Content Document: Fundamentals of the Databricks Lakehouse Platform

Course welcome

Welcome to Fundamentals of the Databricks Lakehouse Platform. This course is designed for everyone who is brand new to the Platform and wants to learn more about what it is, why it was developed, what it does, and the components that make it up.

Our goal is that by the time you finish this course, you'll have a better understanding of the Platform in general and be able to answer questions like:

1) What is Databricks?
2) Where does Databricks fit into my workflow?
3) How have other customers been successful with Databricks?

Specifically, by the end of the course, you will be able to:

- Describe what the Databricks Lakehouse Platform is.
- Explain the origins of the Lakehouse data management paradigm.
- Outline fundamental problems that cause most enterprises to struggle with managing and making use of their data.
- Identify the most popular components of the Databricks Lakehouse Platform used by data practitioners, depending on their unique role.
- Give examples of organizations that have used the Databricks Lakehouse Platform to streamline big data processing and analytics.

Note: This course is not technical in nature. It does not include demonstrations of how to use Databricks functionality, nor does it present deeply technical content on Databricks functionality. However, this course does explain Databricks functionality available for various data practitioners, depending on their unique roles.

If you're ready, click Continue to get started.

Section 1: An Introduction to Databricks

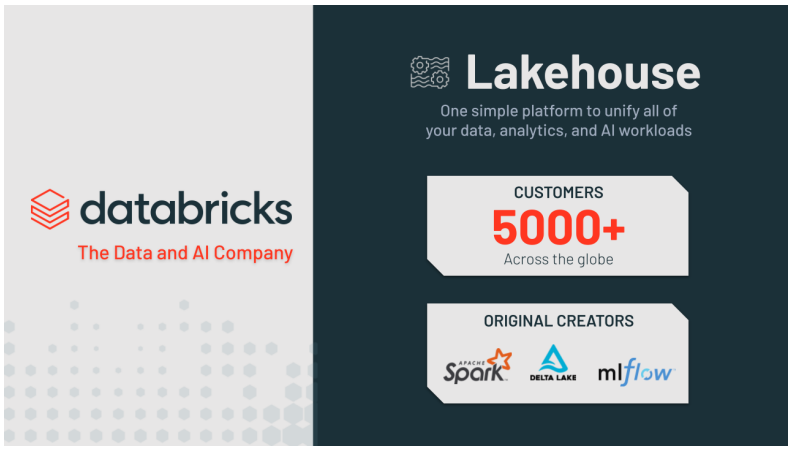
By the end of this lesson, you should be able to:

- Describe what the Databricks Lakehouse Platform is and how it fits into the big data landscape.

What is Databricks?

Databricks is known as the Data and AI company. It is a Software-As-A-Service company that makes big data and AI easier for organizations to manage, enabling data-driven innovation in all enterprises.

To put it simply, the Databricks Lakehouse Platform empowers everyone on a data science team to work together, in one secure platform, from the minute data is ingested into an organization through when it's cleaned up, analyzed, and used to inform business decisions.



Alt-text: An image showing the Databricks logo, the number of customers that Databricks has (5000), and the open-source projects created by Databricks (Apache Spark, Delta Lake, and MLflow).

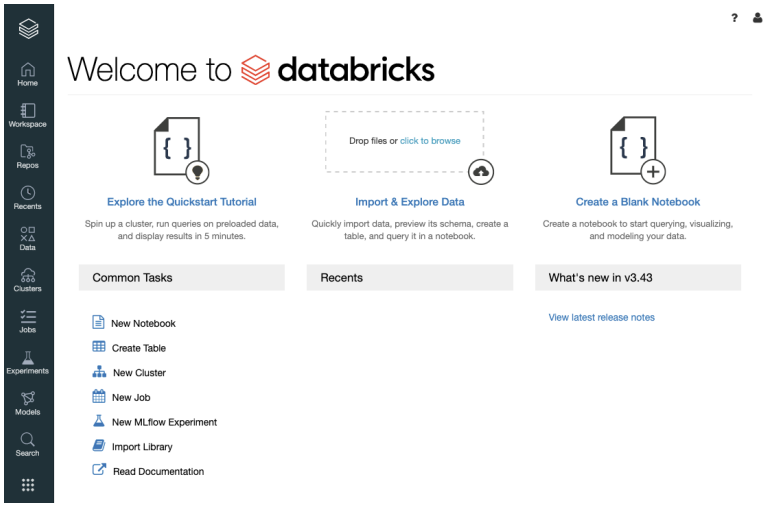
What does Databricks help organizations do?

The Databricks Lakehouse Platform enables organizations to:

- Ingest, process, and transform massive quantities and types of data
- Explore data through data science techniques, including but not limited to machine learning
- Guarantee that data available for business queries is reliable and up to date
- Provide data engineers, data scientists, and data analysts the unique tools they need to do their work
- Overcome traditional challenges associated with data science and machine learning workflows (we will explore this in detail in our next lesson)

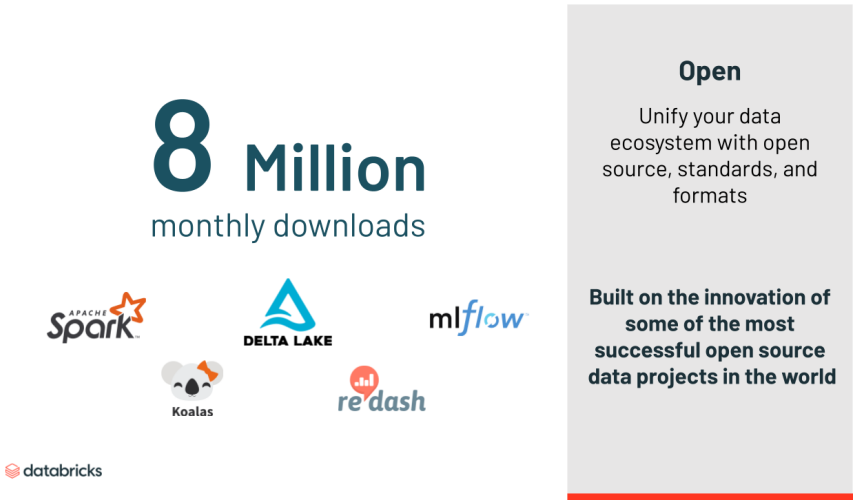
Primarily, data practitioners access Databricks functionality using a custom-built web-based interface. This is an environment for accessing all of your Databricks assets (like notebooks, libraries, experiments and dashboards), as well as the computational resources you need to process data . So, while you might be using different tools/functionality within the Databricks workspace depending on your role, everyone is working in one platform. We'll talk more about the

characteristics of the lakehouse architecture later on in this course. For now, we can say that a single, unified system is one of the core concepts that defines a lakehouse, and this is part of what makes Databricks the Lakehouse Platform.



Caption/Alt-text: An image of what data practitioners see when they log into Databricks. This portal is what they use to access Databricks functionality and do their work.

Another core Lakehouse concept that we'll dig into is that the data in a lakehouse should be in an open format, so that your data belongs to you - always. Databricks has shown a demonstrated commitment to open source technologies. Databricks invented some of the most successful open-source projects in the world, as shown in the image below, and it continues to contribute to these projects regularly. When your organization uses Databricks, you get managed access to open source data projects through the Platform (which makes it straightforward for you to use them). Plus, you are unifying your big data system with open source standards. By aligning to open source standards, you establish building blocks to help make your applications more inoperable with different systems.



Alt-text: An image showing the open source projects created by Databricks.

What types of organizations typically use Databricks?

Currently, Databricks works with over 5000 customers across the globe and with enterprises in every industry - healthcare & life sciences, media & entertainment, financial services, retail, and more.

Caption: A small sample of Databricks customers in a variety of industries.
Alt-text: An image with logos for Databricks customers in various industries, including healthcare & life sciences, manufacturing & automotive, media & entertainment, financial services, and more.

At the end of this course, we'll give you the chance to explore customer use-cases in greater depth. Hopefully, it will help you start thinking about what you can do with the Platform.

What else should I know about Databricks as a Databricks customer?

In addition to access to the Databricks Lakehouse Platform, Databricks customers get:

- Over 450 partners to help you unify your data sources and analytics services and centralize governance.

Caption: The Databricks Partner Network
Alt-text: An image showing select logos from the Databricks Partner network.

- Support from expert services/resources to help enable your users and IT teams to adopt the platform.
- Access to free self-paced training, including a structured set of workstreams and activities that will help you quickly onboard thousands of users while providing the central capabilities to own and run the platform

For questions about any of this - please contact your Databricks account manager or visit help.databricks.com.

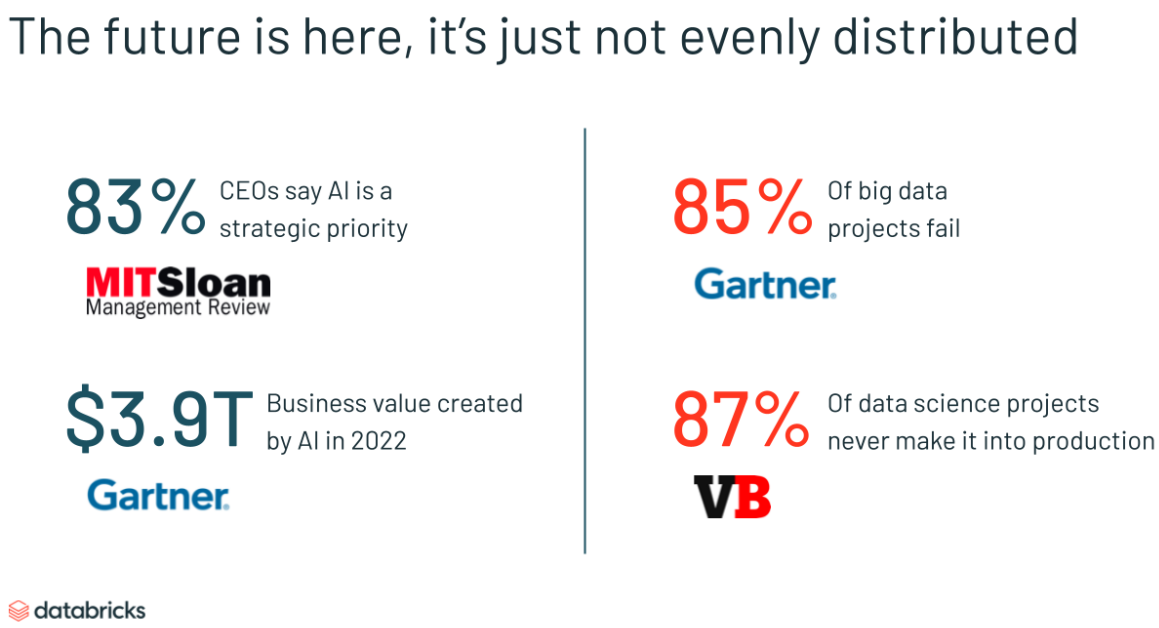
Alright - it's time to move on. At this point, you should have a clear understanding of what Databricks is at a high level. To summarize, Databricks offers a Lakehouse Platform, accessible via an online login, that allows data science teams to collaborate on their work. This Lakehouse Platform gives customers access to Databricks-native tools, managed open-source tools, and technical resources to help customers along the way.

Now, you might be thinking to yourself, "I get what Databricks does, but - why? Why did we move to Databricks? Why are we using Databricks? What business problems are we trying to solve?". To answer this question, go to the next lesson, Big Data, and AI Challenges that Databricks Solves".

Section 2: Big Data and AI Challenges that Databricks Solves

By the end of this lesson, you should be able to:

- Explain the origins of the Lakehouse data management paradigm.
- Outline fundamental problems that cause most enterprises to struggle with managing and making use of their data.



Alt-text: An image showing statistics that highlight the importance of AI as a strategic priority to CEOs (83% of CEOs say AI is a strategic priority), the business value created by AI in 2022 ($3.9 trillion), how many big data projects fail (85%), and the percent of data science projects that never make it into production (87%).

In the early days of data analytics, simple relational databases, historical data, and spreadsheet expertise were used to drive business decisions. Today, with the emergence of big data, these methods are no longer sufficient. Businesses today spend a significant amount of resources trying to piece together solutions for extracting insights from big data. Most organizations (close to 90%) fail on big data and analytics projects despite spending these resources.

Why is extracting insights from big data so complicated? In this lesson, we'll explore just that.

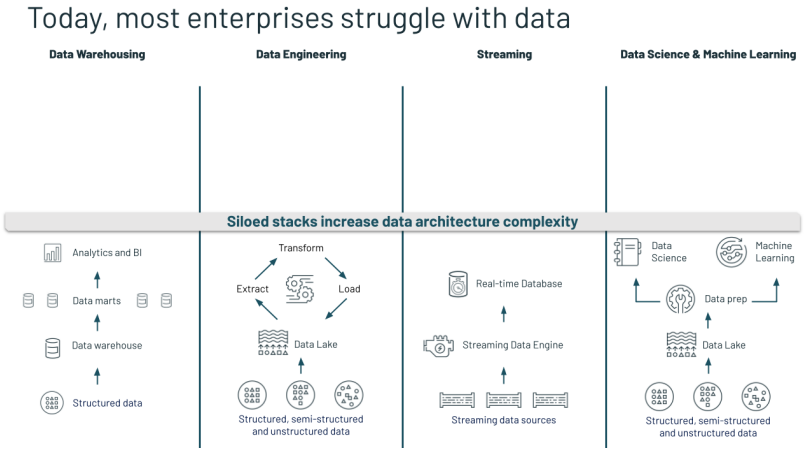
Challenge number one - working with big data is not easy

As explained in previous Databricks Academy courses, big data is hard to manage. This is because it comes into an organization in massive volumes, faster than ever before, and in a wide variety of formats.

As data practitioners work to design their organization's big data infrastructure, they often ask and need to answer questions like:

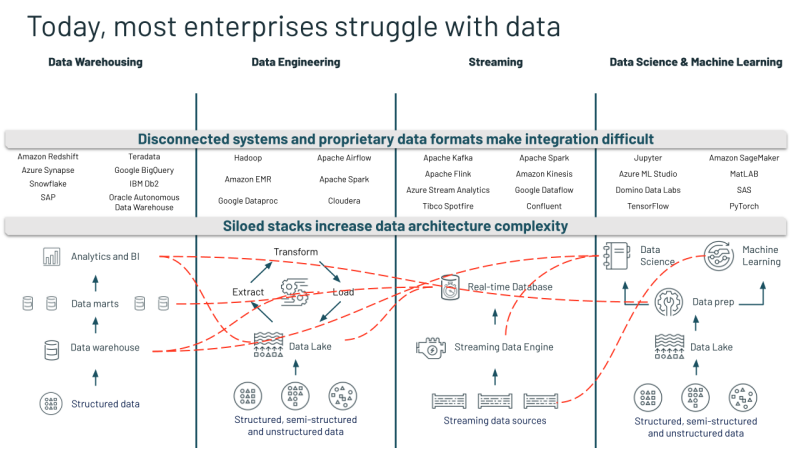
- Where/how will we store our big data?
- How can we process batch and stream data?
- How can we use different types of data together in our analyses (unstructured vs. structured data)?
- How can we keep track of all of the work we're doing on our big data?

As you can imagine, there are many ways that an organization can set up big data infrastructure — getting it right is no easy task. What frequently ends up happening is that organizations set up different technology stacks to handle their data workloads.



Alt-text: An image showing technology stacks divided by data warehousing, data engineering, streaming, and data science & machine learning workloads.

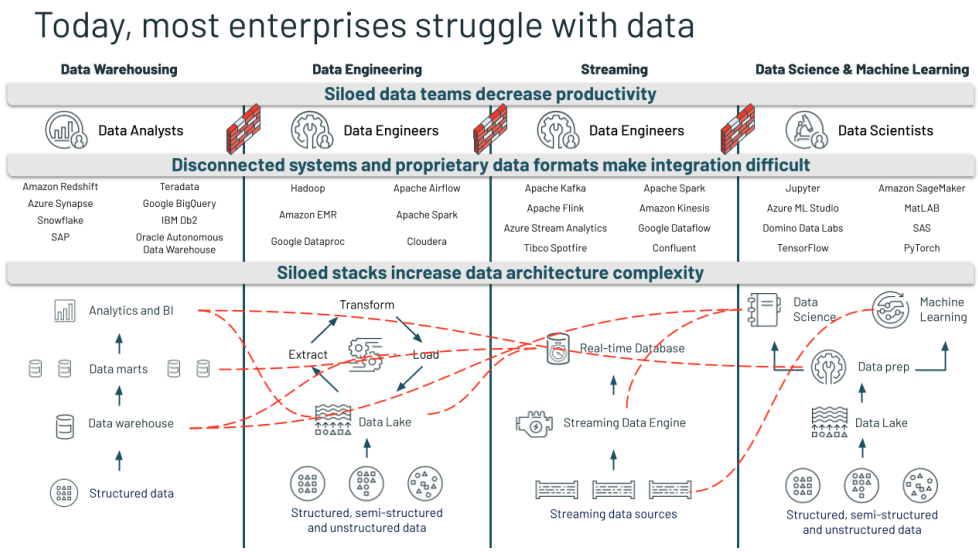
Often, these technology stacks don't work well together due to many different tools introduced that are not compatible with each other, proprietary data formats that are not easy to translate from one tool to another. If you want to enable advanced use-cases, you have to move across these stacks. Think about the work your organization does. You don't typically have a "data warehouse use case" or a "streaming" use case. You most likely have a supply chain logistics use case or a financial risk assessment use case. To address these challenges, you need to move across these technology stacks.



Alt-text: An image showing technology stacks divided by data warehousing, data engineering, streaming, and data science & machine learning workloads, with the addition of dotted lines to represent how organizations must cross into multiple data stacks to get their jobs done.

Siloed roles lead to organizational inefficiencies.

With multiple stacks set up, many organizations suffer from the challenges of having siloed functional roles for individuals on their data science teams. For example, it is not uncommon for a data scientist to build and train a machine learning model in a vacuum on their computer, with little to no visibility to related work being done by, for example, the data engineer preparing that data for them, or the data analysts who might be using results from their experiments to produce dashboards. This leads to communication slowdowns and teams working with different versions of the same data. This affects data security and governance and ultimately results in less productive data teams.



Alt-text: An image showing technology stacks divided by data warehousing, data engineering, streaming, and data science & machine learning workloads, with the addition of dotted lines to represent how organizations must cross into multiple data stacks to get their jobs done, plus dividers between data practitioners roles - this is to showcase that many times, team members end up working in silos because of divided technology stacks.

Data security challenges

Speaking of data security - according to Gartner, 80% of organizations will fail to develop a consolidated data security policy. This leaves them and their data vulnerable to security breaches.

Think about the ramifications of a security breach. Beyond just the immediate monetary cost, there is a long-lasting loss in customer trust and company reputation. If you've ever been a customer of a company that has suffered a security breach, you know first-hand how long it can take to rebuild trust.

Besides protecting data from leaking out, organizations must ensure they're compliant with data protection regulations like GDPR (European Union's General Data Protection Regulation) and HIPAA (Health Insurance Portability and Accountability Act). Plus, they must have the required certifications to run their businesses. There are often hefty penalties if they are not compliant.

Does any of the above resonate with you? If so, this is probably why your company has decided to use Databricks. Through Lakehouse Platform, Databricks helps companies overcome all of these challenges and more.

Next, and for the rest of this course, we will explore Databricks more in-depth. We'll start by describing what Lakehouses are since they sit at the heart of the Databricks Lakehouse Platform.

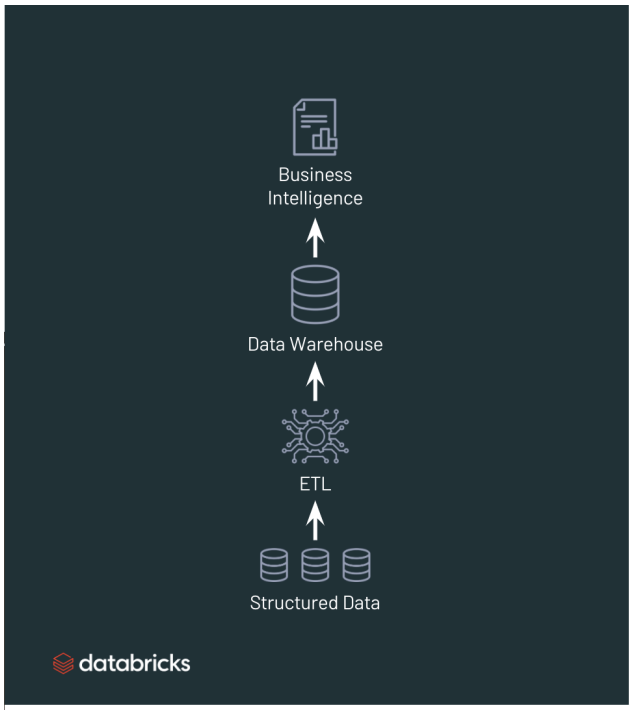
Section 3: The Lakehouse Paradigm

Whether your organization has moved to the cloud for the first time or is reevaluating its current approach, making decisions about the technology used when storing data can have enormous implications for costs and performance in downstream analytics.

As a platform focused on computation and analytics, Databricks seeks to help our customers make choices that unlock new opportunities, reduce redundancies, and connect data teams. That is why we created the Lakehouse - a storage technology that combines the most popular functionality from data warehouses and data lakes.

In this lesson, we'll start to explore the pros and cons of data warehouses and data lakes and discuss how these led to the creation of Lakehouse.

Data Warehouses



Data Warehouses

Pros

- Great for Business Intelligence (BI) applications

Cons

- Limited support for Machine Learning (ML) workloads
- Proprietary systems with only a SQL interface
- Stale data (and no streaming)

Alt-text: A diagram of a data warehouse, with a summary of data warehouse pros and cons reviewed in this course.

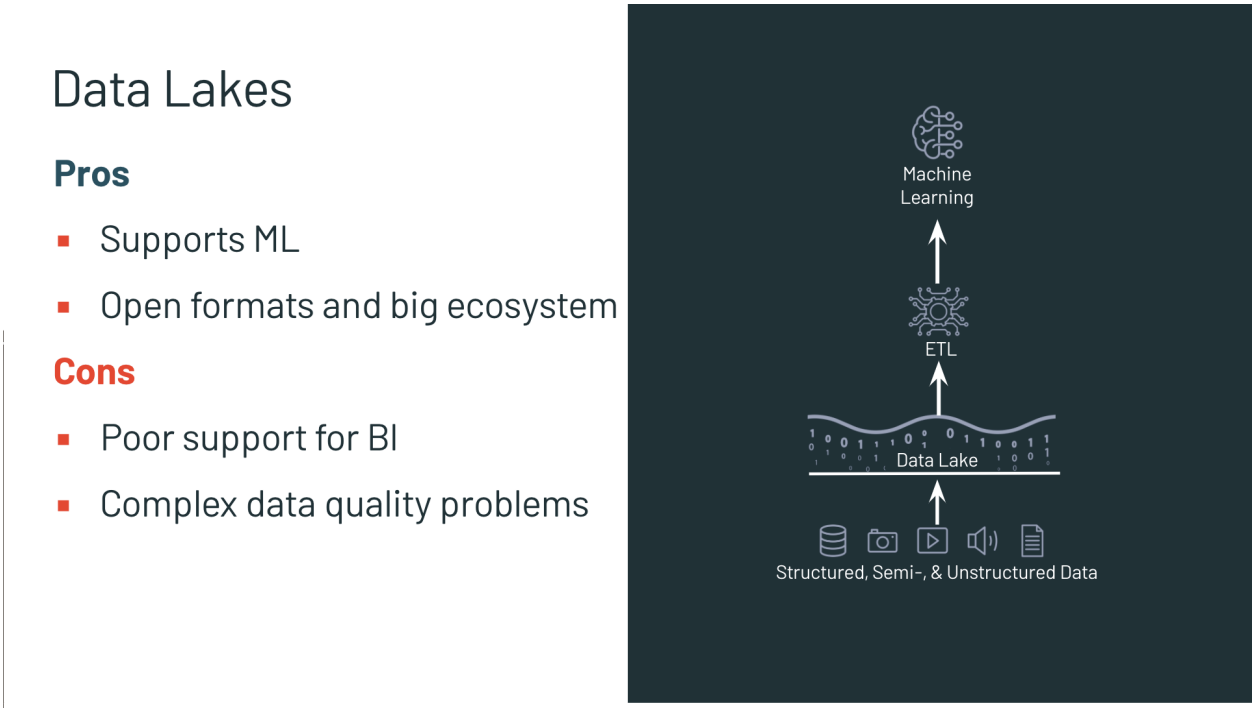
Data warehouses have been powering business intelligence (BI) decisions for around 30 years. Throughout history, data warehouse technologies have evolved in response to changes in the scale and nature of data relevant to decision-makers and the rise of the internet and other technological innovations.

Data warehouses generally follow a set of guidelines to design systems controlling the flow of data used in decision making. They are designed to optimize data queries, prevent conflicts between concurrently running queries, support structured data, and make the assumption that data entered is unlikely to change with high frequency.

Because data warehouses originally were developed using on-premises technology to drive BI decisions, they have had some pain points in adapting to the modern data atmosphere. The proliferation of smart devices and web-based commerce has increased both the volume and the variety of data. Data warehouses traditionally rely on proprietary data formats, which do not support video, image, or free text files. This leads to limited support for machine learning workloads.

Traditional data warehouse technologies force users to invest in enough compute power and storage to handle their peak user traffic. With the exponential growth many companies have seen in their data in recent years, this design can become extremely expensive.

Data Lakes



Alt-text: A diagram of a data lake, with a summary of data lake pros and cons reviewed in this course.

Data lakes have only come into widespread use over the last decade, driven by huge increases in the variety and volume of data and powered by technological advances to store and process data cheaply and efficiently. Data lakes are often defined in opposition to data warehouses: while a data warehouse delivers clean, structured data for BI analytics, a data lake allows an organization to permanently and cheaply store data of any nature in any format - in fact, data lakes allow both structured and semi-structured data to be stored alongside unstructured data like video, images, free text, and log files.

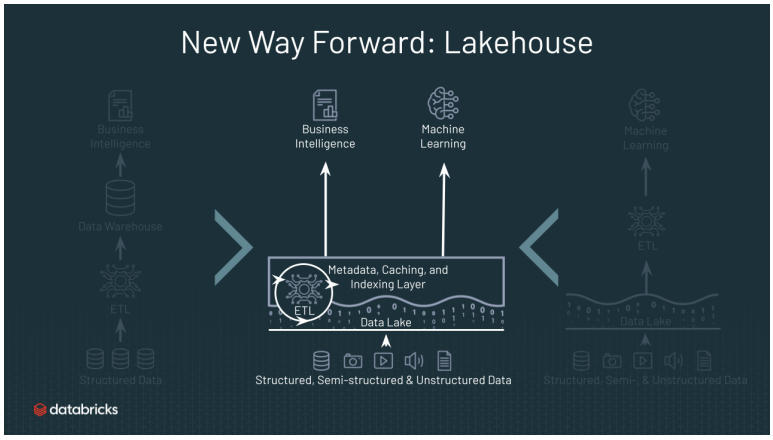
Unlike in data warehouses, data storage and compute and not coupled in data lake design. This has led to several technologies that have evolved around working with data stored in data lakes. Apache Spark, for example, the compute engine used by Databricks, can run analytic queries, extract-transform-load, and machine learning workloads over enormous quantities of data stored in data lakes.

For all their strengths, data lakes have not eliminated the need for data warehouses. For BI analytics, most companies continue to load the relevant data from their data lake into a data warehouse, and it is here that most analysts execute queries and generate reports. Also, because all of an organization's data lands in a data lake, they can quickly become

what is known as a data swamp - a poorly maintained data lake that is difficult to navigate and query.

Lakehouse

A data lakehouse is a new, open data management paradigm that combines the most popular capabilities of data lakes and data warehouses. Per their design, they implement similar data structures and data management features to those in a data warehouse, directly on the kind of low-cost storage used for data lakes. Merging these ideas into a single system means that data teams can move faster as they can use data without accessing multiple systems. Data lakehouses also ensure that teams have the most complete and up-to-date data available for data science, machine learning, and business analytics projects.



Alt-text: A diagram of a Lakehouse.

Data lakehouses have the following key features:

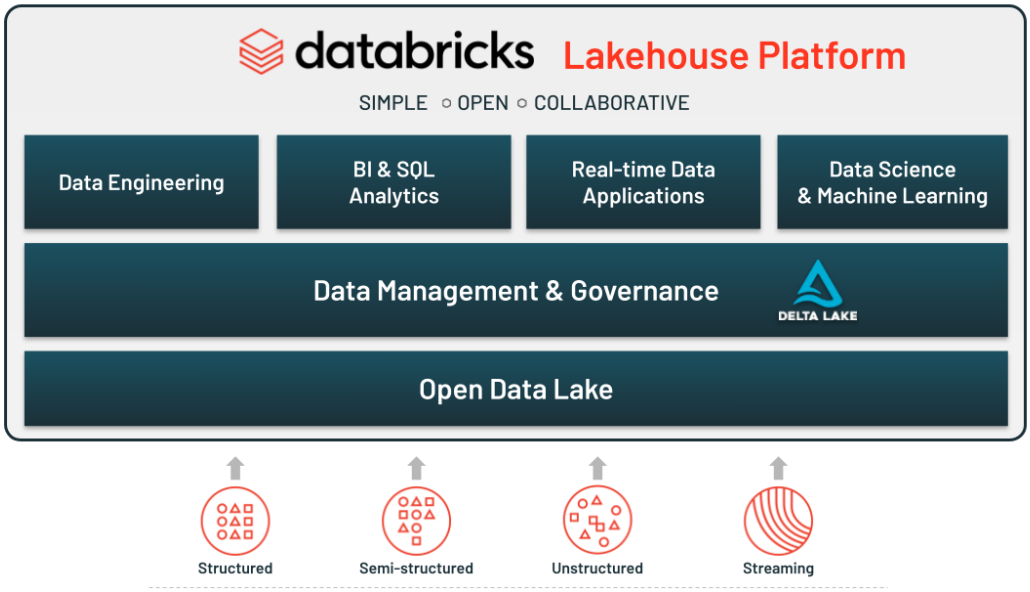
- Transaction support to ensure that multiple parties can concurrently read or write data
- Data schema enforcement to ensure data integrity (writes to a table are rejected if they do not match the table's schema)
- Governance and auditioning mechanisms to make sure you can see how data is being used
- BI support so that BI tools can work directly on source data - this reduces data staleness.
- Storage is decoupled from compute, which means that it is easier for your system to scale to more concurrent users and data sizes.
- Openness - Storage formats used are open and standard. Plus, APIs and various other tools make it easy for team members to access data directly.
- Support for all data types - structured, unstructured, semi-structured

- End-to-end streaming so that real-time reporting and real-time data can be integrated into data analytics processes just as existing data is
- Support for diverse workloads, including data engineering, data science, machine learning, and SQL analytics - all on the same data repository.

Next, we'll explore the last bullet point in more detail and explore what the Platform offers data practitioners unique to their role.

Section 4: Databricks Lakehouse Platform Components

As we mentioned in the previous lesson, the Lakehouse is at the heart of the Databricks Lakehouse Platform. Now that you have a decent understanding of what a Lakehouse is, we can discuss the Platform's capabilities.



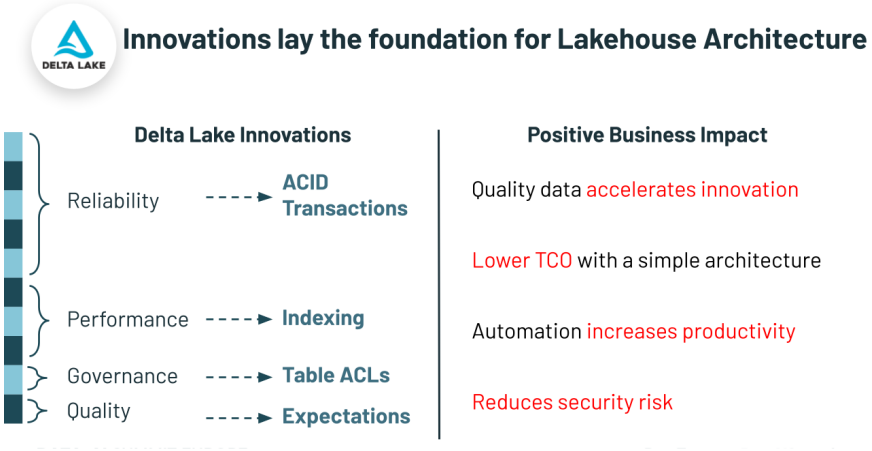
Caption: A conceptual image of the Databricks Lakehouse Platform.
Alt-text: A conceptual image of the Databricks Lakehouse Platform that shows different types of data coming into the Platform and landing in an open data lake. On top of that data lake is the Delta Lake logo. On top of the Delta Lake logo are Platform use-cases including Data Engineering, BI & SQL Analytics, Real-time Data Applications, and Data Science & Machine Learning.

Databricks Lakehouse Platform for Data Engineering

When data arrives into the Lakehouse Platform, it lands into your organization's open data lake. What makes that data lake a Lakehouse is the data management and governance functionality placed on top of that data lake through a product called Delta Lake.

Delta Lake is an open-source storage layer that brings data reliability to data lakes. When we talk about data reliability, we refer to the accuracy and completeness of your data. In other words, Delta Lake working in conjunction with a data lake is what lays the foundation for your Lakehouse - that combination guarantees that your data is what you need for your use-cases via:

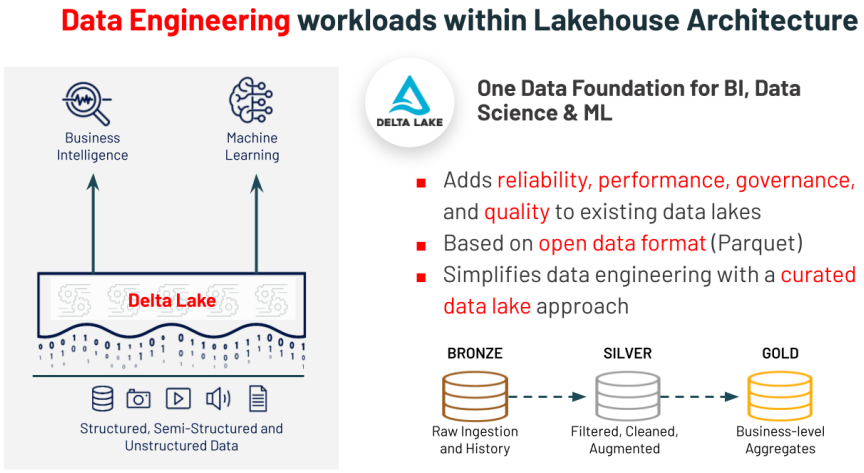
- ACID transactions, which are database transaction properties that guarantee data validity. With ACID transactions, you don't have to worry about missing data or inconsistencies in your data from interrupted or deleted operational transactions because changes to your data are performed as if they are a single operation.
- Indexing, which is a way to get an unordered table (which might be inefficient to query) into an order that will maximize the efficiency of your queries
- Table access control lists (ACLs), or governance mechanisms that ensure that only users who should have access to data can access it
- Expectation-setting, which refers to the ability for you to configure Delta Lake based on your workload patterns and business needs



Alt-text: An image summarizing the concepts presented about Delta Lake in this lesson.

With Delta Lake, data engineers can architect pipelines for continuous data flow and refinement using Delta's architectural model, shown in the image below. It allows data engineers to build pipelines that begin with raw data as a "single source of truth" from which everything flows. Subsequent transformations and aggregations can be recalculated and validated to ensure that business-level aggregate tables are still

reflective of the underlying data, even as downstream users refine the data and introduce context-specific structure.



Alt-text: An image showing the Delta Lake architecture (Raw data into Bronze tables, clean data into Silver tables, and business-level aggregate data in Gold tables). The image also summarizes the concepts presented in this lesson.

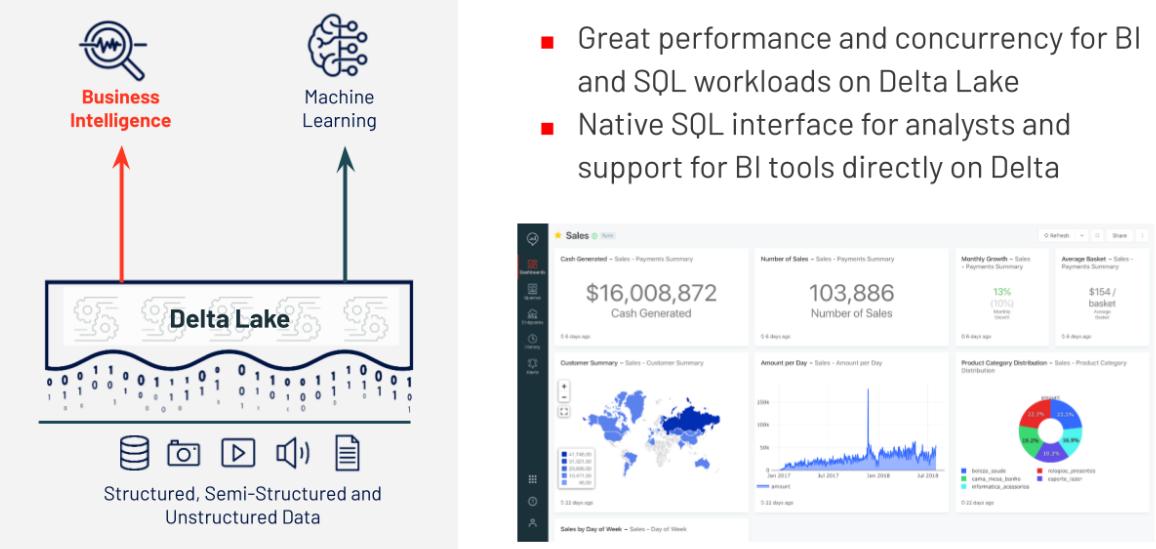
Databricks Lakehouse Platform for BI & SQL Analytics

Databricks SQL Analytics provides SQL users a SQL-native Interface to write queries in a familiar syntax and explore your organization's Delta Lake table. Regularly used SQL code can be saved as snippets for quick reuse, and query results can be cached to keep the query short.

Once queries are built, SQL Analytics also allows analysts to make sense of the results through a wide variety of visualizations. The visualizations can be organized into dashboards through a drag-and-drop interface. Dashboards can be easily shared with other users to provide insights. They are also configurable to automatically refresh and alert your team to meaningful changes in your data.

Also, SQL Analytics was designed to work with Delta Lake, so you're getting all the performance optimizations, governance, and reliability that come with that.

Plus, it supports existing BI tools, which means that you can use your preferred visualization tools to query your organization's Lakehouse.

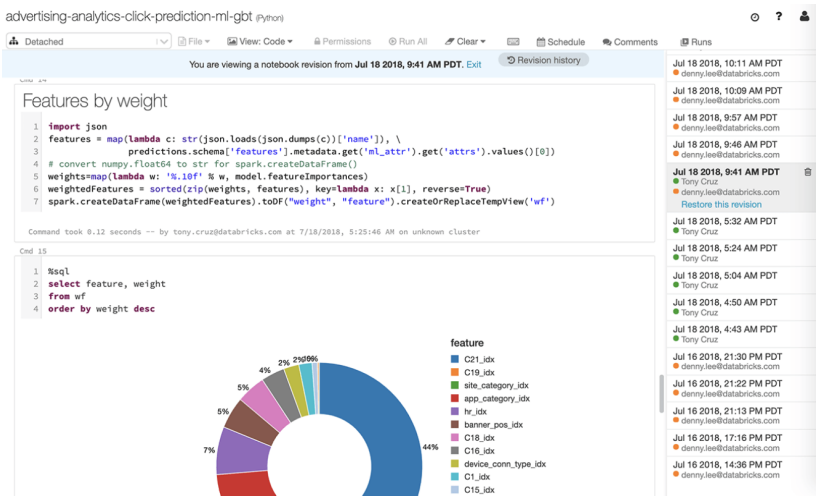


Alt-text: An image summarizing the benefits of SQL Analytics workloads on Delta Lake. It also shows an example of the SQL Analytics user interface open-source.

Databricks Lakehouse Platform for Data Science & Machine Learning

Data science and ML teams use Databricks throughout the entire data science and machine learning life cycles - to explore, prepare and process data, build and test machine learning models, put those models into deployment, and then optimize them. This is done through platform components, including collaborative notebooks, the Databricks Machine Learning Runtime, and Managed MLflow.

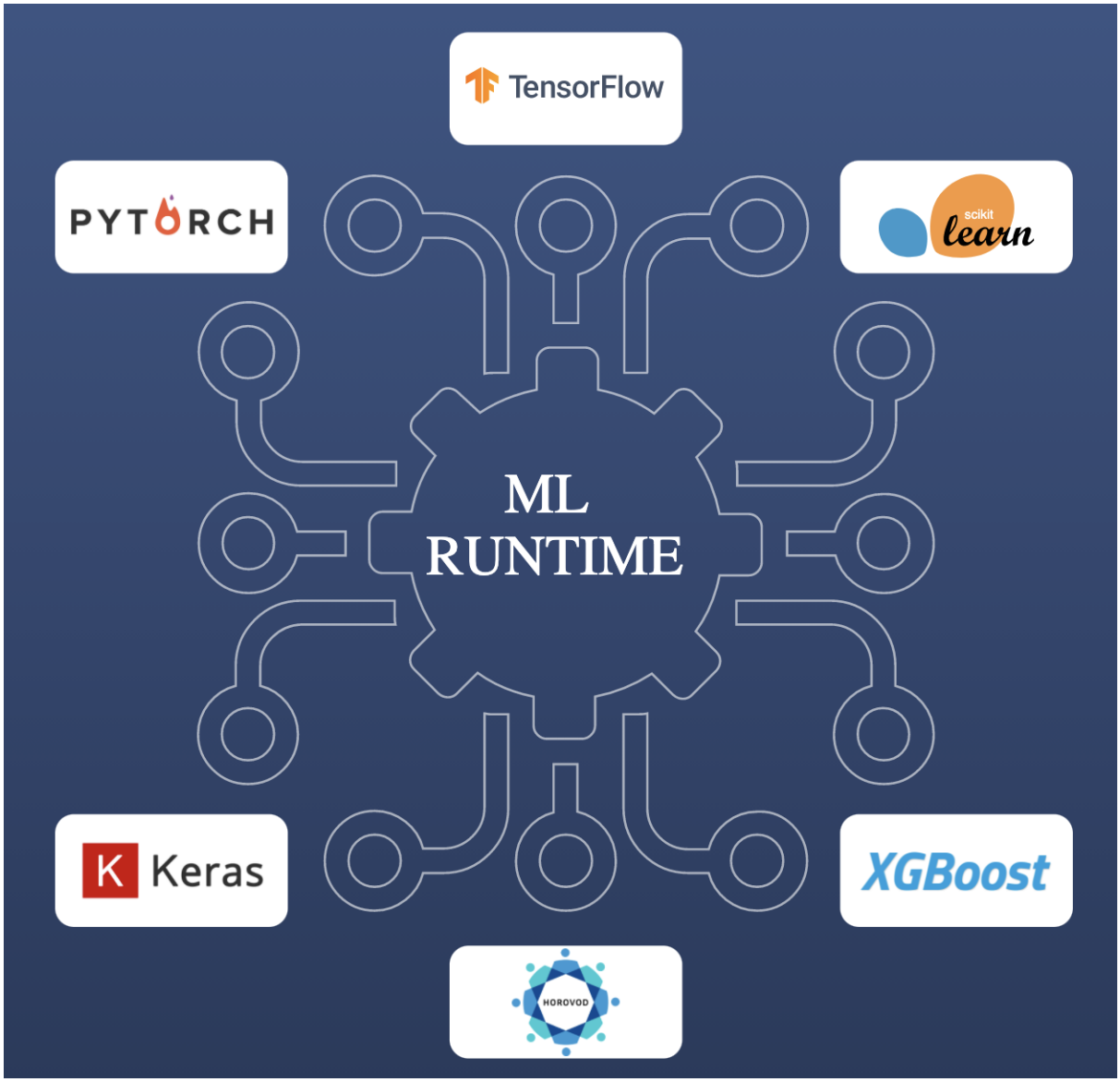
Collaborative notebooks are web-based interfaces that contain runnable code, visualizations, and narrative text. They are used in data science and machine learning to perform exploratory data analysis and build machine learning models. They support multiple programming languages (SQL, Scala, R, Python, and Java), built-in data visualizations, automatic versioning, and the ability to automate processes.



Caption: A screenshot of a Databricks Notebook
Alt-text: An image showing a screenshot of a Databricks Notebook.

Databricks Machine Learning Runtime

The Machine Learning Runtime (MLR) provides data scientists and ML practitioners with scalable computing resources that come with built-in popular data science frameworks (interfaces that help data practitioners quickly build and deploy machine learning models). It provides an optimized computing environment for machine learning workflows.

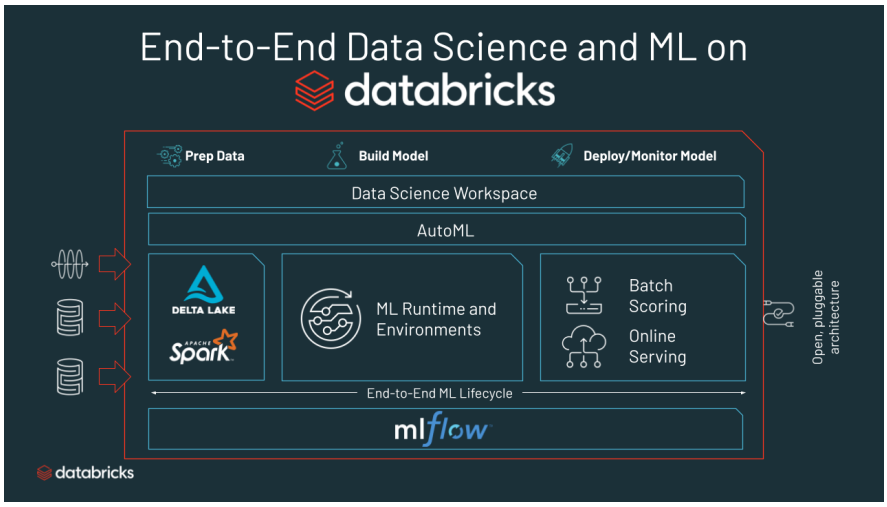


Caption: Popular machine learning frameworks included with the Databricks Machine Learning Runtime.

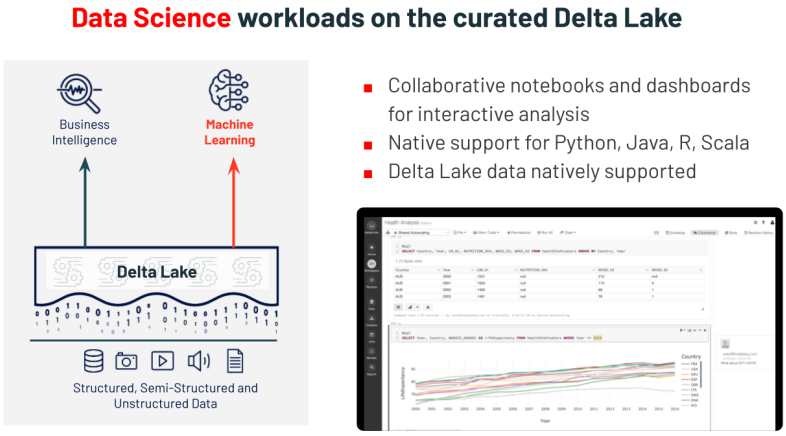
Alt-text: Popular machine learning frameworks included with the Databricks Machine Learning Runtime, including Pytorch, TensorFlow, Scikit Learn, Keras, Horovod, and XGBoost.

Managed MLflow

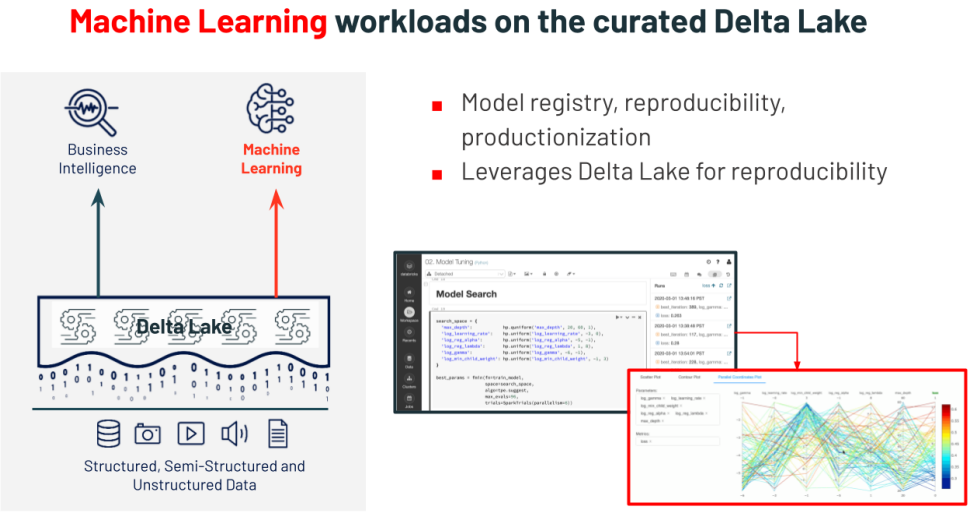
Managed MLflow is built on top of MLflow, an open-source platform developed by Databricks to help manage the complete machine learning lifecycle. With MLflow, you can track machine learning experiments, manage machine learning models, and deploy those models.



Caption: The end-to-end data science and machine learning lifecycle on Databricks.
Alt-text: An image showing Databricks functionality data practitioners can use to work through the end-to-end data science and machine learning lifecycle.

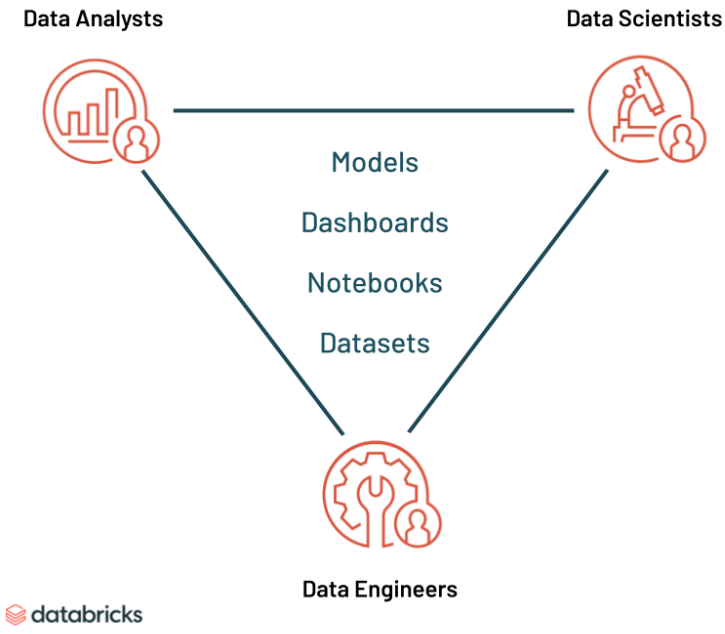


Alt-text: An image summarizing the concepts about data science workloads in this lesson, with a reminder that all of this work is being done on data in your organization's Delta Lake.



Alt-text: An image summarizing the concepts about machine learning workloads in this lesson, with a reminder that all of this work is done on data in your organization's Delta Lake.

Finally, before moving into customer use-cases on Databricks, it's important to stress that all of these capabilities are accessed in the Lakehouse platform. That means that as you and your colleagues use the different tools found within Databricks, you can also share all of the assets you create.



Collaborative

Unify your data teams to collaborate across the entire data and AI workflow

Alt-text: An image showing the assets that data analysts, data scientists, and data engineers can collaborate on as they work in the Databricks Platform. These assets are models, dashboards, notebooks, and datasets.

Next, we'll highlight Databricks customers using the Platform to streamline their big data and AI workflows. In the next lesson, we hope that these stories give you ideas about what you might do in your own organizations now that you are Databricks customers.

Section 5: Customer Success Stories

Before we reach the end of the course, we'd like you to see real-world examples of how Databricks customers have implemented the Lakehouse Platform to help solve business problems. In this lesson, we'll highlight how CVS Health, Scribd, Reckitt Benckiser Group (RB), and Regeneron are using the Lakehouse Platform to streamline their big data and analytics workflows.

CVS Health
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 on the Hadoop environment. The complexity and scale of the diverse data set 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 various segments for personalization at scale.
https://databricks.com/customers/cvs-health

Scribd
Scribd is on a mission to change the way the world reads. With over 60 million titles in its online library, it's focused on leveraging data and analytics to uncover interesting ways that get people excited about reading. Challenged with a legacy Hadoop infrastructure that was too rigid and couldn't scale to meet their real-time needs, Scribd switched to Databricks on AWS and Delta Lake for its performance, elasticity, and ease of use. This migration to the cloud has eliminated infrastructure complexity, allowing their data team to operate with agility, build fast and reliable data pipelines, and easily collaborate on models that deliver an engaging experience to their customers.
https://databricks.com/customers/scribd

RB
As a multinational consumer goods manufacturing company serving millions of retail customers, Reckitt Benckiser Group (RB) struggled with the complexity of forecasting demand, with large volumes of different types of data across many disjointed pipelines. Today, Azure Databricks provides RB with a Unified Data Analytics Platform that enables its data teams to deliver ML-powered insights to the business, improving the support of neighborhood grocery stores through predictive analytics, product placement, and business forecasting.
https://databricks.com/customers/rb

Regeneron
Regeneron'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 discovery and targeted treatments has never been more challenging. With poor processing performance and scalability limitations, their data teams lacked what they needed to analyze petabytes of genomic and clinical data. Databricks now empowers them to quickly analyze entire genomic data sets rapidly to accelerate the discovery of new therapeutics.
https://databricks.com/customers/regeneron

Summary and Next Steps

Congratulations!

You completed the Fundamentals of the Databricks Lakehouse Platform course.

By now, you should be able to:

- Describe what the Databricks Lakehouse Platform is.
- Explain the origins of the Lakehouse data management paradigm.
- Outline fundamental problems that cause most enterprises to struggle with managing and making use of their data.
- Identify the most popular components of the Databricks Lakehouse Platform.
- Summarize Databricks Lakehouse Platform functionality used by data practitioners, depending on their unique role.
- Give examples of organizations that have used the Databricks Lakehouse Platform to streamlining big data processing and analytics.

Next steps:

- Evaluate this course by taking a brief, one-minute survey. Your feedback is valuable and helps us improve our courses.
- Continue your learning journey by visiting the Databricks Academy. The Databricks Academy offers learning plans to help you decide which courses are suitable for you. Current learning plans are created for SQL Analysts, Data Scientists, Data Engineers, and Platform Administrators. Click on the task below that most closely aligns with your responsibilities/interests for a list of self-paced courses you might find helpful.
    - Business management
        - Consider learning more about fundamental concepts about the big data landscape or by Databricks by exploring courses like:
            - Fundamentals of Delta Lake
            - Fundamentals of Structured Streaming
            - Fundamentals of Machine Learning
    - Analysis using SQL
        - Databricks SQL Analytics Fundamentals
        - Data Visualization with Databricks SQL Analytics
    - Data science / Machine learning
        - Apache Spark Programming with Databricks
        - Machine Learning in Production: MLflow and Model Deployment
        - Scalable Machine Learning with Apache Spark
    - Data engineering
        - Apache Spark Programming with Databricks
        - Data Engineering with Databricks
    - Databricks Platform administration
        - Security Fundamentals
        - Identity Access Management
        - Data Access Management
- To learn more about any of the courses listed above, please see the following learning plans:
    - Business Management
    - SQL Analytics
    - Data Science / Machine Learning
    - Data Engineering