 Huntington Avenue, Suite 1703, Boston, MA 02115
Atlanta | 3280 Peachtree Road NE, 7th Floor, Atlanta, GA 30305

tetrascience.com