LEARNING MADE EASY

Databricks Special Edition

Migrating from a Data Warehouse to a Data Lakehouse

for dummies®
A Wiley Brand

Understand the data lakehouse

Modernize your data and AI strategy

Form a migration plan

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 a Data Warehouse to a Data Lakehouse

Databricks Special Edition

by Stephanie Diamond

for dummies®
A Wiley Brand

Migrating from a Data Warehouse to a 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: 978-1-119-89472-8 (pbk); ISBN: 978-1-119-89473-5 (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

Acquisitions Editor: Ashley Coffey

Senior Client Account Manager: Matt Cox

Content Refinement Specialist: Tamilmani Varadharaj

Table of Contents

Introduction

The data lakehouse is a cloud-native platform for data management that provides a powerful engine for data processing and simple and intuitive 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.

With the data lakehouse, you can analyze all your data in one place without moving it to another system first. The key to utilizing this innovative platform is migrating your current system to the data lakehouse. This book looks at why and how to migrate from your enterprise data warehouse (EDW) to the data lakehouse to prepare your organization to meet the future.

About This Book

This book uncovers what leaders need to know to meet new challenges in data management. I cover several topics, including the following:

- The types of unique data that must be managed, including structured, unstructured, and semi-structured data
- How to determine where your company resides on the data and artificial intelligence (AI) maturity curve
- The evolution from EDWs to the data lakehouse
- Benefits that accrue to companies that migrate to the lakehouse
- What to consider when you migrate to the lakehouse
- Ten reasons to migrate to the Databricks Lakehouse

Icons Used in This Book

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

TIP

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

REMEMBER

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

WARNING

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

TECHNICAL STUFF

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

Beyond the Book

This book can help you discover more about migrating from your EDW 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 run SQL workloads on the lakehouse architecture.
- `databricks.com/p/ebook/building-the-data-lakehouse`: Download *5 Steps to a Successful Data Lakehouse*.
- `databricks.com/p/ebook/data-lakehouse-is-your-next-data-warehouse`: Download *Inner Workings of a Lakehouse* to learn what's under the hood of Databricks SQL and how it makes the Databricks Lakehouse your next data warehouse.

IN THIS CHAPTER

- Looking at data management challenges
- Reviewing new data types
- Moving data to the cloud

Chapter 1
Recognizing the End of an Era

Business data continues to be one of the most valuable assets a corporation possesses. As data availability continues to explode, maximizing, optimizing, and refining enterprise data are seen as central to a thriving business. But it's hard for companies to keep up with the growing volume. As a result, you need well-designed data management architecture to help you minimize risks and reach your financial goals.

This chapter looks at the developments driving the adoption of the data lakehouse and reviews the evolution from data warehouses to the lakehouse of today.

Facing Today's Challenges

Effectively managing all forms of data is critically important to leaders who want to future-proof their organizations. They need to integrate their data to accommodate machine learning (ML), artificial intelligence (AI), and data science. When deciding what technology to deploy, they need to ask themselves "Will this get us where we need to go?"

The advent of low-cost cloud storage, open-source software, ML, and AI have allowed for a significant shift in how organizations leverage their data. In addition, the COVID-19 pandemic forced companies to adapt to a remote distributed workforce. As a result, cloud adoption has skyrocketed. The enterprise data warehouse (EDW) of the past was a closed proprietary system not suited to accommodate modern data management challenges that include the ability to

- Perform ML, data science, and AI, and support other new sources of data required to make predictions
- Store audio and video data sets
- Support streaming for real-time operations
- Scale in a flexible manner
- Manage raw data regardless of the format

To understand the scope of the problem, you need to see that as technologies advanced, many different types of data became available and companies recognized their significant value. Businesses realized that they needed a unified place to store and analyze not only their structured data but also an increasing volume of semi-structured and unstructured data.

- **Semi-structured:** This data includes logs, clickstream, CVS, JSON, and XML.
- **Unstructured:** This data comes from internal conversations inside the organization in the form of documents, emails, and letters and includes other unstructured data such as internet of things (IoT) data like temperature gauges, drones, and factory machines, and image, video, and analog-based data.

Looking Back at the Early Days of Data Management

To understand how technology has evolved, consider the early days of data management. There was no real need to house data in a centralized repository. Relational databases using a structured

query language (SQL) were deployed because the available data was hierarchical and stored in database tables. For the longest time, this method was adequate to create the necessary financial and other business reports.

Understanding the origin of data warehousing

As the amount of data grew, the data warehouse was created as a central place that combined data from operational systems and external data sources for reporting and analysis. As a result, IT departments were familiar with their structure and knew how to use them efficiently. Some of the benefits of using a data warehouse included the ability to

- **Consolidate data from various sources:** It brought together data sources that could be optimized and queried. It acted as a single point for all data.
- **Obtain historical intelligence:** It separated analytics processing from transactional databases, thereby allowing the extraction of historical intelligence.
- **Maintain data quality, consistency, and accuracy:** It was consistent in naming conventions codes for various product types, languages, and currencies.
- **Support for reporting and business intelligence (BI) analysis:** Organizations could get the reports and BI they needed by using this structure.
- **Support high-speed retrieval:** They were built around a carefully designed data model that transforms production data from a high-speed data entry design to one that supports high-speed retrieval.

Examining the inadequacies of data warehouses

As businesses began deploying data warehouses, their inadequacies quickly became evident as new forms of data emerged. Consider some of the problems with data warehouses:

- The data was scattered across multiple sources (multiple databases and multiple subject-area-based EDWs and data marts).

- Data from each source had its own schema, and each business application used its own schema. This required extensive and complex extract, transform, load (ETL) to load it in standardized data models — only to be copied again in different formats by different business teams.
- ETL to load data warehousing requires extensive modeling and months of efforts. By the time the data was ready to be analyzed, the business need was either already met or changed, and the data was often outdated.
- Scaling became exponentially more expensive.
- There wasn't support for data science, ML, and real-time analytics, or semi-structured or unstructured data sets.

Adding data lakes

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 ML/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 around 2011 with the advent of Hadoop.

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 it used ETL. 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 there were no bottlenecks between the data lake and its external sources.

Some of the immediate benefits of data lakes included the facts 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 data set.
- 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.

Inevitably, some challenges with data lakes also existed. Those included such things as

- No support for ACID (atomicity, consistency, isolation, and durability) transactions
- No enforcement of data quality or governance
- Failed jobs and missed data
- Poor 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.

Considering Traditional Data Warehouses in the Cloud

A few years into modern enterprise data storage, it became apparent that there were some serious challenges that businesses needed to overcome. Specifically, organizations struggled with large data silos, insufficient storage, and a lack of efficiency, compounded by costly hardware and software. Enter cloud data warehouses.

What spurred the adoption of the cloud data warehouse? It offered several key fundamentals that businesses required for success:

- **Elasticity:** It provided for instant provisioning, which meant you could scale up and down on demand.
- **Easy administration:** In addition to infrastructure cost savings, you could free up your admin resources to focus on what mattered most.
- **Speed of innovation:** You could grow your business with instant environments for innovations.

IN THIS CHAPTER

- Looking at the top goals of data and technology executives
- Developing a data culture
- Evaluating a maturing model

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.

This chapter looks at the top three strategic goals that data and technology executives want to achieve and the benefits of establishing a data culture. You also look at a maturity model to determine where your organization fits and steps you can take to progress.

Examining Top Strategic Goals

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

- Creating new business value from data
- Reducing risks
- Controlling costs

You look at each goal in this section.

Creating new business value from data

Now that there are so many more unique forms of data 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 enterprise data warehouse (EDW) architectures.

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 forms, 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:

- The need to adhere to the evolving privacy regulations landscape, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA)
- The need to adjust data management to contend with new privacy directives like those from Google and Apple
- The need to figure 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 warehouses can get very expensive, very 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 informed business decisions daily by using available, relevant data.
- Data is given greater weight than experience, intuition, or tenure. The data and insights provide the proof.


TIP

Does this sound like your organization? If not, check out the later section "Examining a Data and AI Maturity Curve" for a data maturity model that can guide your progress.

If want to focus on building a better data culture, consider the following ideas:

- **Be clear about how you will achieve business goals and outcomes.** Business leaders today recognize that data and AI help them reach their goals. Make sure to detail how you will support them in your organization.

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.

- **Invest in people.** In today's marketplace, employees are the key to long-term success. Competition is fierce to hire the best talent, so you need to focus on two things:
  - Making your data environment as frictionless as possible so the best talent will come to you
  - Retraining to utilize the skills of your best employees
- **Modernize your technology.** As enterprises move away from complex proprietary solutions, your data environment should support open standards, low-cost storage, and on-demand compute.

Getting What You Need from Your Data

Many people say that data is the new oil. However, oil can only be refined once. The value of data is that it can be re-analyzed in myriad ways to produce the answers you need. For this reason, it's imperative that your enterprise data platform architecture can analyze your data to answer critical questions.

REMEMBER

Customers of data platforms want to know

- **What happened and why?** To answer this question, they need reports and dashboards, analysts, and subject matter experts empowered to self-serve analytics.
- **What will happen?** You need capabilities for data science, machine learning (ML), and artificial intelligence (AI).
- **What timely action can be taken?** They want prescriptive analytics that tell them how to respond (sometimes in seconds, in an automated way; for example, fraud analytics). Business managers, data scientists, and data engineers can derive these answers by working together on a collaborative platform.

Figure 2-1 shows you how these questions work together.

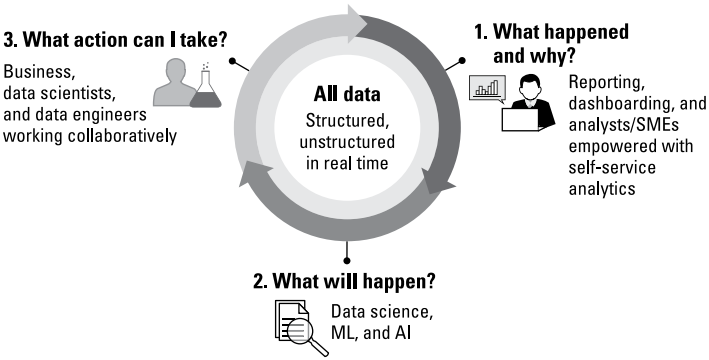

FIGURE 2-1: The questions customers of data platforms ask.

Examining a Data and 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 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`.

IN THIS CHAPTER

- Understanding the evolution of the lakehouse
- Future-proofing your organization
- Looking at lakehouse architecture

Chapter 3
The Dawn 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 and put you at a competitive disadvantage.

According to Statista, research indicates that 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 simplifying 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 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.

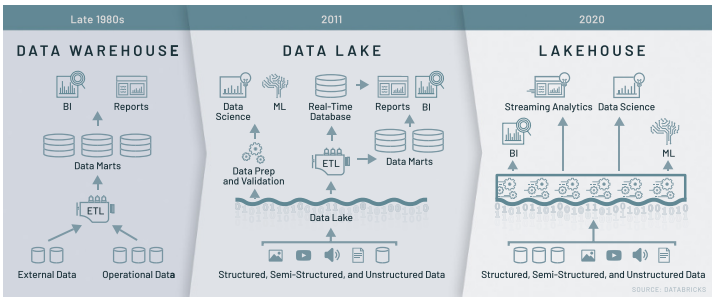

FIGURE 3-1: The evolution of data management.

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 best 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.
- Have everyone 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:** It enables you to unify your data, analytics, and AI on one platform.
- **Open:** It unifies your data ecosystem with open standards and formats, which prevents vendor lock-in.
- **Multi-cloud:** It offers you a consistent management, security, and governance experience across all clouds and enables 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.

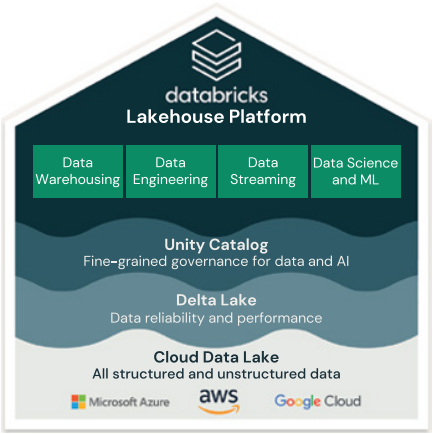

FIGURE 3-2: The Databricks Lakehouse platform.

Understanding Why the Lakehouse Is the Modern Cloud Data Warehouse and More

The lakehouse is a system where all the data from various sources are stored together. This way, a single query can access multiple data points from different business units or points of origins at once.

TIP

The data lakehouse has several critical advantages over traditional data warehouses:

- **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 ELT (extract, load, transform) 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.
- **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

- Looking at the four truths of data leaders
- Improving technology
- Delivering better healthcare

Chapter 4
Benefitting from Lakehouse Migration

Extracting knowledge from all the raw data from your organization's disparate systems provides you with a tremendous competitive advantage. If you could go back in time, you would likely make different decisions about how you manage your data to support today's challenges. This chapter looks at how you can transform your organization by deploying lakehouse architecture and seeing how companies like Bread and Amgen have benefitted from making this transition.

Transforming Your Organization

When making changes that will future-proof their organizations, data leaders are concerned with the following:

- The cost of migrating away from their existing solution
- The learning curve that data teams will encounter
- 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 to take full advantage of ML and AI. See Chapter 2 for more details.
- **Use open source and open formats.** Leaders are uneasy about locking themselves in with any vendor and want to use open source whenever possible.
- **Plan for the cloud.** Most companies move to the cloud to get cost savings, flexibility, and the ability to build to strengths.
- **Use a 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.

Regarding the above 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 cloud.** The lakehouse leverages low-cost cloud object stores to store *all* enterprise data.
- **Use a simple data architecture.** The lakehouse supports all use-case platforms, including data engineering, data warehousing, real-time streaming, data science, and ML.

Customer Success Stories

If you're curious how other companies have used the lakehouse, this section shows you two companies that had significant success when they migrated to the lakehouse, and specifically, the Databricks Lakehouse platform.

Bread

Bread is a division of Alliance Data Systems. It's a technology-driven payments company that integrates with merchants and partners to personalize payment options for their customers. Bread's data warehouse couldn't handle data growth from gigabytes to terabytes. It was taking hours to query data. The company also struggled with switching from batch to streaming data, hindering their ability to deliver real-time insights and results.

Bread migrated to the Databricks Lakehouse Platform on AWS to efficiently ingest transactional data from its point-of-sale systems and extract, load, transform (ELT) systems into Delta Lake. The company did this while supporting its analytics engineers and data scientists to democratize its data sets for credit risk, loss estimation, and fraud use cases. By using this solution, Bread now scales data analytics and ML without worrying about large data volumes slowing data ingestion and performance. As a result, the company can analyze terabytes of data for downstream business reporting, analytics, and ML use cases, all designed to improve business decision making and the customer experience.

In fact, Christina Taylor, staff data engineer at Bread, adds that moving away from the company's old cloud computing–based data warehousing solution to the new Databricks' lakehouse transformed the way Bread drove business actions based on the most complete and current view of its data. This was something that just wasn't possible with its data warehouse before.

Other results included a 90 percent reduction in compute costs compared to its past data warehouse, 140 times the volume of data at only 1.5 times the cost, and a 23 percent performance increase in actionable insights for business reporting. Visit `databricks.com/customers/bread-finance` for the full customer story and journey from a cloud data warehouse to the Databricks Lakehouse.

Amgen

Amgen is the world's largest independent biotech company. Over the last 40 years, its vast amount of data has helped pioneer new drug-making processes and develop life-saving medicines. As the size of its data grew, the company couldn't weave together and scale the various aspects of its business. Amgen needed to expand its cross-functional collaboration to take advantage of

the many perspectives present in its data. To support the digital transformation journey, Amgen chose to use the Databricks Lakehouse Platform.

The implementation of the Databricks Lakehouse Platform helped the company serve its patients and improve the drug development life cycle. Amgen's data ingestion rates increased significantly, improving processing times by 75 percent, resulting in two times faster delivery of insights to the business while reducing compute costs by approximately 25 percent over static Hadoop clusters. Since partnering with Databricks in 2017, over 2,000 data users from data engineering to analysts have accessed 400 terabytes (TB) of data through Databricks to support over 40 data lake projects and 240 data science projects.

SETTING A PERFORMANCE RECORD

Databricks has been rapidly developing full blown data warehousing capabilities directly on data lakes, bringing the best of both worlds in one data architecture: the data lakehouse. It announced its full suite of data warehousing capabilities as Databricks SQL in November 2020. The open question since then has been whether an open architecture based on a lakehouse can provide the performance, speed, and cost of the classic cloud data warehouses.

Barcelona Supercomputing Center, which frequently runs TPC-DS on popular data warehouses, discovered that Databricks SQL set a new world record in November of 2021 in 100TB TPC-DS, the gold standard performance benchmark for data warehousing. Databricks SQL beat the previous record by 2.2 times. The research also highlighted that Databricks outperformed its competitor with the following results:

- **Excellent price/performance as a data warehouse:** Get up to 12 times better price/performance that results in substantial cost savings compared to other data warehouses.
- **Faster time to insights:** Process large volumes of data 2.7 times faster compared to leading cloud data warehouses.

You can read the full story here: `databricks.com/blog/2021/11/02/databricks-sets-official-data-warehousing-performance-record.html`.

IN THIS CHAPTER

- Taking an agile approach to migration
- Seeing why lift and shift doesn't work
- Looking at the five pillars of migration

Chapter 5
Reviewing Why to Migrate to the Lakehouse

Migrating to a new architecture can be a complex process. As you consider your data warehouse modernization strategy, plan out several essential migration factors before proceeding. Understanding the inherent differences in 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.

Using an Agile Approach

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

- **From a platform perspective, go with modernization to lakehouse versus the lift and shift of the legacy EDW to**

yet another Cloud EDW. Lift and shift of the platform results in carrying the same issues and shortcomings to the cloud. From a code and application redesign perspective, have a balanced approach to lift and shift and modernize.

- **Implement 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

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.

- **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 immediately show success to the stakeholders, and the learnings and feedback 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:** *Pro:* It's faster and more critical to do if you have a multi-million data warehouse license renewal coming up. *Con:* You may not take the opportunity to re-engineer and refactor the design and code.
- **Total re-engineering:** *Pro:* It gives you the best quality. *Con:* It can take years to complete and comes at a high upfront cost.

Planning the Migration Journey

When considering a migration journey, carefully plan each step along the way. The journey can be depicted as a set of steps (see Figure 5-1). Here's how each step works:

1. **The discovery phase: Ask internal questions.**

   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. **The assessment phase: Make a migration assessment.**

   Refine and evaluate the solutions on the table. Take an inventory of all migration items and prioritize the use cases. When you complete the migration assessment, you'll have a clearer sense of your timeline and alignment with your original planned schedule.

3. **The strategy phase: Conduct technical planning.**

   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. **The production pilot phase: Complete evaluation and enablement.**

   Understand what your new platform has to offer. Conduct targeted demos or plans to help vet your approach.

5. **The execution phase: Execute your migration.**

   The rubber meets the road — make sure you get this migration right the first time.

REMEMBER

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

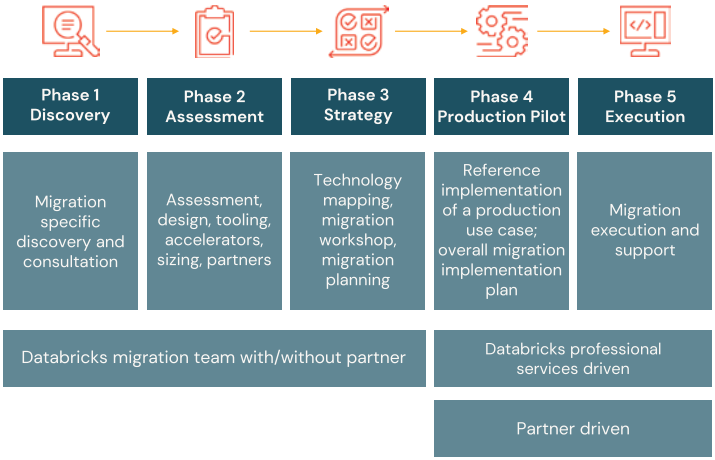

FIGURE 5-1: The phases of migration methodology.

The Five Pillars of Migration

The underlying architecture for the lakehouse follows the five pillars of one framework:

- **Architecture/infrastructure:** Establish the deployment architecture and implement a security and governance framework.
- **Data migration:** Map data structures and layout, complete a one-time load, and finalize an incremental load approach.
- **ETL and pipelines:** Migrate data transformation and pipeline code, orchestration, and jobs in this phase. Speed up your migration by using automation tools and comparing your results with on-premises data and expected results.
- **Analytics:** Repoint reports and analytics for business analytics and business outcomes. Reporting semantic layers and online analytics processing (OLAP) cubes should also repoint to the lakehouse via Open Database Connectivity (ODBC) and Java Database Connectivity (JDBC).
- **Data science/machine learning (ML):** Establish connectivity to ML tools and onboard data science teams.

IN THIS CHAPTER

- Future-proofing your investment
- Extracting the freshest data from anywhere
- Using one source of truth for all your data

Chapter 6
Ten Reasons to Migrate to the Databricks Lakehouse

When you're making your decision to migrate to the lakehouse, Databricks gives you ten reasons to choose the Databricks Lakehouse as your platform. These reasons are

- **You want to have an open and flexible platform.** Platform lock-in is the root cause of migrations every few years. You can avoid vendor lock-in and future-proof your investment by using an open, multi-cloud, highly innovative lakehouse platform that works seamlessly with the modern data stack. Check out Chapter 3 for more on future-proofing.
- **You want to realize faster time to insights.** Enable your business with near real-time, self-served analytics for all, and uncover new insights on your freshest and most complete data.
- **You want to lower operational costs and establish one single governance layer for all of your data.** Establish a unified architecture and governance model.

- **You want the best price/performance.** Operating a lakehouse architecture provides up to 12 times better price/performance than other cloud data warehouses.
- **You want to easily ingest data from anywhere and access the freshest data.** Databricks SQL works with your data, no matter where it is. Databricks has autoloader capabilities for seamless file ingestion and many ingestion partner tooling integrations built-in with PartnerConnect, that provides turnkey capabilities to ingest data from cloud storage and enterprise data to enterprise applications such as Salesforce or Marketo. It's just one click away.
- **You need modern analytics with your tools of choice.** Databricks SQL works seamlessly with the most popular business intelligence (BI) and SQL tools, such as dbt, Tableau, Power BI, and Looker. As a result, analysts can use their favorite tools to discover new business insights on the most complete and freshest data.
- **You experience first-class SQL development experience.** Databricks SQL query editor allows analysts to write queries in a familiar syntax (ANSI SQL) and easily explore data in place in the lakehouse. Analysts can easily make sense of query results through a wide variety of rich visualizations and quickly build and share dashboards with stakeholders.
- **You eliminate infrastructure management.** You experience lower costs and eliminate the need to manage, configure, or scale cloud infrastructure with serverless SQL compute. This frees up your data team to do what they do best.
- **You're able to practice fine-grained governance on the lakehouse.** You can confidently manage and secure data access on your lakehouse with fine-grained governance. In addition, you can meet compliance needs with data lineage, role-based security policies, and table or column level tags for all data assets.
- **You have one source of truth for all your data.** Unlike enterprise data warehouses (EDWs), the Databricks Lakehouse Platform provides one common storage and data management framework for all data types on your existing data lake.

TIP

For Databricks overall performance benchmark, check out the following link: `databricks.com/product/databricks-sql`.

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 →

Evolve your data strategy with the lakehouse

Data is the most valuable resource for modern businesses. As data availability continues to surge, companies are struggling to keep up with the work required to manage massive amounts of data across multiple data sources. 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...

- How to manage unique data types
- Assess your company's data strategy
- The benefits of the lakehouse
- 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 eBooks, including *Facebook Marketing For Dummies*.

ISBN: 978-1-119-89472-8
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.