

# Databricks SQL: Why the Best Serverless Data Warehouse is a Lakehouse?

---

**Miranda Luna**, Senior Staff Product Manager, Databricks

**Cyrielle Simeone**, Principal Product Marketing Manager, Databricks



# Agenda

**Part 1 – Databricks SQL: Why the Best Serverless Data Warehouse is a Lakehouse**

**Part 2 – Databricks SQL: What's new with live demos**

@3:30PM same room



# Product safe harbor statement

- This information is provided to outline Databricks' general product direction and is **for informational purposes only**. Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks discretion and may not be delivered as planned or at all.

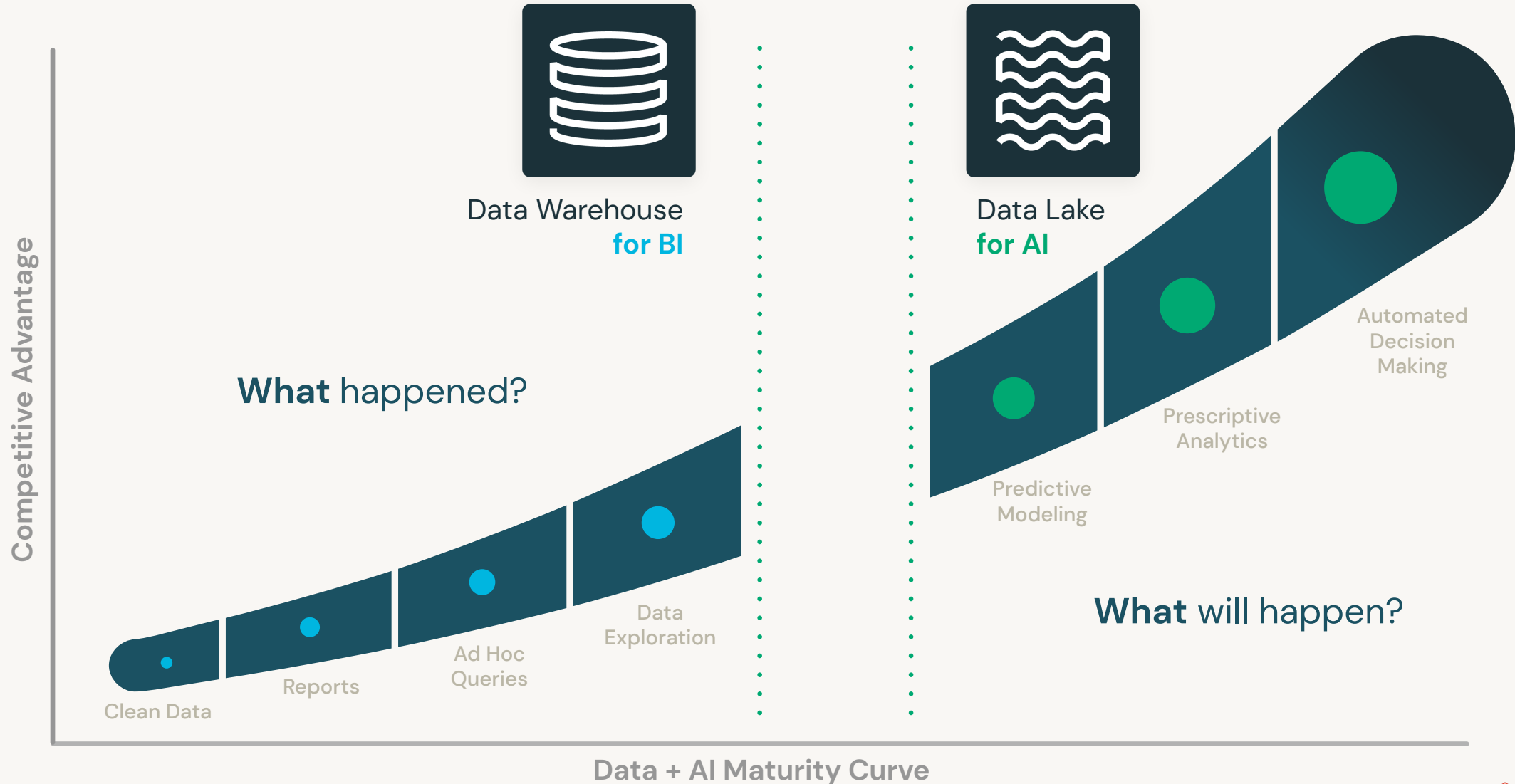


Every company  
wants to become a  
**Data+AI** company





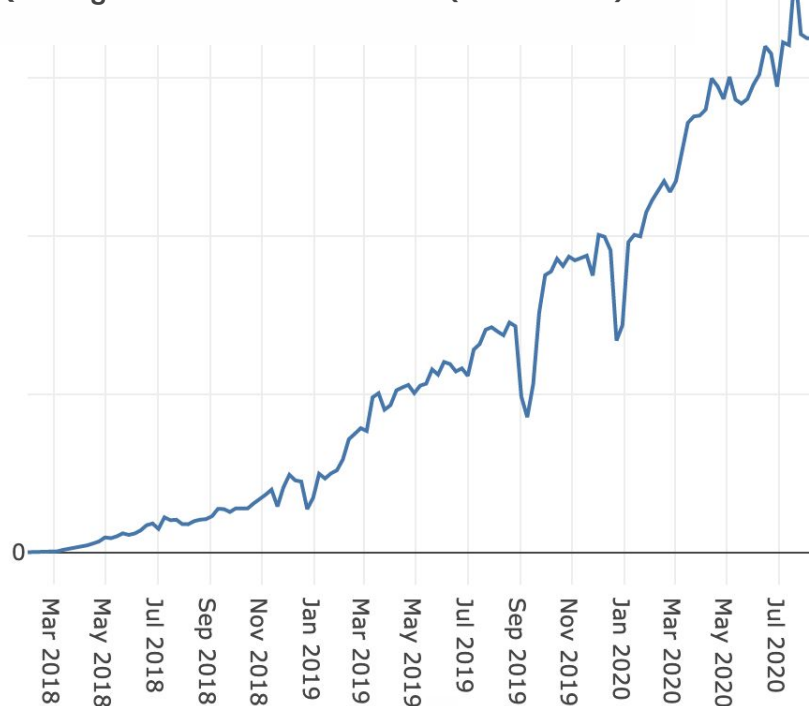
# Two incompatible architectures get in the way



# Customers started using our Spark clusters for SQL

## The origins of the lakehouse

SQL Usage Growth on Databricks (2018–2020)



## What Is a Lakehouse?



by Ben Lorica, Michael Armbrust, Reynold Xin, Matei Zaharia and Ali Ghodsi

January 30, 2020 in [Engineering Blog](#)

Share this post



Over the past few years at Databricks, we've seen a new data management architecture that emerged independently across many customers and use cases: [the lakehouse](#). In this post we describe this new architecture and its advantages over previous approaches.

Data warehouses have a [long history](#) in decision support and business intelligence applications. Since its inception in the late 1980s, data warehouse technology continued to evolve and MPP architectures led to systems that were able to handle larger data sizes. But while warehouses were great for structured data, a lot of modern enterprises have to deal with unstructured data, semi-structured data, and data with high variety, velocity, and volume. Data warehouses are not suited for many of these use cases, and they are certainly not the most cost efficient.



# Data Warehousing on the Lakehouse

Powered by Databricks SQL

Databricks SQL (DB SQL) is a serverless data warehouse on the Databricks Lakehouse Platform that lets you run all your SQL and BI applications at scale with up to 12x better price/performance, a unified governance model, open formats and APIs, and your tools of choice – no lock-in.



**Best price/performance**



**Built-in governance**



**Rich ecosystem**



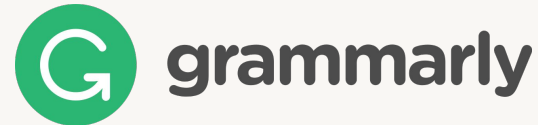
**Break down silos**



# Databricks SQL Momentum



ESTÉE LAUDER



AT&T



COMPASS



CRED



DEVSISTERS

CONDÉ NAST



ABN-AMRO



Abnormal



Radio-Canada

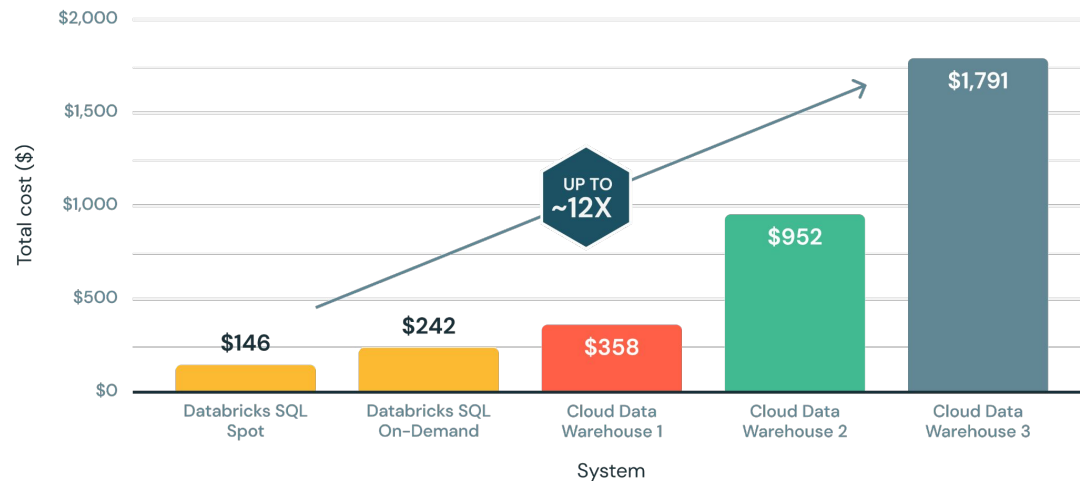


# Built from the ground up for best performance

Lightning fast analytics for all queries

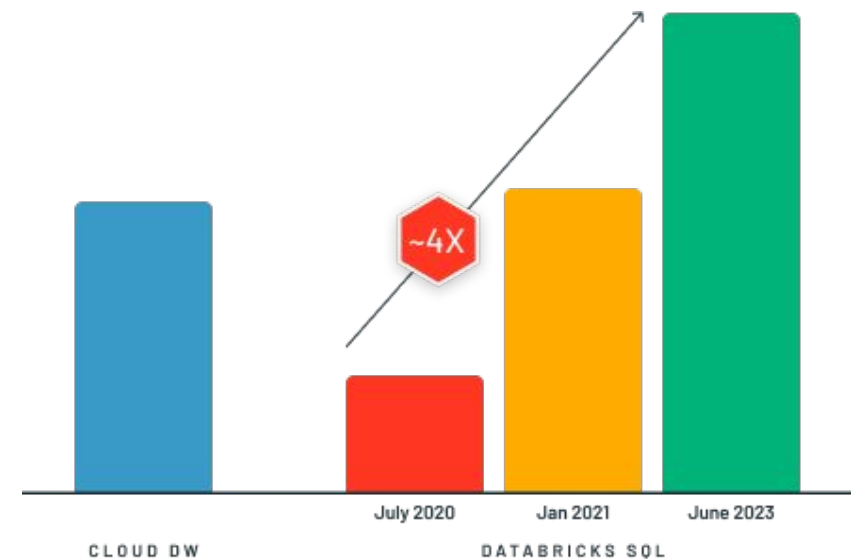
## 100TB TPC-DS Price/Performance

Lower is better



## 10GB TPC-DS @ 32 Concurrent Streams (Queries/Hr)

Higher is better

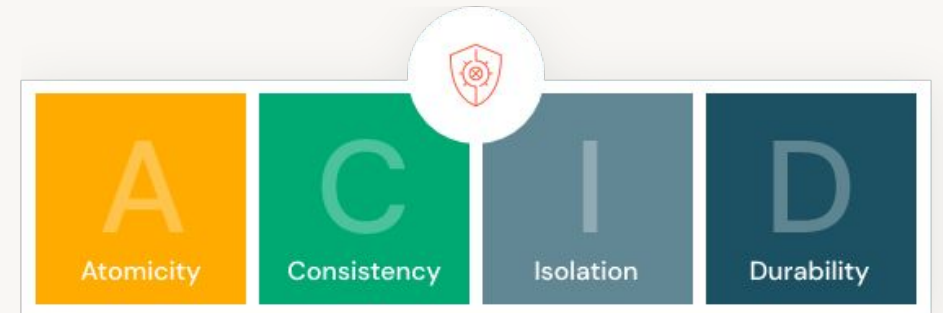


<https://dbricks.co/benchmark>

# One source of truth for all your data

Open format Delta Lake as the foundation

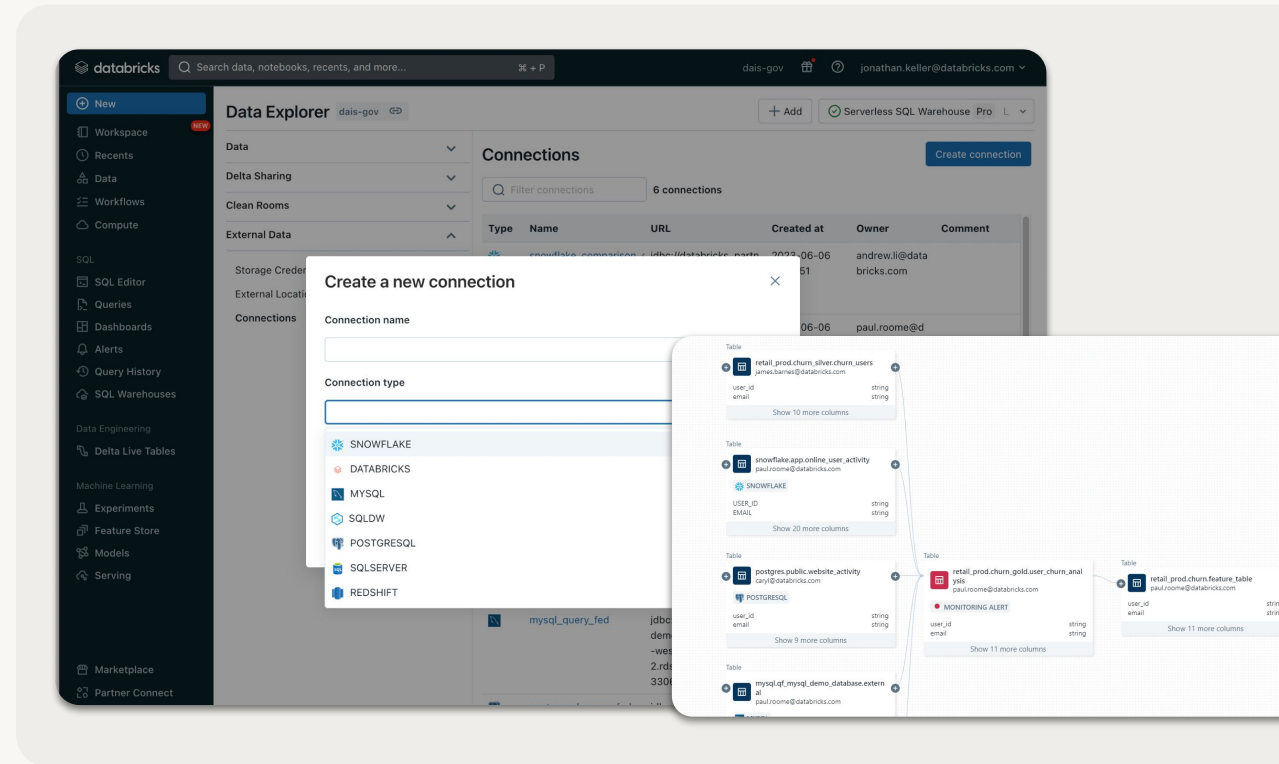
Delta Lake adds **quality, reliability, and performance** to your existing data lakes, and provides **one common data management framework** for data, analytics and AI use cases.



# Seamless integration with Unity Catalog

Unified governance for data and AI

**Securely discover, access and collaborate** on trusted data and AI assets, leveraging AI to boost productivity and unlock the full potential of the lakehouse environment.



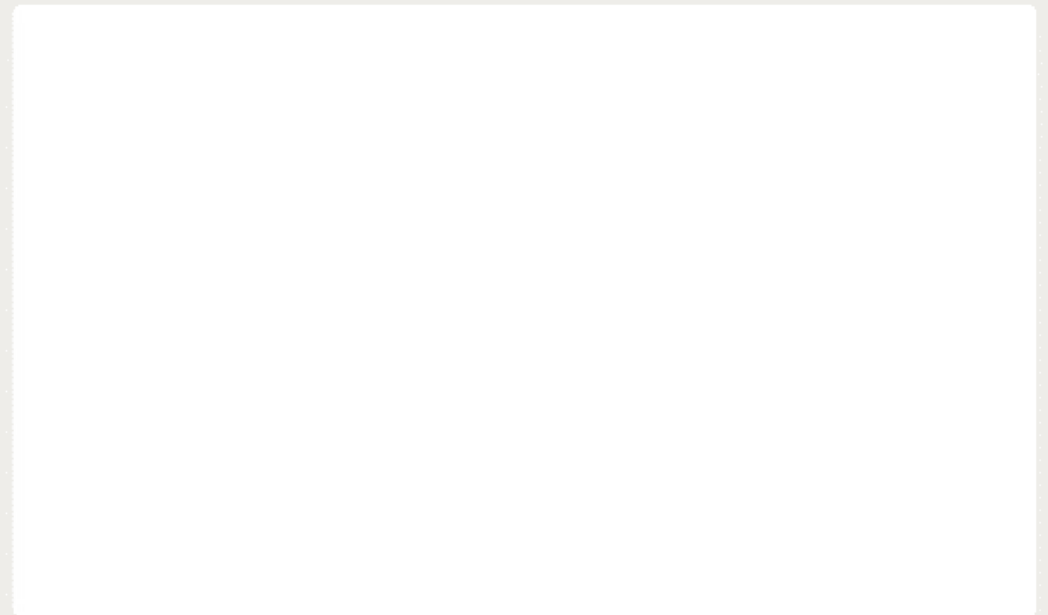
Wednesday, June 28 @11:30 AM | What's New in Unity Catalog -- With Live Demos

Wednesday, June 28 @4:30 PM | Lakehouse Federation: Access and Governance of External Data Sources from Unity Catalog

# Best of the Lake and the Warehouse

Go from BI to AI effortlessly to uncover new insights

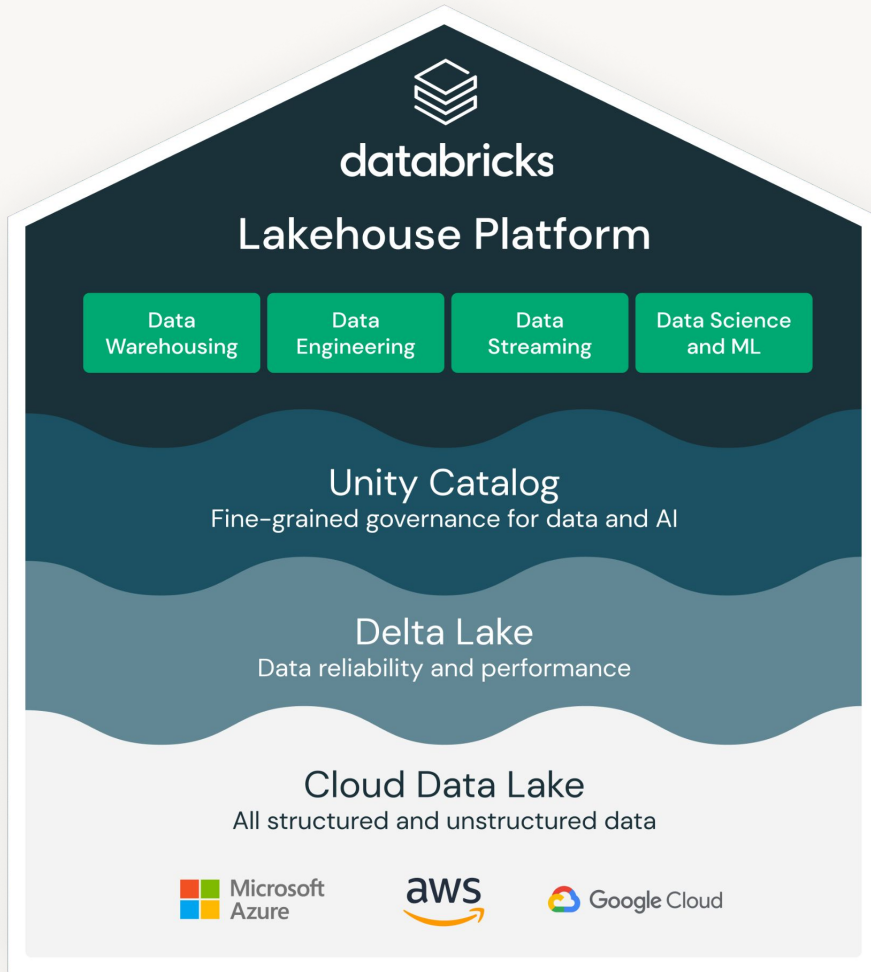
Build and train state of the art LLMs & machine learning models on your most complete data, remove silos, and **democratize AI** across your organization.





# The best data warehouse is a lakehouse

Powered by Databricks SQL



- 1 Data, analytics, and AI in one place
- 2 World-class performance with data lake economics
- 3 One source of truth for all your data

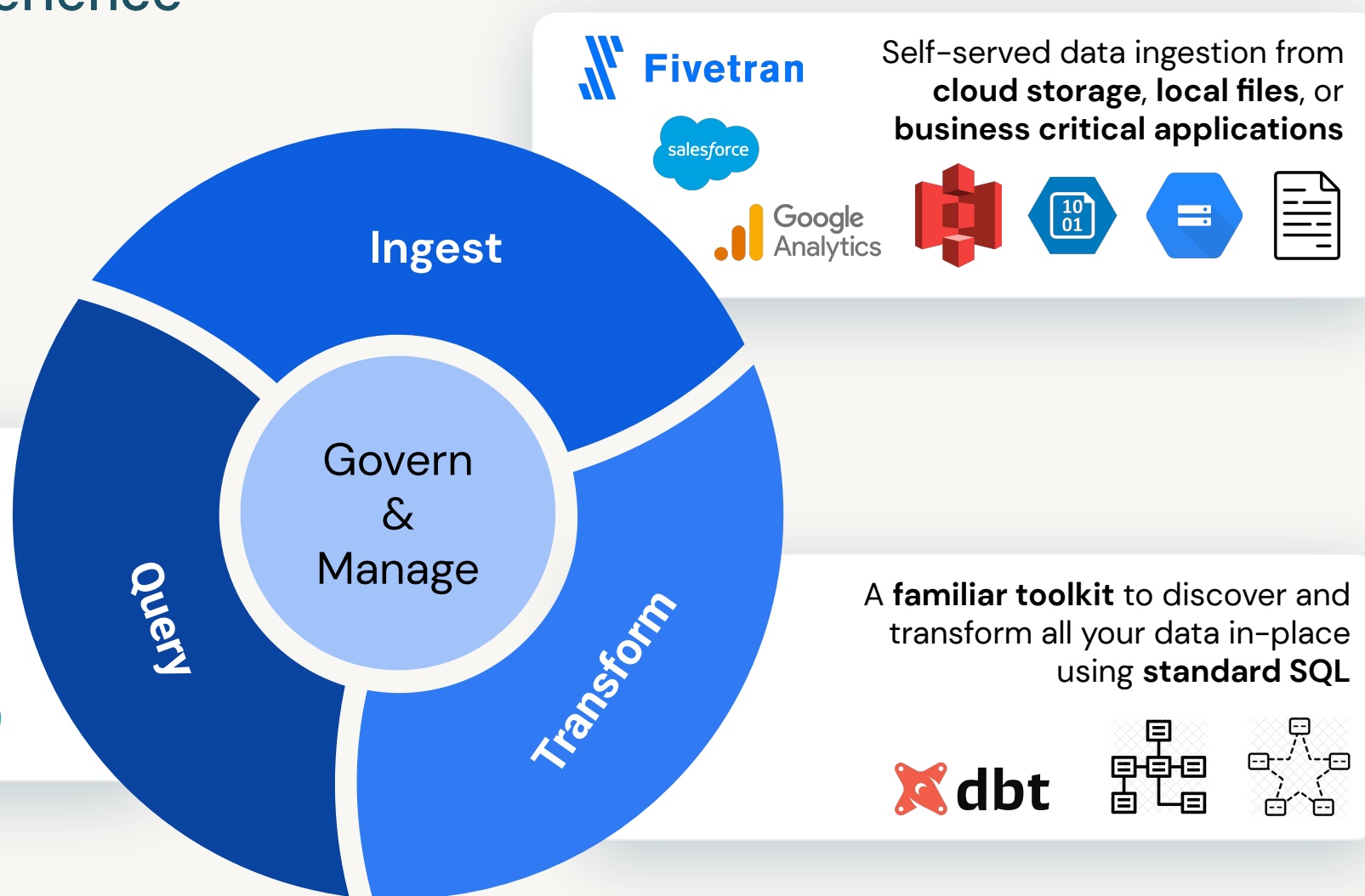


# Product Overview



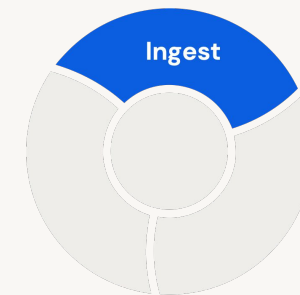
# Ingest, transform, and query with any tool

A first-class SQL experience



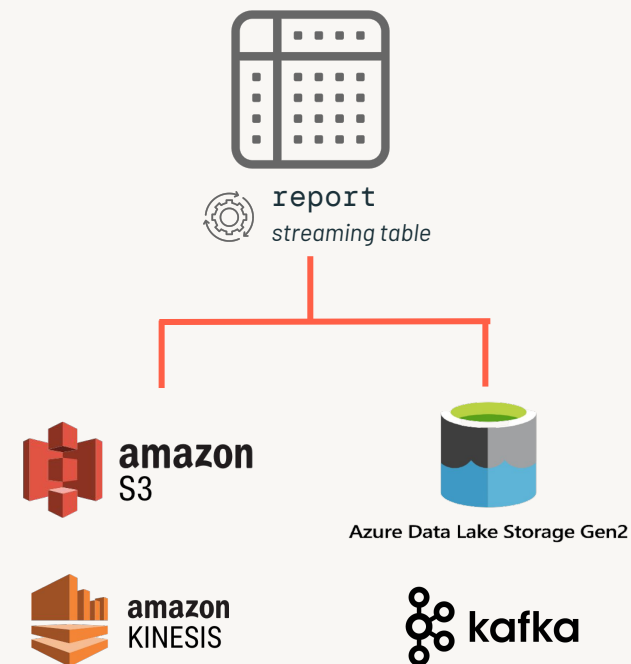
# Data Ingest: Streaming Tables

Efficiently and continuously land data in the bronze layer



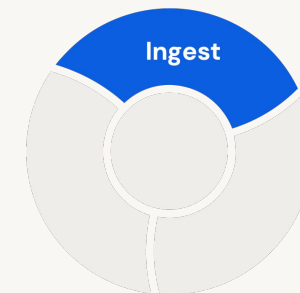
Enable the **continuous, scalable ingestion** from any data source including cloud storage, message buses (EventHub, Kafka, Kinesis) and more

```
CREATE STREAMING TABLE report
AS SELECT SUM(profit)
FROM cloud_files(prod.sales
```

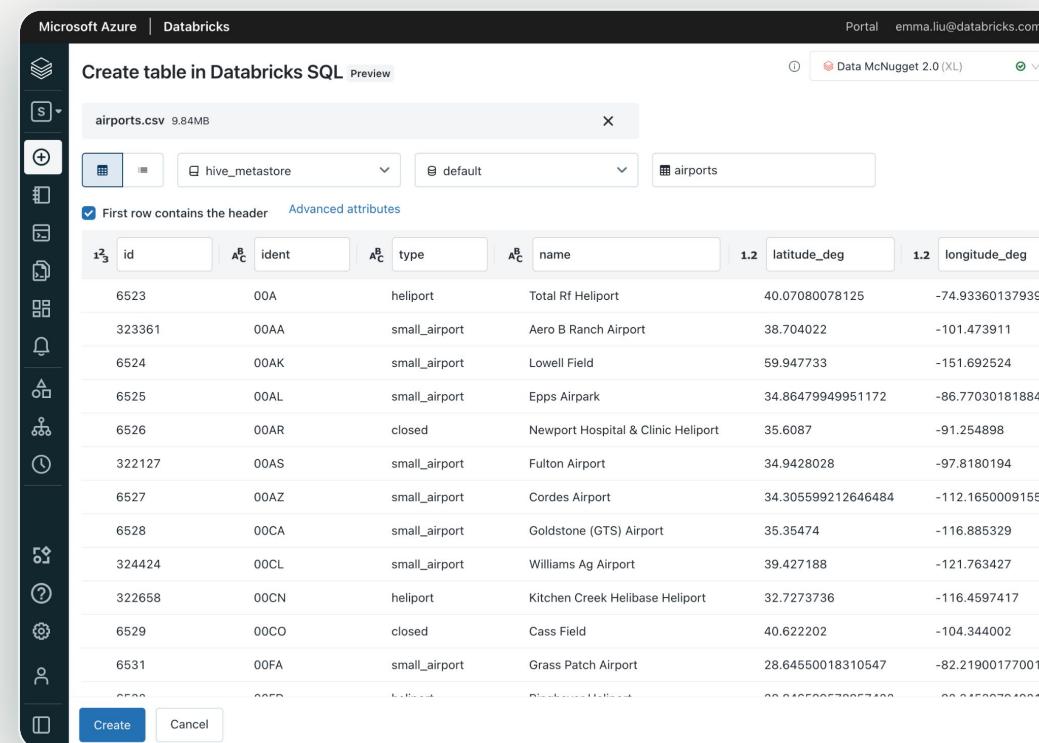


# Data Ingest: Partners

Work with your data no matter where it lives

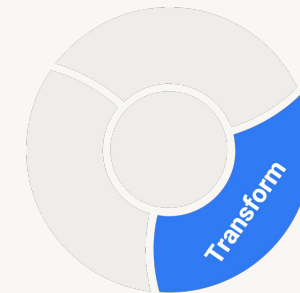


Easily ingest business critical data from Salesforce, Google analytics, Marketo etc... using **Fivetran** or **import files** from your data stores.



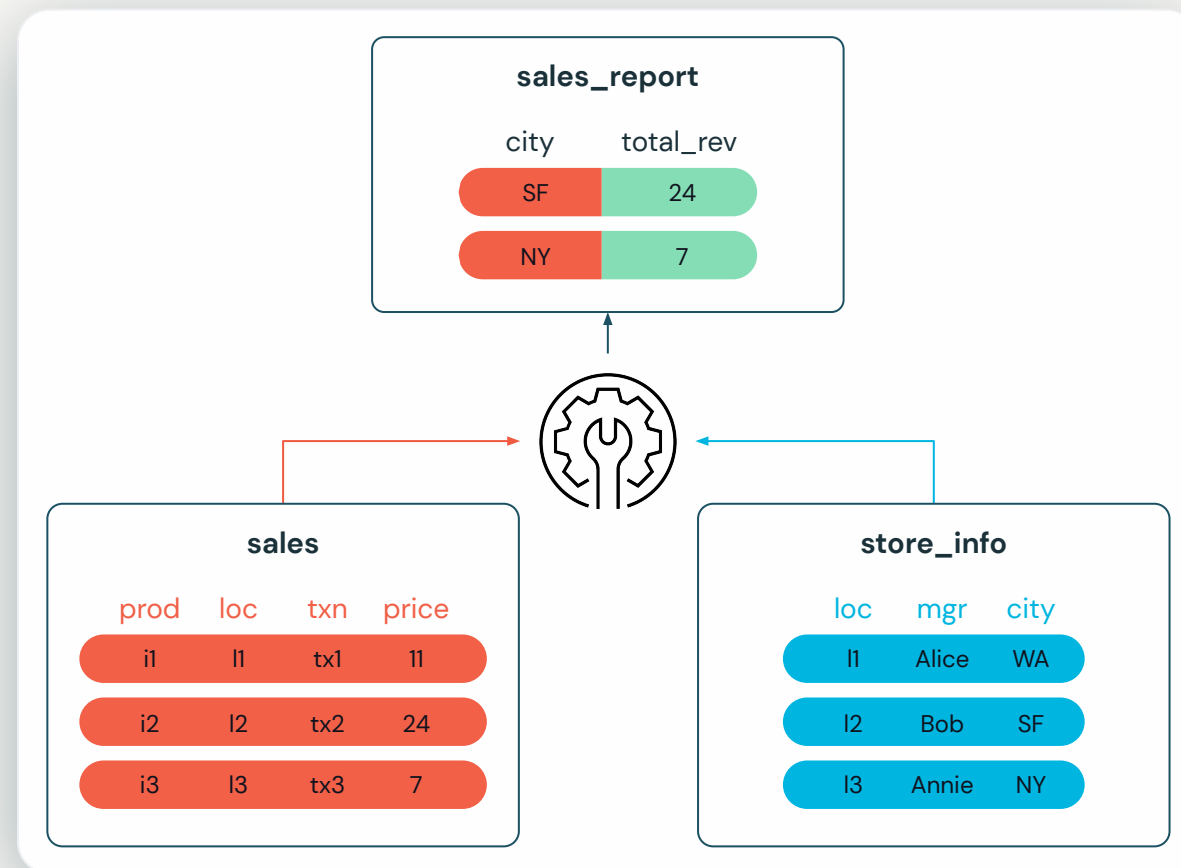
# Materialized Views

Speed up queries with pre-computed results



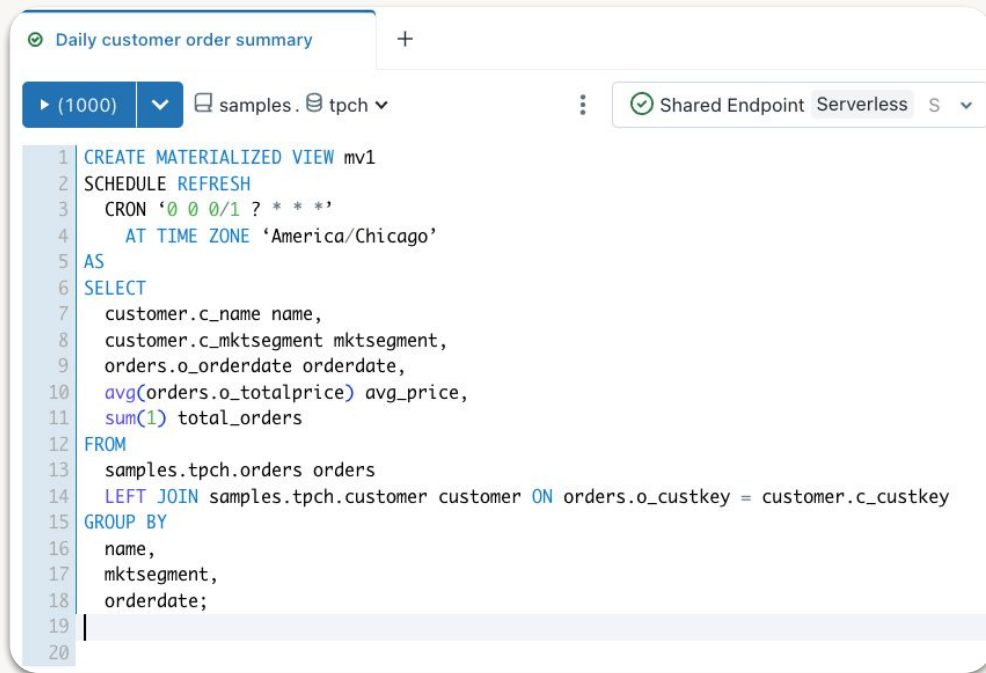
Accelerate end-user queries and reduce infrastructure costs with efficient, incremental computation

- Accelerate **BI dashboards** and **ETL queries**
- **Streaming**: build MVs on top of live tables
- **Easy ELT**: Simplify reporting by cleaning, enriching, denormalizing the base tables



# Materialized Views and Streaming Tables

The best data warehouse gets the best of data engineering



The screenshot shows a Databricks SQL interface. At the top, there's a tab labeled "Daily customer order summary". Below the tab, there's a dropdown menu showing "(1000)" and a button labeled "samples. tpch". To the right, there's a status bar showing "Shared Endpoint" and "Serverless". The main area displays a SQL query for creating a materialized view named "mv1".

```
1 CREATE MATERIALIZED VIEW mv1
2 SCHEDULE REFRESH
3   CRON '0 0 0/1 ? * *'
4   AT TIME ZONE 'America/Chicago'
5 AS
6 SELECT
7   customer.c_name name,
8   customer.c_mktsegment mktsegment,
9   orders.o_orderdate orderdate,
10  avg(orders.o_totalprice) avg_price,
11  sum(1) total_orders
12 FROM
13   samples.tpch.orders orders
14  LEFT JOIN samples.tpch.customer customer ON orders.o_custkey = customer.c_custkey
15 GROUP BY
16   name,
17   mktsegment,
18   orderdate;
```

**Enable your analysts.** SQL and data analysts can easily ingest, clean, and enrich data to quickly meet the needs of your business.

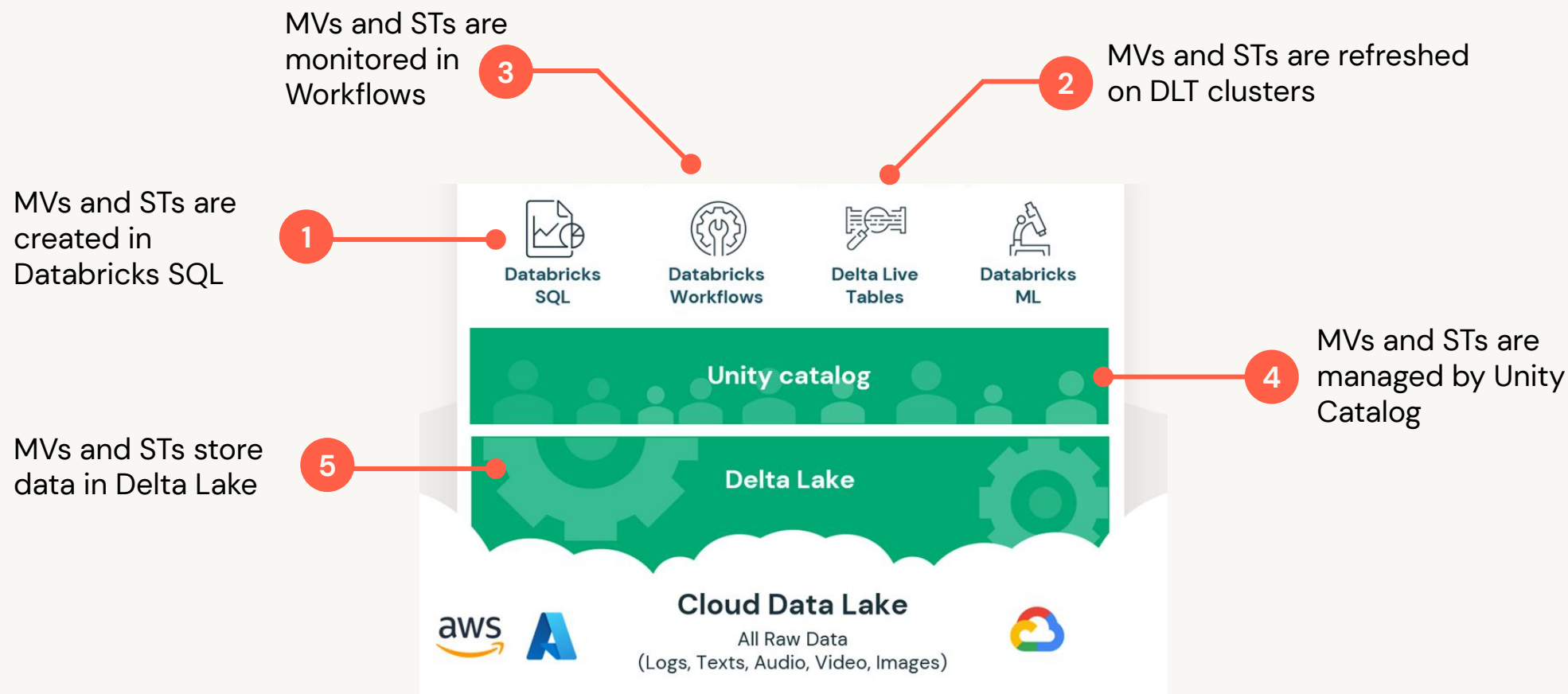
**Speed up BI dashboards.** Create MV's to accelerate SQL analytics and BI reports by pre-computing results ahead of time.

**Move to real-time analytics.** Combine MV's with streaming tables to create fully incremental data pipelines for real-time use cases.



# DBSQL MVs & STs on Databricks Lakehouse

How do MVs and STs fit in the lakehouse architecture?





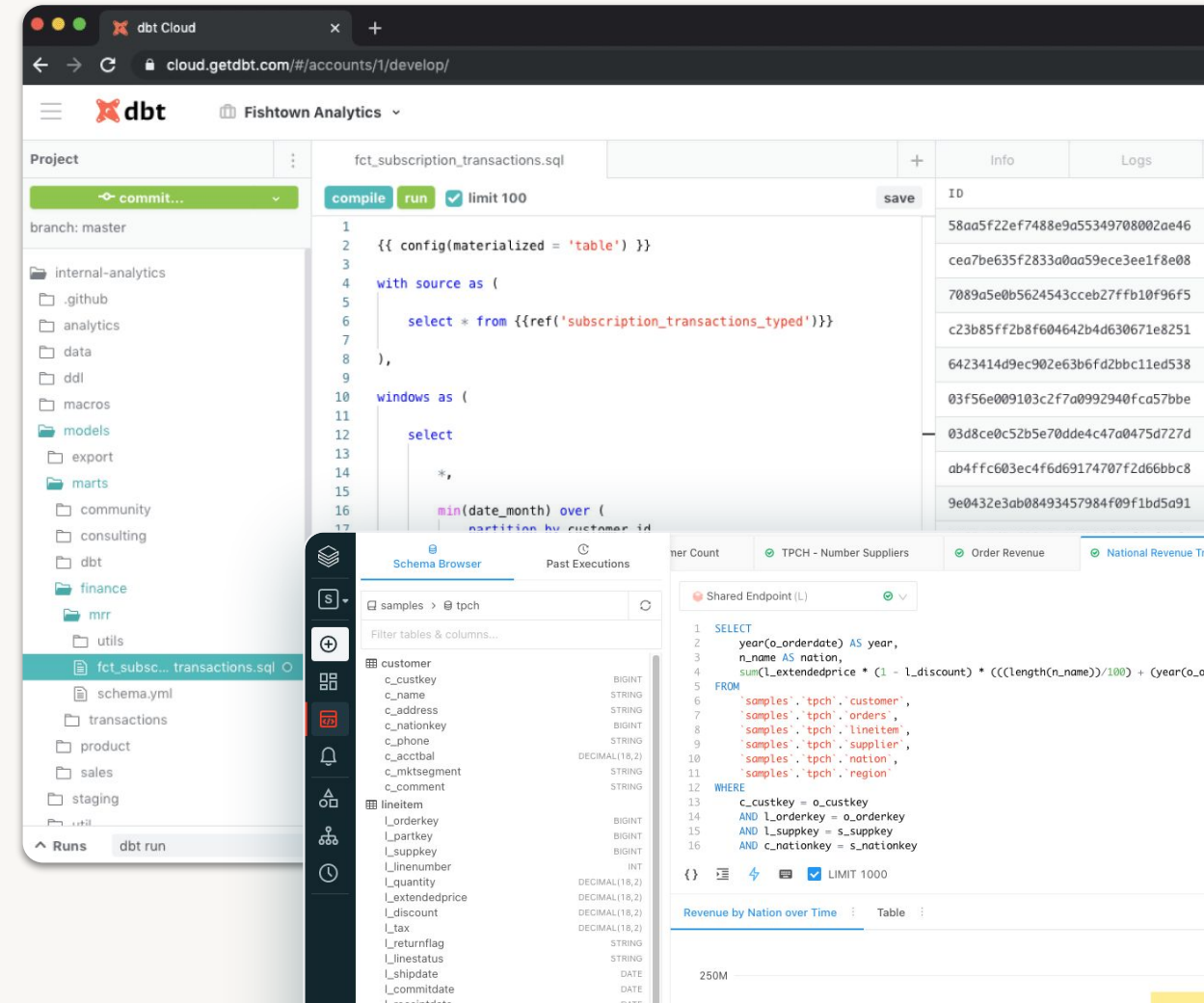
# Data Transformation

Analytics engineering on the Lakehouse made simple



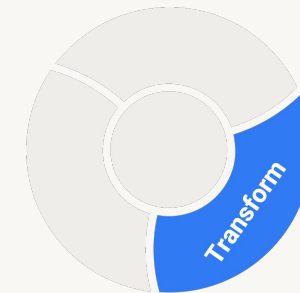
Databricks and dbt Labs simplify analytics engineering on the lakehouse.

Users can ingest and transform streaming data in their dbt pipelines using **Streaming Tables and Materialized Views**.



# Workflows + Databricks SQL

Orchestrate your SQL queries, dashboards, alerts, and more



Schedule and automate your Databricks SQL production workloads

- Easily orchestrate sophisticated workflows with multiple dependencies
- Enhanced monitoring and observability with proven reliability in production
- Up next: Native integration with dashboards, SQL queries and alerts

The screenshot displays the Databricks Workflows interface. The main view shows a workflow named 'Refresh Sales Dashboards' with a task 'Refresh\_SGTM\_Dashboard' (Data Driven SGTM, Serverless Starter Endpoint) that triggers two other tasks: 'Refresh\_Finance\_Dashboard' (Sales Finance Metrics V0, Shared SQL Endpoint - Cutting Edge) and 'Finance\_Alert\_1' (Check that finance and forecasts are..., Shared SQL Endpoint - Stable). The 'Task details' panel on the right shows job ID 1112715633840115, creator Richard Tomlinson, and no lineage information. The 'Schedules & Triggers' section is currently set to 'None'. Below the workflow view, a 'Workflows' summary card shows 'Total runs: 412', 'Active runs: 0', 'Completed runs: 412', 'Successful runs: 180', 'Skipped runs: 0', and 'Failed runs: 232'. A bar chart shows the 'Finished runs count' over time, with a peak in the morning. A table below the chart lists recent runs with columns for Start time, Job, Run as, Launched, Duration, and Status.

Start time	Job	Run as	Launched	Duration	Status
Mar 9 2023, 10:20 AM CET	ingest_sales	jan.vandervegt@databricks.com	By scheduler	34s	Failed
Mar 9 2023, 10:19 AM CET	sales_forecast	jan.vandervegt@databricks.com	By scheduler	48s	Succeeded
Mar 9 2023, 10:15 AM CET	ingest_sales	jan.vandervegt@databricks.com	By scheduler	34s	Failed
Mar 9 2023, 10:10 AM CET	ingest_sales	jan.vandervegt@databricks.com	By scheduler	32s	Failed
Mar 9 2023, 10:09 AM CET	sales_forecast	jan.vandervegt@databricks.com	By scheduler	33s	Succeeded
Mar 9 2023, 10:05 AM CET	ingest_sales	jan.vandervegt@databricks.com	By scheduler	40s	Failed



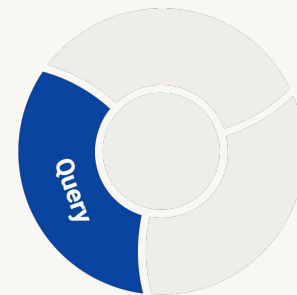
# Data Consumption

## Query from any tool

Connect existing BI tools and dashboards or brand new ones to the freshest data using OAuth or PAT tokens

Leverage your favorite SQL workbenches or IDE to find new insights

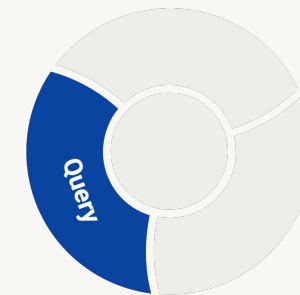
Build custom data apps powered by the lakehouse with tools and languages you already know



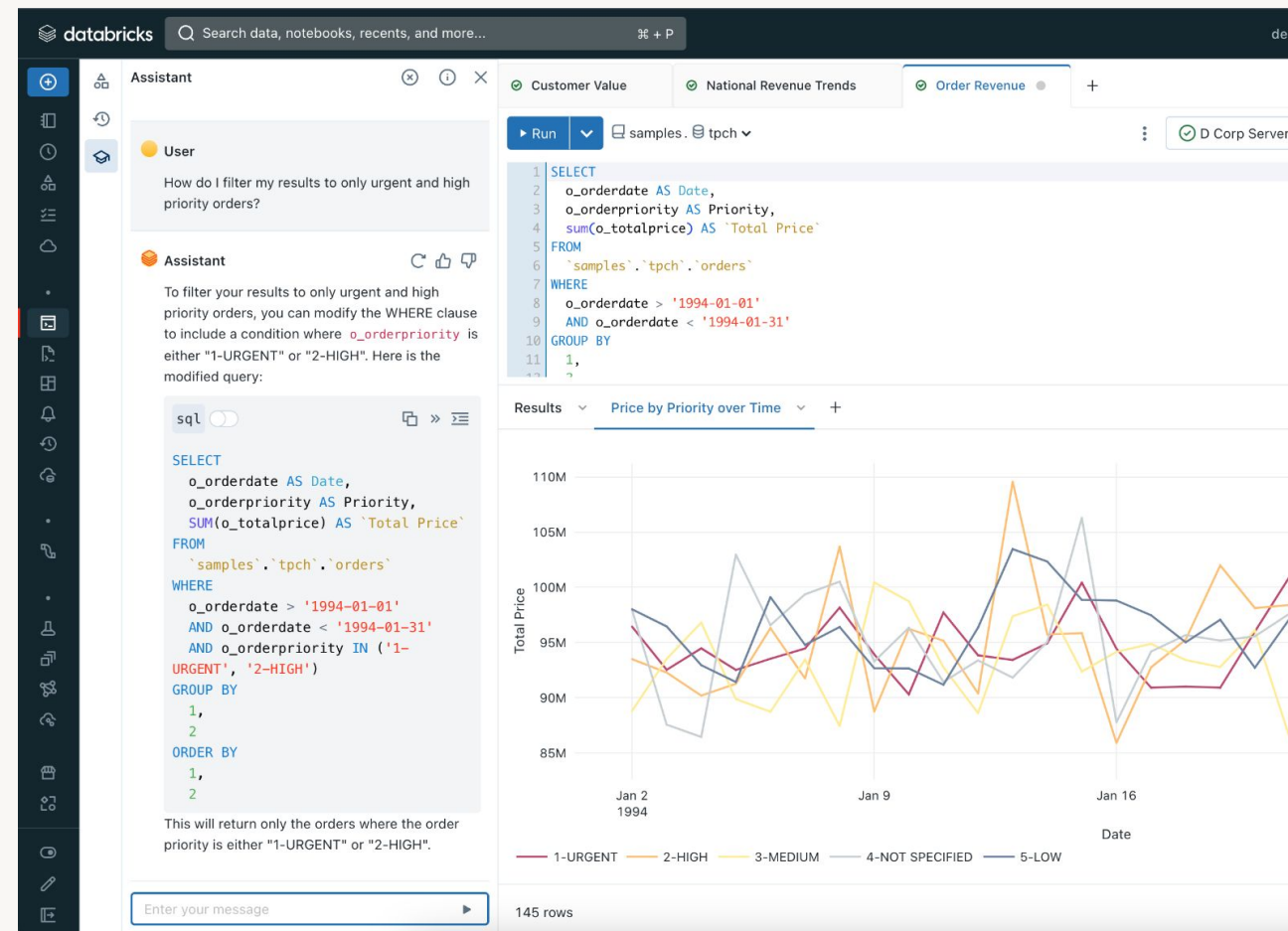
And more...

# Integrated Analytical Tooling

Collaboratively query, explore, and transform data in-place

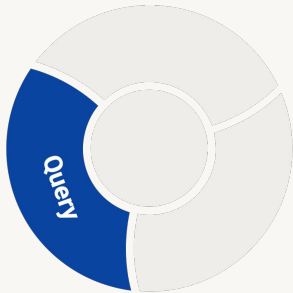


- Discover data, explore database schema, and query data using **ANSI SQL**
- Save, share, and reuse queries across teams to get to results faster
- Up next: Integrated SQL authoring assistant
- Build interactive visualizations and **dashboards**
- Stay up to date with **alerts** and automatic refresh schedules



Private Preview

# Dashboards vNext



## Simple and Beautiful

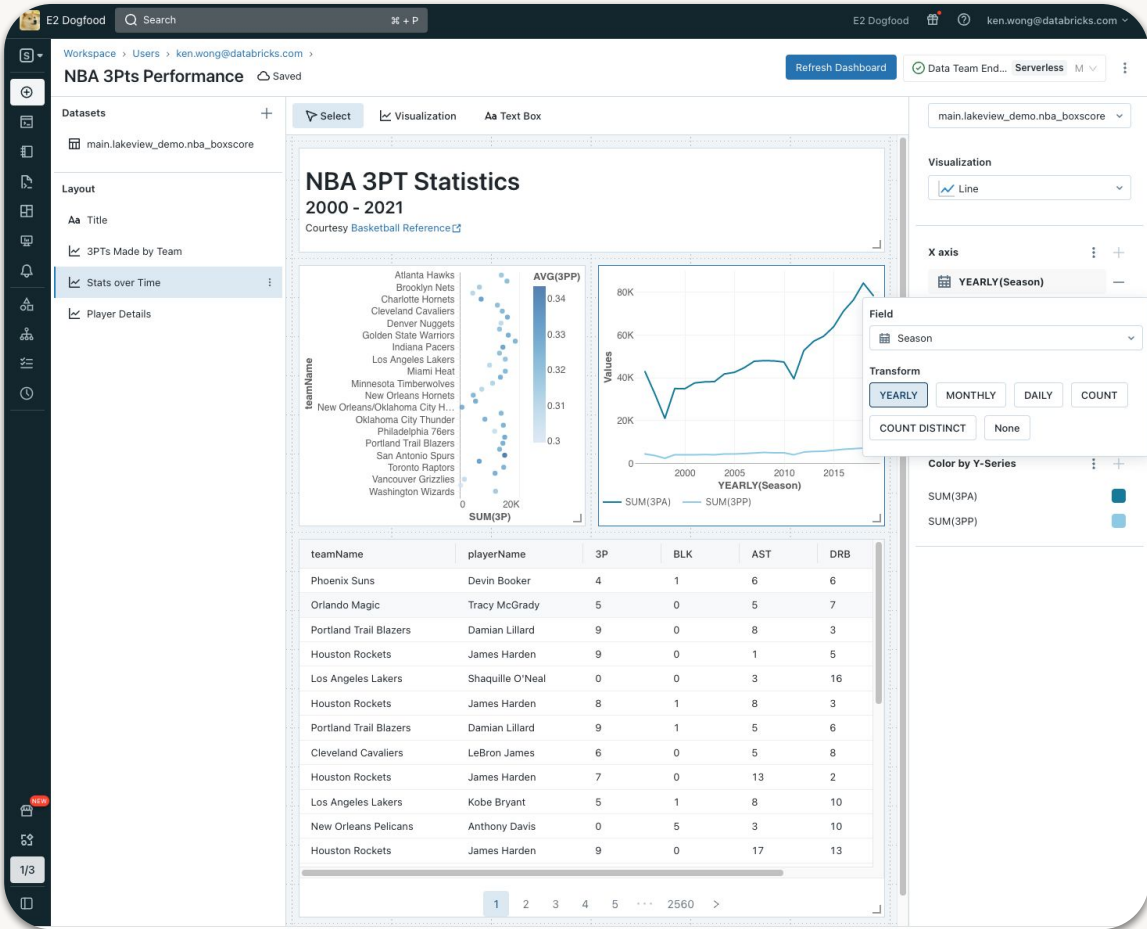
Simplified content model, new visualization library, and **SQL-optional** UX experience

## Optimized for Distribution

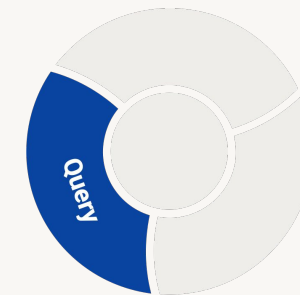
Publish and Share to Org

## Platform Integration

Unity Catalog powered dataset search and lineage



# Write SQL to get insight from unstructured text data via LLMs



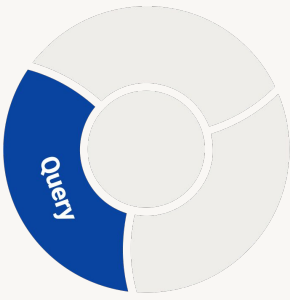
```
1 CREATE
2 OR REPLACE FUNCTION summarize(text STRING) RETURNS STRING RETURN llm_generate(
3   concat('Summarize this to 1 sentence: ', text),
4   'openai/gpt-3.5-turbo',
5   'apiToken',
6   secret('username', 'openai_api_token'),
7   'temperature',
8   0.0
9 );
10 SELECT
11   product_name,
12   summarize(long_product_description) AS product_summary
13 FROM
14   products;
15
```

## Sample use cases

- Extract top product issues from call center transcripts—without manual tagging!
- Tag customers as a potential churn risk based on customer support chat logs
- Generate customized product descriptions for ad campaigns—automatically
- Read product reviews to understand buying decision criteria

...many more...





# Python User Defined Functions (UDFs)

Run Python UDFs from an isolated execution environment

Integrate **Machine Learning** models, **custom logic** & bring the flexibility of Python right into Databricks SQL!

```
CREATE FUNCTION redact(a STRING)
RETURNS STRING
LANGUAGE PYTHON
AS $$
import json
keys = ["email", "phone"]
obj = json.loads(a)
for k in obj:
    if k in keys:
        obj[k] = "REDACTED"
return json.dumps(obj)
$$;
```

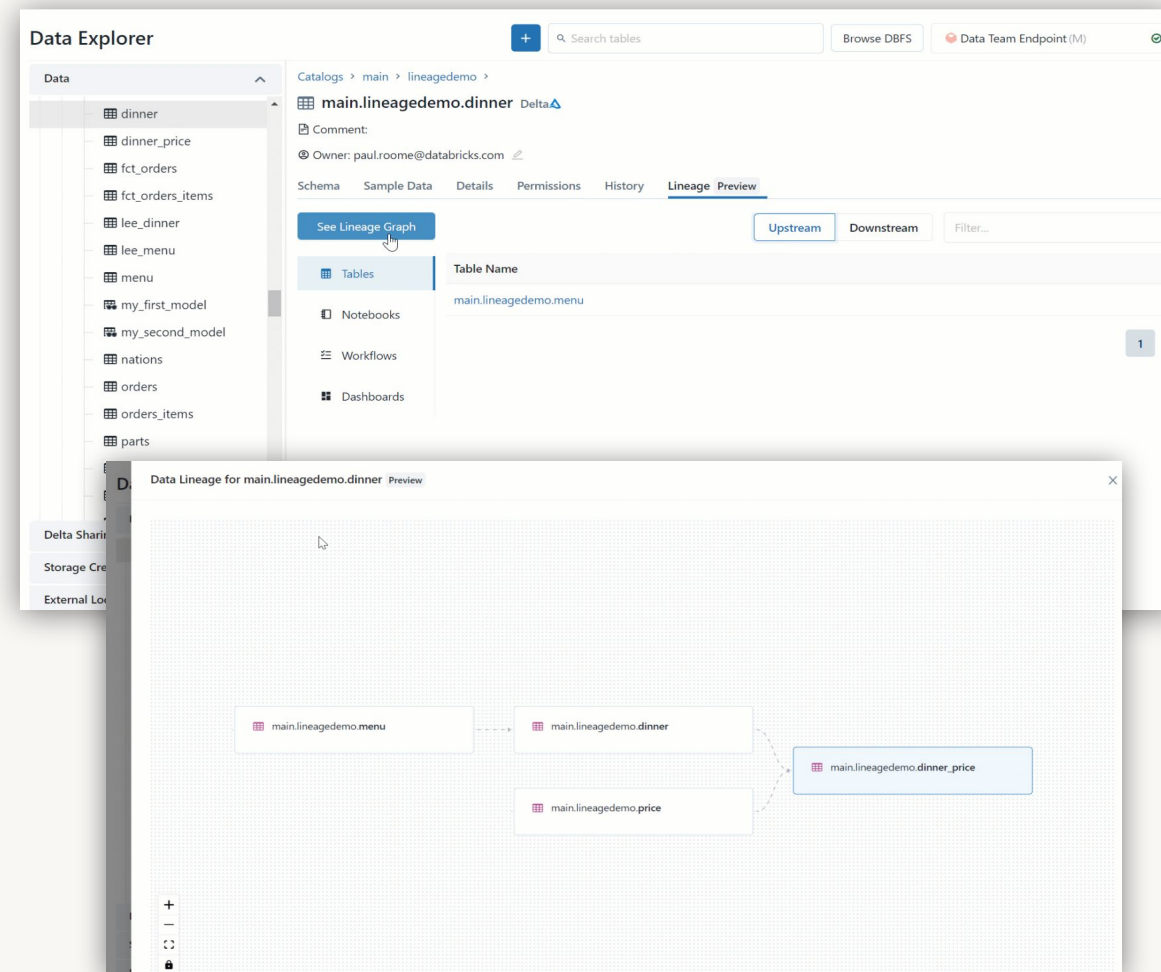


# Unity Catalog

Discover, audit and govern all data assets in one place



- **Fine-grained access and audit controls** using familiar SQL
- **Easily search, discover and access** all data assets from data explorer
- Secure **cross-workspace data access** for SQL warehouses
- **Securely share** live data across platforms, clouds and regions, without data replication
- Automated **data lineage** across tables, columns, notebooks, workflows, dashboards





# Lakehouse Federation

Discover, query, and govern all your data – no matter where it lives



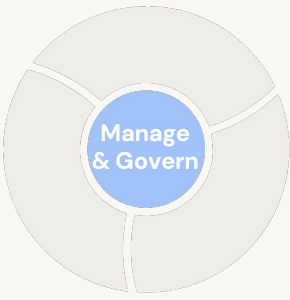
- **Unified view** into all your data
- **Unified engine** for all your data and use cases
- **Unified governance** across all data sources

Sign up @ [databricks.com/qfpreview](https://databricks.com/qfpreview)



# Intelligent Workload Management

## Efficient compute utilization



Workload Management is about efficient compute utilization – when and where to run a query, when to scale up or down, controls for cancelling an execution, etc

- Statement timeouts at workspace and query level already available
- Additional ongoing investment in intelligent auto-scaling, adaptive routing & remote result cache

Mixed Workloads Query Latency (Seconds)  
Lower is better

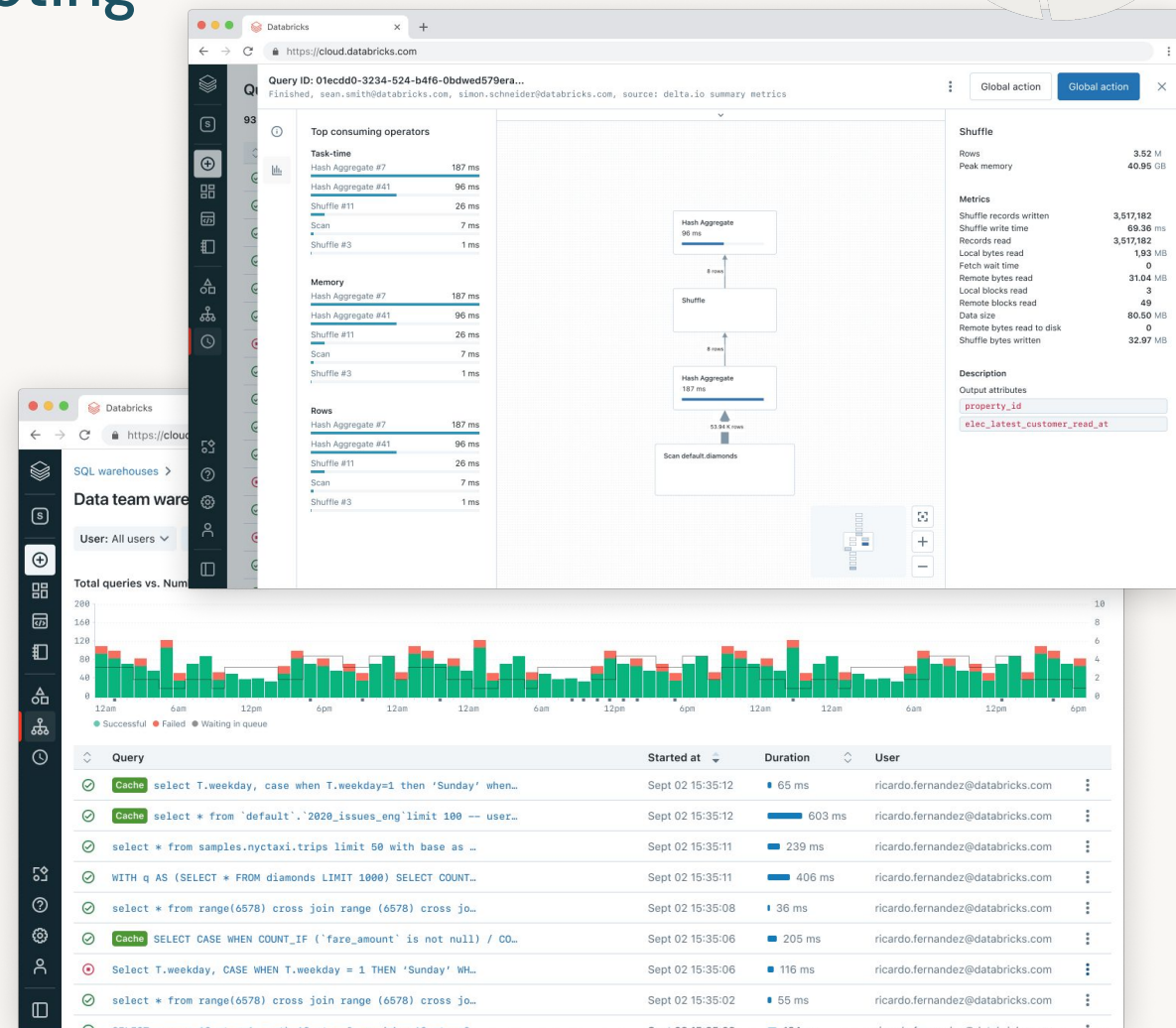


# Query History & Profile

## Execution details review & troubleshooting



- Get full transparency and visibility into query execution with in-depth breakdown at operation level details
- Identify bottlenecks and expensive operations to enhance queries
- Analyze by time spent, data volume and resource usage



# DBSQL System Tables

Bronze tables providing visibility into platform activity



What statements were run  
by whom & when?

```
> warehouse_events  
> warehouses
```

How & when did warehouses scale?

```
> statement_history
```

What was I billed for?

```
> billing
```

Sign up @ [tinyurl.com/sys-tables](https://tinyurl.com/sys-tables)

# Data Warehousing talks

## Wednesday, June 28

- [Databricks SQL: Why The Best Serverless Data Warehouse Is A Lakehouse](#) @2:30 PM
- <<< In this room next! >>> [What's New In Databricks SQL -- With Live Demos](#) @3:30 PM <<< In this room next! >>>
- [Under the Hood: Intelligent Workload Management](#) @4:00 PM
- [Unleashing Large Language Models with Databricks SQL's AI Functions](#) @5:00 PM

## Thursday, June 29

- [Best Practices For Setting Up Databricks SQL At Enterprise Scale](#) @12:30 PM
- [Databricks SQL Serverless Under the Hood: How We Use ML to Get the Best Price/Performance](#) @2:30 PM
- [Building Apps on the Lakehouse with Databricks SQL](#) @2:30 PM
- [AI-Accelerated Delta Tables: Faster, Easier, Cheaper](#) @3:30 PM
- [Unlock The Next Evolution Of The MDS With The Lakehouse Revolution -- With Live Demos](#) @3:30 PM
- [Going Beyond SQL: Python UDFs in Unity Catalog for all your Lakehouse](#) @2:30 PM



# Customer talks at Data+AI Summit

## Tuesday, June 27

- **Rec Room** | [How Rec Room Processes Billions of Events Per Day with Databricks and RudderStack](#) @12:00 PM
- **AT&T** | [Building and Managing Data Platform for 13+ PB Delta Lake and 1000s of Users: AT&T's Story](#) @1:00 PM
- **Collins Aerospace** | [Jet Streaming Data and Predictive Analytics](#) @2:00 PM
- **Michelin** | [Data Democratization at Michelin](#) @3:00 PM

## Wednesday, June 28

- **Banco Bradesco** | [Data Democratization with Lakehouse: An Open Banking Application Case](#) @11:30 AM
- **Zurich Insurance** | [Modernizing the Data Stack: Lessons Learned From the Evolution at Zurich Insurance](#) @12:30 PM
- **S&P GLOBAL** | [Using Databricks to Power Insights and Visualizations on the S&P Global Marketplace](#) @1:30 PM
- **Akamai** | [Internet-Scale Analytics: Migrating a Mission Critical Product to the Cloud](#) @2:30 PM
- **American Airlines** | [Making Travel More Accessible For Customers Bringing Mobility Devices](#) @3:30 PM
- **Land O'Lakes** | [Self-Service Geospatial Analysis Leveraging Databricks, Apache Sedona, And R](#) @4:30 PM

## Thursday, June 29

- **RaceTrac Inc.** | [Unlocking the Power of Real-Time Data to Maximize Data Insights](#) @11:30 AM

**Questions?**



# Learn more at the summit!



Databricks  
Events App



## Tells us what you think

- We kindly request your valuable feedback on this session.
- Please take a moment to rate and share your thoughts about it.
- You can conveniently provide your feedback and rating through the **Mobile App**.



## What to do next?

- Discover more related sessions in the mobile app!
- Visit the Demo Booth: Experience innovation firsthand!
- More Activities: Engage and connect further at the Databricks Zone!



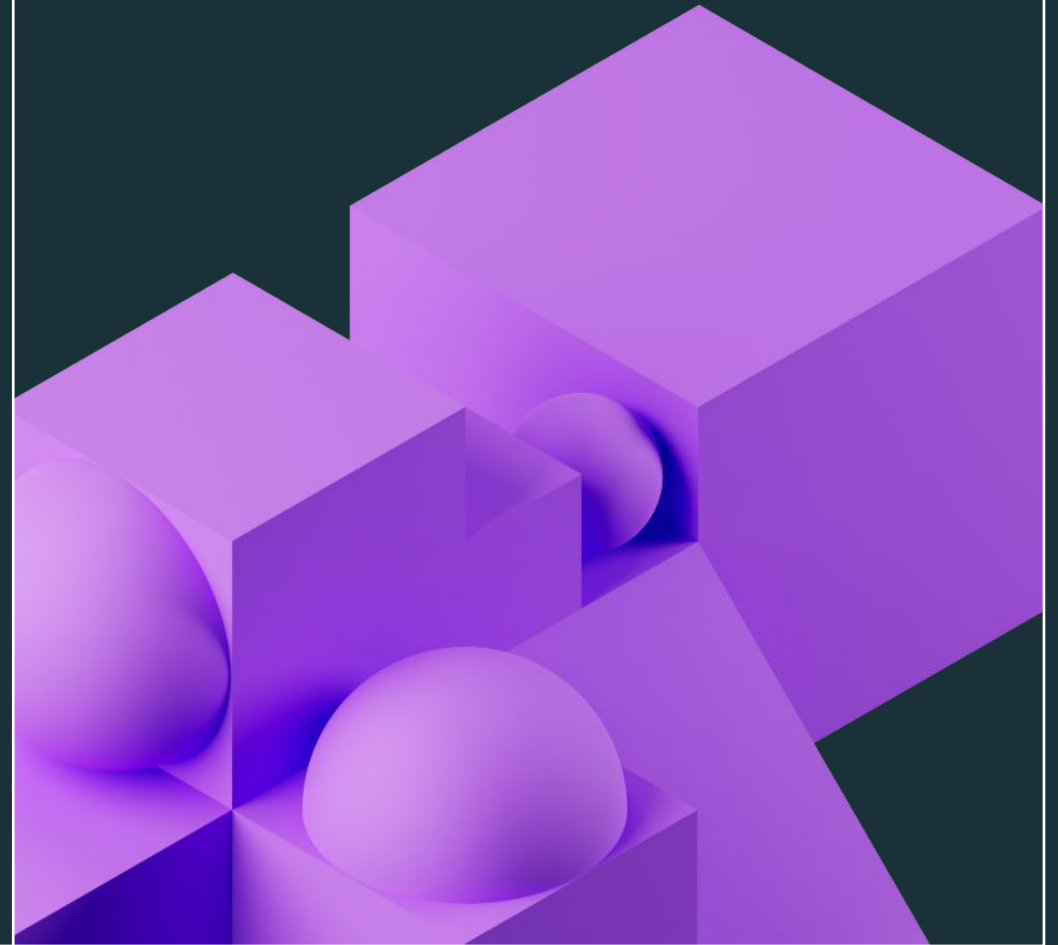
## Get trained and certified

- Visit the Learning Hub at the Databricks Zone!
- Take complimentary certification at the event; come by the Certified Lounge
- Visit our Databricks Learning website for more training, courses and workshops! [databricks.com/learn](https://databricks.com/learn)





Thank you!





# SQL to DLT in three easy steps...

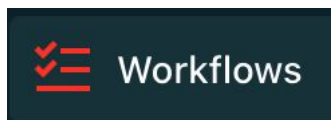
## Write create live table

- Table definitions are written (but not run) in notebooks
- Databricks Repos allow you to version control your table definitions.

```
1 CREATE LIVE TABLE daily_stats
2 AS SELECT sum(rev) - sum(costs) AS profits
3 FROM prod_data.transactions
4 GROUP BY day
```

## Create a pipeline

- A Pipeline picks one or more notebooks of table definitions, as well as any configuration required.



Delta Live Tables

## Click start

- DLT will create or update all the tables in the pipelines.



# Pipeline definition

A **materialized view** (a.k.a live table) is “stateless”:

- Results of the MV are kept up to date by the pipeline

```
CREATE OR REPLACE LIVE TABLE report
AS SELECT sum(profit)
FROM prod.sales
```

A **streaming table** is “stateful”:

- Ensures exactly-once processing of input rows
- Computes results over append-only streams

```
CREATE STREAMING LIVE TABLE report
AS SELECT sum(profit)
FROM cloud_files(prod.sales)
```

