databricks

Big Book of Data Warehousing and BI

Contents

Next Steps 65

databricks

PART 1

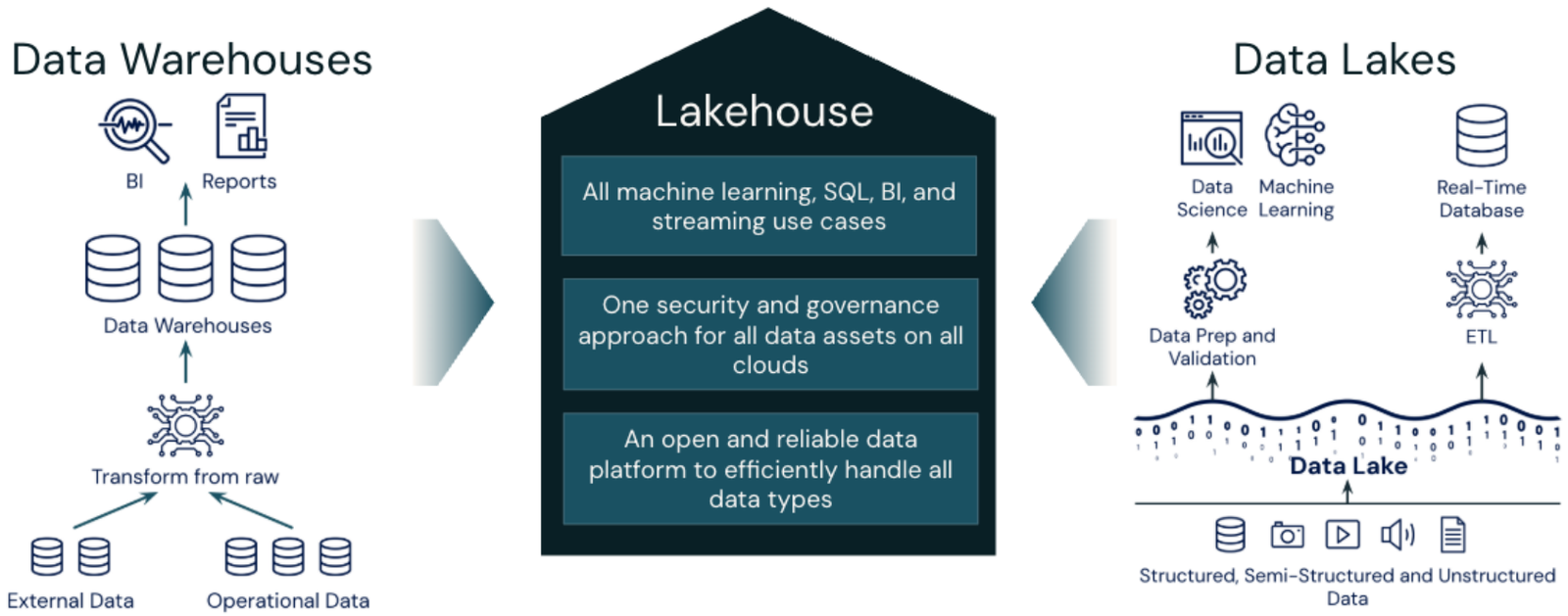
Data Warehousing and BI for What's Next

SECTION 1.1 The Modern Data Landscape

The business world is demanding more from data teams than ever before. Business leaders are asking questions that go beyond "what happened?" They want to know *why* something is happening, *what will happen next* and *what we should do about it*. This shift, powered by the emergence of AI, is disrupting traditional business intelligence (BI). Dashboards alone are no longer enough. Organizations need data platforms that can support predictive insights, dynamic decision-making and proactive recommendations. Moreover, they need these fresh insights served where they are, whether in the data warehouse, in a BI application, in an AI application or in a custom integration.

These new demands are exposing the limits of traditional data warehouse architectures. Many were built to handle structured, historical reporting, but struggle with real-time data, unstructured inputs, AI workloads and large-scale transformations. They're expensive to scale, slow to evolve and often lock customers into proprietary ecosystems that limit innovation.

The lakehouse paradigm offers a modern alternative. It combines the openness and scalability of a data lake with the performance, governance and SQL familiarity of a data warehouse — all in a single open platform. With Databricks' lakehouse architecture, organizations can unify BI and AI, support real-time analytics, and eliminate costly data silos. It enables teams to move faster, stay agile and deliver insights that drive the business forward.



LG Electronics

LG Electronics unified their global supply chain and smart factory operations on Databricks to modernize analytics across 200+ subsidiaries. By replacing fragmented systems with a single lakehouse architecture, they significantly reduced operational complexity and unlocked real-time insights across demand forecasting, manufacturing and logistics. This transformation enabled faster decisions, improved product availability and streamlined their global data estate for AI innovation.

LG customer story

THE BEST DATA WAREHOUSE IS A LAKEHOUSE

Traditional data warehouses are optimized for structured, retrospective reporting — but they fall short when it comes to today's needs: real-time decision-making, AI integration and cost-efficient scale. As organizations face rising data volumes and diverse use cases, a new approach has emerged: the lakehouse.

The lakehouse combines the governance and performance of a data warehouse with the openness and flexibility of a data lake. Built on open formats like Delta Lake, the Databricks Data Intelligence Platform provides a single source of truth for batch, streaming, BI and machine learning (ML) workloads. This eliminates costly duplication, reduces operational overhead and accelerates insight generation.

databricks

Databricks SQL is at the heart of this lakehouse architecture. It enables lightning-fast analytics on massive datasets using Photon, Databricks' massively parallel processing (MPP) query engine, and supports modern BI use cases with materialized views, predictive statistics, and Unity Catalog for semantic consistency and fine-grained access control. Workloads can be powered by serverless SQL warehouses, autoscaling to meet demand while lowering costs.

Critically, the lakehouse model allows organizations to unify structured BI dashboards with ML and AI — without the need to copy or move data across systems. Instead of building separate pipelines for reporting and modeling, data teams use one platform for all personas, and one storage layer that supports both Gold-level dimensional models and raw, unstructured data.

The best data warehouse is a lakehouse — it brings the performance of a warehouse, the scale of a data lake and the intelligence of AI, all on a single platform. This convergence enables enterprises to go beyond answering "what happened" and toward "what's next" — with BI and AI working together on the same foundation.

Learn more:

- Databricks SQL Product Page
- Webinar: The Best Data Warehouse Is a Lakehouse

OPERATIONAL AND ANALYTICAL WORKLOADS CONVERGE: THE ROLE OF LAKEBASE

Traditionally, operational workloads (OLTP) and analytical workloads (OLAP) have been handled by entirely separate systems. OLTP systems — such as application-focused databases — are optimized for fast, granular updates and real-time transactional processing. In contrast, OLAP systems, such as data warehouses and now lakehouses, were designed for aggregations, large scans and historical analytics. Moving data between these systems has historically required custom ETL pipelines, scheduled batch jobs and redundant storage — all of which introduce latency, governance gaps and operational overhead.

Databricks Lakebase helps address this fragmentation by introducing an operational database that runs alongside your existing analytics workflows in the Databricks Data Intelligence Platform. Built on open source Postgres with separated compute and storage, Lakebase offers fully managed transactional capabilities, including low-latency inserts, updates, deletes and point lookup, and is integrated with Delta Lake and Unity Catalog. This opens the door for real-time applications to run side by side with dashboards and ML models without data duplication.

By eliminating the latency and operational overhead of syncing OLTP data into the data warehouse, Lakebase helps modernize transactional workloads for the AI era. Teams can power apps, APIs and real-time decision engines using the same governed datasets trusted by business analysts and data scientists. It's a foundational shift in the modern data landscape.

INTEGRATING OLTP AND DATA WAREHOUSING FOR MODERN BUSINESS NEEDS

Lakebase helps unify operational and analytical use cases by minimizing the friction between databases and the lakehouse. With native support for syncing Delta tables to and from Lakebase, teams can build applications that interact with the same datasets used for analytics and AI, without relying on custom reverse ETL pipelines or duplicate storage.

For example, a retail company can use Lakebase to update inventory transactions in real time as purchases are made while simultaneously allowing analysts to monitor stock levels and sales trends using Databricks SQL. Governance is maintained across both workloads through Unity Catalog, ensuring that compliance and security are never compromised.

Lakebase scales based on workload demand and eliminates the need for provisioning, tuning or infrastructure management. Branching and the separation of compute from storage also support modern application development workflows, something traditional OLTP systems cannot.

As AI-native applications become more prevalent, the importance of integrating transactional systems with analytics will only grow. Lakebase gives organizations the flexibility to build applications and decision systems on a unified data platform, dramatically improving time to insight, operational agility and trust in the data.

Learn more:

- What Is Lakehouse and Lakebase
- Lakebase Public Preview Announcement
- Databricks Launches Lakebase Press Release

SECTION 1.2 Foundations of the Lakehouse for BI and AI

Lakehouse architecture is now a proven pattern that has been adopted by thousands of enterprises globally. At its core, it consists of four components:

1. **Storage:** Open format (Delta Lake, Apache Iceberg™) on cloud object stores (Amazon S3, Azure Data Lake Storage, GCS)
2. **Governance:** Fine-grained access control and auditing via Unity Catalog
3. **Query engine:** Databricks SQL for low-latency BI workloads
4. **Consumption:** Connectors to Power BI, Tableau, Excel, Google Sheets, Looker and AI agents

By consolidating these layers into a single platform, organizations eliminate costly data copies, simplify operations and reduce time to insight. For BI teams, this means faster dashboards and fewer delays. For IT, it means one place to secure, manage and optimize data.

With features like Lakehouse Federation, materialized views and Predictive Optimization statistics, Databricks is delivering warehouse-grade performance without compromise. And because it's all built on open formats and standards, customers retain full control over their data and ecosystem.

ANKER

Anker Innovations migrated to the Databricks Platform to unify their fragmented BI stack. The result: they accelerated BI queries by 94% and reduced time to insight from 30 minutes to 2 minutes, transforming how teams analyze product and customer trends.

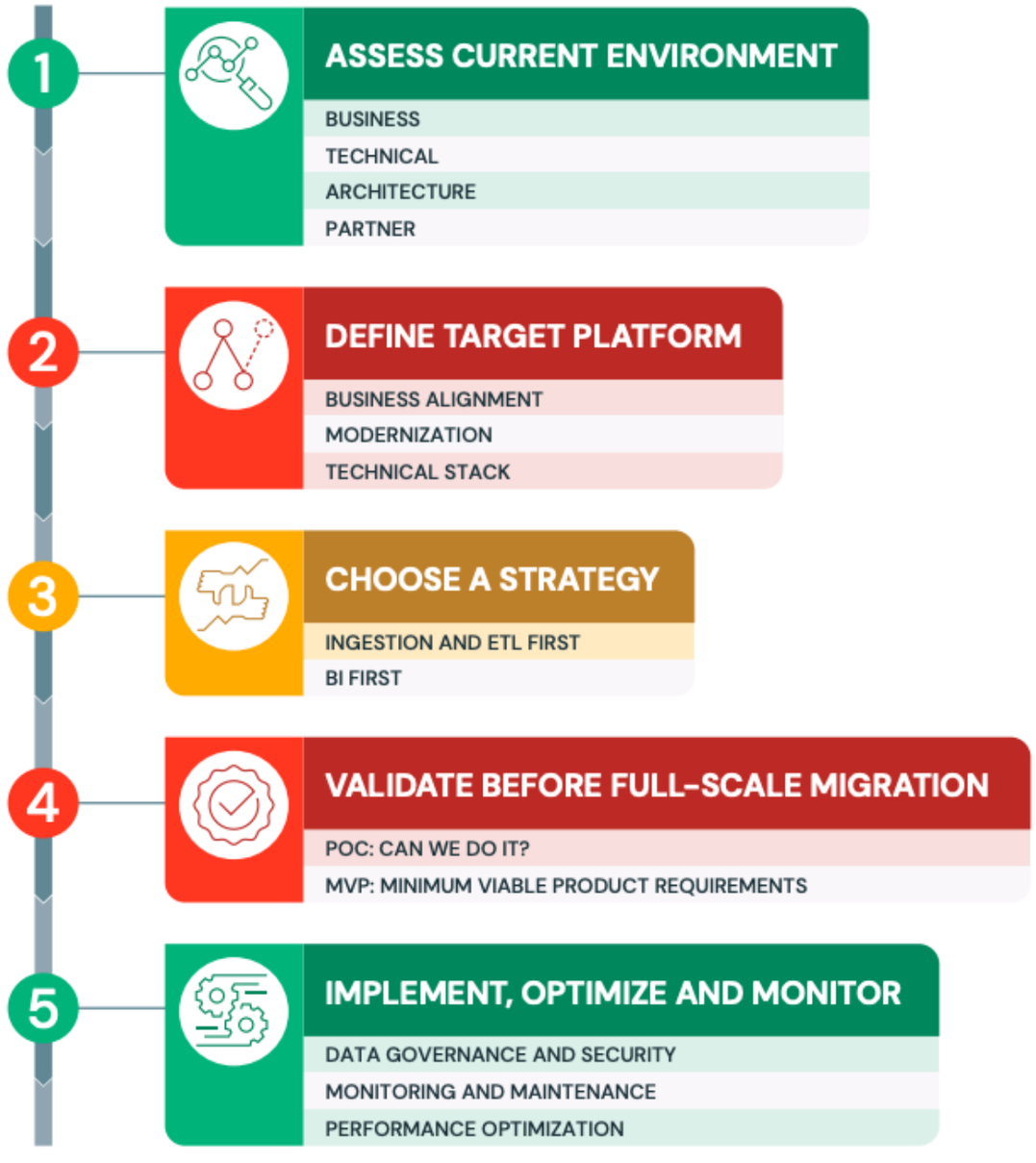
Anker customer story

SECTION 1.3 Migration and Modernization

Enterprises are under increasing pressure to modernize their data infrastructure — not only to reduce costs but also to accelerate analytics and AI. Legacy data warehouses like Snowflake, Oracle, SQL Server and Redshift often create bottlenecks due to proprietary formats, limited scalability and high operational complexity. The Databricks Platform provides a unified destination for both BI and AI, and with the right strategy, organizations can migrate faster and safer than ever before.

Successful modernization starts with understanding your existing landscape. The most effective migrations begin with a discovery phase — inventorying source systems, dependencies, BI dashboards, ETL pipelines and security requirements. Organizations that skip this step often face costly surprises downstream.

Proven steps to a successful data warehouse migration



The Databricks Lakebridge solution accelerates this journey. It automates schema conversion, data movement and workload replatforming. Whether you're migrating dashboards, retiring legacy ETL jobs or modernizing AI/ML pipelines, Lakebridge enables a fast, predictable data warehouse migration. Successful teams focus on phased delivery — targeting high-impact workloads first, validating performance early and establishing strong financial operations (FinOps) visibility throughout.

- Start with data, not dashboards — optimize pipelines first, then BI
- Migrate in logical "waves" by business function (e.g., finance, supply chain)
- Use Unity Catalog to rebuild governance models in parallel
- Validate with shadow testing and cost baselining to prove wins early

Companies that adopt this approach typically cut 20%–50% off total migration effort — and reduce risk to critical business operations.

SNOWFLAKE TO THE DATABRICKS PLATFORM

Lakebridge captures Snowflake metadata, usage patterns and schemas to accelerate migration. Delta Lake's open format ensures compatibility with modern analytics tools. Databricks outlines a practical and scalable approach for migrating from Snowflake to the Databricks Platform, emphasizing that successful migrations require more than simply moving data. Databricks recommends aligning early with Unity Catalog for consistent governance and leveraging native tools such as Auto Loader, Lakeflow Spark Declarative Pipelines and Lakeflow Jobs. The approach prioritizes use-case-driven planning over wholesale replication, enabling faster time to value and reducing technical debt. With support for SQL, Python and other languages, teams can optimize pipelines for performance and cost while maintaining compatibility with BI and data science tools. The result is not just a more efficient platform, but one designed for AI and real-time analytics — unlocking use cases that were previously limited by Snowflake's proprietary model.

> "The Databricks unified lakehouse empowered the teams to build an AI-ready foundation for the future. This wasn't just a migration — it was a strategic leap forward."
>
> — Amit Rustagi, Architect, DeeplearningAPI

Optimizing Analytics Infrastructure: Lessons From Migrating Snowflake to Databricks

ORACLE TO THE DATABRICKS PLATFORM

Legacy Oracle data warehouses often pose licensing and agility constraints. Databricks helps convert PL/SQL and schema logic, enabling customers to move financial analytics and ERP data into a modern lakehouse.

amadeus

Amadeus, a global leader in travel technology, migrated from Oracle to Databricks to modernize their data infrastructure and unlock real-time analytics. The company leveraged Unity Catalog to implement fine-grained access control and standardized governance across business domains.

"What we were excited about is, not just with Databricks SQL but with other pieces and features within Databricks, we were able to consolidate a lot of our resources … and then use DBSQL to do the serving layer with our custom applications."

— Chris Ernst, Data Architect, Amadeus

Amadeus customer story

SQL SERVER TO THE DATABRICKS PLATFORM

Databricks supports T-SQL, SSIS/ADF migration and change data capture (CDC) replication. Many enterprises are migrating from legacy SQL Server environments to Databricks SQL to modernize analytics, cut costs and enable real-time insights. Faster data processing and report runtimes acceleration are made possible by adopting a structured migration plan, including medallion architecture modeling, early use of Unity Catalog for governance, and leveraging Databricks performance features like Photon and Auto Loader. Databricks SQL Server migration guide outlines best practices such as inventorying workloads by use case, starting with high-impact dashboards, and decoupling compute and storage to reduce spend.

abcloudz

Azure Databricks transformed ABCloudz's ability to manage and analyze data.

"Data processing speeds increased by 60%, and real-time report generation moved from hours to just minutes."

— Adam Robertson, Director of Solution Management, ABCloudz

ABCloudz customer story

REDSHIFT TO THE DATABRICKS PLATFORM

Migrating from Amazon Redshift to Databricks is typically more straightforward than other data warehouse migrations because your data is already in the cloud, but it still benefits from a structured, phased approach. Databricks recommends starting with a discovery and assessment phase using tools like the Amazon Redshift Query Profiler and Lakebridge Code Analyzer to inventory workloads, assess complexity and map Redshift features to lakehouse capabilities. Data is then offloaded into Delta Lake's open format — often via Redshift's UNLOAD command or partner tools — organized in a medallion architecture (Bronze, Silver, Gold) to support both parallel runs and incremental cutover. Pipeline and query migration leverage Lakeflow Jobs, Spark Declarative Pipelines and automated SQL conversion to minimize rewrite effort while enabling performance optimizations like Z-Ordering and Intelligent Workload Management. By the end of the process, downstream BI tools are repointed to Databricks SQL, unlocking unified, high-performance analytics at scale without Redshift's concurrency and scaling limitations.

ANYCLIP

AnyClip, a visual intelligence company processing millions of video frames and metadata daily, migrated from Amazon Redshift to Databricks SQL. The switch dramatically improved query performance, stability and cost-efficiency.

- **98% faster query performance** on terabyte-scale datasets
- **~25% lower total cost**, thanks to elastic compute and optimized storage
- **Reliable, self-service analytics** enabled business users to build dashboards independently.

"Previously, when users tried to run queries, it could take hours. With Databricks, it takes something between half a minute to three minutes."

— Gal Doron, Head of Data, AnyClip

AnyClip customer story

SECTION 1.4 Blueprints and Architectures

Databricks provides a library of reference architectures to help data teams accelerate time to value, reduce architectural guesswork and meet the needs of modern analytics and AI workloads. These blueprints have been shaped by experience across thousands of implementations and offer opinionated guidance on building efficient, governed and scalable solutions using the lakehouse architecture.

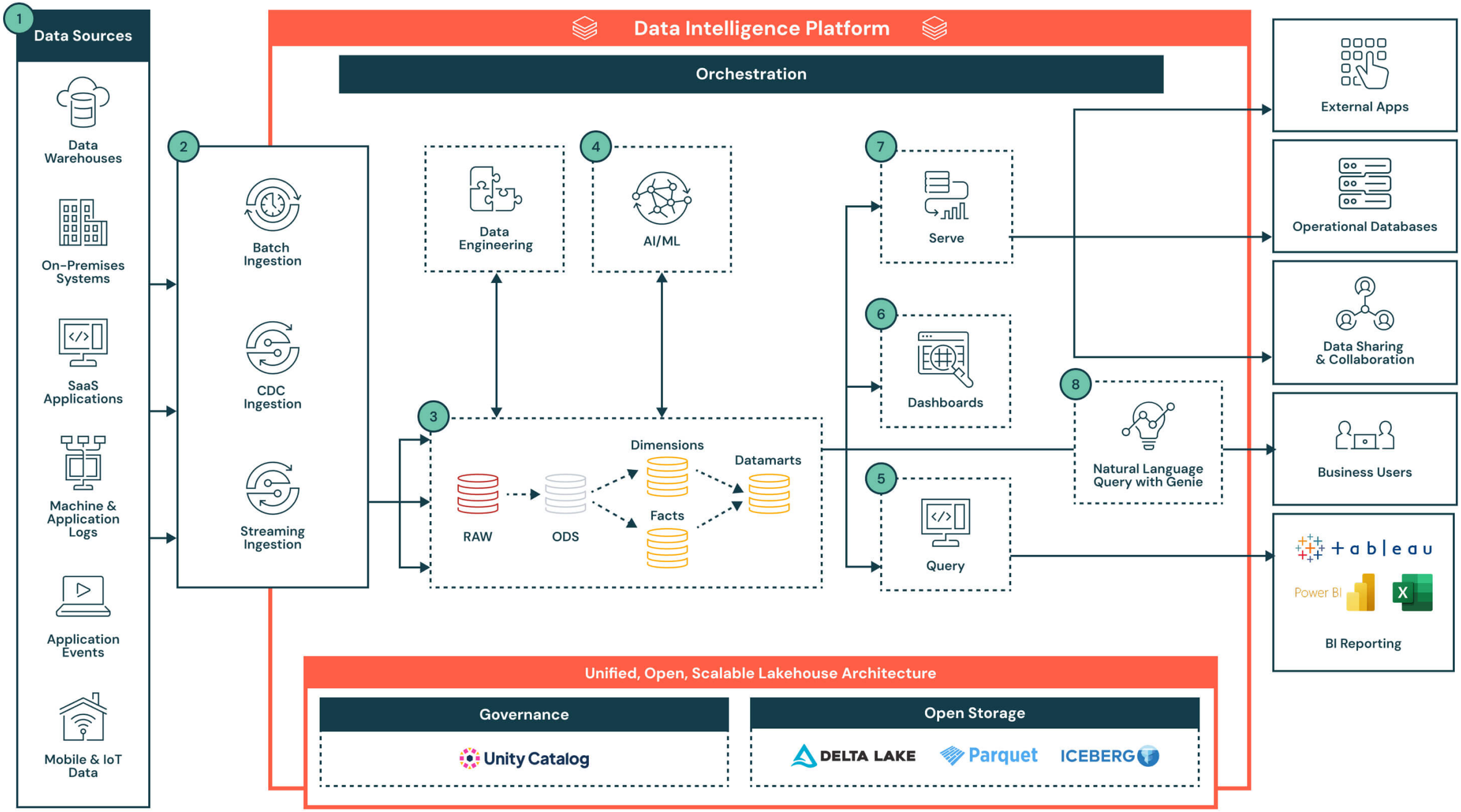
The foundation is the intelligent data warehousing on Databricks reference architecture. This architecture guides organizations in integrating data ingestion, governance, transformation and consumption into a unified workflow. It's structured around Delta Lake storage, Unity Catalog for metadata and security, and Databricks SQL for fast, flexible BI workloads. It supports multicloud ingestion, open data formats and downstream connectors to every major BI and AI tool.

A key benefit of this architecture is modularity. Each layer — whether ingestion with Auto Loader, orchestration with Spark Declarative Pipelines, performance acceleration via materialized views, or consumption through Power BI or Tableau — can be implemented independently, allowing teams to evolve at their own pace. This flexibility is essential for companies migrating from siloed legacy systems and enables phased deployment while maintaining governance.

This architecture isn't theoretical. It's being used in production today by organizations modernizing their BI, scaling embedded analytics or supporting financial reporting at a global scale. The result is not only cost savings but greater speed, agility and readiness for AI.

Learn more:

- Architecture Center
- Intelligent Data Warehousing Reference Architecture


1
Data Sources
Data Warehouses
On-Premises Systems
SaaS Applications
Machine & Application Logs
Application Events
Mobile & IoT Data
2
Batch Ingestion
CDC Ingestion
Streaming Ingestion
Data Intelligence Platform
Orchestration
Data Engineering
4
AI/ML
3
Dimensions
Datamarts
Facts
RAW
ODS
7
Serve
6
Dashboards
5
Query
8
Natural Language Query with Genie
External Apps
Operational Databases
Data Sharing & Collaboration
Business Users
tableau
Power BI
BI Reporting
Unified, Open, Scalable Lakehouse Architecture
Governance
Unity Catalog
Open Storage
DELTA LAKE
Parquet
ICEBERG

REAL-TIME EXECUTIVE DASHBOARDS

Executives demand real-time visibility into key performance indicators (KPIs) that span departments — finance, sales, operations, supply chain — and they want these insights to be trusted, interactive and immediately actionable. Traditional BI environments built on batch pipelines and fragmented data sources fail to deliver on this vision.

With the Databricks Platform, real-time executive dashboards become possible through CDC, streaming ingestion and sub-second performance on fresh data. Technologies like Auto Loader and Spark Declarative Pipelines automate pipeline creation and maintenance, while Unity Catalog ensures governance remains intact across the stack.

Databricks SQL brings low-latency, warehouse-like query performance to real-time data, supporting tools like Power BI and Tableau. This enables executives to interact with dashboards that update continuously, without worrying about stale numbers or inconsistent metrics.

wineshipping

Wineshipping rolled out 140 production AI/BI dashboards, and 105 of ~150 users engaged within just one week (~75% adoption). Their teams — from CEO to hourly staff — use these dashboards daily across finance, sales, operations and customer service.

> "The users range from our CEO to hourly employees. It's really throughout the entire company."

— Christina Mottolo, VP, Enterprise Applications, Wineshipping

Wineshipping customer story

EMBEDDED AND EXTERNAL ANALYTICS

SaaS and data product teams are increasingly expected to deliver analytics as part of the user experience. Whether it's customer-facing dashboards or embeddable insights, the challenge is always the same: scale, governance and customization.

Databricks makes embedded analytics scalable and secure. With Databricks SQL and Unity Catalog, teams can serve data to thousands of users with fine-grained access control. Each customer can receive personalized insights while all users operate on a single governed architecture. Materialized views and performance optimizations ensure speed, while APIs allow seamless embedding into modern front ends.

This model's power lies in its cost-effectiveness. Traditional embedded analytics require provisioning multiple isolated environments or exporting data. With the Databricks Platform, you avoid redundant data pipelines, eliminate data copies and deliver faster time to market for new products.

FINANCIAL ANALYTICS AND REGULATORY REPORTING

Finance teams require more than just visualizations — they need precision, explainability and strict controls over data access. For publicly traded companies or those in regulated industries, the reporting process must meet compliance standards and auditability requirements.

Databricks is well suited to this domain due to its strong support for dimensional modeling, high-performance querying and native governance. Organizations can implement star schemas, slowly changing dimensions, and cross-system joins at scale. Unity Catalog enforces column-level access, while versioned data with Delta Lake ensures full traceability.

With Databricks SQL, finance teams can deliver interactive reporting and complex calculations directly from the lakehouse. Materialized views optimize recurring queries, while the semantic layer provides clarity around key business metrics.

Nasdaq

Nasdaq migrated its enterprise data warehouse to the Databricks Data Intelligence Platform to unify real-time and historical data across its business. This move broke down data silos and enabled agile financial reporting and advanced analytics across trading platforms. With Unity Catalog and MLflow, Nasdaq strengthened governance and accelerated ML use cases.

"We use Unity Catalog for governance and MLflow to track and deploy models, so we have confidence in what we've built and can deliver results quickly."

— Louie Leung, VP, Data Engineering, Nasdaq

Nasdaq customer story

SECTION 1.5 Data Modeling

DATA MODELING ON THE LAKEHOUSE

Data modeling is a foundational discipline for scalable business intelligence, even in the era of cloud-native and AI-driven analytics. With the Databricks Data Intelligence Platform, organizations can modernize how they design and implement data models by drawing on traditional paradigms — like dimensional, 3NF and Data Vault — while also adopting new patterns that align with medallion architecture and real-time ingestion.

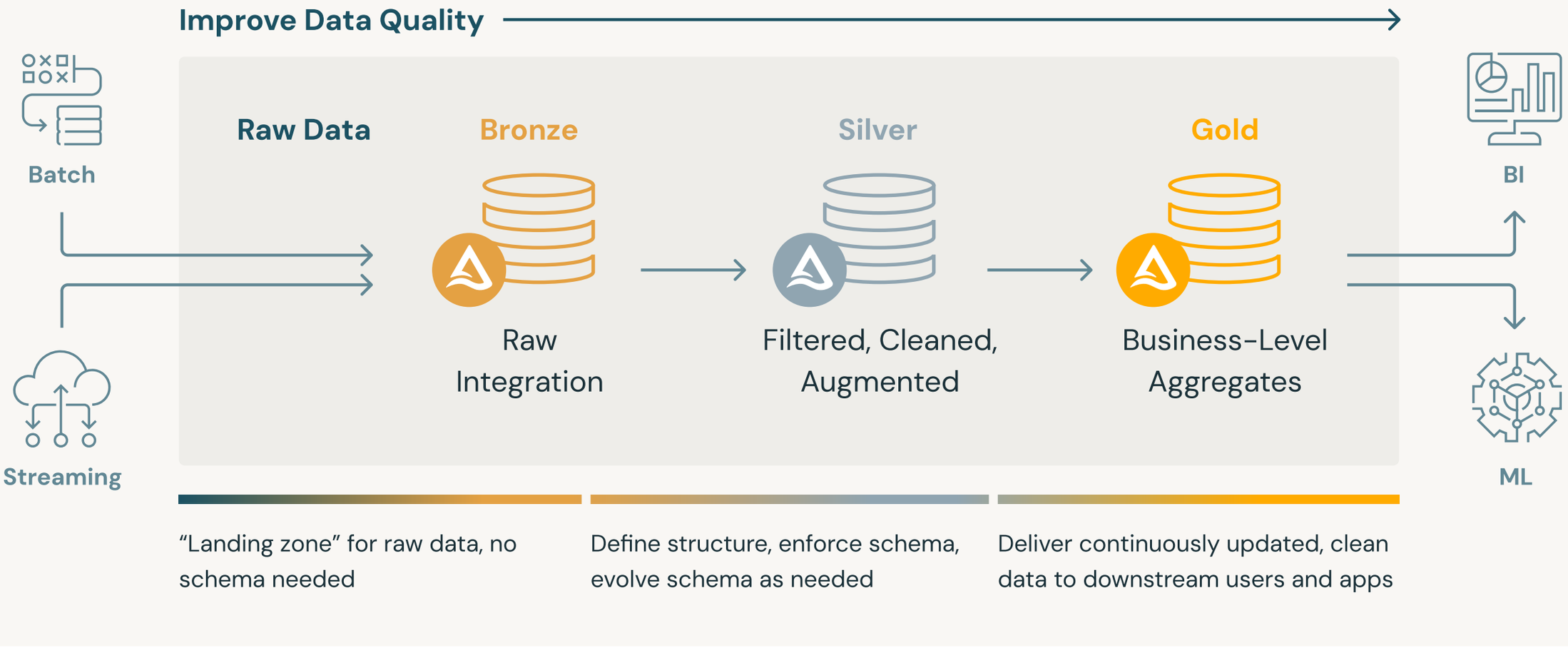
Dimensional modeling in the lakehouse: Dimensional modeling remains one of the most widely adopted paradigms for reporting and BI. On Databricks, star and snowflake schemas can be implemented using Delta Lake tables and optimized with Unity Catalog and materialized views. For teams migrating from traditional enterprise data warehouses (EDWs), this provides familiarity while introducing scalability and openness.

Read this detailed implementation of fact and dimension tables, slowly changing dimensions (SCDs), and surrogate keys in Databricks SQL: Implementing Dimensional Data Warehousing on Databricks.

Data vault and modern modeling: More flexible than traditional normalization, Data Vault offers auditability and agility for teams managing complex data integration pipelines. The lakehouse architecture allows implementation of hubs, links and satellites on Delta Lake, which scale efficiently and support both raw and curated zones. See Data Vault on Databricks for a glossary and architectural reference.

Medallion architecture as a semantic strategy: The medallion architecture — Bronze, Silver, Gold — naturally aligns with data modeling best practices. Raw ingestion lives in Bronze (e.g., CDC feeds from SAP), cleansing and conformance occur in Silver, and business logic is layered in Gold, supporting BI tools and AI/BI dashboards. Medallion Architecture Overview provides guidance on how to align this with modeling layers.

Building reliable, performant data pipelines with Medallion Architecture



Best practices and tooling: A comprehensive guide to modeling trade-offs and implementation techniques is available in the Databricks blog Data Warehousing Modeling Techniques on the Lakehouse and the follow-up on Best Practices. These cover entity-relationship models, hybrid approaches and semantic modeling with tools like dbt and Unity Catalog.

Identity columns: Identity columns allow you to automatically generate surrogate keys in Delta tables — crucial for dimensional modeling in modern data warehouses. Surrogate keys help maintain consistent, immutable identifiers in slowly changing dimensions and complex joins. Previously, generating these required work-arounds or multistep ETL logic. With native identity columns, users can define auto-incrementing columns using standard SQL syntax, simplifying schema design and improving pipeline reliability. Identity Columns to Generate Surrogate Keys Are Now Available in a Lakehouse Near You

Modeling in the lakehouse requires a balance of agility and governance. When designed intentionally, it can serve as the semantic backbone for a high-performing BI platform that feeds traditional dashboards, AI-driven exploration and real-time use cases.

SECTION 1.6 BI in the Era of AI

FROM DASHBOARDS TO DECISIONS: EVOLVING THE ROLE OF BI

The expectations of business intelligence have shifted. Teams no longer want just charts — they want contextual answers, proactive insights and AI-powered assistants that make decisions easier and faster. Databricks is meeting this demand by evolving the BI experience from passive consumption to active conversation.

Databricks SQL integrates directly with AI models, agents and copilots. These intelligent interfaces are built on the same governed data foundation used for traditional dashboards. That means every AI-driven response is as trustworthy and auditable as a BI report. By unifying the analytics experience, organizations can move beyond static KPIs to real-time decisions made in natural language.

AA

The AA, one of the UK's leading motoring organizations, adopted the Databricks AI/BI Genie solution to streamline access to data across teams. By integrating the Genie API into Microsoft Teams, they enabled conversational analytics directly within their business tools — boosting speed, accessibility and adoption. This allowed frontline staff and executives alike to access insights on customer behavior, service KPIs and operational trends without needing to switch platforms. The rollout led to a ~70% reduction in time to insight, significantly improving agility and responsiveness.

"We're unlocking access to customer insights in a faster and more intuitive way. With the Genie API, we're delivering AI-powered analytics right inside our business tools."

— Matt Sanderson, Head of Data Products, Channels, The AA

The AA customer story

MEET GENIE: YOUR AI-POWERED BI ASSISTANT

Genie, Databricks' conversational analytics assistant, allows users to ask questions about their data and access dynamic dashboards, charts and summaries. Its foundation is the Databricks One lakehouse, which means it operates on real-time, governed data from across the enterprise.

As of General Availability, Genie can integrate with existing BI environments, generate new dashboard elements based on intent, and offer explainability features for each response. With Genie, business users don't need to write SQL or wait on analysts — they simply ask.

Learn more: AI/BI Genie Now Generally Available

CONVERSATION APIS: TURNING QUESTIONS INTO PIPELINES

The new Conversation APIs, now in Public Preview, make it possible to embed natural language experiences into any app. Developers can build interfaces where business users, customer support reps or operators can ask questions in plain English and get contextual, governed answers based on warehouse data.

These APIs offer access to metadata, semantic layers and lineage, ensuring that AI answers are accurate and explainable. This opens the door to industry copilots across verticals: finance, supply chain, customer service and more.

Learn more: Genie Conversation APIs: Public Preview

FASTER, SMARTER DASHBOARDS WITH AI-POWERED BI

Modern BI dashboards on Databricks are now dynamically powered by AI. With support for dynamic calculations, model-driven metrics and AI-generated summaries, dashboards do more than display data — they interpret and explain it.

This approach turns a dashboard into a decision workspace. Business users receive alerts, plain-language insights and recommendations — all built on top of governed, real-time data. This creates a smarter, more collaborative BI experience across teams.

Learn more: Author Faster, Smarter AI/BI Dashboards

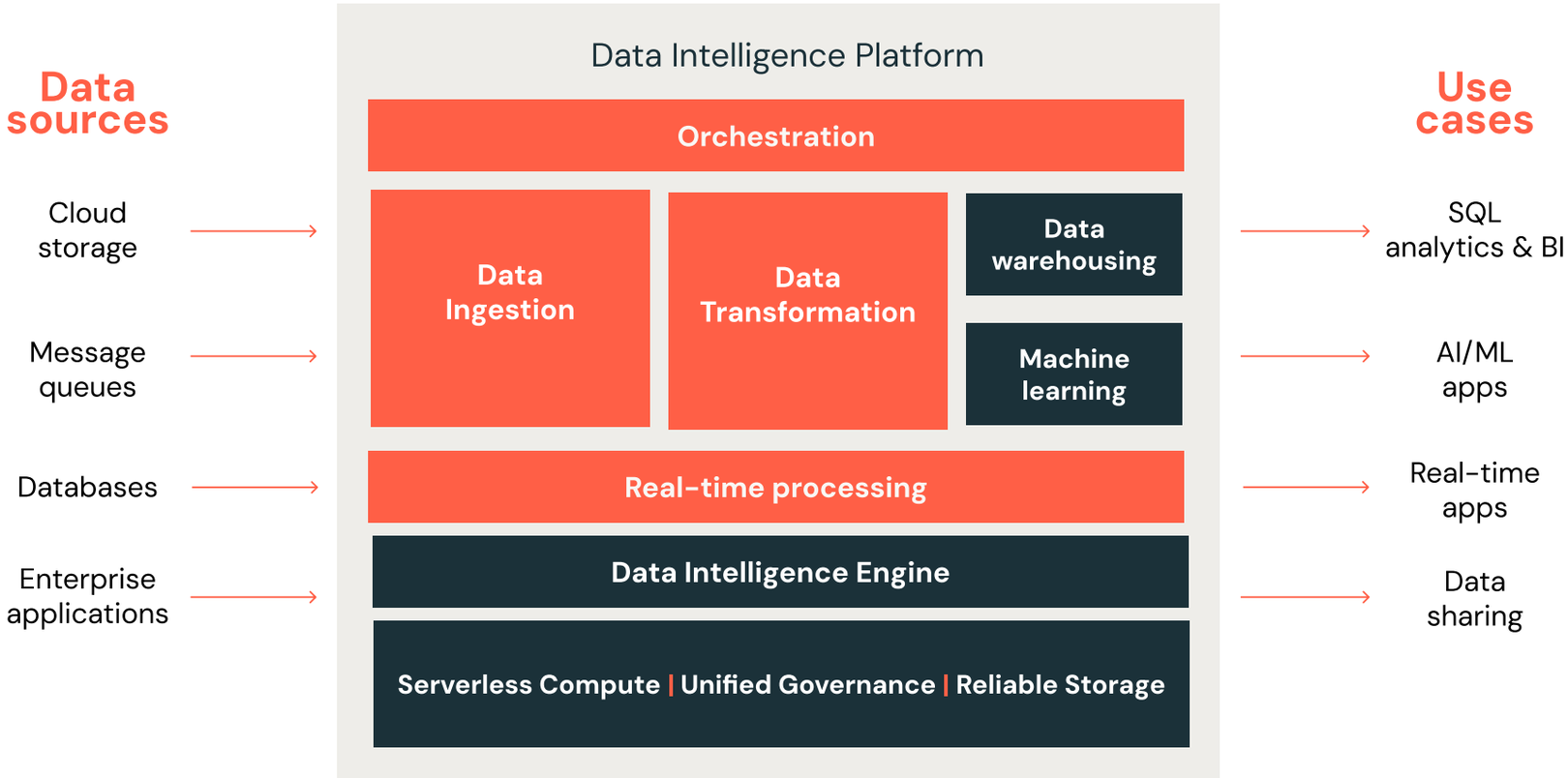
SECTION 1.7 Data Ingestion and Preparation

Modern businesses demand reliable, fresh, clean data as the foundation to power everything from analytics and BI to GenAI. Databricks provides data engineers with the ability to ingest structured and unstructured data all in one place — as part of a unified data engineering solution from Lakeflow. With a wide range of ingestion strategies — from simple ingestion connectors for applications, to databases and file sources, to batch and real-time streaming, to manual uploads and fully automated pipelines — organizations can unify fragmented data sources directly into Delta Lake with minimal friction.

WHAT IS DATA INGESTION?

Data ingestion refers to the process of collecting data from diverse sources and moving it into a centralized system for storage and analysis. In the traditional lakehouse context, this usually means ingesting raw, unstructured, semi-structured and structured data into Bronze tables using Delta Lake, where it can be subsequently processed and modeled for downstream BI and AI use cases. See the Databricks Glossary on Data Ingestion for definitions and examples.

In addition to data ingestion, Databricks Lakehouse Federation helps customers discover, query and govern their data in Unity Catalog. Delta Sharing also expands secure data collaboration across platforms. For example, healthcare and retail customers use open sharing to deliver governed data to partners while maintaining control.



EFFICIENT INGESTION PROVIDES FLEXIBILITY AND SCALE

Databricks offers multiple methods for ingestion:

- **Managed data connectors** for applications, file sources, databases and data warehouses
- **Manual file and table uploads** through the UI
- **Structured streaming and Auto Loader** for real-time pipelines
- **Partner integrations** with tools like Fivetran and Informatica

Lakeflow Connect makes efficient data ingestion easy with a point-and-click UI, a simple API and deep integrations with the Databricks Data Intelligence Platform. These managed connectors provide streamlined ingestion from applications, file sources, databases and data warehouses (see supported data sources below) directly into Delta Lake. Under the hood, incremental updates and optimized API usage provide efficient ingestion with managed schema evolution, seamless third-party API upgrades and comprehensive observability with built-in alerts. It's particularly powerful for teams consolidating data across operational systems into a unified lakehouse.

For database and data warehouse sources, Lakeflow Connect leverages log-based optimal CDC, which accesses the source's transaction logs and other techniques to capture data changes and DML operations like inserts, updates and deletes. When CDC is not available on a source, new query-based connectors can instead leverage a cursor column for incremental ingestion, with no source modifications required.

Supported data sources:

- **Applications:** Salesforce, Workday, ServiceNow, Google Analytics, Microsoft Dynamics 365, Oracle NetSuite
- **File sources:** S3, ADLS, GCS, SFTP, SharePoint
- **Databases:** SQL Server, Oracle Database, MySQL, PostgreSQL
- **Data warehouses:** Snowflake, Amazon Redshift, Google BigQuery

In addition to ingestion via managed connectors, Lakeflow Connect also offers **file and table uploads** directly through the UI, as well as **structured streaming and Auto Loader** for real-time pipelines (see Layers of the ETL Stack below). Databricks also offers third-party **partner integrations** with ingestion tools like Fivetran and Informatica.

In addition to ingestion via managed connectors, Lakeflow Connect also offers **file and table uploads** directly through the UI, as well as **structured streaming and Auto Loader** for real-time pipelines (see Layers of the ETL Stack below). Databricks also offers third-party **partner integrations** with ingestion tools like Fivetran and Informatica.

PORSCHE HOLDING

Porsche Holding — Europe's largest automotive retail company, operating in 23 countries and representing Volkswagen Group brands in wholesale, retail and after-sales — uses Lakeflow Connect to ingest Salesforce data into Databricks. As a result, they were able to implement 85% faster development using Lakeflow Connect instead of a homegrown data ingestion solution.

"Using the Salesforce connector from Lakeflow Connect helps us close a critical gap for Porsche from the business side in terms of ease of use and price. On the customer side, we're able to create a completely new customer experience that strengthens the bond between Porsche and the customer with a unified and unfragmented customer journey."

— Lucas Sulzberger, Project Manager, Porsche Holding

Porsche Holding customer story

NEW DIRECT WRITE API PUSHES EVENT DATA TO YOUR LAKEHOUSE

Part of Lakeflow Connect, Zerobus is a new direct write API that helps simplify ingestion for Internet of Things (IoT), clickstream, telemetry and other similar use cases by bringing you closer to the data source. Zerobus enables data producers, such as cloud apps or IoT devices, to push events directly to Unity Catalog instead of using message buses like Kafka as a transport layer to the lakehouse. When your sole destination is the lakehouse, you can use Zerobus to eliminate data hops and reduce operational burden, as well as deliver high throughput, near real-time performance at scale, and low TCO.

Joby Aviation is using Zerobus to directly push telemetry data into Databricks as part of their broader data and AI strategy for faster insights and data democratization across the company.

"Joby is able to use our manufacturing agents with Zerobus to push gigabytes a minute of telemetry data directly to our lakehouse, accelerating the time to insights — all with Databricks Lakeflow and the Data Intelligence Platform."

— Dominik Müller, Factory Systems Lead, Joby Aviation

What's New: Zerobus and Other Announcements Improve Data Ingestion for Lakeflow Connect

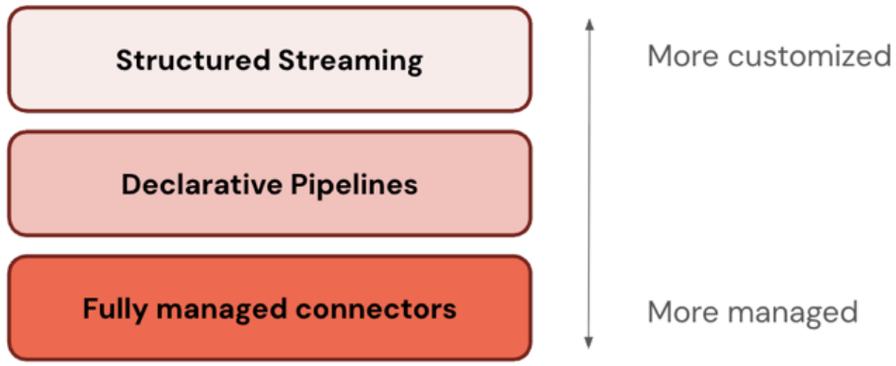
COMMON INGESTION USE CASES

Data ingestion is a core part of Lakeflow, the unified data engineering solution from Databricks covering ingestion, transformation and orchestration. Lakeflow's unified tool stack helps build reliable data pipelines by bringing all your data into one place and helping drive data insights. Some sample use cases include the following:

- **Customer 360:** Measuring campaign performance and customer lead scoring
- **Portfolio management:** Maximizing ROI with historical and forecasting models
- **Consumer analytics:** Personalizing your customers' purchasing experiences
- **Centralized human resources:** Supporting your organization's workforce
- **Digital twins:** Increasing manufacturing efficiency
- **Retrieval augmented generation (RAG) chatbots:** Building chatbots to help users understand policies, products and more

DATA PREPARATION: LAYERS OF THE ETL STACK

According to the Databricks technical documentation, Lakeflow Connect provides a flexible service model offering out-of-the box connectors with simple UIs and powerful APIs, or a custom pipeline with more customization options using Lakeflow Spark Declarative Pipelines or Structured Streaming.



Organizations have the flexibility to select a more managed or more customized experience when ingesting data, depending on their specific needs. Databricks recommends starting with the most managed layer of the ETL stack (see diagram above), and then moving to the next layer if needed.

- **Structured Streaming:** Apache Spark™ Structured Streaming is a streaming engine that offers end-to-end fault tolerance with exactly-once processing guarantees using Spark APIs
- **Lakeflow Spark Declarative Pipelines:** This builds on Structured Streaming, offering a more declarative framework for creating data pipelines. You can define the transformations to perform on your data, and Spark Declarative Pipelines manages orchestration, monitoring, data quality, errors and more. Therefore, it offers more automation and less overhead than Structured Streaming.
- **Managed connectors:** This builds on Spark Declarative Pipelines, offering even more automation for the most popular data sources. They extend Spark Declarative Pipelines functionality to also include source-specific authentication, CDC, edge case handling, long-term API maintenance, automated retries, automated schema evolution and so on. Therefore, they offer even more automation for any supported data sources.

FEDERATED QUERY AND JUST-IN-TIME INGESTION

Ingestion is recommended for most use cases because it scales to accommodate high data volumes, low-latency querying and third-party API limits. However, in scenarios where immediate physical ingestion is unnecessary — or when data must remain in source systems — Lakehouse Federation provides another path. Federation allows Databricks users to query external databases like PostgreSQL, MySQL, Oracle or Snowflake directly, with governance managed through Unity Catalog. Federation can be a precursor to ingestion: once query patterns are established, high-demand data can be ingested into Delta Lake to optimize performance and reduce external query costs. This hybrid pattern — sometimes called "just-in-time ingestion"— blends flexibility with control.

For example, the Salesforce Data Cloud connector helps customers find and query Salesforce data directly from Unity Catalog without needing to replicate pipelines or transform data.

Learn more: What's New: Zerobus and Other Announcements Improve Data Ingestion for Lakeflow Connect

SECTION 1.8 AI Capabilities in the Data Warehouse

AI is becoming a first-class citizen in the modern data warehouse, transforming how teams interact with data, build insights and power intelligent applications. In traditional BI, users queried data manually and relied on dashboards for reporting. In today's lakehouse, AI is embedded directly into the query and dashboarding experience — making analysis faster, smarter and more accessible.

Databricks SQL now offers native large language model (LLM) integration through built-in AI functions. These functions allow users to run LLM queries directly in SQL — enabling natural language processing, summarization, entity extraction and sentiment analysis from within dashboards or scheduled jobs. For example, a product team can analyze customer feedback at scale by extracting themes from unstructured text and joining that insight with usage metrics — all from a single SQL query.

The AI/BI Genie experience takes this further by enabling users to query data using plain English. Business users can ask questions like "What were our top 5 sales regions last quarter?" or "Why did churn increase in June?" and receive accurate answers backed by Databricks SQL and governed data. Genie integrates with Unity Catalog, ensuring access control and lineage are respected.

AI in the warehouse also supports more advanced use cases — such as powering recommendation engines or generating synthetic data for simulations. The lakehouse's unified architecture and AI-native capabilities allow organizations to seamlessly combine traditional BI with LLMs and predictive models, unlocking new possibilities for insight and automation.

FUNPLUS

FunPlus, one of the largest mobile gaming studios in the world, adopted Databricks to bring AI directly into the hands of its analytics and product teams. With a data platform powered by Databricks and integrated AI functions, they dramatically reduced the time it takes to find insights across massive volumes of player and game data. By enabling self-service AI and natural language querying for their analytics users, FunPlus accelerated decision-making across product, monetization and game operations.

"Our data team can create and deliver a Genie to our business team, allowing them to ask questions by themselves to query data."

— Chen Zhao, Product Director, FunPlus

FunPlus customer story

Learn more:

- Query LLMs With Databricks SQL
- AI Functions in Databricks
- YouTube: How Databricks Enables GenAI for BI

VECTOR SEARCH INTEGRATION WITH DATABRICKS SQL

One of the most powerful developments in AI-native analytics is the rise of Vector Search — a capability that allows data warehouses to retrieve results based on semantic similarity rather than exact matches. This technique powers applications like product recommendations, document retrieval, semantic search and AI copilots by comparing embeddings (vector representations) of text, images or structured data. Databricks now natively supports Vector Search in the lakehouse, bridging the gap between LLMs and traditional analytics.

With **Vector Search in Databricks SQL**, users can store, index and query high-dimensional vectors directly in Delta tables. These embeddings are typically generated from LLMs or other AI models and represent customer reviews, product descriptions, support tickets or marketing content. Once indexed, you can use the `VECTOR_SEARCH` SQL function to find the most similar entries to a given query — enabling use cases like "retrieve documents most similar to this user's last inquiry" or "suggest products based on this customer's past behavior."

Because Vector Search is fully integrated into Databricks SQL, it supports the same governance, performance optimizations and security as other lakehouse workloads. You can run hybrid queries that combine traditional filters with vector similarity — e.g., filtering for products in stock and ranking them by semantic match to a search query. This makes it easy to embed AI search and recommendation capabilities into existing dashboards and applications without requiring new infrastructure or duplication of data.

This integration is especially valuable in domains like customer support, marketing and R&D. For instance, support teams can surface similar tickets to reduce resolution time; marketers can discover campaigns that performed well under similar conditions; and product teams can mine user feedback for themes. By combining Vector Search with LLM-powered insights and SQL-native dashboards, the lakehouse becomes not just a repository of data — but a semantic engine for action.

Learn more:

- Create and Query Vector Search With Databricks
- Product Overview: Vector Search
- Azure Databricks Vector Search SQL Reference

SECTION 1.9 Cost Management and Optimization

One of the biggest concerns for analytics leaders today is managing cost without sacrificing performance or innovation. As analytics workloads grow — especially with AI and LLMs in the mix — so do costs. Databricks SQL offers several built-in capabilities to give teams visibility, control and optimization over their usage, supporting modern FinOps best practices.

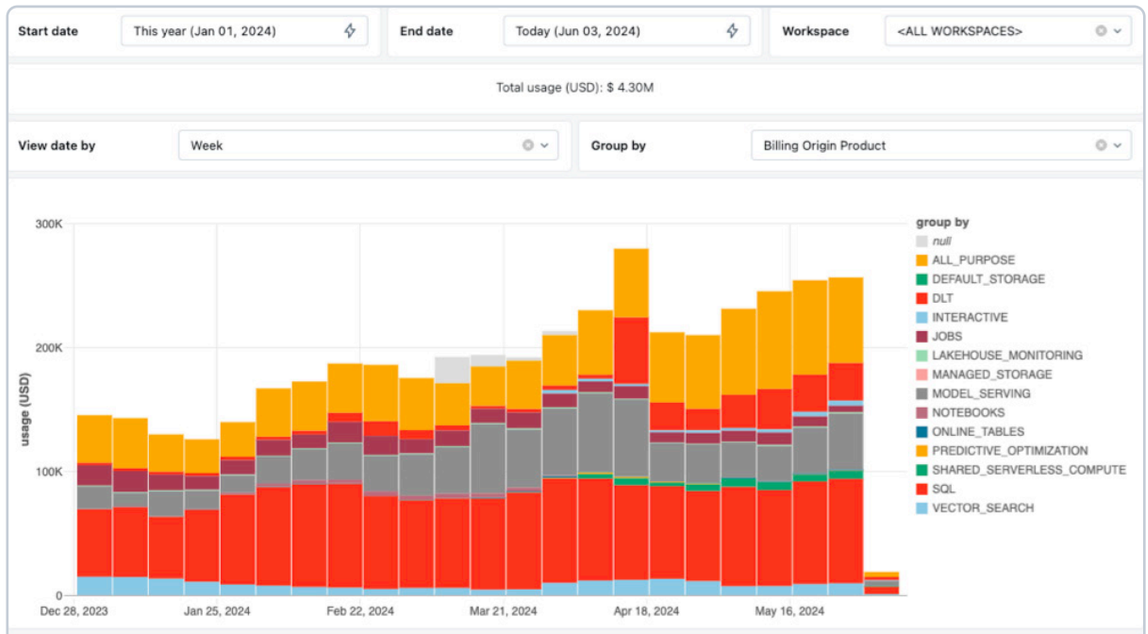
Databricks SQL includes a native cost monitoring UI, which allows administrators to see warehouse usage by persona, workload, dashboard and more. These insights are critical to identifying inefficiencies, eliminating zombie queries and enforcing policies that align with business goals. The integration with Unity Catalog also means usage can be tied directly to data access patterns, giving a more complete picture of spend.

Beyond visibility, Databricks provides automation tools to optimize cost. SQL Serverless warehouses scale automatically and pause when idle, reducing overhead from underutilized compute. Materialized views help cache high-frequency queries, improving both cost and performance. Predictive Optimization statistics use historical data to auto-tune query execution plans, eliminating manual tuning and reducing compute consumption.

The best results come from combining these features with a cost-conscious culture. Organizations are adopting FinOps frameworks that treat cost as a first-class metric — one that engineers, analysts and product teams monitor alongside performance. With Databricks' detailed usage logs and integrations with tools like Monte Carlo and Microsoft Purview, companies are building real-time cost dashboards and alerting systems that empower action.

Learn more:

- Guide to Cost Control for Databricks SQL
- Databricks SQL Cost Monitoring: The Ultimate Guide
- Overview of Cost Management on Databricks SQL
- DAIS 2025 — Cost Management Foundations
- DAIS 2025 — FinOps at Scale



Usage (USD) over past periods

Search group by...

group by	<ALL TIME>	5 periods ago	4 periods ago	3 periods ago	2 periods ago	last period	current period
		Apr 29 - May 06	*May 06 - May 13*	*May 13 - May 20*	*May 20 - May 27*	*May 27 - Jun 03*	*Jun 03 - Now*
<TOTAL>	$4,295,028	210K (-1%)	231K (+10%)	245K (+6%)	254K (+4%)	256K (+1%)	19K (-93%)
SQL	$1,628,471	73K (-2%)	80K (+10%)	77K (-4%)	83K (+7%)	84K (+2%)	6K (-93%)
ALL_PURPOSE	$1,034,418	59K (+5%)	69K (+17%)	79K (+14%)	76K (-3%)	69K (-9%)	4K (-94%)
MODEL_SERVING	$742,363	31K (+0%)	28K (-9%)	28K (-0%)	37K (+31%)	46K (+22%)	4K (-91%)
DLT	$279,231	17K (-23%)	27K (+54%)	33K (+21%)	30K (-8%)	31K (+2%)	3K (-91%)
JOBS	$253,818	9K (+8%)	9K (+2%)	10K (+11%)	8K (-18%)	5K (-36%)	216 (-96%)
VECTOR_SEARCH	$197,426	11K (-14%)	7K (-36%)	7K (+2%)	9K (+20%)	9K (+8%)	740 (-92%)
SHARED_SERVERLESS_COMPUTE	$42,474	5K (+72%)	7K (+23%)	6K (-3%)	5K (-16%)	6K (+18%)	525 (-92%)
NULL	$36,256	0	0	0	0	0	0

SECTION 1.10 Databricks SQL as Your Serving Layer

In the modern data stack, the line between analytical and operational workloads continues to blur. As organizations strive to serve data not only to analysts but also directly into business applications, the need for a reliable, performant serving layer becomes critical. Databricks is increasingly being used not just for data prep and analytics, but also as a serving platform powering real-time operational decision-making.

Databricks provides multiple options for serving data, including materialized views, Spark Declarative Pipelines and Lakehouse Federation — all governed through Unity Catalog. These features enable consistent delivery of data to BI tools, applications and APIs. For high-frequency, low-latency use cases — such as powering dashboards, recommendation systems or real-time operational intelligence — Databricks can serve data directly via Databricks SQL Serverless or REST APIs, without needing to push data into external systems.

One of the most powerful capabilities is integrating streaming and batch data pipelines with serving endpoints. Spark Declarative Pipelines ensures fresh data is always available, while serverless SQL endpoints scale elastically to meet demand. This architecture removes the need for separate serving databases and streamlines governance and lineage.

WORLD BANK GROUP

The World Bank uses Databricks as a core serving layer for delivering economic and development data globally. With Databricks, they manage hundreds of public-facing dashboards and APIs that provide real-time access to vital metrics across countries and sectors. These dashboards serve internal teams, policymakers and the public — all from a unified lakehouse foundation. This model allows the World Bank to modernize data delivery while ensuring governance, performance and scalability. The combination of Delta Lake, Unity Catalog and Databricks SQL ensures that data is accessible where it's needed most — securely and reliably.

"We run the dashboards for our operations from Databricks. We're trying to deliver 100% of our data to clients in a timely fashion using APIs and dashboards powered by Databricks."

— John Tynan, Lead, Development Data Partnership, World Bank

World Bank customer story

Learn more:

- YouTube: World Bank With Databricks
- YouTube: Accelerating Analytics: Integrating BI and Partner Tools to Databricks SQL

Additional integration tools and dashboards

Beyond the sample notebooks and dashboards listed earlier, Databricks SQL now includes several new integrations that make it easier for analysts and business users to interact with governed lakehouse data using familiar tools.

GOOGLE SHEETS ADD-ON

Teams can now connect their Google Sheets directly to Databricks SQL warehouses across all major clouds. The Google Sheets add-on enables spreadsheet users to query governed data in real time without manual exports or scripting. This allows for collaborative, lightweight analysis that's always up to date and backed by the full governance of Unity Catalog.

Technical documentation

AZURE DATABRICKS POWER PLATFORM CONNECTOR

The Azure Databricks connector for Power Platform enables real-time, secure access to Databricks data from Power Apps, Power Automate and Copilot Studio. Organizations can create governed workflows and apps without duplicating data or building custom APIs. The connector supports authentication through Microsoft Entra ID and integrates with Unity Catalog for fine-grained access control.

Power Platform Connector Blog

AUTOMATIC PUBLISHING TO POWER BI

A new Power BI task type in Lakeflow Jobs (Public Preview) allows teams to trigger Power BI dataset refreshes directly from Databricks pipelines. This ensures dashboards are always aligned with the latest data updates — no need to rely on time-based refreshes.

Automatic Power BI Publishing Blog

NEW OPEN SOURCE JDBC DRIVER

Databricks has also released a new open source JDBC driver with support for advanced data types, async execution and clusterless operations. This enhances connectivity for any tool or custom application that supports JDBC and facilitates deeper integration into enterprise data ecosystems.

GitHub: Databricks JDBC Driver

These new capabilities reinforce Databricks SQL's commitment to meeting users where they are — bringing lakehouse-powered analytics to the tools analysts, engineers and business users already know and trust.

SECTION 1.11 Optimizing Outcomes Across Your Data Estate

As organizations adopt Databricks, they're often doing so across a diverse, global data estate — spanning clouds, systems and business units. Databricks' unique strength lies in meeting customers where they are: offering native capabilities on every major cloud, supporting hybrid architectures and enabling interoperability across business suite application data like SAP.

AZURE DATABRICKS

Databricks and Microsoft have partnered deeply to deliver Azure Databricks: a first-party, fully managed service optimized for Azure environments. This tight integration enables organizations to take advantage of Azure-native tools such as Azure Active Directory, Azure AI Foundry, Microsoft Power BI, Azure OpenAI and Microsoft Power Platform.

Azure Databricks offers direct connectivity to Azure Data Lake Storage (ADLS), enabling faster ingestion and query performance. With built-in Unity Catalog integration, customers can apply RBAC and governance policies consistently across their lakehouse. Azure Monitor and Log Analytics help teams track performance, security events and usage — all within the Azure console.

Learn more: Azure Databricks Overview

Azure Databricks for BI and lakehouse workloads

Azure Databricks is a first-party, fully managed service delivered in partnership between Microsoft and Databricks. It combines the reliability, scale and security of Microsoft Azure with the performance and unified analytics experience of Databricks — making it the ideal solution for enterprises standardizing on Microsoft's cloud ecosystem. For BI workloads, Azure Databricks provides native integration with Microsoft tools while delivering powerful AI-native capabilities for the next generation of decision-making.

Azure Databricks offers seamless integration with Power BI. Data professionals can directly connect to Databricks SQL endpoints from Power BI Desktop or Service, using Azure Active Directory (AAD) for authentication. Thanks to Delta Lake and materialized views, dashboards reflect real-time data, and updates to datasets can be scheduled or triggered automatically. Unity Catalog extends governance into Power BI, ensuring that row- and column-level security policies are enforced end-to-end.

Enterprises benefit from a shared security and identity model. Azure Databricks supports native VNet injection, private link and AAD-based access controls, making it compliant with enterprise-grade security and regulatory requirements. Customers can use Microsoft Purview alongside Unity Catalog to manage data lineage and classification across the Databricks Platform and Microsoft services.

For analytics teams, Azure Databricks unlocks full interoperability with Microsoft services such as Azure Synapse, Azure Data Factory, Azure Event Hubs and Azure Machine Learning (AML). This means organizations can orchestrate complex workflows that combine Databricks data processing with Synapse visualization or ML deployments in AML — all under a single Azure subscription with unified billing.

Fonterra, one of the world's largest dairy exporters, harnessed Azure Databricks to unify its global data architecture and modernize analytics workflows. By integrating Delta Lake, Unity Catalog and Databricks SQL, Fonterra built a scalable, open platform that works seamlessly with the broader Microsoft Azure ecosystem — empowering self-service analytics within Power BI and beyond. The redesign led to over 50% faster data ingestion, 20% uplift in engineering productivity, and BI reports that now refresh in seconds rather than minutes.

"Azure Databricks works across the ecosystem. We've got a very rich data platform here that uses so many different aspects of the Microsoft stack. Databricks is open and orchestrates well across the different platforms."

— Helius Guimaraes, Chief Data and AI Officer, Fonterra

Fonterra customer story

Learn more:

- Why Azure Databricks Is the Best Foundation for Azure BI
- Power BI and Azure Databricks Integration
- Security and Compliance on Azure Databricks

DATABRICKS ON AWS

As organizations increasingly pursue multicloud strategies, Databricks on AWS offers a robust foundation for unified analytics, business intelligence and AI workloads in enterprises already invested in the Amazon ecosystem. AWS and Databricks have partnered to deliver a deeply integrated platform that supports scalable lakehouse deployments with high-performance SQL analytics, governed data sharing and advanced ML — all powered by AWS-native infrastructure.

Seamless integration with AWS services

Databricks leverages native AWS capabilities, including S3 for cloud object storage, Glue for metadata integration, and IAM for fine-grained identity and access management. This tight integration enables data teams to launch Databricks workspaces in their existing VPCs, reusing AWS networking, security and billing configurations. Amazon Redshift and Aurora users also benefit from Databricks Lakehouse Federation, enabling direct access to existing data sources through Unity Catalog with full lineage, auditability and role-based controls.

Optimized performance and reliability

Databricks clusters on AWS are optimized for performance and scalability using Amazon EC2 instances, including Graviton-based options for cost efficiency. With autoscaling capabilities, serverless compute (for SQL workloads) and built-in availability zones, Databricks on AWS can handle interactive BI queries as well as massive parallel data transformations. This enables real-time dashboarding in Tableau, Power BI and Databricks SQL without overprovisioning.

Governance and security at scale

Through Unity Catalog, Databricks provides unified governance for AWS-native and external data sources. It supports row- and column-level access controls, audit logs, and data lineage across BI tools and notebook-based analytics. Customers can also leverage AWS-native encryption (KMS), private link and secure networking to meet enterprise security and compliance standards.

bp

BP implemented Unity Catalog to centralize data governance across its Databricks lakehouse environment, enabling consistent access policies and auditability at scale. This allowed teams to manage data permissions, lineage and classification across regions and business units with greater control and transparency. Unity Catalog also laid the foundation for future capabilities like Delta Sharing, helping BP evolve its enterprise data strategy.

"Our Unity Catalog migration project, involving over 200 Databricks workspaces, has achieved outstanding results in cloud cost savings, governance and operational efficiency. By consolidating workspaces under Unity Catalog, we've significantly reduced operational costs, centralized governance, enhanced security and streamlined data sharing and compliance."

— Srinivas Chandolu, Staff Platform Engineer, BP

BP customer story

SAP ON DATABRICKS

SAP systems are at the heart of enterprise operations, from finance to supply chain to HR. Databricks helps unlock SAP data for analytics and AI by ingesting structured and semi-structured SAP data into the lakehouse in near real-time — enabling high-performance reporting and AI-driven forecasting.

HEINEKEN

"Our goal is to make Heineken the best-connected brewer for the digital-age consumer. To do this successfully, we developed a central source of truth for all our data, from consumers to suppliers to products. With SAP and Databricks, we're better able to leverage all of our data to better understand and serve our customers, optimize marketing campaigns, and drive operational efficiency across all markets where we operate."

— Jelle van Etten, Global Head of Data Platforms, Heineken

Introducing SAP Databricks

Learn more: SAP and Databricks at DAIS

SAP AND DATABRICKS: LAKEHOUSE ARCHITECTURE FOR ENTERPRISE APPLICATIONS

SAP systems hold some of the most valuable operational data in the enterprise — covering everything from finance and supply chain to human capital management. However, SAP's closed ecosystem, proprietary data formats and batch-oriented architecture have long limited data access, slowed innovation and restricted the use of modern analytics and AI. Databricks breaks open this black box with a native connection to SAP Business Data Cloud (BDC). SAP BDC is a fully SaaS solution that provides access to all SAP data across your most important business processes with curated and custom data products that eliminate the high costs of data extraction and replication. The connector to Databricks is built on the Delta Sharing protocol, which securely combines business-ready SAP data with the rest of the enterprise data from external data sources on one unified Databricks Platform to better collaborate and unlock value from that data for analytics and AI.

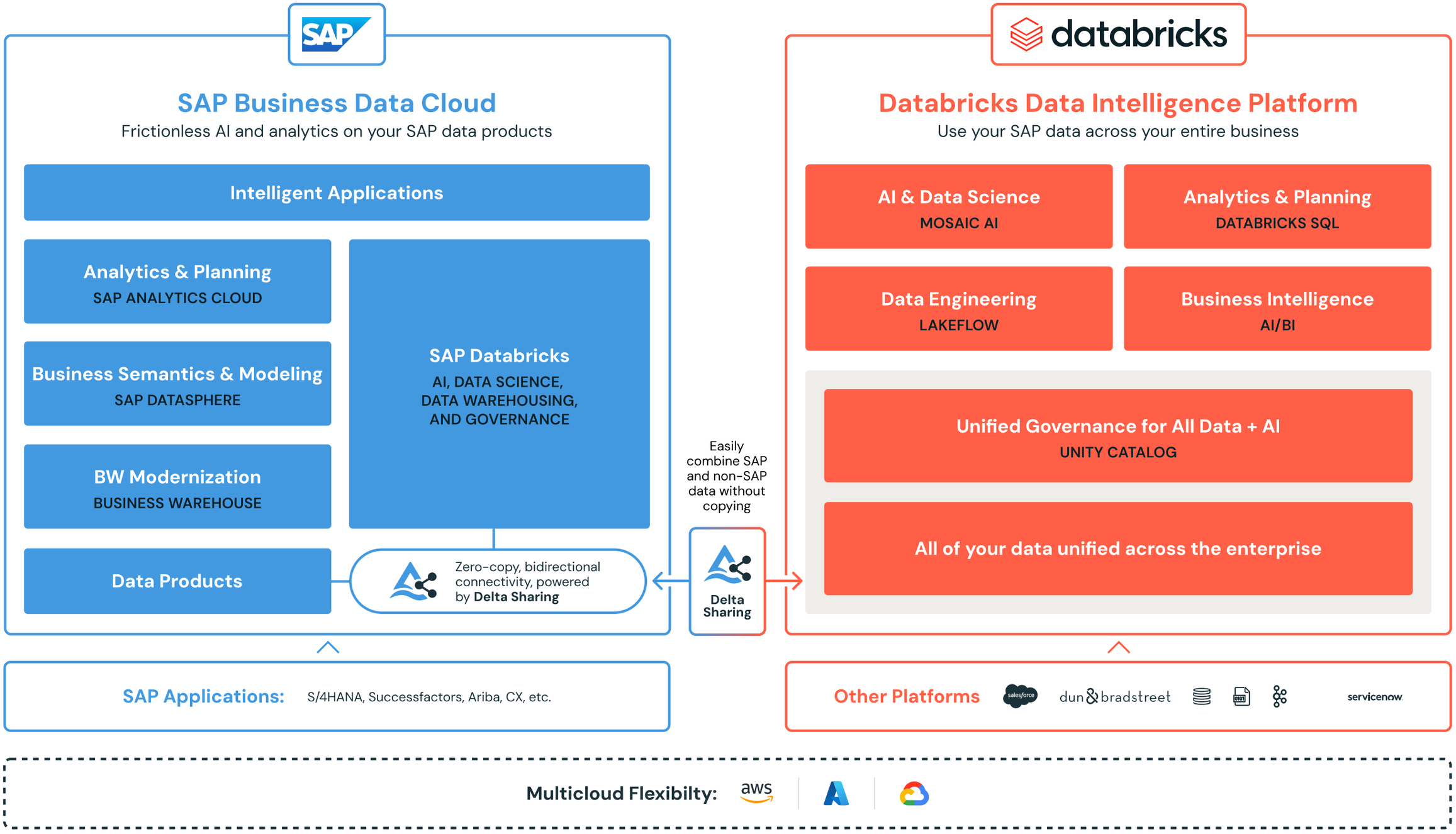
Databricks' approach to SAP unlocks SAP data for BI, ML and analytics across the entire enterprise. With ingestion patterns optimized for SAP BW, ECC and S/4HANA, data can be brought into the lakehouse via partners like Qlik or Fivetran, or custom ODP pipelines. Once ingested, Delta Lake enables high-performance querying, schema enforcement and ACID reliability for formerly locked-down SAP tables.

The real innovation comes from combining SAP data with external data sources. Finance teams can join SAP GL and AR records with CRM and market data. Supply chain teams can enrich SAP inventory and demand signals with IoT and logistics feeds. And all this is done in one platform — governed by Unity Catalog, secured by enterprise-grade controls, and accessible through Databricks SQL, Power BI and other tools.

In 2024, Databricks announced the SAP and Databricks Lakehouse Reference Architecture, which includes modular blueprints for finance analytics, supply chain intelligence and HR optimization. These patterns provide best practices for ingestion, transformation, semantic modeling and dashboard delivery using both SAP-native and Databricks-native tooling.

Learn more: How to Get Real-Time SAP Insights With Databricks SQL

databricks


SAP Business Data Cloud
Frictionless AI and analytics on your SAP data products
Intelligent Applications
Analytics & Planning
SAP ANALYTICS CLOUD
Business Semantics & Modeling
SAP DATASPHERE
BW Modernization
BUSINESS WAREHOUSE
Data Products
SAP Databricks
AI, DATA SCIENCE, DATA WAREHOUSING, AND GOVERNANCE
Zero-copy, bidirectional connectivity, powered by Delta Sharing
Easily combine SAP and non-SAP data without copying
Delta Sharing
Databricks Data Intelligence Platform
Use your SAP data across your entire business
AI & Data Science
MOSAIC AI
Analytics & Planning
DATABRICKS SQL
Data Engineering
LAKEFLOW
Business Intelligence
AI/BI
Unified Governance for All Data + AI
UNITY CATALOG
All of your data unified across the enterprise
SAP Applications: S/4HANA, Successfactors, Ariba, CX, etc.
Other Platforms
Multicloud Flexibilty:

SECTION 1.12 Performance and Workload Acceleration

For data-driven enterprises, performance is not a luxury — it's a prerequisite. Whether it's executive dashboards, financial reports or AI-driven analytics, speed determines how quickly teams can act. Databricks SQL delivers warehouse-grade performance and more through intelligent optimization, caching and a high-speed query engine.

At the core of performance improvements is Photon, Databricks' vectorized query engine built in C++. Photon dramatically accelerates SQL workloads — especially those that rely on aggregations, joins and window functions. Benchmarks show up to 12x speedups compared to legacy query engines. Photon is used automatically when running SQL queries in Databricks SQL warehouses, requiring no user intervention to benefit.

Materialized views are another key optimization feature. They allow teams to cache the results of expensive queries — especially those driving dashboards or executive reports — and serve them instantly to end users. These views are automatically refreshed and managed, reducing both cost and latency. Predictive Optimization statistics further enhance performance by analyzing query patterns and adjusting execution plans, join strategies and partition pruning without any manual tuning.

SQL Serverless is a key enabler of cost-effective performance. It offers instant startup time, autoscaling based on workload concurrency, and automatic suspension when idle. With serverless, data teams don't need to manage infrastructure or overprovision compute just to meet SLAs during peak hours. Databricks handles all provisioning behind the scenes, ensuring fast performance without the administrative burden.

Learn more:

- Databricks SQL Accelerates Workloads 5x in Three Years
- DAIS 2025 — Performance Tuning at Scale
- Announcing GA of Databricks SQL Serverless

SECTION 1.13 Governance and Security at Scale

As enterprises scale their analytics operations, ensuring that robust security, compliance and governance are in place becomes mission-critical. The Databricks Data Intelligence Platform was designed with enterprise-grade governance in mind — balancing flexibility for data teams with rigorous control for security and risk professionals.

Unity Catalog is the centerpiece of Databricks' unified governance model for data and AI. It provides centralized metadata, fine-grained access controls, business semantics, lineage tracking, data quality monitoring, and audit logging for all data, notebooks, AI models, and dashboards across your lakehouse architecture. Teams can define table-, row- or column-level permissions and apply role-based access controls across business domains. With support for dynamic views, tag policies and attribute-based access control (ABAC), it's easier than ever to implement scalable data security policies at the data catalog level.

Crucially, Unity Catalog is not a bolt-on — it's deeply integrated into the compute, storage and query layers of Databricks. This means that every query, every AI workload and every BI dashboard enforces the same governance rules automatically. Unity Catalog also supports open standards and APIs that enable secure data sharing and interoperability across tools, compute engines and platforms, helping teams break free from vendor lock-in and integrate seamlessly into their existing BI and data ecosystems.

With built-in discovery capabilities like business metadata, intelligent search and automated lineage, users can quickly locate, understand and trust the data and AI assets they need. This accelerates time to insight across analytics and BI workloads.

databricks

Databricks also supports modern identity federation and authentication protocols, including SAML, SCIM and OAuth. Integration with tools like Microsoft Entra ID and Okta makes it easy to align platform access with existing enterprise identity frameworks.

PEPSICO

Unity Catalog met PepsiCo's requirements for stringent security and sophisticated access controls. It has streamlined the onboarding process for more than 1,500 active users and enables unified data discovery for PepsiCo's 30+ digital product teams across the globe.

"Based on our experience, Databricks Unity Catalog has proven to be a scalable solution for centralized access management, data governance and data lineage management. Transitioning to Unity Catalog has streamlined our access control processes, reducing onboarding time by 30% and enhancing cost management. Additionally, with comprehensive data lineage capabilities, we have increased confidence in our data by being able to trace its origins and track any changes in real time. This transparency allows us to maintain high data integrity and reliability."

— Bhaskar Palit, Senior Director, Data and Analytics, PepsiCo

PepsiCo customer story

Learn more:

- Databricks Unity Catalog
- A Comprehensive Guide to Data and AI Governance

PART 2

Best Practices and Capabilities

SECTION 2.1 From Dashboards to Decisions: Evolving the Role of BI

The business intelligence landscape is undergoing a radical transformation. Traditional BI tools and dashboards — once the cornerstone of decision-making — are being reimagined to support real-time, AI-infused insights across every part of the business. As data volumes explode and executive expectations rise, dashboards alone are no longer enough.

Organizations are moving toward intelligent BI experiences, powered by the Databricks Data Intelligence Platform. Rather than static reports, today's BI is interactive, dynamic and driven by context. Features like dynamic calculations, GenAI-infused dashboards and natural language query interfaces (like Genie) mean business users can explore data without relying on a data analyst for every change.

This shift requires tight integration between data engineering, data science and business teams. Databricks uniquely supports this convergence by bringing all personas onto one platform — with governed access to shared data, lineage tracking and AI-augmented interfaces. BI users become active participants in generating insights rather than passive consumers.

Modern BI is not just about accessibility — it's about adaptability. With Databricks' AI-native capabilities, teams can build dashboards that not only answer questions but also suggest new ones. Genie leverages LLMs to generate dashboards from plain English, dynamically update visualizations as the user explores the data, and summarize trends or anomalies. This means fewer hours spent dragging charts around — and more time driving outcomes.

Databricks also supports powerful integration with tools like Power BI, Tableau and Looker, ensuring teams can work in familiar environments while benefiting from the Databricks Platform's performance and scale. By leveraging Unity Catalog for consistent definitions and semantic layers, organizations reduce data chaos and foster trust in BI outputs.

BI's role has expanded from rear-view reporting to forward-looking intelligence. Whether it's marketing departments adjusting campaigns in real time or sales teams seeing opportunity pipelines evolve live, modern BI requires agility, trust and intelligence embedded into every step.

Learn more:

- BI in the Era of AI — Webinar
- Introducing Databricks One
- Genie Now Generally Available
- Genie Conversation APIs
- Author Faster, Smarter Dashboards

SECTION 2.2 The Role of the Data Analyst and Data Engineer in the New BI Stack

The rise of the lakehouse and AI-native BI doesn't eliminate the role of the data analyst or data engineer — it elevates it. In modern data platforms like Databricks, these roles are becoming more strategic, collaborative and central to driving business impact.

Data analysts are no longer just dashboard builders. With Databricks SQL, they operate as insight enablers — building semantic layers, shaping models and embedding intelligence into dashboards that are self-service and AI-enhanced. Their role now includes curating metrics with Unity Catalog, collaborating with data engineers on performance tuning, and providing governance guardrails for business teams.

Meanwhile, data engineers are empowered by new declarative tooling like Lakeflow Spark Declarative Pipelines, which reduces the overhead of writing and maintaining brittle pipelines. They design robust streaming- and batch-capable pipelines that power both real-time and scheduled reporting. With strong integration into Unity Catalog, data engineers help enforce lineage, quality and access policies without friction.

Collaboration between these two roles has never been more important. Analysts bring context and business logic; engineers bring scalability and data quality. In Databricks, shared notebooks, managed governance and a single runtime mean these teams operate together on the same data, breaking down silos that existed in traditional BI environments.

Learn more:

1. **Databricks SQL Product Page**
   Learn how analysts and engineers collaborate with Databricks SQL for governed self-service analytics.
   https://www.databricks.com/product/databricks-sql
2. **Unity Catalog Overview**
   Explore how Unity Catalog enables fine-grained access control and shared metrics governance.
   https://www.databricks.com/product/unity-catalog
3. **Introducing Lakeflow Spark Declarative Data Pipelines**
   Blog post on how data engineers can use Spark Declarative Pipelines to simplify extract, load, transform (ELT).
   https://www.databricks.com/product/data-engineering/spark-declarative-pipelines
4. **Databricks Notebooks**
   Documentation showing how shared notebooks drive collaboration between analysts and engineers.
   https://docs.databricks.com/en/notebooks/index.html

SECTION 2.3 Migration Strategies and Lessons Learned

Migrating to the Databricks Data Intelligence Platform can unlock significant improvements in performance, scalability, governance and cost-efficiency — but successful outcomes require careful planning, best-practice architecture and organizational alignment. Whether moving from a traditional data warehouse or a cloud-native solution, Databricks offers prescriptive strategies and tooling to streamline the journey.

Migration starts with understanding the current state of your data warehouse: what workloads exist, who the users are, what data models are in play and how governance is enforced. From there, you can develop a migration roadmap that targets high-impact use cases first — typically analytical workloads, batch pipelines, or dashboards with performance or scale issues.

Databricks provides migration tooling through Lakebridge, which accelerates the conversion of schema, lineage, data and queries from legacy data warehouses like Snowflake, Redshift, SQL Server and Oracle. The tooling supports bulk export/import, syntax translation and automated testing, enabling teams to validate outcomes before decommissioning legacy platforms.

Best practices:

- **Phased migration:** Start with analytical and reporting workloads that don't require real-time operational dependencies. These are easier to lift and shift to the lakehouse without major application refactoring.
- **Schema translation and modeling:** Use Databricks' dimensional modeling guidance to translate star/snowflake schemas to Delta Lake formats. Tools like dbt and Unity Catalog make this process maintainable and secure.
- **Test early, test often:** Implement test harnesses that compare row counts, aggregates and business KPIs between legacy and lakehouse systems. Ensure each migrated workload is functionally equivalent.
- **Business alignment:** Involve business stakeholders from day one. Validate that the migrated data meets their reporting and compliance requirements.
- **Training and enablement:** Prepare analysts and engineers with training on Databricks SQL, Unity Catalog and lakehouse best practices to ensure post-migration success.

Learn more:

- Databricks Migration Strategies eBook
- Lakebridge Migration Tooling

SECTION 2.4 Dimensional Modeling in the Lakehouse

Dimensional modeling has long been the backbone of business intelligence, allowing organizations to build intuitive, performant and explainable data models. Traditionally associated with enterprise data warehouses, dimensional modeling is now being reimagined for the modern lakehouse. With Databricks SQL, you can apply best practices from Kimball-style modeling — like star schemas and slowly changing dimensions (SCDs) — while taking advantage of the scale and flexibility of Delta Lake.

The Databricks Platform supports both batch and streaming data sources, making it possible to model facts and dimensions that reflect real-time business events. With Spark Declarative Pipelines, data engineers can define dimensional pipelines declaratively, automating the orchestration of Type 1 and Type 2 SCDs. These patterns are reusable, testable and optimized for reliability at scale.

Data analysts benefit from a governed semantic layer on top of these dimensional structures. Using Unity Catalog and dbt, teams can define canonical metrics (e.g., revenue, churn, margin) once and reuse them across BI tools like Power BI, Tableau and Looker. This eliminates metric drift, increases stakeholder trust and improves self-service analytics across the business.

Best practices:

- Use CDC and Spark Declarative Pipelines to ingest source system changes in near real-time
- Implement SCD Type 2 for history-aware dimensions where auditability is important
- Use surrogate keys to maintain clean joins and track lineage across transformations
- Separate business-friendly naming (presentation layer) from raw staging tables for flexibility and security

Learn more:

- Blog: Implementing Dimensional Modeling on Databricks
- YouTube: Dimensional Modeling With Databricks SQL

SECTION 2.5 Data Quality and Observability in BI Workloads

As BI workloads grow more complex and real-time, the need for trust in data becomes paramount. Dashboards and reports are only as good as the data powering them. Errors in upstream pipelines or unnoticed schema changes can result in broken visuals, incorrect metrics or worse — business decisions based on false premises. That's why data quality and observability are essential to modern BI.

Databricks provides built-in tools and ecosystem integrations to monitor, alert and remediate data quality issues across the Databricks Platform. Spark Declarative Pipelines includes declarative quality checks that allow engineers to define expectations — such as null constraints or value ranges — directly within pipelines. Failures trigger automatic alerts or quarantine logic, keeping bad data from propagating downstream.

For more advanced observability, customers are integrating tools like Monte Carlo, Datafold and Great Expectations with the lakehouse. These tools plug into Delta logs and Unity Catalog metadata to detect schema drift, freshness anomalies or unexpected volume changes in near real-time. With Databricks' open APIs and lineage features, observability platforms can map data issues to affected tables, dashboards and business owners.

Unity Catalog helps enforce trust at the governance level. By tracking data lineage, catalog changes and user access events, organizations can ensure that only authorized, high-quality data is available to BI users. When issues arise, teams can quickly trace errors to source systems or recent code changes.

Learn more: Data Quality at Scale — DAIS Session

SECTION 2.6 Managing the Semantic Layer With Unity Catalog

The semantic layer is the connective tissue between raw data and business understanding. It provides the consistency, clarity and trust needed to scale self-service analytics. In traditional environments, semantic layers were often managed separately within each BI tool — resulting in metric drift, duplicate logic and governance gaps. With Unity Catalog, Databricks offers a unified semantic layer directly within the lakehouse.

Unity Catalog allows teams to define business entities — such as revenue, active users or gross margin — as curated views or tables. These assets are version controlled, governed and discoverable across the enterprise. Analysts can access them from tools like Power BI, Tableau or Databricks SQL, knowing the logic behind the metrics is standardized and certified.

Because Unity Catalog enforces role-based and attribute-based access control at the column level, it enables fine-grained security while supporting broad self-service. Teams can define dynamic views that restrict data by geography, role or customer segment — ensuring the right users see the right data, and only that data.

Best practices:

- Centralize metric logic in dbt models or views published to Unity Catalog, and annotate with descriptions and ownership metadata
- Use tags and labels to support discovery and impact analysis across BI and ML
- Establish a semantic review process to validate critical KPIs before they are made widely available
- Implement dynamic filters to personalize dashboards while enforcing privacy policies

Learn more:

- **Unity Catalog Product Page**
  Core overview of Unity Catalog's features, including centralized governance, data lineage and access control.
  https://www.databricks.com/product/unity-catalog
- **Introducing Lakehouse Federation Capabilities in Unity Catalog**
  Blog post covering semantic layer unification and federation across external systems.
  https://www.databricks.com/blog/introducing-lakehouse-federation-capabilities-unity-catalog
- **Unity Catalog Technical Documentation**
  Deep dive into creating and managing views, access control, row- and column-level security, and tagging.
  https://docs.databricks.com/en/data-governance/unity-catalog/index.html
- **Best Practices for Data Governance in Unity Catalog**
  Practices for structuring semantic assets, managing tags and ensuring consistent governance.
  https://docs.databricks.com/aws/en/data-governance/unity-catalog/best-practices

SECTION 2.7 Best Practices: End-to-End Streaming ELT on Databricks SQL

Streaming ELT on Databricks SQL empowers data teams to deliver fresh, low-latency insights while simplifying orchestration and reducing pipeline complexity. Drawing from the *Guide to Data Warehousing in the Lakehouse*, this section highlights design patterns and operational principles for streaming use cases that feed BI dashboards.

Unified streaming and Databricks SQL now support continuous ingestion of streaming data into Delta Lake tables using Auto Loader and Structured Streaming. Combined with Unity Catalog and materialized views, teams can build streaming pipelines that publish real-time insights directly into queryable SQL endpoints. Business users can interact with these datasets in tools like Power BI and Tableau without needing to understand the complexity behind them.

ARCHITECTURE BLUEPRINT

An end-to-end streaming ELT architecture in the lakehouse typically consists of:

- Data ingestion from Kafka, IoT and CDC systems using Auto Loader
- Bronze layer for raw storage with schema evolution
- Silver layer for transformation and normalization using SQL queries
- Gold layer with materialized views or streaming tables for BI consumption

This layered model follows the medallion architecture and aligns with modular, testable data engineering principles.

OPERATIONAL BEST PRACTICES

- Use Auto Loader for incremental ingest, minimizing file listing overhead
- Implement schema inference and evolution to handle upstream changes
- Create streaming tables or materialized views for low-latency query performance
- Apply Unity Catalog for permissioning, lineage and auditability across layers
- Monitor query freshness and SLAs using Databricks SQL dashboards

Shell has implemented streaming ELT pipelines on Databricks to monitor IoT sensor data from their energy assets. With Bronze-to-Gold workflows built in Databricks SQL, analysts across operations teams gain live access to metrics like equipment health and energy output, reducing downtime and enabling real-time alerts.

Shell customer story

This pattern is extensible across industries — from digital ads and fintech to healthcare, where latency and freshness matter. By leveraging SQL-native constructs, teams reduce complexity, avoid duplicate logic and democratize access to streaming insights.

SECTION 2.8 Operational Analytics and Supply Chain Intelligence

Modern supply chains are complex, multitiered and dynamic. Operational leaders need real-time visibility into production, inventory, shipping and demand forecasting — yet most traditional BI systems struggle to deliver the agility and integration required. With Databricks, operational analytics becomes a proactive capability, not just a rear-view report.

Lakehouse architecture allows organizations to combine streaming IoT data from factories and warehouses with historical ERP records and third-party logistics (3PL) feeds. Auto Loader and Spark Declarative Pipelines automate the ingestion and transformation of these diverse data sources into a unified Delta Lake format. Business units can then build analytics dashboards, root-cause reports or even AI-based anomaly detectors on top of the same governed datasets.

For example, predictive analytics and ML models can forecast demand spikes or supplier delays. By training models on historical supply chain data, macroeconomic indicators, and weather or geopolitical signals, organizations can take preemptive action — rerouting shipments, accelerating inventory staging or adjusting production plans. These outputs feed directly into Databricks SQL dashboards that guide daily operations.

Unity Catalog ensures secure access to sensitive operational data. Attribute-based access control can restrict views by plant, vendor or product line while preserving a global view for corporate HQ. With the semantic layer in place, supply chain KPIs like fill rate, on-time in-full (OTIF) and average order cycle time become universally understood and trusted.

Learn more:

- **Real-time Analytics**
  Understand real-time analytics and data streaming
  https://www.databricks.com/glossary/real-time-analytics
- **Lakeflow Spark Declarative Pipelines Product Page**
  Learn how to simplify batch and streaming ETL with automated reliability and built-in data quality.
  https://www.databricks.com/blog/whats-new-lakeflow-declarative-pipelines-july-2025
- **Auto Loader Documentation**
  Learn how Auto Loader incrementally and efficiently processes new data files as they arrive in cloud storage without any additional setup.
  https://docs.databricks.com/en/ingestion/auto-loader/index.html
- **Unity Catalog for Secure Data Access**
  Unity Catalog overview, including attribute-based access controls for operational and manufacturing datasets.
  https://docs.databricks.com/en/data-governance/unity-catalog/index.html

SECTION 2.9 BI for Product and Growth Teams

Product and growth teams are constantly pressured to deliver results — optimize conversion rates, boost user engagement, reduce churn and personalize the experience. However, their success hinges on rapid, trusted access to user behavior and experimentation data. With Databricks, product-led organizations can harness the full spectrum of data — from raw event logs to predictive models — without waiting for days of ETL or analyst backlogs.

The Databricks Data Intelligence Platform supports product analytics at scale. By ingesting clickstream, mobile and server-side events using tools like Auto Loader and Kafka, teams can create sessionized, structured datasets in Delta Lake. These datasets power funnel analysis, cohort tracking, retention dashboards and attribution models. With Unity Catalog, these assets can be shared securely across product, data and marketing teams — governed and standardized.

Beyond retrospective analysis, growth teams are increasingly relying on real-time decisioning. Personalization models — whether for homepage content, recommendation engines or email campaigns — can be trained and deployed directly within Databricks. With features like feature store and MLflow, teams can iterate faster and deploy models with lineage and performance tracking baked in.

Experimentation is another core capability. Databricks supports A/B testing at scale, with statistical significance modeling, experiment assignment tracking and outcome attribution. By unifying experimentation data with behavioral telemetry, teams can identify winning variants, run more tests and feed insights back into product loops.

Learn more:

- **Auto Loader for Ingesting Event Streams**
  Documentation on using Auto Loader to ingest clickstream, mobile and server-side events at scale.
  https://docs.databricks.com/en/ingestion/auto-loader/index.html
- **Streaming Analytics on the Lakehouse**
  Learn how to build real-time pipelines that support product metrics, funnel tracking and user telemetry.
  https://www.youtube.com/watch?v=MAOG1qvE4Bk&t=38s
- **Unity Catalog for Secure Data Sharing**
  Enable governed, shared access to product data and metrics across product, marketing and engineering.
  https://www.databricks.com/product/unity-catalog
- **A/B Testing With Databricks SQL and ML**
  Example approaches to scalable A/B experimentation using notebooks, SQL analytics and ML.
  https://docs.databricks.com/aws/en/machine-learning/mlops/mlops-workflow

SECTION 2.10 BI for Marketing, Sales and Customer Success

In today's competitive landscape, go-to-market (GTM) teams depend on data to drive decisions — across campaign targeting, lead qualification, customer segmentation and churn prevention. Yet many organizations still struggle with siloed CRM systems, inconsistent reporting definitions and slow dashboard performance. The Databricks Data Intelligence Platform changes the game by unifying GTM data, enabling predictive insights and democratizing access across teams.

With Databricks, marketing teams can ingest data from tools like Marketo, Google Ads and LinkedIn Ads, and combine it with Salesforce, Zendesk and product usage data — all in Delta Lake. This gives a 360-degree view of the customer journey: from anonymous visitor to closed-won opportunity to support interactions. With Unity Catalog, GTM leaders can create governed datasets that fuel dashboards, models and AI assistants.

Predictive models built on Databricks help prioritize high-intent leads, recommend the next best actions and forecast pipeline health. For instance, ML models can identify customer segments most likely to respond to campaigns or accounts most likely to churn, triggering personalized outreach. These models can be embedded into CRM tools or consumed via dashboards.

Databricks SQL and materialized views ensure marketing, sales and customer success teams get fast access to insights. Teams can drill into funnel stages, attribution paths or NPS trends without waiting for a data analyst. AI-powered tools like Genie let GTM users ask questions in plain language, summarizing account health, campaign ROI or regional sales trends.

Learn more:

- **Databricks for Marketing Analytics**
  Solution overview of how Databricks enables modern marketing analytics across attribution, segmentation and campaign optimization.
  https://www.youtube.com/watch?v=WO3msSDE7zY
- **Unity Catalog Product Page**
  Central hub for governed data sharing, lineage and fine-grained access controls across BI and ML use cases.
  https://www.databricks.com/product/unity-catalog
- **AI/BI Genie Blog: Onboarding Your New AI/BI Assistant**
  Describes how Genie transforms the way GTM teams interact with data through conversational analytics.
  https://www.databricks.com/blog/onboarding-your-new-aibi-genie
- **Databricks Connector for Salesforce and Marketing Platforms**
  Information on how Lakeflow ingests CRM and marketing data into Delta Lake.
  https://www.databricks.com/product/data-engineering/lakeflow-connect
- **Materialized Views for Dashboard Acceleration**
  Blog post explaining how materialized views improve dashboard performance for business teams.
  https://www.databricks.com/blog/introducing-materialized-views-and-streaming-tables-databricks-sql

SECTION 2.11 BI for Finance and HR Teams

Finance and human resources teams are increasingly leaning on data to deliver strategic insights, not just operational reports. They require accurate, secure and timely access to KPIs around cash flow, head count, payroll, compliance and forecasting. However, many finance and HR teams remain burdened by siloed systems like SAP, Workday, Oracle and Excel spreadsheets, which limit agility and trust in the numbers. Databricks solves this with a unified, governed analytics platform tailored for enterprise needs.

With lakehouse architecture, finance and HR data from disparate systems can be ingested into Delta Lake, cleaned and modeled into dimensional structures that support both reporting and predictive analytics. Spark Declarative Pipelines simplifies complex logic like payroll period transformations, cost center mappings and multi-currency normalization. Once transformed, these datasets are governed using Unity Catalog and served securely to analysts and stakeholders through Databricks SQL.

Finance teams use Databricks to build robust forecasts — covering revenue, expense and cash flow — leveraging historical GL data, macroeconomic indicators and scenario modeling. ML models built on Databricks enable rolling forecasts, budget variance analysis and fraud detection. HR teams, meanwhile, use the platform to monitor head-count changes, predict attrition, optimize recruiting pipelines and analyze DEI metrics. Genie allows both teams to ask natural language questions about KPIs, improving adoption and reducing dependence on ad hoc reports.

Security and compliance are critical in this domain. Unity Catalog provides fine-grained access control over sensitive fields like compensation, benefits and personally identifiable information (PII). Audit logging, data lineage and tokenization ensure compliance with GDPR, SOX and other regulations.

Learn more:

- **The Data Intelligence Platform for Financial Servicess**
  Unleash data-driven decisions and AI innovation for a competitive edge in financial services
  https://www.databricks.com/solutions/industries/financial-services
- **Lakehouse Architecture Overview**
  Understand the data lakehouse design pattern for analytics across enterprise domains, including finance and HR.
  https://www.databricks.com/product/data-lakehouse
- **Lakeflow Spark Declarative Pipelines Product Page**
  Learn how to simplify batch and streaming ETL with automated reliability and built-in data quality.
  https://www.databricks.com/blog/whats-new-lakeflow-declarative-pipelines-july-2025
- **Unity Catalog Product Page**
  Detailed explanation of governance and compliance tooling, like access control, lineage and audit logs.
  https://www.databricks.com/product/unity-catalog
- **AI/BI Genie for Business Teams**
  How finance and HR teams can interact with KPIs and dashboards using natural language with Genie.
  https://www.databricks.com/blog/onboarding-your-new-aibi-genie

SECTION 2.12 BI for IT, Support and Service Teams

IT, customer support and service delivery teams generate massive amounts of data — from ticketing systems and call logs to system telemetry and configuration changes. Making sense of this data requires a unified platform that can correlate operational events with user impact and enable proactive incident prevention. Databricks empowers these teams to move from reactive troubleshooting to predictive operations.

With Databricks, organizations can ingest data from tools like ServiceNow, Jira, Zendesk, Splunk and Datadog into Delta Lake. Combining this data with product telemetry or CRM records allows support teams to understand incident root causes, identify common patterns and prioritize high-impact resolutions. By correlating service tickets with system performance, teams can proactively address emerging issues before they escalate.

ML models built on Databricks can predict ticket volume spikes, detect anomalies in support metrics (like resolution time or CSAT) and even classify tickets for triage. This reduces response times, improves SLAs and ensures customers receive consistent support. Genie adds another layer by allowing agents and managers to query support metrics using natural language, without needing SQL or dashboard skills.

IT teams also use Databricks to optimize infrastructure. By analyzing logs, configuration changes and usage patterns, they can right-size cloud resources, detect inefficient services and identify underutilized assets. Databricks SQL supports fast dashboarding on this data, making it accessible to business and IT stakeholders alike.

Learn more:

- **Unity Catalog Product Page**
  Covers governance and fine-grained access controls critical for secure IT and support analytics.
  https://www.databricks.com/product/unity-catalog
- **Genie Blog: Author Faster, Smarter AI/BI Dashboards**
  Demonstrates how nontechnical users, like support managers, can use Genie to ask natural language questions of support KPIs.
  https://www.databricks.com/blog/author-faster-smarter-aibi-dashboards-dynamic-calculations
- **Databricks Blog: 5 Lessons From Implementing AI/BI Genie for Self-Service Marketing Insights**
  Shows how customer-facing teams apply Genie and natural language BI to reduce ticket backlog and drive faster answers.
  https://www.databricks.com/blog/5-key-lessons-implementing-aibi-genie-self-service-marketing-insights
- **Auto Loader Documentation**
  Learn how Auto Loader incrementally and efficiently processes new data files as they arrive in cloud storage without any additional setup.
  https://docs.databricks.com/aws/en/ingestion/cloud-object-storage/auto-loader/
- **Lakeflow Spark Declarative Pipelines Product Page**
  Learn how to simplify batch and streaming ETL with automated reliability and built-in data quality.
  https://www.databricks.com/product/data-engineering/spark-declarative-pipelines

SECTION 2.13 Embedded Analytics and Customer-Facing Insights

Many modern software companies and digital businesses are embedding analytics directly into their products — giving customers, partners and external stakeholders access to reports, dashboards and insights within the context of their daily workflows. Embedded BI not only enhances user value but also creates new monetization opportunities and deepens product stickiness. Databricks provides a robust foundation for delivering these experiences securely and at scale.

With Databricks, teams can power embedded analytics using the same governed lakehouse data used internally. Dashboards and queries created in Databricks SQL can be shared externally using APIs or embedded securely within applications via iframe or REST endpoints. Because these visualizations sit atop Unity Catalog-managed datasets, customers can be granted role- or attribute-based access to only their relevant data — whether it's scoped to tenant, geography or product tier.

Materialized views and streaming tables allow embedded experiences to reflect real-time data without degrading performance. Whether it's usage analytics for SaaS customers, order histories for suppliers, or campaign results for advertisers, Databricks ensures responsiveness and reliability. Organizations can also expose metrics programmatically to third-party applications or customer-facing APIs.

Security and scalability are key differentiators. With serverless Databricks SQL warehouses, teams can support high concurrency without overprovisioning compute. Built-in authentication support (including OAuth and token-based access) enables seamless SSO for end users. Dashboards remain interactive, governed and performant — no matter how many tenants are online.

Learn more:

- Embedding Dashboards With Databricks SQL
- Scaling BI with SQL Serverless

SECTION 2.14 Onboarding Your AI/BI Genie

Databricks AI/BI Genie redefines the analyst experience by making it easier to build, explore and iterate on dashboards using conversational interfaces. For technical and nontechnical users alike, Genie acts as a productivity multiplier — speeding up insights with natural language prompts, dynamic calculations and embedded guardrails.

Key onboarding practices:

- Set up a discovery layer using Unity Catalog to expose high-value Gold tables
- Align prompt engineering with key BI use cases to generate relevant visualizations
- Leverage Genie's explainability tools to help users trust and understand its outputs

Teams report that AI/BI Genie reduces time to dashboard by over 60% and accelerates self-serve access to KPIs across sales, marketing and finance.

Learn more:

- Onboarding Your New AI/BI Genie

SECTION 2.15 Lessons Learned From AI/BI Genie in Marketing

At Databricks, the internal marketing analytics team embarked on a mission to democratize data through a natural language BI assistant called Genie. Built on Databricks SQL and powered by generative AI, Genie allows nontechnical marketers to ask questions in plain English and receive governed, accurate answers from a shared semantic layer. This effort aimed to reduce the analytics bottleneck by empowering users with self-service access to insights — without needing to learn SQL or rely on analysts for ad hoc requests.

One of the most important lessons was the need for a strong semantic foundation. Genie uses Databricks SQL dashboards and Unity Catalog's semantic layer to ensure consistent business logic across questions and users. Rather than generating answers from raw tables, Genie was trained on preapproved queries and dashboards, improving answer quality and reliability. The team emphasized investing up front in modeling the data into a curated structure of metrics, dimensions and KPIs that business users actually care about.

The project also required close collaboration between data engineers, analysts and marketers to co-develop prompts and iterate quickly. Genie's answers were grounded in specific dashboards so users could click into the source and validate the results. This transparency helped build trust in the AI assistant's responses and made it easier to debug or refine questions. One practical example included a marketing director asking about the impact of email campaigns — Genie translated the query, surfaced key engagement metrics and linked directly to the dashboard showing performance trends.

Training and onboarding were essential to drive adoption. The team ran enablement sessions to teach marketers how to write effective natural language prompts and what types of questions Genie could answer. They also emphasized communicating Genie's limitations — it's not a magic box but rather a new interface for interacting with existing governed data. Success hinged on aligning expectations, continuously improving based on feedback, and encouraging curiosity among business users.

Overall, Genie's implementation led to faster decisions, reduced dependency on analysts for simple questions and more consistent use of trusted data sources. It demonstrated the power of combining Databricks SQL, Unity Catalog and generative AI to make analytics more accessible and impactful. As the system evolves, the team plans to expand Genie's reach to more departments and continue iterating on prompts, performance and data coverage. This project shows what's possible when AI meets well-modeled data and cross-functional collaboration.

Learn more:

- 5 Key Lessons From Implementing AI/BI Self-Service for Marketing Insights

SECTION 2.16 AI-Powered Customer Sentiment Analysis: Best Practices

Understanding customer sentiment is critical to enhancing product development, marketing and support. Databricks provides an end-to-end solution for customer sentiment analysis powered by AI — allowing teams to ingest, process and analyze customer feedback at scale.

The process begins by ingesting raw data from surveys, reviews and support channels using Auto Loader. This data is transformed into structured formats using Spark Declarative Pipelines (formerly known as DLT), which handles parsing, deduplication and normalization. Next, sentiment classification is performed using AI functions in Databricks SQL or fine-tuned models in MLflow, enabling the detection of positive, negative or neutral sentiments.

To contextualize sentiment, Databricks recommends extracting product and service topics using named entity recognition (NER) or custom text parsing. This allows teams to see which aspects of their offerings are driving sentiment. These enriched datasets are stored in Unity Catalog and visualized in Databricks SQL dashboards with trend lines, keyword clouds and customer journey analysis.

Best practices include establishing a feedback ingestion pipeline using Spark Declarative Pipelines and Auto Loader, implementing AI models with governance (including human review where necessary), and exposing curated outputs through semantic views. Teams should also implement alerting on sentiment dips and tie sentiment scores to NPS or CSAT metrics to detect systemic issues.

Learn more:

- Step-by-Step Guide to AI-Powered Customer Sentiment Analysis

SECTION 2.17 Best Practice: Real-Time Population Health Monitoring With NHS England

NHS England, the publicly funded healthcare system for over 55 million people, partnered with Databricks to modernize its national data infrastructure into a unified, cross-cloud virtual data layer. This architecture integrates diverse sources across Azure and AWS using Unity Catalog, Delta Sharing and Lakehouse Federation, enabling real-time access to operational, clinical and administrative datasets. The project established a single access point for the Federated Data Platform (FDP) used by Palantir, minimizing data duplication while preserving governance.

Databricks handles data ingestion, transformation and processing — acting as the "kitchen" that prepares datasets to be consumed by FDP dashboards and analytics. This setup provides secure, low-latency analytics across the NHS's sprawling estate. Data teams benefit from centralized governance and scalable compute while maintaining the flexibility to work with the tools and clouds of their choice.

> "Databricks is at the NHS already involved in data collection services, data processing services, secure research environments. … We adopted the role of the back-end kitchen — handling raw ingestion, transformation and processing — and handed over ready-made meals to the FDP to serve up to analysts for consumption."
>
> — Mike Dobing, Specialist Architect, Databricks (speaking as former Lead Architect for the NHS project)

YouTube: How NHS England Uses Databricks

This collaboration between Databricks, NHS England and Palantir showcases a successful model for secure, interoperable data platforms in public health.

SECTION 2.18 Best Practice: Scalable, Governed Product Analytics at PicPay

PicPay, one of Brazil's largest digital wallets, serves over 30 million users across a rapidly evolving suite of financial services. With growth came scale — and with scale came complexity. The company faced a growing backlog of analytical requests, data duplication across teams and increasingly siloed insights.
To transform data into a competitive advantage, PicPay adopted the Databricks Data Intelligence Platform to unify its data and democratize analytics across the business.

By centralizing raw and curated data in Delta Lake and governing it with Unity Catalog, PicPay established a shared foundation for product, engineering and analytics teams. What had previously been a fragmented ecosystem of isolated BI pipelines and inconsistent metrics became a single source of truth. Data engineers created reusable Gold tables, version-controlled with Git and orchestrated through Spark Declarative Pipelines, while business analysts queried these assets using Databricks SQL.

One of the most transformative outcomes was self-service. With the Data Intelligence Platform, PicPay empowered business teams to explore data on their own, leveraging fast dashboards, materialized views and clear ownership of metadata. According to PicPay's team, the shift allowed product teams to answer questions that once required weeks of engineering support — like user journey drop-offs, experiment analysis and feature adoption rates — directly within minutes.

PicPay

"Now our users were running queries much faster. The performance is up six times better compared to the classic setup, and we saw a reduction of up to 44% in the average daily costs. That not only made our users happy but gave us real financial efficiency."

— Gustavo Baião, Data Engineering Leader, PicPay

YouTube

The platform also paved the way for responsible AI adoption. By layering ML models directly on top of trusted BI data, PicPay unlocked advanced use cases like churn prediction, real-time recommendations and fraud detection — all with auditable lineage and reproducibility.

This end-to-end modernization allowed PicPay to:

- Cut time to insight from days to minutes across product and GTM teams
- Reuse over 70% of pipeline logic via centralized models and Unity Catalog tables
- Scale analytics across thousands of internal stakeholders with confidence
- Lay a secure, governed foundation for AI-powered growth

SECTION 2.19 Best Practice: Real-Time Operational Insights at Walmart

As one of the world's largest retailers, Walmart faces enormous complexity in managing its global supply chain, in-store operations and e-commerce platform. To remain competitive and operationally agile, Walmart turned to the Databricks lakehouse architecture to power real-time data products that deliver trusted insights at scale.

Walmart's data team adopted a domain-oriented architecture, organizing development efforts into reusable, secure and governed data products. These products serve internal customers across merchandising, finance, store operations and supply chain. By building shared assets on top of Delta Lake and managing them with Unity Catalog, Walmart enables thousands of stakeholders to access consistent, high-quality data while maintaining granular access control.

One of the key innovations was developing a centralized metrics platform that serves over 50 different dashboarding and analytical use cases. Powered by Databricks SQL and built for reusability, this platform allows teams to build once and consume everywhere — cutting down redundant data work and aligning KPIs across departments. Near real-time updates enable operational decisions to be made faster, from shelf restocking to regional logistics adjustments.

Walmart's success with Databricks highlights how organizations can move beyond siloed reporting into a unified, governed and real-time analytics ecosystem that enables decisions at the speed of retail.

Source: YouTube: Walmart and Databricks — Scaling Real-Time Decisioning

SECTION 2.20 Best Practice: Financial Forecasting and Compliance at SMBC

Sumitomo Mitsui Banking Corporation (SMBC), one of Japan's largest financial institutions, relies on Databricks to power its critical financial forecasting and compliance workloads. As a highly regulated global bank, SMBC must maintain precise, auditable reports while also adapting quickly to market changes. To achieve this, the organization modernized its data architecture using the Databricks Data Intelligence Platform, bringing together disparate data systems into a unified environment.

By consolidating data pipelines and analytical workflows on Databricks, SMBC was able to streamline the process of generating financial risk and liquidity reports, while also increasing the granularity and flexibility of its analyses. The bank's teams built scalable data products on Delta Lake and used Databricks SQL to enable governed access for analysts and regulators across jurisdictions.

Critically, the adoption of Unity Catalog helped ensure compliance with strict internal controls and external regulatory requirements. The platform enabled centralized governance and auditing of sensitive datasets while facilitating collaboration across business units. Databricks' native support for Spark, Python and SQL empowered a wide range of users — from engineers to financial analysts — to develop and operationalize models faster.

> "Our goal here is … a well-governed platform. Govern means we have all the controls, we can identify lineage. … It should be something that everyone knows — they can push a button and get it. Everyone just wants things to work. … The engineering behind it is critical."
>
> — Gordon Wilson, CIO, SMBC Americas

YouTube

With Databricks, SMBC has established a flexible, AI-ready analytics foundation that supports both compliance and innovation — critical for staying competitive in the evolving Financial Services landscape.

SECTION 2.21 JLL — Upskilling Program for Data Warehouse Migration

JLL, a global leader in commercial real estate services, needed to unify data across 80+ countries to deliver real-time business intelligence and AI-driven insights. Facing increasing data volumes, performance bottlenecks and legacy BI limitations, JLL pivoted from Snowflake to a strategic partnership with Databricks. They centralized analytics on Databricks SQL, modernizing reporting and analytics pipelines for scale, speed and governance.

To ensure successful adoption across its 120+ global analysts, JLL launched the Databricks Odyssey program — a gamified, role-based training framework designed to onboard and certify users through hands-on learning. This change management initiative paired technical enablement with community engagement, reducing migration friction and empowering analysts to build and launch dashboards faster. The program included learning tracks, real-world challenges and public recognition mechanisms that helped foster motivation, consistency and accountability across a globally distributed team.

By consolidating their analytics environment onto a unified lakehouse architecture, JLL eliminated data silos, improved governance with Unity Catalog and enabled governed self-service analytics. Their analysts and data engineers now collaborate in Databricks Notebooks and SQL dashboards to model global property data, market trends and client KPIs. The transformation also paved the way for broader adoption of AI and ML, with future plans to build AI copilots, automate forecasting, and support interactive property intelligence experiences for clients and internal teams alike.

> "We formed a strategic partnership with Databricks. … 94% of our immediate team has completed the Databricks fundamentals training — establishing the essential foundation for our data migration efforts."
>
> — Kris Curtis, Global BI Technology Director, JLL Technologies

YouTube

SECTION 2.22 HP Delivers Enterprise-Scale BI With Databricks SQL

HP, a global technology leader operating in over 170 countries, faced growing demands for timely, consistent insights across its diverse product lines, supply chain operations and sales channels. The company had long operated in a fragmented analytics environment — multiple regional data marts, siloed ETL pipelines and dashboard tooling spread across departments. This disjointed approach made it difficult to establish consistent KPI definitions, ensure governance or scale analytics in a cost-effective way.

To modernize its data architecture, HP adopted the Databricks Data Intelligence Platform, with Databricks SQL serving as the core for enterprise analytics. By moving from isolated data silos to a centralized, governed lakehouse, HP unified operational, sales and financial data into a common Delta Lake format. This architectural shift simplified data access, improved reliability and enabled near real-time analytics for faster decision-making.

With Databricks SQL warehouses, HP teams delivered fast, cost-effective dashboards that scaled to global usage. Teams across functions — from supply chain to product to customer support — built self-service dashboards powered by certified datasets in Unity Catalog. Materialized views and serverless SQL endpoints allowed for low-latency query performance without manual tuning, and automated usage optimization features helped control warehouse costs even at scale.

HP invested in a semantic layer strategy using Unity Catalog to curate datasets with business-friendly names, consistent logic and strict access controls. These curated tables fed directly into Power BI and internal reporting tools, giving thousands of employees governed access to shared metrics. Analysts no longer had to redefine KPIs or duplicate joins for every report — which led to a significant reduction in reporting backlog and rework.

Beyond static dashboards, HP leveraged Databricks to support streaming telemetry and predictive analytics. By combining sensor data from global manufacturing systems with real-time sales feeds and inventory tracking, the company created dashboards that could proactively surface bottlenecks, flag product availability issues or identify emerging customer satisfaction risks. Spark Declarative Pipelines and automated workflow orchestration ensured this insight was always fresh and reliable.

The shift had measurable outcomes. Time to insight dropped from hours to seconds. Business teams gained confidence in a single source of truth. Governance teams simplified audit and access compliance through Unity Catalog lineage tracking. HP also reduced infrastructure complexity by consolidating redundant data systems and dashboard engines into a single lakehouse architecture.

> "Databricks helped us to unify all of our data across HP into one place. We use Databricks SQL for enterprise dashboards — it gives us faster time to insight and the governance we need."
>
> — Ravi Yadav, Head of Data and Analytics Engineering, HP

YouTube

Today, Databricks SQL powers decision-making across HP's finance, operations, supply chain and executive teams. As HP expands into AI-driven insights and ML model development, it continues to rely on the same lakehouse architecture to ensure scalability, consistency and security across all analytics use cases.

PART 3

Next Steps

SECTION 3.1 Transforming Data Warehousing and Business Intelligence With Databricks

Today's business landscape demands more than static dashboards and siloed reports. Organizations need insights that move at the speed of their operations, power AI-native applications and scale securely across the entire enterprise. The Databricks Data Intelligence Platform is purpose-built to meet this moment — combining the best of data warehousing performance with the openness and intelligence of a modern data platform.

Throughout this guide, we've explored how Databricks is helping data teams, business leaders and IT organizations rethink what's possible for data warehousing and BI — from operational analytics and marketing intelligence to embedded dashboards and predictive models. Whether you're migrating from legacy systems, modernizing SAP workloads or building BI into your product, Databricks provides the blueprint for intelligent decision-making at scale.

With Unity Catalog as the foundation for governance, Delta Lake for real-time data reliability and serverless Databricks SQL for cost-effective performance, your teams gain the agility and trust they need to turn data into action. Add AI assistants like AI/BI Genie and an open ecosystem that spans clouds and industries, and the result is a single platform for your entire data estate — today and tomorrow.

TRY DATABRICKS FREE EDITION

Ready to get hands-on with the Databricks Platform? **Databricks Free Edition** is the fastest way to start exploring Databricks SQL, Delta Lake, notebooks and AI/BI capabilities — at no cost.

Designed for data professionals, analysts and engineers, Free Edition provides access to a shared, collaborative environment where you can:

- Run **Databricks SQL queries and dashboards**
- Explore **notebooks and Delta Lake tables**
- Test **AI functions** and **Vector Search**
- Use the **Unity Catalog** for governance basics
- Experiment with **AI/BI Genie** and **natural language querying**

Whether you're evaluating Databricks for BI, building proofs of concept, or learning lakehouse best practices, Free Edition is an easy entry point — no credit card required.

Sign up now

databricks

TAKE THE NEXT STEP

- **Start a free trial:** Explore Databricks SQL and lakehouse capabilities in a no-risk environment
  databricks.com/try-databricks
- **Watch product tours for end-to-end capabilities:**
  Databricks SQL: https://www.databricks.com/resources/demos/tours/governance/dbsql
  AI/BI: https://www.databricks.com/resources/demos/tours/bi/introducing-databricks-aibi-genie-enduser
- **Watch demos to see how the products work:**
  Databricks SQL: https://www.databricks.com/resources/demos/library?q=SQL
  AI/BI: https://www.databricks.com/resources/demos/library?q=ai%2Fbi
- **Explore migration services:** Accelerate your move from Oracle, SQL Server, Redshift or Snowflake with lakehouse migration tools
  databricks.com/solutions/migration/lakebridge
- **Getting started with SQL and BI:**
  databricks.com/resources/learn/training/get-started-sql-analytics-and-bi
- **Talk to a lakehouse architect:** Get expert guidance on designing your data warehousing strategy
  databricks.com/contact
- **Take free training on getting started:**
  https://www.databricks.com/resources/learn/training/get-started-sql-analytics-and-bi
- **Stay connected:** Subscribe to the Databricks blog and YouTube channels for best practices, customer stories and product updates
  databricks.com/blog | youtube.com/databricks

SECTION 3.2 Dashboards and Notebooks

HOW TO ACCESS DASHBOARDS AND NOTEBOOKS

These demos are accessible via the Databricks Demo Center or by installing the dbdemos package in a workspace. Each includes step-by-step instructions, example dashboards and guided notebooks.

COMPLETE LIST

- **Lakehouse C360 Retail Churn Demo**
  A full-stack demo (DBDemos) showcasing ingestion, governance, BI dashboards and ML for churn prediction. Link

  Install it via:

PYTHON

```
CopyEdit
pip install dbdemos
dbdemos.install('lakehouse-retail-c360')
```

- **AI/BI: Sales Pipeline Overview With Dashboards and Genie**
  An interactive tutorial to explore the Sales Pipeline analytics workflow. Available in the Demo Center under AI/BI tutorials.
- **Data Quality Monitoring Demo**
  Explore Lakehouse Monitoring notebooks to profile and detect data anomalies for BI workloads. Link
- **Notebooks-to-Dashboards Integration Tutorial**
  Walks through executing SQL in notebooks, seamlessly converting results into interactive dashboards, and leveraging AI/BI Dashboards for sharing. Link
- **Structured Streaming Tutorial Notebook**
  Run your first Structured Streaming workload using Auto Loader and Delta Lake. Ideal for Bronze table streaming ingestion. Link
- **Spark Declarative Pipelines Example Notebooks**
  Real-time analytics, CDC and streaming workflows using Spark Declarative Pipelines and SQL/Python. GitHub Repo
- **CDC + Lakeflow Spark Declarative Pipelines Tutorial**
  Tutorial on building CDC-driven pipelines using Auto Loader and Lakeflow. Demonstrates Bronze-Silver-Gold pattern. Link
- **Built-In AI/SQL Functions Demo Notebook**
  A community-shared Jupyter notebook demonstrating Databricks SQL's built-in AI functions like summarization, translation and sentiment analysis directly within SQL queries. Link
- **Vector Search Support Demo**
  Part of the "genai-cookbook" project, this Databricks Notebook showcases how to build and query vector indexes using Databricks' Vector Search client APIs. It includes setup, indexing and query examples that can seamlessly integrate into SQL workflows. Link

PYTHON

```
CopyEdit
from databricks.vector_search.client import VectorSearchClient
```

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.

databricks