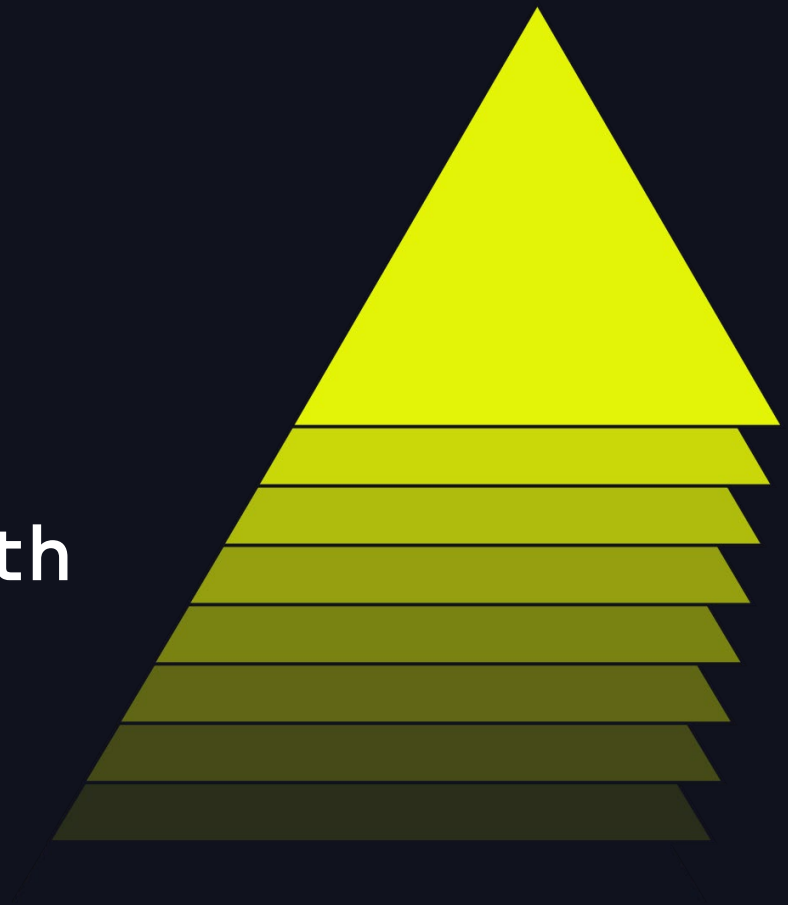
DATA+AI SUMMIT BY databricks

DATABRICKS SQL

Addressing Data Warehousing's Biggest Challenges with Data Intelligence



Kevin Clugage, Principal Product Marketing Manager, Databricks
Gaurav Saraf, Sr. Staff Product Manager, Databricks

Complementary Sessions

Part 1 **This session**

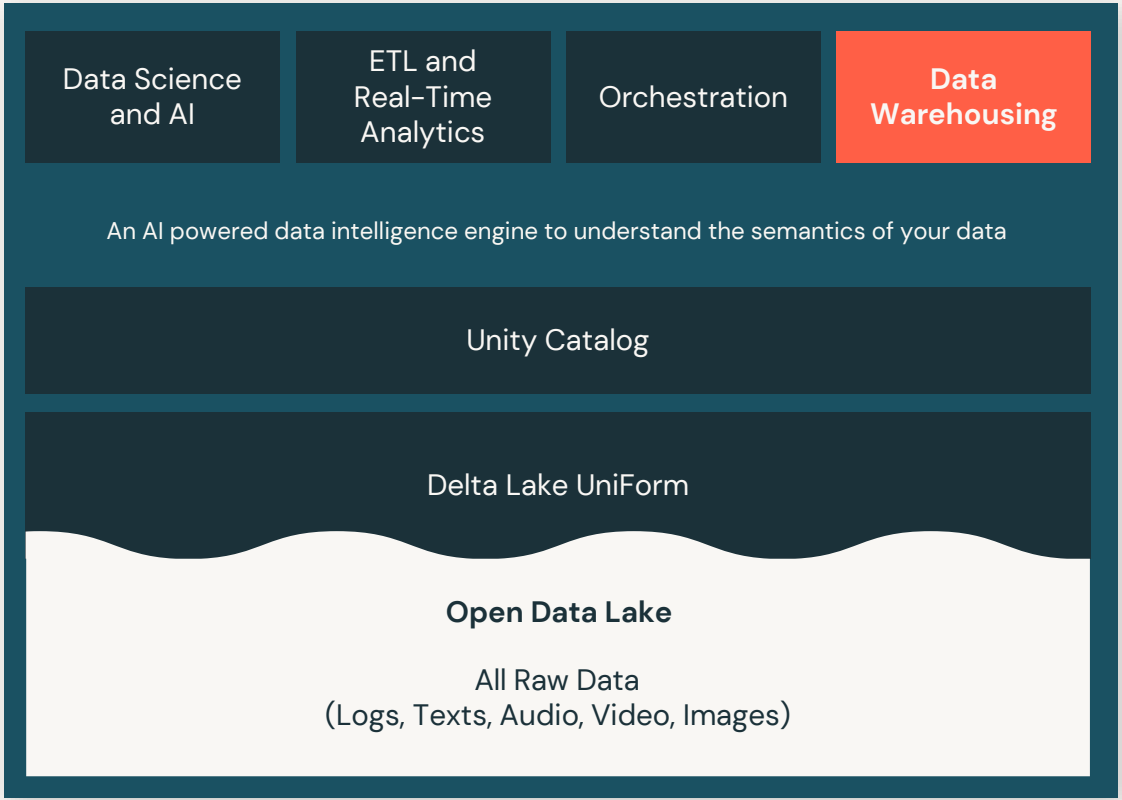
Part 2 **What’s new in Databricks SQL with live demos**

@12:30pm Moscone South, Rm 209 (down 1 floor)

Product safe harbor statement

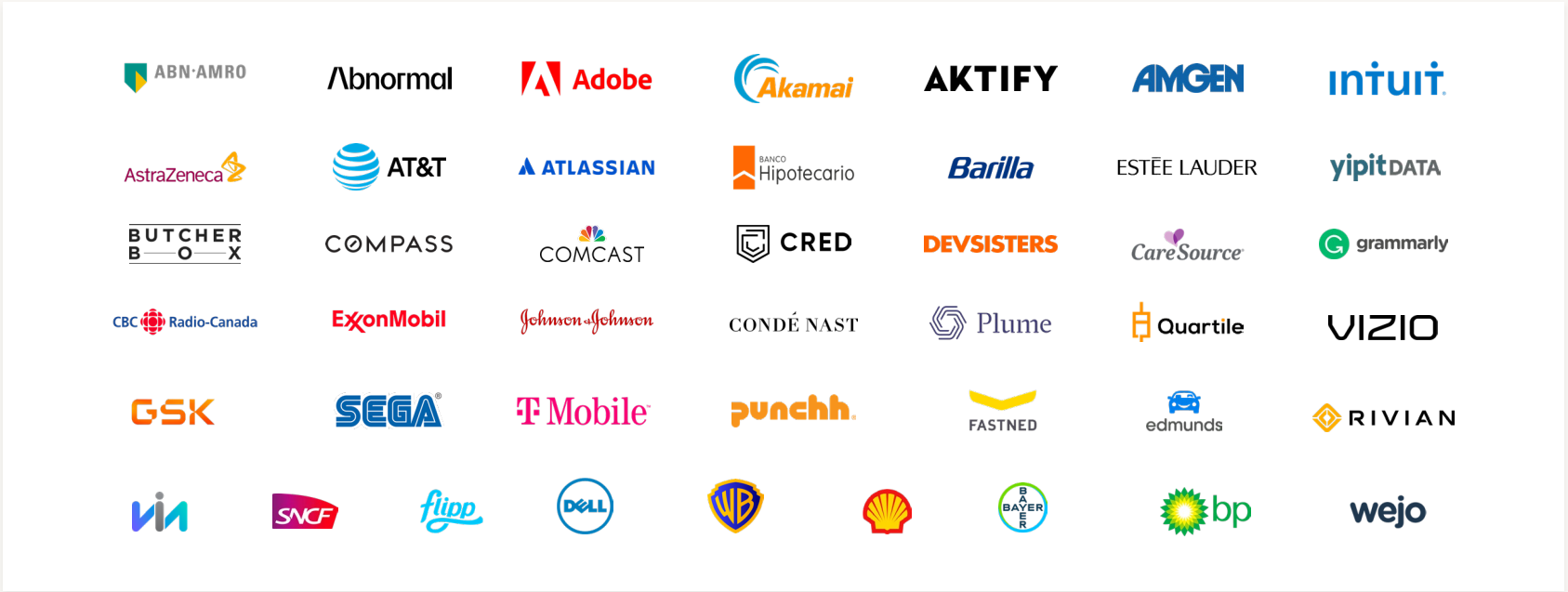
This information is provided to outline Databricks' general product direction and is for **informational purposes only.** Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features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

Databricks SQL
intelligent data warehousing on the Data Intelligence Platform



Trusted by organizations of all sizes

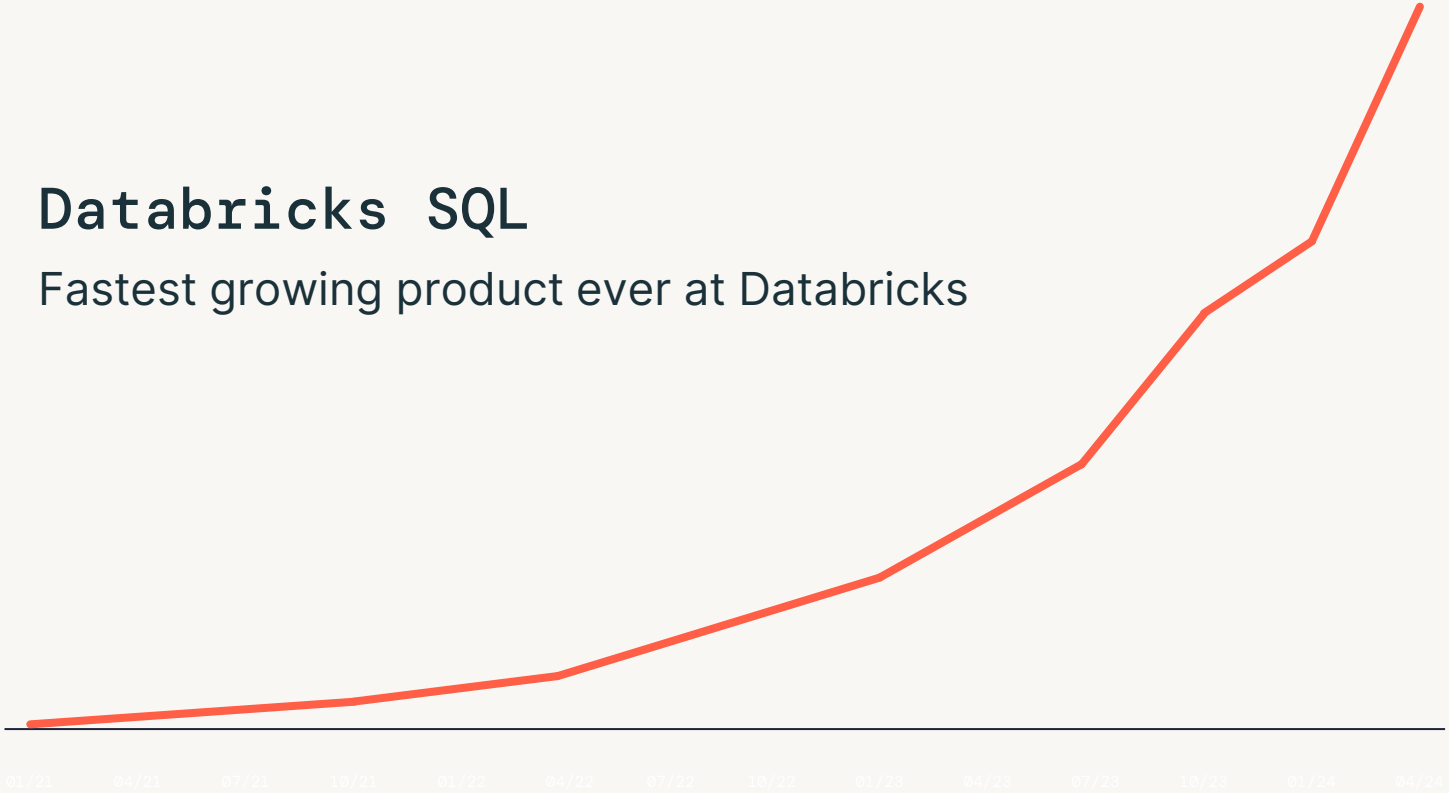
7,000+ data warehouse customers on the lakehouse



©2024 Databricks Inc. — All rights reserved

Databricks SQL

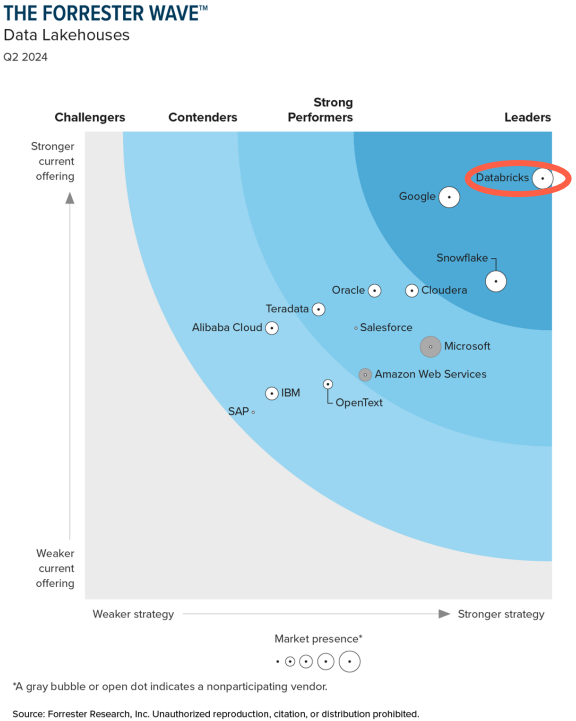
Fastest growing product ever at Databricks



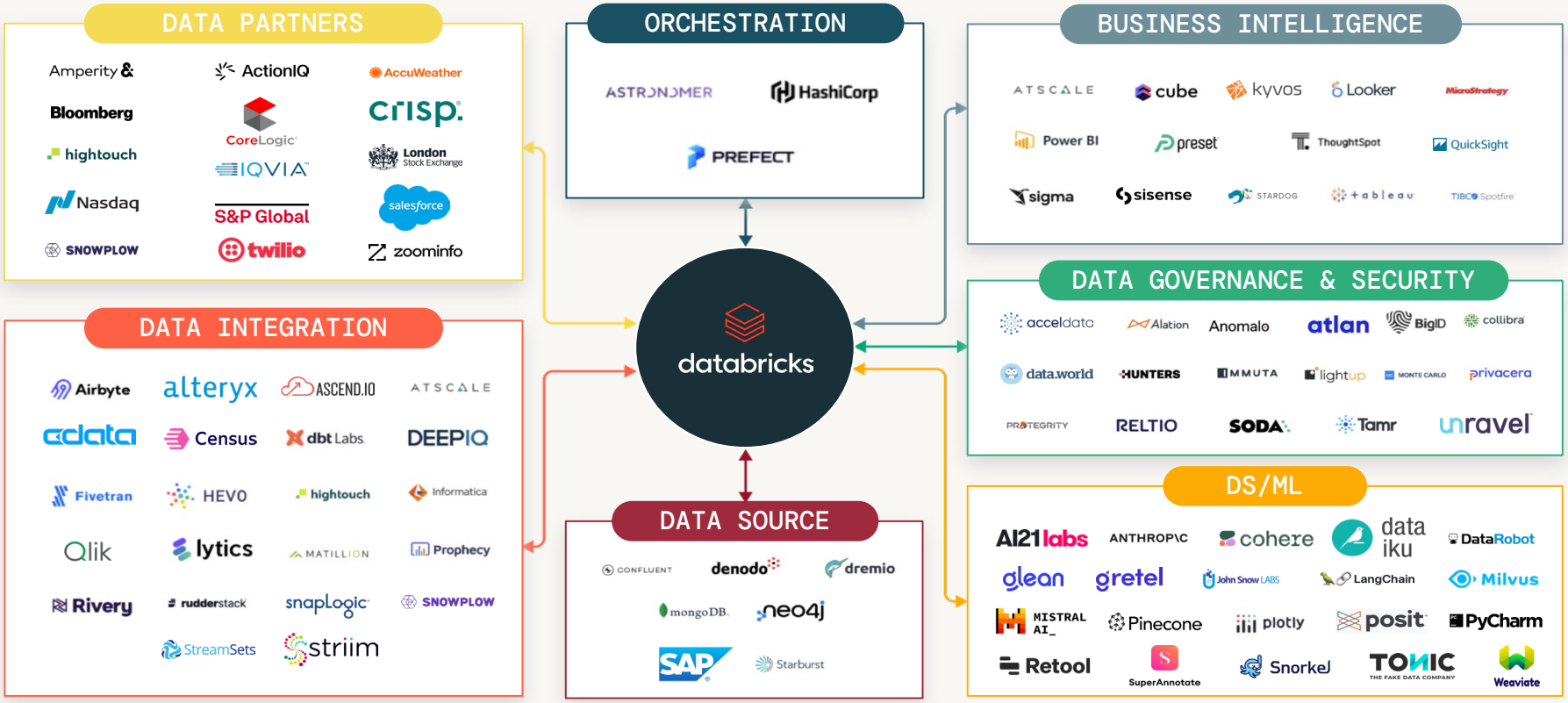
Recognized as a leader in the industry

Gartner MQ Leader
Database Management Systems

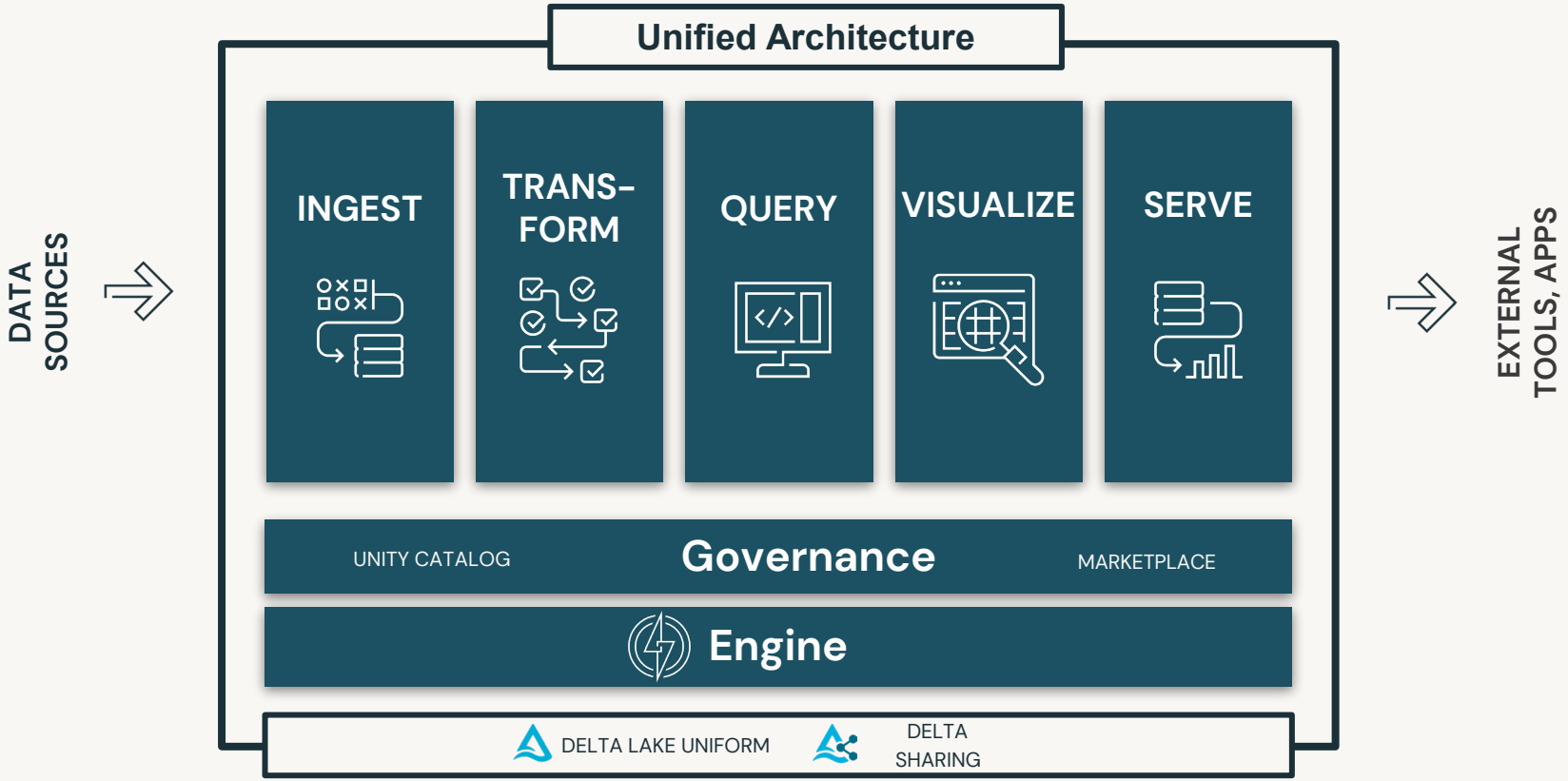
Forrester Wave Leader:
Data Lakehouses



Integrated with the tools you know and love



Databricks SQL



Complete Data Warehousing Capabilities

Ingest & Transform

- Streaming Tables
- SQL Workflows
- Partner Connectors
- 3P Orchestration
- COPY INTO
- DBX Orchestration
- Materialized Views
- Structured Streaming
- Autoloader for Ingest
- dbt-databricks
- Add Data via UI
- FiveTran connectors

Query, Visualize & Serve

- SQL Editor
- Dashboards
- Serverless warehouses
- ANSI SQL
- SQL Alerts
- JDBC/ODBC Drivers
- SQL UDFs
- H3 GeoSpatial
- SQL Rest API
- Python UDFs
- Query Profiler
- Py/Go/Node.js Conn
- Temp Views
- PK/FK support
- Rich Visualizations
- Array Functions
- JSON functions

Govern & Manage

- Table ACLs
- Data Quality Monitor
- Warehouse Monitoring
- Schema Browser
- Ext HMS support
- Query Duration Limits
- Table Lineage
- Query Federation
- Query History
- Delta Sharing
- Marketplace
- Billing System Tables
- OAuth
- Warehouse APIs
- OAuth

Engine

- Photon (MPP)
- Intelligent WL Mgmt
- Liquid Clustering
- Predictive I/O
- Predictive Optimization
- Cloud Fetch
- Automated Data Layout
- Query Results Cache
- Adaptive Routing
- Join Hints

New Capabilities and Enhancements

Ingest & Transform

- Native Connectors
- SQL Workflows
- Native CDC Support
- DBT w/ MVs & STs
- Streaming Tables
- Materialized Views

Query, Visualize & Serve

- SQL WH in Notebooks
- AI/BI Dashboards
- AI/BI Genie
- New Results Table
- Session Vars
- Publish to Tableau
- SQL Scripting
- PK/FK Rely
- Publish to PowerBI
- Vector Search via SQL
- Variant Data Type
- Collations
- AI/LLM Functions
- Spatial SQL
- AI Forecasting

Govern & Manage

- Row level security
- Serverless WH
- WH Events Sys Table
- AI Catalog Comments
- Entity Relations
- Query Hist Sys Table
- Column level Masking
- LH Monitoring
- Warehouses Sys Table
- ABAC
- Query Federation
- Cost Dashboard
- Monitor Permissions
- UC HMS Federation
- Budgets

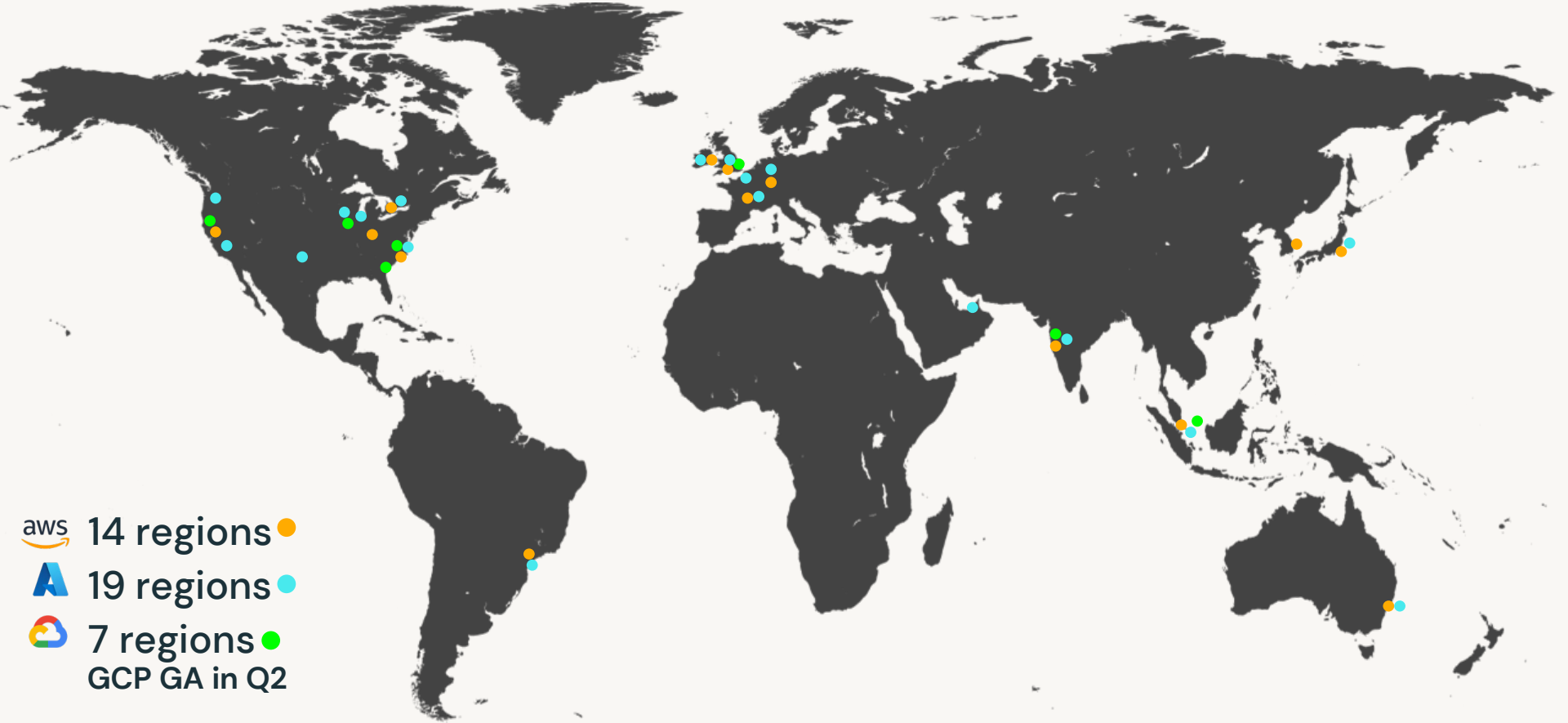
Engine

- Predictive Optimization
- Intelligent WL Mgmt
- Liquid Clustering
- Deletion Vectors
- Automatic Statistics
- Row Concurrency

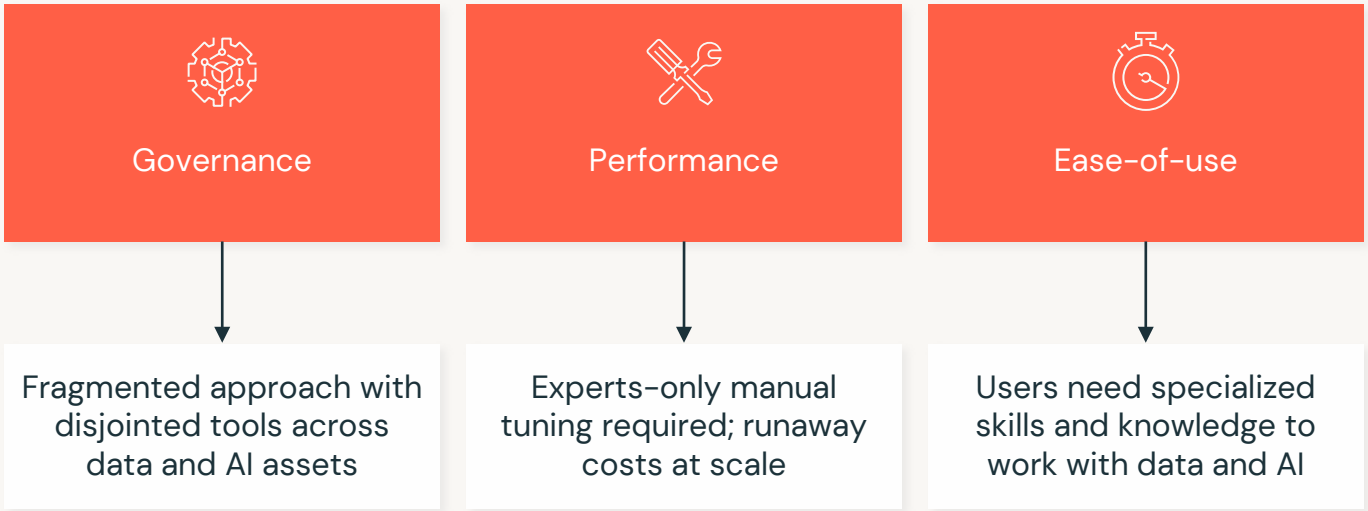
New feature

Enhancement, Release progression

SQL Serverless available globally



Biggest challenges for data warehousing



Governance

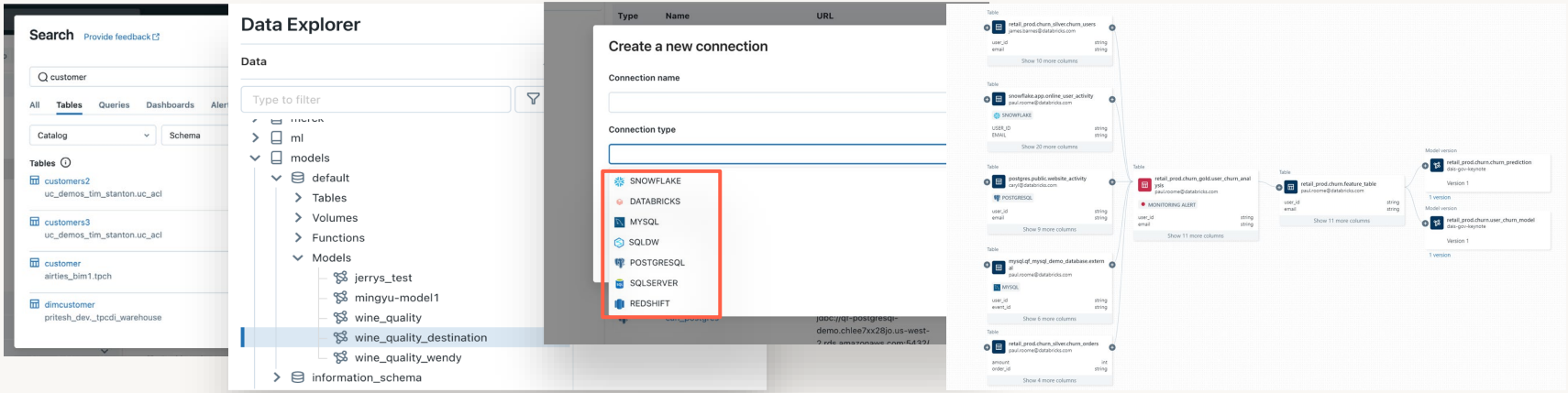
Discovery | Access | Federation

Sharing | Lineage | Auditing

Administration

Governed and secured by Unity Catalog

Governance for all your data and AI assets



Simplified **data discovery**, **federation**, **lineage**, and **compliance** with enhanced **security** and **auditing** with Unity Catalog and Databricks SQL

Announcing GA

Lakehouse Federation: Databases & Warehouses

Unify your data estate with the Lakehouse
Discover, query, and govern all your data – in any system

Database & DW - General Availability

What's new?

- Improved pushdown coverage & performance for Snowflake, SQL Server, Postgres, Redshift & Synapse.
- OAuth support for Snowflake connections.
- Azure AD support for Azure ecosystem connections.
- Case sensitive namespace support
- Salesforce Data Cloud Connector (Preview)

Public Preview

Row Level Security and Column Level Masking

Provide differential fine grained access to file based datasets and foreign tables

Only show specific rows

```
CREATE FUNCTION <name> ( <parameter_name >
<parameter_type> .. )
RETURN {filter clause whose output must be a boolean}

CREATE FUNCTION us_filter(region STRING)
RETURN IF(IS_MEMBER('admin'), true, region="US");

ALTER TABLE sales SET ROW FILTER us_filter ON
region;
```

Test for group membership

Assign reusable filter to table

Specify filter predicates

Mask or redact sensitive columns

```
CREATE FUNCTION <name> (<parameter_name>,
<parameter_type>, [, <column>...])
RETURN {expression with the same type as the first
parameter}

CREATE FUNCTION ssn_mask(ssn STRING)
RETURN IF(IS_MEMBER('admin'), ssn, "****");

ALTER TABLE users ALTER COLUMN table_ssn SET
MASK ssn_mask;
```

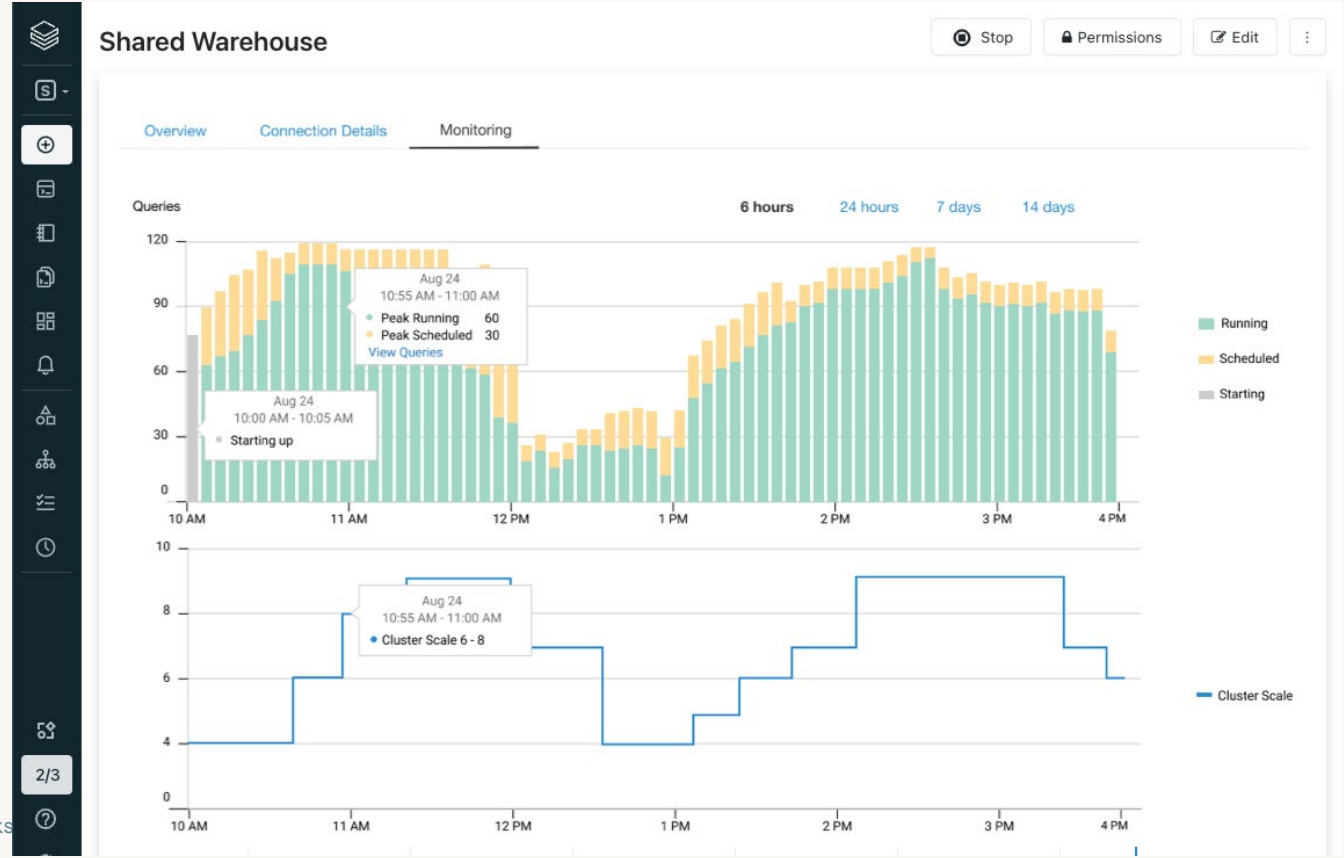
Test for group membership

Assign reusable mask to column

Specify mask or function to mask

SQL Warehouse Monitoring

Real Time UI



Public Preview

System Tables -> Monitor & Alert

Public Preview

Billing

Warehouse events

Query history

Q3

Warehouses

Answer complex questions:

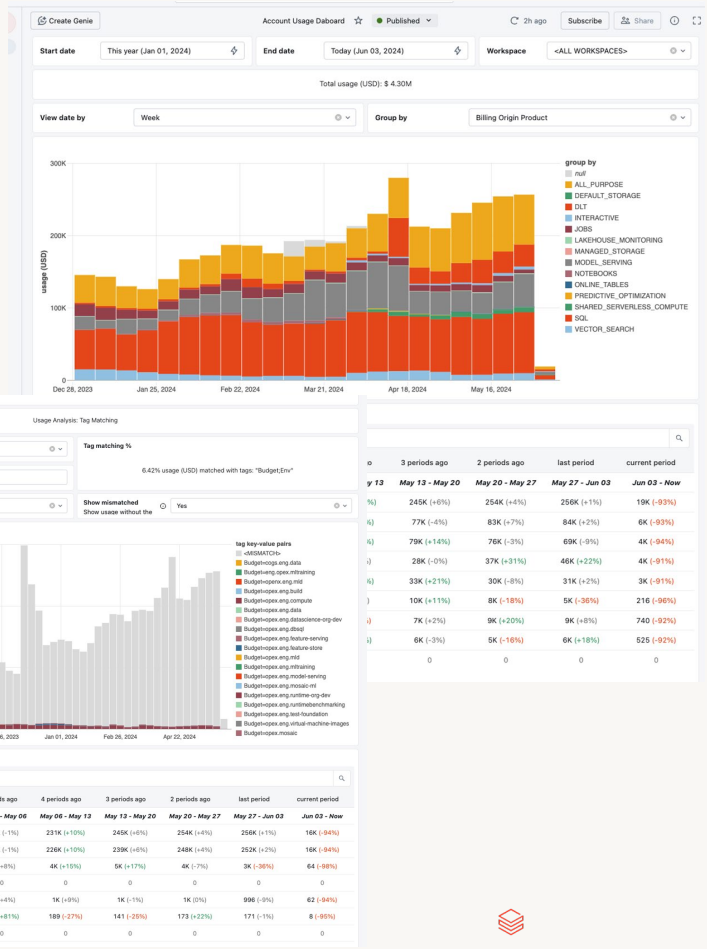
Visualize spend per warehouse per hour/day/week
Identify queries took the longest to run
Attribute warehouse spend by user or source tool
Track New Warehouses created (or ones without a tags!)
Changes to Warehouse settings (T-shirt size etc)

DBSQL Performance, Scale and Security | Thu @5pm | Jeremy Lewallen, Alex Esibov

Public Preview

Cost Tracking Dashboard

- View usage trends in your Account or Workspaces
- Easy-to-use UI filters to drill down by product, Workspace and more
- Quickly identify the workloads, users, endpoints, etc. with the highest spending
- Attribute costs using tags and check for completeness
- Share with others by publishing it with embedded credentials

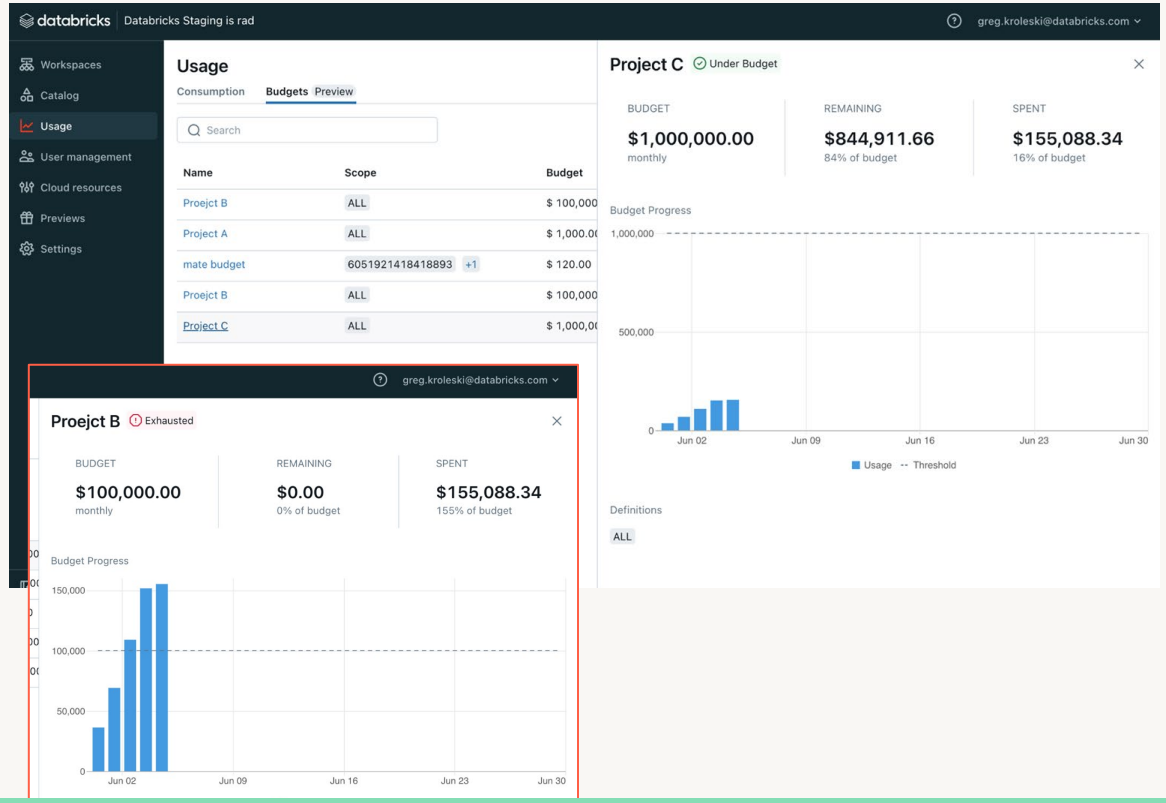


Public Preview

Databricks Budget Alerts

Manage your DBX spend!

- Use Tags & Workspaces to allocate & track budgets
- Get Alerted when spend surpasses threshold
- Easy UI interface and API access



Performance & TCO

AI systems at every layer of the stack

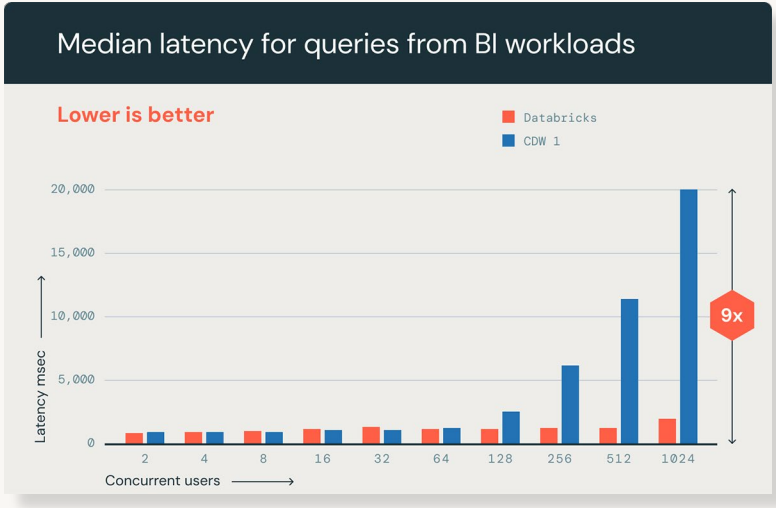
Telemetry feeds models that replace the classic tuning knobs

	~~Knobs~~	Telemetry	AI Models
 **Data Layout**	*Partitions* *Analyze* *Vacuum* *Optimize*	*Data layout* *Query patterns* *Table statistics*	**Intelligent & Predictive Data Layout Optimization for fast data access**
 **Query Engine**	*Indexes* *Query optimization* *Hints*	*Query plans* *Query patterns* *Table statistics*	**Fast planning & execution with intelligent data retrieval & writes**
 **Workload Mgmt**	*Scheduling* *Max Concurrency* *Max clusters* *Cluster sizes*	*Compute demand* *Query queue* *Query size* *Cluster utilization*	**Intelligent routing of queries to compute with Instant & Elastic compute capacity**

GA

Intelligent Workload Management

- Uses machine learning to efficiently route queries and autoscale clusters based on actual workloads
- Benefits
  - Protects query latency by routing queries to best cluster and/or upscaling quickly when queueing occurs
  - Reduces costs by minimizing always-on clusters and scaling down quickly.



GA

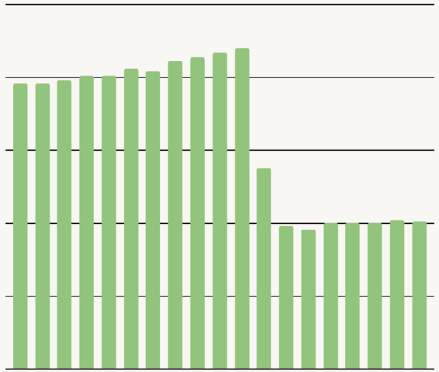
Predictive Optimization

AI-Optimized Delta table layouts for best price-performance

Runs OPTIMIZE, VACUUM, ANALYZE, Liquid clustering

AI model prioritizes tables to maximize ROI

Data is intelligently optimized and clustered to make querying faster and reduce storage cost



"Databricks' Predictive Optimizations intelligently optimized our Unity Catalog storage, which **saved us 50% in annual storage costs** while **speeding up our queries by >2x**. It learned to prioritize our largest and most-accessed tables. And, it did all of this **automatically,** saving our team valuable time."

—Anker

Public Preview

Liquid Clustering

High-performance, easy- to-use clustering

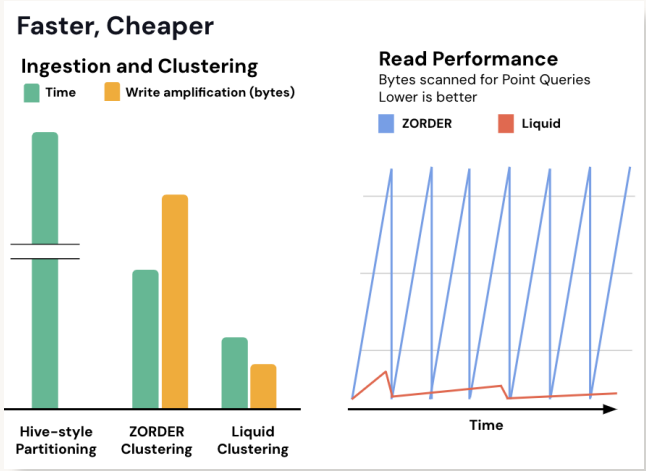
Hassle free clustering column selection

Performant, cost-effective incremental clustering

Change clustering columns at any time

Coming soon: Auto Column Selection

No more manual partitions & Zordering!



"Liquid clustering has greatly improved the ability of our researchers to investigate complex datasets for specific trends and events. It's a great option for optimizing point lookups across multiple columns. We look forward to watching this feature grow and be adopted as a key feature of the Delta ecosystem."

—Cisco

Private Preview

Automatic statistics

Efficient statistics collection for improving query performance

Runs during ingest + via Predictive Optimization

AI model prioritizes tables to maximize ROI

More efficient than running ANALYZE separately

33% faster*

On average, for queries with statistics

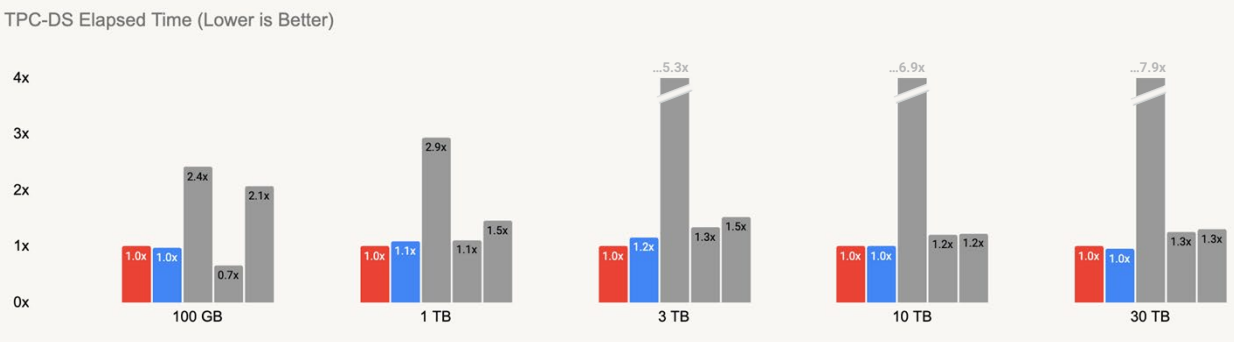
*TPC-DS 100TB overall benchmarks

Impact?

Best Perf/Performance as data scales

Meets or beats major CDWs across scales!

PERF

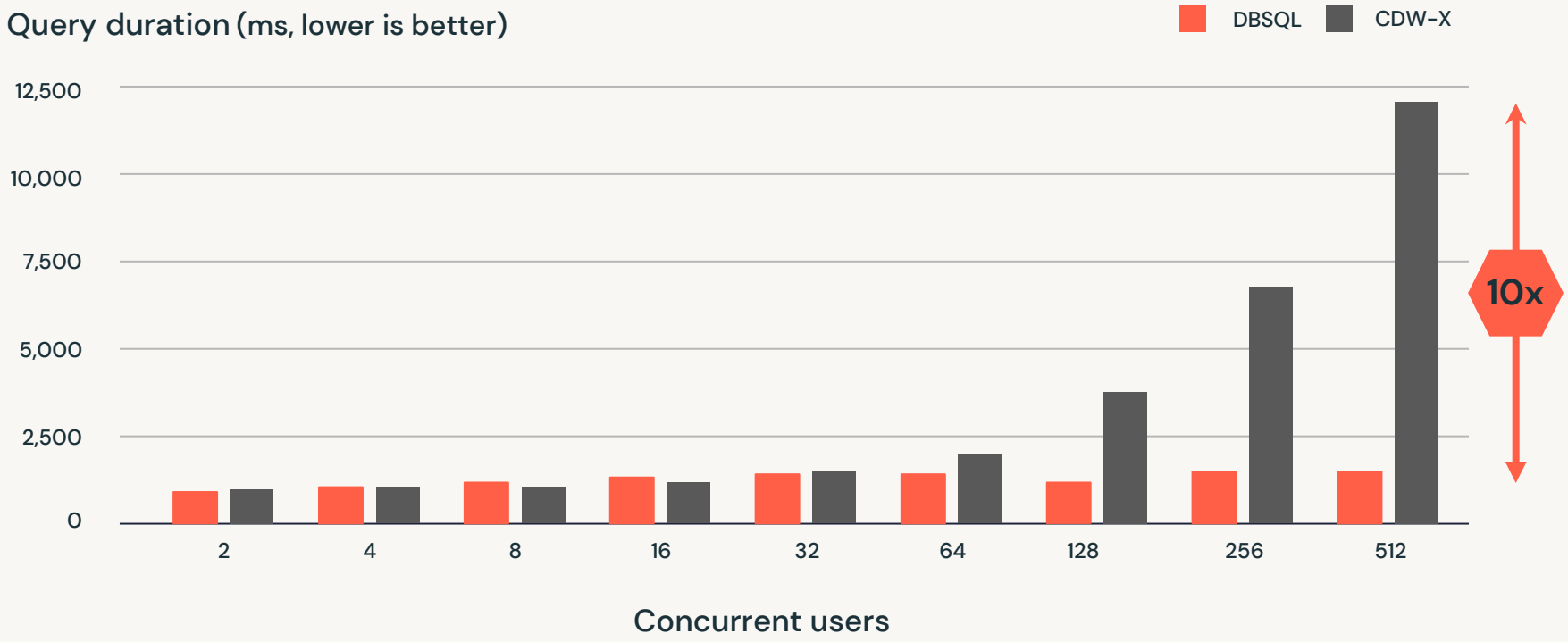


COST

Databricks
CDW 1
CDW 2
CDW 3
CDW 4

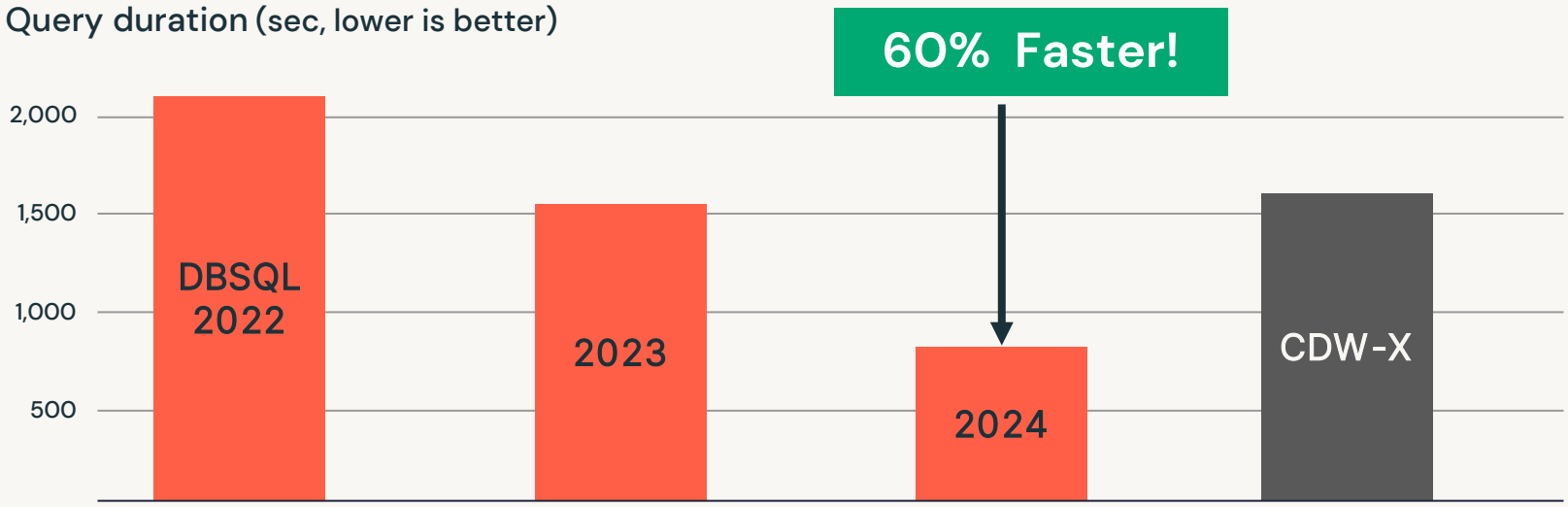
Highly Concurrent BI Queries

Latency remains flat as number of users (or queries) go up!



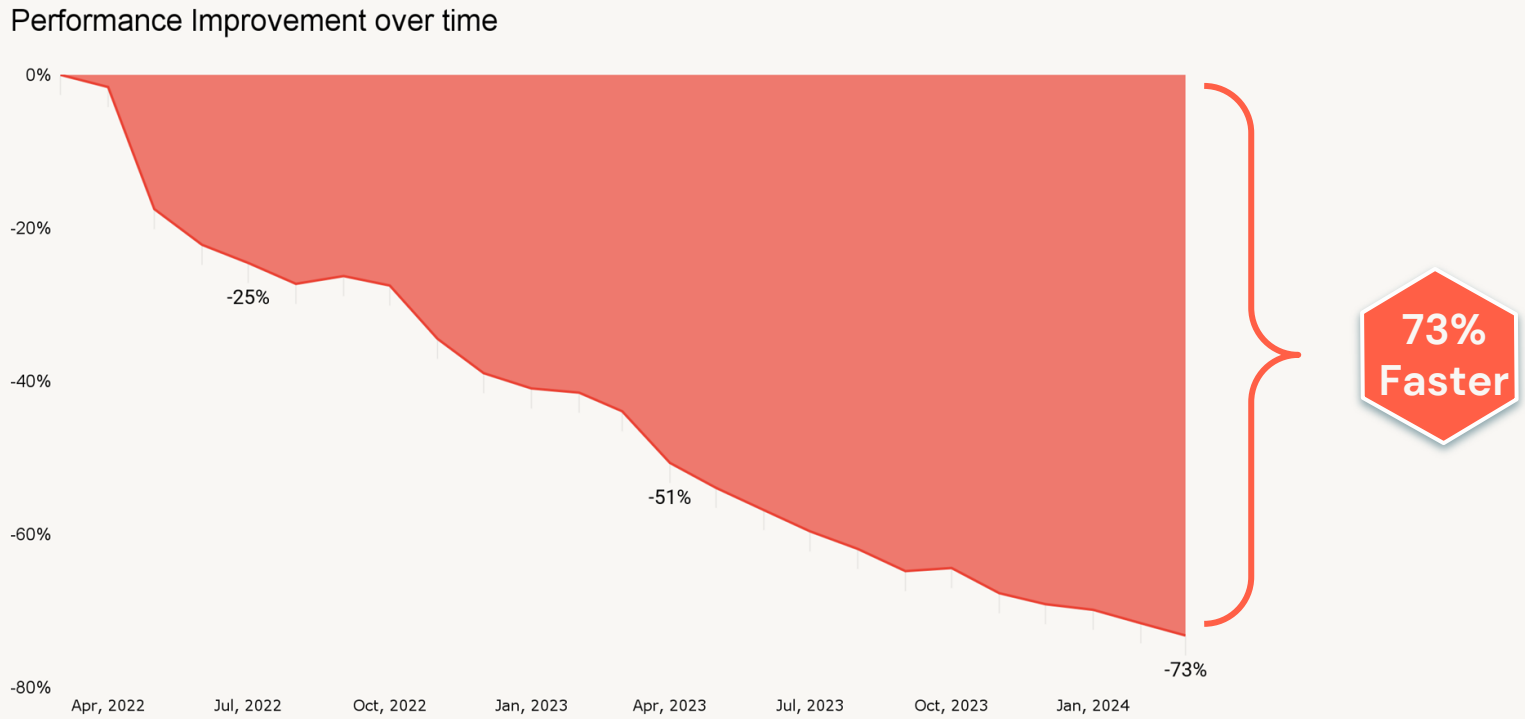
Out-of-the-box Performance

Automatic Performance improvement over time! No knobs needed



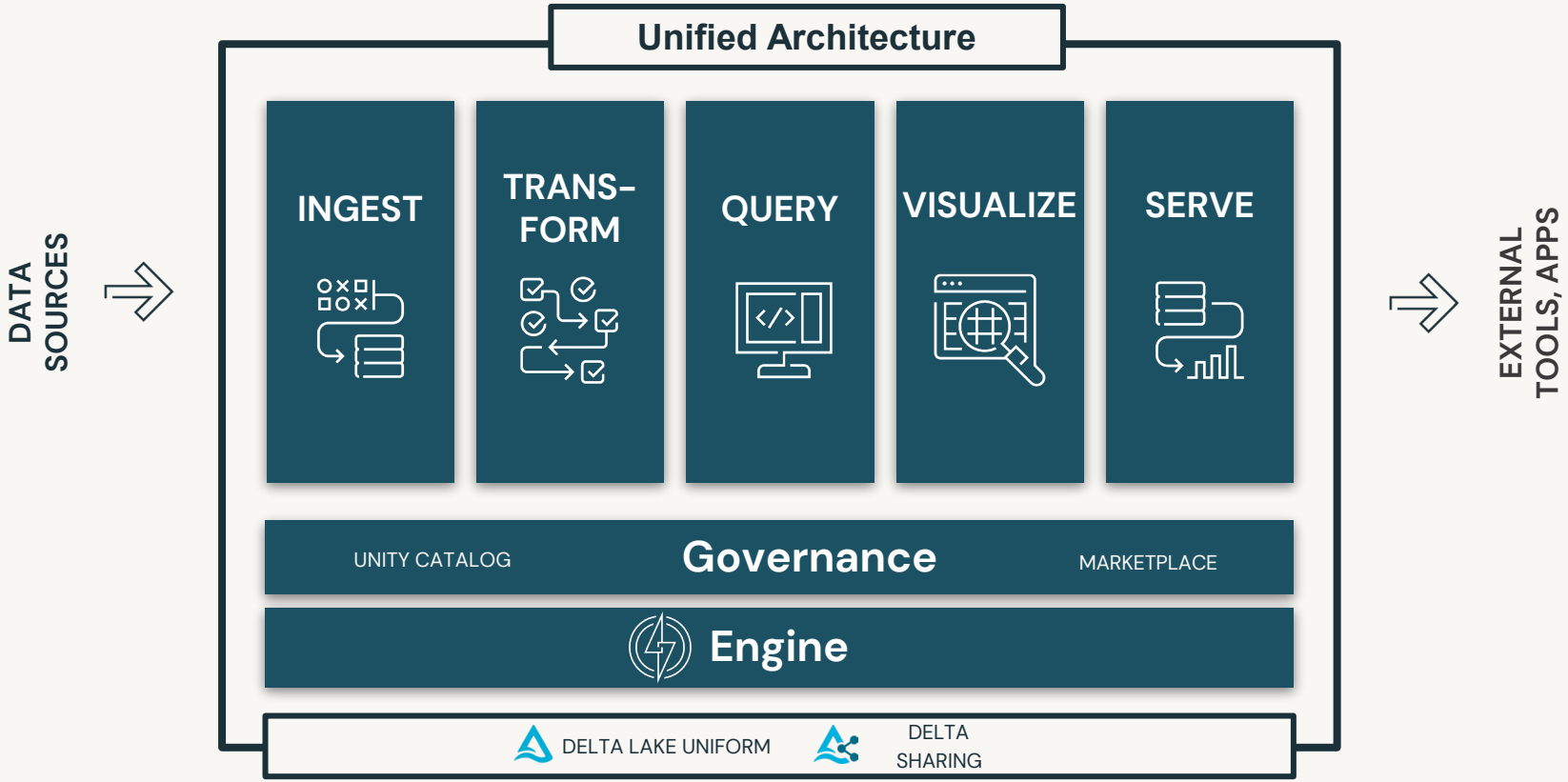
Performance improvements time

Customer BI queries – improved 73% over last 2 years

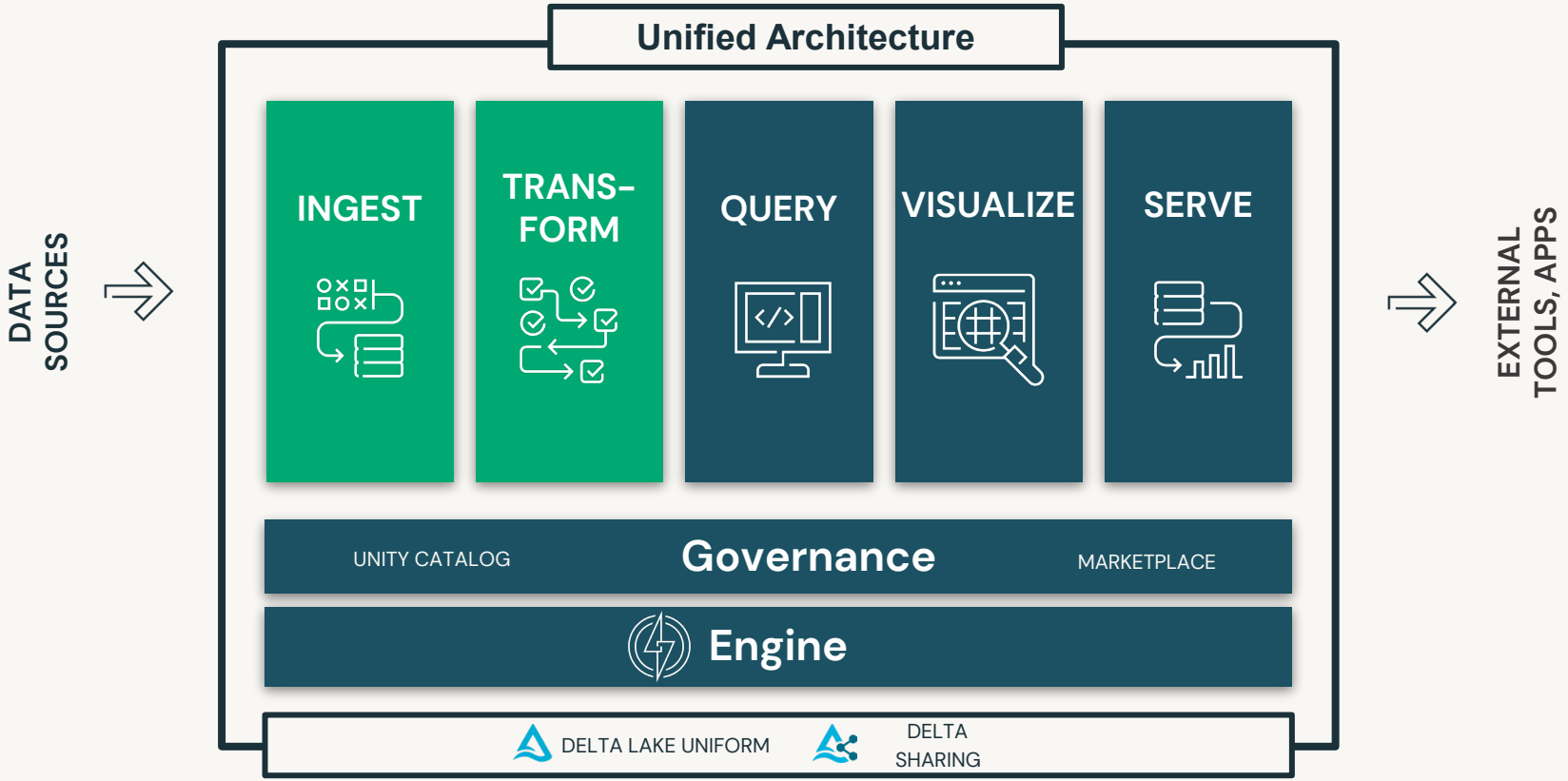


Ease-of-use

Databricks SQL



Simplifying the user experience end-to-end



Pvt Preview

Native Connectors for Ingestion

Applications

Private Preview

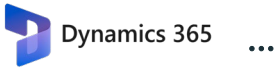
Private Preview

Coming soon

Coming soon

Coming soon

Coming soon



...

Databases

Private Preview

Private Preview

Coming soon

Coming soon

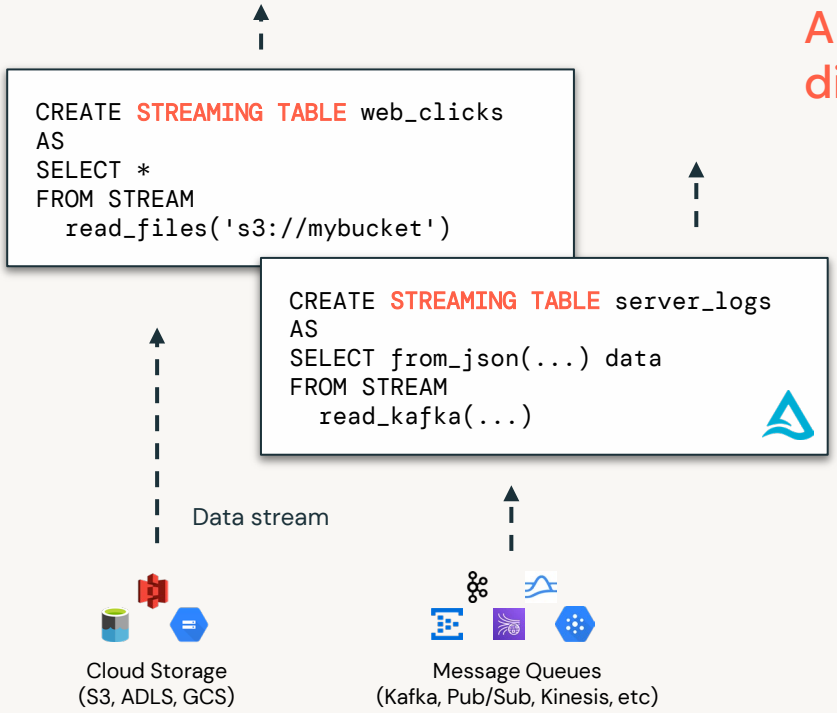
Coming soon

...

Keynote | Wed/Thu @ 8:30 AM

GA Q2

Streaming Table



A simple way to stream any data directly into the Lakehouse.

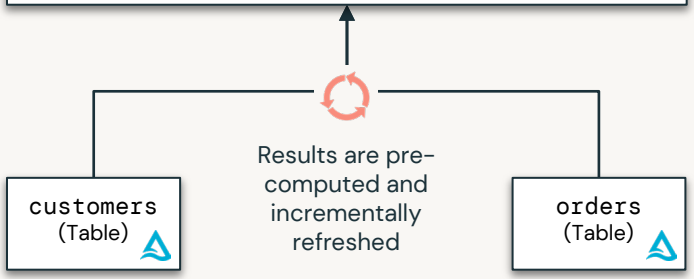
Benefits:

1. Enable more practitioners. Simple SQL syntax makes data streaming accessible to all data engineers and analysts.
2. Better scalability. More efficiently handle high volumes of data via incremental processing vs large batches.
3. Unlock real-time use cases. Ability to support real-time analytics/BI, machine learning and operational use cases with streaming data.

GA Q2

Materialized View

```
CREATE MATERIALIZED VIEW customer_orders
AS
SELECT
  customers.name,
  sum(orders.amount),
  orders.orderdate
FROM orders
  LEFT JOIN customers ON
    orders.custkey = customers.c_custkey
GROUP BY
  name,
  orderdate;
```

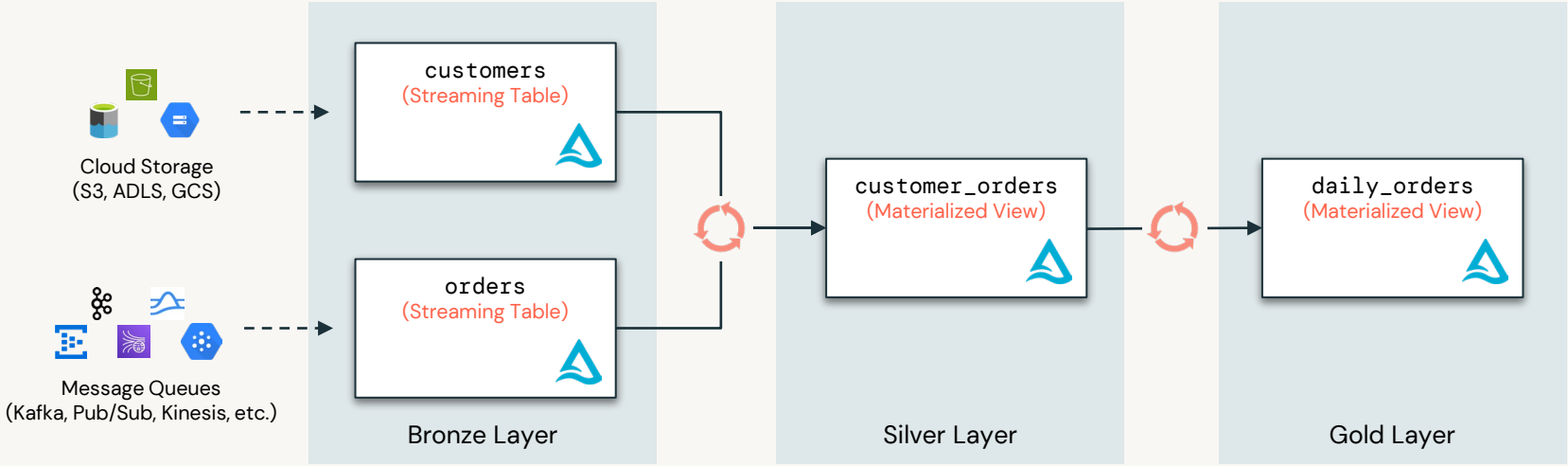


Perform complex transformations for ETL and accelerate end-user queries for dashboards/BI.

Benefits:

1. Simple ETL. Transform and process data in a declarative way.
2. Improve data freshness. MVs can be incrementally refreshed when new data arrives, avoiding time-consuming full recomputes
3. Accelerate BI dashboards. Much faster to query data that is pre-computed vs querying base tables.

Linking STs and MVs to build data pipelines



Intelligent ETL optimizations for efficient, cost-effective, and incremental table updates.

Data Transformation: dbt Labs

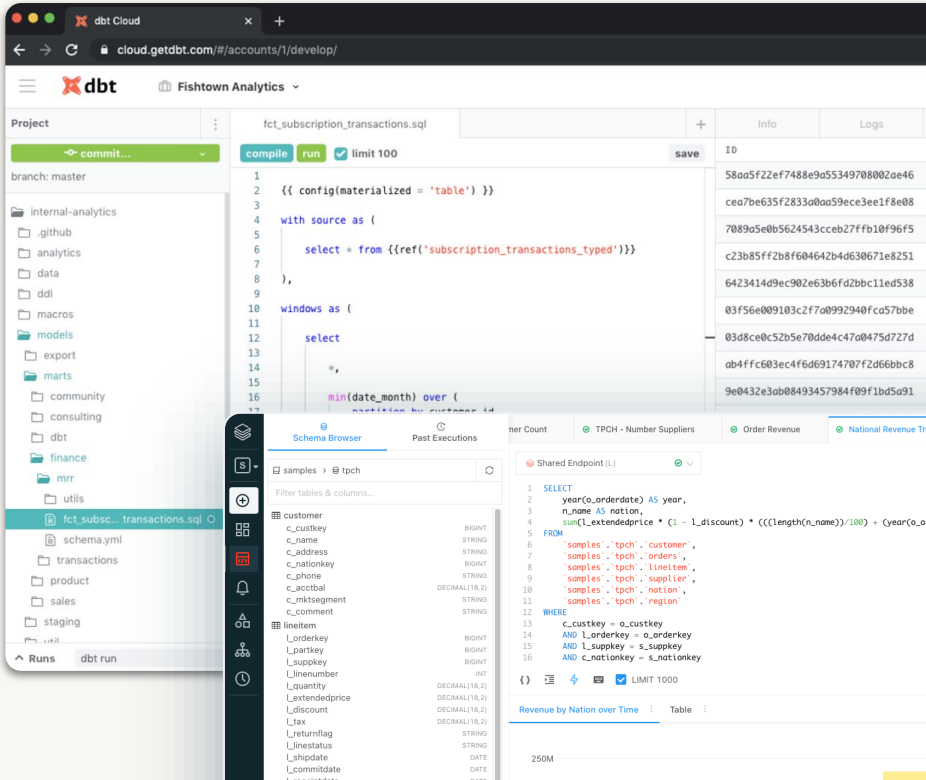
Analytics engineering on the Lakehouse made simple

Databricks and dbt Labs **simplify analytics engineering on the lakehouse.**

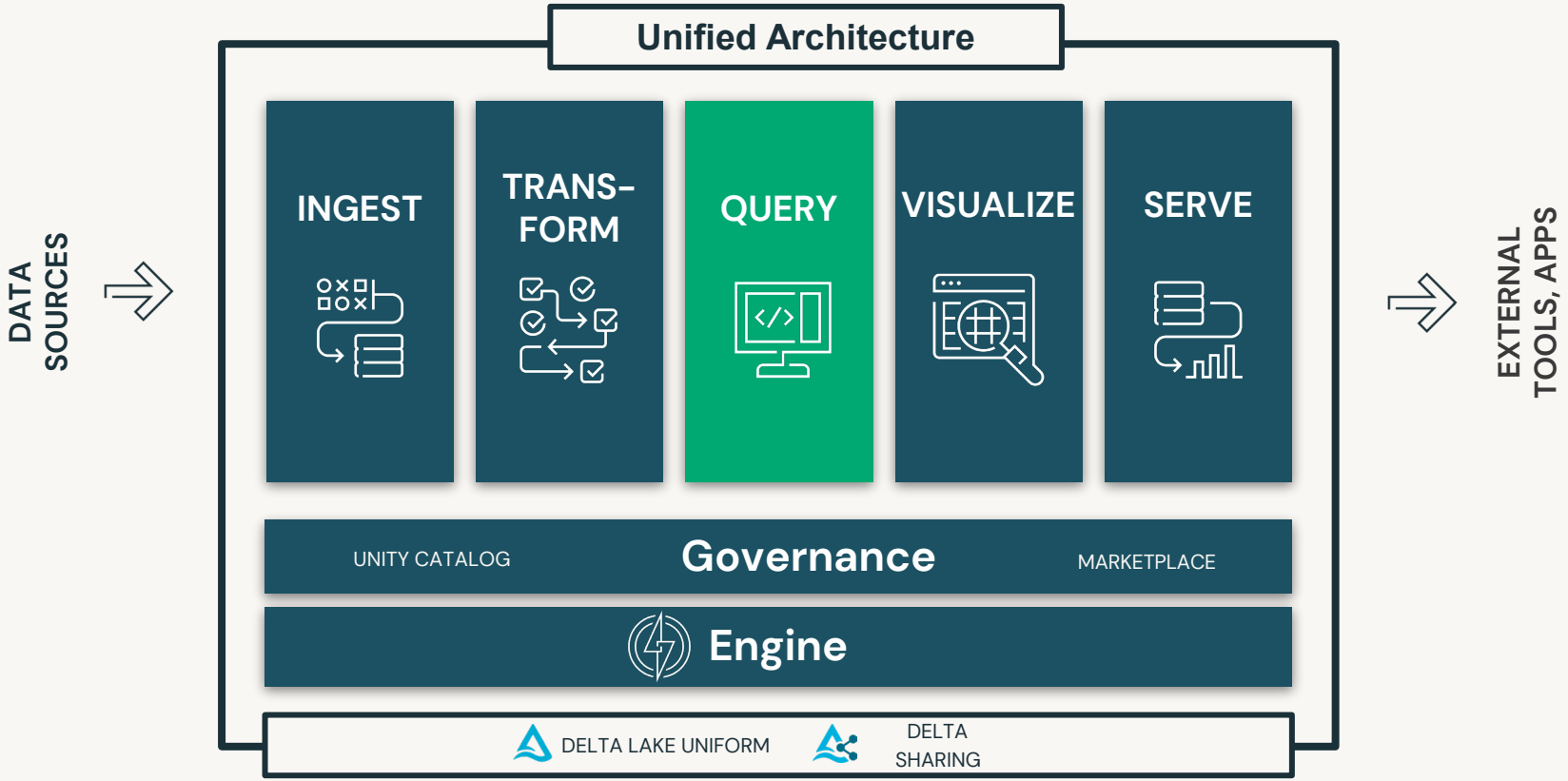
[NEW] Users can ingest and transform streaming data in their dbt pipelines **using Streaming Tables and Materialized Views**.

Run your dbt projects as a task in a Databricks job to automate tasks and schedule workflows

databricks + dbt



Simplifying the user experience end-to-end

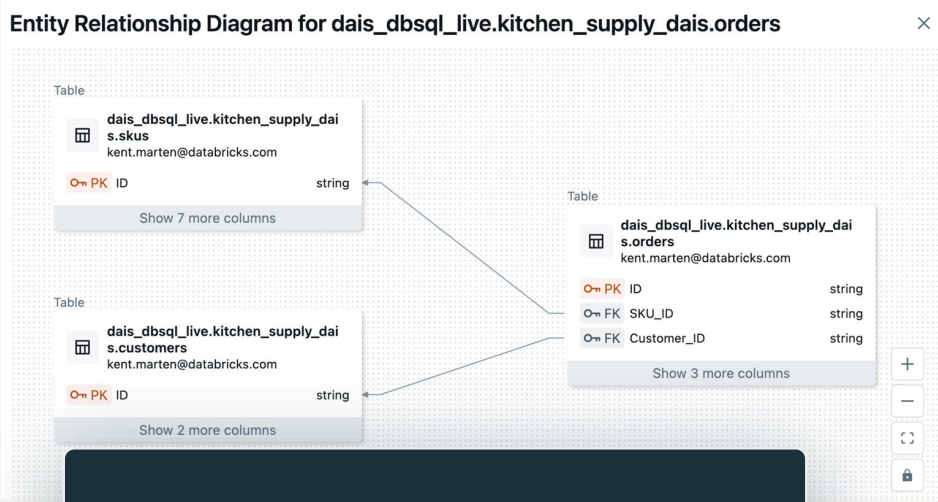


Public Preview

PK/FK + Entity-Relationship Diagram (ERD)

Improve relational data management

- Easily understand table relationships with ERD
- Relationships used by apps and tools like Tableau and PowerBI
- Leverage optimizations (RELY) to speed up queries




```
-- ADD PRIMARY KEY WITH RELY
ALTER TABLE customers ADD PRIMARY KEY
(customer_id) RELY;
```

Private Preview

SQL Scripting

More power to control your data with SQL

Newly supported statements streamline migrations to Databricks and efficiently **write complex logic** with SQL.

- Process multiple statements as a single transaction using BEGIN/END
- Build efficient business logic with looping statements (WHILE, IF/ELSE, FOR, etc.)
- Support for Exception Handling, and easy debugging

```
DECLARE count INT;
SET count = 1;

DECLARE Sum INT;
SET Sum = 0;

BEGIN
    WHILE count <= 10 DO
        INSERT INTO tab_script VALUES (count);
        SET Sum = Sum + count;
        SET count = count + 1;
    END WHILE;
END;

SELECT Sum AS sum_of_numbers;
```

Private Preview

Variant

The Open Data Type for Semi-Structured Data

Ingest JSON into an efficient and flexible format, powering **massive performance** improvements over JSON as string!

Up to **20X Performance gains**

Fully **flexible type** can handle schema changes

Open format in **Apache Spark and Delta Lake** - no proprietary vendor lock-in

```
INSERT INTO variant_tbl (event_data)
VALUES (PARSE_JSON('{"level": "warning",
"message": "invalid request",
"user_agent": "Mozilla/5.0 ..."}'));

SELECT * FROM variant_tbl
WHERE event_data:user_agent ilike
'%mozilla%';
```

Variant Data Type | Thu @ 4pm | Gene & Chenhao

Private Preview

Spatial SQL

Supercharge your geospatial analysis

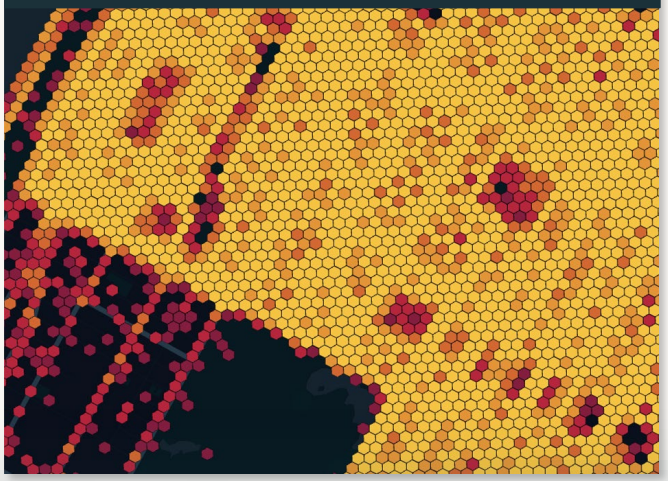
60+ Spatial Functions - broad set of ST_ expressions provide flexibility for working with Vector data

Fast Spatial Joins - efficient spatial query execution

Geometry / Geography types - read and write spatial data to native types, easily convert between WKT, WKB, GeoJson

H3-based indexing (already GA) makes it easy to see spatial patterns, combine disparate data, visualize and integrate with ML

Rideshare pick-up locations in New York City visualized in a Databricks Notebook using Kepler.gl



```
with wkt_poly as ( select
   'POLYGON((-115.42 32.57, -115.42 32.57,-115.42
      32.57, -115.42 32.57, -115.42 32.57))' as g )
select
 st_geogarea(g) as dbx_area_meters,
 st_area(g) as dbx_area_units
from wkt_poly
```

AI Functions for SQL Analysts

Use SQL functions to call AI/LLM models without needing Py/ML skills!

Built-in LLM Functions

Leverage built-in LLMs with SQL

LLM prompts + 9x functions

No setup! Uses DBX default LLM model

10x perf improvement

GA Q2

AI_QUERY()

Query any ML Model

MLFlow model or DBX LLM models or External Models e.g. Llama3

Custom models built by AI team can now be used by SQL team

GA Q2

VECTOR_SEARCH()

Query Vector DB with SQL

Perform KNN searches

Enables easy out-of-the-box RAG!

Pvt Preview

AI_FORECAST()

Metric Forecasting

Forecast business metrics without knowing any ML!

Simultaneously evaluates many models and picks the best

Pvt Preview

AI Functions for SQL Analysts

Use SQL functions to call AI/LLM models without needing Py/ML skills!

Built-in LLM Functions

Leverage built-in LLMs with SQL

LLM prompts
+9x functions

No setup! Uses DBX default LLM model

10x perf improvement

GA Q2

AI_Gen

Sentiment Analysis

Text Classification

Fix Grammar

Sensitive Data Masking

Text Similarity

Text Summarization

Language Translation

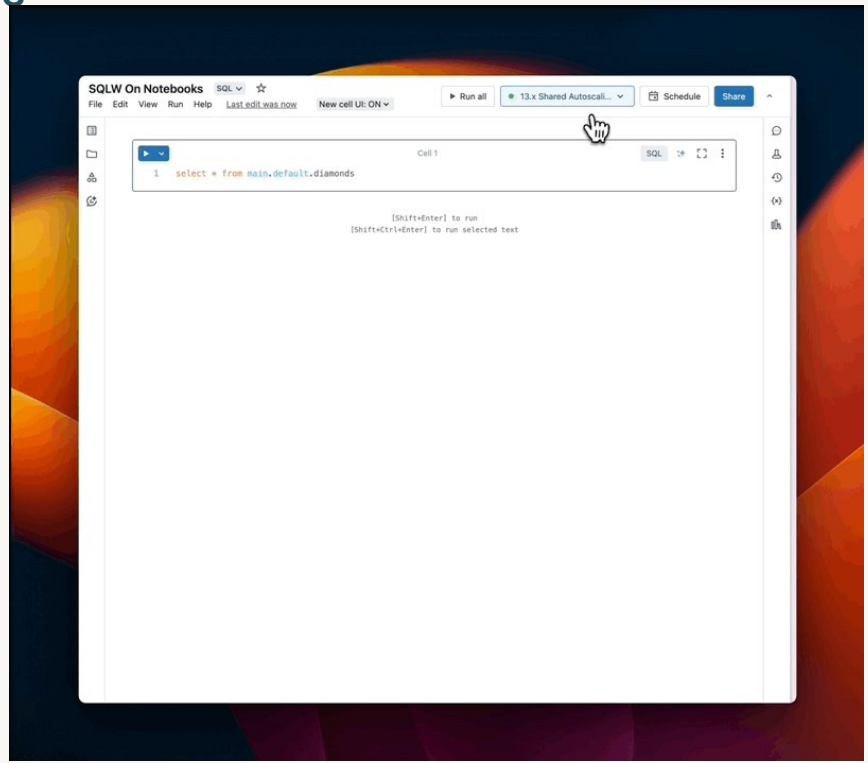
Information Extraction

GA

SQL Notebooks GA

+ Schedule SQL notebooks in Workflows

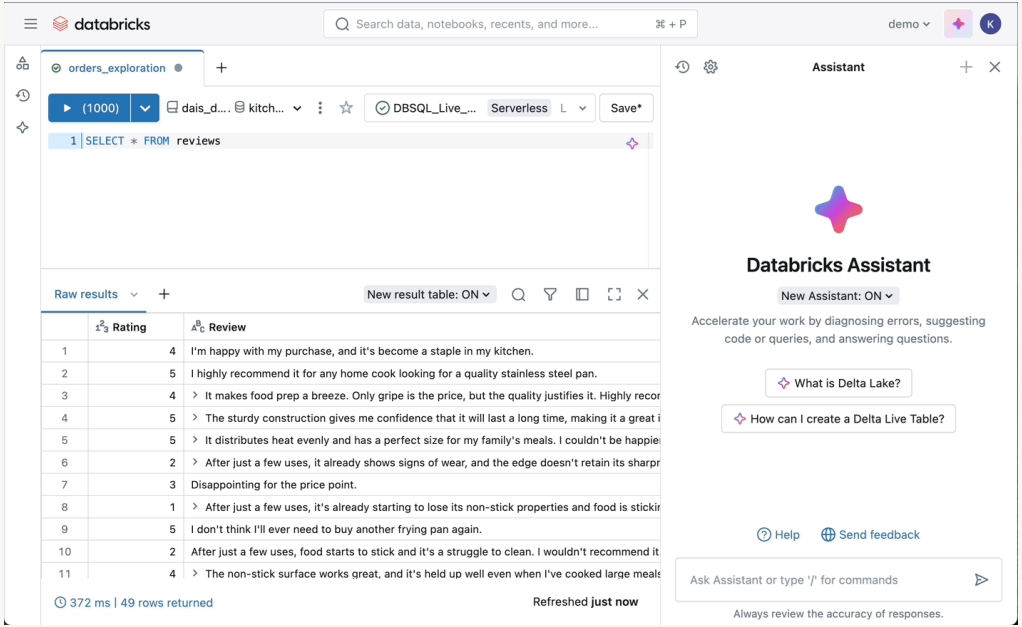
- Run multiple SQL Statements, see multiple results in Notebooks
- Native integration with Databricks Git-folders allows version control, collaboration, and CI/CD
- SQL-optimized compute provides up to 12x price performance



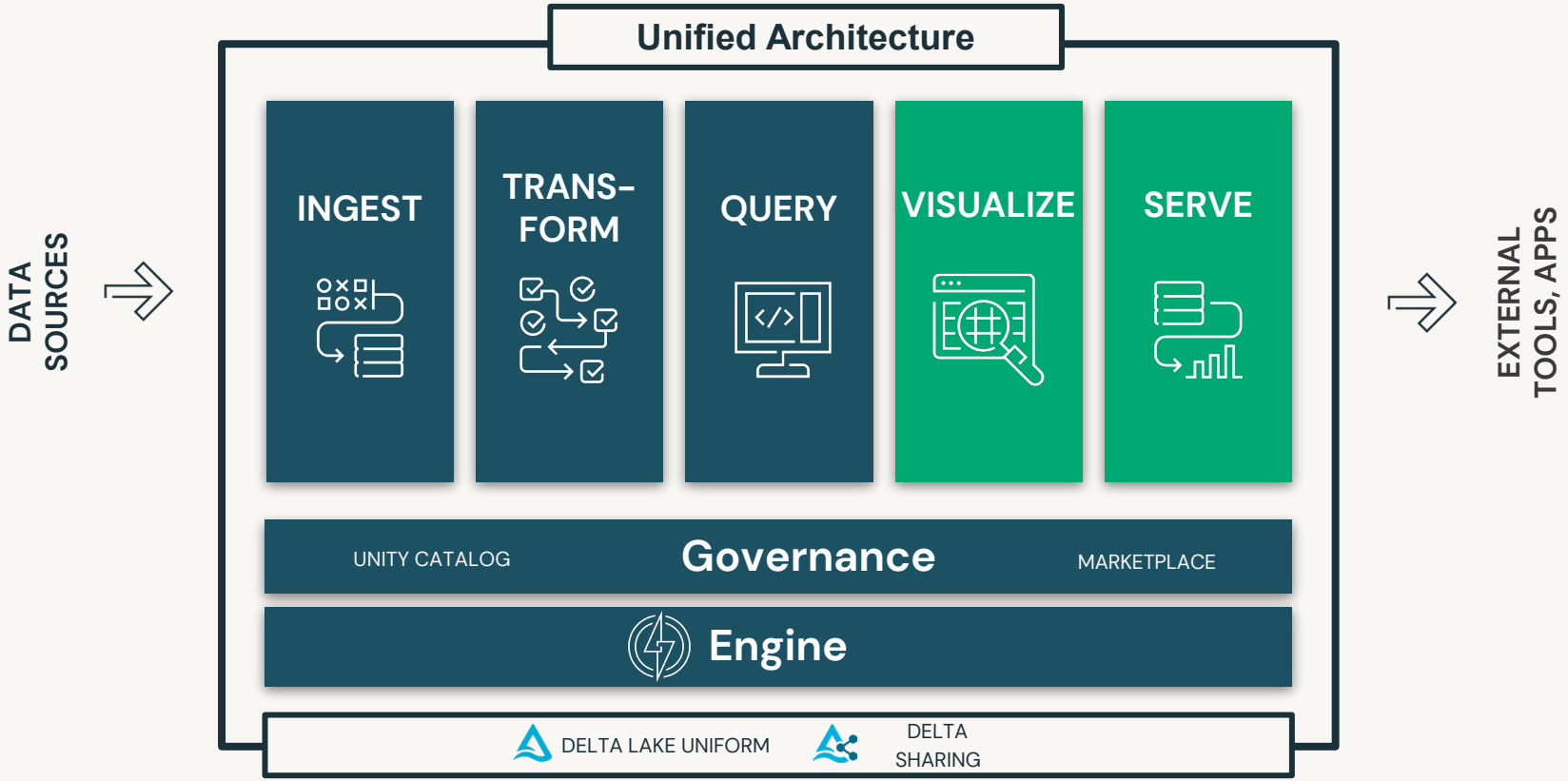
AI-Assistant in Query Editor & Notebooks

Code & debug faster with AI helping you!

- **English instructions for**
  - Automatic SQL generation
  - Code explanation
  - Error correction
- Integrated with Unity Catalog → **Knows your data, schema and context!**
- **Gets better automatically over time!**



Simplifying the user experience end-to-end



Preview

Deep Power BI & Tableau Integrations

Seamless catalog integration & data model sync

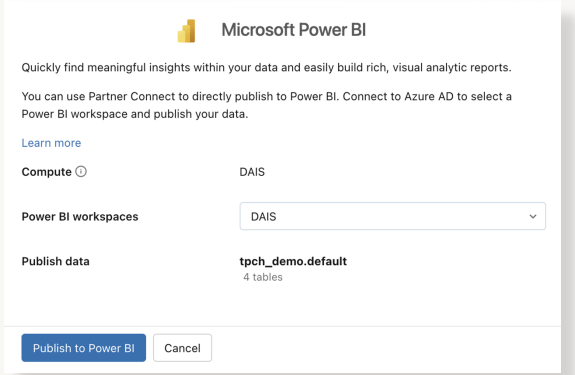
Power BI Integration

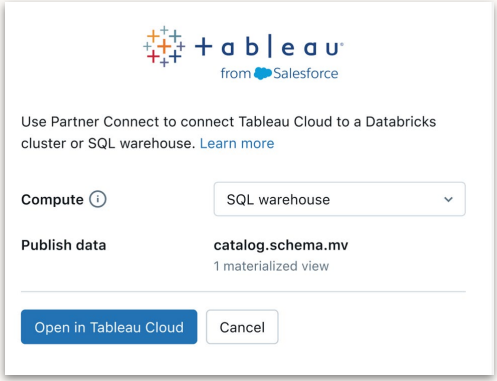
Publish UC datasets from Databricks UI, without PBI Desktop to Power BI Online.

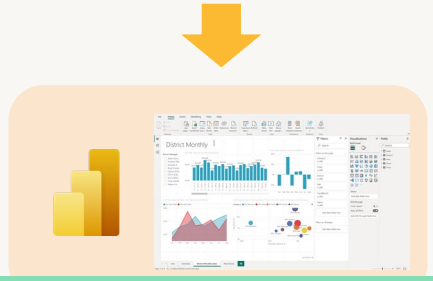
Sync entire schemas including table relationships (PK/FK) to save time.

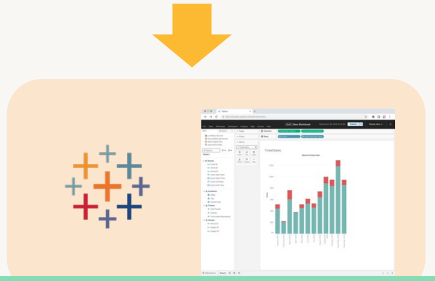
Tableau Integration

Easily explore Unity Catalog datasets in Tableau Online with a single click from Data Explorer.









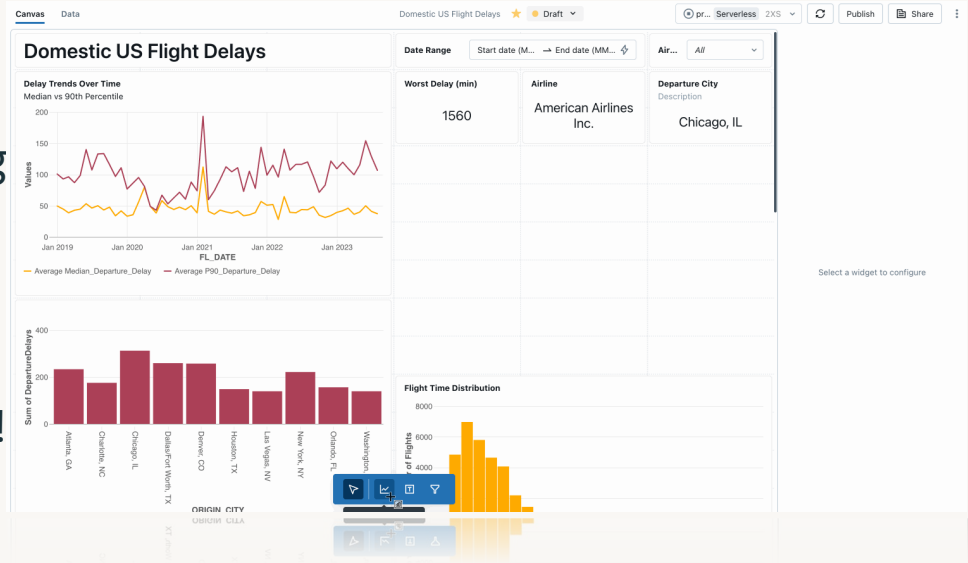
Deeper BI Integrations with Databricks SQL | Tue 4pm | Can & Jade

GA

AI/BI Dashboards

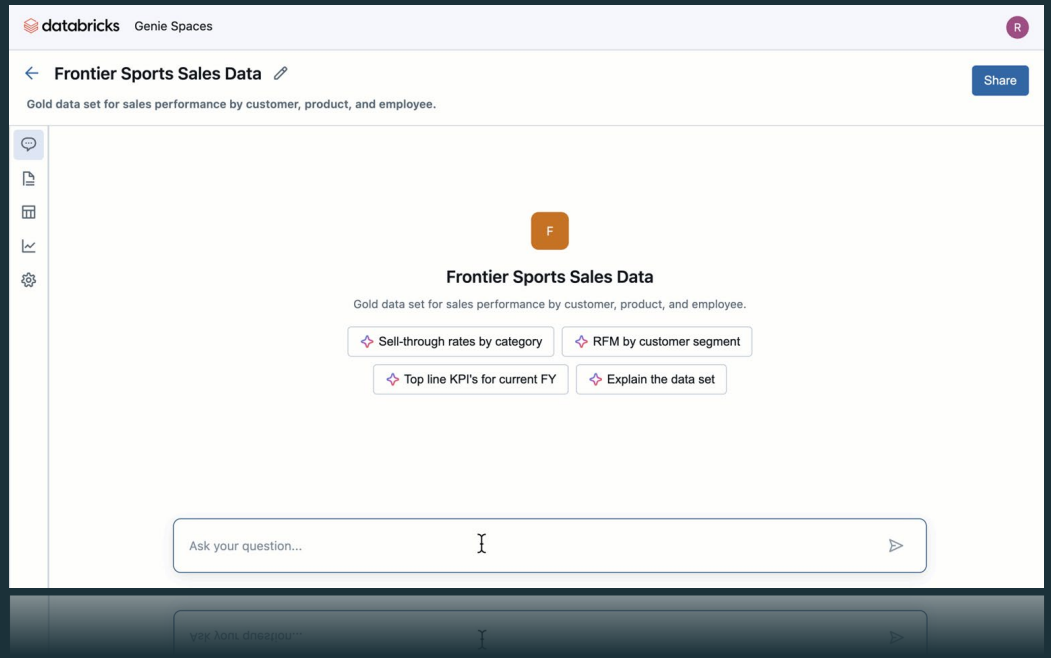
New Dashboarding experience. Now GA with more functionality!

1. Improved Performance (Caches, Filters, Scheduling)
2. Publish Externally and Share to Org
3. Assistant for English to Viz
4. UC Dataset Search & Lineage
5. Simple and Beautiful! SQL optional!



Public Preview

Introducing AI/BI Genie

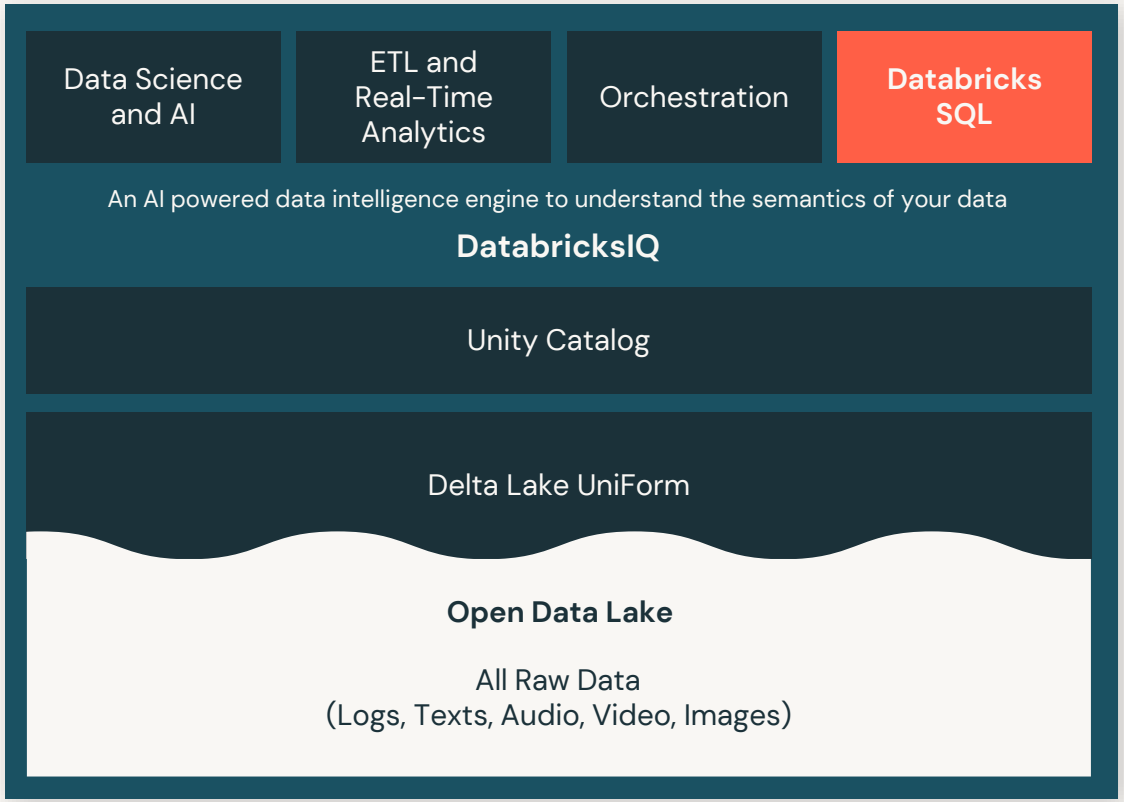


Natural Language Analytics

Direct Guidance/Control

Learns Over Time

Self-Service Data Exploration for Business Users | Wed 4pm & Thu 5pm | Hanlin & Chao



Databricks SQL
Intelligent data warehousing on the lakehouse architecture

Unified governance
for all your data + AI asset

World-class price / performance
with lowest TCO

Simplified ease-of-use
that boosts productivity for every user

DATA+AI SUMMIT