tetrascience

TetraScience Lakehouse Architecture

SOLUTION BRIEF

As scientific data continues to grow in complexity and volume, organizations face increasing challenges in managing, analyzing, and deriving insights from this data. TetraScience's adoption of a lakehouse architecture offers a transformative solution that seamlessly bridges the gap between data lakes and data warehouses. This architecture enables the efficient querying and transformation of scientific data, unlocking advanced analytics and Scientific AI use cases—from exploratory data analytics using SQL to integrating scientific data into business intelligence (BI) dashboards to training and running ML models.

By leveraging this architecture, TetraScience empowers its customers to more effectively automate data engineering processes and enable robust data analytics, machine learning, and Scientific AI capabilities. With TetraScience's lakehouse-based solution, customers can now:

- Efficiently manage datasets for both data lake and data warehouse use cases via SQL, using an open table format with Delta Lake.
- Create analytics-optimized datasets tailored to specific scientific use cases.
- Share datasets across data catalogs and platforms, such as Databricks, without incurring additional governance burdens or costs associated with data duplication.

The Challenges

The ability to leverage scientific data is fraught with challenges, primarily due to the diverse and complex nature of the data generated:

- **Inconsistent Data Formats:** Scientific data is often generated as files in various formats, making it difficult to conduct comparable analyses across experiments or runs. Instrument and experimental assay data come in a plethora of vendor-proprietary formats, complicating data analysis.
- **Fragmented Data Outputs:** Analysis outputs and results are typically exported into spreadsheets, images, and graphs, leading to fragmented data storage that hinders comprehensive analysis. Furthermore, the context and design of experiments are not file-based and often exist in inconsistent formats, creating additional hurdles in data management.
- **Limited Analytical Scope:** The current approach to scientific data analysis is largely limited to individual assays and the associated files, resulting in missed opportunities to conduct broader analyses across assays, experiments, and studies, whether longitudinally or across an organization's scientific activities.
- **Siloed Compute and Storage:** Various compute engines and analytical tools need seamless access to scientific data to power analysis, BI, reporting, and data science use cases. The inherent gravity of data, which refers to the tendency of large datasets to attract related applications and services, adds challenges when moving data between systems or architectures. Current data management approaches often hinder quick and effective access to workable, analyzable datasets. They rely on ETL processes that copy data, complicating governance and regulatory compliance and driving up storage costs due to data duplication and inefficient storage practices.

To address these challenges, it's imperative to shift from a file-based to a dataset-based paradigm, enabling data to be engineered into formats that are more suitable for advanced analytics, machine learning, and Scientific AI.

Corporate Headquarters | 177 Huntington Avenue, Suite 1703, Boston, MA 02115
© 2024 TetraScience, Inc.

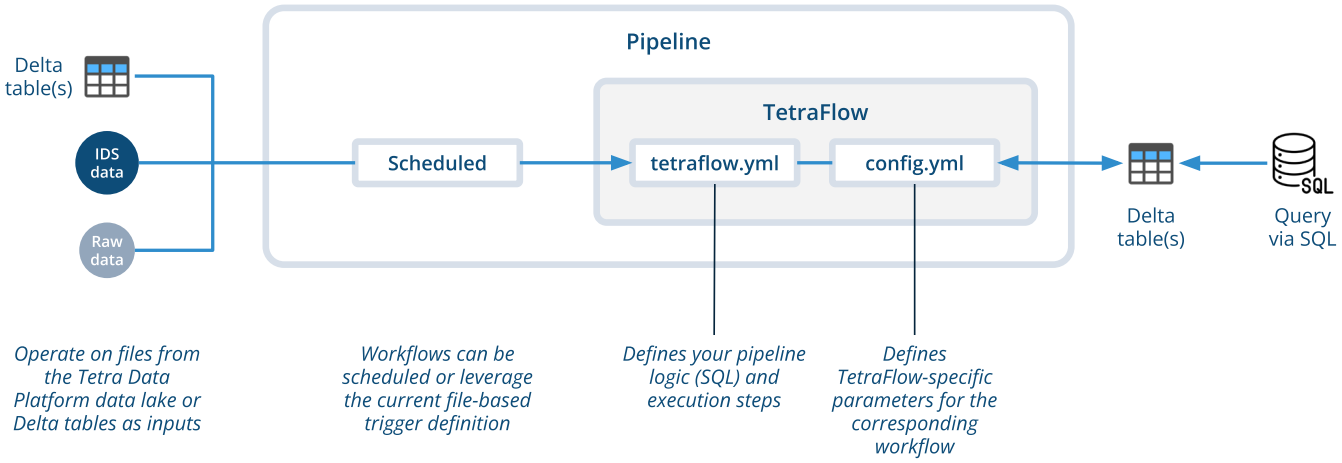
The Solution

Version 4.1 of the Tetra Data Platform introduces three key components that support a lakehouse architecture. They equip customers with the tools to create AI-ready datasets for analytics and Scientific AI use cases, leveraging both new scientific data and existing historical data.

1. **Data in an Open Storage Format Accessible via SQL:** The Intermediate Data Schema (IDS), a component of Tetra Data, is transformed into an open table format using Delta Lake, enabling high-performance access via SQL. This transformation is achieved in two ways: by creating copies of existing SQL tables that were previously backed by a less optimal storage format, and by generating flattened tables that are optimized for aggregate query performance, with each row representing a single record. These flattened tables are designed to facilitate both data analytics and further data transformation into analytics-ready datasets, functioning as the equivalent of a Bronze table in a medallion architecture.

2. **Data Transformation Engine Supporting a Medallion Data Architecture:** The new transformation engine supports SQL-to-SQL transformations and scheduled workflows for batch processing, combining multiple datasets into one and enriching data with reference metadata. This engine simplifies the process of creating analytics-optimized datasets and ensures that data is ready for advanced use cases, including Scientific AI. It also enables data teams to easily manage and promote their datasets through the different phases of the medallion architecture, with the ultimate goal of Gold datasets that are enriched and optimized for analytics.

3. **Data Catalog Integration:** All Delta tables are registered in data catalogs accessible via the Tetra Data Platform. These catalogs facilitate data sharing with platforms like Snowflake and Databricks Unity Catalog. This integration eliminates the need for data duplication and simplifies data governance.



TetraFlow pipelines powering the new lakehouse architecture

© 2024 TetraScience, Inc.

Implementation on the Tetra Data Platform

As part of the Tetra Data Platform deployment, starting with version 4.1, customers can take the following steps:

For Existing SQL-Based Analytics

1. Replay existing Tetra Data into the new SQL tables using the bulk processing functionality on the Tetra Data Platform. Coordinating with the support team ensures a smooth transition.
2. Confirm that the table structure and data remain consistent.
3. Update downstream consumer queries to point to the new tables, with example table name differences provided for clarity.

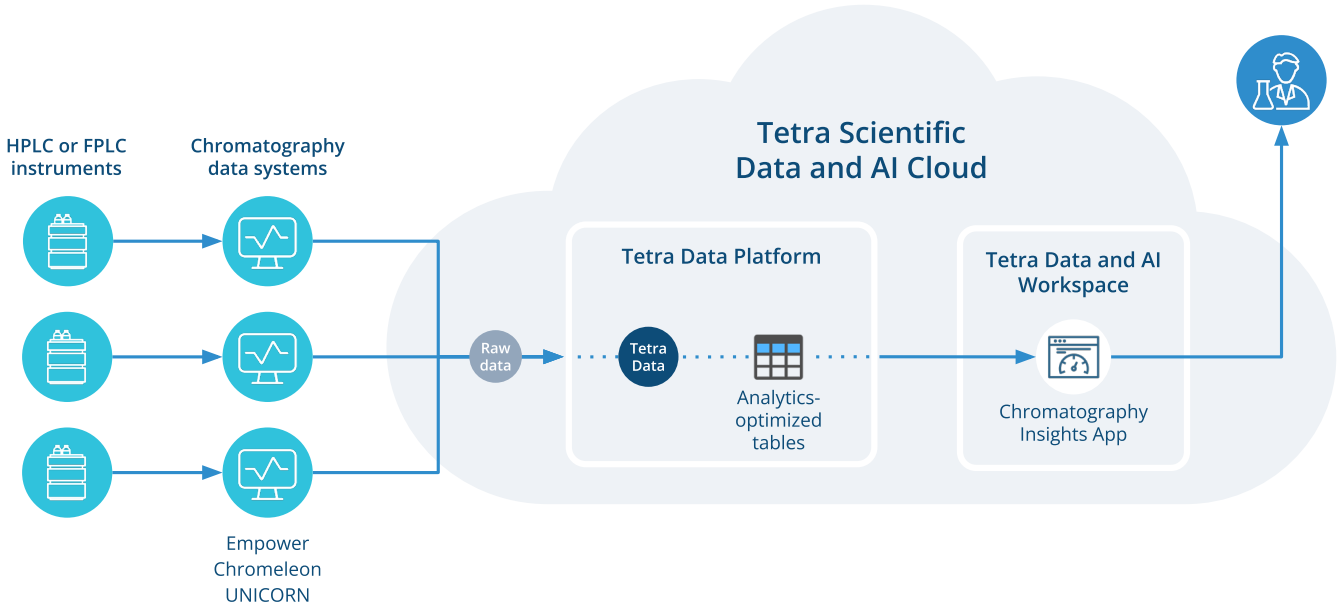
For New Datasets and Analytics-Ready Data

1. Set up an IDS-to-Delta pipeline to facilitate the transformation.
2. Build and test TetraFlow pipelines for SQL-to-SQL transformations.
3. Reprocess existing Tetra Data (IDS) through the new pipelines for historical data as needed, ensuring that all data is up-to-date and ready for analysis.

Example Reference Solutions

Merging Datasets

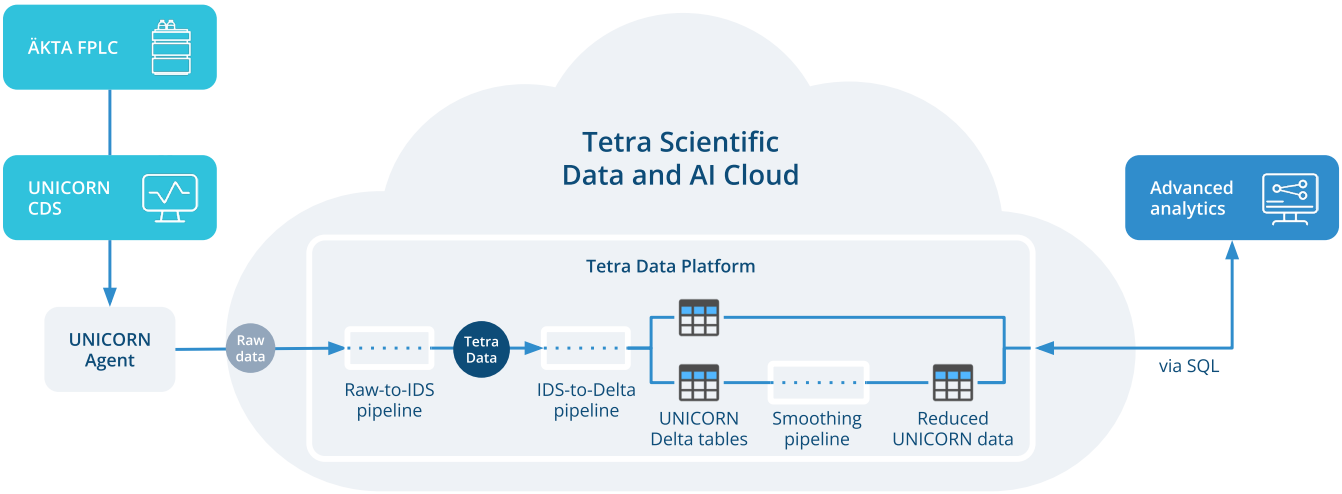
The solution diagram below shows how the lakehouse architecture processes multiple source chromatography schemas into a single, optimized table using the new TetraFlow pipeline artifact. This approach streamlines analysis and provides a comprehensive view of the data.



© 2024 TetraScience, Inc.

Creating Analytics-Ready Datasets

By reducing the resolution of raw chromatography data, the lakehouse architecture enables faster analyses, particularly when overlaying more than 10 runs with millions of data points. The transformed SQL table provides an algorithmic sample of measurements, while the original data remains accessible in another table. This balance between performance and data integrity is crucial for effective scientific analysis while remaining compliant.



The Benefits

- **Faster SQL Performance:** By transitioning to an open storage data format, SQL performance improves significantly, particularly when querying large datasets or performing aggregate queries. This enhancement is particularly evident when using AWS Athena, Databricks, or Snowflake.
- **Creation of Analytics-Ready, Use Case–Specific Datasets:** The new capabilities allow customers to create optimized SQL tables from existing Tetra Data schemas, simplifying the analysis process. For instance, transforming a schema for a chromatography data system to focus on injection information, samples, and demarcated peaks eliminates the need for complex SQL statements.
- **Merging Multiple Datasets for Comprehensive Analysis:** The ability to merge data from different sources into a single analysis table streamlines the analytical process. For example, analyzing lab trends across multiple instruments becomes more straightforward with a unified users table, eliminating the need for data analysts to have to manage complex queries.
- **Simplified Data Transformations with Scheduled Workflows:** Scheduled workflows reduce the complexity of data transformations by eliminating the need for multiple pipelines and triggers. A single pipeline can manage the retrieval and combination of data from various sources, ensuring consistent and timely data availability.

Conclusion

TetraScience's lakehouse architecture represents a significant advancement in leveraging scientific data, enabling customers to accelerate their scientific outcomes through replatforming, engineering, and harnessing scientific data with analytics and Scientific AI. To learn more about incorporating this solution into your scientific software landscape, please reach out to TetraScience.

Corporate Headquarters | 177 Huntington Avenue, Suite 1703, Boston, MA 02115
© 2024 TetraScience, Inc.

tetrascience
tetrascience.com