Technical Migration Guide

Strategies to Evolve Your Data Warehouse to the Databricks Lakehouse

databricks

Contents

Lakehouse Architecture

Data warehouses were designed to provide a central data repository with analytic compute capabilities to help business leaders get analytical insights, support decision-making and business intelligence (BI). Legacy on-premises data warehouse architectures are difficult to scale and make it difficult for data teams to keep up with the exponential growth of data. Oftentimes data teams publish and use a subset of well-defined data for development and testing. This slows down both innovation and time to insight.

Cloud data warehouses (CDW) were an attempt to tackle the on-premises data warehouse challenges. CDWs removed the administrative burden of tasks such as setup, upgrades and backups. CDWs also improved scalability and introduced cloud's pay-as-you-go model to reduce cost. CDWs leverage a proprietary data format to achieve cloud-scale and performance; however, this also leads to customers locked into these formats with difficult paths to support use cases outside the data warehouse itself (i.e., machine learning). Customers often find themselves with a bifurcated architecture, which ultimately leads to a more costly and complex data platform over time.

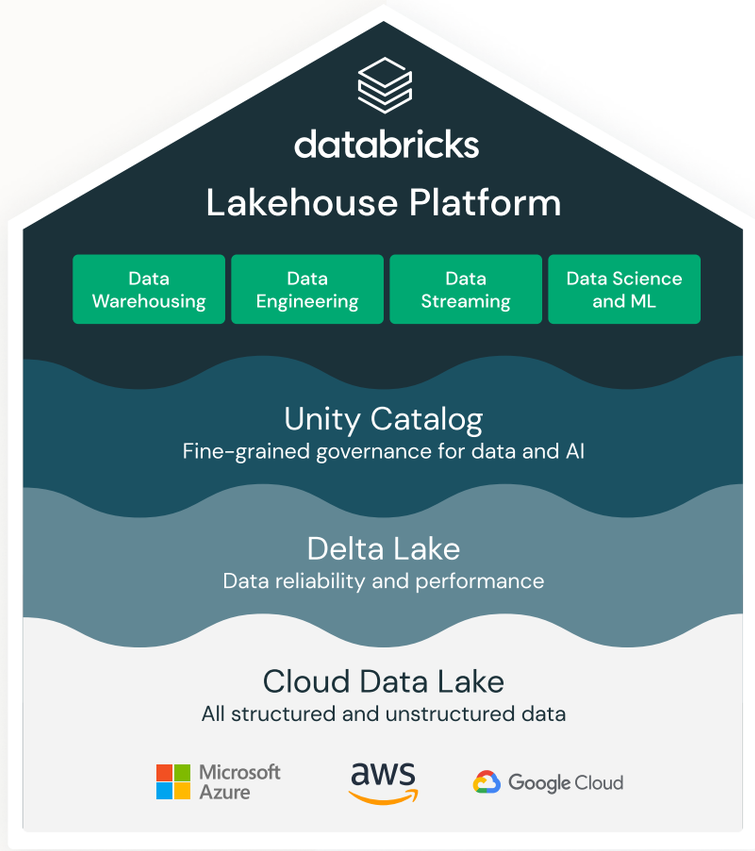
But enterprise data teams don't need a better data warehouse. They need an innovative, simple solution that provides reliable performance, elastic scale and allows self-service to unblock analytics to access all data at a reasonable cost. The answer is the lakehouse.

The lakehouse pattern represents a paradigm shift from traditional on-premises data warehouse systems that are expensive and complex to manage. It uses an open data management architecture that combines the flexibility, cost-efficiency and scale of data lakes with the data management and ACID semantics of data warehouses. A lakehouse pattern enables data transformation, cleansing and validation to support both business intelligence and machine learning (ML) users on all data. Lakehouse is cloud-centric and unifies a complete up-to-date data set for teams, allowing collaboration across an organization.

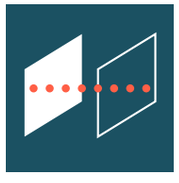
The Databricks Lakehouse Platform

The Databricks Lakehouse Platform is **simple**; it unifies your data, governance, analytics and AI on one platform. It's **open** — the open source format Delta Lake unifies your data ecosystem with open standards and data formats. Databricks is **multicloud** — delivering one **consistent experience across all clouds** so you don't need to reinvent the wheel for every cloud platform that you're using to support your data and AI efforts.

Databricks SQL stores and processes data using Delta Lake to simplify and enhance data warehousing capabilities. Analysts can use their favorite language, SQL, popular transformation tools such as dbt, and preferred BI tools like Power BI and Tableau to analyze data. The built-in query editor reduces contextual switching and improves productivity. Administrators enjoy simplified workload management via serverless compute and auto-scaling to meet high-concurrency workload needs. All this at a fraction of the cost of traditional data warehouses.




Simple


Open

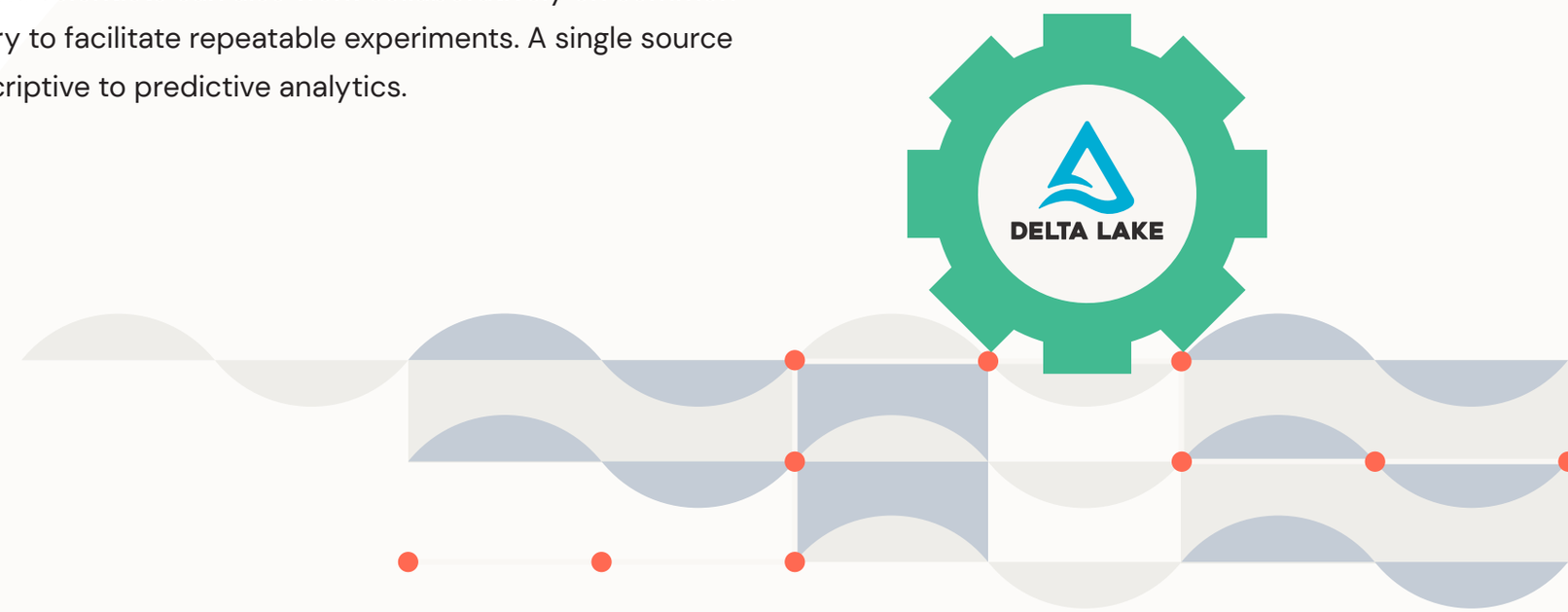

Multicloud

Business Value

Single source of truth

Databricks Delta Lake leverages cloud-based blob storage to provide an infinitely scalable storage layer where you can store all your data, including raw and historical data, alongside structured data tables in the data warehouse. The lakehouse pattern avoids data silos and shares the same elastic scale and governance across all use cases: BI, data engineering, streaming and AI/ML. This means that data engineering teams don't have to move data to a proprietary data warehouse for business analysts or create a separate data store to support data science.

Instead, data teams can access the open format Delta tables directly and combine data sets in the lakehouse, as needed. Data scientists can also work collaboratively on common data with access to versioned history to facilitate repeatable experiments. A single source of truth facilitates moving from descriptive to predictive analytics.



Data team

With central data governance and fine-grained access control capabilities to secure the lakehouse, you can enable self-service SQL analytics for everyone on the Databricks Lakehouse Platform. This allows each team to be more agile and innovate faster.

Data Analysts — Using the Databricks SQL editor or their tools of choice (DBT, Power BI, Tableau), SQL analysts can leverage familiar toolsets.

Data Engineers — Utilizing Delta Lake as a unified storage layer, data engineering teams can eliminate duplicate data and ETL jobs that move data across various systems. Databricks supports both batch and streaming workloads to reduce bottlenecks and serve the most up-to-date data to downstream users and applications.

Administrators — The pay-as-you-go, decentralized compute resource allows each team to run their workload in isolated environments without worrying about contention. Serverless SQL endpoint frees your team from infrastructure management challenges.

The Databricks Lakehouse Platform provides a reliable ETL and data management framework to simplify ETL pipelines. Data teams can build end-to-end data transformations in a single pipeline instead of many small ETL tasks. Databricks supports data quality enforcement to ensure reliability with auto-scalable infrastructure. Your teams can onboard new data sources quickly to power new use cases with fresh data. This not only allows your team to efficiently and reliably deliver high-quality data in a timely manner, it also reduces ETL workload cost significantly.

Future-proof

Unlike CDWs that lock customers in, Databricks offers an open platform with open standards, open protocols and open data formats. It supports a full range of popular languages (SQL, Python, R, Scala) and popular BI tools. You can leverage the performant and low-cost distributed compute layer for data processing — or use a variety of tools and engines to efficiently access the data via Databricks APIs. Databricks also allows data consumption with a rich partner ecosystem. Teams can handle all existing BI and AI use cases with the flexibility to support future use cases as they emerge.

Migration to Lakehouse

Overview

A lakehouse is the ideal data architecture for data-driven organizations. It combines the best qualities of data warehouses and data lakes to provide a single solution for all major data workloads and supports use cases from streaming analytics to BI, data science and AI. The Databricks Lakehouse Platform leverages low-cost, durable cloud storage and only consumes (charges for) compute resources when workloads are running. This pay-as-you-go model means compute resources are automatically shut down if no processing is needed. Data teams can use small clusters that can power individual workloads they plan to migrate. They can make the choice to leverage serverless SQL endpoints and completely free data teams from infrastructure capacity planning and cluster maintenance. The auto-scaling, elastic nature of Databricks clusters leads to significant savings on infrastructure cost and maintenance. Organizations typically achieve 50% TCO savings compared to other cloud data warehouses.

Data warehouse migration is never an easy task. Databricks aims to mitigate the things that can go wrong in these demanding migration projects. The Databricks Lakehouse Platform provides many out-of-the-box features to mitigate migration risks.

CUSTOMER STORY

Building the Lakehouse at Atlassian

Watch now →

CUSTOMER STORY

Driving Freight Transportation Into the Future

Read more →

Migration strategy

Migration is a huge effort and very expensive. Yet, almost every enterprise has to migrate to new platforms every 3–5 years because the old platform cannot support new use cases, catch up with data growth or meet scaling needs. To get better ROI on migration, implement a migration strategy that can reduce future re-platform needs and extend to your future data and AI strategy.

Use the opportunity of a data migration to standardize your data in open Delta format to allow existing and future tools to access it directly without moving or converting it. Merge your siloed data warehouses into the unified storage layer in the Databricks Lakehouse Platform — without worrying about storage capacity.

The unified storage layer allows your team to deploy a unified data governance on top to secure all data access consistently. Simplify your data governance story with Databricks Unity Catalog.

Move toward a single, consistent approach to data pipelining and refinement. Merge batch and streaming into a single end-to-end pipeline to get fresher data and provide more real-time decisions. Take a metadata-driven approach to align the dataflow with business processes and have data validation and quality check built-in. Through a series of curation and refinement steps, the output results in highly consumable and trusted data for downstream use cases.

The lakehouse architecture makes it possible for the organization to create “data assets” by taking a stepwise approach to improving data and serving all essential use cases. Encourage your BI/analyst team to leverage Databricks serverless endpoints for self-serve and agility. Each team can evaluate their top priority workloads and migrate them in parallel to speed up migration.

Take advantage of Databricks’ rich partner ecosystem. Your favorite partners are likely already integrated via Partner Connect and can be set up with a few clicks. There are also many ISV and SI consulting partners who can help your migration journey.

Migration planning

Migrating a data warehouse to the cloud can be time consuming and challenging for your data teams. It's important to agree on the data architecture, migration strategy and process/ frameworks to be used before undertaking a data migration. Databricks provides Migration Assessment and Architecture Review sessions to develop a joint migration roadmap. This process is designed to help organizations to successfully migrate to a lakehouse architecture. Based on information collected and business objectives, the Databricks team will work with customers to propose a target architecture and provide a tailored migration roadmap.

These assessments help get a full picture of current data systems and the future vision. They clarify what you are migrating and do proper use case discovery. This includes identifying workloads and data source dependency, for example:

Sample migration assessment checklist:

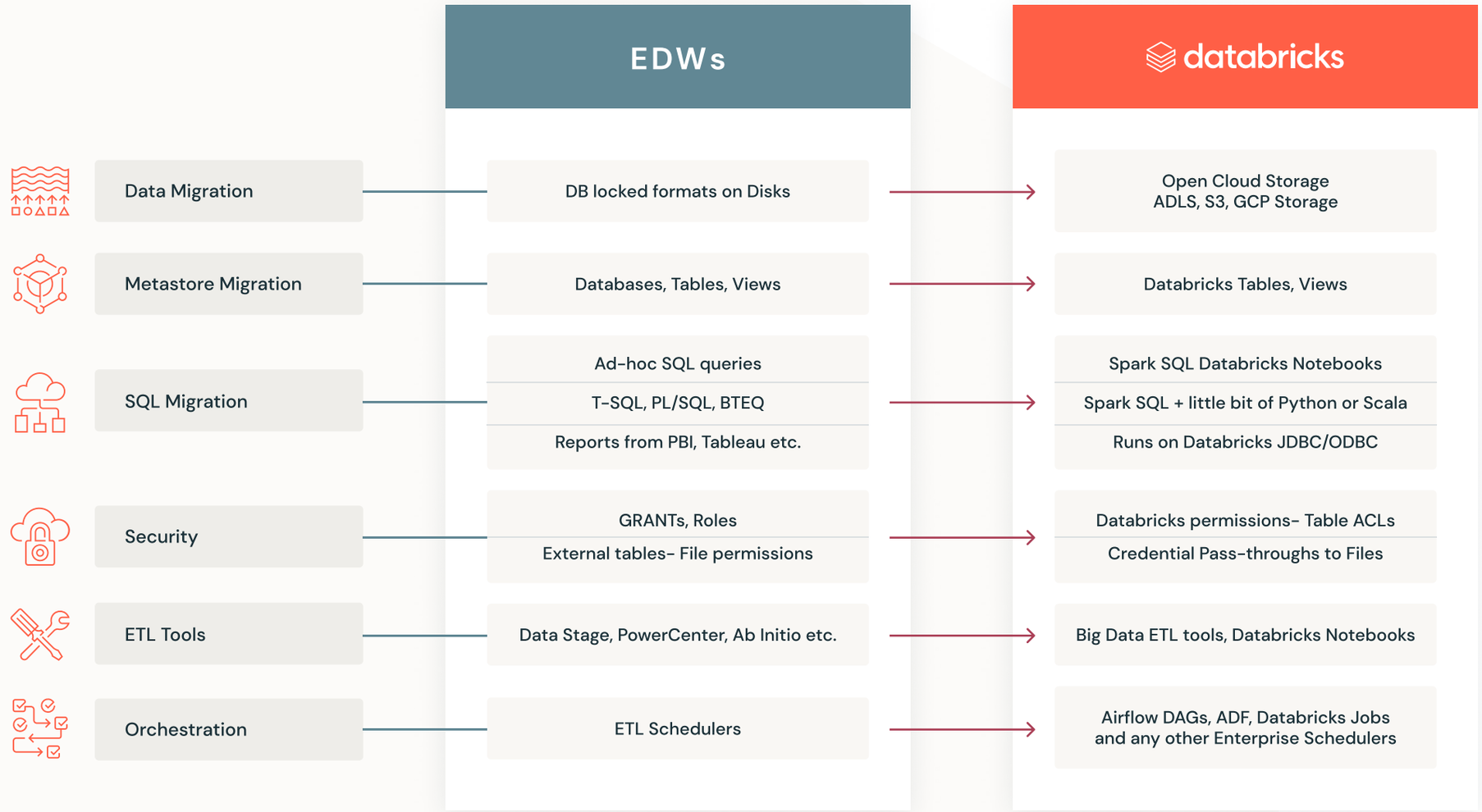
- [ ] Identify upstream data sources and workload dependencies
- [ ] Identify active/inactive data sets and database objects
- [ ] Identify downstream application dependencies and data freshness requirements
- [ ] Define a cost-tracking mechanism, such as tag rules for chargeback and cost attribution
- [ ] Define security requirements and data governance
- [ ] Clarify access management need, document needed permissions per user/group
- [ ] Outline current tooling (ingestion, ETL and BI) and what's needed

It's important to identify key stakeholders and keep them engaged during the migration to make sure they are aligned with the overall objectives. The workload assessment result will be reviewed with key stakeholders. Through the review process, data teams can get a better understanding of which workloads can most benefit from modernization.

Databricks often works with partners to provide a workload assessment and help customers understand their migration complexity and properly plan a budget. Databricks also partners with third-party vendors that provide migration tools to securely automate major migration tasks. Databricks Partner Connect makes it easy to connect with this ecosystem of tools to help with the migration, including:

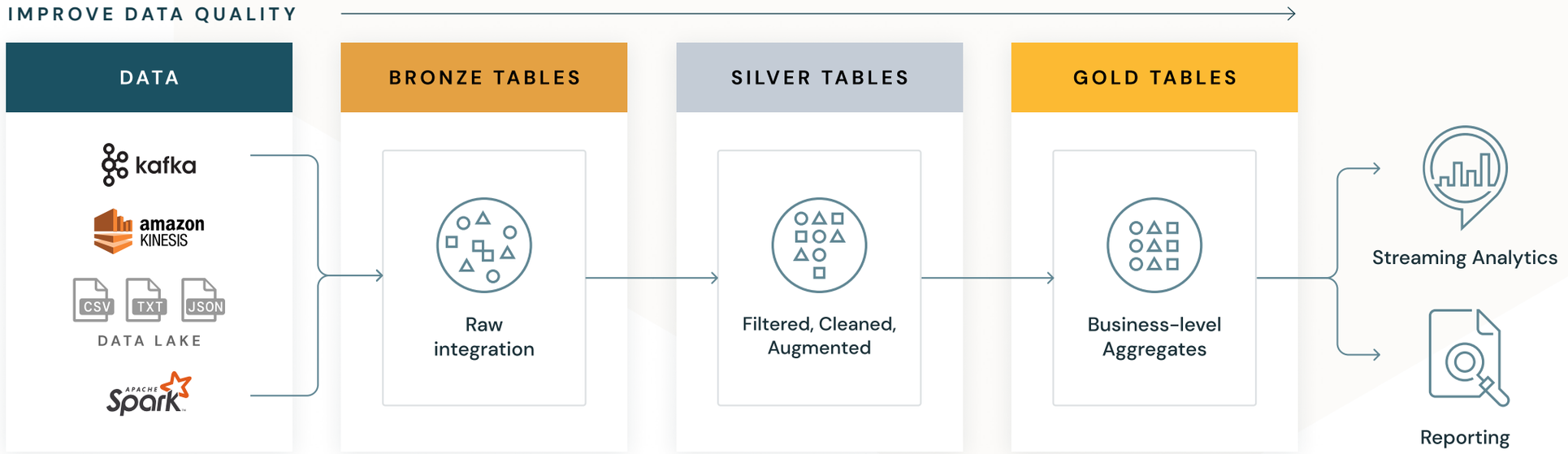
- Code conversion tooling that can automatically translate 70%–95% of the SQL code in your current system to Databricks optimized code with Delta and other best practices
- Converters that automate multiple GUI-based ETL/ELT platform conversion to reduce migration time and cost
- Data migration tools that can migrate data from on-premises storage to cloud storage 2x–3x faster than what was previously possible

We can use Automated conversion for most workload types



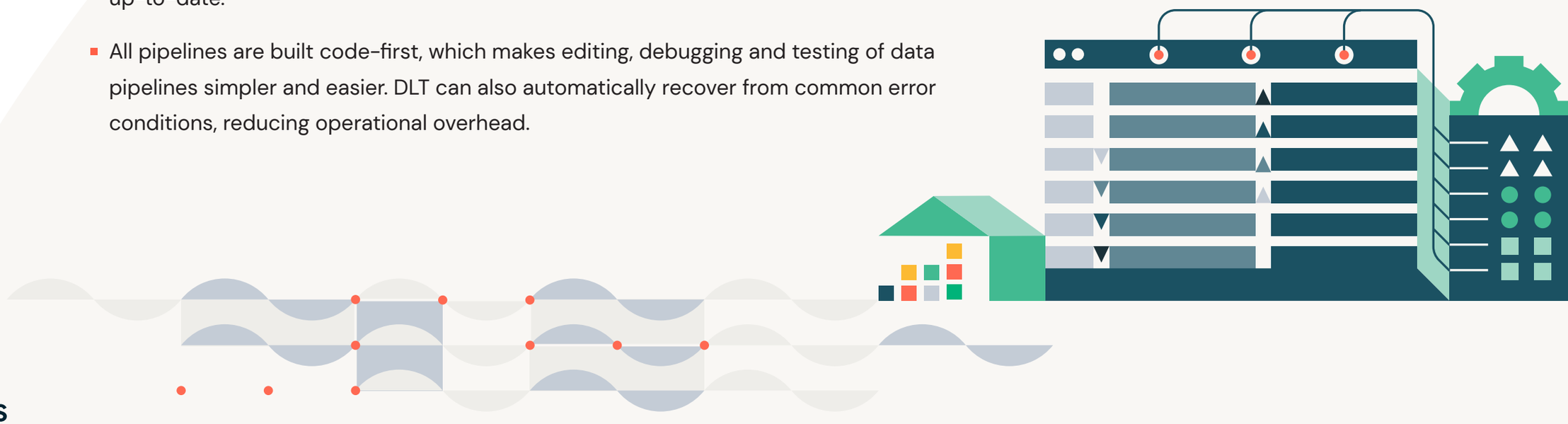
ELT approach

The separation of storage and compute makes ELT on lakehouse a better choice than traditional ETL. You can ingest all raw data to Delta Lake, leverage low-cost storage and create a Medallion data implementation from raw/Bronze to curated/Gold depending on what's needed to support use cases. During ingestion, basic data validation can occur, but establishing a Bronze data layer is the foundation of a single-pane-of-glass for the business. Teams can leverage compute resources as needed without a fixed compute infrastructure. Establishing a Silver layer further enriches data by exploring and applying transformations. ELT allows data teams to break pipelines into smaller "migrations," starting with a simple workload, then improving the pipeline design iteratively.



We highly recommend leveraging Delta Live Tables (DLT), a new cloud-native managed service in the Databricks Lakehouse Platform that provides a reliable ETL framework to modernize your data pipeline at scale. Instead of migrating multiple ETL tasks one by one in a traditional data warehouse, you can focus on source and expected output, and create your entire dataflow graph declaratively. Delta Live Tables offers:

- A metadata-driven approach — You just specify what data should be in each table or view rather than the details of how processing should be done
- An end-to-end data pipeline with data quality and freshness checks, end-to-end monitoring/visibility, error recovery, and lineage, which reduces the strain on data engineering teams and improves time-to-value in building data pipelines
- Automatic management of all the dependencies within the pipeline. This ensures all tables are populated correctly, whether continuously or on a regular schedule. For example, updating one table will automatically trigger all downstream table updates to keep data up-to-date.
- All pipelines are built code-first, which makes editing, debugging and testing of data pipelines simpler and easier. DLT can also automatically recover from common error conditions, reducing operational overhead.



Agile modernization

Agile development allows teams to move quickly knowing migrated pipelines can be revisited at a later cycle and evolving data models are supported within the architecture. Allowing business impact to drive priorities via an agile approach helps mitigate migration risks. Prioritizing and selecting use cases where modernization brings business benefits quickly is a good starting point. Focus on the 20% of workloads that consume 80% of budget. By breaking workflows down into components and managing data stories, teams can adjust priorities over time. Changes can be made in collaboration with the user community to fit the business definition of value.

Migrating to a lakehouse architecture leverages separation of storage and compute to remove resource contention between ETL and BI workloads. As a result, the migration process can be more agile, allowing you to evolve your design iteratively without big-bang effort:

- Reduce time during the initial phase on full capacity plan and scoping
- Flexible cloud infrastructure and unlimited, autoscaling storage
- Workload management is much simpler, you can isolate each workload with a dedicated compute resource, without worrying about managing workload contention
- Auto-scale and tear down the compute resources after the job is done to achieve cost efficiency

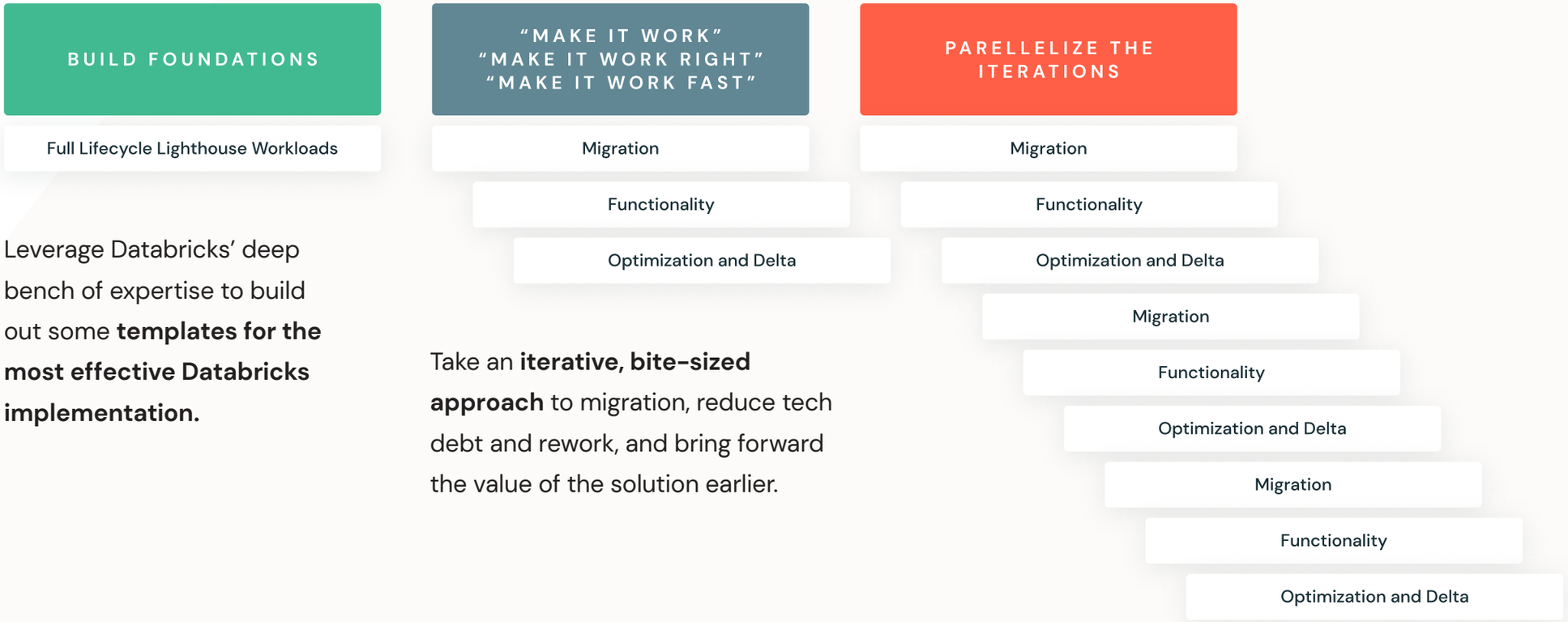
All of this allows you to take a more iterative and business-focused approach for migration instead of a full planning, execution, test/validation approach. Here are more approaches that help facilitate this phased implementation:

- Leverage Databricks Auto Loader. Auto Loader helps to ingest new data into pipelines quicker to get data in near real-time.
- Delta Live Tables (DLT) improves data quality during data transformation and automatically scales to address data volume change. DLT can also support schema evolution and quarantine bad data or data that needs to be reprocessed at a later stage.
- Use dedicated clusters to isolate workloads, lower the total cost of ownership and improve overall performance. By using multiple clusters, we can shut down resources when not in use and move away from managing fixed resources in a single large cluster.

Leverage Databricks' deep bench of expertise to build reusable assets along the migration:

- Create a migration factory for iterative migration process
- Determine and implement a security and governance framework
- Establish a to-be environment and move use cases/workloads in logical units
- Prove business value and scale over time
- Add new functionality continuously so important business requirements are not left on hold during migration

Take this iterative and templated approach. Migration speed will accelerate. Customers can finish migration 15%–20% faster and reduce the amount of tech debt created during the migration.



Leverage Databricks' deep bench of expertise to build out some **templates for the most effective Databricks implementation.**

Take an **iterative, bite-sized approach** to migration, reduce tech debt and rework, and bring forward the value of the solution earlier.

To maximize the value of your lakehouse, you should consider retiring some legacy architecture design patterns. Leverage the migration process to simplify data warehousing tasks. Regardless of how you complete your migration, you could utilize lakehouse strengths to improve architectural patterns:

- Merge your siloed data warehouses on your unified lakehouse platform and unify data access and data governance via Unity Catalog. The lakehouse architecture provides a unified storage layer for all your data where there is no physical boundary between data. There is no need to keep data copies for each system using the data set. Clean up and remove jobs that are created to keep data in sync across various data systems. Keep a single copy of raw data in your lakehouse as a single source of truth.
- The Databricks Lakehouse Platform allows you to merge batch and streaming into a single system to build a simple continuous data flow model to process data as it arrives. Process data in near real-time and enable data-driven decisions with the most recent updates.
- Simplify your workload isolation and management by running jobs in dedicated clusters. Separating storage and compute allows you to easily isolate each task with isolated compute resources. There is no need to squeeze them into a single large data appliance and spend lots of time managing and coordinating resources. Leverage the elasticity of the Databricks compute layer to automatically handle workload concurrency changes at peak time instead of paying for over-provisioned resources for most of the time. This greatly simplifies the workload management effort the traditional data warehouses require.
- Simplify disaster recovery. Storage and compute separation allows easy disaster recovery. The cloud storage provides very good data redundancy and supports automated replication to another region. Customers can spin up compute resources quickly in another region and maintain service availability in case of an outage.

Security and data governance

Security is paramount in any data-driven organization. Data security should enforce the business needs for both internal and external data, so the lakehouse should be set up to meet your organization's security requirements. Databricks provides built-in security to protect your data during and after migration.

- Encrypt data at rest and in-transit, using a cloud-managed key or your own
- Set up a custom network policy, use IP range to control access
- Leverage Private Link to limit network traffic to not traverse the public internet
- Enable SSO, integrate with active directory and other IdPs
- Control data access to database objects using RBAC
- Enable audit logs to monitor user activities

The challenge with the traditional data warehouse and data lake architecture is that data is stored in multiple stores and your data team also needs to manage data access and data governance twice. The lakehouse pattern uses unified storage which simplifies governance. The Databricks Lakehouse Platform provides a unified governance layer across all your data teams. Migrating to Databricks Unity Catalog provides data discovery, data lineage, role-based security policies, table or row/column-level access control, and central auditing capabilities that make the data platform easy for data stewards to confidently manage and secure data access to meet compliance and privacy needs, directly on the lakehouse.


Audit Log
Metastore
Account Level
User Management
Lineage Explorer
Credentials
Centralized
Governance
Data Explorer
ACL Store
Access Control

Team involvement

Plan to educate and train your team iteratively throughout the migration process. As new workloads are migrated, new teams will gain exposure to the lakehouse pattern. Plan to ramp up new team members as the migration process progresses, developing a data Center of Excellence within the organization. Databricks provides a cost effective platform for ad hoc work to be performed. A sandbox environment can be leveraged for teams to get exposure to Databricks technology and get hands-on experience. Databricks also provides learning path training for customers. Encourage teams to get hands-on experience relevant to their immediate tasks, gain exposure to new things and try new ideas.

Conclusion

Data warehouse migration touches many business areas and impacts many teams, but the Databricks Lakehouse Platform simplifies this transition, reduces risks and accelerates your ROI. The Databricks Business Value Consulting team can work with you to quantify the impact of your use cases to both data and business teams. And the Databricks team of solution architects, professional services, and partners are ready to help.

Reach out to your Databricks account team or send a message to sales@databricks.com to get started.

Additional resources

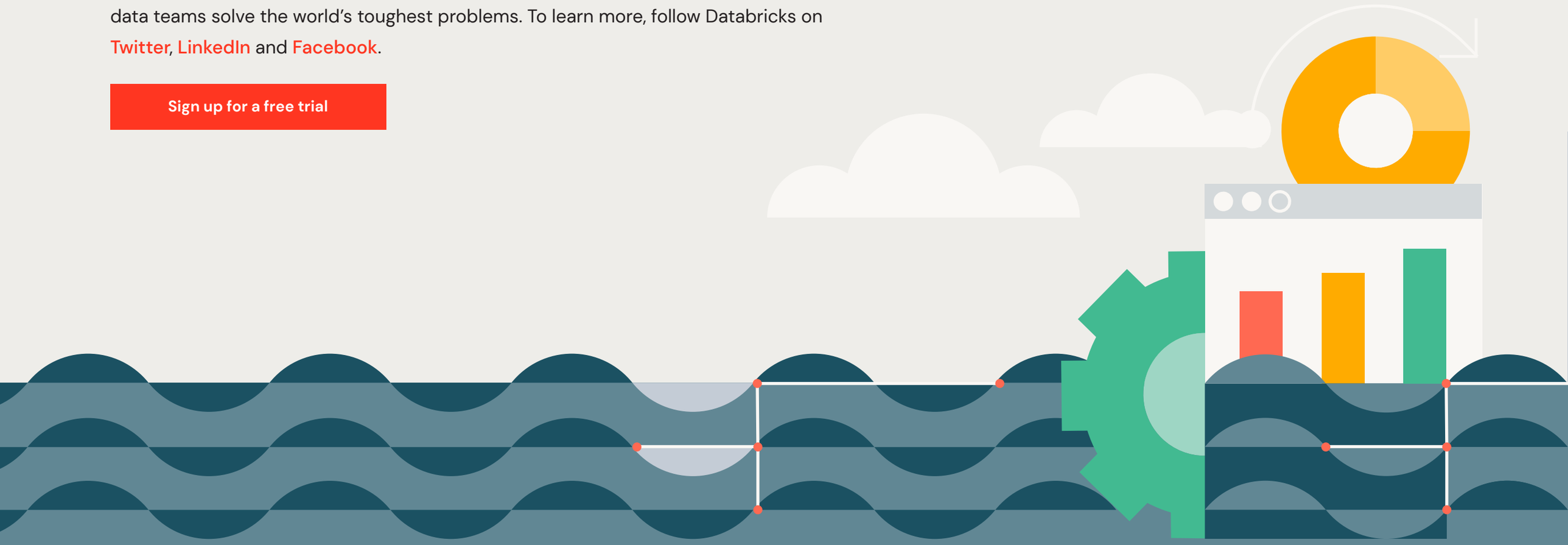
Migrate to Databricks →

Modernize Your Data Warehouse →

About Databricks

Databricks is the data and AI company. More than 7,000 organizations worldwide — including Comcast, Condé Nast, H&M and over 40% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics and AI. Databricks is 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 Twitter, LinkedIn and Facebook.

Sign up for a free trial



databricks