Ardigen

Enabling Real-Time Analytics in Clinical Trials

Ardigen: Łukasz Kozak, Bohdan Bondar, Filip Chrzan, Błażej Szczerba, Anna Sanecka-Duin, Piotr Faba
Ryvu Therapeutics: Kamila Olech, Przemysław Szczygieł, Karol Rorat, Joanna Macnar, Aleksandra Mazgaj, Daniel Grzywczak, Rafał Wojdan

ABSTRACT

Ryvu Therapeutics, in collaboration with Ardigen, transitioned from traditional data silos to a unified, AI-ready platform using the Databricks Lakehouse for Clinical Trials Management. The key drivers behind this transformation like near real-time analytics, scalable data integration, and AI readiness are explored. The solution unifies diverse clinical datasets, ensures regulatory compliance, and leverages cloud-based scalability across AWS, GCP, and Azure. This case study showcases how modern data architectures empower clinical teams and senior management to make faster, data-driven decisions while laying the groundwork for AI innovations in clinical research.

RYVU THERAPEUTICS

Ryvu Therapeutics is a clinical-stage drug discovery and development company focused on advancing novel small molecule therapies targeting emerging areas in oncology.

Founded in 2007 and headquartered in Kraków, Poland, Ryvu employs approximately 300 professionals, including 90 Ph.Ds. Originally known as Selvita, the company adopted its current name following the spin-off of its services division.

Ryvu's lead candidate, RVU120, is a selective CDK8/CDK19 kinase inhibitor in Phase II development for relapsed/refractory acute myeloid leukemia (r/r AML), high-risk myelodysplastic syndromes (HR-MDS), and low-risk MDS. Another asset, SEL24 (MEN1703), a dual PIM/FLT3 kinase inhibitor, is licensed to the Menarini Group.

The company's pipeline also includes candidates in kinases, synthetic lethality, and immuno-oncology, supported by global collaborations with BioNTech, Merck KGaA, Exelixis, Menarini Group, and the Leukemia & Lymphoma Society.

ARDIGEN

Ardigen is a leading AI-driven Contract Research Organization (CRO) dedicated to transforming drug discovery with precision and innovation. Recognized among the top 5% of AI-focused CROs, Ardigen combines expertise in biology, bioinformatics, machine learning, and software engineering to accelerate drug development and enhance the probability of clinical success.

With over nine years of experience, Ardigen has partnered with more than 100 clients, including 15 of the world's top pharmaceutical companies, delivering tailored solutions to address the industry's most complex challenges.

Guided by the mission to deliver medicines to patients faster, Ardigen turns biomedical data into actionable insights, empowering biotech and pharma companies with advanced tools and technology to drive effective treatments forward.

At its core, Ardigen's work is driven by a "People First" philosophy — ensuring every AI algorithm, discovery solution, and partnership contributes to improving human health and well-being.

METHODS / TECHNOLOGIES

The solution discussed in the webinar is built on the Databricks Data Intelligence Platform, leveraging its Lakehouse architecture to unify diverse clinical datasets into a single, AI-ready environment. This platform combines the flexibility of data lakes with the performance and governance of data warehouses, enabling real-time analytics and scalable data integration. Deployed on AWS, it ensures cloud-based scalability, security, and compliance with regulatory standards. A crucial component of the solution is a suite of interactive dashboards developed in Power BI and Spotfire, tailored to the needs of various stakeholders — from clinical operations teams tracking patient progress to senior management reviewing high-level trial outcomes. These dashboards provide intuitive, real-time visualizations, empowering teams to make faster, data-driven decisions. The platform also includes automated data ingestion pipelines, collaborative workspaces for data science teams, and AI-driven insights, laying the foundation for future innovations in clinical trial management.

databricks aws Power BI Spotfire

RESULTS AND DISCUSSION

The implementation of the Clinical Data Lakehouse has delivered significant results, both measurable and operational. The platform has ensured on-time data delivery, enhanced decision-making processes, and fostered the development of data engineering and analytics skills. Operational efficiency has improved, and data literacy across teams has increased substantially. Notably, data maturity advanced from level 1 to level 3 within just 10 months.

From a quantitative perspective, the internalization of CRO services has yielded an estimated 10x ROI. Data recency now stands at 1 day, and the time spent on manual data preparation has been reduced to zero. The platform enjoys full adoption, with 100% positive stakeholder feedback and 100% team utilization. Additionally, 15 diverse data sources have been successfully integrated into the system.

Looking ahead, the solution offers strong potential for expansion by incorporating Generative AI (GenAI) agent systems, leveraging Databricks' end-to-end AI capabilities. This includes tools such as APIs for seamless interaction, unstructured retrieval through vector search, and structured retrieval with the Genie API. Advanced LLM serving, external LLM integrations, and robust guardrails ensure safe AI operations. Furthermore, the platform supports agent logging, evaluation, serving, and continuous monitoring—paving the way for AI-powered clinical insights and decision support in the future.

The challenge

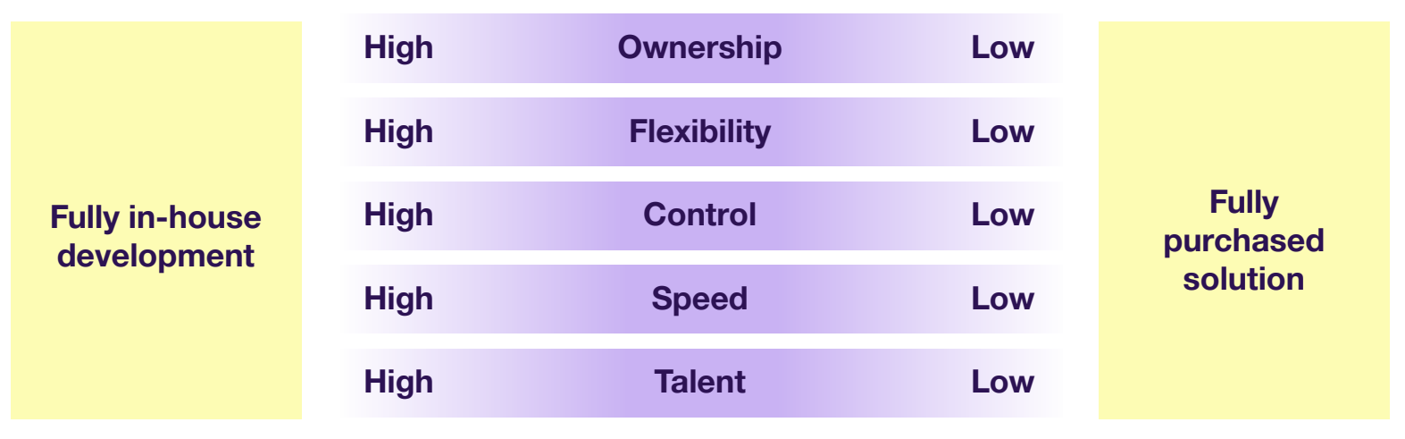
1. 10-20 fold increase from Phase I to Phase II in clinical data volume makes manual data wrangling not feasible anymore.
2. Increased risk of errors.
3. Time consuming Excel-based wrangling.
4. Data analysis reproducibility and data linage out of control.

Business needs

1. Tracking of patient's safety and complying with reporting requirements.
2. Data summarisation for scientific discussions, publications, regulatory interactions.
3. Improve data analytics and reporting.
4. Automate visual analytics of data e.g. correlations.

Desired Characteristics of the Solution

1. Data integration, centralisation and unification under one platform.
2. Data automation, orchestration and harmonisation ensuring daily update.
3. Visual analytics via dedicated dashboards.
4. Faster insights extraction and decision making
5. Improved data quality monitoring.



Fully in-house development

High Ownership Low
High Flexibility Low
High Control Low
High Speed Low
High Talent Low

Fully purchased solution

Figure 1. Diagram represents parameters considered when analysing Buy vs Build strategies

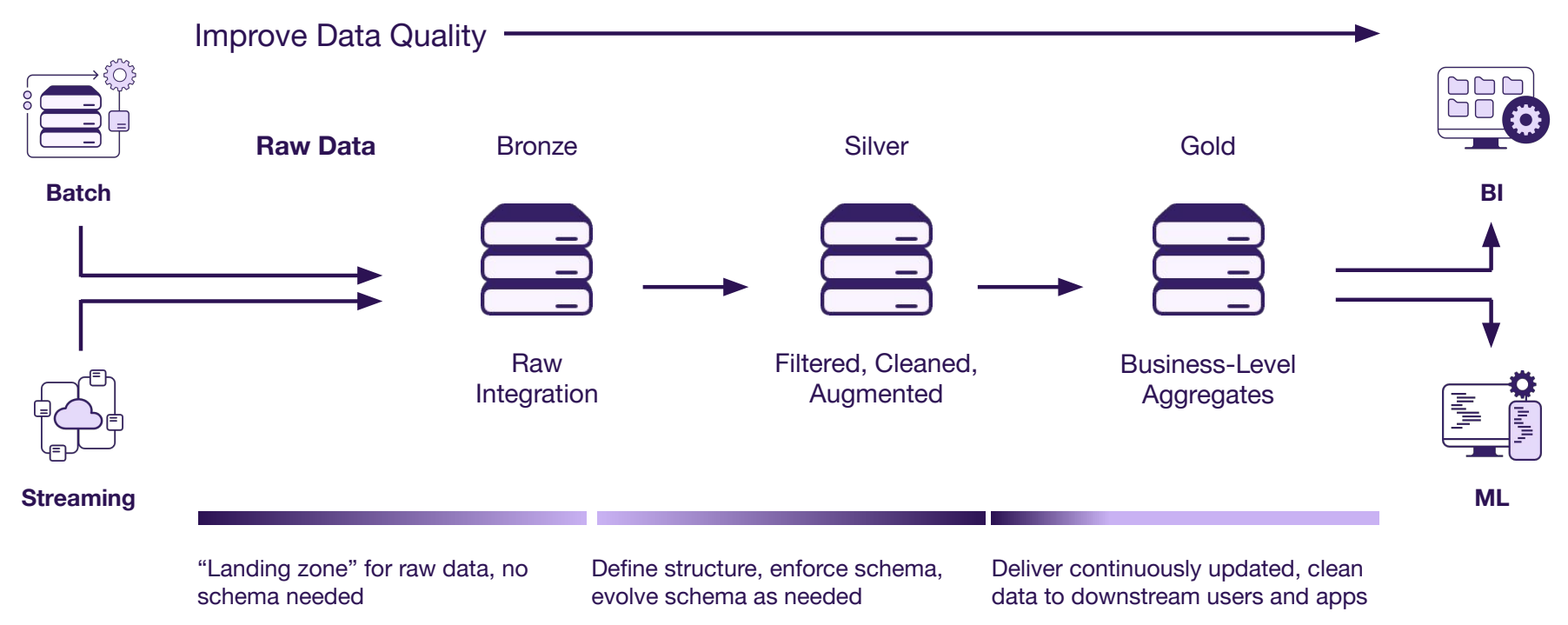
Management of buy vs build dillema: Hybrid model works best

1. Internal PoC Development: Launched to better understand data challenges.
2. Needs Specification: Requirements discussed with a cross-functional team.
3. Buy Option Considered: Evaluated due to limited resources and priority on speed.
4. Market Analysis: Platforms had strong analytics but weak data integration, high costs, and an unfavorable pricing model.
5. Informed Decision: Insights helped secure CIO and CEO support for an internal build with consulting assistance.

Medallion Architecture for Organizing Clinical Data in the Lakehouse

Medallion Architecture was used to structure and manage data within the Databricks Lakehouse. The architecture organizes data into three progressive layers, ensuring data quality, traceability, and AI readiness:

- Bronze Layer (Raw Data): Ingests and stores raw clinical and genomic data from diverse sources, preserving its original format for auditability.
- Silver Layer (Cleansed Data): Processes and refines the raw data by removing duplicates, handling missing values, and ensuring data consistency — preparing it for downstream analysis.
- Gold Layer (Business-Ready Data): Provides curated datasets tailored for clinical reporting, real-time dashboards, and AI-driven analytics.



Improve Data Quality

Batch

Streaming

Raw Data

Bronze

Raw Integration

Silver

Filtered, Cleaned, Augmented

Gold

Business-Level Aggregates

BI

ML

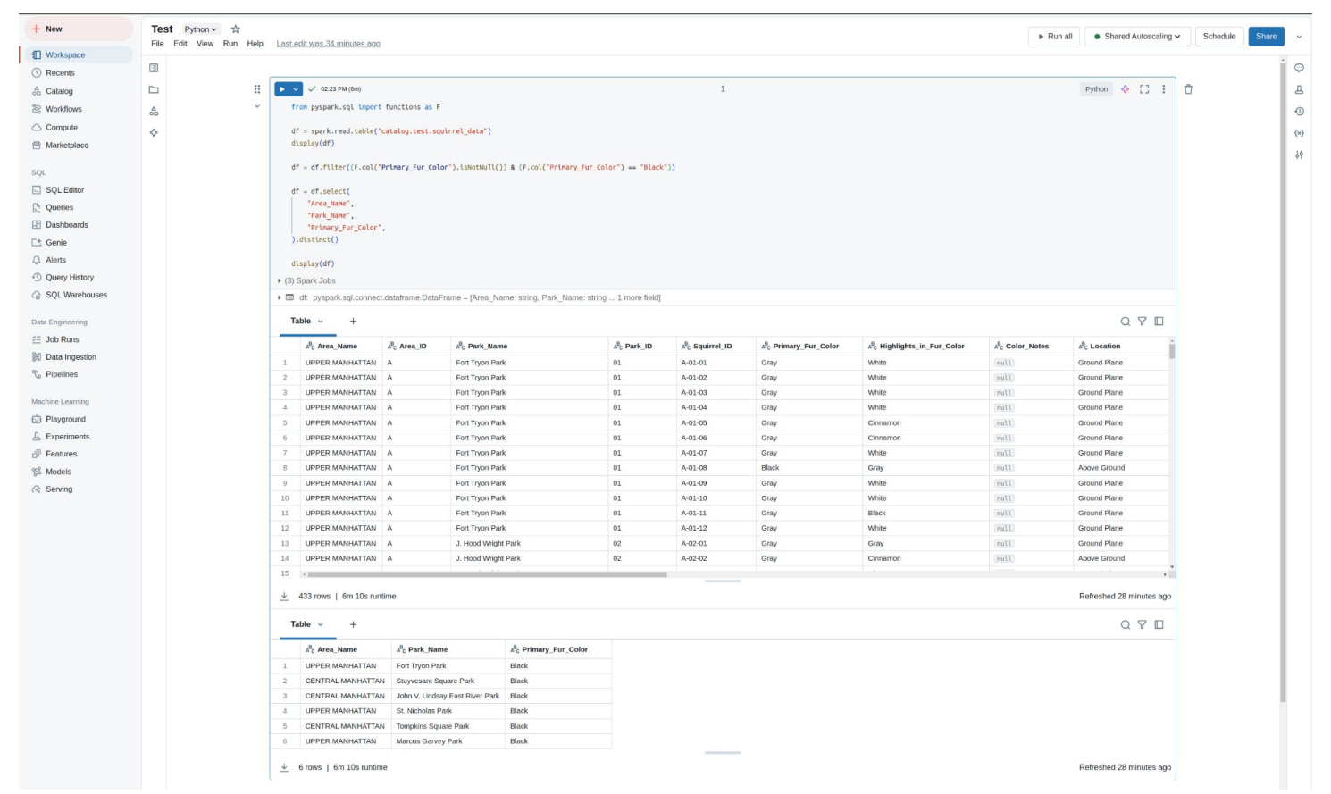
"Landing zone" for raw data, no schema needed

Define structure, enforce schema, evolve schema as needed

Deliver continuously updated, clean data to downstream users and apps

Figure 2. The diagram illustrates the Medallion Architecture, organizing data within the Lakehouse to ensure a structured flow from raw data to refined, analytics-ready datasets.

The Medallion Architecture improves data quality, lineage, and governance by establishing a clear progression from raw to fully-processed data. This structured approach supports scalable data integration, real-time insights, and regulatory compliance — all essential for modernizing clinical trial data management.



```
from pyspark.sql import functions as F

df = spark.read.table("catalog.test.squirrel_data")
display(df)

df = df.filter((F.col("Primary_Fur_Color").isNotNull()) & (F.col("Primary_Fur_Color") == "Black"))

df = df.select(
    "Area_Name",
    "Park_Name",
    "Primary_Fur_Color",
).distinct()

display(df)
```

df: pyspark.sql.connect.dataframe.DataFrame = [Area_Name: string, Park_Name: string ... 1 more field]

	Area_Name	Area_ID	Park_Name	Park_ID	Squirrel_ID	Primary_Fur_Color	Highlights_in_Fur_Color	Color_Notes	Location
1	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-01	Gray	White	null	Ground Plane
2	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-02	Gray	White	null	Ground Plane
3	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-03	Gray	White	null	Ground Plane
4	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-04	Gray	White	null	Ground Plane
5	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-05	Gray	Cinnamon	null	Ground Plane
6	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-06	Gray	Cinnamon	null	Ground Plane
7	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-07	Gray	White	null	Ground Plane
8	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-08	Black	Gray	null	Above Ground
9	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-09	Gray	White	null	Ground Plane
10	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-10	Gray	White	null	Ground Plane
11	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-11	Gray	Black	null	Ground Plane
12	UPPER MANHATTAN	A	Fort Tryon Park	01	A-01-12	Gray	White	null	Ground Plane
13	UPPER MANHATTAN	A	J. Hood Wright Park	02	A-02-01	Gray	Gray	null	Ground Plane
14	UPPER MANHATTAN	A	J. Hood Wright Park	02	A-02-02	Gray	Cinnamon	null	Above Ground

433 rows | 6m 10s runtime — Refreshed 28 minutes ago

	Area_Name	Park_Name	Primary_Fur_Color
1	UPPER MANHATTAN	Fort Tryon Park	Black
2	CENTRAL MANHATTAN	Stuyvesant Square Park	Black
3	CENTRAL MANHATTAN	John V. Lindsay East River Park	Black
4	UPPER MANHATTAN	St. Nicholas Park	Black
5	CENTRAL MANHATTAN	Tompkins Square Park	Black
6	UPPER MANHATTAN	Marcus Garvey Park	Black

6 rows | 6m 10s runtime — Refreshed 28 minutes ago

Figure 3.
Databricks provides an interactive environment for data exploration and analysis through its intuitive, cloud-based notebooks, allowing users to work directly with data in place. These notebooks support multiple programming languages, including Python, enabling seamless integration of data science libraries and real-time collaboration. With built-in support for data visualization, machine learning, and distributed computing.

Databricks as technological fit

Unified Data and AI Platform

1. Combination of data engineering, data science and machine learning workflows in a single platform.
2. Elastic scalability adjusts resources based on workload needs (cost-effective).

Automation

1. Automated execution of data pipelines, machine learning models, and analytics workflows.
2. Flexible Scheduling (time-based, event-based).

Monitoring of Pipelines and Data Flow

1. Job execution tracking with logs and metrics.
2. Alerts for failures and performance bottlenecks.

Integration & Centralization

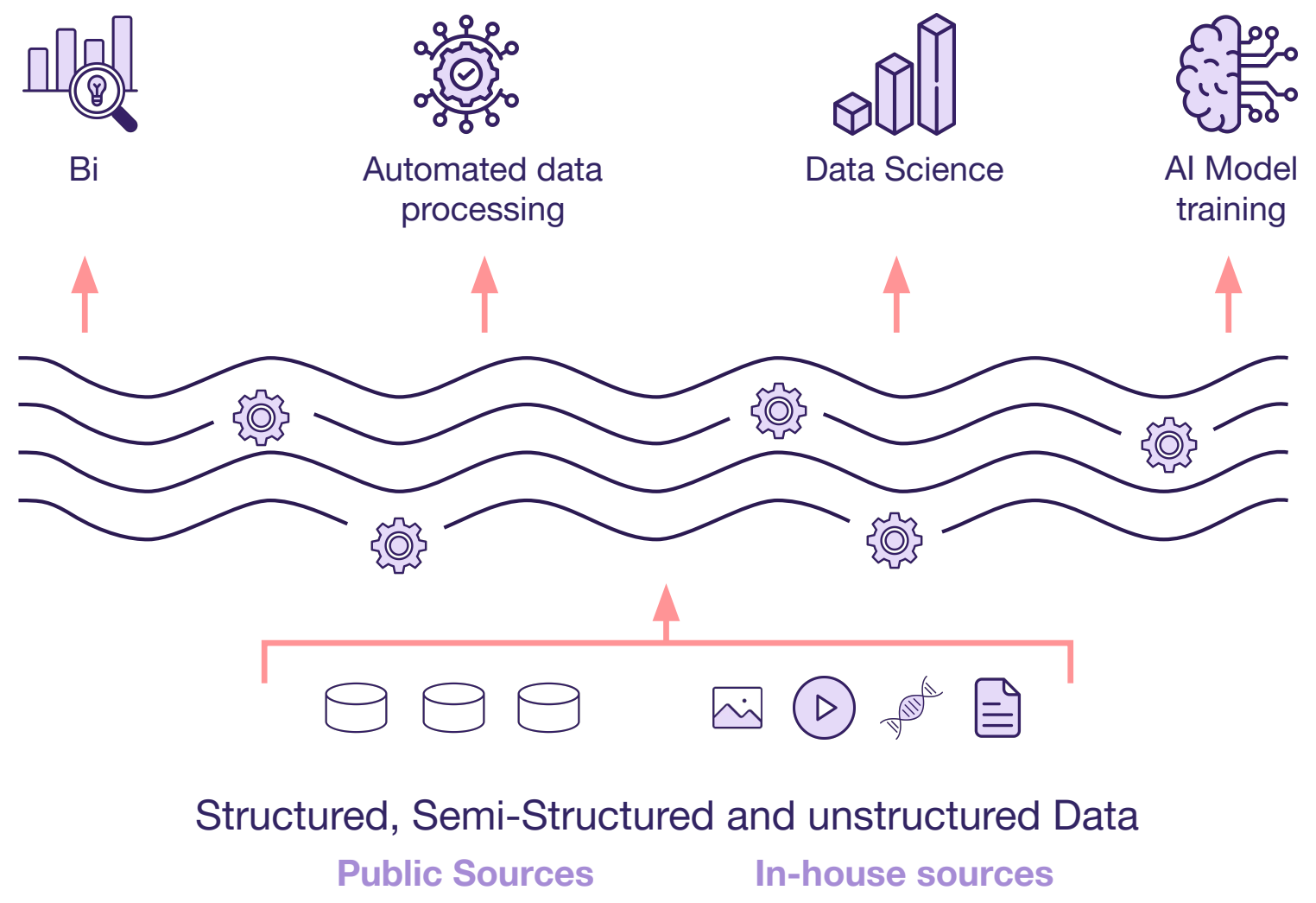
1. Seamless integration of structured and unstructured data sources (e.g. EDC, External lab AWS S3, Internal SharePoint).
2. A unified point of access designed to accommodate the diverse needs of users (Power BI, Spotfire, raw data).

Exploration

1. Built-in data explorer for viewing schemas, sample data and lineage of tables.
2. Shared notebooks for real-time collaboration.

End-to-end support for Agent Systems:

1. Unstructured Retrieval (Vector Search)
2. Structured Retrieval (Genie API)
3. LLM Serving, External LLM Integrations, guardrails
4. Agent Logging, Evaluation, Serving,
5. Monitoring



Bi

Automated data processing

Data Science

AI Model training

Structured, Semi-Structured and unstructured Data

Public Sources

In-house sources

Figure 4.
The diagram illustrates the data lakehouse concept, which supports both structured and unstructured data. It features functional components (represented by cogs) that enable strict data organization (see Figure 2) and interfaces that facilitate data serving to BI applications, data exploration, AI-ready data utilization for AI applications, and automated data processing.

QUANTITATIVE BENEFITS

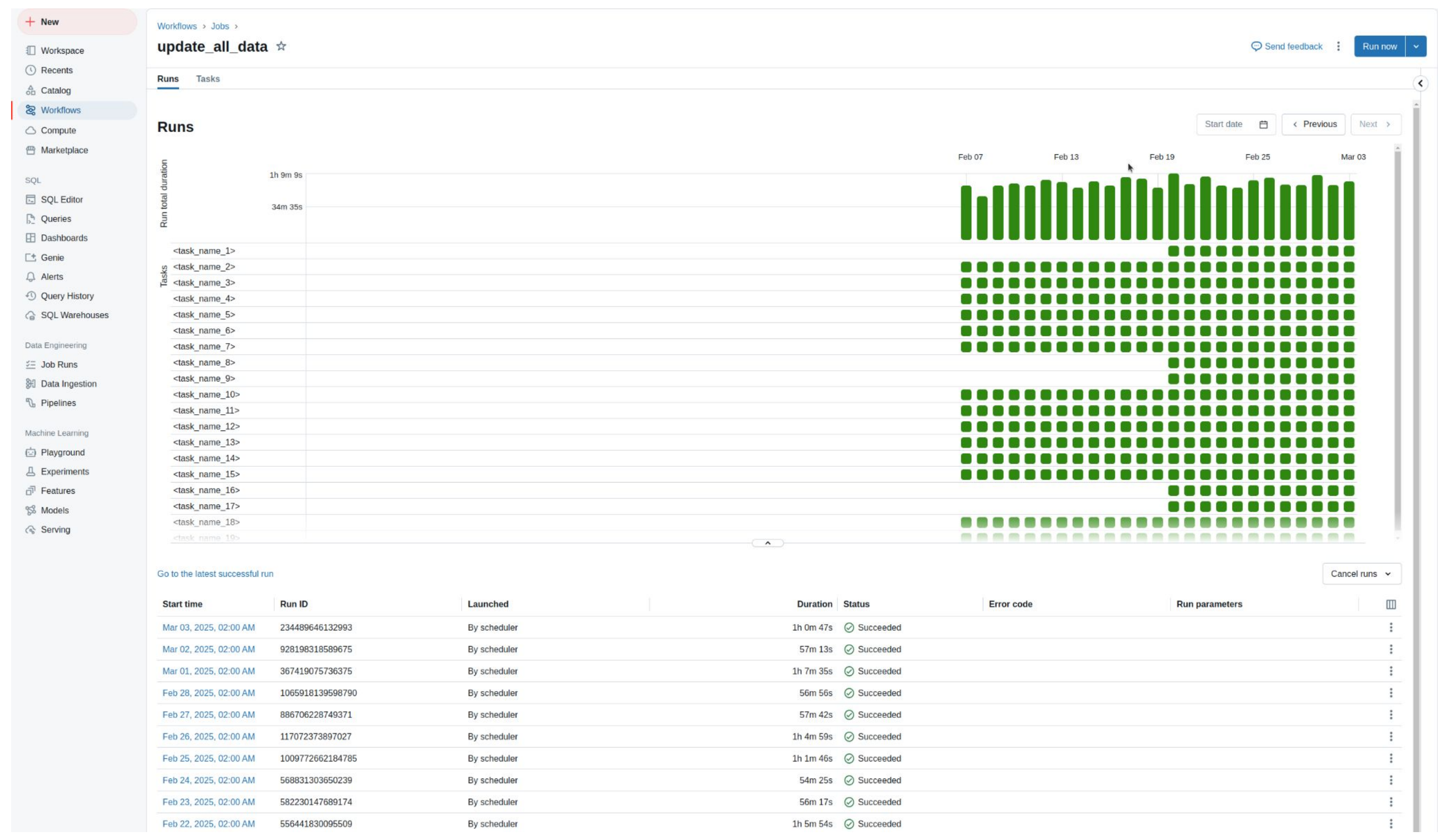
Return on Investment	Data recency
~10x	1 day
Time spent on data preparation reduced to	Stakeholders positive feedback
0 days	100%
Number of integrated data sources	Utilisation by teams
15	100%

Qualitative benefits

1. On-time delivery of trials
2. Enhanced decision-making
3. Data engineering and analytics skill development.

1. Improved operational efficiency
2. Increased data literacy
3. Data Maturity* jumped from level 1 (Basic) to 3 (Systematic) within 10 months

Figure 5: Quantitative and qualitative results as reported by Rafał Wojdan from Ryvu Therapeutics; * According to Gartner Analytics Maturity Model



Start time	Run ID	Launched	Duration	Status	Error code	Run parameters
Mar 03, 2025, 02:00 AM	234489646132993	By scheduler	1h 0m 47s	Succeeded		
Mar 02, 2025, 02:00 AM	928198318589675	By scheduler	57m 13s	Succeeded		
Mar 01, 2025, 02:00 AM	367419075736375	By scheduler	1h 7m 35s	Succeeded		
Feb 28, 2025, 02:00 AM	1065918139598790	By scheduler	56m 56s	Succeeded		
Feb 27, 2025, 02:00 AM	886706228749371	By scheduler	57m 42s	Succeeded		
Feb 26, 2025, 02:00 AM	117072373897027	By scheduler	1h 4m 59s	Succeeded		
Feb 25, 2025, 02:00 AM	1009772662184785	By scheduler	1h 1m 46s	Succeeded		
Feb 24, 2025, 02:00 AM	568831303650239	By scheduler	54m 25s	Succeeded		
Feb 23, 2025, 02:00 AM	582230147689174	By scheduler	56m 17s	Succeeded		
Feb 22, 2025, 02:00 AM	556441830095509	By scheduler	1h 5m 54s	Succeeded		

Figure 6: The left side of the menu displays a clear and organized interface for managing data automations, overseeing data access, accessing workspaces for data wrangling, managing compute resources, and utilizing machine learning functionalities. The selected Workflows screen provides an overview of job statuses, enabling users to monitor and manage automated jobs, as well as investigate any failures when necessary.

Figure 7:
To the right is the Swimmer Plot, one of several Power BI dashboards that are pulling data from Databricks. On the left side, a menu displays an overview of dashboard categories. The main screen offers comprehensive visualizations, keeping medical officers and management well-informed about progress.

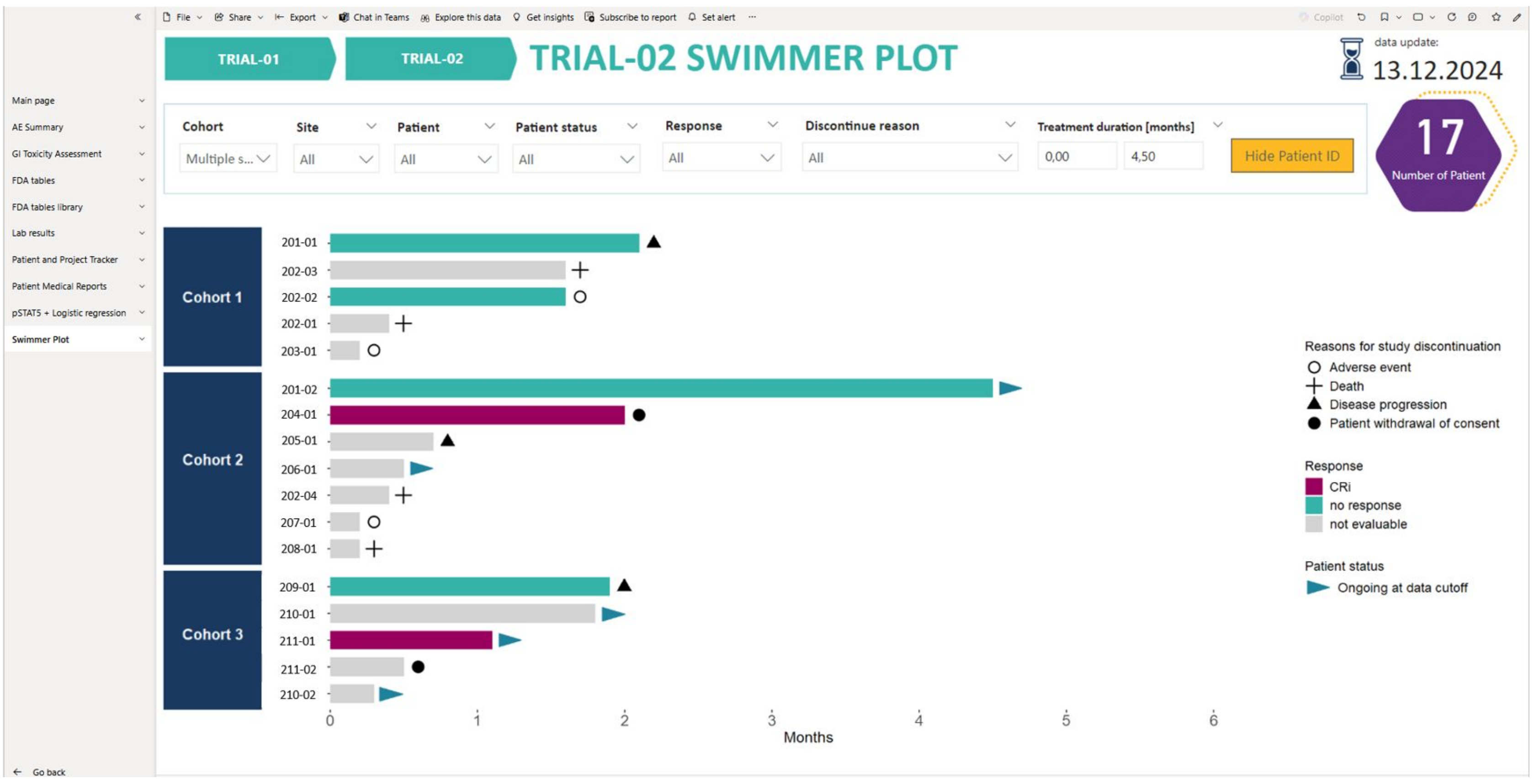
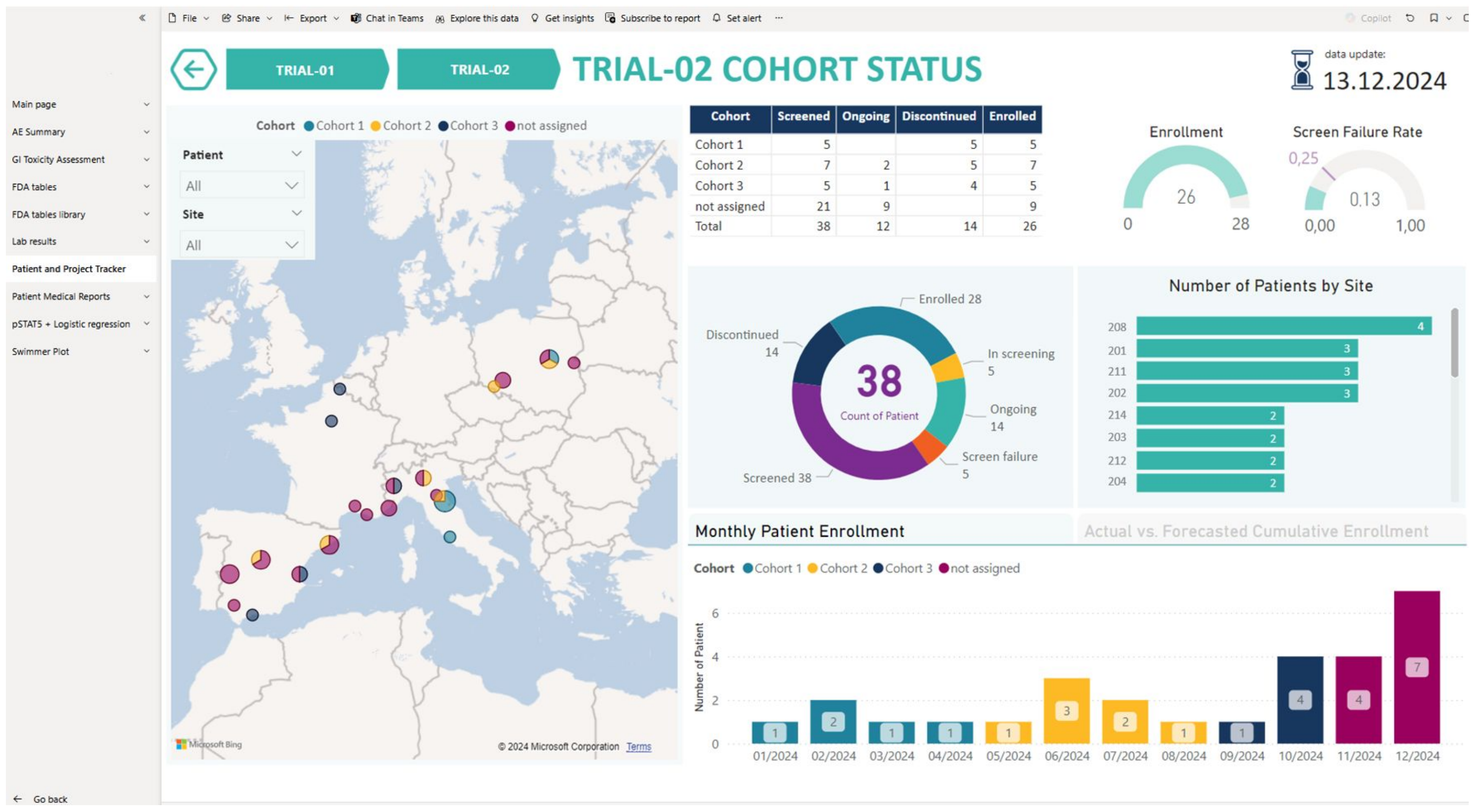


Figure 8.
The Cohort Status dashboard displays patient recruitment status per trial, providing valuable insights for monitoring the progress of multi-site and multinational clinical trials. The dashboard is implemented in Power BI and is particularly useful for the Clinical Operations team.



Cohort	Screened	Ongoing	Discontinued	Enrolled
Cohort 1	5		5	5
Cohort 2	7	2	5	7
Cohort 3	5	1	4	5
not assigned	21	9		9
Total	38	12	14	26