

eBook

Transforming Legacy Data Warehouses: A Strategic Migration Blueprint

Proven strategies to achieve seamless transitions for scalability, cost efficiency and AI innovation



Contents

Introduction.....

Why migrate your data warehouse?

Proven steps to a successful data warehouse migration.....

Step 1: Assess your legacy data warehouse

Step 2: Define Your target platform.....

Step 3: Choose a migration strategy

Step 4: Validate the data warehouse and migration capabilities.....

Step 5: Implement, optimize and monitor

Conclusion.....

3

4

6

7

8

10

14

16

17

Introduction

Organizations must continuously modernize their data infrastructure to stay competitive. Traditional enterprise data warehouses (EDWs) often struggle with scalability, flexibility and high costs, making migration to a modern data platform an attractive option. However, migration is a complex process that requires careful planning and execution to ensure minimal disruption and maximum business value.

This eBook describes the essential steps to migrating your data warehouse, including best practices, strategies and key considerations. Following a structured approach, businesses can complete a low-risk, predictable migration to a fast, open data warehouse while preparing for innovative capabilities like enhanced performance, real-time analytics, artificial intelligence (AI) and machine learning (ML).



Why migrate your data warehouse?

Migrating to a modern data platform offers several advantages that can significantly impact business operations and decision-making:

Scalability and performance

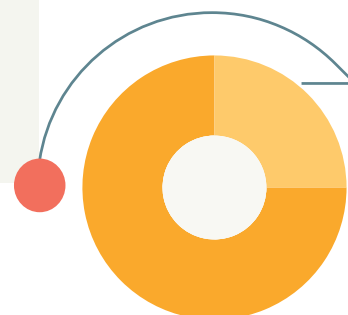
Cost efficiency







Advanced analytics and AI

Governance and security

Flexibility and agility

Open source technologies



-  **Scalability and performance:** Traditional EDWs often require expensive hardware upgrades to handle increased workloads. Ordering this hardware, racking and stacking it in data centers sometimes takes months before you can initiate a new data project. Cloud-based solutions provide elastic scalability instantaneously, allowing businesses to scale up or down as needed without significant capital investment.
-  **Cost efficiency:** Cloud-native data warehouses typically operate on a pay-as-you-go model, eliminating the need for large upfront costs. Optimized storage and computing resources also contribute to overall cost savings.
-  **Advanced analytics and AI:** A modern data warehouse supports cutting-edge analytics, enabling real-time data processing, AI-powered insights and predictive modeling.
-  **Data governance and security:** Modern platforms include built-in security features such as encryption, access controls and compliance tools for regulatory requirements.
-  **Flexibility and agility:** Unlike rigid legacy systems, cloud-based platforms enable businesses to quickly adapt to new use cases and integrate diverse data sources.
-  **Using open source technologies and avoiding vendor lock-in:** Open source technologies are built on open standards, allowing you to mix and match best-of-breed tools without being tied to a single vendor. Then you can tailor your architecture to your needs and easily integrate with other systems. Getting locked into one proprietary EDW after another is the root cause of migrations every few years.

Proven steps to a successful data warehouse migration



STEP 1



ASSESS CURRENT ENVIRONMENT

Assess your current data warehouse to identify potential challenges, optimize strategies and ensure a smooth transition.

Before embarking on the migration journey, conducting a thorough assessment of your existing data environment is essential. This evaluation helps identify potential challenges, optimize migration strategies and ensure a smooth transition.

We complete the assessment at three different levels:

Business assessment

- Identify key business use cases that rely on the data warehouse. This exercise is key to building the target architecture and choosing the right features to implement these new uses.
- Work with stakeholders, such as data analysts, engineers and IT teams, to understand current usages and future needs. The Databricks' account team will lead this while working closely with your teams.
- Assess the current system's ability to support emerging trends like real-time analytics, machine learning and self-service BI.

Architecture evaluation

- Document the existing data architecture, including data models, storage methods and integration points.
- Identify dependencies between applications, databases and workflows.
- Understand how data flows from ingestion to transformation and consumption.

Technical assessment

- Evaluate performance metrics, including query execution times, data refresh rates and system uptime.
- Analyze the complexity of ETL (Extract, Transform, Load) processes and the dependencies on legacy systems.
- Identify security risks, compliance requirements and access control mechanisms.

STEP 2



DEFINE TARGET PLATFORM

Define your target platform by designing the future state of your data warehouse. It is OK to set current goals small, but talk about future end-state.

After the assessment is complete, the next step is designing the future state of your data warehouse. This step involves selecting the right platform and defining an architecture that aligns with business goals.

You must align with the target business and technical goals with your partner.

Business assessment

- Ensure the new platform meets strategic business objectives such as faster reporting, reduced operational costs and improved decision-making.
- Define the key performance indicators (KPIs) that will measure the success of the migration.

Modernization goals

- Decide whether to execute a simple lift-and-shift migration, optimize existing processes or fully modernize data workflows. Let's examine what this means in detail.

After you set your goals, we will assess the source platform, which will help you choose the best migration approach among the following:



A lift and shift migration: this approach will be the fastest way to decommission the existing platform. It will maximize the usage of automated conversion tools like BladeBridge (now a part of Databricks). We will maintain Ingestion phases, data sources, data models and exposition layers in the existing data warehouse. One of the main drawbacks resides in adding new business cases and usages, which will remain very limited.



A full modernization. We can modernize the data platform at various levels. It could be a new greenfield implementation of the data warehouse that will reimplement the existing logic and add some new use cases (e.g., machine learning-based use cases that will enable more predictive analytics). It could be a change in the data architecture to add some new data sources to modernize the existing data, etc. Solutions are multiple and will mostly depend on the assessment results and your new usage of the latest data platform. Modernizing will typically take longer, but it will leverage all the opportunities to implement a complete modern data platform beyond using the data warehouse.

We can also choose something in between. For example, if you want to quickly migrate using a lift and shift and then modernize by adding new use cases, you can update the architecture during the migration phase to allow it to accept new use cases in the future. We call this a hybrid approach.

Technology stack selection

- Consider adopting a cloud-native architecture that leverages serverless computing, automated data pipelines and AI-driven insights.
- Evaluate data storage solutions such as Delta Lake Uniform, columnar storage and distributed computing frameworks for better scaling capabilities and interoperability
- Explore **our ISV partners** to enhance your stack:
 - Ingestion partners: Qlik, Fivetran etc.
 - ETL partners: Informatica, Prophecy etc.
 - Reporting and semantic layers partners: Microsoft PowerBI, Tableau, AtScale etc.

STEP 3



CHOOSE A STRATEGY

Choose a migration strategy: ETL first, or BI first.

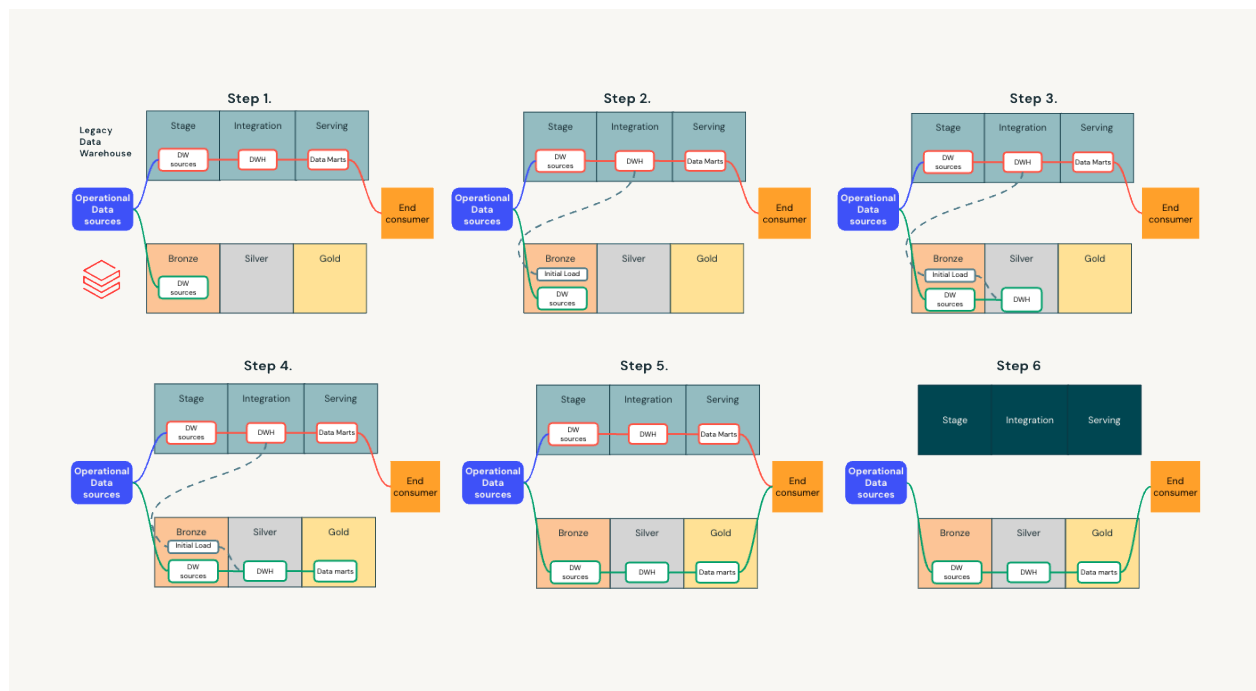
After we have assessed the current platform and designed the target platform, it's critical to select the right migration strategy to minimize downtime and ensure data consistency. Here are the common approaches:

- **Ingestion and ETL first migration**
- **BI-first migration**

Ingestion and ETL first migration (back-to-front)

This strategy involves opening a new branch from the data sources and ingesting them into the Lakehouse. Once ingested, we gradually implement the ETL logic to build a new Data Warehouse. Optionally, we could create an operational data store to support tactical decisions. Then, we implement the serving layer with the data marts in the data warehouse. Finally, after the serving layer has been built and is ready to use, we repoint existing dashboards and reports to the new platform, or they could also be migrated or modernized to a new cloud-based dashboarding solution.

This strategy's main advantages include ensuring a solid data foundation and allowing gradual optimization. However, business users may not see immediate benefits of this migration, which could slow adoption.

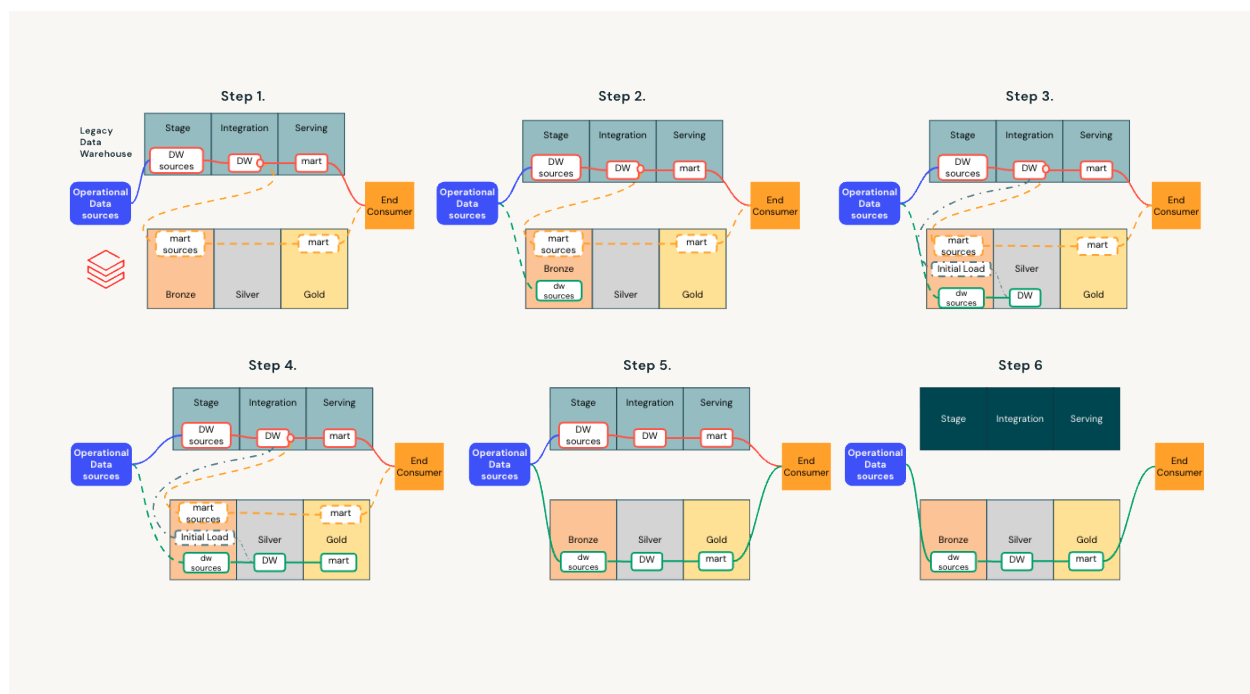


BI-first migration (front-to-back)

The other strategy for implementing your data warehouse modernization to Databricks consists of replicating the serving layer as a first step. This first exposes the data sets tailored to the specific needs of a particular business function and then implements the modernization of the ingestion and ETL afterward.

By doing this and using Lakehouse Federation, it's possible to quickly make the benefits visible to business users, increasing confidence in the new system. It also allows quickly modernizing (for example, "creating new data products, building a semantic layer in Databricks or leveraging Genie spaces to expand capabilities to query data and simplify the decision making process using natural language)."

However, this strategy must manage multiple layers simultaneously, requiring a strict plan.



Databricks' features to help accelerate the migration process

These strategies are possible by leveraging multiple Databricks features tailored to support the migration of legacy data warehouses:



Lakeflow Connect will support the ingestion of data coming from legacy database systems but also by ingesting new data coming from modern SaaS platforms like Salesforce or Workday.



Databricks Workflows to support complex pipelines, streaming ingestion, dependency management, version control etc.



Streaming tables and Materialized views are used to automatically refresh the serving layer and implement real-time decision-making scenarios.



Lakehouse Federation to easily and quickly replicate data marts coming from the legacy data warehouse to Databricks.



Unity catalog to ensure the governance of the modernized platform, but also expand capabilities to data monetization of customers' assets.



BladeBridge to support automatic code conversion coming from the legacy (e.g., Teradata BTEQ, Oracle stored procedures, Third party ETL pipelines).



Genie, Mosaic AI and Mlflow to enable predictive and prescriptive analytics scenarios.

STEP 4**VALIDATE BEFORE FULL-SCALE MIGRATION**

Validate the data warehouse and migration capabilities to identify potential risks.

We highly recommend running some evaluation procedures before executing a full-scale migration to validate assumptions and identify potential risks.

There are two types of procedures: the proof of concept (POC) or minimal viable product (MVP).

The POC

During a POC, it's essential to validate the feasibility of a specific solution or technology. A POC will answer "Can we do it?". For example, a POC will validate the cost/performance of a query that requires constant performance tuning. It could validate the capability to transform a strategic ETL pipeline to Databricks, etc.



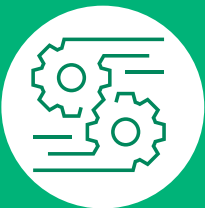
The MVP

An MVP will be a functional version of the migrated data warehouse that includes the essential features or business cases necessary to meet the needs of early users. A MVP will answer the question, “Should we do it?”. In a migration context, an example of MVP will reproduce the whole migration process on a subset of existing use cases and potentially new ones.

Whether POC or an MVP is required, it’s important to define clear success criteria, which can vary depending on the business, customer’s expectations, SLAs etc. Running these processes will give confidence in the migration process and the Databricks’ capabilities to deliver the best cost/performance ratio and the highest business value realization.



STEP 5



IMPLEMENT, OPTIMIZE AND MONITOR

Implementation includes optimizing and monitoring your new Databricks SQL data warehouse.

With a set of successful POCs optionally completed by an MVP, the final step is the full-scale implementation. Implementation involves:

- **Data governance and security:** Enforcing policies for access control, data lineage and compliance.
- **Performance optimization:** Implementing indexing, caching, partitioning and auto-scaling mechanisms.
- **Monitoring and maintenance:** Continuously tracking system performance, optimizing costs and addressing issues proactively.

A strong testing policy will complete all of these operations, whether it concerns unit tests that validate the data ingested has been well transformed and produced the same data quality as in the previous environment or compliance tests of the overall process. Various stakeholders, such as Databricks, the customer or the partner, will define this testing phase.

Conclusion

Migrating to a modern data warehouse is a complex but rewarding journey that can transform an organization's ability to harness data effectively. By following a structured approach and leveraging best practices, businesses can achieve enhanced performance, lower costs and unlock new data-driven opportunities.

Contact your Databricks representative or reach out to us using this [form](#) for expert guidance on your data warehouse migration. Our specialists are ready to help you every step of the way.



About Databricks

Databricks is the data and AI company. More than 10,000 organizations worldwide — including Block, Comcast, Condé Nast, Rivian, Shell and over 60% of the Fortune 500 — rely on the Databricks Data Intelligence Platform to take control of their data and put it to work with AI. Databricks is headquartered in San Francisco, with offices around the globe, and was founded by the original creators of Lakehouse, Apache Spark™, Delta Lake and MLflow.

To learn more, follow Databricks on [LinkedIn](#), [X](#) and [Facebook](#).

[Start your free trial](#)

