Brought to you by:

databricks®

Unified Analytics

for dummies®

A Wiley Brand

Accelerate AI initiatives

Embrace Apache Spark™, TensorFlow®, and more

Unify data science and engineering

Databricks Special Edition

Ulrika Jägare

About Databricks

Databricks' mission is to accelerate innovation for its customers by unifying data science, engineering, and business. Databricks' founders started the Spark research project at UC Berkeley that later became Apache Spark. Databricks provides a Unified Analytics Platform powered by Apache Spark for data science teams to collaborate with data engineering and lines of business to build data products. Users achieve faster time-to-value with Databricks by creating analytic workflows that go from ETL and interactive exploration to production. The company also makes it easier for its users to focus on their data by providing a fully managed, scalable, and secure cloud infrastructure that reduces operational complexity and total cost of ownership. Databricks, venture-backed by Andreessen Horowitz, NEA, and Battery Ventures, among others, has a global customer base that includes Viacom, Shell, and HP. For more information, visit **www.databricks.com**.

Unified Analytics

Databricks Special Edition

by Ulrika Jägare

for dummies®
A Wiley Brand

Unified Analytics For Dummies®, Databricks Special Edition

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

ISBN: 978-1-119-54597-2 (pbk); ISBN: 978-1-119-54598-9 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

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

Project Editor: Carrie A. Burchfield
Editorial Manager: Rev Mengle
Acquisitions Editor: Katie Mohr
Business Development Representative: Karen Hattan
Production Editor: Siddique Shaik

Table of Contents

Introduction

The world has come a long way since the early days of data analysis where a simple relational database, point-in-time data, and some internal spreadsheet expertise helped to drive business decisions. Today, enterprises focus significant resources to tap into the enormous promise of artificial intelligence (AI) to drive disruptive innovation and transform their businesses. According to a recent survey and research report commissioned with IDG's CIO, nearly 90 percent of enterprises are investing in data and AI technology.

Data is at the core of how these modern enterprises are unlocking the potential of AI to change their business. The challenge most enterprises face is how to succeed with both their data and AI. How can the gap between managing the data (data engineering) and developing and managing the algorithms (data science) be bridged and what impact can that have on an organization's ability to succeed with AI? That is the major question addressed in this book.

About This Book

In this book, you not only discover how to avoid and overcome the most common challenges impacting AI success, but a new concept is also introduced. Unified Analytics is a concept that brings together solutions that unify data science and data engineering, making AI much more achievable for enterprise organizations and enabling them to accelerate their AI initiatives.

This short book is packed with useful information about how to accelerate AI initiatives. The first chapter outlines the major challenges enterprises face when investing in AI. The following three chapters focus on explaining how to tackle or avoid these challenges through the concept of Unified Analytics and give examples on how this has been addressed in different companies using a Unified Analytics approach. The last and final chapter offers information on what's available as well as some useful tips on how to get started.

Icons Used in This Book

I occasionally use special icons to focus attention on important items. Here's what you'll find:


REMEMBER

This icon with the proverbial string around the finger reminds you about information that's worth recalling.


TIP

Expect to find something useful or helpful by way of suggestions, advice, or observations here.


WARNING

Warning icons are meant to get your attention to steer you clear of potholes, money pits, and other hazards. Soft clouds can deliver hard knocks!


TECHNICAL STUFF

This icon may be taken in one of two ways: Techies will zero in on the juicy and significant details that follow; others will happily skip ahead to the next paragraph.

Beyond the Book

This book can help you explore general strategies for how to overcome the most common challenges of implementing AI and how to optimize your AI investments through a new concept called Unified Analytics. If you want more deep dives into resources beyond what's offered in this book, additional reading that's chock-full of useful info can be found at the following links:

- **Research report by CIO/IDG:** `https://databricks.com/CIO-survey-report`
- **Learn about the Databricks Unified Analytics Platform:** `https://databricks.com`
- **Free trial of Databricks Unified Analytics Platform:** `https://databricks.com/try-databricks`

IN THIS CHAPTER

- Understanding why teams working in silos are hindering AI success
- Identifying when data preparation becomes too complex
- Managing agility in AI ecosystems
- Avoiding infrastructure complexity

Chapter 1
Identifying Challenges That Impact AI Success

Despite the allure of artificial intelligence (AI), most enterprises are struggling to succeed with AI. Why is that? Well, there's not just one answer to that question. One major challenge is that data science and data engineering are siloed in different systems and different organizations. Enterprise data is itself siloed across hundreds of systems such as data warehouses, data lakes, databases, and file systems that aren't AI-enabled.

This means an enormous amount of time is spent on preparing the data for deeper analysis. It includes activities like cleaning data from errors, duplicates, and missing fields and combining different data types into new groups for added understanding. After those activities, other activities are needed, such as verifying that the data is actually correct and enriching the data with additional attributes, giving it a context. All these activities and many more are aimed at getting the data ready to be part of a model that can execute an analysis.

Teams Working in Silos

Companies usually have siloed teams for data engineering and data science. Where data engineers deal with large scale data preparation and deploying models into production, data scientists deal with AI — exploring data and building, training, and validating models. This organizational separation creates friction and slows projects down, which then becomes an impediment to the highly iterative nature of AI projects. In fact, 80 percent of organizations cite challenges with data engineering and data science collaboration due to siloed teams. So, in order to avoid a decrease in AI productivity due to siloed teams, make sure to be aware of the following indicators:

- Lack of an integrated environment for data engineers to iteratively create and provide high quality datasets to data scientists
- Lack of collaboration capabilities limiting knowledge sharing among data scientists, and the ability to iteratively explore data, train, and fine-tune models as a team
- Complex procedures when deploying models into production leading to multiple hands-off between data scientists, data engineers, and developers, slowing down processes and increasing risks to introduce errors

REMEMBER

The limited availability in the market of talent and skills in Data Science and Data Engineering to help solve AI makes it critical to improve internal efficiency, collaboration, knowledge sharing, and employee satisfaction in every company aiming to utilize the AI potential.

Preparing Data Is Too Complex

As you might understand, preparing data for AI is a major bottleneck. By the time the data has made it all the way to predictive models, data scientists often find out that the results aren't good enough and have to go back to square one.

Given the fundamental need for an iterative approach in AI development, companies need to cycle through the life cycle of preparing data, training the models, and deploying the model into

production quickly. However, because data systems in general aren't enabled for AI and AI technologies such as TensorFlow, PyTorch, and SciKit-Learn don't do data processing, it's very hard for enterprises to succeed in AI.

WARNING

In order to succeed, enterprises have to invest in an army of highly sophisticated data engineers and data scientists, which is both expensive and time consuming, and on top of that, talent could prove difficult to find and recruit due to the increasing demand on the market. Another challenge is that data engineers are expected to continuously equip the business with high-quality datasets in real time while keeping costs low, data secure, and managing complex data. It's no wonder 96 percent of companies cite data-related challenges as the number one obstacle to AI success.

REMEMBER

Empowered data engineers are successful data engineers. Make sure to invest time in defining an integrated and flexible data strategy to avoid

- A data structure where data is spread across multiple disparate systems across the organization, such as data warehouses, data lakes, databases, and file systems
- An increased complexity due to equal priority and combination of both streaming datasets (IoT, social) and historical datasets for real-time analytics purposes
- Too high demands on high-quality datasets, causing too much time being spent on data preparation tasks such as combining data, cleaning and verifying data, enriching data, labeling data, and so on

Lack of an Agile AI Ecosystem

Another major challenge to AI success is how to efficiently set up and maintain proper machine learning environments from an AI ecosystem perspective. There has been an explosion of machine learning frameworks and technologies that further impedes an organization's ability to unlock the promise of AI. In fact, organizations are using on average seven different tools within their AI technology stack. At the same time, this explosion of options is a good thing because data scientists should have the choice to use the right framework to solve the right problem.

A disjointed ecosystem lacks the ability to secure sufficient capacity and relevant data feeds for model training. That procedure requires multiple hand-offs, and productivity is limited without efficient integration between environments and the ability to leverage new specialized hardware for training purposes to accelerate development.

REMEMBER

Data scientists should be able to choose their favorite languages to visualize data and train models; this is the only way enterprises can truly solve the talent gap, which makes data scientists productive in their existing skills.

When the Infrastructure Gets in the Way

Implementing a secure but enabling infrastructure is key to succeed with your AI investments. However, when approaching AI infrastructure work, many companies tend to forget to define a clear and aligned data strategy first. The impact of lacking a unified data strategy, which is communicated across the company and reflected in how the infrastructure is approached, shouldn't be underestimated.

A poorly thought through infrastructure strategy will inevitably increase complexity in development and operations (DevOps) in terms of setting up and maintaining a big data infrastructure, managing upgrades and fixes, and scaling the infrastructure with growing data volumes as well as in providing high performance infrastructure for large teams of data scientists.

Furthermore, to increase processing efficiency, distributed computing is needed to ensure performance at scale. However, the complexity of managing distributed computing on CPU/GPU to reduce time needed to train sophisticated models is another challenge. It requires highly specialized skills, which could be difficult to get hold of, as well as it tends to slow down projects and their ability to get results faster.

WARNING

On-premises infrastructures often lack the ability to automatically scale up and down resources based on business needs. This means that the on-premises approach doesn't allow fast and flexible response to new and changing demands while maintaining costs low. There's also an increasing pressure on the business to keep data safe and secure, which many times tends to go overboard where an unnecessarily secure infrastructure tends to hinder productivity instead of enabling it.

IN THIS CHAPTER

- Analyzing the cornerstones of Unified Analytics
- Explaining the role of Apache Spark
- Understanding the fundamentals of Apache Spark

Chapter 2
Understanding Unified Analytics

Overcoming the inefficiencies of scattered IT environments where data is spread out over multiple systems and architectures and not prepared for artificial intelligence (AI) is a major challenge for most enterprises. The overall target is usually to accelerate company innovation in this field, but the truth is that few companies are prepared to accelerate when their starting point isn't solid enough. In this chapter, I introduce the concept of Unified Analytics, which aims to unify data science, engineering, and business perspectives. The idea with Unified Analytics is to address these challenges and make it easier for enterprises to build data pipelines across various siloed data storage systems to prepare labeled datasets for model building. This, in turn, allows organizations to explore data, train models, and deploy machine learning applications by leveraging massive datasets.

The Cornerstones of Unified Analytics

Unified Analytics is a concept that brings together data and AI under a singular environment and workflow, making AI much more achievable for enterprise organizations and enabling them to accelerate their AI initiatives.

Unified Analytics has four key parts:

- **Unifying all your siloed datasets:** This is mainly addressed with highly performant and reliable data pipelines. When using a Unified Analytics approach in your company, you can speed up the process to explore, prepare, and ingest massive datasets for best-in-class AI applications. You can simplify data management and easily connect data pipelines with machine learning (ML) to quickly fit the models to the data. The idea is to separate compute from storage for the best performance at lower costs.
- **Unifying data processing and AI technologies:** Setting up Unified Analytics enables companies to continuously train, track, and deploy AI models on big data faster, from experimentation to production. Ideally, prepackaged and ready-to-use ML frameworks should be available out of the box. Unified Analytics should also simplify model deployment and management to various platforms, so you can take advantage of hardware support for techniques such as deep learning.
- **Unifying data engineering and data science organizations:** A fundamental part of a Unified Analytics approach is to foster a collaborative environment for data scientists and data engineers to work effectively across the entire development-to-production life cycle. Providing a collaborative and interactive workspace makes it easier to automate and manage production pipelines, and build and train models, as well as visualize and share insights between key stakeholders.
- **An efficient infrastructure supporting DevOps:** In this context, DevOps refers to unifying software development (Dev) and software operation (Ops) in order to improve monitoring of all steps of software construction, from integration, testing, and releasing to deployment and infrastructure management. The role of Unified Analytics is

to help companies reduce infrastructure complexity, particularly when delivered as a fully managed cloud service, which offers reliability, scale, security, and cost efficiency.

TIP

It is probably safe to say that the companies and organizations that succeed in unifying their data engineering at scale and unifying that data with the best data science techniques will be the ones that succeed with AI going forward.

Figure 2-1 shows how AI can fundamentally change how enterprises run their businesses. Modern and data driven enterprises need to run fast, iterative experiments to test and refine learnings supported by a strong collaboration between data science, engineering, and the business. Machine learning requires collaborative experimentation on big data.

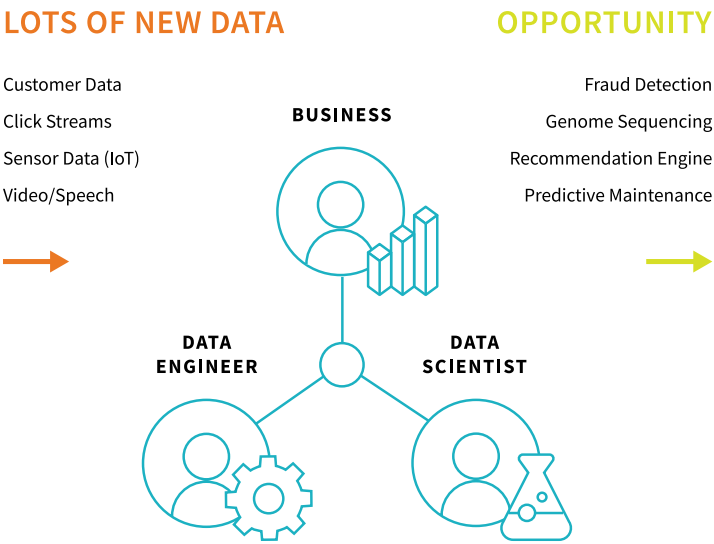

FIGURE 2-1: A Unified Analytics approach is the key to unlocking the potential of AI.

The Role of Apache Spark

Initially Apache Spark started as a research project at UC Berkeley, where students were working with large web companies building some of their first applications on big data. Initially they worked with earlier big data technologies, such as the other open-source

solution, Apache Hadoop. However, they soon came to realize that although there's a lot of potential in terms of the applications you can run with Apache Hadoop, the technologies were both difficult to use, often inefficient, and expensive to actually run at scale.

TIP

So over time as the big data use cases became more detailed, Apache Spark proved a simpler and more efficient programming model for writing these applications — so efficient in fact, that many times more people could begin to write applications on large clusters and reliably build them to power parts of their business.

The main difference between Apache Spark and other compute engines for big data is actually the focus on unification. Before Apache Spark really came to be, people used dozens of different distributed computing engines for big data, and they each specialized on a specific kind of workload, such as machine learning algorithms, graph processing, streaming analytics, or SQL queries. But unfortunately, even though each engine could do each workload, combining these into an end-to-end application was complex and error-prone.

REMEMBER

Many things can go wrong as you hook these up with each individual engine. And it's also fundamentally inefficient because you have different tools where you have to convert data between them.

As shown in Figure 2-2, the Apache Spark open-source Unified Analytics engine is built around speed, ease of use, and sophisticated analytics and powers a stack of libraries, including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these libraries seamlessly in the same application.

Today Apache Spark is the most actively developed open-source project in big data and probably the most widely used as well. Hundreds of people contribute to it every year — more than 1,000 since the project first started. Apache Spark continues to rapidly add new features and new capabilities such as libraries with algorithms to do different types of analytics and AI. Apache Spark has also set a few records in terms of performance.

From a Unified Analytics perspective, Apache Spark was the first engine to unify data engineering with data science. Over the last years, Apache Spark has become the de facto data processing and AI engine in enterprises due to its speed, ease of use, and sophisticated analytics.

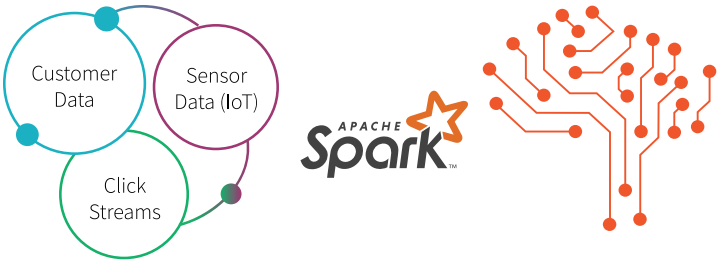

FIGURE 2-2: Apache Spark combines data and AI technologies in one platform.

The fundamental components of Apache Spark

In this section, you dive into the components in Apache Spark to truly understand why Apache Spark is so fundamental for the Unified Analytics concept. Apache Spark has five main components:

- **The core engine:** The core engine is called *Spark Core.* This is the lowest level of the engine. It handles efficiently scheduling work across the nodes of a cluster and distributes computation load to the nodes, manages data, and sends the results back. Everything else in Apache Spark is built on top of this.
- **SQL Analytics:** On top of the core engine, there are a few higher level libraries or components. The first is Spark SQL and DataFrames. This is a widely used library for working with structured data. It just makes it easy to do data transformations all the way from simple selections and filtering, to joins and aggregations and more complex SQL operations. And it also provides a DataFrame API that's easy to use for data scientists when scripting languages like Python and R. It kind of mirrors the tools that they're used to on a single node. This is an important capability to give data scientists quick access to Spark.
- **Streaming:** There is Structured Streaming, which is based on Spark SQL and also the older Spark Streaming, which is slightly lower-level. What these components do is let you do

real-time continuous processing on top of the core engine where you receive data continuously and update some state and produce new results. And it's pretty straightforward to take an Apache Spark program that works for you in a batch mode where it runs once and turn it into a streaming program and then just run that continuously to get new results.

- **Machine Learning Library (MLlib):** MLlib comes with a wide variety of scalable ML algorithms that can run over all the Apache Spark data sources and take advantage of a cluster. This also comes with an easy way of combining the algorithms into pipelines and trying different models and evaluating the results of different pipelines so you can select the best performing model for a task. This is widely used across the industry for doing machine learning at scale.
- **GraphX:** This component is for graph computation, which enables users to interactively build, transform, and reason about graph structured data at scale. It comes complete with a library of common algorithms.

REMEMBER

Although Apache Spark successfully addresses the two first cornerstones of a Unified Analytics approach, the siloed data challenge, and the need to merge data processing and AI technologies, enterprises need more than the Spark engine. They need an Apache Spark–based platform that also provides a collaborative environment for hundreds of internal users across organizational borders, simplifies and automates the management of clusters at scale, and delivers enterprise-grade data security and compliance.

Learning from others

Companies use Apache Spark in a variety of ways that combine data analytics and AI. Apache Spark is used by essentially all the top web companies to power key features of their products, including recommendations, fraud detection, spam filtering, and everything you see that personalizes the site for you as a user and makes the experience great. Beyond that, it's used in some of the largest enterprises, including financial institutions; tech companies; biotech companies; scientific groups, such as NASA; and essentially any organization that works with large datasets.

IN THIS CHAPTER

- Understanding key benefits of cross-team collaboration
- Exploring ways to accelerate innovation
- Managing the life cycle of machine learning
- Learning from others

Chapter 3
Unifying Data Science and Engineering Organizations

Apache Spark is a powerful Unified Analytics engine, but more is needed to unify the data science and engineering organizations. As the amount of data in an organization grows, more and more engineers, analysts, and data scientists need to analyze this data. Today, IT teams constantly struggle to find a way to allocate big data infrastructure budgets among different users and optimize performance. End-users like data scientists and analysts also spend enormous amounts of time tuning their big data infrastructure, which isn't their core expertise. This chapter presents practical examples on how unification of data science and data engineering helps solve these problems and accelerate innovation.

Cross-Team Collaborative Workspace Accelerates Productivity

The importance of an efficient cross-team collaborative workspace shouldn't be underestimated. Such a collaborative workspace should be able to handle all analytical processes end-to-end

across different systems and organizations and ensure that productivity isn't lost along the way.

As you might understand, this isn't an easy task. But with a unified workspace approach you minimize the inevitable hand-offs between data engineers and data scientists, creating a seamless workflow from data ingest to deployment of models into production. They can also share insights with the business, all from the same environment. Figure 3-1 gives you the elements of a well-functioning collaborative workspace.

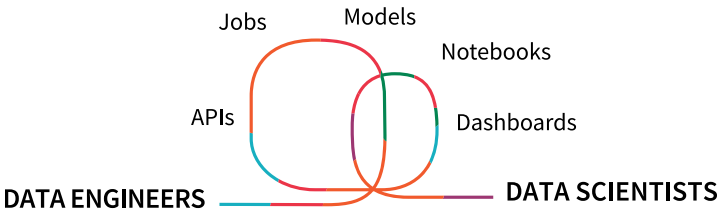

FIGURE 3-1: The ingredients of a well-functioning collaborative workspace.

TIP

The unified workspace should also have sufficient ecosystem support for the most popular languages and tools, such as R, Python, Scala, Java, and SQL, as well as integrations with RStudio, DataRobot, Alteryx, Tableau, Power BI, and more, which allows practitioners to use their preferred toolkit. A good cross-team collaborative workspace should also foster teamwork between data engineers and data scientists via interactive notebooks, APIs, or their favorite Integrated Development Environments (IDEs), all backed with version control and change management support.

Regarding data management, practitioners must be able to access all needed data in one place and automate the most complex data pipelines with job scheduling, monitoring, and workflows as notebooks or APIs. That access gives the teams full flexibility to run and maintain data pipelines and machine learning (ML) at scale.

Managing the ML Life Cycle

Reproducibility, good management, and tracking experiments are necessary for making it easy to test others' work and perform analysis. ML isn't easy, but creating a good workflow that you can

reproduce, revisit, and deploy to production is even harder. It is about how to record and query experiments, as well as packaging ML models so they can be reproducible and ran on any platform. ***Note:*** The ML life cycle is *not* the data science life cycle, which is more complex and has many more parts. In principle, there are three main areas addressed as part of the ML life cycle: ML frameworks, ML life cycle management, and Distributed Computing.

TECHNICAL STUFF

ML frameworks are used in order to quickly set up a productive ML environment. By having preconfigured and optimized ML environments, packaged with the most popular ML libraries and frameworks, such as TensorFlow, Horovod, Keras, XGBoost, MLlib, GraphX, and sparklyr, you can significantly speed up the process.

REMEMBER

However, it is equally important that the environment simplifies activities for the data science teams during the entire ML life cycle, not just to speed up the initial phase. The environment must enable the teams to iteratively run, track, share, and reuse ML models and quickly operationalize models in production on any platform throughout the life cycle of the algorithm. This includes reuse of models locally or at scale in the cloud, as well as good iterative support for moving them between different stages of their life cycle, from experimentation to production.

Finally, it is vital to address how to secure sufficient processing power for the different activities during the ML algorithm life cycle. By using Distributed Computing, you get the ability to scale advanced analytics at lower cost. This is done by decoupling compute and storage to reduce data movements, and significantly simplify parallelization and distributed computing on both CPU and Accelerated Hardware (Cuda GPU) via built-in optimizations (HorovodEstimator).

REMEMBER

Key features in a well-managed ML life cycle environment include the following:

- Prepackaged and ready-to-use ML frameworks out of the box
- Reproduced runs using any ML library or framework, locally or in the cloud
- Simplified management and keeping track of experiments and results

- Streamlined model deployment and management to various platforms
- Making distributed deep learning simpler with built-in optimizations
- Getting results faster with accelerated hardware support for deep learning

Learning from Others

Enterprises that want to be at the forefront of ML and want to leverage artificial intelligence (AI) should invest in getting a solid data strategy in place. The combination of data with ML is key. Organizations have to unify these two, which means that they must be prepared to adjust their organizational structures. If you separate the data engineering and data science teams and get silos, progress will be difficult.


TIP

Companies that succeed in AI have a Chief Data Officer (CDO) who's building unified teams with well-established data management and governance, together with data science. With that approach, you can move past the politics and siloed data that are separating teams in different departments.


REMEMBER

How you organize your business is important if you want to be successful with your AI investment. When it comes down to the core, it's not just about the technology; it's about the people.

In this section, I give you two examples of companies that have implemented a Unified Analytics approach supported by the Databricks Unified Analytics Platform, which I further explain in Chapter 4.

Retail

A leading Internet retail company sold furniture and home décor. Its main use case was creating a personalized experience based on users' propensity to purchase to drive conversion and customer lifetime value.

Facing the challenges

The company faced many challenges:

- Huge amount of new customers with more than 5 billion unique pay views per year
- Low conversion and low web activity but many visitors hitting one page and leaving
- ETL using an enormous amount of resources that were already scarce and stretched to their limits
- Data science teams spending too much time on DevOps and not iterating on models
- Differing programming language experience among team members, making it difficult to share data and collaborate on insights

Finding a solution

The retailer needed to find a solution to close the gap between development and production of new features. Databricks offered that solution by doing the following:

- Simplifying ETL processes and increasing runtime performance
- Supporting elastic compute scalability
- Improving data science collaboration
- Introducing serverless for efficient resource usage
- Offering more efficient training
- Establishing model iteration

Achieving results

After implementing the solution from Databricks, the company achieved the following results:

- 50 percent lowered cost of moving models to production
- Five times faster deployment of new models to production
- More efficient use of cloud resources at any scale by going serverless
- Improved productivity with collaborative and interactive workspace

Media

One of the world's largest broadcasting and media companies wanted to drive customer retention by improving streaming video quality and identifying ways to increase viewer engagement. The overall objective was to maximize customer retention and loyalty.

Facing the challenges

One challenge for the company was that videos were failing to load or were constantly rebuffering, impacting customer user experience. The company wanted to grow its audience, and in order to do that, it needed to mine massive amounts of data for new insights. Finally, the targeted advertising was failing and ad sales were falling due to poor audience engagement.

Finding a solution

The solution that the company decided to go for was to establish real-time analytics leveraging machine learning on Databricks' Unified Analytics Platform (see Chapter 4 for more details). The targeted solution enabled the company to accomplish the following:

- Constantly monitor the quality of video feeds and reallocate resources in real time when needed to ensure that performance issues didn't impact the viewer
- Leverage predictive analytics techniques to extract insights from customer viewing trends to inform smarter customer marketing decisions

Achieving results

The implementation was successful and resulted in the following:

- Video start delay was reduced by 33 percent by predicting trends and streaming issues to provide superior viewing experience.
- Customer retention levels increased up to seven times by leveraging viewership data to understand the best strategies to drive engagement.
- Conversations on ads were increased by targeting customers with personalized ads based on comScore ratings and viewing behavior.

IN THIS CHAPTER

- Introducing the Unified Analytics Platform
- Considering the implementation
- Learning from others

Chapter 4
Accelerating Innovation with the Databricks Unified Analytics Platform

REMEMBER

Succeeding with your artificial intelligence (AI) investment depends heavily on your ability to bridge the gap between the data engineers and the data scientists. However, understanding that this is important and getting advice on what you need to do is one thing. To actually deploy a productive environment and get it up and running is another. This chapter focuses on practical infrastructure recommendations and presents the Unified Analytics Platform from Databricks, built on Spark, as a candidate to achieve this.

Unified Analytics Platform

The Unified Analytics Platform from Databricks was developed by the original creators of Apache Spark. The platform is designed with one idea in mind: to accelerate innovation by unifying data

science, engineering, and business. With the Unified Analytics Platform, Databricks address all the challenges around the divide between data and AI — even the ones that Apache Spark doesn't cover, such as unifying data engineering and data science organizations as well as an efficient infrastructure supporting DevOps.

The Unified Analytics Platform improves collaboration across the ML life cycle, speeding up the process to prepare quality data at scale essential for best-in-class AI applications, and allowing teams to iteratively train and deploy cutting-edge AI models on big data faster, while keeping data safe and reducing DevOps complexity.

As you can see in Figure 4-1, the Unified Analytics Platform is made up of three main parts: the collaborative workspace, the runtime environment, and the cloud service.

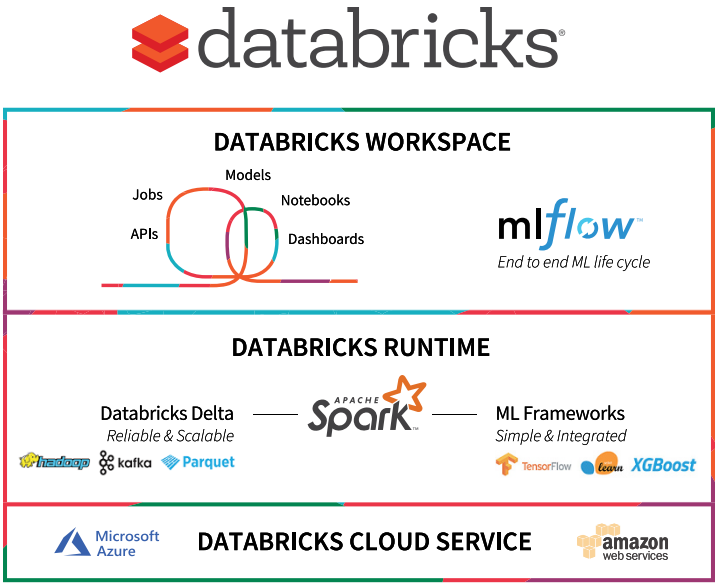

FIGURE 4-1: An overview of the Unified Analytics Platform from Databricks.

The collaborative workspace

Databricks Workspace empowers data scientists and engineers to streamline collaboration across the ML life cycle — from ETL to models training and deployment — leveraging familiar tools and skills, via interactive notebooks or APIs. The end-to-end life cycle

capability is called ML Flow, which is an open-source platform to manage the machine learning life cycle, including experimentation, reproducibility, and deployment. It currently offers three components to log and query experiments, package machine learning code for a reproducible way, and send models to a variety of downstream deployment tools.

REMEMBER

When it comes to data management, the Unified Analytics Platform lets you access all your data in one place and automates the most complex data pipelines. It supports jobs scheduling, monitoring, and workflows as notebooks or APIs, giving teams full flexibility to run and maintain data pipelines and ML at scale.

TECHNICAL STUFF

The Databricks Workspace has extensive ecosystem support. Databricks supports SQL, R, Python, Java, and Scala and provides native integration with popular IDEs like RStudio or BI tools with ODBC/JDBC connections, allowing data engineers and data scientists to use familiar languages and tools within the environment.

The runtime environment

Databricks runtime environment is powered by Apache Spark that unifies batch and streaming pipelines to accelerate the time to insight for reliable and current data. The process is to prepare clean data at massive scale, and continuously train and deploy state-of-the-art ML models for best-in-class AI applications.

Apache Spark could be seen as the Unified Analytics main engine, however; the runtime environment is also equipped with a data management system called Databricks Delta. The added capability with Delta is to ensure data reliability of the pipelines by enforcing transactionality.

TECHNICAL STUFF

The Databricks Runtime with Delta is 100 times faster than Apache Spark, and makes advanced analytics teams more effective. Delta enables fast queries at massive scale with I/O optimizations, indexing, and caching, and allows teams to manage continuously changing data reliably, with transactional guarantees and automatic schema validation for real-time analysis. Delta also provides massive scale and cost efficiency by leveraging cloud blob storage while ensuring data portability with open file formats.

The cloud-native service

The third and last part of the Unified Analytics Platform is Databricks' elastic cloud service. This is designed to reduce operational

complexity while ensuring reliability and cost efficiency at scale, with a unified security model featuring fine-grained controls, data encryption, identity management, rigorous auditing, and support for compliance standards.

Implementation Considerations

One of the most difficult aspects to tackle when investing in AI is how to manage that everything is moving and evolving so fast in this area. What is a safe bet, when everything is constantly changing? Everything from the amount of data that can be processed, to the variety of software, methodologies, and technologies available is enhancing fast, and to a great extent driven open source wise. Because of this, it is of growing importance to consider the context of your AI solution. For example, how will the data flow to and from your systems and other external systems? Your AI implementation will inevitably have to execute in a rapidly changing ecosystem of data and AI solutions. Some aspects on what to consider in your AI investment are addressed in this section.

An efficient and reliable data pipeline

It all starts with the data, so investing in a highly performant and reliable data pipeline is key. Make sure that the data flow is continuous and correct. However, there are as many ways of solving this puzzle as there are implementations in the market. From a data pipeline perspective, you should consider three main aspects:

- **Data reliability:** You must be able to trust your data. Make sure the infrastructure can handle automatic schema validation and optimization as data flows from various storage sources into the processing engine.
- **Query performance:** You need a solution that's fast at data processing, including scale indexing and caching, as well as the ability to manage query executions at scale.
- **Simplified architecture:** Don't make things too complicated. If possible, select an infrastructure setup that supports you to unify batch and streaming functionality, which decreases the complexity of big data processing by unifying real-time and batch analytics at massive scale and across various sources and data types. Finally, don't forget to prioritize access to early data availability for analytics.

Simplified infrastructure in the cloud


TIP

Part of the important investment to consider for your AI environment is the decision to invest in a cloud-based environment. Taking the step from an on-premises solution to an environment in the cloud comes with many benefits:

- A cloud-based environment makes it easy to scale analytics to petabytes of data in a fast, reliable, and cost-efficient manner. It also makes it possible to share resources across users and automatically scale resources for best-in-class performance at lower costs. Evaluate analytics platforms that are cloud-based because they're easier and cheaper to get up and running.
- A cloud environment simplifies infrastructure management by abstracting the complexity of the data infrastructure and DevOps. Instead, you gain elasticity, reliability, and better performance.
- A cloud vendor should be able to promise 99.9 percent availability via Service Level Agreements (SLAs). The provider should provide auto-configuration, auto-scalability, and high-reliability for its preconfigured clusters. This allows teams to focus on innovation rather than management of data systems.
- Security and compliance aspects should be guaranteed. The cloud provider shall promise enterprise-grade security with encryption, auditing, role-based control, and compliance to important international laws and regulations.


TIP

Consider the cloud provider's global presence of data centers. You'll have the possibility to compute data locally without moving raw data out of the country it has been generated within. Laws and regulations could be very strict in terms of how you're allowed to move data across country borders.

Making security and flexibility go hand-in-hand

Data is key to making machine learning (ML) work, and enterprises need to utilize a lot of different data sources. Data already available to enterprises often needs to be enriched with data from other sources, meaning that the different data types need to be

combined. In reality, this often poses a major problem because these data assets are often located in different places like different internal systems, data warehouses, and data lakes.

The main aspect to consider is how to combine the data assets in a flexible way while still leveraging the value through ML and staying secure. From that perspective a cloud environment is key.


TIP

Being security-first aware is all about making sure that everything is secure directly from the beginning. Make sure that everything comes with role-based access control. Through and through it should be built for enterprises where different stakeholders and different departments are able to access and combine datasets in a secure way.


WARNING

If you can't get the security right in your infrastructure, it could be very costly for the company. The regulations around data are just growing in importance and so is user and society awareness. Don't risk your business or customer trust by not investing enough time and effort in a secure solution. However, don't forget to balance your investments. If you invest too much control into your infrastructure, it might tip over and become so rigid that it prevents agile execution and critical cross-team collaboration.

Learning from Others

There are many pitfalls to run into during your AI implementation. This section presents two different but very relevant use cases from the travel industry and the oil and gas sector. It will give you the opportunity to learn more about challenges other enterprises are facing, and also how Databricks has supported their journey using the Unified Analytics Platform and what that has resulted in.

Travel and hospitality

A leading global travel technology company wanted to increase bookings by using ML to deliver a personalized shopping experience, with the overall objective to improve the hotel booking experience.

Facing the challenges

The main challenge was the requirement of massive compute and analytics capabilities to ensure a targeted and satisfying customer experience when booking travel. Other challenges included the following:

- **The data pipeline was slow and unstable.** The on-premises Hadoop cluster was used to do data science at scale, taking two hours to process the data pipeline on only 10 percent of the data.
- **Data wasn't structured and organized properly.** Massive volume of image files corresponding to each property listing lacked organization for ranking and classification, making it impossible to leverage machine learning to drive consumer experience.
- **No real visibility into consumer trends in real time.** This was impacting the ability to understand customer trends in real time to develop strategies to drive conversion and lifetime value.

Finding a solution

Databricks managed to help the travel company realize its goal of becoming "data science focused" so that it now can anticipate customer behavior and provide a more optimized user experience through machine learning.

Achieving results

The client achieved the following results:

- **Accelerated ETL at scale:** Ability to increase the volume of data processed by 20 times without impacting performance
- **Optimized user experience:** Highly accurate and effective display of images within the context of property searches by customers
- **Increased sales efficiency:** Providing the right hotel with the right images based on searches resulted in higher number of customer conversions

Oil and gas

The world's largest energy and petrochemical company's overall target was to improve management efficiencies of global inventory of drilling machinery parts.

Facing the challenges

The company faced many challenges:

- **Disjointed inventory distribution:** Stocking practices were often driven by a combination of vendor recommendations, prior operational experience, and "gut feeling."
- **Limited DSS (Decision Support System) data availability:** There was limited focus directed toward incorporating historical data and doing advanced analysis to come up with decisions.
- **Lost business agility:** This was causing excessive or insufficient stock being held at the company's different locations, and with oil rigs, that has significant business implications.

Finding a solution

The solution setup by Databricks included the ability to run distributed data processing at scale and train and deploy models for predictive allocation and planning of inventory.

Achieving results

The main result achieved was improved operational efficiency, which is saving the company millions of dollars per year. This was achieved through reduced runtime of the inventory analysis from 48 hours to less than 45 minutes. The company has seen an improvement in time-to-value by 32 times!

IN THIS CHAPTER

- Enjoying a collaborative workspace
- Getting faster processing time of big data
- Sharing resources across users
- Scaling advanced analytics at lower cost

Chapter 5
Ten Ways a Unified Analytics Platform Can Help You

Each *For Dummies* book ends with a Part of Tens chapter. This book is no different. Here, I give you ten ways a Unified Analytics Platform can help you.

A Unified Analytics Platform can

- **Unify your big data processing** for both batch and real-time data analytics from various data sources. Whether it is on-premises or in the cloud, you can manage ETL on structured and unstructured data with speed and scale.
- **Ensure data quality and data lineage over time** by keeping your data pristine with transactional guarantees and automatic schema validation as data is ingested.
- **Manage end-to-end data pipelines,** from data ingest, through ETL, to storage with code or clicks directly in notebooks or APIs. The extensibility of the platform lets you use familiar languages and tools to connect to data.

- **Enjoy a collaborative workspace** that accelerates teamwork and fosters collaboration between data engineers and data scientists through shared and interactive notebooks.
- **Support Agile development** with ready-to-use machine learning (ML) frameworks that allow ML practitioners to get started fast. Get easy access to on-demand and preconfigured ML clusters and accelerated hardware support, as well as a wide range of low-level to high-level APIs for all your ML needs.
- **Utilize distributed training** with easy-to-scale-out computation on Spark Dataframes with deep and native integration with Apache Spark. Simplify distributed training for the most demanding applications as well as customized optimizations.
- **Benefit from complete life cycle support** with a unified experience to help collaboratively build models with shared notebooks, track and share experiments across frameworks, and quickly deploy dashboards or models for production use.
- **Simplify your infrastructure** with auto-configuration and management of clusters to reduce DevOps and infrastructure complexity, giving you full flexibility with cross-cloud support on both CPUs and GPUs.
- **Manage scalability in your cloud service** with auto-scale/termination of clusters based on your needs, and decoupling compute from storage resources, which allows you to get the performance needed when you need it, while keeping costs under control.
- **Keep your data safe, secure, and compliant** with enterprise-grade security and common regulatory standards such as the Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR).

Accelerate innovation

Today, enterprises direct significant resources to tap into the promise of artificial intelligence (AI) to transform their businesses. You can manage the gap between data engineering and data science. Unified Analytics brings together solutions that unify data science and data engineering, making AI much more achievable for your enterprise organization and helping you accelerate your AI initiatives. Get started reading to find out how!

Inside...

- Challenges impacting AI success
- Overcome data and ML technology silos
- Best practices for team collaboration
- Learn how Databricks accelerates AI
- Ten ways Unified Analytics can help

databricks®

Ulrika Jägare is an M.Sc. Director at Ericsson AB with over 18 years of experience in telecommunications, including ten years in analytics and machine intelligence. She's held various leadership positions, mostly in R&D and product management. Ulrika is also an appreciated conference speaker.

Go to Dummies.com®
for videos, step-by-step photos, how-to articles, or to shop!

ISBN: 978-1-119-54597-2
Not For Resale

for dummies®
A Wiley Brand

Also available as an e-book

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.