LEARNING MADE EASY

Databricks Special Edition

Migrating from Hadoop to Data Lakehouse

for dummies®
A Wiley Brand

Evaluate your data and AI strategy

Prepare your organization

Migrate to the lakehouse

Brought to you by

databricks

Stephanie Diamond

About Databricks

Databricks is the data and AI company. Thousands of organizations worldwide — including Comcast, Condé Nast, Nationwide, and H&M — rely on Databricks' open and unified platform for data engineering, machine learning, and analytics. Databricks is venture-backed and headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark, Delta Lake, and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on social media:


twitter.com/databricks


www.linkedin.com/company/databricks


www.facebook.com/databricksinc

Migrating from Hadoop to Data Lakehouse

Databricks Special Edition

by Stephanie Diamond

for dummies®
A Wiley Brand

Migrating from Hadoop to Data Lakehouse For Dummies®, Databricks Special Edition

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

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: 978-1-119-89420-9 (pbk); ISBN: 978-1-119-89421-6 (ebk). 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:
Carrie Burchfield-Leighton

Sr. Managing Editor: Rev Mengle

Managing Editor: Camille Graves

Acquisitions Editor: Ashley Coffey

Sr. Client Account Manager:
Matt Cox

Table of Contents

Introduction

The cost, complexity, and viability of existing Hadoop platforms have failed to deliver on business value due to the lack of data science capabilities, the high cost of operations, the inability to scale, the lack of agility, and poor performance. As a result, enterprises are looking to migrate their existing Hadoop platforms to the data lakehouse.

The *data lakehouse* is a cloud-native platform for data management that provides a powerful engine for data processing and simple tools for developers, analysts, data scientists, and business users in an intuitive user interface (UI). It enables you to build, deploy, scale quickly, and manage analytical applications in minutes instead of hours or days. The data lakehouse is an open data architecture that combines the best of data warehouses and data lakes on one platform.

About This Book

Welcome to *Migrating from Hadoop to Data Lakehouse For Dummies,* Databricks Special Edition. This book looks at why and how to migrate from Hadoop to the lakehouse to prepare your organization to meet the future. It also uncovers what your leaders need to know to meet new challenges in data management. This book covers several topics:

- The history of data lakes
- Determining where your company resides on the data and artificial intelligence (AI) maturity curve
- Reviewing what a data lakehouse is
- Accruing benefits of migrating from Hadoop to the data lakehouse
- What to consider when you migrate to the lakehouse
- Ten steps to enable and scale your data and AI strategy

Icons Used in This Book

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

The Tip icon adds information to help you manage processes faster and easier.

The Remember icon points out content to remember when searching your memory bank.

The Warning icon alerts you to information that you should be aware of that can be harmful to you or your company.

Sometimes I give you a few tidbits of research or facts beyond the basics. If you want to know the technical details, watch out for this icon.

Beyond the Book

This book can help you discover more about migrating from Hadoop to the data lakehouse, but if you want resources beyond what this book offers, check out the following links:

- `databricks.com/discoverlakehouse`: See why your legacy data-warehouse can't support the advanced needs you have today.
- `databricks.com/product/data-lakehouse`: See how the lakehouse platform can support all your data, analytics, and AI use cases on a simple, open, multicloud platform.
- `databricks.com/product/Databricks-sql`: Discover how Databricks SQL allows you to SQL workloads on the lakehouse architecture.
- `databricks.com/try`: Try out the Databricks Lakehouse.
- `databricks.com/p/ebook/migration-guide-hadoop-to-databricks`: A step-by-step technical guide on how to migrate from Hadoop to Databricks.

IN THIS CHAPTER

- Looking at the history of data lakes
- Migrating away from Hadoop
- Dealing with changing regulations

Chapter 1
Understanding the Shortfalls of Legacy Data Lakes

Traditional data lakes such as Hadoop (and even their cloud variants) are coming to the end of their life cycles. As a result, data lakes are primed for the next phase in their evolution. Businesses have realized that to use their data and analytics as the premier assets they are, migrating to the data lakehouse is the next logical step from the Hadoop data lake. This chapter looks at the evolution of the data lake (including Hadoop) and how it's poised to play a role in the data lakehouse architecture.

Reviewing the History of Data Lakes

Before the creation of data lakes, companies relied on enterprise data warehouses (EDWs) to store data. This storage worked for a time. However, as the volume and types of data grew exponentially, aggregating databases in EDWs produced a collection of multiple disconnected databases for different business units and purposes.

WARNING

The main drawback of this approach was that they couldn't be queried across business unit boundaries to deliver enterprise-level actionable insights. As a result, the data lake was created around the year 2011 to address this problem.

REMEMBER

The *data lake* is a repository for raw unstructured data and is usually a collection of stored files created for various purposes. Apache Hadoop was deployed to lower costs and scale out the resources using commodity hardware, albeit using tightly coupled storage and compute resources. Now Apache Spark runs data lakes in the cloud. Data lakes were cheap, and they could store all kinds of data based on open-standard formats. That meant that no bottlenecks existed between the data lake and its external sources.

The benefits

EDWs were formalized data models that aggregated the data at an enterprise level for certain reports and dashboards. They usually didn't have the raw, granular data that business teams needed for self-service analytics and exploratory and advanced machine learning (ML)/artificial intelligence (AI) needs. They also didn't have the capacity to scale in storage and compute to house all the data — structured and unstructured — for the whole enterprise. As a result, the data lake was created.

TIP

The immediate benefits of data lakes included the fact that

- All data was up to date, and it was easy to add new sources of data.
- You didn't need to maintain multiple copies of the database.
- Large-scale data cleansing and transformations were possible.
- You could run ad hoc queries against the entire data set.
- You could easily extract data from the data lakes and send it to other locations.
- It supported open-source ML libraries.

The challenges

Inevitably, data lakes also had some challenges:

- No support for atomicity, consistency, isolation, and durability (ACID) transactions

- No enforcement of data quality or governance
- Failed jobs and missed data
- Poor business intelligence (BI) support
- Poor performance


WARNING

Another problem with data lakes was that they often became a dumping ground for any available data. As a result of significant shortcomings such as inadequate data governance and the lack of attention to data quality, data lakes were sometimes referred to as *data swamps.* Read more about how data lakes evolved to support the new data lakehouse in Chapter 3.

Recognizing the Limitations of Hadoop


WARNING

Hadoop solved a data problem. However, as time went on and organizations faced new challenges, several limitations of Hadoop came to the fore. They included

- **Wasted hardware capacity:** Over-capacity is a given in on-premises implementations so you can scale up to your peak time needs, but the result is that much of that capacity sits idle but continues to add to the operational and maintenance costs.
- **DevOps burden:** You can assume roughly four to eight full-time employees (FTEs) for every 100 nodes. (This is based on the experience of Databricks' customers.)
- **Increased power costs:** Expect to pay as much as $800 per server annually based on consumption and cooling. That's $80,000 per year for a 100-node Hadoop cluster.
- **Software version upgrades:** These upgrades are often mandated to ensure that the support contract is retained. Those projects take months at a time, deliver little new functionality, and take up precious bandwidth.

The Divided Worlds of Data

Cloud data lakes, by design, require customers to use EDWs for data analytics and BI. Data lakes live in parallel to EDWs but are ultimately two competing systems that require different storage,

management, and governance of data. Data handling is the key to why the lakehouse is so valuable. The value of using the lakehouse architecture is that you can

- Use data lakes as a landing zone for all your data for SQL, BI, data science, and ML use cases.
- Easily meet regulatory requirements to mask data that contains private information before it enters your data lake.
- Secure your data lake with role-and-review-based access controls.
- Catalog the data in your data lake.
- Get the same price performance of an EDW with the scale of a data lake.

Keeping Up with Changing Regulatory Requirements

One of the crucial things that organizations must do is ensure that they comply with changing regulations. Regulatory change has increased 500 percent since the 2008 global financial crisis, and it's boosted regulatory costs. Two key regulations include the General Data Protection Regulations (GDPR) and the California Consumer Privacy Act (CCPA).


TECHNICAL STUFF

Delta Lake on Databricks manages GDPR and CCPA compliance for your data lake. Delta Lake adds a transactional layer that provides structured data management on top of your data lake. As a result, it can dramatically simplify and speed up your ability to locate and remove personal information (also known as *personal data*) in response to consumer GDPR or CCPA requests. For more information on Delta Lake, check out these resources:

- `databricks.com/product/delta-lake-on-databricks`
- `docs.databricks.com/security/privacy/gdpr-delta.html`

IN THIS CHAPTER

- Looking at top strategic goals
- Developing a data culture
- Using the Vs of data
- Increasing your business impact
- Looking at tech options
- Getting useful insights

Chapter 2
Prioritizing Your Data and AI Strategy

In today's competitive environment, it's not enough to have the right architecture to support your organization's data. You also need a comprehensive strategy that serves all the essential components of your organization. This strategy should include leveraging people, business goals, and technology. It's the key to long-term business success. Ultimately, the technology should be an enabler of the strategy and not the other way around.

Examining Top Strategic Goals

Data and technology executives who want to nail their data and artificial intelligence (AI) strategy will prioritize three goals:

- Moving forward on the AI maturity curve
- Reducing risks
- Controlling costs

You look at each goal in this section.

Moving forward on the AI maturity curve

To meet the future and to move along a defined maturity curve, companies need to adopt a data asset and AI mindset. The goal is to go from being descriptive to prescriptive. To this end, Databricks created a maturity model that your organization can use to understand the current state of your journey to data and AI maturity. The model is as follows:

- **Explore:** At this stage, your organization is beginning to explore Big Data and AI and understand the possibilities and potential of a few starter projects and experiments.
- **Experiment:** Organizations at this stage are building the basic capabilities and foundations to begin to explore a more expansive data and AI strategy, but they're lacking vision, long-term objectives, or leadership buy-in.
- **Formalize:** At this stage, data and AI are budding into a driver of value for business users aligned to specific projects and initiatives, as the core tenets of data and AI are integrated into corporate strategy.
- **Optimize:** Data and AI are core drivers of value across the organization — structured and central to the corporate strategy with a scalable architecture that meets business needs and buy-in from across the organization.
- **Transform:** At this stage, data and AI are at the heart of the corporate strategy and an invaluable differentiator and driver of competitive advantage.

Do you recognize your organization at one of these stages? If so, do you know how to move forward to achieve a higher maturity? You can schedule a custom business value assessment at `databricks.com/p/business-value-assessment-databricks`.

Reducing risks

Another strategic goal for leaders of organizations is to reduce several potential risks such as weak data management, failed IT projects, missing out on innovation due to the lack of advanced analytics platforms, and the ever-present threat of cyberattacks. These threats make it imperative to have a consistent way

to store, process, manage, and secure data. However, this goal is made more complex by the following needs:

- Adhering to the evolving privacy regulations landscape, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA)
- Adjusting data management to contend with new privacy directives such as those from Google and Apple
- Figuring out how to take advantage of new data sources that can supersede typical behavioral, demographic, and engagement data

Controlling costs

Leaders must always contend with the need to control costs. Data lakes can get expensive fast as the amount of data it manages grows. On top of it, there are overheads from data center equipment, database administration operations and maintenance, and many locked-in vendor agreements.

The solution to this problem (accommodating current data and AI initiatives) is implementing a cloud architecture that's elastic and flexible to adapt to the changing business needs and has good price performance in the cloud as data size increases.

REMEMBER

Utilizing simpler architectures also produces more agility for data and technology executives to integrate and have actionable insights without delays or IT intervention.

Focusing on Creating a Data Culture

Taking advantage of the value of data to organizations has created the need for companies to establish and maintain a data culture. A robust data culture ensures that your organization will meet the future. You can spot a company that has a strong data culture in two key ways:

- You see the entire organization making daily informed business decisions by using available, relevant data.
- Data is given greater weight than experience, intuition, or tenure. The data and insights provide the proof.

This section looks at two key ingredients necessary for a data-driven culture.

Improving data literacy

According to an Accenture study, many data workers feel ill-prepared to work effectively with data. As a result, organizational productivity is being held back. The study suggests providing a data literacy training program to create confident data workers. You can find out more about this study by visiting `newsroom.accenture.com/news/new-research-from-accenture-and-qlik-shows-the-data-skills-gap-is-costing-organizations-billions-in-lost-productivity.htm` and downloading the report, *The Human Impact of Data Literacy.*

Collaborative tooling

Another essential way to build a robust data culture is to ensure that you evaluate and improve data tools to provide for actual user requirements. Collaborative tooling enables data to be easily manipulated and consumed.

DEMOCRATIZING YOUR DATA

In partnership with Databricks, MIT Tech Review conducted a global survey (2021) of 351 chief data officers, chief analytics officers, chief information officers, and other senior technology executives to determine how they succeeded (or didn't) at building a high-performance data and AI organization.

Among their key findings was the need to democratize the data. To accomplish this, they recommended the following:

- **Embedding data scientists directly into business units:** This process enables them to interact directly with data users.
- **Putting access to the analytics in the hands of the users:** This action permits users to draw insights themselves.
- **Providing senior leaders with access to visual tools with simple interfaces:** This provision allows your senior leaders to get data insights as needed.

Capitalizing on the Vs of Data

Coping with the wave of data that hits companies every day is impossible to manage without the right data management solution. Known as the five Vs of big data, you can describe data in five ways:

- **Volume:** The size of the data
- **Veracity**: The extent to which you can rely on the data
- **Velocity:** The speed at which the data is consumed

  The assumption is that it should be analyzed as close to receipt as possible.
- **Variety:** The complexity of the data (different formats)
- **Value:** The usefulness of data in decision making (not all data is equal)


TIP


TIP

If you're using Databricks Lakehouse, all your data (both structured and unstructured) can be processed without moving or copying any data. This process allows you to include all properties of unstructured data sets from social media posts and metadata to catalog images in their analysis and models.

Focusing on Business Value

Now that so many more unique forms of data are available (for example, semi-structured data that includes customer interactions from the web and mobile devices or social media posts), businesses realize that their legacy platforms can't scale and meet the increasing demands for better data analytics.

Data and technology executives want to use that data to get better insights to increase business impact. Specifically, they seek a lower-cost approach that improves the user experience and increases collaboration across data personas. This goal moves them away from complex and expensive on-premises architectures.

Reviewing Options in Future Technology

Lakehouse architecture allows you to future-proof your technology with open standards and the use of multi-cloud. Many drivers exist for a multi-cloud approach, but one of the biggest is economic leverage.

REMEMBER

As cloud adoption and data needs grow, spending on cloud infrastructure will be one of many organizations' most significant line items. You should adopt multi-cloud architecture for a variety of reasons:

- Data is naturally multi-cloud.
- It improves negotiation leverage.
- You get the best technology for the workload.
- Cost is optimized.

Getting Better Insights from Data and Using AI/ML at Scale

Using AI and machine learning (ML) at scale provides you with a competitive advantage. It helps you find patterns and trends in the data that provide insights useful for decision making. AI and ML help your organization make better decisions by analyzing data collected from different sources. It's key to being successful in today's marketplace.

IN THIS CHAPTER

- Future-proofing your organization
- Understanding the evolution to the lakehouse
- Looking at lakehouse architecture and data lakes

Chapter 3
Dawning of the Lakehouse

Lessons learned from working with enterprise data warehouses (EDWs) and data lakes have paved the way for the lakehouse's modern cloud-based data architecture. It combines both the best properties and capabilities to provide a far more powerful and flexible data platform than possible in the past. This chapter looks at the evolution of the modern cloud-based lakehouse and the need to future-proof your organization.

Future-proofing Your Organization

One of the critical requirements of enterprise leaders is that they successfully prepare their organizations to meet the future. Relying on antiquated processes to manage data could overwhelm your organization, put you at a competitive disadvantage, and waste critical human and financial capital.


TECHNICAL STUFF

According to Statista, the total amount of data created, captured, copied, and consumed worldwide is forecast to increase by 152.5 percent from 2020 to 2024 to 149 Zettabytes.

Is your organization prepared to operate and maintain a complex technology stack of data lakes, EDWs, business intelligence (BI), data science, machine learning (ML), and streaming platforms and the complexity of moving data between them and managing different security paradigms of each? Or would you rather consider streamlining it to one lakehouse platform that's simple to manage so you're prepared and focused on solving the business challenges with data versus complex platform and security management? Consider migrating to the lakehouse to ensure that you're prepared for the new challenges ahead.

Evolving from the EDW to the Data Lake to the Lakehouse

The lakehouse's modern data architecture can be seen as the evolution of the EDW from the 1980s and the Hadoop-style data lakes from the mid-2000s, as shown in Figure 3-1.

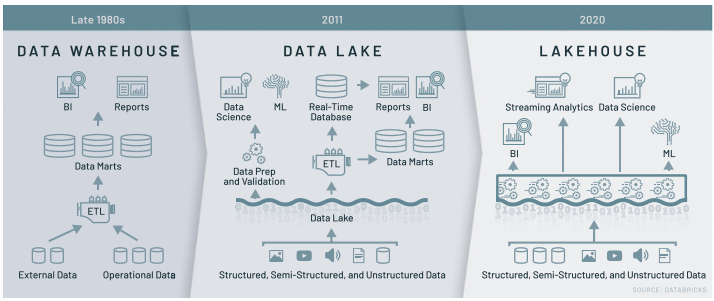

FIGURE 3-1: The evolution of data management.

Early data warehouses were optimized for analytics but not for unstructured data. Likewise, data lakes were traditionally used to store unstructured data but weren't optimized for analytics. The result: You had to choose between agility and governance. The value of the lakehouse architecture is that data teams can now store all their data on *one* platform, with the speed and governance of a data warehouse and the flexibility of a data lake.

Reviewing What a Lakehouse Is

The lakehouse takes the greatest elements of data warehouses and data lakes and combines them into a single platform that gives you the best of both worlds. Operating a lakehouse architecture is the foundation that enables you to

- Manage all data use cases on one single source of truth for all your data.
- Be more responsive and find new insights faster.
- Allow everyone to look at the same version of the data.
- Simplify existing architectures and security by reducing silos and the number of systems and tools that you need to manage.
- Have the ability to consolidate and tie your data marts and EDWs with other unstructured data for enrichment and create innovative data products.
- Perform extract, transform, load (ETL) operations on the data within the data lakehouse.

REMEMBER

Lakehouse architecture is

- **Simple:** Unify your data, analytics, and AI on one platform.
- **Open:** Unify your data ecosystem with open standards and formats, which prevent vendor lock-in.
- **Multi-cloud:** Maintain consistent management, security, and governance experience across all clouds and enable your teams to focus on putting all your data to work to discover new insights.

As an example, the Databricks Lakehouse platform is shown in Figure 3-2.

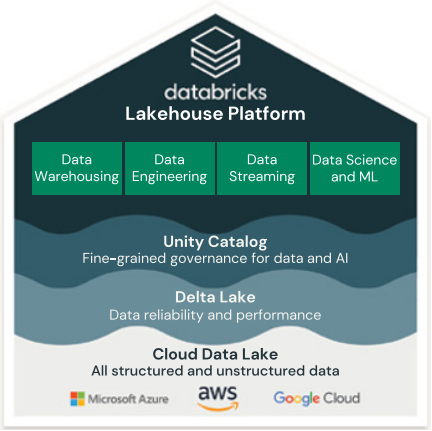

FIGURE 3-2: The Databricks Lakehouse platform.

Looking at Legacy Data Lakes

Data lakes, as they were previously developed, are considered legacy architecture. The idea behind them was to store all the data generated by various sources, whether structured or unstructured, in a single repository and handle computations where data is located (tightly coupled storage and compute). If you want to build a modern data lake, you need to rethink your approach.

TIP

Compared to traditional data warehouses and data lakes, the data lakehouse has several critical advantages:

- **It's more flexible.** You can store data in whatever format makes sense for your business.
- **It's always up-to-date.** Because of the extract, load, transform (ELT) patterns, analysts get access to the freshest data without additional ETL. Real-time streaming pipelines are natively supported in lakehouse.
- **You only need a single copy of the data.** You don't need to maintain multiple copies of the data and in different schemas and formats for facilitating BI or ML workloads.
- **It's easy to add new data sources.** It's easy to ingest data without extensive ETL.
- **It's truly enterprise scale.** Unlike maintaining many subject-area-wise data warehouses and data marts, the data lakehouse houses all your EDWs, data marts right next to your raw, and unstructured and semi-structured data.

IN THIS CHAPTER

- Transforming your organization
- Looking at customer success stories

Chapter 4
Benefitting from Lakehouse Migration

Legacy on-premise analytics architectures aren't keeping up with the needs of modern organizations. Things like the inability to use real-time data, getting predictive insights, and Hadoop's rising cost prevent teams from driving high-impact business outcomes.

As organizations look to do more with their data, empower their data teams to do more analytics and artificial intelligence (AI), and reduce infrastructure maintenance and data management costs, the world of data and AI needs a Hadoop alternative. Organizations worldwide have realized that it's no longer a matter of *if* migration is required to stay competitive and innovate but a matter of *when.*

This chapter looks at the many benefits of migrating to the data lakehouse and gives you some customer success stories.

Transforming Your Organization

When making changes that will future-proof their organization, data leaders are highly concerned about the following:

- The opportunity cost of not migrating
- The cost of migrating away from their existing solution
- The learning curve that data teams will encounter and the availability of skills for building solutions on the new platform
- The breadth of the ecosystem built around the solution

They believe that four truths should guide their decisions:

- **Machine learning (ML) and artificial intelligence (AI) are the future.** To gain better insights from their data, leaders want to move up the maturity curve so they can take full advantage of ML and AI. See Chapter 2 for details on the maturity curve.
- **Use open source and open formats.** Leaders are uneasy about locking themselves in with any vendor and want to use open source whenever possible. Using open source and open formats also helps with more readily available skilled resources compared to using proprietary technologies and formats.
- **Plan for multi-cloud.** Most companies are moving to the cloud for cost savings, flexibility, and scalability and to build to strengths. However, in adopting a multi-cloud strategy, it's important to have the ability to run your analytics and AI workloads on a single unified platform.
- **Use simple data architecture.** Leaders want to eliminate complexity, minimize silos, and avoid multiple copies of the same data. They want one governance, security, and lineage model for all their data across all clouds.

Regarding these truths, consider how the lakehouse architecture meets each of these needs:

- **ML and AI are the future.** The lakehouse uses ML and AI from the ground up.
- **Use open source and open formats.** The lakehouse uses open formats and standards to provide greater data portability and avoid vendor lock-in.

- **Plan for the multi-cloud.** The lakehouse leverages low-cost cloud object stores to store *all* enterprise data across the major cloud providers.
- **Use simple data architecture.** The lakehouse supports all use-case platforms, including data engineering, data warehousing, real-time streaming, data science, and ML across the major clouds.

When you migrate to the lakehouse, you derive several beneficial outcomes. These include the ability to

- Continue leveraging low-cost cloud object stores.
- Run different data workloads from BI to AI cost-effectively and efficiently.
- Utilize open formats and standards. You won't experience vendor lock-in.
- Future-proof your investment in modern cloud data architecture, including the ability to leverage multi-clouds.

Databricks Lakehouse Customer Success Stories

If you're curious how other companies have migrated from Hadoop to the lakehouse, this section provides you with two companies that realized significant success when they migrated to the Databricks Lakehouse platform.

CBC Radio-Canada

CBC Radio-Canada's mission is to enlighten and entertain its diverse audience. However, its legacy Hadoop infrastructure prevented the company from using its diverse data to personalize its customer interactions in the way it wanted to. Personalization was a driving goal. CBC Radio-Canada turned to Databricks for help and leveraged the efficiencies of the cloud and the power of a lakehouse architecture to give data teams direct access to all of the company's data.

Databricks SQL empowered CBC Radio-Canada to quickly run SQL queries to derive insights into its digital audiences and behaviors. The reduction in time-to-insight was lowered by 50 percent. The company is now able to provide more visibility into its digital audiences and develop strategies and services that will boost engagement and retention.

H&M Group

H&M Group is well-known as a major disruptor and innovator in the fashion and retail industry. However, its on-premise Hadoop system crippled its ability to ingest and analyze data generated by millions of customers needed to power predictive models. To solve this problem, H&M Group moved to the Databricks Lakehouse Platform to simplify infrastructure management, enable performant data pipelines at scale, and simplify the ML life cycle. This allowed the group to make data-driven decisions that accelerated business growth.

The results were dramatic:

- A 70 percent reduction in operational costs
- Better cross-team collaboration and a reduction in the number of components needed to go in production with easy setup and management
- Faster time-to-insight

IN THIS CHAPTER

- Taking an agile approach to migration
- Detailing the Hadoop migration

Chapter 5
Reviewing Why to Migrate to the Lakehouse

Migrating to a new architecture can be a complex process. If you've decided to migrate from your legacy Hadoop, think through several essential considerations before proceeding. Understanding the tradeoffs inherent in different architectural choices helps you make well-informed decisions about how best to proceed with your modernization initiatives. This chapter looks at the value of taking an agile, iterative approach to migration and suggests how to plan and execute the migration journey.

Reviewing an Agile Iterative Approach to Migration

How you migrate your data to the lakehouse is a critical decision. Regardless of the path you follow, balance is required. Databricks recommends that you do all your migrations in a phased agile manner:

- **Make the decision.** Migrating from Hadoop to a modern cloud data platform can be daunting, but staying with

existing solutions can be even worse. The technical debt you accrue and pain of staying where you are can be significantly worse than the costs of migrating.

- **Use a balanced approach to lift and shift and modernization.** Lift and shift the code as well as modernize in one iteration. Use lift and shift with automated code convertors and immediately modernize to optimal Databricks patterns.


TECHNICAL STUFF

Here *lift and shift* refers to the movement of an application design and code from one environment to another without making massive changes. But don't wait to modernize and redesign later — immediately redesign and apply all best practices. Decide what needs redesign and what code can benefit from a lift-and-shift approach. As the old adage goes, don't try to fix something if it's not broken.

- **Learn what worked and didn't and iterate.** Add on additional use cases and workloads as you go.
- **Show success in shorter sprints and adapt.** This way you accelerate the value delivered and demonstrate immediate success to the stakeholders. The learnings and feedback also help you improve the next iteration of migration.


WARNING

Just simple lift and shift is rarely the answer; if you just lift and shift, you don't get these three essential benefits:

- You won't get to fully utilize the lakehouse.
- You lose out on any innovation you may discover if you just move everything as is.
- You won't have the chance to improve your tech strategy to optimize for cost savings, agility, and scale.

Both lift and shift and total re-engineering have pros and cons:

- **Lift and shift:** On the pro side, it's faster and more critical to do if you have an upcoming license renewal deadline. The downside is that you may not take the opportunity to re-engineer and refactor the design and code.
- **Total re-engineering:** One benefit is that it gives you the best quality. On the other hand, it can take a long time to complete and comes at a high upfront cost.

Moving Hadoop Architecture to the Cloud

Moving Hadoop to the cloud is an excellent first step in your migration journey. However, it doesn't unlock your data like moving to the lakehouse does. Consider what the impact of moving your data to the lakehouse affords you. You can

- Deliver data faster to business users for better and more timely business decisions.
- Deliver consistent data from a shared data lake that's properly governed to ensure the entire organization is working off the same data.
- Achieve greater scalability to take on the largest AI and ML use cases and enable more effective analytics and AI by finding new markets, increasing revenue, reducing costs, and lowering risk.
- Make larger data sets available to business decision makers and give them the ability to visualize your entire data lake while keeping costs low through a pay-for-what-you-use architecture.

Planning the Migration Journey

When considering a migration journey, carefully plan each step along the way. Therefore, the journey can be depicted as a set of steps. Here's how each step works:

1. **Ask internal questions.**

   This step covers the discovery phase. The key to this step is to answer two questions: Where am I now, and where do I need to go? Make sure that you collect questionnaires from all your data teams, chief information officers, and other relevant stakeholders. Be prepared for a lot of new learning and self-discovery as teams test and validate assumptions.

2. **Make a migration assessment.**

   During the assessment phase, you want to refine and evaluate the solutions on the table. Take an inventory of all workloads and prioritize the use cases.

TIP

When you complete the migration assessment, you'll have a clearer sense of your timeline and alignment with your original planned schedule.

3. **Conduct technical planning.**

   The strategy phase is critical. Think through your target architecture and make sure it supports the business in the long term. You make crucial decisions in this phase on your ingestion strategy and technologies, extract, transform, load (ETL) patterns and tools, data organization principles in the lakehouse, and semantic and reporting layer architectural and tool choices.

4. **Complete evaluation and enablement.**

   At the production pilot phase, you need to understand what your new platform has to offer. Conduct some targeted demos or plans to help you finalize your approach.

5. **Execute your migration.**

   The rubber meets the road at execution. Make sure you get this migration right the first time.

This migration methodology is shown in Figure 5-1.

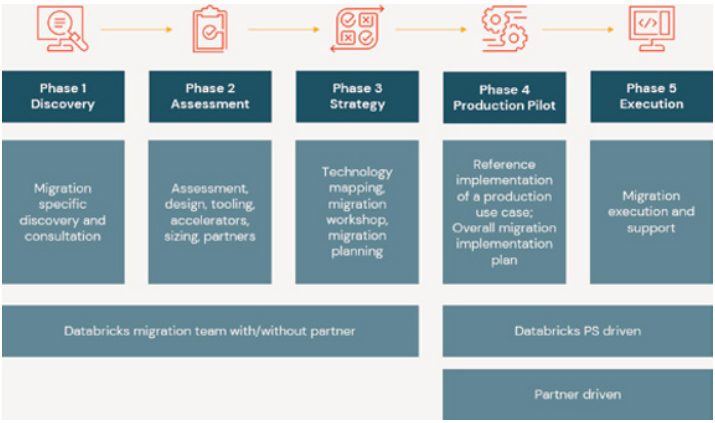

FIGURE 5-1: The phases of migration methodology.

REMEMBER

The sooner you execute your migration, the quicker you can start to scale your analytics practice, cut costs, and increase overall team productivity.

Migrating from Hadoop to the Lakehouse

No matter what architecture you're working with, migrations are never easy. Databricks suggests you consider five areas of focus when migrating from Hadoop to minimize adverse impact, ensure business continuity, and manage costs effectively. Those areas are covered in this section.

REMEMBER

Don't forget to bring key business stakeholders along on this journey. It's as much a technology decision as it is a business decision. You need your business stakeholders brought into the journey and its end state to have a successful migration.

TECHNICAL STUFF

For a technical deep dive on the five migration areas covered in this section, visit `databricks.com/blog/2021/08/06/5-key-steps-to-successfully-migrate-from-hadoop-to-the-lakehouse-architecture.html`.

Figuring out what you're administering

Hadoop environments have been confined by single compute environments. As migrations happen, consider how to take advantage of a multi-compute environment with customization to size, run time, and more — all while gaining greater control on who has access to what data.

Migrating your data

Consider a balanced approach to your data migration to ensure business continuity while enabling new use cases. Dual ingestion allows for organizations to feed new data to your preferred cloud storage, and historical data can be migrated in parallel or at a later stage. Also, with data migration, you need to decide between a push-oriented data migration and a pull data migration. A *push-oriented data migration* is when the source pushes data to the target system, and a *pull data migration* is when the target systems pull data from the source system.

You choose between the two by determining how critical the current workloads are to the business and if the business can afford downtime with the migration. Push migrations tend to be easier and more automated, whereas pull migrations (maybe needed for a subset of your data) can be more complex and intrusive as you invite third parties to extract the data.

Processing your data

Data processing is transferred over to a data lakehouse across the full stack. From data engineering, ETL, batch processing, to SQL and ML, the data lakehouse architecture easily allows organizations continuity with their data processing needs for like-to-like capabilities — the biggest difference is that it's now all on a modern cloud platform.

Reaping the benefits of security and governance

As you migrate, you can enjoy the benefits of familiar security and governance capabilities to meet your organization's requirements. The benefits include

- **Authentication:** Single sign on (SSO) with SAML 2.0 supported corporate directory
- **Authorization:** Role-based and table row/column access control to secure your data
- **Metadata management:** Integration with your preferred third-party tools
- **Simplified management:** Quick and easy curation of data for your organization

  You can also define who can access what and under what purpose. Search and quickly find data assets through enhanced discovery capabilities, and know your data lineage.

Considering your SQL and BI workloads

Don't just transfer your Hadoop workloads; consider your SQL and BI workloads as well. Easily transfer over users to Databricks SQL with familiar experiences, tools, and third-party visualization integrations.

IN THIS CHAPTER

- Identifying business value
- Building successful teams
- Democratizing access to quality data

Chapter 6
Ten Points to Enable and Scale Your Data and AI Strategy

To be successful in scaling your data and artificial intelligence (AI) strategy, you need to enable and scale your data and AI strategy with a modern data stack. To achieve that goal, Databricks suggests you keep in mind the following:

- **Move to production and drive adoption.** Establish a clear set of metrics to measure adoption and track the net promoter score (NPS) so the user experience continues to improve over time.
- **Establish goals and your business value.** Most organizations establish goals for their data, analytics, and AI journey across business outcomes, people, and technology. Modernization efforts support your organization's ability to be more agile, flexible, and secure to build new products/offerings and drive operational efficiencies.
- **Identify and prioritize use cases.** Use cases should avoid the trappings of "cool" data science and machine learning (ML) projects and focus on delivering business value.

- **Build successful data teams.** To succeed with data, analytics, and AI, find and organize the right talent into high-performing teams that can execute against a well-defined strategy with the proper tools, processes, training, and leadership.
- **Deploy a modern cloud data stack.** The capabilities, limitations, and lessons learned from working with two legacy and disparate data architectures inspired the next generation of data architectures — what the industry now refers to as the *data lakehouse.*
- **Improve data governance and compliance.** We recommend adopting data policies and practices that help the business to realize value through centralizing storage, metadata, and governance on a single platform. This enables you to ensure your data governance rules scale with your needs, regardless of the number of workspaces or the business intelligence tools your organization uses.
- **Democratize access to quality data.** Legacy data platforms lack the ability to democratize access to quality data. Effective data, analytics, and AI solutions rely more on the amount of quality data available than on the sophistication or complexity of the model or algorithm.
- **Increase the productivity of your workforce.** What tools should you provide to the user community so it can be the most effective at using the new data ecosystem? When thinking about this question, consider working backward from the user experience, move beyond best-of-breed tools, unify the platform and personas, and make sure the tools are simple and self-service.
- **Make informed build versus buy decisions.** When you're going through modernization initiatives, your engineers may want to build versus leveraging existing vendor solutions. When faced with this dilemma, consider if you can afford to become your own software vendor for the long term, how long it will take, and if your organization can afford the wait. Also consider if the technology can evolve quickly as your data needs change.
- **Allocate, monitor, and optimize costs.** Any decision to modernize your data platform requires you to reduce complexity and the costs of managing legacy platforms to reinvest resources into new business value initiatives.

EBOOK

Why the Data Lakehouse Is Your Next Data Warehouse

Ready to explore the inner workings of the lakehouse?

It's time to go under the hood. See how Databricks SQL delivers up to 12x better price/performance than legacy cloud data warehouses.

In this eBook, you'll learn how to:

- Ingest, store and govern business-critical data at scale to build a curated data lake
- Get started in seconds with instant, elastic SQL compute to process all query types with best-in-class performance
- Quickly find and share new insights with a built-in SQL editor, visualizations and dashboards or your favorite BI tools

databricks

Get the eBook →

Revamp your data strategy with the lakehouse

Traditional data lakes are primed for the next phase in their evolution. As the volume and types of data available continue to grow, quality of data within data lakes becomes more difficult to manage. With the data lakehouse, you can manage and analyze all your data in one place, allowing you to build, deploy, and scale analytical applications in minutes instead of days.

Inside...

- The history and evolution of data lakes
- Why the lakehouse is the next step
- Assessing your company's data strategy
- Planning your migration

Stephanie Diamond is a former AOL marketing director and founder of Digital Media Works, an online marketing company that helps businesses discover their hidden profits. She has authored over 25 marketing books and custom e-books, including *Facebook Marketing For Dummies*.

ISBN: 978-1-119-89420-9
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

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