eBook

Guide to Data-Driven Innovation in the Life Sciences Industry

Improving the entire drug lifecycle with data and AI

databricks

Contents

Introduction 3

The opportunity for data and AI in life sciences 4

What challenges are holding back innovation? 6

How Databricks helps life sciences realize the power of data 8

Deliver better outcomes with data 12

Helping life science companies across the R&D lifecycle 13

Conclusion 17

databricks

Introduction

Bringing new treatments to market is a risky endeavor marked by high costs and even higher failure rates. For drugs that successfully make it to market, it can be equally as challenging to get the right treatment to the right patient at the right time. The opportunity to innovate has never been greater.

Organizations that are making data and AI a priority are leading the transformation to a new world of personalized medicine and care delivery. For example, genetic data, combined with robust phenotypic data, is leading to the discovery of novel medicines that help reduce the burden of disease. In other areas, social and behavioral data are being used to tailor interventions and improve access to and equity of care. Innovators that make data a priority are more equipped to predict disease onset, progression and intervention response, which is driving the triple aim: lower costs, better outcomes and improved patient experiences.

In order for life science organizations to truly realize the value of their data, they need technology and infrastructure that is capable of integrating and harmonizing vast and disparate data sources to feed downstream analytics. Additionally, companies must instill a culture of open collaboration across typically siloed functional areas to ensure optimal productivity.

Deploying predictive analytics during drug discovery, development and distribution can help life science organizations:

- More efficiently discover targets by using genetic markers
- Lower development costs by reducing the number of experiment runs
- Increase the efficiency and success rates of clinical trials
- Enhance supply chains with predictive maintenance
- Improve success rates of interventions in the real world
- Measure the safety and effectiveness of treatments post-launch

Massive opportunity to reduce costs and drive efficiencies

$1.4 billion and 12.5 years
average investment to launch a single successful drug

90%*
of potential treatments fail to reach trial stage

*www.lek.com/sites/default/files/insights/pdf-attachments/2060-AI-in-Life-Sciences.pdf

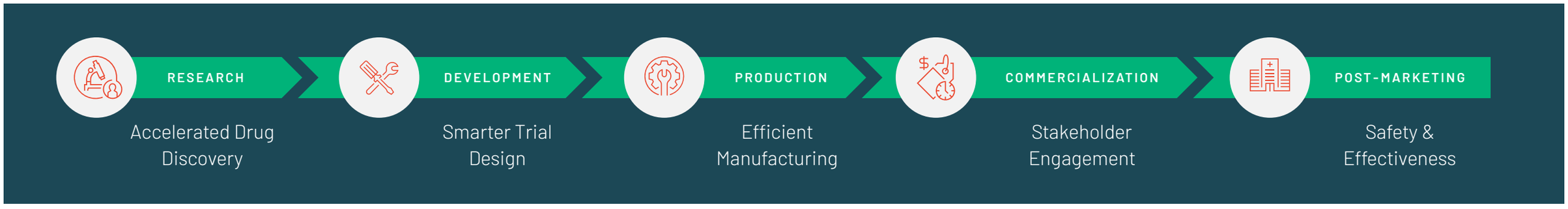
The opportunity for data and AI in life sciences

Data and AI will be critical success factors for life science organizations that aim to reduce the costs of drug development while also increasing the efficiency of delivering a drug or device to market. Bringing a new therapeutic to market even a month faster lowers trial costs, generates new revenue and, most importantly, gets life-saving medications into the hands of providers and patients quicker. The use of AI and machine learning (ML) can also enable organizations to achieve the promise of drug design that is more patient-centric: understanding how real-world human biology and drugs work outside a clinical trial, identifying how patients adhere to regimens, and identifying ways to better engage patients and providers to ensure better outcomes. There is an opportunity to realize benefits across the entire R&D and commercialization and post-launch lifecycle.

AI use cases across the drug development lifecycle

Life science companies see that their future depends on maximizing the value of the data that they and their strategic partners collect, and technology investments are critical to enabling this. AI spending alone in the healthcare and life sciences industries is expected to increase from the $463 million spent in 2019 to more than $2 billion over the next five years, according to ABI Research.

For life science organizations, the majority of their investments are going to be in R&D efforts, analysis of the massive volumes of data they collect from proprietary and public sources and, effectively taking their treatments to market so as to realize a maximum return on their development investments.



The opportunity for data and AI in life sciences

There are many opportunities for life science companies to leverage analytics and AI across the entire drug development lifecycles. Common use cases include:

RESEARCH

Greatly enhance the ability to discover new drugs and therapeutics faster and more cheaply

- Identify and validate high-quality targets using genotype-phenotype and gene expression screens
- Automate hit identification and profile using ML on high-throughput screens
- Generate efficient lead optimization using ML to build quantitative structure–activity relationship (QSAR) models

DEVELOPMENT

Optimize clinical trial protocols for speed and success

- Efficiently design optimized trials using real-world data
- Compute complex biomarkers in trials by using ML on imaging and IoMT data
- Manage clinical trial supply chains in real-time

PRODUCTION

Improve operational efficiencies to boost time-to-market and profitability

- Forecast seasonal demand for specific lines of medicines
- Identify and prescriptively mitigate manufacturing issues with ML on signals from IoT devices
- Monitor fulfillment systems to identify bottlenecks in dispensing medicines

COMMERCIALIZATION

Leverage actionable insights to augment the performance of marketing and sales

- Integrate CRMs with real-world data for an ML-based system that identifies next-best steps for sales teams
- Use real-world data to identify underdiagnosed patients and line extension opportunities
- Improve ad spend efficiency with ML models

POST-MARKETING

Ensure the safe and effective delivery of treatments

- Study populations to better understand real-world effectiveness
- Automate signal detection by analyzing population-scale electronic healthcare records

databricks

What challenges are holding back innovation?

The opportunity to drive innovation in life sciences is massive. But there are several common barriers that are holding organizations back from forward delivering those innovations.

It starts with the data

For many organizations, the challenges that are tied to data are twofold: volume and organization. It's estimated that a single individual will generate a staggering 1 million gigabytes of health-related data in their lifetime. This includes clinical data, imaging and genomic data. When you also begin to include the collection of streaming data from wearables, which are becoming ubiquitous, it's easy to see that this number is going to continue to grow.

From an organization standpoint, much of this data is often disorganized. Data may exist in different systems that are walled off from each other. There may be separate data storage systems for different therapeutic areas. Even if all data is stored in the same location, it isn't necessarily curated well. Data integration is one of the biggest challenges facing pharmaceutical teams, who usually have limited-to-no access to the data they need and often have to rely on IT teams to manually fulfill data requests, slowing the progress of work.

The ability to integrate disparate and often siloed sets of structured and unstructured data, regardless of the source, is fundamental to deriving maximum benefit from it. Yet most life science organizations still focus their attention on data warehouses designed for structured data. Unstructured data — such as medical imaging, genomics, IoT data and research reports — are oftentimes ignored, limiting the real-world insights that teams can feed back into their R&D efforts. Without access to complete and consistent data, downstream analytics and data science teams are unable to effectively innovate.

What challenges are holding back innovation?

Collaboration bottlenecks slow innovation

Collaboration is going to be just as important as data integration for life science organizations to fully realize a “digital transformation” of their industry. The challenge here is that it is very common for researchers, analysts and informatics teams — who should openly communicate — to find themselves working in silos. Analysts and data scientists are often stuck using homegrown infrastructure that limits the data sets they can explore and leads them to run experiments in isolation. These infrastructure problems make it hard to share analyses, combine data sources, reuse code and audit completed analyses. On top of that, most legacy platforms that serve as the analytics backbone for life science organizations lack the modern machine learning capabilities and workflow management tools that data scientists need to build, iterate and track AI models.

By eliminating these silos and adopting analytical tools that support seamless cross-team collaboration among stakeholders, life science companies can realize benefits across their organizations including in the areas of drug research, development, commercialization and delivery.

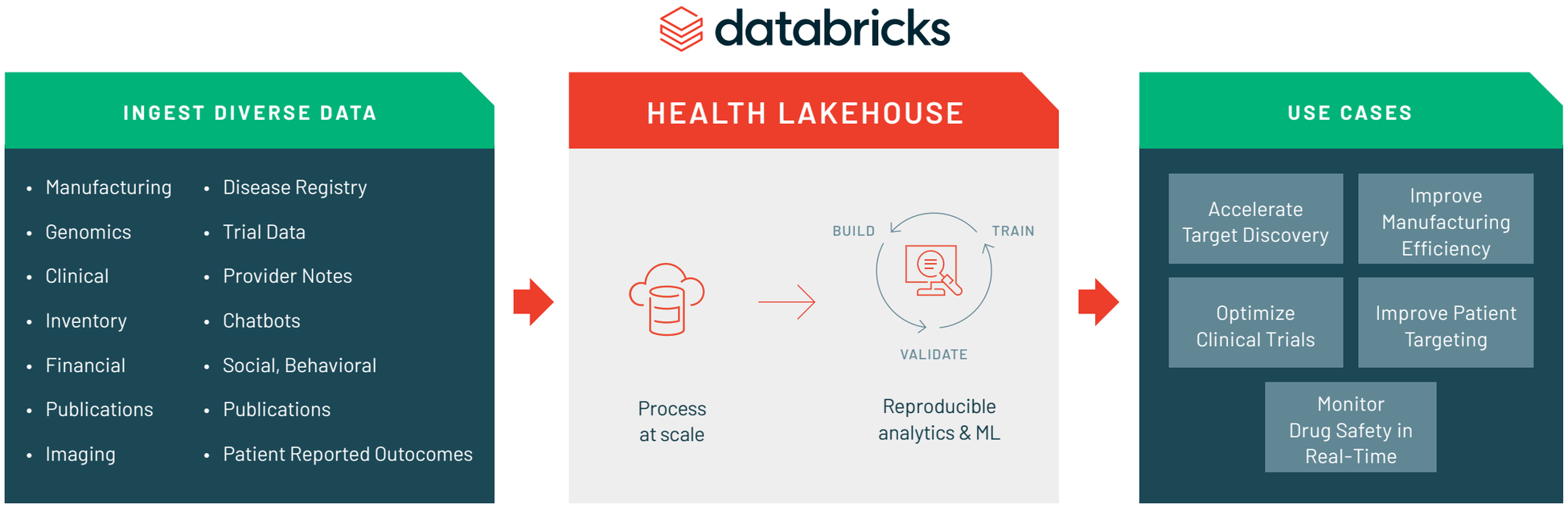
Managing R&D reproducibility

Reproducibility of findings from research and experimentation are a constant challenge for life science organizations. Being able to reproduce the results of long and expensive R&D efforts is critical to the success of pharmaceuticals and treatments.

Pharmaceuticals are high tech products that are the result of many complicated chemical and biological inputs, and much of the experimentation done in the research of these products occurs through intense computational methods. Unfortunately, while a scientist would normally never be caught in a lab without a lab notebook, a similar analog doesn’t exist for computational work.

How Databricks helps life sciences realize the power of data

Databricks provides life science companies with a Lakehouse Platform that combines the best of data warehouses and data lakes to store and manage all your data for all your analytics workloads. Databricks federates all data and democratizes access for downstream use cases. This in turn unlocks new ways to explore and analyze data to enable discovery, foster collaboration and maximize efficiencies across the drug development lifecycle.



How Databricks helps life sciences realize the power of data

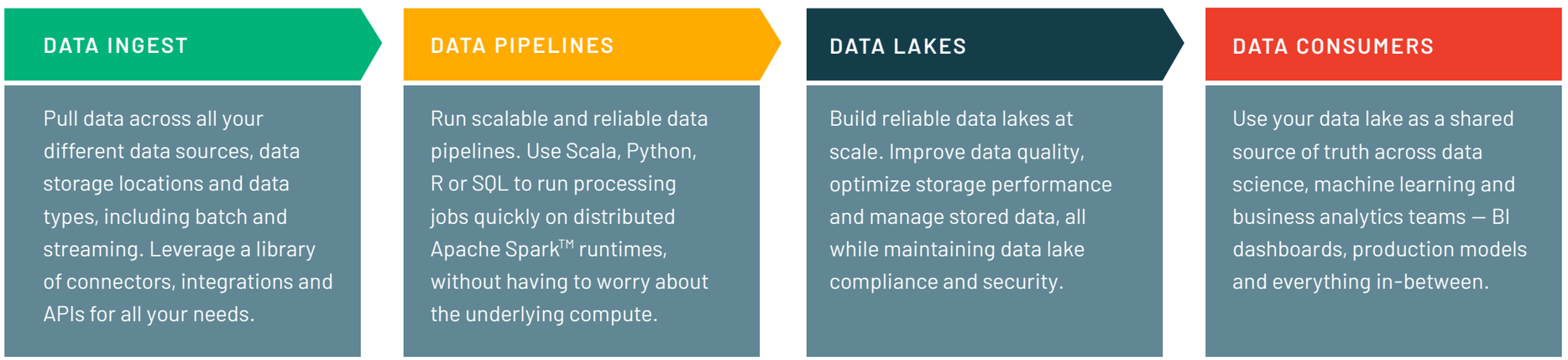
Bring your data into focus

Databricks enables the aggregation and processing of the massive collections of real-world data that currently exists in silos, both structured and unstructured. For many organizations, this requires a large infrastructure build out. With Databricks, you get an open, simple and collaborative platform for all your data and analytics. Built in the cloud, it lets you manage the full data analytics lifecycle in one single platform from ingest through insight without a costly and complex infrastructure footprint.

Building performant and reliable data pipelines at scale

One of the most common problems organizations face when dealing with massive volumes of real-world data across disparate sources is that it can become unreliable, low quality and challenging to manage. Many organizations turn to data lakes to aggregate their big data cost-effectively, but this creates a new set of challenges around data management and governance.

Databricks enables organizations to overcome these challenges with a lakehouse architecture powered by Delta Lake and Apache Spark.™ Delta Lake, an open source technology, is natively integrated within Databricks to provide reliability and performance. It acts as a storage layer on top of your data lake that enforces data quality with ACID transactions. Life science organizations can land structured, semi-structured and unstructured data, both in batch and streaming, to a single Delta Lake to ensure that the supply of data is clean and usable. The scalability of Databricks enables organizations to then process and query this data for near real-time insights.



Databricks empowers life science organizations to unify all forms of data for downstream analytics and machine learning.

databricks

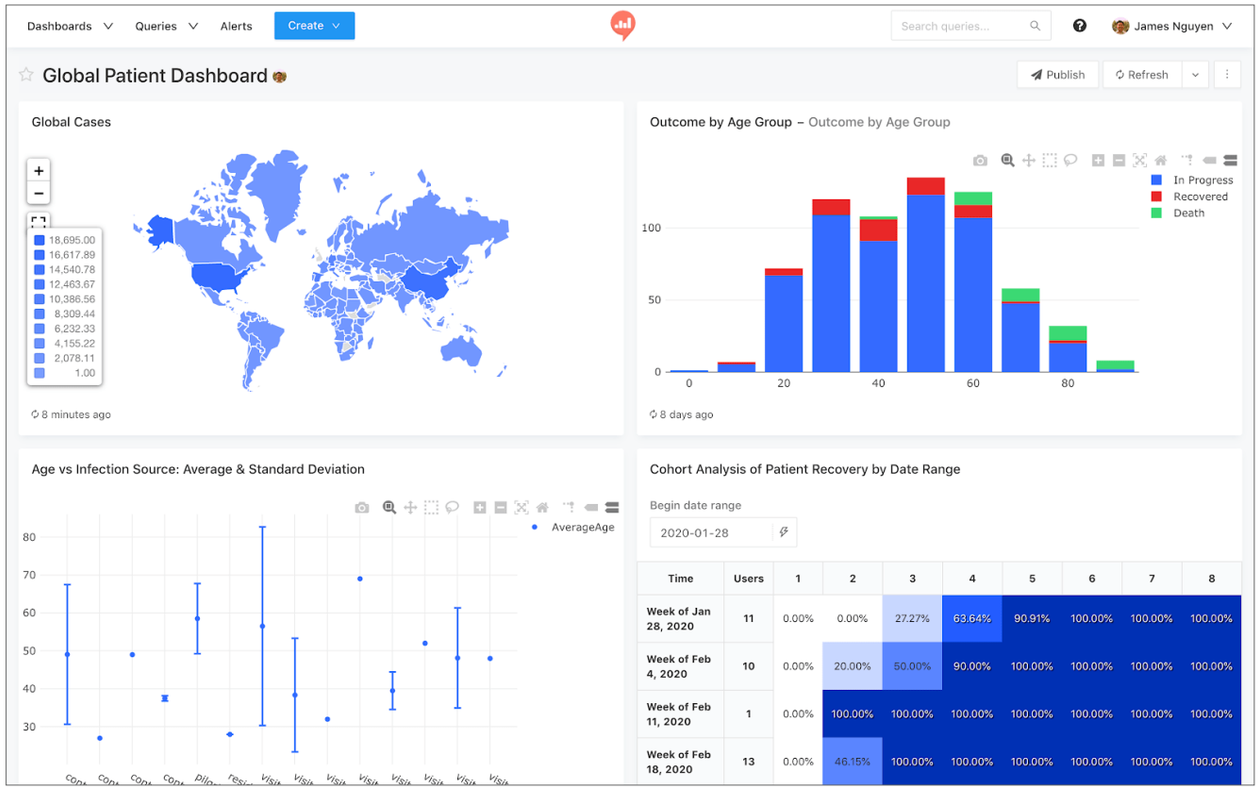
How Databricks helps life sciences realize the power of data

Extract scientific insights with collaborative analytics and AI

With all your data centralized, data teams can now work together to build powerful analytics and predictive models. Databricks provides shared interactive workspaces that enable teams — including data scientists, engineers, researchers and business analysts — to collaborate on data products with a wide range of analytics capabilities.

Data analysts can easily access the information they need in real-time using SQL Analytics. They can take advantage of built-in visualization capabilities or integrations with popular BI tools — including Power BI and Tableau — to publish and share dashboards across the business.

Additionally, data scientists can use popular languages that they are familiar with such as Python, R and Scala and built-in machine learning libraries to quickly build and iterate on models. Code and models can easily be shared in Databricks' collaborative environment making data science a team sport. With Databricks, data scientists can turn insights from real-world data into powerful visualizations designed for machine learning. And in turn, visualizations can be turned into interactive dashboards to share with clinicians, research scientists and decision makers, across the drug development lifecycle.



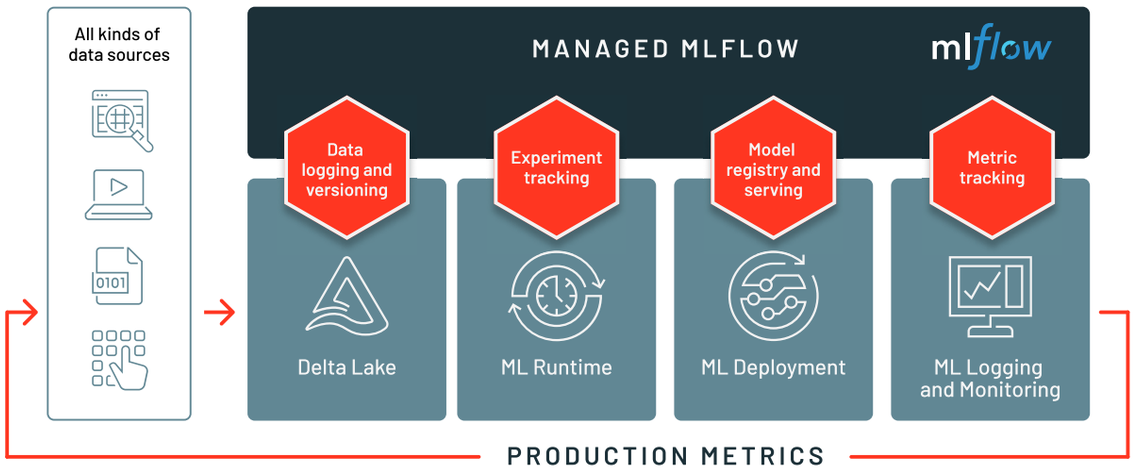
databricks

How Databricks helps life sciences realize the power of data

Automate tracking for clinical-grade reproducibility

Databricks workspaces include a managed version of MLflow, an open-source framework that streamlines the machine learning lifecycle. MLflow automatically tracks the computational lineage of all experiments, while Delta Lake tracks updates to data. This combination allows data scientists to reproduce a pipeline, compare the results of different versions, track what's running where, and explore past experiments to ensure reproducible results, which is critical when developing pharmaceuticals and therapeutics.

Data Science and ML Lifecycle, from data to deployment and back



Keep data secure and stay compliant

Databricks understands the sensitive nature of the proprietary and private data that life science organizations are responsible for. Data is protected at every level of the platform through fine-grained access control and deep integration with cloud-provider access control mechanisms. The independently audited Databricks cloud-based platform is compliant with FedRAMP High security assessment protocols on the Azure cloud and can provide a HIPAA-compliant deployment on both AWS and Azure clouds, and numerous global pharmaceutical companies use the Databricks platform to run GxP-validated workloads. The Databricks platform includes comprehensive technical safeguards covering access, encryption, auditing and many other stringent controls.

databricks

Deliver better outcomes with data

When life science organizations are fully able to realize the power of data, they will have the potential to transform their industry and have a massive impact on global health. The key is to unify both the data science and engineering challenges, enable collaboration and bring reproducibility to both experimental and clinical models. Stakeholders that are committed to innovation, who are willing to build their capabilities, and invest in the right infrastructure and tooling will be able to speed the process of drug discovery, drive innovation across the industry, reap the rewards of big data and help patients achieve better outcomes.

> “Databricks enables everyone in our integrated drug development process — from physician-scientists to computational biologists — to easily access, analyze and extract insights from all our data.”
>
> **JEFFREY G. REID, PH.D.**
> VP, Head of Genome Informatics at Regeneron

The Databricks Impact

Databricks helps companies automate infrastructure management, increase ETL performance at scale, and accelerate machine learning and analytics initiatives.

12x faster ETL pipelines

Impact: Faster time-to-market of new analytics insights and models

+25% gain in productivity

Impact: More productive data scientists result in more AI innovation

+47% Overall cost savings

Impact: Lower infrastructure costs boost operational margins

databricks

Helping life science companies across the R&D lifecycle

Early Research

Biogen

USE CASE: GENETIC TARGET IDENTIFICATION

Identify high-quality genes to target for neurodegenerative diseases like Alzheimer's and Parkinson's

CHALLENGE

Legacy on-premises infrastructure was unable to process the billions of data points contained within the UK Biobank data set

WHY DATABRICKS

Significantly accelerates data pipeline performance and supports the full spectrum of Biogen's needs, from data processing to large-scale statistical analysis and ML

OUTCOME

- 2 million genomic variants analyzed in 15 minutes
- 2 drug targets discovered for neurodegenerative diseases like Alzheimer's and Parkinson's

> "We really needed a new data paradigm for Biogen. Moving to Databricks and the cloud has helped us visualize and analyze our genomic data at petabyte scale."

DAVID SEXTON

Senior Director of Genome Technology and Informatics, Biogen

databricks

Helping life science companies across the R&D lifecycle

Translational Research

AstraZeneca

USE CASE: RESEARCH KNOWLEDGE BASE

Provide researchers with immediate access to a knowledge graph annotating targets and compounds with research papers and internal genomics/chemistry research findings

CHALLENGE

- Struggled to ingest, parse and make available to researchers their millions of data points across hundreds of data sources, including internal data sources and public sources, such as technical literature, public databases, etc.
- Unable to scale operations to support data science efforts with open source notebooks

WHY DATABRICKS

- Able to ingest a rapidly changing and diverse corpus of internal research data and public publications using NLP
- Built a reproducible, ML-based system for hypothesis generation and validation

OUTCOME

- Processed millions of data points from thousands of sources in minutes
- Built a recommendation engine that has improved their ability to make more informed hypotheses and accelerate time-to-market for novel drugs

> “By moving to Databricks, we have seen an order of magnitude improvement in performance.”

ELISEO PAPA
Computational Biologist, AstraZeneca

databricks

Helping life science companies across the R&D lifecycle

Clinical Trials

USE CASE: CLINICAL TRIAL OPTIMIZATION

Use real-world data to model patient selection criteria and site placement for clinical trials with a goal of designing smaller trials and improving prioritization of trial candidates

CHALLENGE

Legacy Hadoop-based architecture was costly and hard to manage requiring five admins focused solely on administration. Large effort spent on managing clusters, upgrades and disjointed tools instead of analytics.

WHY DATABRICKS

- Provides one, secure platform that accelerates both ETL and BI workloads for all their data (RWD and beyond)
- Shared notebook environment centralizes and accelerates experimentation while providing required traceability

OUTCOME

- Reduced infrastructure costs by 25% and FTEs required to manage the platform by 50%
- Improved recruitment with better models informing trial eligibility and real-world patient populations with the potential to provide millions in cost saving

> “Databricks has provided us with a platform that brings together our enterprise data lake and data science workbench in a way that is secure, easy-to-use and flexible enough to integrate with our technology ecosystem without compromising on performance or costs.”

DEEPAK ABBURI
Director of Data and Analytics, Amgen

Helping life science companies across the R&D lifecycle

Personalized Patient Care

USE CASE: PERSONALIZED PRESCRIPTION RECOMMENDATIONS

Make recommendations to patients to improve patient medication adherence, while increasing stability and predictability of CVS Health's prescription supply chain

CHALLENGE

CVS Health serves over 80 million customers across their 10,000 stores. They started their personalization efforts by targeting 1% of their customers. When they tried to scale to 5% of their customers, they hit a roadblock due to the lack of processing power and data storage provided by their legacy Hadoop-based architecture.

WHY DATABRICKS

- Cloud-based analytics platform that enables CVS Health to expand the number of use cases that they can support and scale their personalization efforts
- Collaborative environment with a full set of data analytics and ML tools improves team productivity and agility across data science, engineering and analytics

OUTCOME

By scaling their personalization efforts, CVS Health improved medication adherence by 1.6%, meaning an increased number of patients are now taking their medication on time and as directed.

> "Through Azure Databricks we have the flexibility to spin up clusters that meet our unique business needs and various business use cases. We're also not restricted by any more physical hardware constraints."

MICHELLE UN
Director of Enterprise Analytics, CVS Health

databricks

Conclusion

Integrating and deriving insights from complex data is critical to innovation in the life sciences. Databricks enables organizations across pharmaceuticals, biotech, genomics, infectious disease and even insurance to harness the power of data and analytics to solve the problems that are impacting the health of millions of people and to ultimately save lives.

Get started with a free trial of Databricks and start building data applications today

START YOUR FREE TRIAL

Learn more about our Life Sciences solutions: **dbricks.co/LifeSciences**

databricks