EBOOK

Improving Health Outcomes With Data and AI

Accelerate innovation in research and care with a modern data Lakehouse for Healthcare and Life Sciences

databricks

Contents

INTRODUCTION

An industry in flux: Patient-centric and data-driven

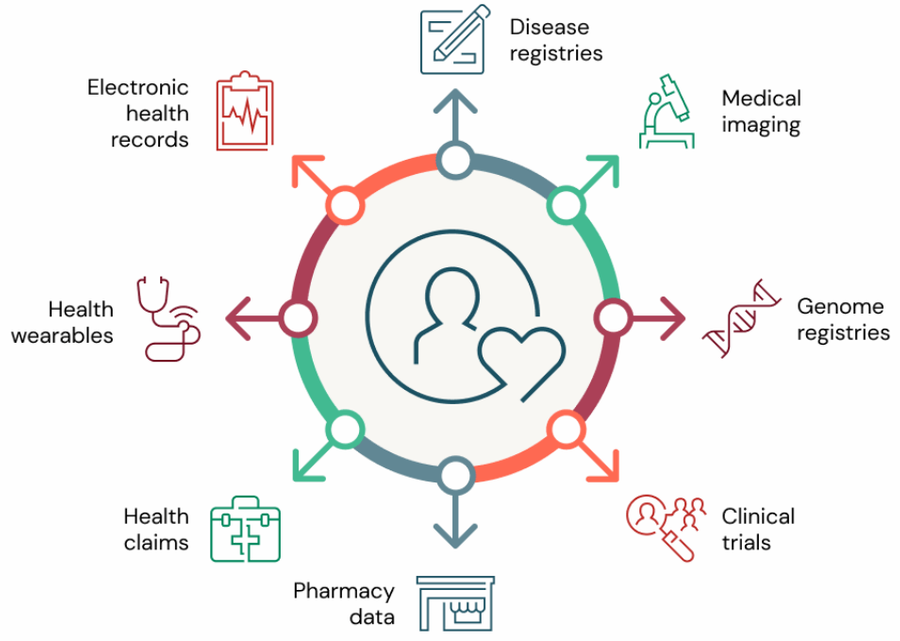
More than most industries, healthcare and life sciences has been constantly revolutionized by technology. Rapid advances in diagnostics, patient care and drug development have led to remarkable progress in improving patient outcomes and treating diseases — but there's still so much more to do.

Over the last decade, organizations across the health ecosystem have placed a major emphasis on lowering the cost of care and improving patient outcomes. Delivering on this promise requires a holistic view of the patient and innovation at every level, from drug discovery through the delivery of care.

Now we're on the cusp of a breakthrough in healthcare and life sciences. The explosion of digital technologies — such as electronic medical records, genome sequencers, IoT sensors, wearables and medical imaging — has ushered in a new era of big data. For example, a single patient produces over 80 megabytes of EMR and imaging data a year. Add in all the other types of health-related data such as wearables, claims, lab reports and genetics and the number is massive. It's no surprise that the average healthcare provider is sitting on over 9 petabytes of digital information. This data, while large and unwieldy, is critical to building a holistic view of the patient. By bringing this data together, organizations can better understand the factors that contribute to a positive or negative health outcome.

The key is making sure that teams that need this data have it at their fingertips — along with the analytics and machine learning tooling necessary to unlock innovative use cases. The need for a modern data analytics and AI platform has never been greater, driven by tectonic market shifts — many driven by the global pandemic.

It's estimated that a single person will generate more than 1 million gigabytes of health-related data in their lifetime



Healthcare and Life Sciences Transformation Trends

Healthcare is in a state of change and today's most successful organizations are tapping into the power of data and AI to respond to these five tectonic shifts:

TREND #1

Increasing Regulatory Pressure for Interoperability

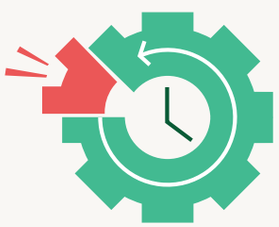
Healthcare interoperability is at the center of all transformation initiatives as it enables different IT systems to exchange health information. Simply put, interoperability is table stakes for building a holistic view of the patient.

Pressure from regulators to adopt industry standards for interoperability is on the rise. For example, the 21st Century Cures Act, passed by the U.S. Congress in 2016, requires every payer and provider to make health data accessible via real-time APIs. Additionally, the U.S. Centers for Medicare & Medicaid Services made the Fast Healthcare Interoperability Resources (FHIR) its standard for data exchange, helping drive adoption across the industry. And it's not just in the United States. Similar global regulations and adoption of FHIR are helping to unlock interoperability across borders.

All organizations — whether they are looking to innovate drug R&D or patient care — need to consider how their data platforms support these standards and work toward data consolidation internally and with their partners.

Interoperability continues to be a challenge

Only 18% of providers and 36% of payers have expertise in FHIR



80% of providers say the pandemic has impacted their ability to comply with new rules

TREND #2

Growing Acceptance of Real-World Evidence

Real-World Evidence (RWE) is generated from any data collected outside a clinical trial, such as claims, clinical notes and medical images. This data, called Real-World Data (RWD), has huge potential to provide new insights into the usage, effectiveness and risks of a drug.

Regulators around the world recognize the benefits of RWE. The U.S. Food and Drug Administration published guidance in September 2021 on the proper use of real-world data. This builds on the 21st Century Cures Act, which signaled U.S. regulatory acceptance of RWE back in 2016. In Europe, the Heads of Medicines Agencies – European Medicines Agency (HMA-EMA) are working on similar initiatives and workshops to govern the proper use of real-world data.

With RWE entering the mainstream, life sciences organizations need to invest in flexible data platforms that support unstructured RWD, and data standards (e.g., OMOP common data model) to facilitate population-level analysis. Additionally, RWE requires governance with lineage for all data transformations so regulators can recreate analyses. Organizations adopting RWE need to ask if their data platforms provide this level of transparency.

TREND #3

Advancements in Personalized Medicine and Prevention

The pace of personalization in healthcare and life sciences is accelerating. Consider: The first human genotype cost more than $1 billion to sequence. Today, individual whole-genome sequencing is available to consumers via mail for $300. And in 2021, the UK Biobank released hundreds of thousands of whole-genome sequences for research.

Genomic analysis is essential for drug development and increasingly important in informing patient care. In short, genomics is core to personalization. Like genomic data, image data is also becoming more readily available and is essential to the diagnosis and treatment of the fastest-growing therapeutic areas such as oncology, immunology and neurology.

Another major trend in personalized care are liquid biopsies. These tests are done on samples (typically blood) to detect the presence of disease markers. Early, noninvasive detection is one of the best approaches to managing cancer and will likely grow in importance. With all these advancements in personalized care, healthcare and life sciences organizations should expect patient data to grow tremendously in the future.

TREND #4

Displacement and Digitization of Patient Care

Accelerated by the pandemic, the displacement of care from traditional settings — like primary care offices and hospitals — to the home is on the rise. At the same time, the digital health market, experienced both through telemedicine as well as wearable devices, is exploding. Regulators are getting on board too with the U.S. Centers for Medicare & Medicaid Services' expanding reimbursable services for telehealth.

Livongo, for example, offers behavioral recommendations through an ML-engine running on Databricks that pulls in insights from connected health devices like continuous glucose monitors. Innovative technologies and services are becoming the norm as we move to virtual care settings. Underpinning these initiatives is the need for data platforms that can collect, curate and analyze streaming patient data in real time.

CHANGING PERCEPTIONS OF TELEHEALTH

38x
Telehealth services are up 38x compared to pre-pandemic

40-60%
of consumers want broader virtual health solutions depending on the service

58%
of physicians now view telehealth more favorably

TREND #5

Unpredictable Demand and Supply Chain Disruptions

Perhaps more so than other industries, COVID-19 upended the historical dynamics of supply and demand in healthcare. Providers witnessed unmanageable spikes in ICU bed demand, nursing shortages and massive gaps in COVID-19 diagnostic tests. At the same time, the biopharmaceutical industry has grappled with the challenges of an inadequate cold chain ill-equipped for mass vaccine distribution. Most organizations struggle to even do real-time reporting on data from legacy systems like SAP and Oracle ERP, creating gaps in inventory moving through distribution centers.

Barriers to Data-Driven Innovation

There's no shortage of data or the need for data-driven innovation. Yet, most healthcare and life science organizations struggle to tap into the full potential of data and AI. There are five common challenges hindering their success:

CHALLENGE #1

Scaling for Rapidly Growing Health Data

Healthcare data is growing exponentially. For example, just one human genome sequence produces approximately 200 gigabytes of raw data. Multiply this across millions of patients and that number is massive. And it's not just genomics data that is huge. A single individual is expected to generate a staggering 1 million gigabytes of health-related data in their lifetime. This includes clinical data, imaging health wearables and more.

Unfortunately, legacy on-premises data architectures are complex to manage and costly to scale for today's massive volumes of healthcare data. Scaling analytics across large population data sets is nearly impossible with today's limitations, inhibiting innovations in care and research.

CHALLENGE #2

Building a Holistic View of the Patient

Healthcare and life science organizations deal with a tremendous amount of data variety, each with its own nuances. It is widely accepted that over 80% of medical data is unstructured, yet most organizations still focus their attention on data warehouses designed for structured data and traditional SQL-based analytics.

In the context of healthcare, unstructured data includes medical images, which are critical to diagnose and measure disease progression, and narrative text in clinical notes, which are critical to understanding the complete patient health and social history. Ignoring these data types, or setting them to the side, is not an option.

Some organizations have invested in data lakes to support unstructured data and advanced analytics, but this creates a new set of issues. In this environment, data teams need to manage two systems — data warehouses and data lakes — where data is copied across siloed tools resulting in data quality and management issues. Without a unified platform for all data, organizations cannot create a holistic view of patient health.

CHALLENGE #3

Delivering Real-Time Insights

Healthcare decisions can be a matter of life and death, and health conditions can change quickly. Weekly or daily batch data processing — which is commonplace for most organizations — is not good enough. Access to the latest, up-to-the-second information is critical for interventional care. Streaming data needs to be made available for everything from predicting sepsis to real-time forecasting of ICU bed usage.

The use cases for real-time data extend to life sciences as well. Consider temperature sensitive medications like vaccines. Streaming data from IoT sensors is critical to monitoring the production, storage and transportation of these life-saving treatments. An unnoticed outage of refrigeration equipment along the distribution line can mean those in need never receive their medication.

Unfortunately, traditional data platforms, such as data warehouses, were not designed to operate in real time. And disconnected platforms for data processing, analytics and AI further slow down the analysis and dissemination of critical information.

CHALLENGE #4

Overcoming the Complexities of ML

Big data and machine learning have demonstrated their promise in enhancing multiple areas of care and pharmaceutical research, such as population health and precision medicine. Unfortunately, many organizations still struggle to embed ML within their operations.

The challenges are twofold. For one, the legacy analytics platforms that underpin healthcare and pharma organizations typically lack robust data science capabilities. Secondly, the tools used in machine learning are complex and hard to manage. Organizations brave enough to build a modern data science architecture need to support the many open source frameworks (TensorFlow), libraries (matplotlib), scripting languages (R, Python, Scala or SQL) and IDEs (JupyterLab, RStudio) that data science teams require at each stage of the machine learning lifecycle.

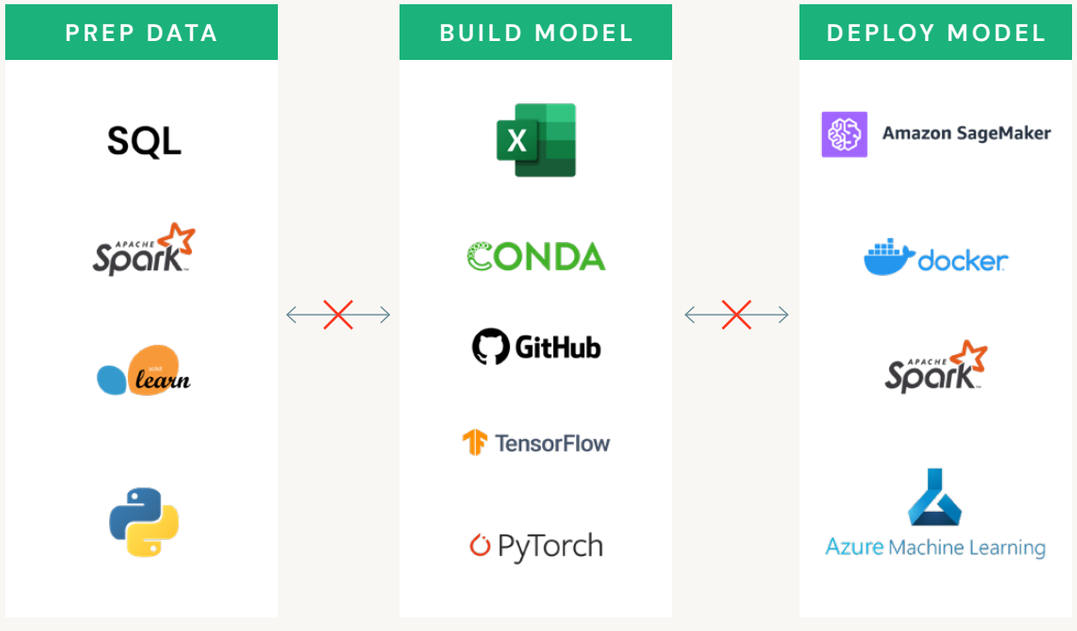
CHALLENGE #5

Ensuring Clinical Data Governance and Reproducibility

With patient lives on the line, clinical and regulatory standards demand the utmost level of data accuracy. Healthcare and life science organizations are required to meet stringent public health compliance requirements. Data that is used to inform healthcare decisions must be tracked, transparent and well-governed. Additionally, organizations need good model governance when bringing machine learning into a clinical or research setting.

Unfortunately, most organizations have separate platforms for data science workflows that are disconnected from their data warehouse. Data is copied across systems and then transformed to meet the needs of the requesting team. This creates significant data lineage challenges when trying to build trust and reproducibility in AI-powered applications.

Managing the diversity of ML tools is overwhelming

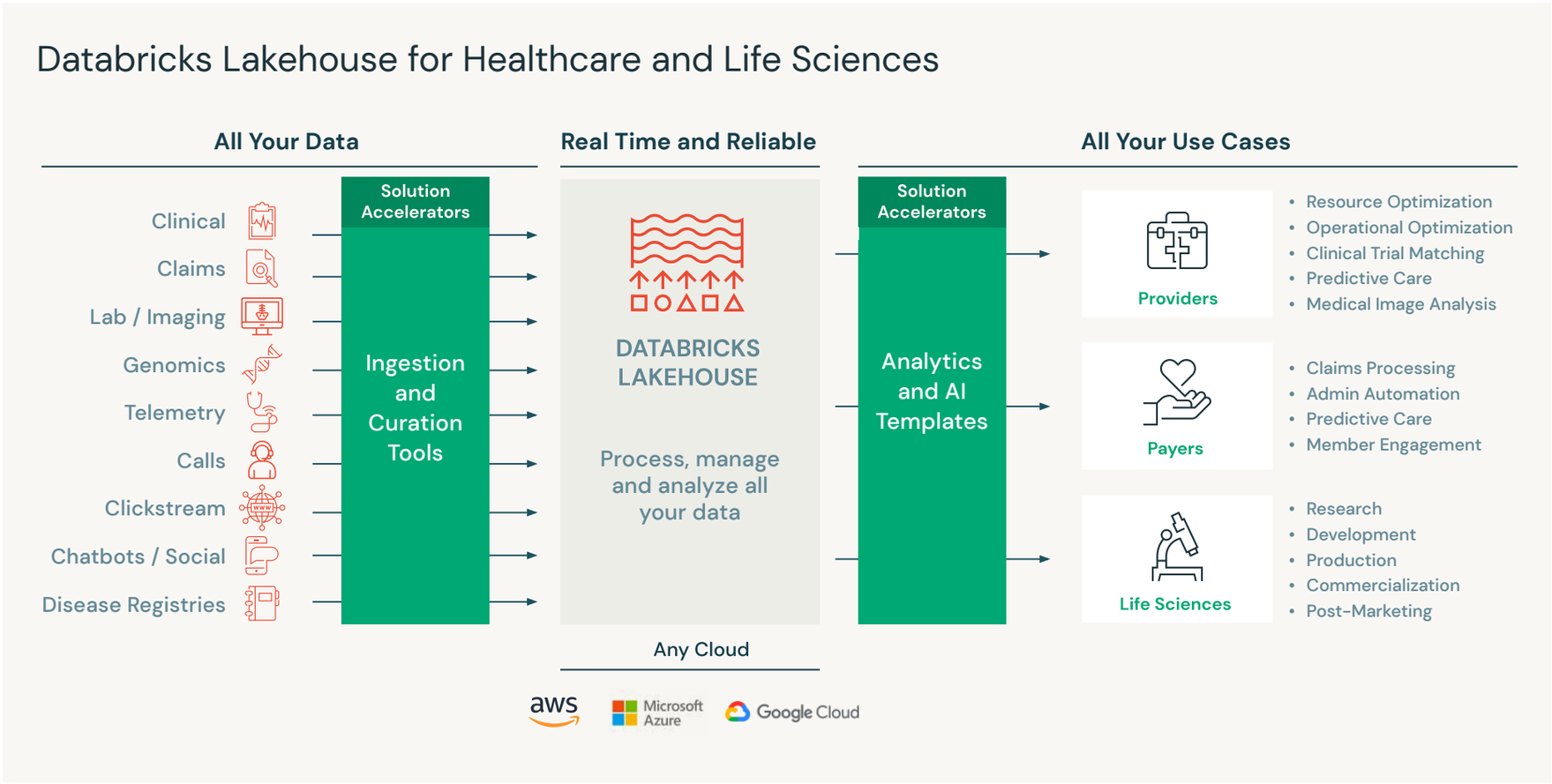


Unlocking Innovation With the Lakehouse for Healthcare and Life Sciences

For organizations interested in data-driven transformation, there is a path forward — introducing the Databricks Lakehouse Platform for Healthcare and Life Sciences.

The Lakehouse for Healthcare and Life Sciences enables organizations across the health ecosystem to work together to improve health outcomes with a single and collaborative platform for data, analytics and AI. With these capabilities, organizations can leverage all their data to build a holistic view of the patient, make real-time decisions and drive innovation with advanced analytics.

Building on this foundation are Solution Accelerators for common analytics and AI use cases. These solutions are developed by Databricks and our ecosystem of partners to accelerate the delivery of analytics projects and provide measurable outcomes.



databricks

Platform Benefits

Infinite scale for population-level studies

Built in the cloud and designed for high performance, the Lakehouse for Healthcare and Life Sciences supports the largest of data jobs at lightning-fast speeds. For example, Regeneron reduced data processing from 3 weeks to 5 hours, and genotype-phenotype queries from 30 minutes to 3 seconds for workloads that scaled to 1.5M exomes. With these capabilities, organizations can quickly and reliably analyze data for millions of patients.

DIFFERENTIATED CAPABILITIES:

- **Performance at scale:** With Apache Spark™ and Delta Lake — the leading open source engines for large-scale data processing and data management — under the hood, the lakehouse delivers massive scale and speed. And because it's optimized with performance features like indexing and caching, Databricks customers have seen ETL workloads execute 48x faster.

- **Elastic compute:** Scalable cloud compute resources are available at the click of a button to meet the demands of any size job. Autoscaling clusters scale up or down based on the size of your workload so you only use as much processing power as needed to meet the demands of your workloads.

360° view of the patient

Unify all your data — patient, R&D and operations — in a single platform to unlock innovations in personalized care and therapeutic design. Unlike a traditional data warehouse, the lakehouse supports all types of structured and unstructured data enabling organizations to build a holistic view of patient health. To make data curation easy, Databricks and our partners built data ingestion tools for domain-specific data types.

DIFFERENTIATED CAPABILITIES:

- **All data types:** Support for all types of structured, unstructured and semi-structured data with Delta Lake and Apache Spark™ at the foundation.

- **Ingestion and curation Solution Accelerators:** Databricks and partners like John Snow Labs provide a suite of notebook templates that make it easy to ingest common data formats, such as HL7 messages, FHIR bundles, medical text and imaging. The lakehouse also supports common data models such as OMOP for real-world data.

- **Open formats and open data sharing:** Data is stored in Delta Lake using an open source data format that prevents vendor lock-in. Additionally, Delta Sharing provides an open source data-sharing capability that promotes collaboration.

> **Databricks is enabling 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 Reid**
> Vice President, RGC Chief Data Officer, Regeneron

databricks

Real-time insights for real-time operations

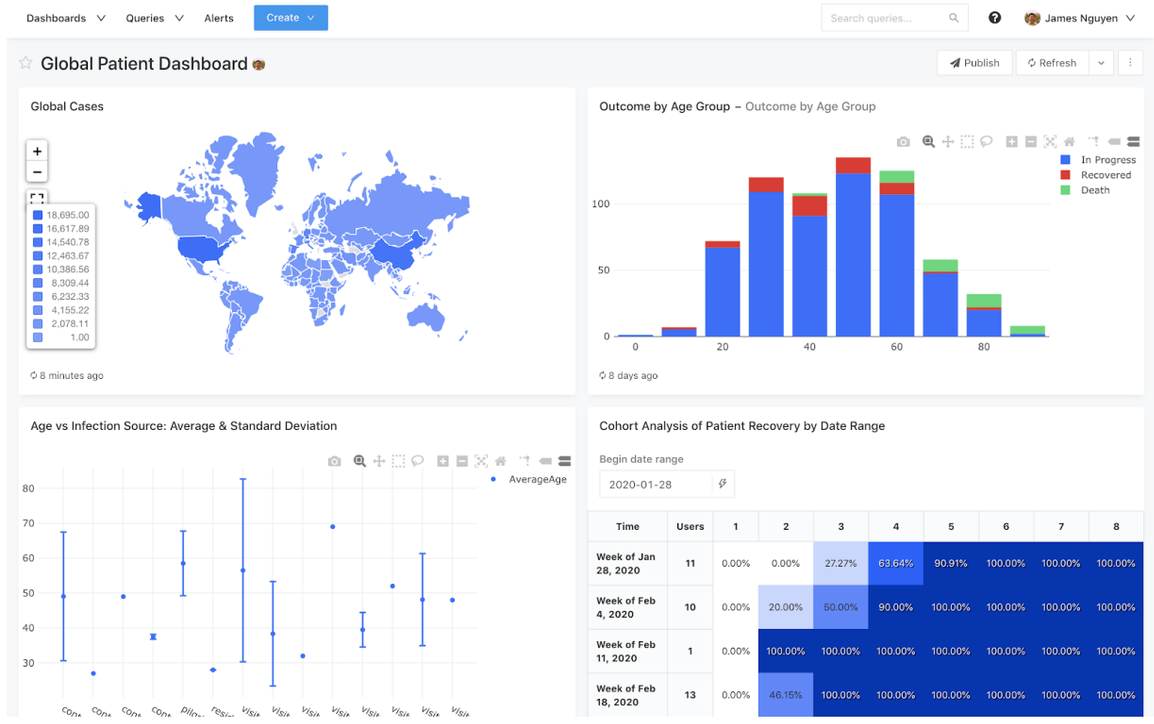
In no other industry, are real-time insights more critical than healthcare. Take advantage of rapid data ingestion at scale to support real-time analytics use cases from interventional care (e.g., predicting sepsis) to optimizing supply chains on the fly.

DIFFERENTIATED CAPABILITIES:

- **Real-time data ingestion:** The lakehouse allows for true real-time stream ingestion of data, and even analytics on streaming data. Data warehouses require the extraction, transformation, loading, and then additional extraction from the data warehouse to perform any analytics.
- **Simplified streaming and batch architecture:** The lakehouse event-driven architecture provides a simpler to develop and manage method of ingesting and processing batch and streaming data than legacy approaches, such as lambda architectures. This architecture handles the change data capture and provides ACID compliance to transactions.
- **Interoperability Solution Accelerators:** Solution Accelerators and open source libraries — such as Smolder for HL7 messages — built by Databricks and our ecosystem of partners support real-time patient insights.

ML-powered drug discovery and patient care

Unlock the power of machine learning to better understand disease and predict health needs. With all your data centralized and seamlessly connected with a full suite of collaborative analytics and machine learning tools, data teams can work together to build powerful predictive models that drive new innovations in care delivery and drug research.



- **Collaborative data science:** The lakehouse provides an interactive notebook environment that enables cross-functional teams — including data scientists, engineers, researchers, clinicians and business analysts — to collaborate on data products with a wide range of analytics and ML capabilities, including support for multiple languages (R, Python, SQL and Scala) and popular ML libraries.
- **Publish powerful dashboards:** Make insights consumable with interactive visualizations and publish as dashboards to teams in the field — be it in the ER or conducting clinical research — so they can stay abreast of the latest findings.

databricks

- **Easily manage the ML lifecycle:** Manage the complete ML lifecycle from model development through deployment with Managed MLflow, an open source platform developed by Databricks to help streamline machine learning. Centralize models and features in the registry to help teams collaborate, iterate and reuse existing work.

- **Analytics and AI Solution Accelerators:** Solution Accelerators built by Databricks and our partners help teams deliver value faster with quick-start notebook templates. Solutions are available in a number of areas including medical image analytics, healthcare natural language processing, drug R&D, and population health.

Healthcare analytics you can trust

Deliver analytics and AI with data governance, lineage and model reproducibility. The Databricks Lakehouse Platform includes capabilities missing from traditional data lakes like schema enforcement, auditing, versioning and fine-grained access controls. This helps bring data governance to your big data projects.

An important benefit of the lakehouse is the ability to perform both analytics and ML on this same, trusted data source. Additionally, Databricks provides ML model tracking and management capabilities to make it easy for teams to reproduce results across environments and help meet compliance standards. All these capabilities are provided in a secure analytics environment.

databricks

Solution Accelerators

Databricks and our ecosystem of partners have packaged Solution Accelerators to help organizations derive value from their lakehouse projects faster.

- **Data Ingestion and Curation Tools:** Easily ingest domain-specific data modalities into your lakehouse and prepare for analytics at scale with common data models like OMOP.
- **Analytics and AI Templates:** Quick-start templates for high-value analytics and AI use cases, such as digital pathology, genetic association studies, and drug repurposing.

Featured Partner Solutions

accenture	Lovelytics	John Snow LABS	ZS
Intelligent Drug Repurposing	Interoperability	Natural Language Processing for Healthcare	Biomedical Research Intelligent Data Management
Identify new therapeutic uses for existing drugs with the power of data and machine learning.	Automate the ingestion of streaming FHIR bundles into your lakehouse for patient analytics at scale.	Extract insights from unstructured text for use cases such as PHI removal, adverse drug event detection, and oncology evidence generation.	Improve biomarker discovery for precision medicine with a highly scalable and extensible whole-genome processing solution.

> **One of the biggest challenges facing healthcare organizations today is building a comprehensive view of the patient. The Databricks Lakehouse for Healthcare and Life Sciences is helping GE Healthcare with a modern, open and collaborative platform to build patient views across care pathways. By unifying our data in a single platform with a full suite of analytics and ML capabilities, we've diminished costly legacy data silos and equipped our teams with timely and accurate insights.**
>
> **Joji George**
> Chief Technology Officer
> LCS Digital, GE Healthcare

Real Examples of Lakehouse Success

Healthcare and life sciences organizations are unified around a single goal: improve health outcomes. The role of data and AI in delivering on that mission has never been more critical.

The Lakehouse Platform for Healthcare and Life Sciences is enabling organizations across the health ecosystem to collaborate and unlock data-driven innovation from drug discovery through patient care and beyond.

Featured Customers

Biogen

Analyzed 2 million genomic variants on Databricks in under 15 minutes. This research led to the discovery of 2 new drug targets for neuro-degenerative diseases like Alzheimer's and Parkinson's.

SANFORD HEALTH

Modeled large volumes of patient data (e.g., images, genomics, EHR) on Databricks to provide clinicians with genetically informed disease risk reports that improve care planning.

SANOFI

Applied machine learning to 17M+ electronic health records to identify new treatment indications for approved therapies while reducing data processing costs by 30%.

REGENERON

Built one of the most comprehensive genetics databases on Databricks with 1M+ exomes. Reduced data processing time from 3 weeks to 5 hours and accelerated genotype-phenotype queries by 600%.

Livongo

Delivering real-time recommendations to patients using streaming data from connected health wearables for diabetes management.

databricks

Databricks customers across the industry

AstraZeneca · Bayer · Cerner · gsk · Merck · novo nordisk · Eisai · Genome Institute of Singapore, GIS · CRISP · Zeiss · healthgrades. · healthdirect · Austin Health · Humana

About Databricks

Databricks is the data and AI company. More than 5,000 organizations worldwide — including Humana, Sanford Health, Amgen and over 40% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark,™ Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.

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

START YOUR FREE TRIAL

To learn more, visit us at:

Healthcare and Life Sciences Industry Solution

databricks