Brought to you by:

databricks

The Data Intelligence Platform

for dummies®
A Wiley Brand

Democratize data & AI with intelligence

Understand enterprise data with AI

Accelerate innovation with ETL, DW, BI, & AI

Ari Kaplan
Stephanie Diamond

Databricks Special Edition

About Databricks

Databricks is the Data and AI company. Thousands of organizations worldwide — including Comcast, Condé Nast, Grammarly, and over 60 percent of the Fortune 500 — rely on the Databricks Data Intelligence Platform to unify and democratize data, analytics, and AI. Databricks is headquartered in San Francisco, with offices around the globe, and was founded by the original creators of Lakehouse, Apache Spark™, Delta Lake, and MLflow. To learn more, follow Databricks on social media:

x.com/databricks

linkedin.com/company/databricks

facebook.com/databricksinc

The Data Intelligence Platform

Databricks Special Edition

by Ari Kaplan and Stephanie Diamond

for dummies®
A Wiley Brand

The Data Intelligence Platform For Dummies®, Databricks Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

ISBN: 978-1-394-28407-8 (pbk); ISBN: 978-1-394-28408-5 (ebk); ISBN: 978-1-394-28409-2 (ePub). Some blank pages in the print version may not be included in the ePDF version.

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Project Manager and Editor: Carrie Burchfield-Leighton

Sr. Managing Editor: Rev Mengle

Acquisitions Editor: Traci Martin

Sr. Client Account Manager: Matt Cox

Table of Contents

Introduction

Your organization's success relies on the effective use of data to drive intelligent decision-making and business growth. This involves converting raw data into strategic assets that can be used for analytics or artificial intelligence (AI). Data intelligence with AI gives organizations the power to make smarter decisions and achieve success for your company.

The Databricks Data Intelligence Platform provides a unified platform for data and AI, allowing organizations to democratize data and build AI applications. Teams can collaborate and break down data silos, creating a culture of data-driven decision-making. Fully capitalize on your own data assets, leveraging traditional and generative AI, data warehousing, business intelligence, and governance.

About This Book

The Data Intelligence Platform For Dummies, Databricks Special Edition, explores significant ways companies can shift from reactive to proactive strategies and use their data as a competitive asset. This book covers

- The value of data intelligence and the power of AI
- The Databricks Data Intelligence Platform
- Using traditional and generative AI to build applications
- Reasons why you need a data intelligence platform

Foolish Assumptions

In writing this book, we made a few assumptions about you:

- You want to leverage AI to solve complex problems, and you want solutions that integrate AI with data.
- You're a decision-maker, looking for a unified, open, and scalable platform to drive efficiency, innovation, and a competitive advantage.

- You're responsible for ensuring data governance, security, and regulatory compliance. You want to know how Databricks supports this.
- You want solutions that can scale effectively as data volumes and processing needs grow.
- You want new ways to tackle challenges, whether they're in operational, strategic, or mission-critical domains.
- You're curious how the Databricks platform integrates with existing systems, data infrastructures, and analytics tools.

If any of these assumptions describe you, you've come to the right place.

Icons Used in This Book

Throughout this book, different icons are used to highlight important information. Here's what they mean:

TIP

The Tip icon highlights information that can make doing things easier or faster.

REMEMBER

The Remember icon points out things you need to remember when searching your memory bank.

WARNING

The Warning icon alerts you to things that can harm you or your company.

Beyond the Book

This book can help you discover more about data intelligence platforms, but if you want resources beyond what this book offers, here are more insights:

- See demos, product tours, and tutorials on Databricks. To get started, visit `databricks.com/resources/demos`.
- Join your peers in the 100,000+ strong Databricks community: `community.databricks.com`.

IN THIS CHAPTER

- Exploring the value of data intelligence
- Learning about the key features of data intelligence platforms
- Finding out about use cases in several industries

Chapter 1 Understanding Data Intelligence

When utilized effectively, data intelligence has the potential to make data more accessible to everyone, revolutionizing decision-making and data interaction within organizations. It takes data analytics to the next level by combining generative artificial intelligence (AI) to get greater insights and deliver strategic decision-making. It allows nontechnical users to ask questions in natural language about their organization.

Data intelligence refers to the application of AI to understand the uniqueness of your organization's data to get greater insights and deliver actionable insights. This process involves the use of generative AI to sift through and make sense of vast amounts of your data, enabling organizations to derive intelligent insights that can inform decision-making and improve services, investments, and overall business strategies.

This chapter explores the definition, benefits, and impact of data intelligence, which empowers organizations to turn data into actionable knowledge.

Learning about Data Intelligence

Companies who want to gain a competitive advantage need to make sense of their data. Augmented with generative AI, data intelligence encompasses a range of activities, from data gathering and analysis to applying data insights to solve real-world problems. By leveraging data intelligence, companies can uncover patterns, predict trends, and make evidence-based decisions. This section examines how data intelligence gives companies the tools they need to succeed.

Intelligent

Data intelligence combines generative AI with the unification benefits of a lakehouse architecture to power data intelligence that understands the unique semantics of your data. This allows the Databricks Data Intelligence Platform to automatically optimize performance and manage infrastructure in ways unique to your business.

Simple

Natural language simplifies the user experience. Data intelligence understands your organization's language, so search and discovery of new data is as easy as asking a question like you would to a coworker. Additionally, developing new data and applications is accelerated through natural language assistance to write code, remediate errors, and find answers.

Private

Data and AI applications require strong governance and security, especially with the advent of generative AI. Databricks provides an end-to-end MLOps and AI development solution that's built upon our unified approach to governance and security. You're able to pursue all your AI initiatives — from using APIs like OpenAI to custom-built models — without compromising data privacy and IP control.

REMEMBER

Generative AI is any type of AI capable of interpreting or creating new content by itself. Generative AI content includes text, images, videos, music, translations, summarizations, and code. It can also complete certain tasks, such as answering open-ended questions and participating in chats. The general public was introduced to the meaning of generative AI by solutions such as ChatGPT and DALL-E, which also greatly raised the popularity of the technology.

TIP

Without data intelligence, platforms aren't smart. Platforms like data warehouses (DWs) aren't intelligent because they require highly skilled engineers to manually maintain and optimize the infrastructure. In this example, data intelligence learns about the usage and trends of the platform and applies the learnings to make the platform better and more efficient.

Gathering, analyzing, and applying insights

Companies seeking a competitive edge must effectively gather and interpret their data. Data intelligence helps better collect and combine data into robust datasets, analyzes the data with analytics and AI, and helps with real-world decisions.

Helping companies understand data

Data intelligence is key for companies to make sense of their data. Using advanced tools helps businesses to better understand their customers and the market. It helps them to make smarter decisions. Here are a few examples of how it facilitates this:

- **Enhancing data democratization:** Through the use of natural language, decision-makers can now independently ask questions about their data without the help of more technical programmers. This enables significantly more people to get value from a company's data assets.
- **Streamlining operations:** Data intelligence automatically optimizes performance and manages infrastructure in ways unique to your business.
- **Ensuring data governance and compliance:** Understanding your data also means knowing where it comes from, figuring out how you'll use it, and making sure it complies with legal and ethical standards. Data intelligence provides the tools for effective data governance, helping companies manage data quality and security to ensure it complies with regulations.

Building on the lakehouse architecture

Unified, open, and scalable lakehouse architectures built with data intelligence serve as a comprehensive system that integrates data-related functions into a single, cohesive environment.

TIP

Organizations benefit from a more efficient data management and analysis approach by leveraging unified platforms. It eliminates data silos and provides a single, centralized repository for all data assets. This ensures consistency, accuracy, and governance across the organization.

Maximizing Benefits from Data Intelligence

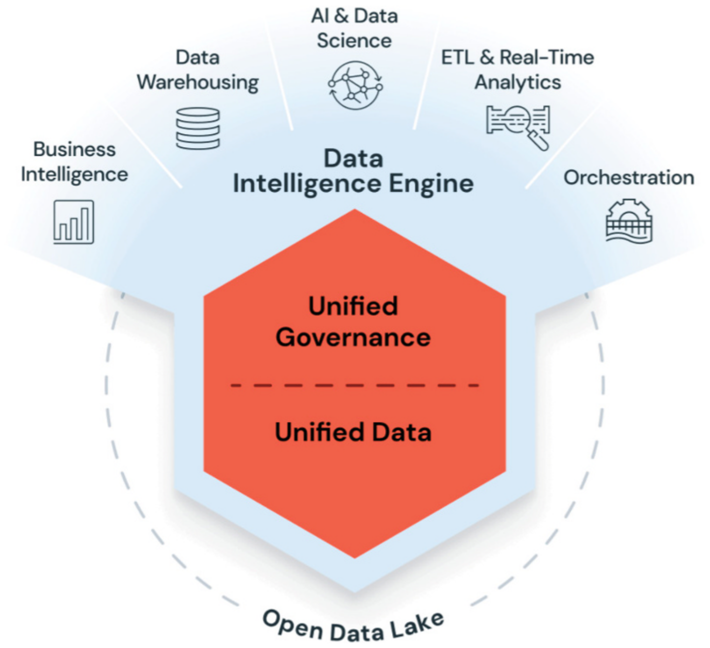
Data intelligence has developed as a key strategy for organizations, helping them utilize the power of their data. Data intelligence simplifies development that doesn't require skilled personnel. This section outlines some benefits these initiatives can provide organizations.

Making data easily searchable and understandable

Data intelligence understands your organization's language, so searching and discovering new data is as easy as asking a question like you would to a coworker. As shown in Figure 1-1, data intelligence makes information easily discoverable and searchable by understanding the context of the search — and not just a simple keyword match. Additionally, adopting natural language processing (NLP) tools can allow users to query data using plain language.

REMEMBER

Natural language is a critical feature of data intelligence and enables the system to understand and interpret language. This allows for extracting important information from large volumes of text more easily with semantics, which you can use to enhance decision-making and customer insights.



SOURCE: DATABRICKS

FIGURE 1-1: Data intelligence engines add unified governance and data in every step of the end-to-end data platform.

Unifying siloed data into a single platform

Unifying siloed data into a single platform addresses an issue many organizations face — data fragmentation across different systems, departments, and locations. When data is siloed, it's isolated from other relevant data, making it almost impossible to understand concepts such as customer behavior or market trends.

WARNING

Fragmented data can lead to inefficiencies and, more importantly, missed opportunities because managers can't see the big picture. Combining data into a single unified platform helps companies break down these silos, allowing data to flow and be analyzed.

REMEMBER

A unified data platform provides a centralized repository where all data (structured or unstructured) can be stored and analyzed. This ensures that data is accurate and allows advanced analytics and AI applications to be used more effectively. As a result, organizations can leverage their data assets to the fullest.

Empowering non-technical users to get data insights

Providing simpler access to data means making it more accessible for nontechnical users. This enables them to get insights without relying on IT departments. Make sure that all employees understand the basics of data analysis and the tools available that make it so simple.

Streamlining company operations and cost savings

Data intelligence can streamline company technology operations that lead to cost savings. Predictive analytics forecast trends and allow companies to adjust their strategies.

TIP

AI can uncover new opportunities for technology efficiencies and cost reduction, such as automating time-consuming manual processes. Examples include identifying and reallocating improperly utilized resources, or reorganizing how data is stored so it can scale larger and faster. The key is that AI can now improve the operations of every technology step from raw data software engineering to analytics.

Fostering collaboration

Data intelligence tools facilitate collaboration across different teams by providing a common environment into all tasks. Teams can work simultaneously on the same datasets, develop code together, share dashboard and reporting insights, and make collective decisions. This collaborative environment encourages people to work toward common goals.

Impacting the Entire Business

Data intelligence improves the functionality and efficiency of every aspect of business. It drives their evolution, ensuring that they respond more to human needs and ethical standards. This enhances the entire data and AI ecosystem, making it capable of addressing complex challenges. This section discusses key ways data intelligence shapes the landscape.

Improving data quality and integrity

The integrity and quality of data are foundational to the effectiveness of both AI systems and analytical processes. Data intelligence enhances these aspects by providing mechanisms for data validation, cleansing, and consistent management across different data sources.

Driving innovation and new business models

Data intelligence is crucial in delivering innovation and shaping new business models. Companies can identify emerging trends and underserved market needs by analyzing their data. This opens up opportunities for new and innovative products and services.

TIP

A data-driven approach allows businesses to experiment with business models, such as subscription services or on-demand platforms, which can provide a competitive edge. Insights gained from data can lead to new revenue streams and transformative strategies.

Accelerating AI and ML

Data intelligence provides the foundation for AI and ML by preparing and transforming data into a format these technologies can use. High-quality, well-governed data is essential for training accurate and reliable AI models.

Evaluating Key Features of Data Intelligence Platforms

Data intelligence platforms give businesses the tools they need to make their data a valuable business asset. They bring data together in a unified platform to analyze it and formulate effective strategies. Knowing what these platforms can do helps you choose the right one for your data needs and goals.

TIP

When evaluating data intelligence platforms, organizations should consider factors such as scalability, performance, ease of use, and integration capabilities to ensure that the chosen platform aligns with their specific business requirements and technical infrastructure.

Using NLP

TIP

NLP technology is at the core of tools like translation software, chatbots, and search engines. By leveraging NLP, data intelligence platforms can unlock the full potential of your company's unstructured data such as customer reviews, social media posts, and support tickets.

Ensuring data security and scalability for growth

Security and scalability are vital for the growth of any organization. Data intelligence provides strong security features to safeguard private information and ensure compliance with data regulations. Additionally, it must be scalable to accommodate the growing data and the organization's expanding needs.

Making the platform usable for diverse skill levels

Data intelligence platforms should be accessible to people with varying levels of technical expertise. This ensures that a range of users, from data scientists to business analysts, within your organization can leverage the platform's capabilities.

TIP

Simplifying the user experience, through no-code tools and intuitive interfaces, can widen data usage and empower more stakeholders to make more informed data-driven decisions.

Automating data processes

Automation in data intelligence platforms transforms how companies handle their vast amounts of data. Businesses can significantly enhance efficiency, accuracy, and speed by integrating automation into their data processes. It streamlines workflows, reduces manual intervention, and improves the overall data management experience.

REMEMBER

One of the most important benefits of automation is that it reduces the need for manual data handling. Manual tasks are time-consuming and also prone to errors. Automation minimizes the need for human input, reducing the risk of data handling mistakes.

TIP

Businesses can improve operational efficiency, save time, and apply resources to strategic initiatives by automating tasks like data collection, cleaning, and processing.

Examining Data Intelligence Use Cases in Diverse Industries

Data intelligence is used in many industries from finance to healthcare to energy. Using data-driven insights changes how businesses operate. In this section, you look at the various use cases across sectors to see how data intelligence helps companies learn about their customers, improve processes, and spot fraud. A few examples include

- **Finance:** This sector uses data intelligence to handle financial risks, predict economic trends, and follow regulations. Banks and other financial institutions use data analysis to assess creditworthiness, spot fraud, and categorize customers.
- **Retail and CPG:** They apply data intelligence to learn about customer preferences, better manage stock, optimize supply chain, and tailor marketing to individual purchasers.
- **Public sector:** In the public sector, data intelligence is important for improving services and making policy choices. Government agencies use data to monitor the changing economy and make services more helpful.
- **Insurance:** Insurance companies use data intelligence to evaluate risks, set prices for insurance plans, and find false claims. By studying large amounts of data, they can get a clearer picture of risks and make filing claims more efficient.
- **Healthcare:** Healthcare organizations use data intelligence to improve patient care, control costs, and do research. Data analytics helps make medical decisions and find treatments that work.
- **Energy:** Energy companies use data analysis to track and forecast energy use and improve the efficiency of the power grid.

REMEMBER

Data intelligence applications may vary across industries, but the common goal remains the same. It’s to extract valuable insights from data and leverage them to drive business growth and enhance customer experiences.

IN THIS CHAPTER

- » Looking at challenges of not having a lakehouse
- » Exploring lakehouses, data warehouses, and data lakes
- » Looking at generative AI and traditional AI
- » Understanding the potential of AI
- » Using lakehouse architecture and generative AI
- » Deploying data intelligence platforms

Chapter 2
Exploring the Lakehouse and Generative and Traditional AI

Data intelligence platforms build on lakehouse architecture with generative artificial intelligence (AI) and offer powerful ways to democratize data and AI across an organization. Lakehouse architectures store and process huge amounts of structured and unstructured data together into one unified environment, bringing data warehousing, business intelligence, traditional AI, and generative AI to new heights.

This chapter examines the differences among lakehouses, data warehouses (DWs), and data lakes, and how adding generative and traditional AI to a lakehouse enhances its value to organizations.

Experiencing Challenges without a Lakehouse

Most companies find it challenging to effectively combine data and AI to achieve their business objectives. These challenges incorporate various components, each critical to the data intelligence ecosystem. To combine data management and AI, you face these problems:

- Your data and AI are siloed. Data silos drive high operational costs.
- Your data privacy and control are challenged. Inconsistent policies reduce the trust in the data.
- You depend on highly technical staff. Disparate tools slow down cross-team production.

You have to stitch together several services to make things function. Each component has its challenges. Take a look at Figure 2-1; starting with data lake and going clockwise each component is explained with its challenges:

- **Data lake:** The challenge lies in storing and managing huge, unstructured datasets.
- **Machine learning (ML):** The challenge is to experiment, develop, apply, and monitor the accuracy of complex algorithms.
- **Streaming:** The technical demand for processing continuous data streams in real-time is high.
- **Generative AI:** There is complexity in generating new, realistic content with AI technologies.
- **Data warehouse:** The issue is centralizing structured data for analysis, which can be complex and costly.
- **Business intelligence (BI):** The difficulty is in visualizing data effectively and being able to extract business insights.
- **Orchestration and extract, transform, load (ETL):** This involves coordinating data preparation and movement.

- **Governance:** The challenge is to navigate regulations and implement strong data management and security controls.
- **Data science:** The task is complex when exploring and analyzing data using scientific methods.

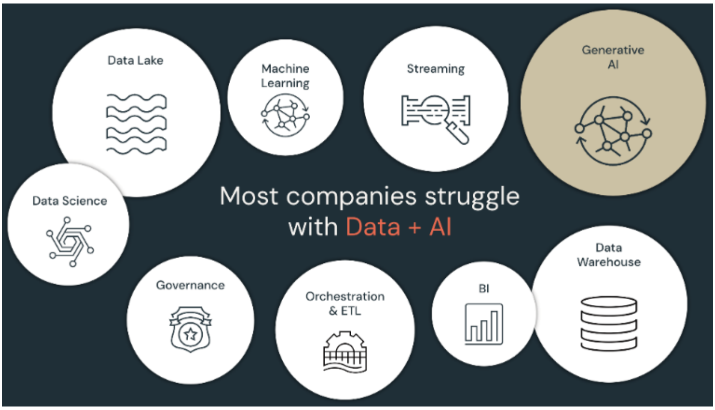


FIGURE 2-1: The challenges of the data intelligence ecosystem.

For more details about these components and how the Databricks Data Intelligence Platform solves these challenges, see Chapter 3.

REMEMBER

A data intelligence platform has many elements that help an entire organization: an open data lake for storing and managing all your data types; unified data storage for reliability and sharing; a unified security, governance, and catalog environment; and an AI-powered engine to understand the semantics of your data. A data intelligence platform improves the experiences of data science and AI, ETL, and real-time analytics, orchestration, and data warehousing.

Comparing Lakehouses with Data Warehouses and Data Lakes

Lakehouses represent a distinct approach to data storage and analytics from DWs and data lakes. The transition from DWs and data lakes to lakehouses was prompted by the need for more scalable, open, and cost-effective solutions for managing huge amounts of structured and unstructured data. In this section, you look at the differences.

Open architectures

With Databricks, your data is always under your control, free from proprietary formats and closed ecosystems. The Databricks Lakehouse is underpinned by widely adopted open source projects Apache Spark, Delta Lake, and MLflow. On top of this, Delta Sharing provides an open solution to securely share live data from your lakehouse to any computing platform without costly replication and complicated ETL.

Unified architecture

Lakehouse architecture unifies all integration, storage, processing, governance, sharing, analytics, and AI. It's one approach to working with structured and unstructured data, one end-to-end view of data lineage and provenance, one notebook for all Python, R, Scala, and SQL, one source for batch and streaming, and one platform for all three major cloud providers.

Scalable

Lakehouses are more scalable than traditional DWs and data lakes; they scale up to trillions of records with lower cost and higher performance. They offer automatic optimization for performance, and storage ensures the lowest total cost of ownership (TCO) of any data platform together with world-record-setting performance.

Improving data governance and security

Lakehouses improve how data is governed and protected by using one security and governance model for all data and AI access across the organization. Having one unified governance platform makes it easier to follow regulations and keep data safer than traditional DWs and data lakes because they have multiple disjointed governance solutions and different kinds of data, making it more difficult to apply consistent policies and protections.

Distinguishing between Generative and Traditional AI

Generative AI and traditional AI are two branches of AI. Traditional AI is useful in fields where you need to make numerical predictions or classify items, such as predicting future sales across each store or grouping your millions of customers into different segments. On the other hand, generative AI offers value in making text summaries or answering general questions from unstructured data such as PowerPoints, PDFs, and Word documents.

Generative AI generates new content based on the patterns it learned from its training data. Instead of simply analyzing data, generative AI systems can interpret and search through text, images, audio, video, or other media. Generative AI can also be used internally to create new text and to write and edit software code.

REMEMBER

A key capability of generative AI is its ability to produce new content that's like the training data in style and structure but isn't a direct copy of anything. By identifying and learning from the underlying patterns in the data, generative models such as Large Language Models (LLMs) can combine concepts to create original creations. Some examples of generative AI for business include

- **Generating text:** Can write human-like sentences and explanations and is trained on a company's own data
- **Summarizing text:** Can take large amounts of documents and give a synopsis or a grade for easier interpretations from humans
- **Writing software code:** Can take simple prompts and write code in SQL, Python, Scala, and R
- **Documenting data assets:** Can describe the contents of a table and columns for better semantic searches

Realizing the Significance of AI in Enhancing Data Intelligence

Combining AI with data intelligence creates great advancements in how businesses analyze, understand, and leverage their data. This impact boosts the strategic capabilities of organizations. It allows companies to adapt more quickly to market changes and consumer needs.

REMEMBER

The importance of AI in improving data intelligence lies not just in its technological ability but in its ability to drive innovation across various sectors.

Employing the Use of Lakehouse Architecture and Generative AI

Integrating lakehouse architecture with generative AI enhances the capabilities of both technologies. This integration creates a more powerful environment for data processing and analytics.

Leveraging a lakehouse architecture

Lakehouse architecture provides a data storage and management foundation by combining the best features of data lakes and data warehouses. It enables organizations to store structured and unstructured data in a single repository while still being able to perform analytics and ML tasks.

REMEMBER

Incorporating generative AI introduces new ways to analyze and generate data-driven insights. This technology can help organizations improve data quality and develop more accurate predictive models, which is a competitive advantage.

By integrating lakehouse architecture and generative AI, organizations can manage their data more effectively and find new possibilities for data-driven decision-making.

Utilizing open data storage

Utilizing an open data storage helps with reliability and data sharing. It’s essential for employing lakehouse architecture and incorporating generative AI capabilities. It's a technology that provides an efficient data storage layer that helps organizations develop and deploy generative AI applications easily and efficiently. By leveraging this organizations can use the power of generative AI to best drive innovation.

Integrating generative AI capabilities into a lakehouse

Integrating generative AI capabilities into a lakehouse architecture enhances data analysis. By leveraging the combined strengths of a lakehouse, organizations can elevate their data intelligence efforts. A few key benefits of this integration are

- **Automating data tasks:** Generative AI can streamline data operations within a lakehouse. While traditional AI may automate data cleansing, generative AI can further assist by generating artificial data for testing and training models, ensuring robust analytics and AI applications.
- **Enhancing search functions:** AI enhances intelligent search capabilities within a lakehouse. Users can utilize natural language queries to discover and comprehend the relationships between data assets efficiently. This simplifies data discovery beyond a mere keyword search and ensures that the right datasets are easily accessible for analysis.
- **Developing custom AI applications:** Integrating AI into the lakehouse framework allows organizations to create applications tailored to specific needs. Examples include making LLMs on your own company's data, developing predictive models, customizing recommendation engines, or automating complex reporting tasks.

REMEMBER

Integrating generative AI capabilities into a lakehouse architecture empowers organizations to extract more value from their data.

Enabling data teams to collaborate

The combination of lakehouse architecture and generative AI capabilities significantly improves the collaborative potential of data teams. It enables a more dynamic exchange of ideas. This fosters a culture of innovation. Data teams can work together to build, train, and deploy AI models more efficiently. This leverages the strengths of a lakehouse's data management and the creative potential of generative AI to drive business growth.

Enhancing data analysis and insights with AI

To enhance data analysis with AI, leverage various AI-powered tools and techniques to automate and streamline multiple stages of the data analysis process. AI can be integrated into different phases of data analysis in the following ways:

- **Data preparation:** AI can automate the data preparation phase, which includes cleaning, organizing, and preprocessing data. AI tools can detect and correct data quality issues, extract information from unstructured data, and combine data from different formats.
- **Data exploration:** AI algorithms can explore your data by using natural language. This helps uncover insights that may not be apparent to humans.
- **Data interpretation:** AI can improve data interpretation by generating summaries, insights, or stories from your data. It can identify causal relationships and predict future outcomes or actions based on the data.
- **Data quality:** AI can help detect when data and model quality is skewed, automatically flag it, and assist in remediation.

Automating complex data tasks and processes

Automating complex data-related tasks and processes means making the handling and analysis of data more efficient by using technology to do the work. This organizes large amounts of unstructured data in data lakes, builds and applies generative AI and ML models, and deals with the continuous data flow in real time.

For orchestrating jobs, AI can automatically select the right instances and start time to hit your requirements. It handles tasks like auto-scaling and error remediation for you.

Many aspects of data engineering, such as optimizing file sizes for tables, can benefit from AI and be automated as well. Engineers typically invest a lot of their time and expertise to figure out the optimal file sizes for reading or writing data, which can lead to significant performance improvements. Automating this complex task is a game-changer.

Intelligent autoscaling in ETL processing optimizes cluster utilization and minimizes end-to-end latency for streaming workloads by automatically adjusting resources based on data volumes and processing needs, up to a specified limit. It efficiently scales up when data arrival outpaces processing and scales down during low load, ensuring task completion before shutting down to save on infrastructure costs.

Deploying a Data Intelligence Platform

A data intelligence platform helps organizations innovate with a unified platform that combines uses across personas such as data scientists, data engineers, architects, and business analysts. It combines different stages of data into a single, integrated environment. The following enables this integration:

- **Integrating data:** You can bring data from different sources into a single place. It supports data integration from databases, data warehouses, data lakes, and streaming data sources, making working with all your data in one platform easier.
- **Processing and analysis:** After your data is in the platform, you can process and analyze it. The platform supports all major languages you may prefer, such as Python, R, Scala, and SQL. Using built-in functions and libraries, you can easily clean, transform, and analyze your data.
- **Collaborating in a workspace:** The platform gives you shared workspaces where different team members can work together on the same data and projects. Data engineers, data scientists, and analysts can create visualizations and dashboards, all on the same platform. Shared workspaces help maintain version control, so everyone uses the most current data and analyses.

- **Using a unified place:** You can manage your entire data analytics workflow from a single place to control access to data, manage resources, and monitor jobs in one place. This allows better resource allocation and monitoring of ongoing jobs.
- **Deploying seamlessly:** After you've built your data analytics solution, you can easily deploy it to production. Seamless deployment lets organizations move data projects from development to production without problems typically found without a data intelligence platform.

Check out Chapter 3 for more information on this platform.

IN THIS CHAPTER

- » Reviewing the Databricks Data Intelligence Platform
- » Looking at the Databricks Platform architecture
- » Assisting developer efforts

Chapter 3 Getting Started with the Databricks Data Intelligence Platform

Companies are constantly looking for ways to simplify their data architecture while improving their ability to gain meaningful insights. This chapter looks at the foundational aspects of getting started with the Databricks Data Intelligence Platform.

Introducing the Databricks Data Intelligence Platform

The Databricks Data Intelligence Platform allows your entire organization to use data and artificial intelligence (AI). It's built on a lakehouse to provide an open, unified foundation for all your data, AI, and governance needs and is powered by a data intelligence engine that understands your data's uniqueness.

From extract, transform, load (ETL) to data warehousing to generative AI, Databricks helps you simplify and accelerate your data and AI goals.

Delivering data intelligence with DatabricksIQ

Databricks combines the power of generative AI with the comprehensive features of a lakehouse architecture to create a data intelligence engine called DatabricksIQ. DatabricksIQ learns the unique nuances of your business and data to power natural language access to it for a wide range of use cases. Any employee in your organization can search, understand, and query data in natural language. DatabricksIQ uses information about your data, usage patterns, and trends to understand your business's jargon and unique data environment and give significantly better answers than the naive use of Large Language Models (LLMs) — a form of generative AI that can perform a wide range of language-related tasks, including translation, summarization, question-answering, and text generation.

LLMs, of course, promised to bring language interfaces to data, and many data companies are adding an AI assistant, but in reality, many of these solutions fall short on enterprise data. Every enterprise has unique datasets, jargon, and internal knowledge that's required to answer its business questions, and simply answering questions through an LLM trained on the Internet often gives wrong results. Even something as simple as the definition of a customer or the fiscal year varies across companies.

DatabricksIQ is a data intelligence engine that directly solves this problem by automatically learning about business and data concepts throughout your enterprise. It uses signals from across the Databricks Platform, including Unity Catalog (UC), dashboards, notebooks, data pipelines, and docs, leveraging the unique end-to-end nature of the Databricks Platform to see how data is used in practice. This lets DatabricksIQ build highly accurate specialized models for your enterprise.

Simplifying the user experience through natural language

Using natural language processing (NLP) greatly simplifies the user experience on Databricks. The Databricks Data Intelligence Platform is designed to understand the specific terminology used in your organization. This makes searching for and discovering data as straightforward as asking a coworker a question.

TIP

NLP enables the system to understand and interpret language. This capability extends to developing new data applications. It assists in writing code, correcting errors, and providing answers. This speeds up the development process.

Ensuring privacy and governance

The need for strong governance and security in data and AI applications has never been more important. Databricks offers a comprehensive solution for Machine Learning Operations (MLOps) and AI development supported by a unified approach to governance and security. This solution allows you to pursue a wide range of AI initiatives, which enables you to maintain privacy and control over your intellectual property.

REMEMBER

MLOps is a function of ML engineering. It's focused on streamlining the process of taking ML models to production and maintaining and monitoring them as your data changes.

Using the Data Intelligence Platform

Databricks has created a data intelligence platform that utilizes the power of the data lakehouse and generative AI. Significant strides have been made in exploring the potential of AI within data lakehouse platforms. The Databricks Data Intelligence Platform stands out because its unified governance layer covers both data and AI. It also has a single query engine that spans ETL, SQL, ML, and business intelligence (BI). In addition, the integration of Mosaic AI helps to develop AI models that support DatabricksIQ. This integration is crucial in making data accessible to everyone in your organization.

REMEMBER

Databricks pioneered the lakehouse concept. It provides an open, unified architecture for all data and governance needs. It also enables organizations to store and manage structured and unstructured data in one system.

TIP

Databricks developed performance enhancements like Photon (the next-generation engine that provides extremely fast query performance at low cost), which makes the platform more scalable and efficient. This ensures that Databricks can handle even the largest-scale data workloads.

A visual representation of the architecture of the Databricks Data Intelligence Platform is shown in Figure 3-1. Starting at the bottom of the figure, this section looks at each component to see how everything fits together.

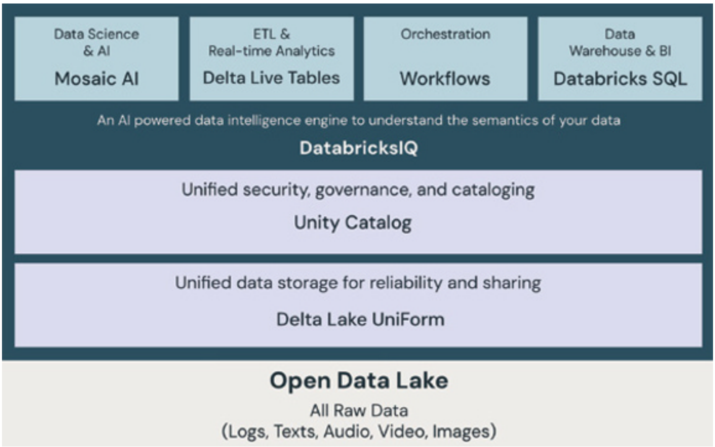

Data Science & AI
Mosaic AI
ETL & Real-time Analytics
Delta Live Tables
Orchestration
Workflows
Data Warehouse & BI
Databricks SQL
An AI powered data intelligence engine to understand the semantics of your data
DatabricksIQ
Unified security, governance, and cataloging
Unity Catalog
Unified data storage for reliability and sharing
Delta Lake UniForm
Open Data Lake
All Raw Data
(Logs, Texts, Audio, Video, Images)

FIGURE 3-1: The Databricks Data Intelligence Platform.

Open Data Lake

With Databricks, your data is always under your control, free from proprietary formats and closed ecosystems. The data lake can be used to store, refine, analyze, and access data types needed for many new data applications, including images, video, audio, semi-structured data, and text.

Delta Lake UniForm

With Delta Lake Universal Format (UniForm), you use your favorite Iceberg or Hudi client to read your Delta tables through the UC endpoint. DatabricksIQ uses AI models to solve common data storage challenges, so you get faster performance without having to manually manage tables, even as they change over time.

REMEMBER

Storage has three major formats: Delta Lake, Apache Iceberg, and Apache Hudi. In the past, companies duplicated and had multiple copies of their data in several places and formats. This practice is costly and time-consuming and doubles your cost and effort.

Databricks introduced Delta Lake UniForm so that your data can be stored in any of these three formats and still be processed (for business intelligence, AI, and so on) without copying data over and over again.

Unity Catalog

Databricks UC offers a unified governance layer for data and AI within the Databricks Data Intelligence Platform. With UC, organizations can seamlessly govern their structured and unstructured data, ML models, notebooks, dashboards, and files on any cloud or platform. Data scientists, analysts, and engineers can use UC to securely discover, access, and collaborate on trusted data and AI assets, leveraging AI to boost productivity and unlock the full potential of the lakehouse architecture. This unified approach to governance accelerates data and AI initiatives while simplifying regulatory compliance.

DatabricksIQ

DatabricksIQ is built on and governed by UC. DatabricksIQ improves governance in UC by automatically inserting descriptions and tags of all data assets in UC. These assets are then leveraged to make the whole platform aware of jargon, acronyms, metrics, and semantics. This process enables better semantic search, better AI assistant quality, and improved ability to do governance.

DatabricksIQ also significantly enhances Databricks' in-product Search. The new search engine doesn't just find data; it interprets, aligns, and presents it in an actionable, contextual format, helping all users get started faster with their data.

TIP

After assets are registered in UC, DatabricksIQ significantly improves the discoverability of data by allowing users to search for data using natural language (NL) and company-specific terminology, making it easier for users to use data assets within the organization.

Mosaic AI

Databricks' acquisition and integration of Mosaic into its Data Intelligence Platform has significantly boosted its capabilities around LLMs. It lets users fine-tune or create customized generative AI applications that fit their specific needs. This integration enables users to either start from scratch or refine preexisting models, all while ensuring the privacy and control of their proprietary data.

REMEMBER

The platform's utilization of generative AI enhances data comprehension, providing a semantic understanding that powers intelligent search functions, assists in the creation and modification of SQL code, and automatically generates detailed descriptions for data tables and columns. Combining features from Mosaic and Databricks' generative AI capabilities creates a powerful environment for developing AI applications focusing on data security and user autonomy.

For more about Mosaic AI and how the platform enables you to build your own generative AI applications see Chapter 4.

Delta Live Tables

Delta Live Tables (DLT) is a declarative ETL framework for the Databricks Data Intelligence Platform that helps data teams simplify streaming and batch ETL cost-effectively. Simply define the transformations to perform on your data and let DLT pipelines automatically manage task orchestration, cluster management, monitoring, data quality, and error handling.

With DLT, you describe what your ETL should do, and the data intelligence engine understands the data and transformations and autoscales the workload for processing. DatabricksIQ handles everything, updating only what's necessary for optimal total cost of ownership. Additionally, when new data is added, the engine figures out the best way to update the underlying table, making streaming/real-time ETL affordable. Built-in data quality and monitoring are vital to enable downstream business apps.

REMEMBER

ETL is the process data engineers use to extract data from different sources. Then, they transform the data into a usable and trusted resource. Finally, they load that data into the systems that end-users can access and use downstream to solve business problems.

Databricks Workflows

Databricks Workflows orchestrates data processing, ML, and analytics pipelines on the Databricks Data Intelligence Platform. With a wide range of task types, deep observability capabilities and high reliability provide your data teams with the tools to better automate and orchestrate any pipeline on serverless compute.

With data intelligence at its core, Databricks Workflows not only simplifies debugging and alerting by suggesting potential resolutions but also allows you to analyze all interactions, making it easy to identify which job and team is responsible for processing your data. This simplifies data processing and observability. When a job fails, the workflow intelligently recovers tasks and reruns only the necessary parts, resulting in a much lower total cost of ownership and enhanced intelligence.

Databricks SQL

Databricks SQL is one of the leading serverless data warehouses. A few of its capabilities include

- Running your ETL workloads and BI, with the added benefit of governance through UC
- Using open source foundational architecture that scales with optimal price and performance
- Optimizing how quickly queries run, which makes data analysis easier
- Using advanced techniques to speed up data access, including tools for building queries and reports

Databricks SQL utilizes a next-generation vectorized query engine called Photon and is packed with thousands of optimizations to provide you with the best performance for all your tools, query types, and real-world applications. This includes the AI-powered predictive I/O that eliminates performance tuning like indexing by intelligently prefetching data using neural networks.

SQL is crucial for data analysis due to its versatility, efficiency, and widespread use. Its simplicity enables swift retrieval, manipulation, and management of large datasets. Incorporating AI functions into SQL for data analysis enhances efficiency, which enables businesses to swiftly extract insights.

AI Functions is a built-in Databricks SQL function, allowing you to access LLMs directly from SQL. AI Functions abstracts the technical complexities of calling LLMs, enabling analysts and data scientists to start using these models without worrying about the underlying infrastructure.

LOOKING AT DATABRICKS AI/BI

Databricks AI/BI is a first-of-its-kind analytics offering built on data intelligence to democratize BI for everyone in your organization. Powered by DatabricksIQ, the Data Intelligence Engine for Databricks, AI/BI understands your unique data and business concepts. It does this by automatically capturing signals from your data platform along with curated instructions and proactively seeking and incorporating clarifications, ensuring users receive relevant and accurate AI-generated insights from their complex, real-world data.

Dashboards help analysts quickly build highly interactive data visualizations for their business teams by using natural language, and Genie offers a conversational experience for business users to self-serve their own analytics. They can ask questions in the same manner they would ask an experienced coworker, enabling them to get trusted answers directly from their data without relying on technical experts.

Databricks AI/BI is native to the Data Intelligence Platform, providing instant insights without trading off interactive performance for data scale while ensuring unified governance and fine-grained security through Unity Catalog.

REMEMBER

Databricks is built on open source technology, which differentiates it from some competitors that use proprietary systems, potentially leading to locking you into a specific vendor. This open approach encourages innovation with contributions from the open source community.

Using DatabricksIQ to Assist Programmers

Databricks Assistant is a context-aware AI assistant available natively in Databricks notebooks, SQL editor, and file editor. Databricks Assistant lets you query data through a conversational interface, making you more productive inside Databricks. You can describe your task in English and let the Assistant generate SQL queries, explain complex code, and automatically fix errors. The Assistant leverages UC metadata to understand your tables, columns, descriptions, and popular data assets across your company to provide responses that are personalized to you.

Generating SQL, Python, R, and Scala code

Generative AI assists in generating code, which streamlines the process of data querying. It can automatically generate code by understanding the semantics of the data and the intent behind a user query. This reduces the time and effort required to perform data operations.

For example, write a SQL program to find which ten cities sold the most bicycles or write a Python program that divides annual salaries into bi-weekly salaries for employees.

Transforming code to different languages

One of the capabilities of generative AI is its ability to transform code from one programming language to another, which is useful in environments where you use multiple programming languages. This feature enables seamless integration and interoperability between systems and applications.

Documenting or explaining existing code

Understanding existing code, especially in complex projects, can be overwhelming. Generative AI aids in documenting or explaining code and can provide clear and concise explanations of what specific code segments do. This not only helps in onboarding new team members but also in maintaining and updating the codebase.

Debugging and fixing issues and errors

Generative AI can identify potential issues and errors in code, offering suggestions for debugging and fixing them. This approach to error detection and resolution can reduce development time and improve the quality of the software. It can be as easy as typing “/fix” into the prompt, and the code is fixed along with handy documentation and links to learn more.

Getting contextualized responses

With generative AI, responses can be a transformative tool for developers. By adapting to individual coding habits, project specifics, and data meanings, generative AI makes sure that every piece of advice and support is relevant and directly applicable to the business and its up-to-date data. This makes the development process easier, more intuitive, and more relevant.

IN THIS CHAPTER

- » Using traditional AI to build applications
- » Managing traditional AI development challenges
- » Getting into model management
- » Building applications with generative AI
- » Seeing it all come together

Chapter 4 Building AI Applications on a Data Intelligence Platform

With the integration of traditional and generative artificial intelligence (AI) into almost every facet of business technology, companies must develop their own custom AI applications to better serve their customers and stay ahead of the competition.

This chapter explores how the Databricks Data Intelligence Platform supports traditional and generative AI application development, and their management through Machine Learning Operations (MLOps) and Large Language Model Operations (LLMOps). It delves into the platform's tools and features that enable seamless feature engineering, model creation, model experiment tracking, automation of ML, and deployment of AI applications. The Databricks Data Intelligence Platform provides unified tooling to build, deploy, and monitor AI and ML solutions, ranging from building predictive models to the latest generative AI and LLMs.

Developing Traditional AI Applications

REMEMBER

Traditional AI uses models built on explicit programming algorithms. This form of AI depends on human guidance for its logical rules and decision-making processes. Traditional AI models are optimized for specific, structured tasks such as predictions and classifications. One of the strengths of data intelligence platforms is MLOps and LLMOps, which helps you with the entire life cycle of AI models from developing models, tracking experiments, managing models, deploying models, and monitoring the health of all AI models as the underlying data changes.

Using Delta Live Tables and Databricks Workflows

Delta Live Tables and Databricks Workflows enhance the development and deployment of traditional AI applications:

- **Delta Live Tables:** This feature simplifies the building and maintaining of reliable data pipelines at scale. It automates many aspects of the extract, transform, load (ETL) process, ensuring data integrity and reducing the need for manual oversight. Delta Live Tables lets you define tasks to orchestrate in the right order; it lets you set data quality rules to handle issues with data as it comes in; it provides robust error handling to quickly find root causes if something doesn't go as planned; and it helps you monitor your entire pipeline with event logs. It's easy, it's scalable, and it ensures high data quality in a distributed environment.
- **Databricks Workflows:** This capability allows you to run your jobs without pre-configuring and managing the underlying infrastructure. It automatically optimizes and scales resources for your workloads and starts up environments nearly instantly, making implementing data processing and analysis pipelines easier. You also reduce costs and keep performance high.

Ensuring governance, security, and compliance

AI applications are becoming increasingly important to business operations, and the need grows for strong governance, security, and compliance measures. Through Unity Catalog (UC), Databricks provides comprehensive governance and security features to ensure data privacy and regulatory compliance.

REMEMBER

These features are vital for both traditional and generative AI applications that handle sensitive or proprietary information. This ensures that everything from raw data to AI models, notebooks, and applications is used responsibly and protected against unauthorized access.

Addressing the Challenges of Traditional AI Development

Developing traditional AI applications comes with its own set of challenges:

- **Poor data quality and availability:** The foundation of any AI model is data. Poor data quality and insufficient data can impede the performance, accuracy, and trust of AI models.
- **Model complexity:** Traditional AI models can become complex, making them difficult to understand, trust, manage, and scale.
- **Integration with a broader ecosystem:** Integrating AI applications with multiple internal and external business systems and workflows can open up broader customization and configuration.

Databricks addresses these challenges by using a suite of features designed to streamline the AI development life cycle:

- **Managing data:** Databricks provides comprehensive tools for data integration, processing, and quality management, ensuring that AI models have access to high-quality data.
- **Simplifying AI workflows:** Databricks' unified approach simplifies the management of complex AI models and workflows. Generative AI can automate tasks such as data preparation and initial data analysis, making these tasks more accessible.

- **Seamless integration:** Databricks offers various application programming interfaces (APIs) and connectors that facilitate the integration of AI applications with existing systems, which ensures a smooth deployment.

Considering Model Management and MLOps/LLMOps

AI models don't just pop into existence. There is an end-to-end process for building, deploying, and managing models, and the Databricks Data Intelligence Platform helps every step of the way. This section delves into the world of MLOps/LLMOps.

Refining feature engineering

Models are built on data in the lakehouse environment — either directly in the Databricks Data Intelligence Platform or through connections to third-party environments. The Databricks Data Intelligence Platform ensures that data is quality, which leads to quality models. After your data has been transformed and loaded into the lakehouse, you can begin to make models.

These models can often be refined and improved by making new features (another word for variables) based on processing existing features. For example, in finance you can make a new price-per-earning feature based on the ratio of price and earning. The resulting ratio can be even more predictive than the original price and earning features.

Developing your models

After you have your data, your model can be developed. MLflow is the leading open-source model development solution and comes with Databricks. You can choose what type of model is to be used with your specific data. Some models such as linear regression are better for predicting the optimal price of a product, for example, while other models such as logistic regression are better at predicting if a person should get a loan at a given rate or not.

TIP

Hyperparameters are instructions on how each model should be run. These are like dials you can turn and levers you can pull, in order to instruct the model to behave in a certain way. For example, should you classify customers into 8 segments or 20 segments?

After you select the type of model and its hyperparameters, MLflow runs the model and provides information, such as how accurate the model prediction is for the given data. Is it 80 percent accurate or 99 percent accurate?

Documenting experiment tracking

After you develop a model, you aren't done. You want to run many experiments, trying different features and hyperparameters and seeing which approach is most accurate. MLflow can provide different metrics on the accuracy of the model at the time it's created.

REMEMBER

When you like your model, register it into UC through the MLflow tracking server. The server stores information such as the date the model was created, its version number, the hyperparameters used, and the accuracy metric.

After many experiments, you can determine which challenger model outperforms the others and eventually replace the champion model. If you have sufficient permissions, you can make this comparison and let the best model win.

Streamlining with AutoML

Automation becomes vital for scaling, and automated machine learning (AutoML) can run through many model scenarios, trying different types of models and combinations of hyperparameters. Hundreds and thousands of combinations are tried — all automatically. In the end, there's a leaderboard where AutoML can decide which model is most accurate for the given data and objective. AutoML in Databricks saves data scientists from time-consuming and repetitive tasks, freeing them to perform more complex tasks. AutoML enables nontechnical business analysts to create models that used to be the sole domain of expert PhD programmers.

Clarifying model explainability and transparency

People in the real world shouldn't blindly implement a recommendation from a model without understanding what drove the model's decision. A business demands to have visibility into which features are most important to drive the outcomes. For example, what were the reasons for and against the recommendation that a given prospect should be given a bank loan at a specific interest rate?

Without this explainability, the decisions are like black boxes, and human trust in the model is understandably limited. This often prevents models from ever seeing real-life practical use.

REMEMBER

Databricks offers full transparency from the data to the end results of the model with lineage tracking. This sheds transparency on every step of the data journey.

Deploying models

After you have the model, a data scientist or ML engineer can easily use Databricks to deploy it into production. This is done by creating a model serving endpoint — something that doesn't require much experience anymore to achieve.

ML workflows help you move the model from experiment and development to staging and ultimately to production for use in the real world.

Observing model governance

Over time, a business grows from its first model in production to its second, to its 50th, and then before long, hundreds or thousands are in production. Being able to manage the life cycle of models is key, and Databricks makes this task simple.

Also, governing model life cycles through UC is critical. You need to make sure that only the right people can place models into production and that only the allowed users are able to access the data on which these models are built. Unity Catalog's audit logs and system tables show all these details, including who ran which model with what data and when.

Monitoring models and data drift

Your data is likely to change over time — interest rates fluctuate, shopping patterns change, and your customers' spend varies. This is known as *data drift,* and if data changes enough, it could make a model become less accurate or even lose all its value.

With the Databricks SQL dashboard, you can see the full health of each model, whether models are failing (for example, a data source stopped working), and whether a model needs to be refreshed. Metrics can be set during development with the option for custom metrics with lakehouse monitoring.

TIP

If the model becomes out of date, Databricks can send SQL alerts to notify data scientists when their models should be recalibrated, or the models can be set to automatically refresh without human intervention.

Developing Generative AI Applications

Generative AI is revolutionizing the way businesses in every industry develop new applications. This technology enables companies to speed up innovation, customize products, and tackle complex challenges. By adopting generative AI, businesses can shorten development times and enhance the scalability of their solutions. This makes it possible to deliver better services and products.

Crafting custom generative AI applications

Databricks' Mosaic AI is part of the Databricks ecosystem that lets you build generative AI applications from the ground up. You can start with raw data and develop AI models specifically designed to address the context of your business, and on your proprietary data, without leaking confidential information outside your data perimeter.

Some key features of Mosaic AI on the Databricks Data Intelligence Platform include

- **Customizing LLM training:** Mosaic AI enables LLMs to be customized by using an organization's proprietary data. This ensures that the knowledge of the model is closely aligned with your specific domain, providing more relevant and accurate outputs.

REMEMBER

LLMs use big data to understand and generate human-like text. They learn from vast amounts of information to perform language tasks, creating text based on the patterns they recognize in the data they've trained on.

- **Reducing training costs:** The platform offers an optimized training solution that significantly lowers the costs of training custom LLMs. This makes it feasible for more businesses to invest in custom AI solutions without compromising on the quality of the models.
- **Supporting comprehensive models:** After the models are trained, Mosaic AI provides a unified service for deploying, governing, and querying these AI models. This includes custom ML and foundation models, ensuring they're integrated seamlessly into business applications and workflows.
- **Enhancing data security and governance:** Mosaic AI ensures that all data and intellectual property remain within your organization's control, reducing data privacy and compliance risks. This is particularly important for enterprises that handle sensitive information, like medical or financial organizations. Furthermore, organizations gain strong access controls, end-to-end lineage, and auditing from data to production model.
- **Complete control:** Maintain ownership over both the models and the data. Databricks enables organizations to use their unique enterprise data to build generative AI solutions.
- **Supporting various GenAI architecture patterns:** Databricks supports multiple generative AI architectures — from prompt engineering, Retrieval Augmented Generation (RAG), fine-tuning, and pre-training custom LLMs. This flexibility allows businesses to choose the best approach for their specific use cases and develop as their requirements change.

Designing RAG applications

RAG applications combine LLMs with custom enterprise data to improve the accuracy and relevance of AI-generated responses.

It retrieves data relevant to a query and provides it as context to the LLM.

TIP

RAG has successfully supported chatbots and Q&A systems that need to maintain up-to-date information or access domain-specific knowledge. RAG provides a cost effective and efficient solution compared to other methods like fine-tuning an entire model to adapting language models to domain-specific applications. With RAG, organizations can utilize external data without modifying the underlying LLM model, which is especially beneficial when data needs to be frequently updated. RAG also ensures that the model's answers are based on current information instead of potentially outdated training data.

Fine-tuning existing models

If you already have open source LLM models, Databricks Mosaic AI enables you to fine-tune these models with your data. It adapts a pre-trained generative AI model to specific datasets or domains. This means you can improve models' performance by adjusting them to better reflect your datasets. Fine-tuning allows you to maintain control over your data and ensures privacy because the data never leaves your secure environment.

Building models from scratch

Databricks Mosaic AI allows you to build custom LLM models from scratch to create AI solutions that align with your data's unique characteristics. This process involves using your own datasets to train entire new models, ensuring that the resulting AI applications are integrated with your business processes and can provide insights.

Training LLMs is usually complex and difficult and requires extensive expertise. However, Mosaic AI Foundation Model Training lets anyone easily and efficiently train their own custom LLMs by simply pointing to their data sources. Foundation Model Training handles the rest: scaling to hundreds of GPUs, monitoring, and auto-recovery. Training multibillion parameter LLMs can be completed in days, not weeks.

As an example of what's possible, Databricks trained a state-of-the-art LLM, called DBRX, using Foundation Model Training and the power of Mosaic AI. DBRX was built with a Mixture of

Experts (MOE) architecture and at the time of announcement was a leading open source LLM for quality and price/performance. With all the techniques and optimizations in DBRX, and with Foundation Model Training, now every organization can build their own LLM, fully customized on their data, at a reasonable price. The result is a customized model trained on the organization's IP that is uniquely differentiated.

Putting It All Together

Using the Databricks Data Intelligence Platform greatly simplifies the development of enterprise AI applications. DatabricksIQ is integrated directly with the AI platform, Mosaic AI, to make it easy for enterprises to create AI applications that understand their data. To directly integrate enterprise data into AI systems, Mosaic AI offers multiple capabilities:

- End-to-end RAG to build high-quality conversational agents on your custom data
- Training custom models either from scratch on an organization's data or by continued pre-training of existing models, such as DBRX, MPT, and Llama 3, to further enhance AI applications with a deep understanding of a target domain
- Efficient and secure serverless inference on your enterprise data and connected into UC's governance and quality monitoring functionality
- End-to-end MLOps based on the popular MLflow open source project, with all produced models and data automatically actionable, tracked, and monitorable in the lakehouse

IN THIS CHAPTER

- Benefitting from a unified data platform
- Uncovering data insights
- Owning your own data and intellectual property
- Saving money

Chapter 5

Ten Reasons Why You Need a Data Intelligence Platform

Businesses collect huge amounts of data from various sources every day. However, having access to vast quantities of data isn't enough; you need a powerful tool to employ the full potential of your data assets. You need a data intelligence platform now because it

- **Has a unified data platform:** This central place serves as one location for all your data types. It enables consistent data management and eliminates data silos.
- **Enhances data and AI security:** With a data intelligence platform, you get enhanced data and artificial intelligence (AI) security. This platform provides robust security features, helping your organization protect sensitive data and meet compliance requirements.

  REMEMBER

  When your company works with AI and analytics, you must avoid leaking information.
- **Allows you to fully own your data and intellectual property (IP):** With the unification of the platform, you can

create or enhance applications, solutions, or analytics based on your own data. This provides you with the strongest competitive and economic advantages.

- **Improves your searches:** A data intelligence platform makes searching through your data assets easier and helps find a better context for your data's meaning. In this scenario, all data insights are more comprehensive and accessible to various users within your organization.
- **Finds more intelligent data-driven insights.** A data intelligence platform enables your organization to uncover insights and trends hidden in your data. AI drives better context-aware information, which is important because it ensures data quality by providing data cleansing, validation, and enrichment capabilities.
- **Accelerates your data tasks with automated and unified workflows.** You work within a single platform that speeds up tasks such as data engineering, data science, and machine learning — enabling seamless collaboration and efficient workflows. The data intelligence platform automates data pipelines and infrastructure management, which reduces manual effort, minimizes errors, and improves scalability.
- **Makes data more accessible to everyone in your organization.** Data intelligence makes it possible for nontechnical people to query their own data with natural language, without even knowing how to write software code. This opens the possibilities for everyone — business analysts, executives, line-of-business managers — to get more intelligent insights easier and faster than ever before.
- **Offers better collaboration:** The platform facilitates easy partnerships among teams and users, helping them share insights, code, and results. This teamwork accelerates data-driven decision-making, which greatly benefits your business.
- **Massively scales:** The platform handles large-scale data processing, allowing your organization to work more efficiently with its huge amount of data. You get structured, semi-structured, and unstructured data all in one platform.
- **Improves ROI:** Everyone loves cost savings, right? By consolidating data management and analytics tools into a single platform, your organization can reduce costs and simplify your data infrastructure.

Maximize your company's potential for data+AI

Maximize your organization's potential for data and AI with data intelligence. The Databricks Data Intelligence Platform is built on an open, unified, and governed lakehouse architecture and powered by a data intelligence engine. By using AI, the platform can reason on your own enterprise data, enabling you to harness the full value of your unique data assets. Whether it's ETL, data warehousing, BI, traditional or generative AI, data intelligence streamlines and accelerates your path to data-driven success.

Inside...

- The value of data intelligence
- The power and potential of AI
- Features of data intelligence platforms
- Building AI applications
- Why you need a data intelligence platform

databricks

Ari Kaplan is Databricks' Head of Technical Evangelism. He is Caltech's "Alumni of the Decade" and created the Chicago Cubs' & Baltimore Orioles' analytics departments. **Stephanie Diamond,** former AOL marketing director, is founder of Digital Media Works. She's authored dozens of marketing and custom e-books.

ISBN: 978-1-394-28407-8
Not For Resale

9 781394 284078

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT