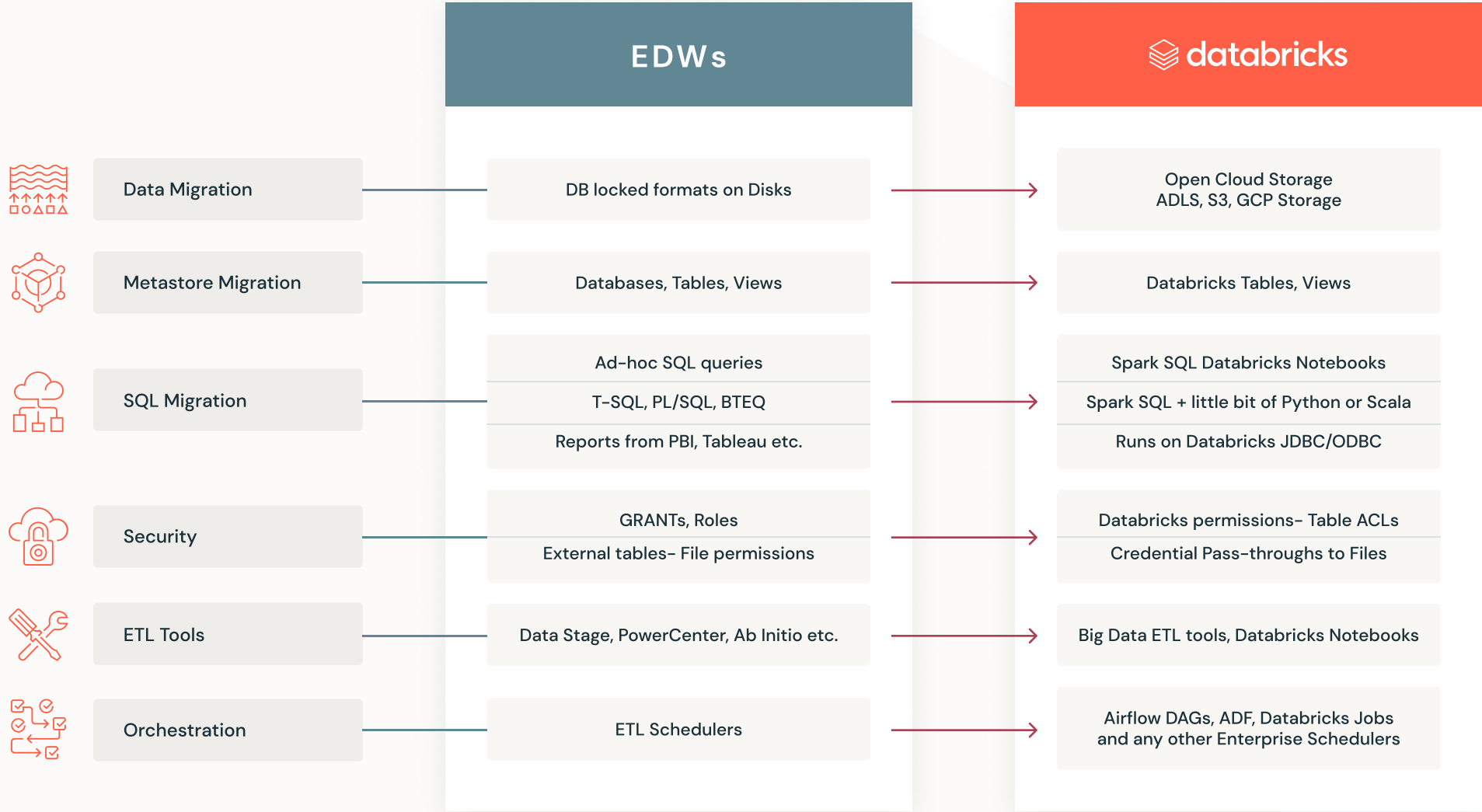


It's important to identify key stakeholders and keep them engaged during the migration to make sure they are aligned with the overall objectives. The workload assessment result will be reviewed with key stakeholders. Through the review process, data teams can get a better understanding of which workloads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

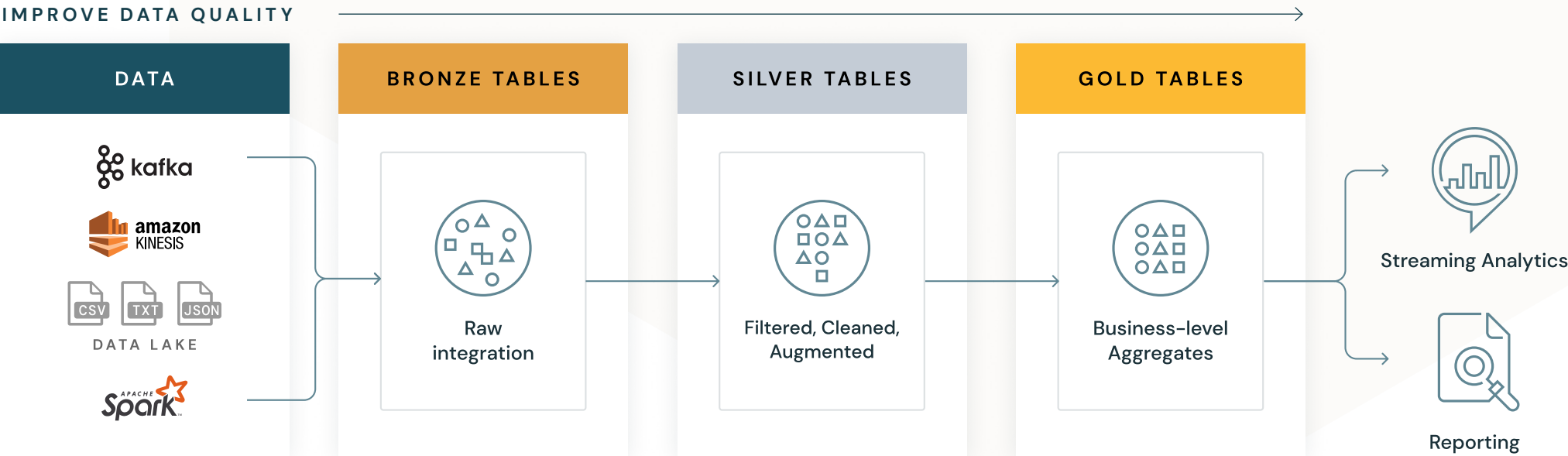
- Code conversion tooling that can automatically translate 70%–95% of the SQL code in your current system to Databricks optimized code with Delta and other best practices
- Converters that automate multiple GUI-based ETL/ELT platform conversion to reduce migration time and cost
- Data migration tools 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



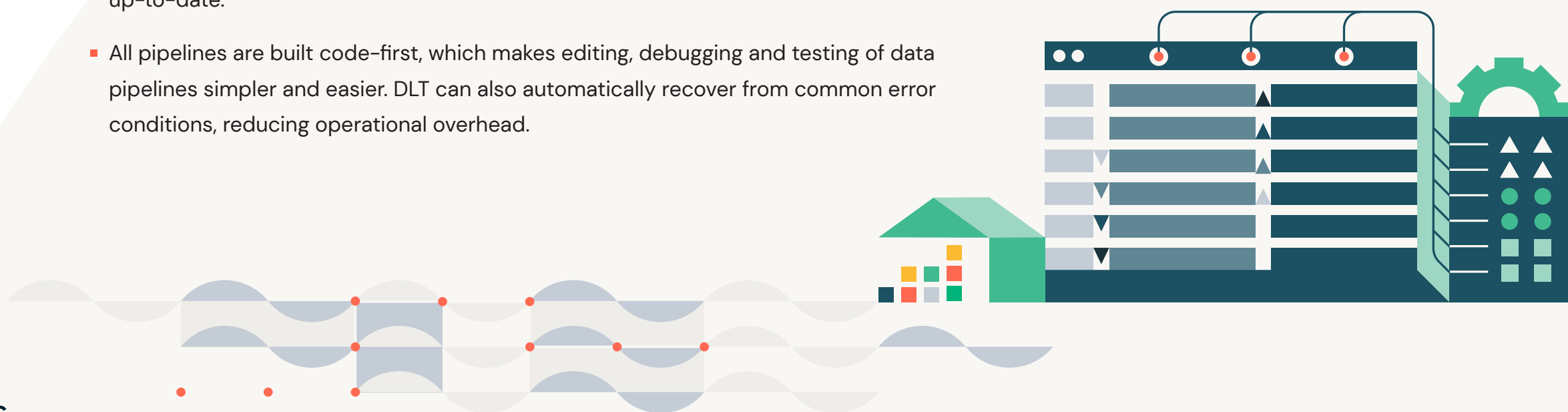
ELT approach

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

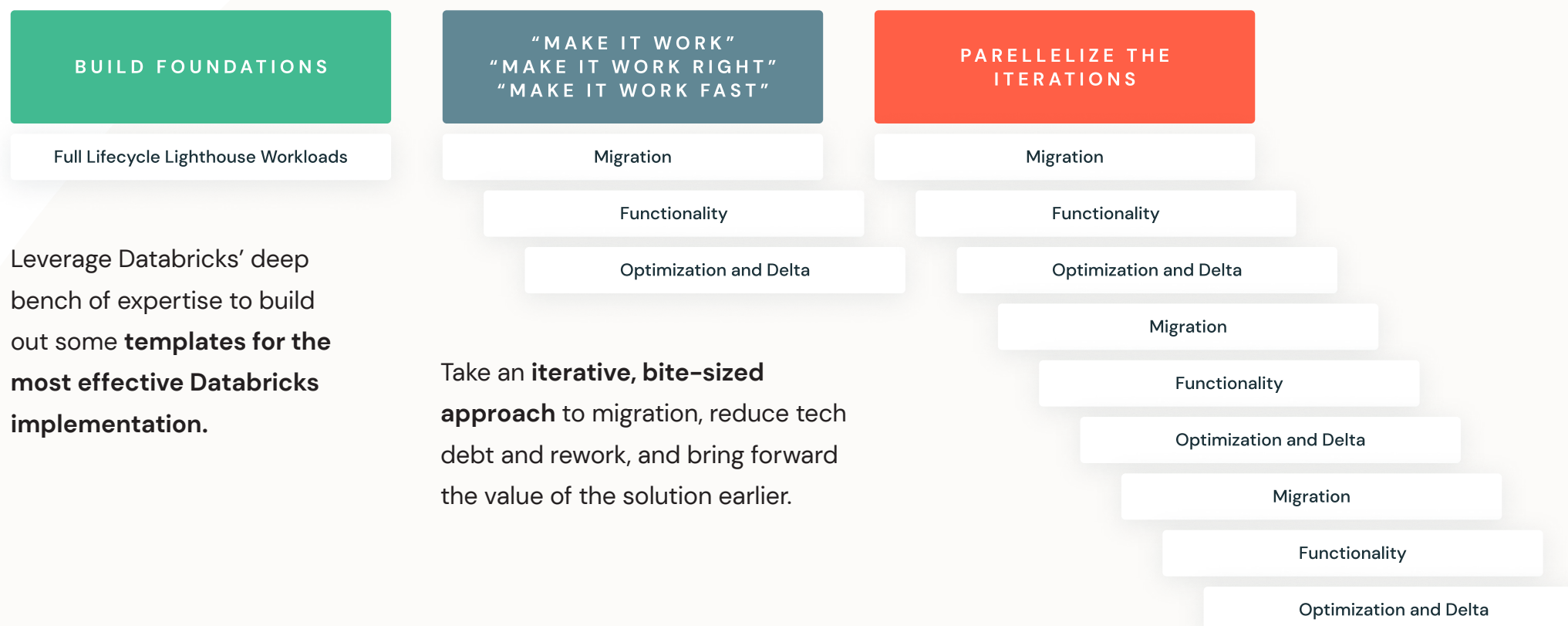
All of this allows you to take a more iterative and business-focused approach for migration instead of a full planning, execution, test/validation approach. Here are more approaches that help facilitate this phased implementation:

- 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.



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.

