USE CASE GUIDE

Eight Use Cases for Healthcare and Life Sciences

databricks

Contents

Introduction

The adoption of data, analytics and AI is helping to drive massive transformation among healthcare providers and within life sciences companies. The range of possibilities across sectors is pushing the boundaries of how healthcare is delivered — from genomics and real-world evidence to predicting disease risk and personalizing patient experiences — as these organizations look to streamline R&D, mitigate barriers to equitable treatment and improve patient outcomes.

The Databricks Lakehouse Platform for Healthcare and Life Sciences allows organizations to bring together all their patient, research and operational data using powerful analytics and AI capabilities to deliver real-time insights at population scale. As one of the fastest-growing segments at Databricks, HLS customers on Databricks Lakehouse span all sectors of the industry, including pharmaceuticals, healthcare providers and payers, and public health agencies. In this guide, we highlight some of these use cases across the industry, and how Databricks helps these organizations get value from their data and AI initiatives.

Lakehouse for Healthcare and Life Sciences

Overview

The healthcare industry is one of the biggest producers of data. The average healthcare organization owns nearly 9 petabytes of medical data. The rise of electronic health records (EHR), digital medical images and wearables is contributing to this data explosion. With the availability of advanced computational tools and cloud resources, the potential to learn from this population-scale data is massive. For example, a patient's longitudinal records can be used to illuminate the patient's health history as well as predict health outcomes. The lakehouse architecture enables healthcare and life sciences organizations to combine all their data in a cost-efficient and reproducible manner to support use cases ranging from automated reporting to machine learning (ML). Access to such fine-grained data ensures a patient's individual needs are taken into account when planning treatment, resulting in better health outcomes.

Common challenges

As healthcare and life sciences organizations accumulate petabytes of heterogeneous data, they face many challenges in getting value from these data sets. Some of the most common are:

Fragmented patient data
Data silos and limited support for unstructured data prevent organizations from understanding the patient journey

Rapidly growing health data
Legacy on-premises data architectures are complex to manage and costly to scale for today's massive volumes of healthcare data, including the growth in imaging and genomics

Real-time care and operations
Data warehouses and disjointed sets of tools impede the delivery of real-time insights needed for critical care decisions and the safe production and delivery of important therapeutics

Complex advanced health analytics
Lightweight ML capabilities are preventing organizations from tackling everything from next-gen patient care models to predictive analytics for drug R&D

Lakehouse overview

The lakehouse paradigm, by design, brings the best of the two worlds of data lakes and data warehouses together: Data lakes meet the No. 1 requirement for healthcare and life sciences organizations, which is to store and manage different data modalities and allow advanced analytics methods to be performed directly on the data where it sits in the cloud. Data warehouses, which allow for quick access to data and support basic analytics and reporting applications, provide governance and compliance. The lakehouse architecture allows organizations to perform descriptive and predictive analytics on the vast amounts of healthcare data directly on cloud storage while ensuring regulatory-grade compliance and reproducibility.

Value with Databricks

By design, the Databricks platform is optimized to support an enterprise data lakehouse:

Collaborative environment

The Databricks platform allows scientists and engineers to use a common workspace that promotes collaboration and creates a smooth handoff experience. For example, clinical data scientists can create a model for patient-risk scoring based on a common data model produced by the data engineering team.

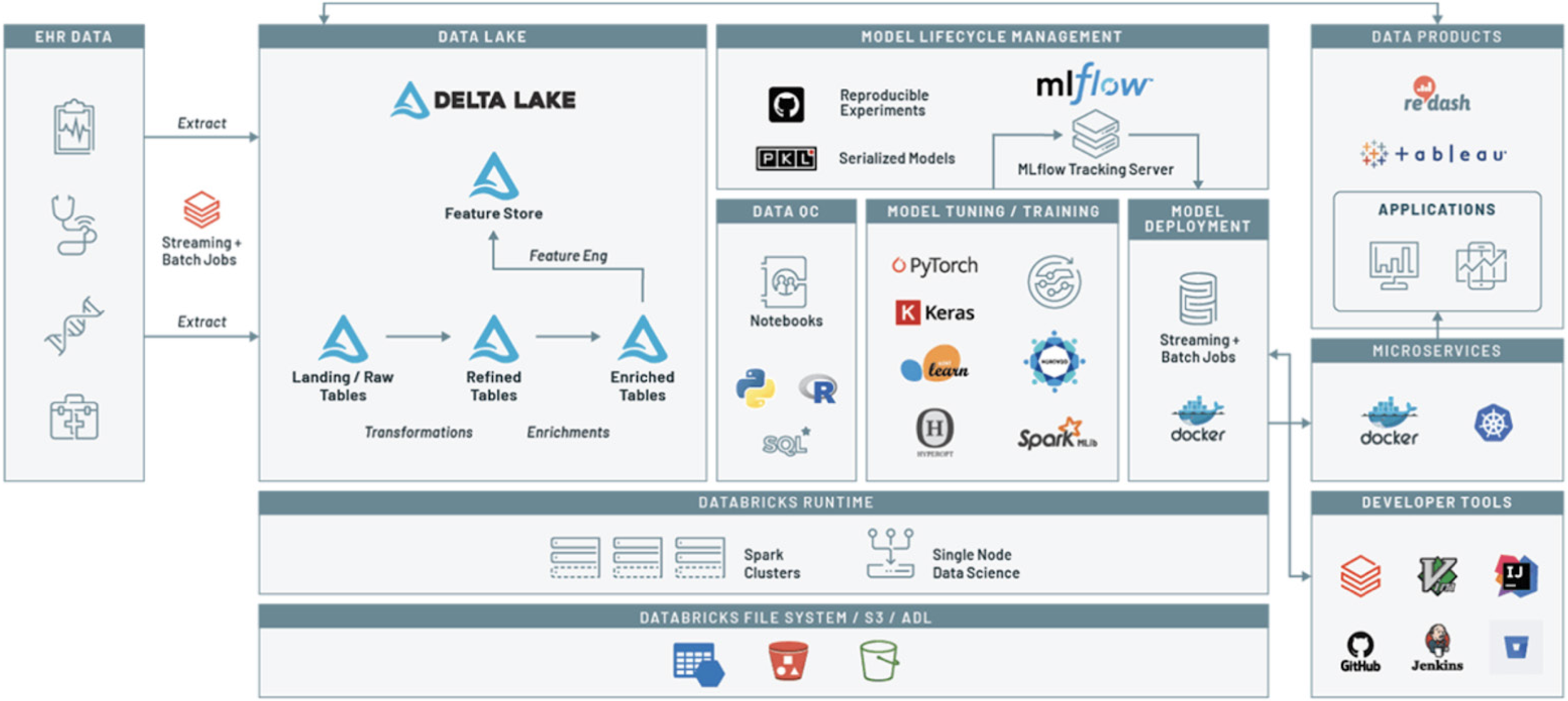
Fully managed platform

Databricks is fully integrated with all of the components of the lakehouse architecture. This integration helps free up resources, such as MLflow, to focus on solving business problems instead of hosting software. For example, Delta Lake's time travel features, combined with integrated MLflow APIs that allow for experiment tracking and logging, as well as managed runtimes, ensure the highest level of reproducibility of experiments, without the need for writing additional code.

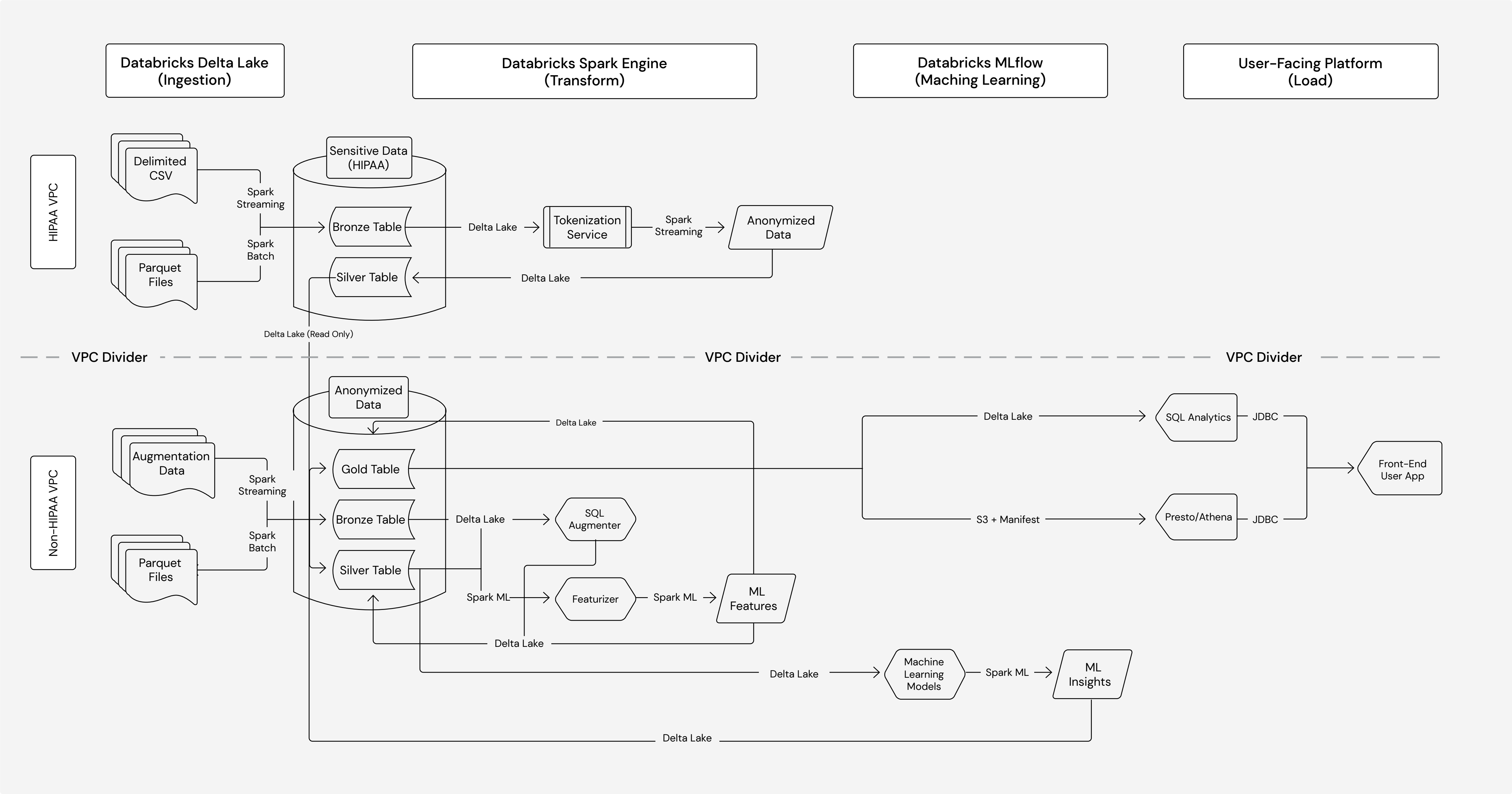
Reduced infrastructure costs

Databricks is deployed on your existing cloud infrastructure and provides you with the tooling needed to manage your compute resources effectively (e.g., auto-scaling)

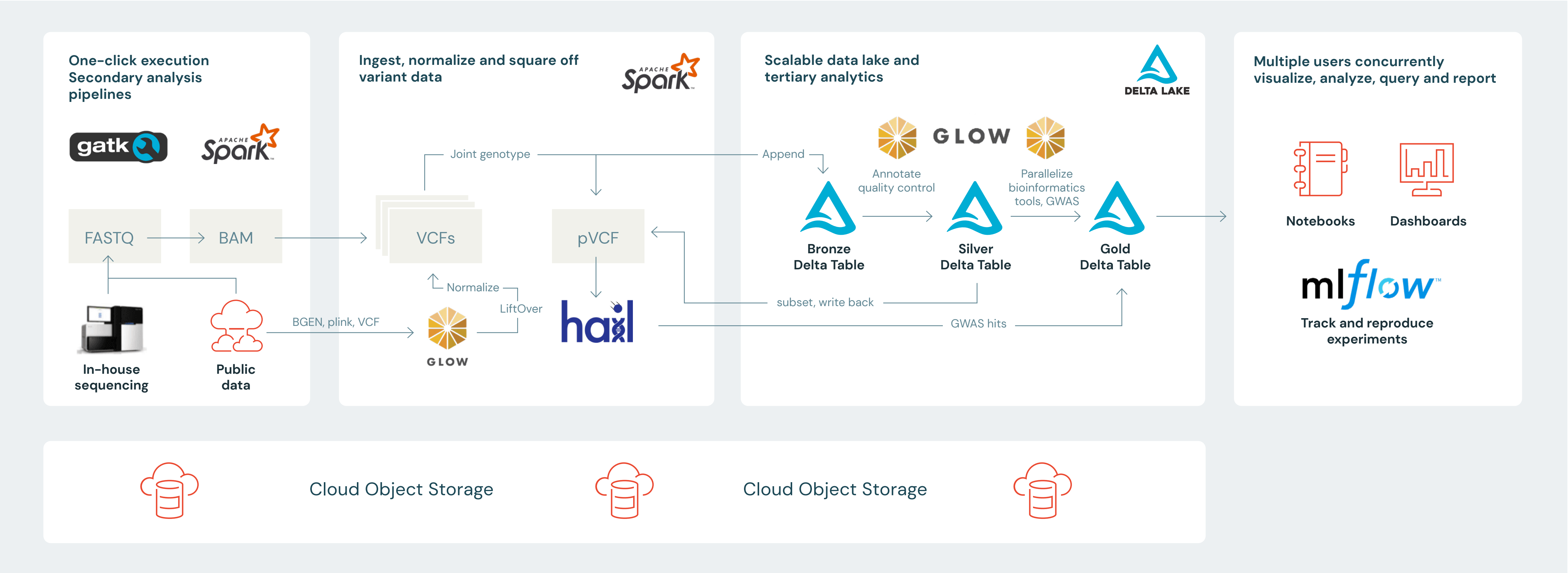
Reference architecture



HIPAA reference architecture



Genomics reference architecture



One-click execution Secondary analysis pipelines

Ingest, normalize and square off variant data

Scalable data lake and tertiary analytics

Multiple users concurrently visualize, analyze, query and report

Track and reproduce experiments

Cloud Object Storage

Cloud Object Storage

USE CASE: Genetic Target Identification

Overview

Research and development is one of the most costly stages of drug development. On average, every new drug costs up to $2.6 billion to develop, of which more than 50% is attributed to R&D. Typically, researchers discover new drugs through new insights into a disease process, which in turn allows them to design a product to stop or reverse the effects of the disease.

One of the approaches in understanding the disease mechanism is to identify the genes underlying the disease, which can lead to the characterization of the molecular mechanisms addressed by the target. Studies have shown that drugs that have been developed starting from a genomics approach are twice as likely to succeed in phase II clinical trials. Taking into account the fact that approximately 70% of phase II trials fail, this increased success rate translates into a huge R&D cost savings.

Challenges

- **Lack of scalable workflows**
  Legacy tools for multiomics are primarily designed to run on single-node machines that do not scale for population-wide analyses
- **Rigid tools that are hard to use**
  Traditional bioinformatics tools have a steep learning curve with little flexibility, making them hard to adopt and use
- **Limited support for modern data science workflows**
  Genomic analysis tools rely on file formats without explicit data schemas, which means they're designed for a limited set of genomic analyses. Integrating novel genomic methods and machine learning is not an option with such tools, preventing teams from building powerful predictive models.

Value with Databricks

Many pharmaceutical companies are now able to analyze genomics data and combine it with other data sources such as EHR and imaging data sets. Among the results:

- Improved productivity thanks to a collaborative research environment and standardized APIs
- Accelerated drug target identification due to reduced analysis time for tertiary analysis and GWAS
- R&D teams better able to rapidly search and access analysis results, leading to faster identification of potential drug target and reducing time to market

How to get started

Databricks, in partnership with Regeneron's Genomics Center, has created an open source tool called Glow. By design, Glow seamlessly integrates with Apache Spark™ SQL and the high-performance Delta Lake storage layer. These tools transparently manage, cache and process large volumes of data, making it possible to both query petabytes of genomic data in near real-time and run thousands of data processing tasks with high reliability and scalability. Databricks genomics solutions aim to simplify analysis and integrate with common programming environments such as Python, R, Scala and SQL for common genomic analyses. Databricks also empowers downstream workflows with open source integrations to take advantage of machine learning with native integrations with popular open source technologies for machine learning.

One important type of genomic analysis is genome-wide association studies (GWAS). GWAS identify genetic variations associated with a target disease or trait. Researchers and clinicians can use this information to better detect, treat and prevent chronic health conditions. This Solution Accelerator notebook builds on top of Glow, an open source project for genomic analysis, to provide a fast, scalable and easy-to-implement method for whole genome regression tests.

Proof points

REGENERON

The company's mission is to tap into the power of genomic data to bring new medicines to patients in need. Yet, transforming this data into life-changing discoveries and targeted treatments has never been more challenging. With poor processing performance and scalability limitations, Regeneron's data teams lacked what they needed to analyze petabytes of genomic and clinical data. Databricks now empowers them to analyze entire genomic data sets to accelerate the discovery of new therapeutics.

Discovering new treatments with AI

Biogen

Massive genomics data sets are transforming how pharmaceutical companies like Biogen pinpoint new targets for therapeutic approaches to patient care and enhance the efficacy of existing treatments. But as Biogen's portfolio of research programs grew, their infrastructure and analytics capabilities weren't prepared to manage the immense genomics data sets comprising billions of findings for neurological disorders.

Biogen turned to Databricks to move their on-premises data infrastructure into the AWS cloud, which drastically reduced data processing time and increased bandwidth across collaborating teams. With this enhanced scalability and speed, disease biologists are now able to deepen their understanding of genetic variants, human longevity, and neurological statuses to develop therapies and treatments for patients around the world.

Advancing disease therapies through cloud-based AI

USE CASE: R&D Optimization With Research Knowledge Graph

Overview

Pharmaceutical companies increasingly rely on a massive corpus of scientific publications, internal documents and other assets to find answers to critical questions by extracting pieces of information from these documents. By connecting data sets through knowledge graphs, users can quickly access critical information hidden within an organization's proprietary knowledge assets. For example, target compounds can be annotated with all relevant internal and external research, which in turn accelerates the rate of discovery.

Challenges

- **Data silos**
  Internal data is sparse and every new technology requires a different data processing pipeline
- **Massive volumes of disjointed data**
  Limitations in ingesting, parsing and analyzing millions of data points across hundreds of data sources, including internal data sources and public sources, such as technical literature, public databases, etc.
- **Scale**
  Scaling operations to support data science efforts with open source Python notebooks

Value with Databricks

Many pharmaceutical companies are now able to analyze genomics data and combine it with other data sources such as EHR and imaging data sets. Among the results:

- **Fully managed platform**
  Simplified cluster management and maintenance of analytics resources at scale
- **Scalable, performant data pipelines**
  Ability to leverage natural language processing (NLP) across a huge library of scientific literature and data sources for downstream analysis
- **Accelerated machine learning innovation**
  Data scientists are empowered to build and train models that provide ranking predictions that will help them make smarter decisions

How to get started

In this solution, we demonstrate how to use a pretrained language model to extract relevant clinical entities from a collection of scientific publications and construct a graph of connected clinical entities, such as genes and molecular compounds. We also show how to use GraphFrames to analyze resulting large graphs by applying techniques such as community detection on graphs and graph embedding to build a recommender system providing relevant information on a given entity of interest.

Learn from leading pharmaceutical organizations like AstraZeneca and GlaxoSmithKline that are supporting scientists by integrating all internal and external research information. These graphs power a recommendation system that enables scientists to generate novel target hypotheses using all of these data.

Proof points

AstraZeneca

AstraZeneca discovers, develops and commercializes groundbreaking drugs for some of the world's most serious diseases. The biggest obstacle to new innovations is the inability to tap into all of the scientific information available to them faster than the pace of new data coming in. AstraZeneca needed a platform that allowed them to build scalable, performant data pipelines that feed machine learning models designed to help their scientists make targeted decisions. With Databricks, they are able to leverage data and machine learning to build a recommendation engine that empowers scientists to more easily uncover novel drugs quicker, cheaper and more effectively. Learn more about how they used Databricks to build a knowledge graph here.

USE CASE: Real-World Evidence (RWE) Generation

Overview

Real-world data (RWD) are data generated from routine clinical care as opposed to data gathered in an experimental setting such as a randomized controlled trial (RCT). RWD sources include data derived from electronic health records (EHRs), claims and billing systems, and wearable devices.

Biopharma and device companies are increasingly using RWD to support clinical trial designs, such as large simple trials or pragmatic clinical trials, and to monitor drug performance in a post-market setting.

Challenges

- RWD encompasses large heterogeneous data (EHR, streaming, image, text, etc.), and traditional legacy systems don't scale
- Most existing software and tools designed for supporting clinical trials rely on proprietary formats that do not scale
- Massive volumes of disjointed data — tasked with ingesting, standardizing and analyzing millions of data points across hundreds of data sources

Value with Databricks

Databricks provides the ability to store and analyze all types of structured and unstructured data (e.g., clinical documentation, imaging), to build integrated evidence plans. Organizations can scale with a cloud-native platform designed to process and analyze massive volumes of data at the fastest speeds. The collaborative nature of the platform (multilingual, notebook interface) enables multiple personas, including business analysts, data scientists and statisticians, to work together. Additionally, Databricks supports clinical-grade reproducibility by enabling teams to track data and ML models from source through transformations in one place.

How to get started

In this RWE solution, Databricks shows how to standardize real-world data to the OMOP (Observational Medical Outcomes Partnership) Common Data Model. In addition, Databricks offers cohort building capabilities, like propensity score matching, and shows common SQL queries like drug incidence calculations.

Databricks also has partnered with John Snow Labs to develop solutions around extracting oncology RWD from pathology reports, and detecting adverse events in clinical documentation and conversational text.

Proof points

At Amgen, RWD assets are used to model inclusion/exclusion criteria and site selection for clinical trials. As a result, Amgen has been able to accelerate recruitment and save trial costs. "Adopting the Databricks Lakehouse Platform has enabled a variety of teams and personas to do more with our data. With this unifying and collaborative platform, we've been able to utilize a single environment for all types of users and their preferred tools, keeping operations backed by a consistent set of data."

Webinar with Fierce Pharma: Featuring Biogen, AstraZeneca, Syneos Health: Learn more about how these organizations integrate large volumes of real-world data, ensure reproducibility and scale analytics across teams.

USE CASE: Commercial Effectiveness

Overview

Commercial models in the biopharma industry are changing. The era of "me too" products, large sales forces and lots of face time with prescribers is over. Biopharma manufacturers have more specialized products, which require more customized education and messaging. To make the most of their limited time with prescribers and healthcare institutions, commercial teams must become more predictive and prescriptive in their detailing with digital engagement and next-best-action (NBA) models.

NBA models require massive data engineering pipelines across prescription data, prescriber CRM sources, marketing sources and more. Models predict which prescribers are most likely to prescribe products, what is the best method of engagement (email, phone, in-person visit, etc.), and what message is most likely to resonate.

Applications of advanced analytics that improve pharmaceutical companies' commercial effectiveness also include demonstrating economic value for treatment to payers, targeting underdiagnosed patients, and building clinical-decision support systems with provider recommendations.

Challenges

- **Managing large data volumes**
  Large volumes of prescription, CRM and marketing data require quality controls and complex data pipelines to ensure successful downstream analytics; ETL workloads can be very slow, resulting in stale data
- **Extraction of information from unstructured data**
  Commercial data includes narrative documentation and transcribed calls with prescribers, which contain value insights for NBA models
- **Lack of machine learning capability**
  Existing data architectures do not support performing ML and advanced analytics directly on commercial data sources

Value with Databricks

- **Integrated platform**
  Databricks is fully integrated with all of the components of the lakehouse architecture, which brings the flexibility of data lakes and the reliability of data warehouses together in one unified platform
- **Accelerating ETL pipelines**
  Leveraging Databricks' managed Spark runtime, ETL workflows run orders of magnitude faster; in addition, Databricks' integrated, collaborative platform improves productivity by up to 30%
- **ML is a first-class citizen**
  Unlike with traditional data platforms, where AI and advanced analytics are not directly supported, AI and ML are Databricks first-class citizens: managed ML runtime in combination with Managed MLflow ensures reliable and reproducible management of end-to-end machine learning workflows

How to get started

Using different connectors, data can be ingested into Delta Lake's bronze layer. Pipelines allow for easy implementation of data transformations and robust QA. Using ML runtime and MLflow, data scientists and ML practitioners can interactively explore different approaches and manage and access thousands of pretrained models directly within the platform.

Proof points

Using Databricks, Amgen built a robust approach to Customer 360, providing reps with next-best-action recommendations and enabling the commercial analytics team to improve sales forecasting. Databricks — by providing a single, secure platform — accelerates both ETL and BI workloads. Shared notebooks centralized and accelerated experimentation while providing required traceability. Key ETL workflows went from 5 days to under 8 hours, ensuring data is always fresh. Support for AI/ML at scale allows 80 different data feeds to be integrated into NBA models.

At GSK, there are three global businesses that discover, develop and manufacture innovative pharmaceutical medicines, vaccines and consumer healthcare products. Watch this talk to learn how GSK delivered commercial analytics and went from hackathon to MVP on Azure Databricks.

USE CASE: Interoperability to Automate Healthcare Reporting

Overview

Interoperability policies set by Centers for Medicaid & Medicare Services (CMS) in the United States started to go into effect in 2021, changing how clinical and administrative information is exchanged between payers, providers and patients.

In its policies, CMS has adopted the standards of Health Level Seven (HL7) International, a not-for-profit, ANSI-accredited standards developing organization dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing and retrieval of electronic health information. The most important of these standards in terms of recent policies is the Fast Healthcare Interoperability Resources, commonly known as FHIR (pronounced "fire"). FHIR is a relatively new specification, built on resources spanning domains such as clinical, administrative and financial. A FHIR bundle, or collection of resources, can represent a longitudinal patient record.

As CMS goes, commercial health plans follow. Moreover, FHIR is a global standard that many countries and healthcare organizations have adopted.

Challenges

While FHIR solves the challenge of having a standard format with which to exchange healthcare information, it was not designed for analytics.

The following challenges exist in working with FHIR bundles:

- Converting FHIR (often serialized in JSON format) to tables for longitudinal analytics
- Supporting streaming, real-time data to reflect the dynamic nature of patient health
- Combining FHIR with unstructured data, or other structured data models, outside of those bundles in a common data model
- Connecting data to advanced analytics and machine learning tooling

Value with Databricks

Delta Lake is an open format that ensures data is easily accessible from many analytical systems. Additionally, Delta Lake is designed to support cascading streams, meaning that data can stream through the bronze layer, into the silver layer and finally the gold layer. Delta Lake also provides numerous ways to optimize tables to improve query performance — for example, querying rapidly across both patient ID and date of an encounter. Delta Lake supports Z-ordering to do multidimensional data clustering and provide performance on both of these query patterns.

How to get started

Databricks has two open source projects to support healthcare interoperability standards. Project Smolder is designed to stream and parse HL7v2 messages in Delta Lake. Project dbignite is designed to easily extract resources from FHIR bundles and store the resulting tables in a Databricks SQL database, which can be readily queried using simple SQL statements.

Proof points

With more than 50 hospitals across seven U.S. states, Providence Health wanted to simplify and modernize its technology ecosystem to serve more people and reduce hospital overcrowding. More specifically, Providence built a streaming solution for hospital overcrowding (incorporating HL7v2 messages) on Azure Databricks, automating the previously manual process and improving the National Emergency Department Overcrowding Score for each of its emergency departments.

USE CASE: Disease Risk Prediction

Overview

In the United States, 7 out of 10 deaths and 85% of healthcare spending are driven by chronic conditions. Noncommunicable diseases are generally preventable through patient education and by addressing underlying issues that drive the chronic condition.

Precision prevention is focused on using data to identify patient populations at risk of developing a disease, and then providing interventions that reduce disease risk. An intervention might include a digital app that remotely monitors at-risk patients and provides things like lifestyle and treatment recommendations, monitoring of disease status and supplemental preventative care. However, deploying these interventions first depends on identifying the patients most at risk.

Challenges

- **Data volumes**
  Healthcare data is large and heterogeneous (e.g., clinical documentation, claim data, image, etc.), and traditional legacy systems don't scale
- **Proprietary legacy systems**
  Existing software and tools for managing healthcare data rely on proprietary formats that do not scale
- **DevOps overhead**
  Data scientists' time is mostly spent managing resources and infrastructure rather than building models
- **Ensuring reproducibility**
  When it comes to novel insights based on patient data, it is very important to be able to reproduce results: Current solutions are mostly ad hoc and do not allow for efficient ways to track experiments and the versions of the data used during machine learning model training.

Value with Databricks

- **Reduced DevOps overhead**
  Databricks managed runtimes provide an easy-to-use, robust environment for managing compute resources
- **Improved data team productivity**
  Databricks data science workspace provides collaborative notebooks that allow different teams to rapidly develop models
- **Reproducibility**
  Delta Lake and Managed MLflow allow for experiment tracking and logging, and managed runtimes ensure the highest level of reproducibility of experiments, without the need to write additional code

How to get started

Databricks' solution relies on Delta Lake technology for data transformation and storage. Building on top of the lakehouse for healthcare, we track a patient's journey and train a machine learning model that assesses the risk of a patient for a given condition. Using this model, given a patient's encounter history and demographics information, we can assess the risk of a patient for a given condition, such as drug overuse, within a given window of time. By tracking our models using MLflow, we make it easy to track how models have changed over time, adding confidence about the process of deploying a model into patient care. Learn more about How to Detect At-Risk Patients With Real-World Data.

Proof points

MUSC
Medical University
of South Carolina

The Medical University of South Carolina (MUSC) is the state's only integrated, academic health sciences center, with a unique charge to serve the state through education, research and patient care. With more than 1,600 beds and nearly 275 telehealth locations, MUSC is focused on delivering the highest-quality care. The core of their strategy is leveraging data analytics and AI to innovate the patient experience. MUSC built a modern unified data analytics architecture that allows their teams to unlock insights buried within their data and build powerful predictive models. More specifically, this strategy enabled MUSC to build robust sepsis risk prediction models, and prepared MUSC to quickly respond to the dynamic environment of COVID-19: Building a Modern Unified Data Analytics Architecture for Real-Time COVID Response.

Humana strives to help the communities it serves and individual members achieve their best health. Humana had the opportunity to rethink its existing operations and reimagine what a collaborative ML platform for hundreds of data scientists might look like. The primary goal of its ML platform, named FlorenceAI, is to automate and accelerate the delivery lifecycle of data science solutions at scale. Learn more about how Humana deployed FlorenceAI and predictive models at scale in this presentation.

USE CASE: Patient Personalization

Overview

A central tenet of modern medicine is patient-centric care — which engages a patient in their treatment plan and includes the principle that healthcare should be based on the unique person's needs and their right to health. A growing body of evidence shows that patients who have been actively involved in their treatment planning have better outcomes and incur lower costs. As a result, many public and private healthcare organizations are employing strategies to better engage patients, such as educating them about their conditions and working with patients to plan medical treatments that accommodate their unique needs.

Challenges

- Gathering data securely to enable patient communication
- Providing end-to-end reliability and correctness guarantees
- Performing complex transformations on a diverse range of data sets
- Unifying data and ML for enhanced patient communications

Value with Databricks

- Databricks utilizes Delta Lake to provide a robust data solution for huge data merges, reducing complexity when integrating with the broader electronic medical record (EMR) system for data reporting, patient contact and data flow
- Databricks provides a separate ML workspace that is designed for data scientists to retrain models and rapidly evaluate new ideas codesigned with the medical teams
- Shared workspaces and collaborative notebooks facilitate data team productivity and accelerating time-to-insight
- Role-based access controls and network security controls ensure the safety and confidentiality of patient data

How to get started

Building on top of a lakehouse architecture, with Delta Lake at the center, healthcare data such as HL7 message streams, EMR and claim data can be ingested and securely tokenized. Using bronze, silver and gold architecture framework, data is ingested and stored in the bronze layer, while tables based on common healthcare data models are stored in silver tables. Using features extracted from silver tables and taking advantage of machine learning tooling, personalized recommendations can be generated and monitored for performance throughout the model lifecycle.

Proof points

With over 80 million customers passing through their pharmacies every day, CVS Health is always striving to provide more meaningful interactions that put customers on a path to better health. In 2018, they embarked on a journey to personalize experiences through machine learning in their Hadoop environment, but the complexity and scale of the diverse data sets prohibited understanding the behaviors of a large number of micro-segments of customers. With Databricks, CVS Health was able to analyze their customer data to implement different experiences, experiments and a variety of segments for personalization at scale. You can learn more about CVS's journey here.

USE CASE: Automating De-Identification

Overview

Under the Health Insurance Portability and Accountability Act (HIPAA) minimum necessary standard, HIPAA-covered entities (such as health systems and insurers) are required to make reasonable efforts to ensure that access to protected health information (PHI) is limited to the minimum necessary information to achieve the intended purpose of a particular use, disclosure or request.

In Europe, the GDPR lays out requirements for anonymization and pseudo-anonymization that companies must meet before they can analyze or share medical data. In some cases, these requirements go beyond U.S. regulations by also requiring that companies redact gender identity, ethnicity, and religious and union affiliations. Almost every country has similar legal protections on sensitive personal and medical information.

Challenges

Minimum necessary standards create obstacles to advancing population-level healthcare research. This is because much of the value in healthcare data is in the semi-structured narrative text and unstructured images, which often contain personally identifiable health information that is challenging to remove. Such PHI makes it difficult to enable clinicians, researchers and data scientists within an organization to annotate, train and develop models that have the power to predict disease progression, as an example. De-identification historically has required heavy manual intervention and review.

Value with Databricks

John Snow Labs, the leader in healthcare natural language processing (NLP), and Databricks have teamed up to help organizations process and analyze their text data at scale.

At the core of the Databricks Lakehouse Platform is Delta Lake, an open source storage layer that brings performance (via Apache Spark), reliability and governance to a data lake. Healthcare organizations can land all of their data — including raw provider notes, radiology reports and PDF pathology reports — into Delta Lake. This preserves the original source of truth before applying any data transformations. By contrast, with a traditional data warehouse, transformations occur prior to loading the data, which means that all structured variables extracted from unstructured text are disconnected from the native text.

Building on this foundation is John Snow Labs' Spark NLP for Healthcare, the most widely used NLP library in the healthcare and life science industries. Optimized to run on Databricks, Spark NLP for Healthcare seamlessly extracts, classifies and structures clinical and biomedical text data with state-of-the-art accuracy at scale. It is the only native distributed open source text processing library for Python, Java and Scala, and since every Spark NLP pipeline is a Spark ML pipeline, it is particularly well suited to building unified NLP and machine learning pipelines. Spark NLP provides Python, Java and Scala libraries with the full functionality of traditional NLP libraries (like spaCy, NLTK, Stanford CoreNLP and Apache OpenNLP) and adds additional functionality, such as spell-checking, sentiment analysis and document classification.

How to get started

Databricks developed a joint Solution Accelerator with John Snow Labs to automate the detection of sensitive information contained within unstructured data using NLP models for healthcare. Extracted data is stored within the lakehouse, where teams can use the pretrained models to easily remove, obfuscate or mask data for downstream analytics at massive scale.

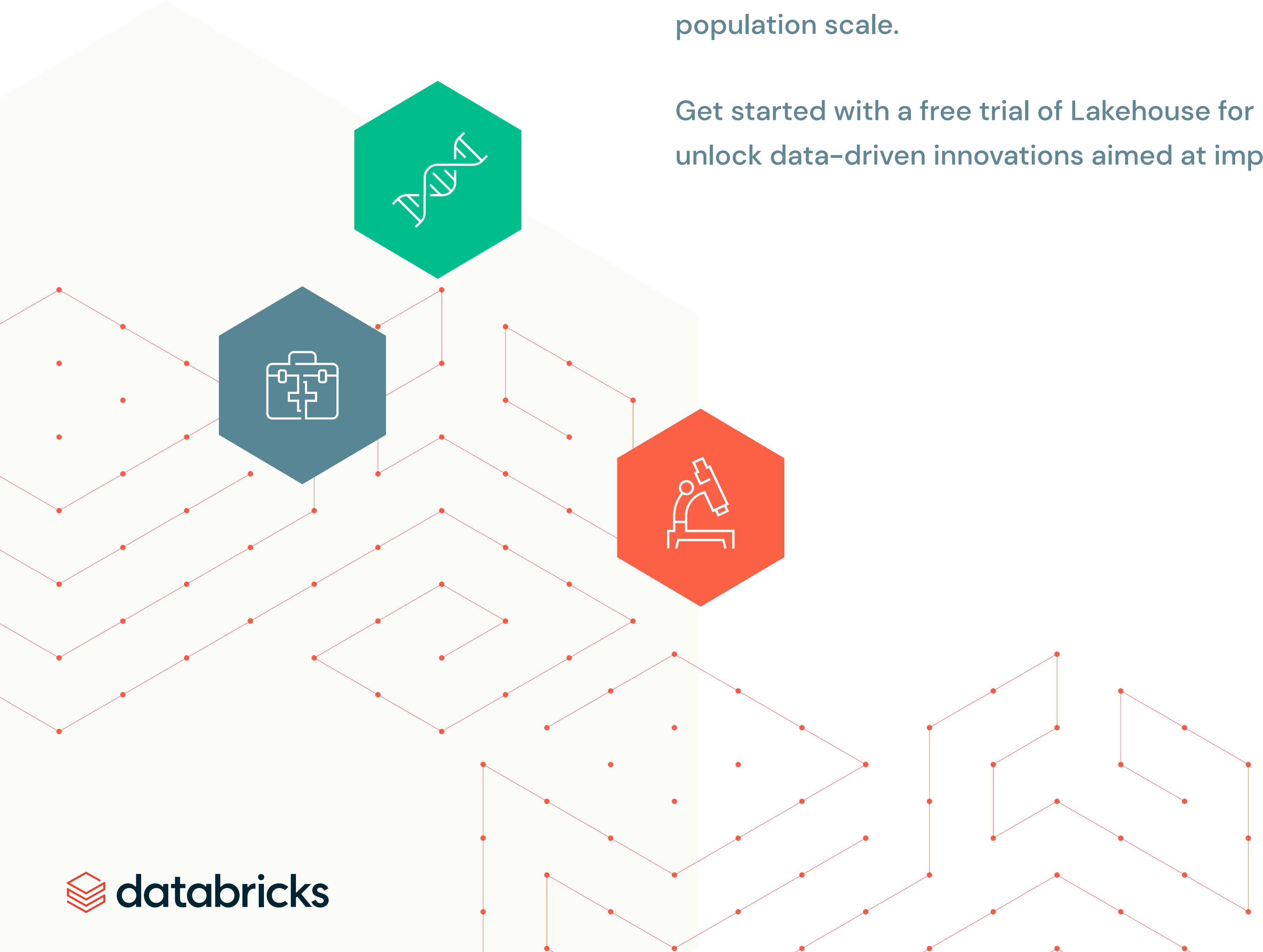
- Convert unstructured data like PDFs to structured text with OCR models
- Easily detect PHI using pretrained NLP models for healthcare
- Automatically remove or de-identify PHI for downstream analysis

Proof points

Providence

Providence Health embarked on an ambitious journey to de-identify all of its clinical electronic medical record (EMR) data to support medical research and the development of novel treatments. This required building a de-identification pipeline using pretrained deep learning models, fine-tuned to its own data. Ultimately, through experimentation and iteration, Providence achieved a level of performance that safeguards patient privacy while minimizing information loss. Learn more about Providence's de-identification here.

Conclusion



Today, data is at the core of every innovation in the healthcare and life sciences industry. The Databricks Lakehouse Platform for Healthcare and Life Sciences enables organizations to bring together all their patient, research and operational data with powerful analytics and AI capabilities to deliver real-time insights at population scale.

Get started with a free trial of Lakehouse for Healthcare and Life Sciences today to unlock data-driven innovations aimed at improving health outcomes.

About Databricks

Databricks is the lakehouse company. More than 7,000 organizations worldwide — including Comcast, Condé Nast, H&M and over 50% 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.

START YOUR FREE TRIAL

Contact us for a personalized demo at
databricks.com/contact

databricks

© Databricks 2022. All rights reserved. Apache, Apache Spark, Spark and the Spark logo are trademarks of the Apache Software Foundation. Privacy Policy | Terms of Use