databricks

Databricks Roadmap

Q2 FY24

May 10, 2023

DATA+AI SUMMIT

World's largest data, analytics and AI conference

Explore sessions and register at
databricks.com/dataaisummit

IN-PERSON | JUNE 26-29 | SAN FRANCISCO

- Registration NOW OPEN!
- 10k people onsite
- 250+ breakout sessions
- 20+ hands-on-training sessions
- 2 keynotes featuring Databricks founders and guest speakers from DuckDB Labs, LangChain, PyTorch and more
- More meetups, parties and and fun than ever before

Hear from data and AI thought leaders about latest trends and innovations, including Apache Spark, Delta Lake, MLflow, Presto, dbt and more

Learn how others are applying the data lakehouse paradigm to unify data, analytics, and AI on one platform

Confidential & Subject to Change

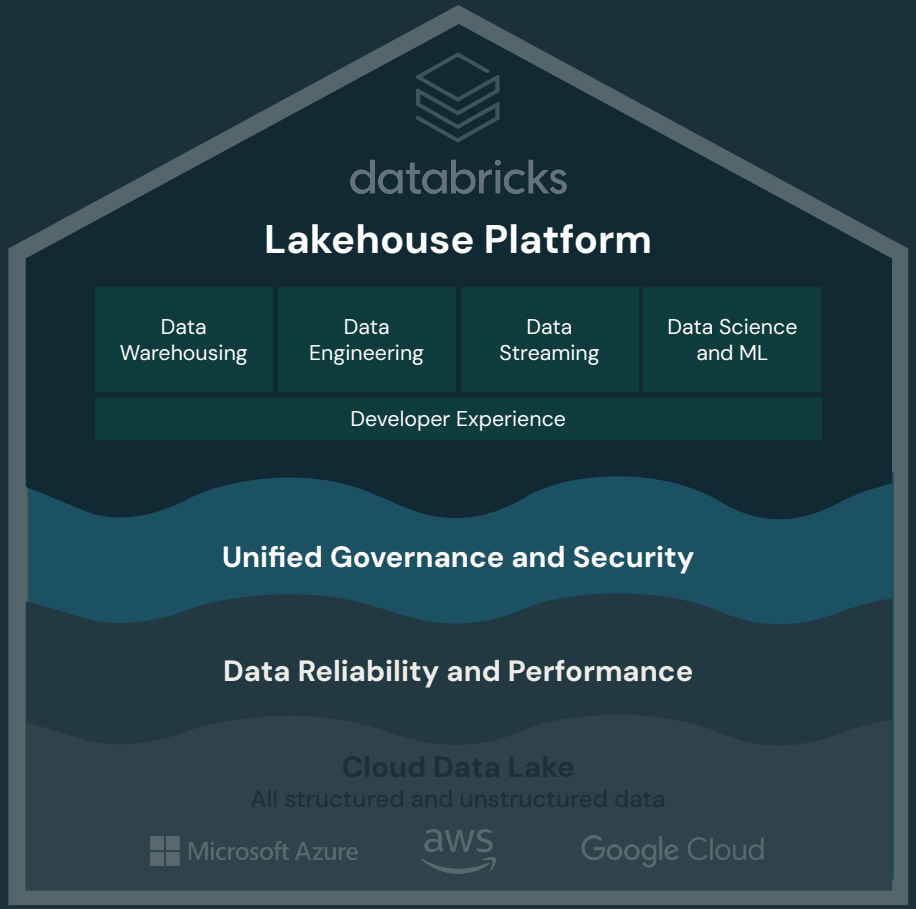
The following is intended to outline our general product direction. We constantly innovate with our customers, and the specifics may change. Do not share this information without explicit permission from Databricks.

"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"

Agenda

Databricks Lakehouse Platform—Q2 Roadmap

- Unified governance and sharing with Unity Catalog
- Enterprise security and compliance
- Orchestration with Databricks Workflows
- Data streaming with Delta Live Tables and Spark Structured Streaming
- AI & Machine Learning, including LLMs
- Data warehousing with Databricks SQL
- Developer Experience
- Delta Lake



Unified Governance Sharing, and Security

Today's Topics

Unity Catalog

Security and Compliance

Unity Catalog

Unified governance for data, analytics, and AI

Now GA on AWS/Azure/GCP

What's coming this quarter?

- Advanced Access Control with row and column filtering (Public Preview)
- Tagging/Classification (Public Preview)
- Volumes (Public Preview)
- Hive Interface to Unity Catalog (Public Preview)
- Unity Catalog Integration with Delta Live Tables (Public Preview)
- Databricks Marketplace (GA)
- Databricks Clean Room Preview (Private Preview)

Unity Catalog—row and column filtering

Use standard SQL functions to **define row filters and column masks**

Fine-grained access controls on rows and columns using SQL UDFs

Less overhead of creating views on the data for granular access controls

```
// Row filtering

CREATE FUNCTION us_filter(region STRING)
RETURNS BOOLEAN
  RETURN if(is_member('admin'), true, region='US')

ALTER TABLE sales
SET ROW FILTER us_filter ON (region)
```

```
// Column masking

CREATE FUNCTION ssn_mask(ssn STRING)
RETURNS STRING
  RETURN if(is_member('admin'), ssn, '****')

ALTER TABLE users
ALTER COLUMN ssn SET MASK ssn_mask
```

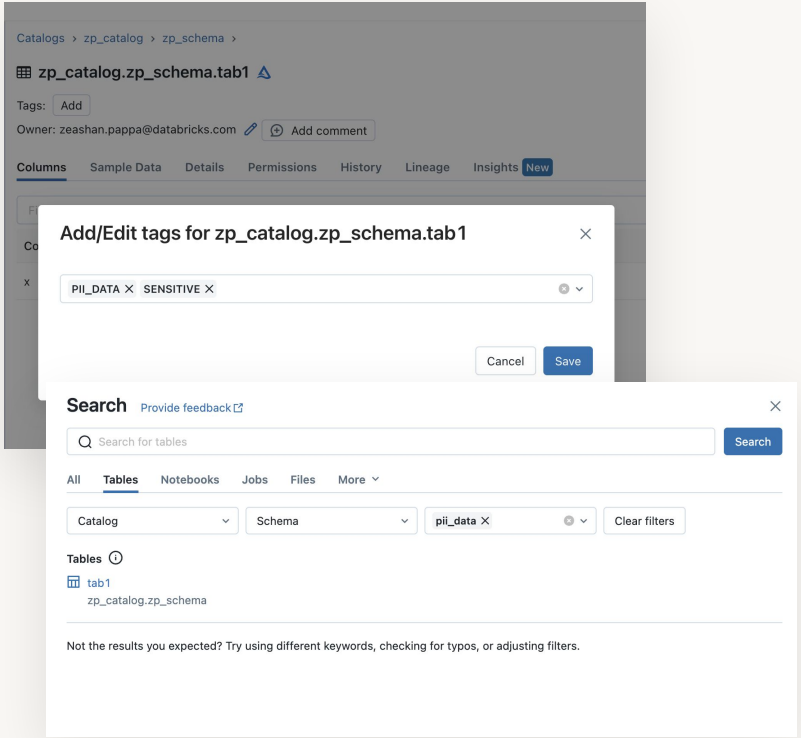
Public Preview in Q2

Unity Catalog—tagging/classification

Classify your sensitive data and simplify understanding of the underlying data

Assign tags to entities in Unity Catalog (catalogs, schemas, tables, and columns)

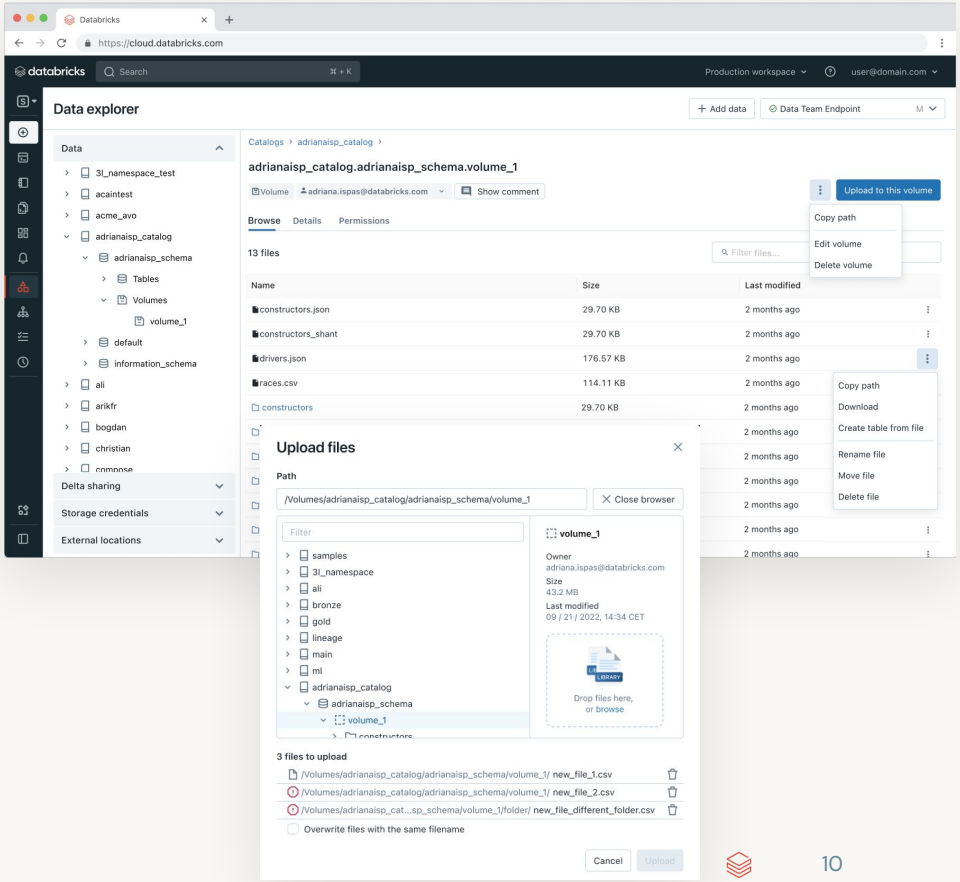
Search and filter the entities based on the tags



Public Preview in Q2

Volumes

Access, store, organize and process **non-tabular data** with Unity Catalog governance



Unlock new processing capabilities for arbitrary files, including data science and machine learning

Any file format; data can be structured, semi-structured, or unstructured

Files accessible via UI, Spark APIs, FUSE, dbutils, REST, SQL, Databricks CLI, Terraform.

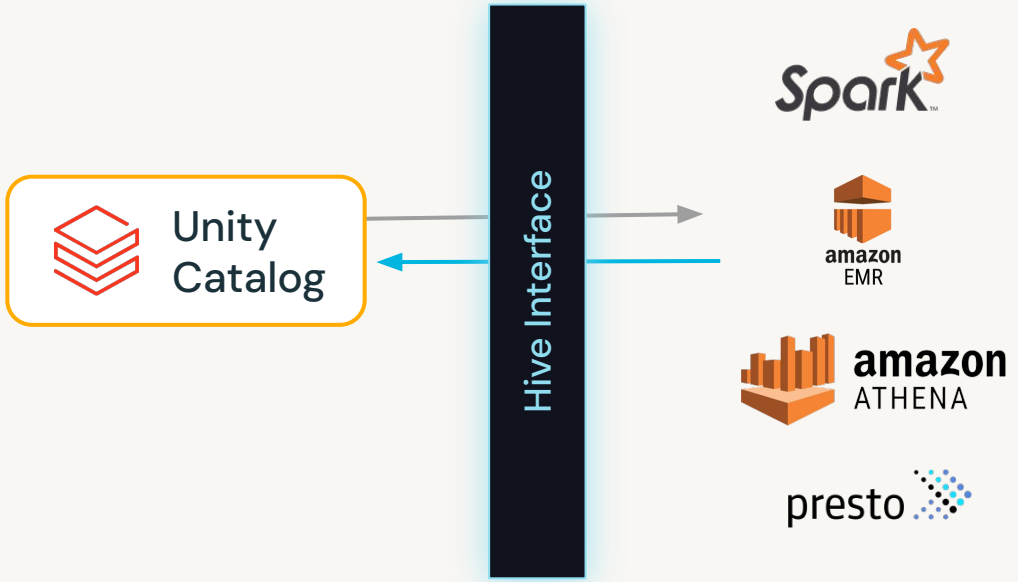
Public Preview in Q2

Unity Catalog—Hive metastore interface

Extend Unity Catalog
beyond Databricks

Single source of truth for your data, eliminate the overhead of maintaining multiple metastores

Allowing external connectivity to Unity Catalog from EMR, Athena and other tools



Public Preview in Q2

Databricks Marketplace

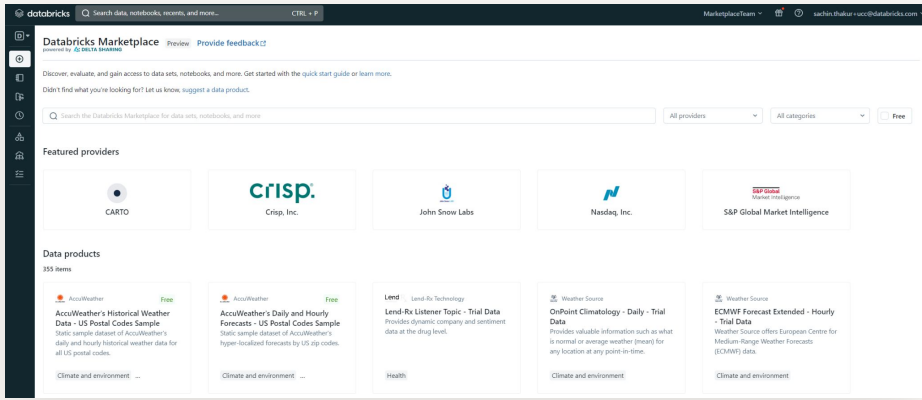
Open marketplace for data, analytics, and AI

Share complete solutions:
not just data, but notebooks, models, applications, dashboards and more

Public marketplace with top data & solution providers

Private marketplaces with same UI

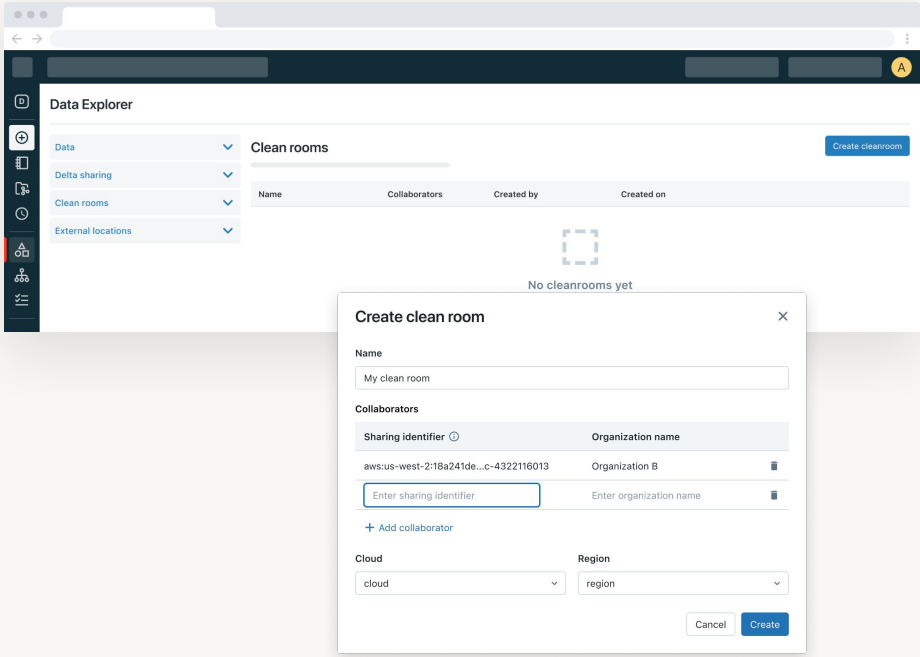
Notebook, stream, model and app sharing



GA in Q2

Databricks Clean Room

Confidential, multi-party computation on existing lakehouse data



Run any computation in Python, SQL, R, Java, etc.

No data replication of existing lakehouse tables with Delta Sharing

Scale to multiple collaborators and any data size

Private Preview in Q2

Enterprise security

Your data security is our priority

Build on a secure and trusted platform

Compute, network and workload security controls with best practices guidance

- Private connectivity and network security
- Enhanced Security Monitoring

Meet regulatory requirements

A broad set of compliance controls for regulated and sensitive workloads

- HIPAA
- PCI-DSS
- FedRAMP Moderate

Protect and control your data

Secure data with your encryption key, get granular access control and audit logs, and govern your data with Unity Catalog

- Customer Managed Keys
- Unity Catalog

What's coming this quarter?

Azure Databricks support on Azure confidential computing (ACC) (GA)

Enhanced Security and Compliance on Azure Databricks (GA)

Serverless on AWS certified on HIPAA, PCI and FedRAMP Moderate (GA)

Serverless on Azure certified on HIPAA (GA)

IP Access Control Lists (ACLs) for Account Console on AWS and GCP (GA)

Customer Managed Keys on GCP (Public Preview)

Private Link support for Serverless SQL on Azure Databricks (Private Preview)

Azure Storage firewall for DBFS (Private Preview)

Azure Databricks support for Azure confidential computing (ACC)

Enable data, analytics, and AI use cases for confidential data

Build your Data and AI strategy for sensitive datasets with increased security from confidential computing

Enhance data confidentiality with encryption in memory through keys secured at the hardware level

Complement your encryption at rest with customer-managed keys

Get up and running quickly by leveraging your existing Spark and Databricks workloads

GA in Q2

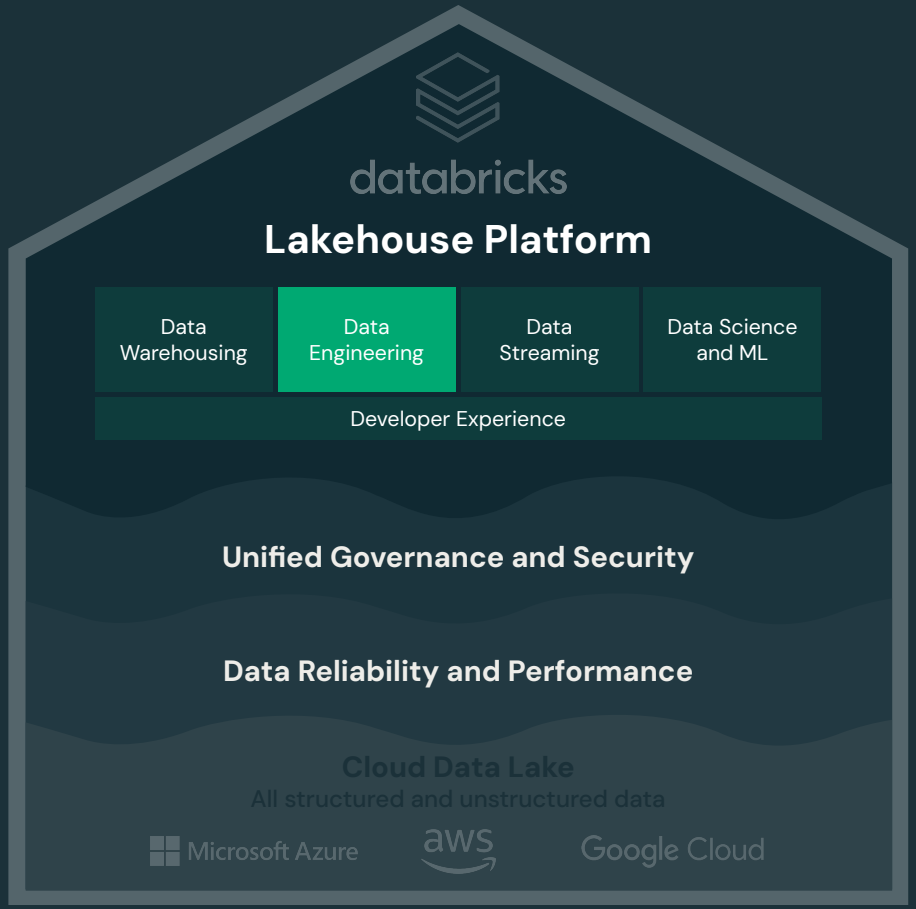
Enhanced Security and Compliance expanded availability

Simplify the complexity of meeting security and regulatory requirements

Get increased visibility, threat protection and security hardening for your workloads with Enhanced Security Monitoring

Run cloud-ready HIPAA, PCI-DSS and FedRAMP Moderate workloads with the Compliance Security Profile

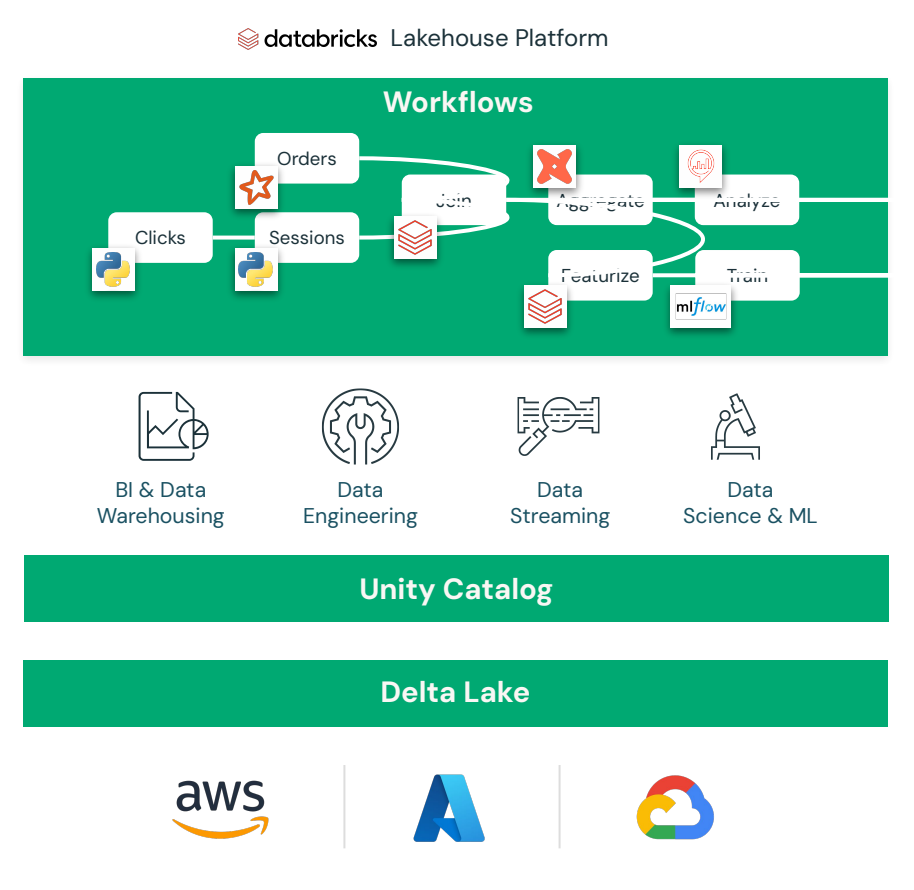
GA in Q2 | Azure | Serverless on AWS



Data Engineering

Today's Topics

Orchestration with Databricks Workflows



Databricks Workflows

Unified orchestration for data, analytics, and AI on the Lakehouse Platform

Simple authoring

Actionable insights

Proven reliability

What's coming this quarter?

- Enhanced Control Flow (GA)
- Run a Workflow as a Task (GA)
- Visual Monitoring (Public Preview)
- Job-level Parameters (Public Preview)
- Late Job Notifications (GA)
- Webhook Alerts (GA)
- Job Queueing (GA)

Enhanced control flows

Adding enhanced control flow logic to workflows

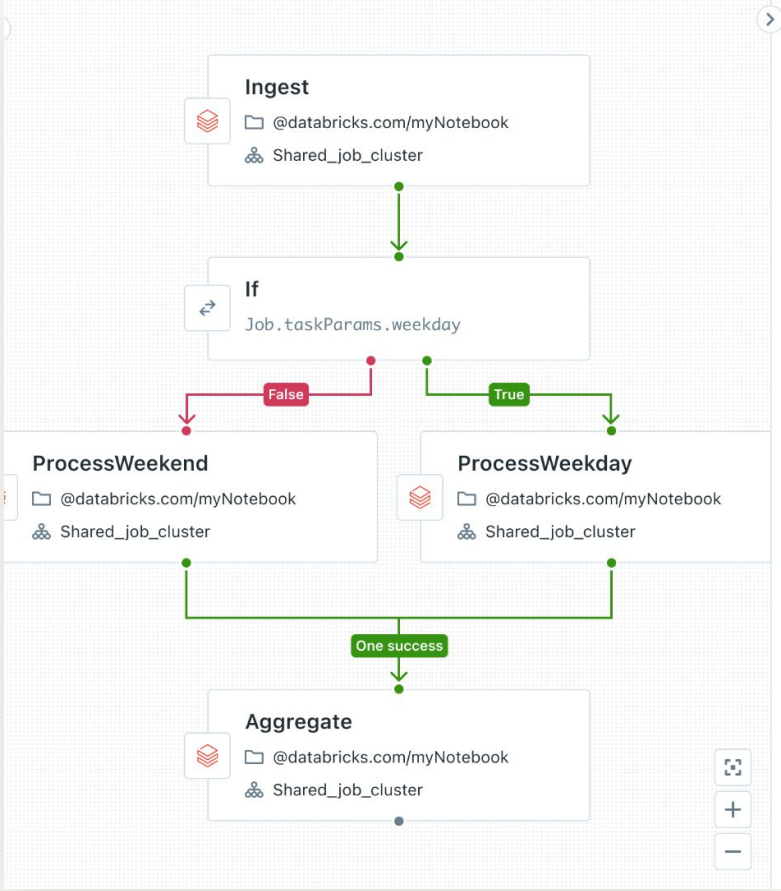
Define sophisticated multi-task workflow logic that executes different tasks according to runtime conditions

Incorporate error handling in workflow definitions

Exclude unnecessary tasks during runtime for higher efficiency

Full visibility to what tasks executed in every run

GA in Q2



Workflows as tasks

Run a workflow as a task in another workflow

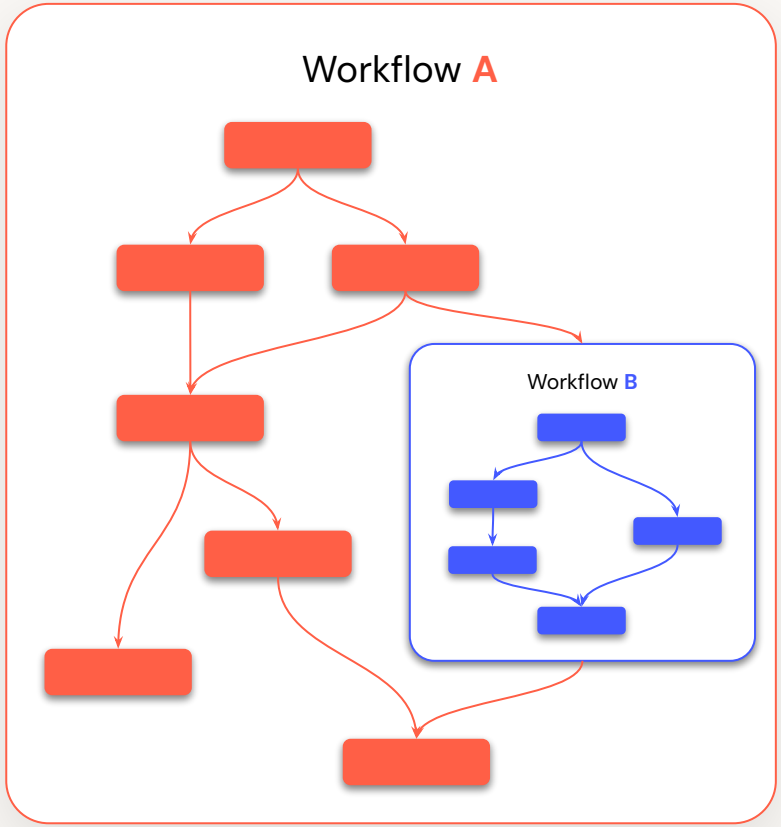
Break down complex DAGs into smaller, repeatable units of execution that can be triggered as tasks

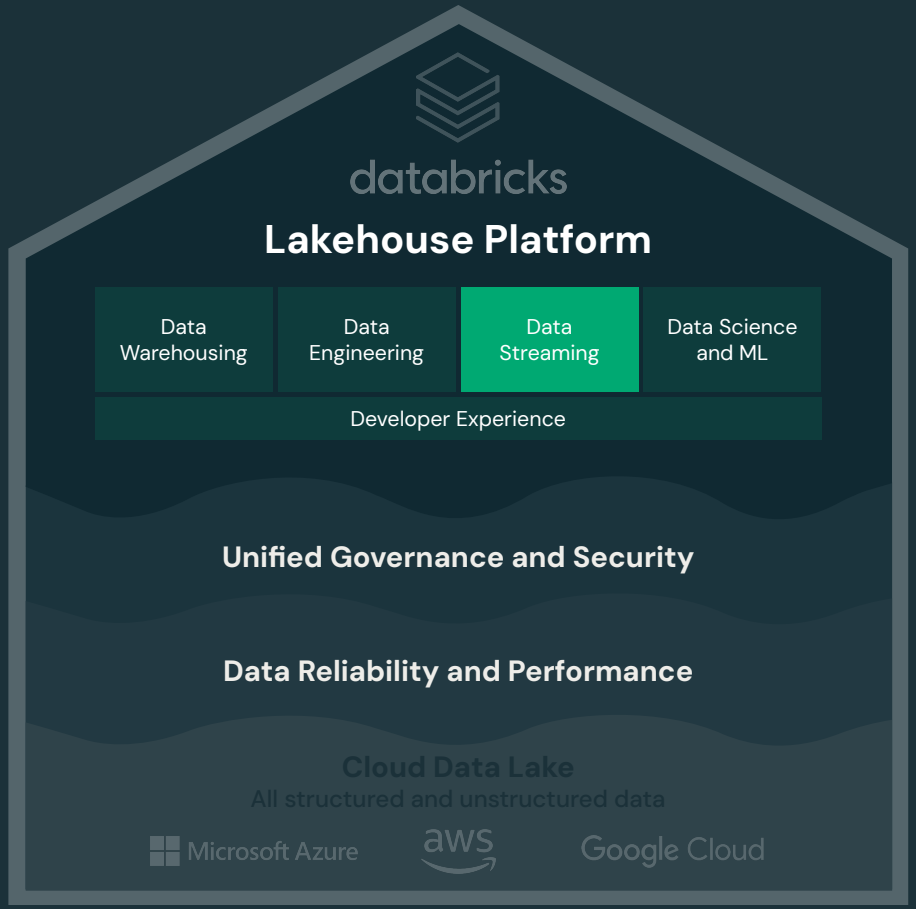
Define tasks that trigger other workflows

Simplify authoring, testing and troubleshooting

Reuse workflow definitions to save time

GA in Q2

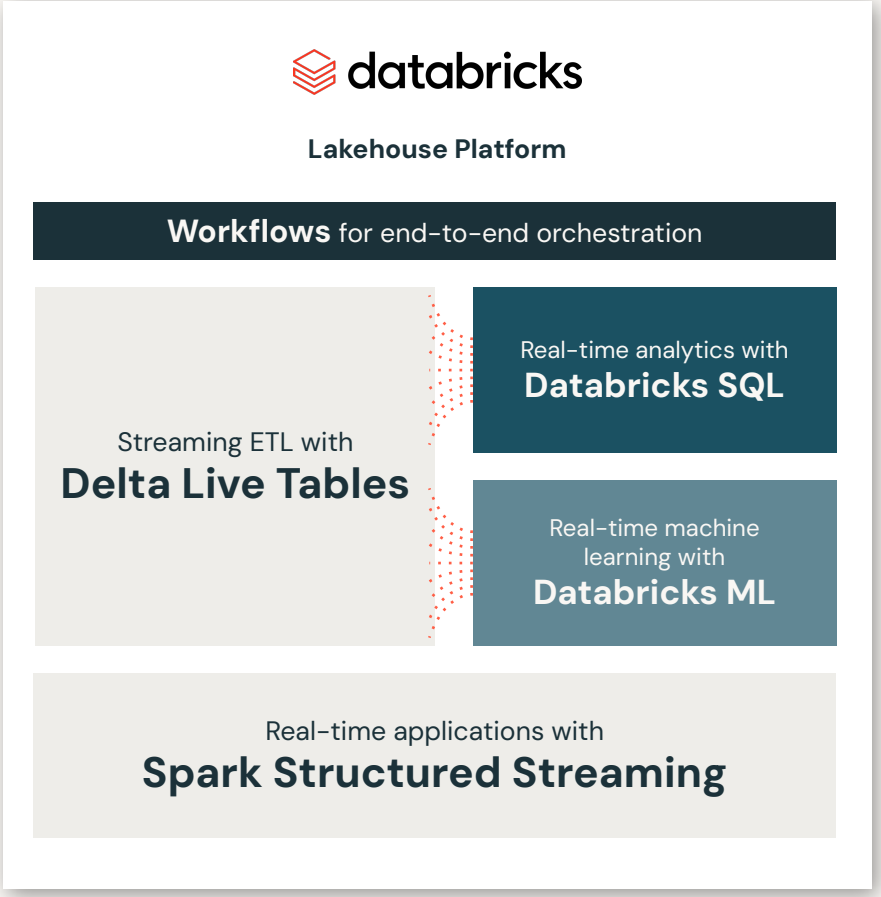




Data Streaming

Today's Topics

- Streaming ingestion and transformation with Delta Live Tables
- Real-time applications with Spark Structured Streaming



Data streaming

Data streaming made simple

Simplified development

Simplified operations

Unified governance

What's coming this quarter?

Delta Live Tables for GCP (GA)

Append Flows API for DLT (Private Preview)

IAM Auth support for Amazon MSK Connector

Serverless Compute for Delta Live Tables (Private Preview)

Custom Schemas for DLT (Private Preview)

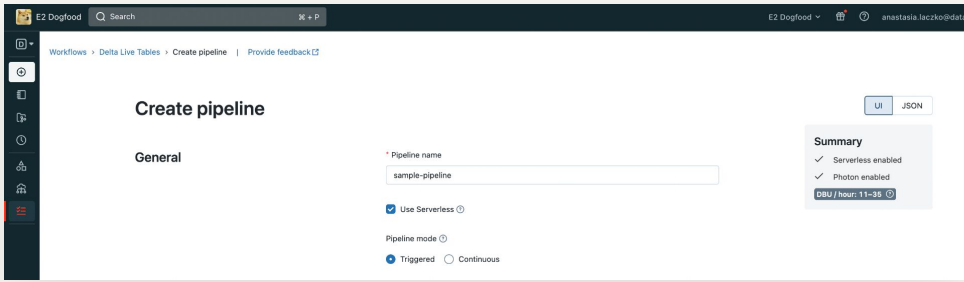
Unity Catalog Integration with Delta Live Tables (Public Preview)

Apache Pulsar Connector for Structured Streaming

Serverless compute for Delta Live Tables

Simplified, fully managed compute

Cheaper and simpler data pipelines



Auto optimized compute that only runs when needed

Reliable, fully managed compute resources

Simple configuration for any use case

Private Preview in Q2

Delta Live Tables support for Unity Catalog

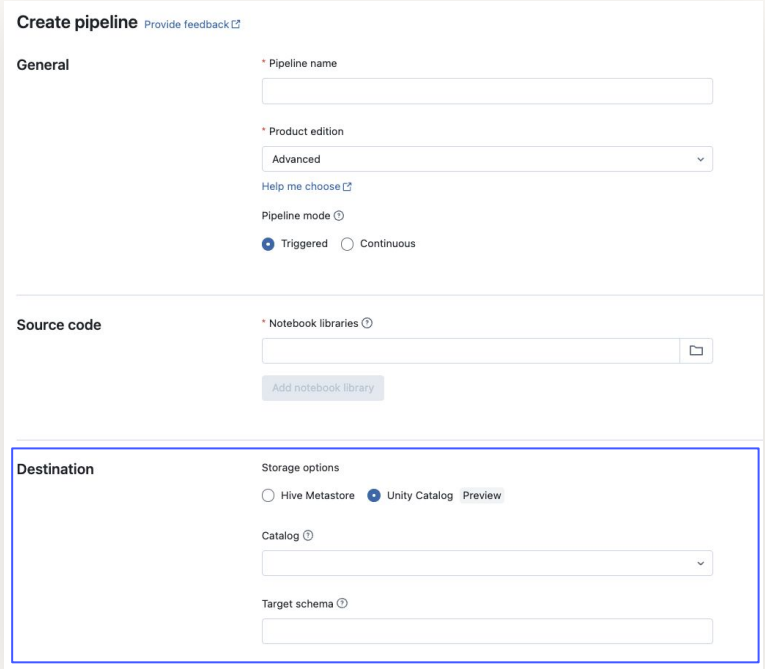
Unified governance for DLT pipelines

Delta Live Tables may now benefit from UC management and governance features

Publish DLT tables to Unity Catalog

Read from Unity Catalog managed tables

Define and manage ACLs on Live Tables and Streaming Live Tables



Public Preview in Q2

Append Flows API

Append data to existing streaming tables

Seamlessly evolve sources and perform backfills without requiring a full refresh

Add and remove data sources without a full refresh

Backfill (append only) data in streaming tables

Works with any streaming source supported by DBR

```python
import dlt

@dlt.append_flow(target = "tgt")
def f1():
return spark.readStream.load(src1)

@dlt.append_flow(target = "tgt")
def f2():
return spark.readStream.load(src2)
```

Private Preview in Q2

Custom schemas

Flexible schemas for DLT tables

Organize DLT tables into different schemas based on their contents

Maintains environment independence for CI/CD

Change schema locations anytime

Define schemas entirely in source code

```
/* Create in a bronze schema */
CREATE OR REFRESH STREAMING TABLE bronze.table1 AS
SELECT * FROM STREAM read_files('s3://mybucket/analysis')

/* Create in a silver schema */
CREATE or REFRESH MATERIALIZED VIEW silver.table2 AS
SELECT count(foo) from LIVE.bronze.table1;
```

Private Preview in Q2

Custom sinks

Persist data to Kafka or other targets natively

Achieve low-latency for operational workloads

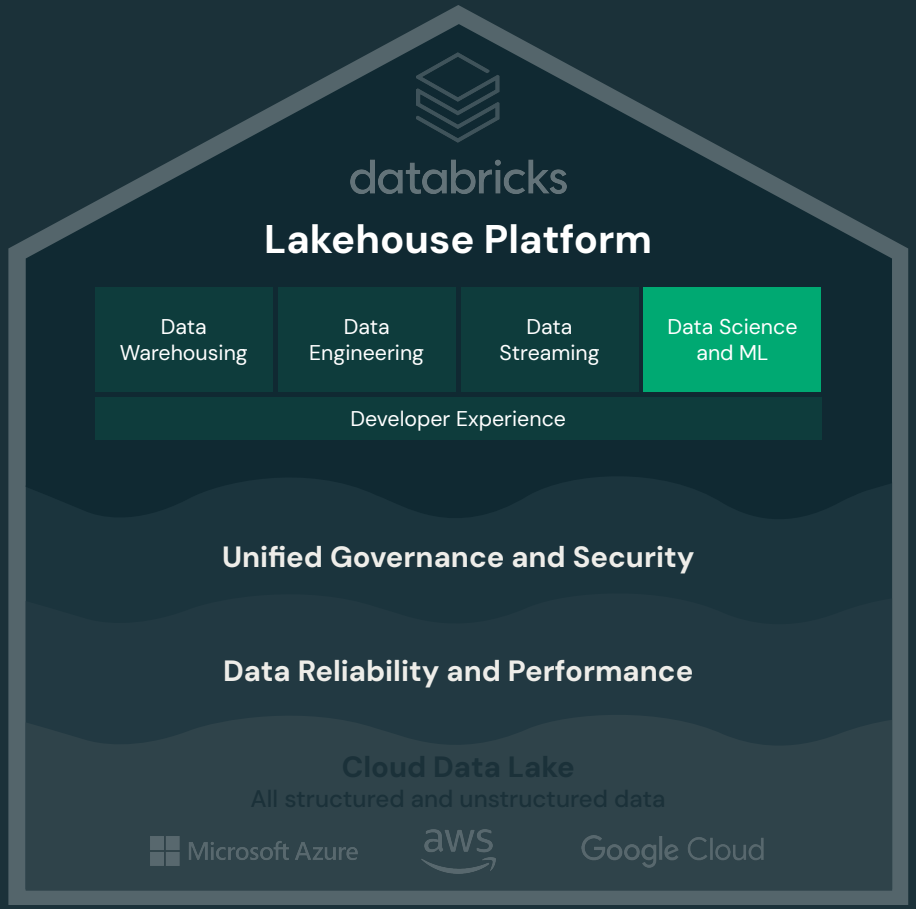
Supports low-latency operational use cases

Output data from DLT using any of the predefined Structured Streaming sinks

Granular observability

```
CREATE SINK mySinkName
FORMAT kafka
[FORMAT_OPTIONS (
  `kafka.opt` = 'foo',
  `kafka.flag` = 'true',
  `kafka.secret` = secret('myScope', 'mySecret')
)]
AS ...
```

Private Preview in Q2



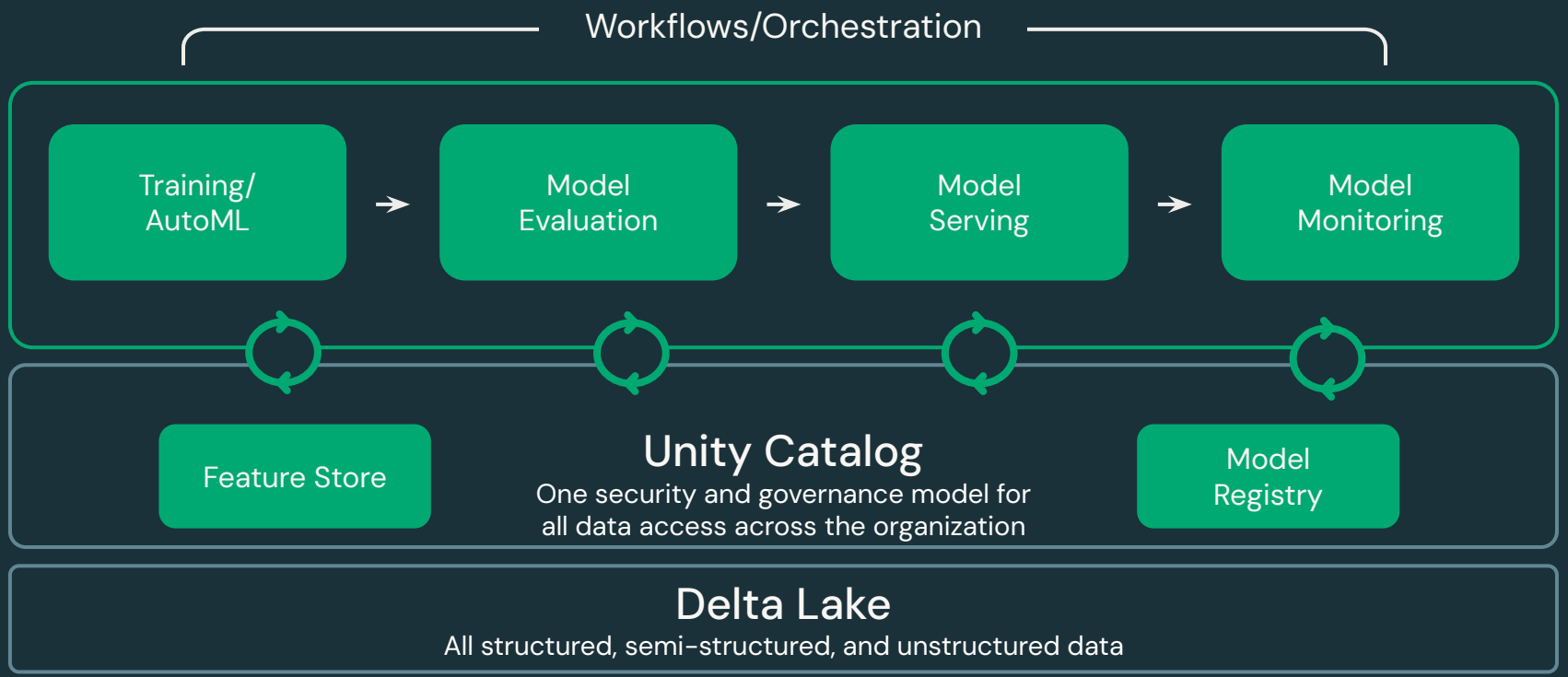
Data Science & ML

Today's Topics

Large Language Models

Model Serving

Accelerate Machine Learning with a unified data and ML platform



What's coming in ML this quarter?

- Large Language Models (LLM)— AI Functions
- Model Serving—GPU Support
- LLM—MLflow Upcoming Improvements
- Model Serving—Payload Logging
- LLM— Spark Support for Hugging Face
- Feature Store—DLT Integration

Large Language Models (LLM)— AI functions

Easy path for analysts to use OpenAI within DB SQL to integrate LLM services in their queries

Simple functions replace complex API calls

Native in Pro and Serverless SQL Warehouse

```sql
CREATE
OR REPLACE FUNCTION summarize(text STRING) RETURNS STRING RETURN llm_generate(
  concat('Summarize this to 1 sentence: ', text),
  'openai/gpt-3.5-turbo',
  'apiToken',
  secret('username', 'openai_api_token'),
  'temperature',
  0.0
);
SELECT
  product_name,
  summarize(long_product_description) AS product_summary
FROM
  products;
```

Gated Public Preview in Q2 (on AWS)

LLM—MLflow upcoming improvements

MLflow increases support for popular LLM frameworks such as Hugging Face Transformers, LangChain, and OpenAI

Customize a model on your data for your specific task and use it easily

MLflow support drastically improves lakehouse integration

```
import transformers
import mlflow

summary_pipeline = transformers.pipeline(model="databricks/dolly-v2-12b")

with mlflow.start_run() as run:
    mlflow.transformers.log_model(
        transformers_model=summary_pipeline,
        artifact_path="your_path",
        input_example="Some kind of long form text as example")
```

Preview in Q2 (on AWS)

LLM—Spark support for Hugging Face

Easily load Spark data into Hugging Face for **up to 40% speed improvements for AI model training**

Use your Spark data and start fine tuning your LLMs

Expanded support for Spark Streaming coming as well

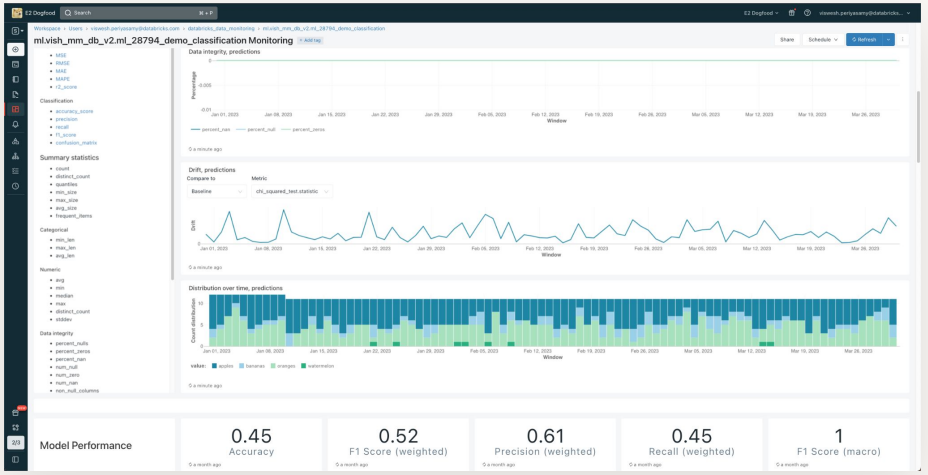
GA in Q2

Model Serving—payload logging

Automatically track model serving requests and responses to delta tables in the user's account

Built-in dashboard and notifications

Join the inference logs with other data for deeper analysis



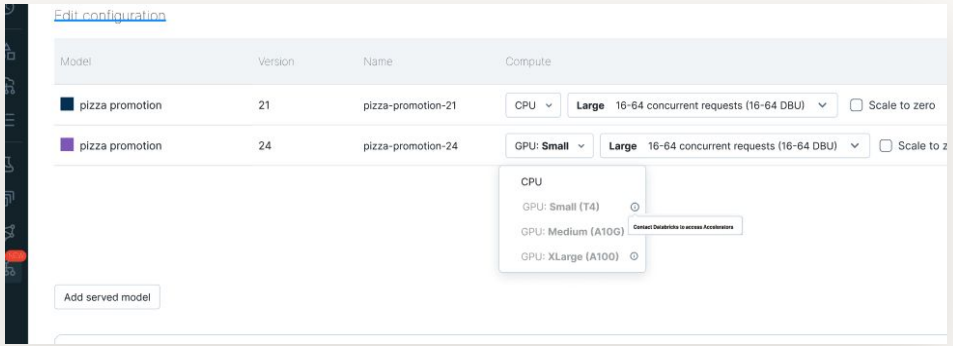
Private Preview in Q2

Model Serving—GPU support

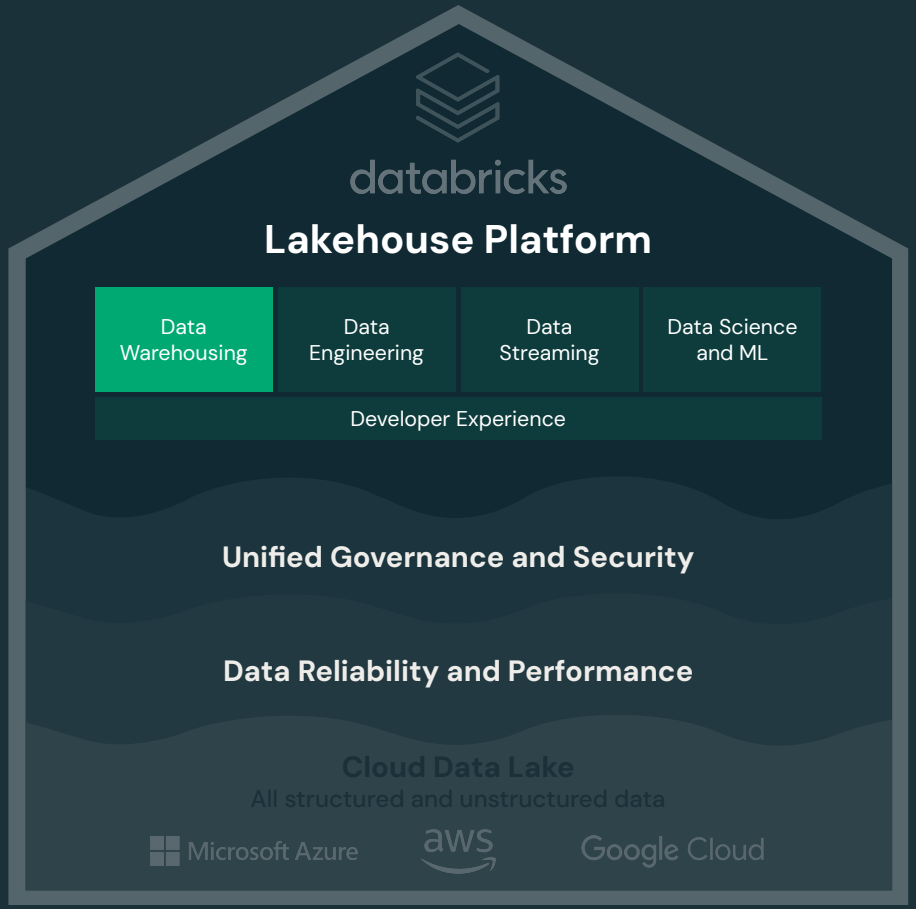
Fully managed serving endpoint with GPU support to enable inference of large models, including LLMs

Simplified, easy single-click experience to add GPUs

Real-time inference with auto-scaling, all the way to zero



Gated Public Preview in Q2 (on AWS)



Data Warehousing

Today's Topics

Databricks SQL

Data Warehousing on the Lakehouse

Powered by Databricks SQL

DB SQL is a serverless data warehouse on the Databricks Lakehouse Platform that offers high scalability, up to 12x better price/performance, unified governance, open formats, APIs, and tool flexibility with no lock-in.

Best price/performance

Built-in governance

Rich ecosystem

Unified SQL experience

	AWS	Azure	GCP
Databricks SQL Classic	GA	GA	GA
Databricks SQL Pro	GA	GA	GA Q2
Databricks SQL Serverless	GA Coming Soon	GA Coming Soon	Coming Soon

What's coming this quarter?

- Databricks Serverless (GA)
- Intelligent Autoscaling & Adaptive Routing (GA)
- Materialized Views (Public Preview)
- SQL Execution API (GA)
- Account Level Automation with Tokens (Public Preview)
- Python UDFs (Public Preview)
- Databricks SQL on Azure China (GA)
- dbt: OAuth, better incremental models & performance
- Databricks SQL System Tables (Private Preview)
- Enterprise Scale User Support Increase 10k limit/workspace or account (Private Preview)

Databricks SQL Serverless

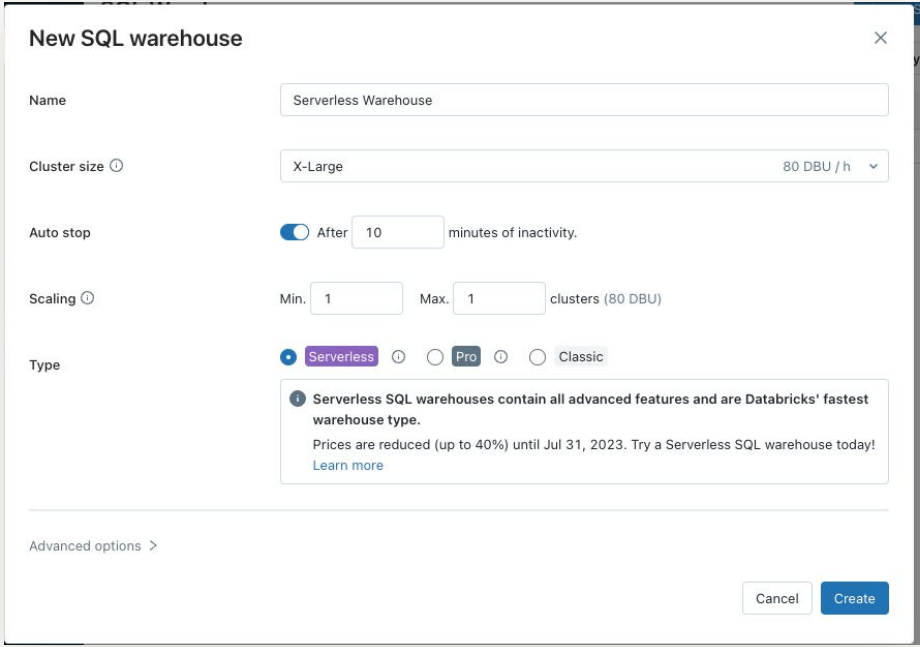
Instant, scalable compute for all DW/BI workloads

Get best performance, lower costs, and focus on delivering value rather than managing infrastructure.

Instant, elastic compute decoupled from storage

Eliminate management overhead

Lower TCO with AI powered optimizations



GA Q2 on AWS & Azure

Adaptive routing

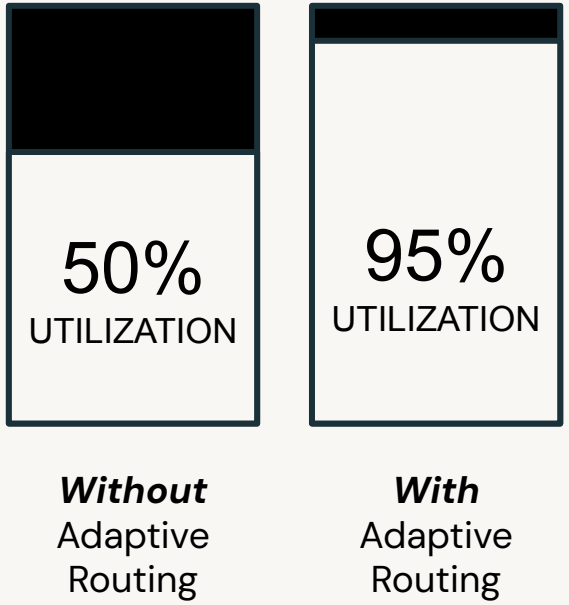
Efficient compute utilization for faster results

Deliver faster response times at a lower cost by optimizing warehouse utilization

AI-driven intelligent query allocation prevents cluster over-subscription, keeping queries fast; meanwhile, ensuring maximum utilization of the underlying hardware

Lower latency and queuing without the need to upscale, resulting in cost savings

GA in Q2 on AWS & Azure



Intelligent Autoscaling

Enabling BI users to stay in the flow

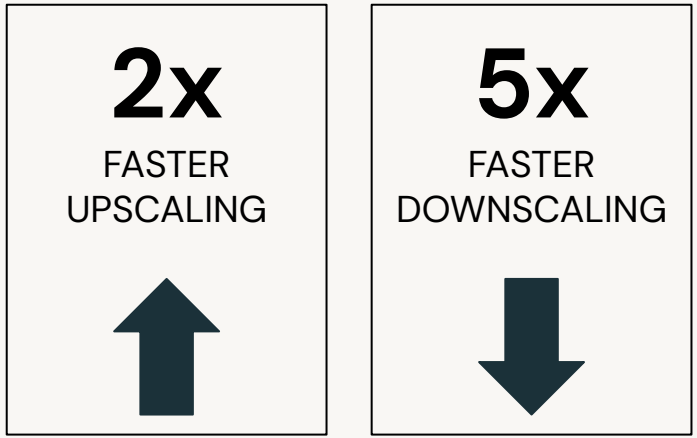
Deliver faster response times at a lower cost

2x faster upscale time and 5x faster downscale time compared to our 2022 baseline algorithm

Improved flexibility and responsiveness keep users engaged in their data analysis tasks

Dynamic resource allocation for optimal efficiency and cost-effectiveness

GA in Q2 on AWS & Azure



Materialized views

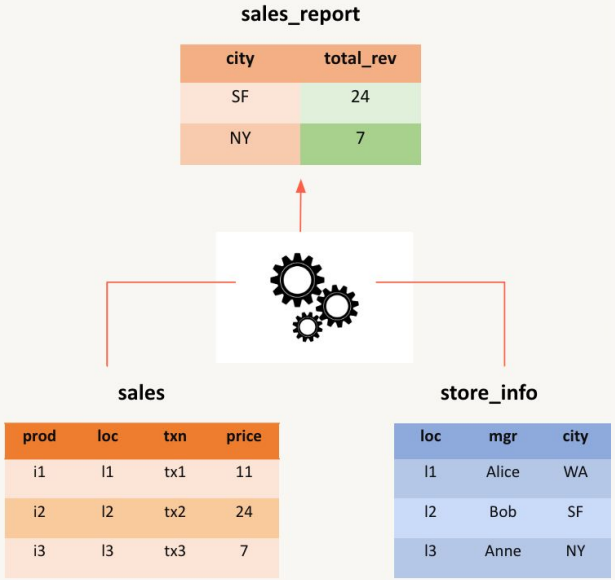
Speed up queries with pre-computed results

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

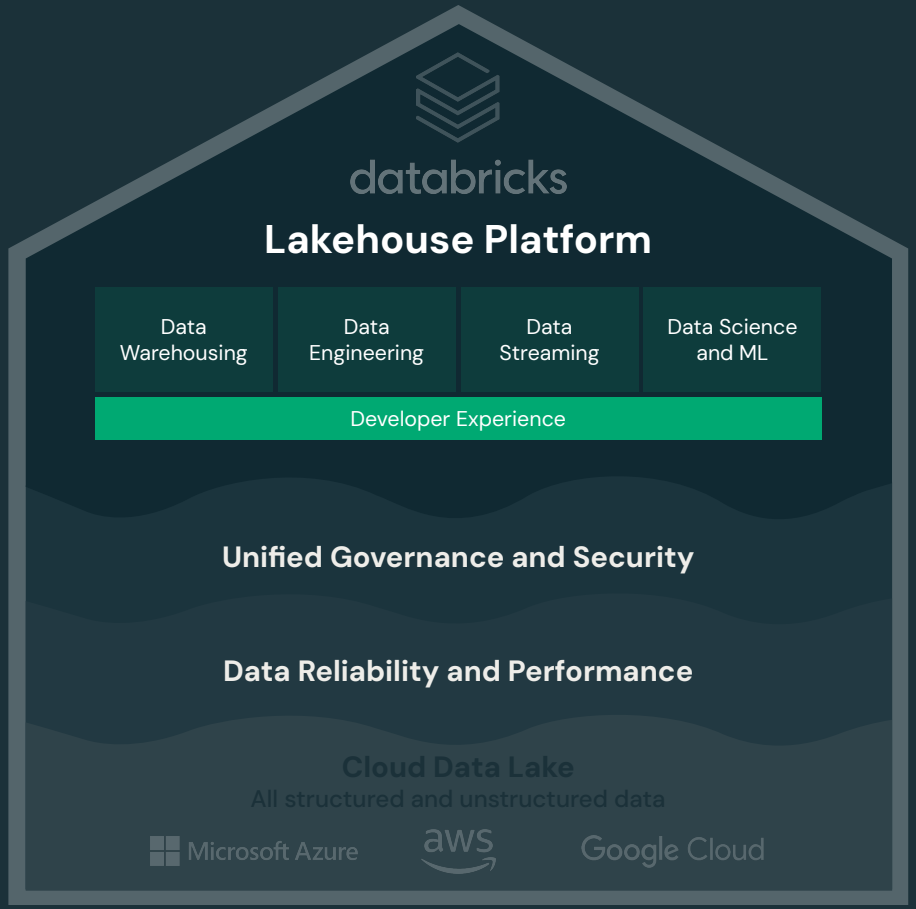
Accelerate BI dashboards and ETL queries

Easily clean, enrich, and denormalize base tables

Built in data sharing & access control



Public Preview Q2 on AWS & Azure



Developer Experience

Today's Topics

New Navigation UI

New Notebooks Editor

SQL Warehouses

DB Connect v2

What's coming in this quarter?

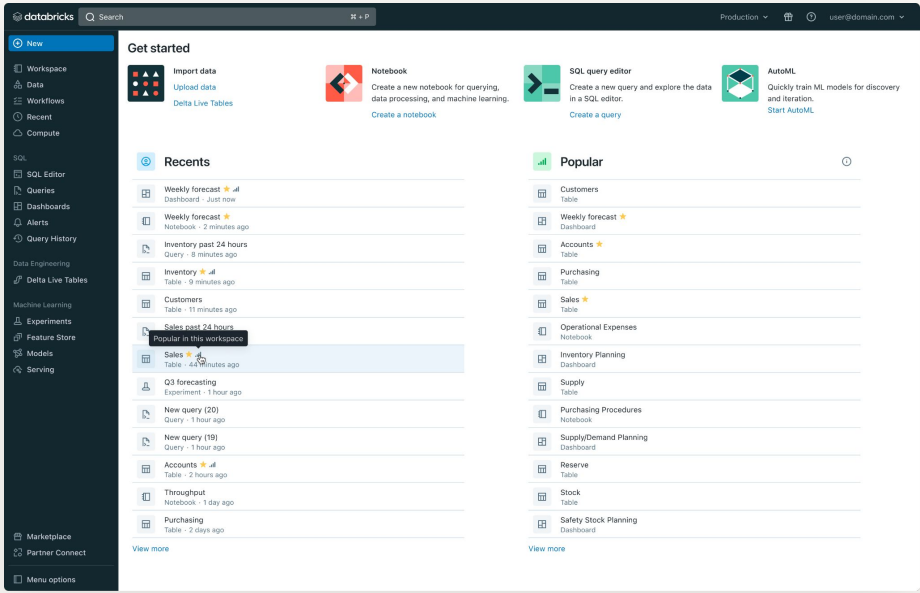
- New Navigation UI
- New Notebooks Editor
- SQL Warehouse
- DB Connect v2
- Full Page Workspace Browser
- Python SDK
- Databricks Asset Bundles (DABs)

New Navigation UI

Navigate Databricks faster and easily find what you are looking for

Fewer clicks to complete tasks

Discover the latest capabilities



Public Preview & GA in Q2

New Notebooks editor

Easier to write, read, and review code in Notebooks

Faster autocomplete-as-you-type

Run selected text from a cell

GA in Q2

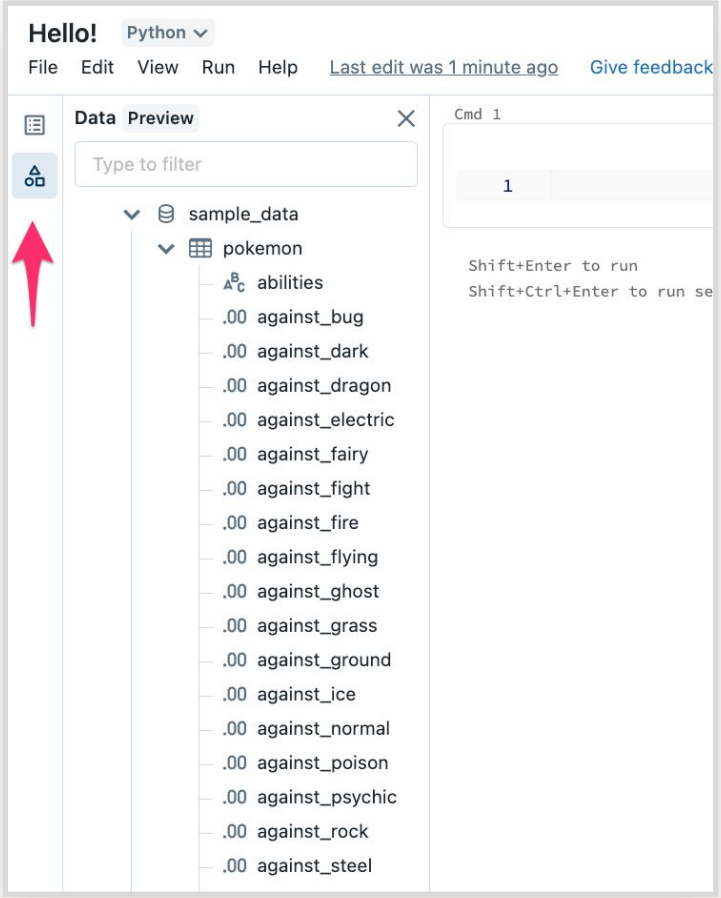
SQL warehouse with Notebooks

Take advantage of SQL warehouses from Notebooks for optimized analytics and serverless capabilities

Better price performance for SQL executions

Use serverless warehouses from Notebooks

GA in Q2

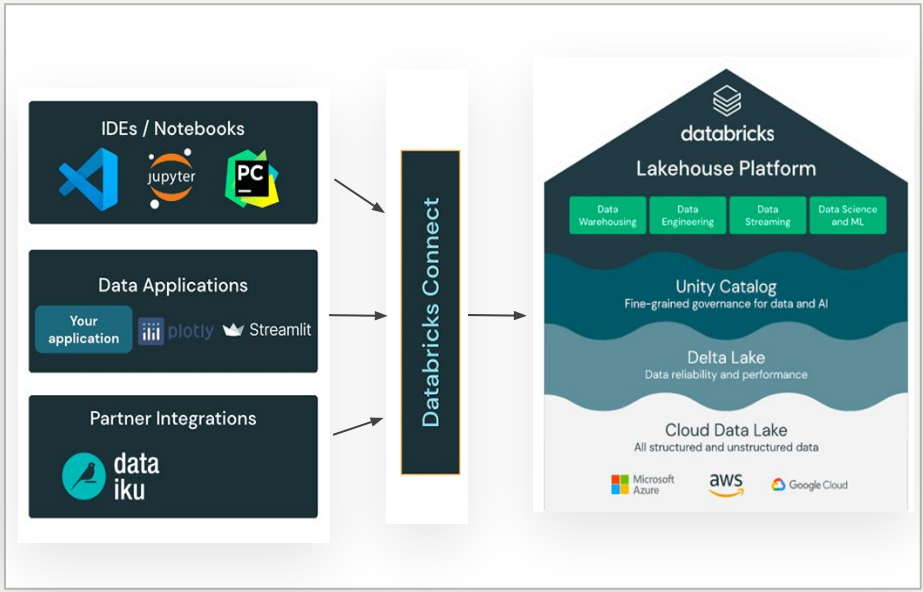


DB Connect v2

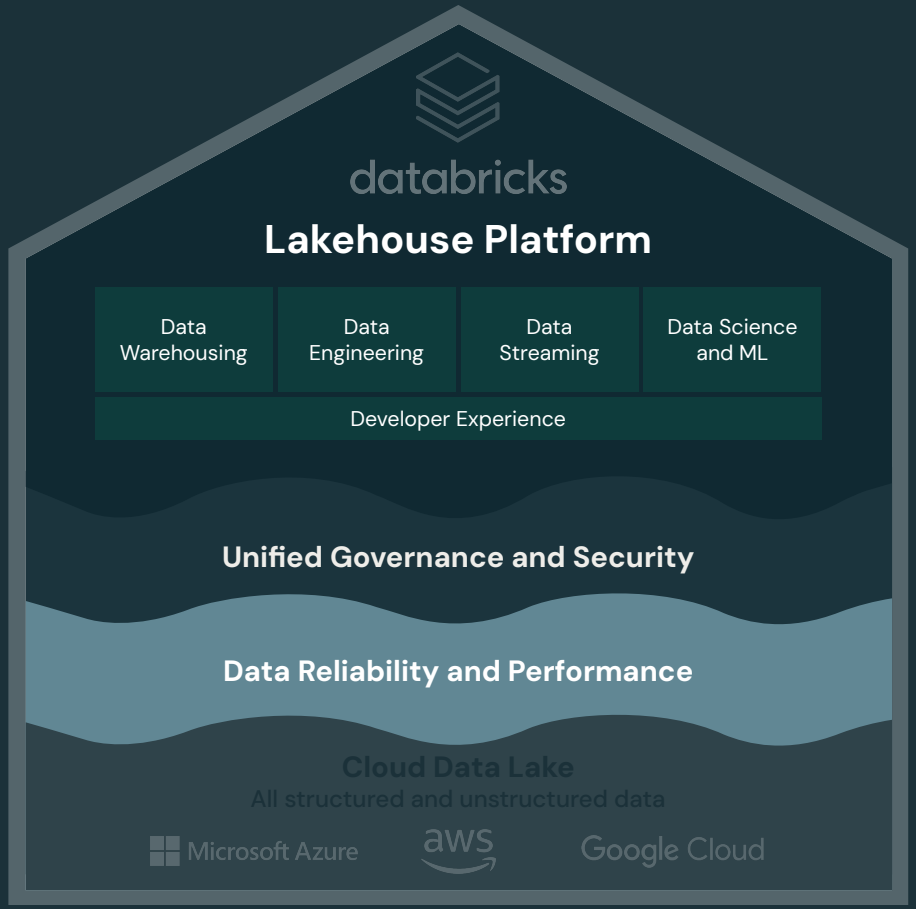
Use Databricks from anywhere

Connect to any IDE or data app

Interactive development and debugging



GA in Q2



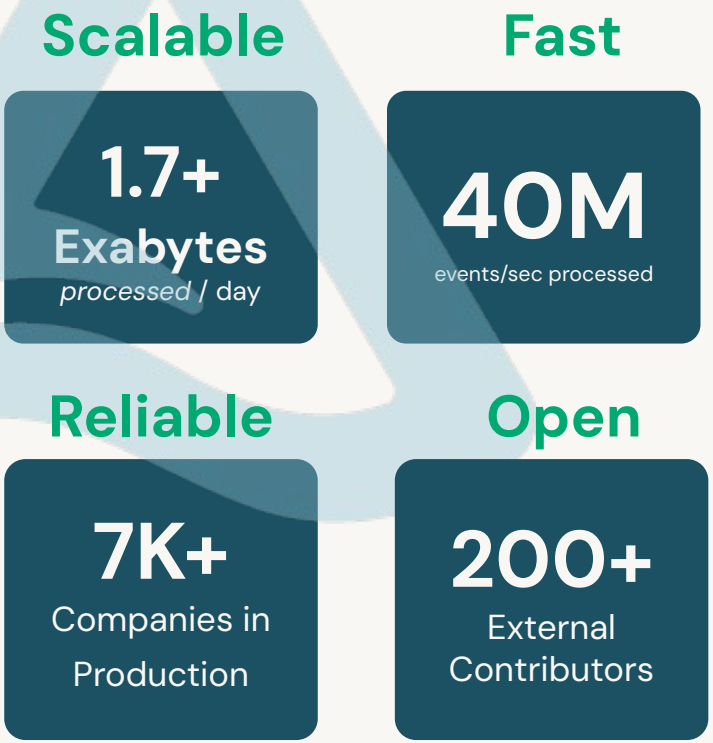
Data Reliability and Performance

Today's Topics

Delta Lake

Photon

Delta Lake



The best foundation for the lakehouse

Open format

First class Streaming support

Secure and open data sharing

Fastest format in the world

What's coming in Delta this quarter?

https://delta.io/roadmap/

Easier Statistics Collection

Spark 3.4

Delta rs0.10: Arrow and DataFusion

Flink SQL and Catalog Support

Flink Scalability Improvements

Optimization to https://github.com/kotosiro

Delta 2.4 on Spark 3.4

Improved SQL functionality and expanded support for deletion vector and streaming

Support for WHEN NOT MATCHED BY SOURCE in SQL MERGE

Support for all writes on tables with Deletion Vectors and for purging Deletion Vectors

Support for reading append-only commits during Streaming Reads

Public Preview in Q2

Fine-grained Statistics collection

Collect fine-grained statistics and improve ingestion performance

```
SELECT * FROM events
WHERE year=2020 AND uid=24000
```

Alter Table events
SET TBLPROPERTIES
(columnstats = [year, uid])

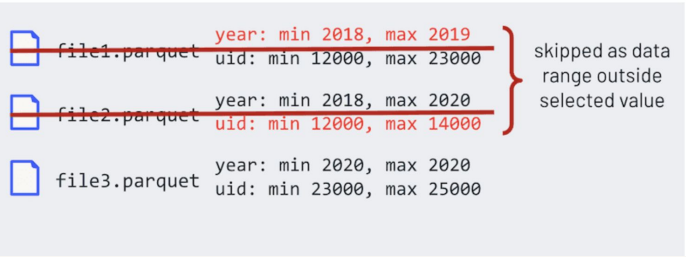


Table property to specify exact columns to collect statistics

Eliminate stats collection on unnecessary columns

Eliminate metadata bloat due to stats on unnecessary columns

Public Preview in Q2

Data management capabilities

Auto Tune for UC Tables

Auto Maintenance

Enhanced Shallow Clones

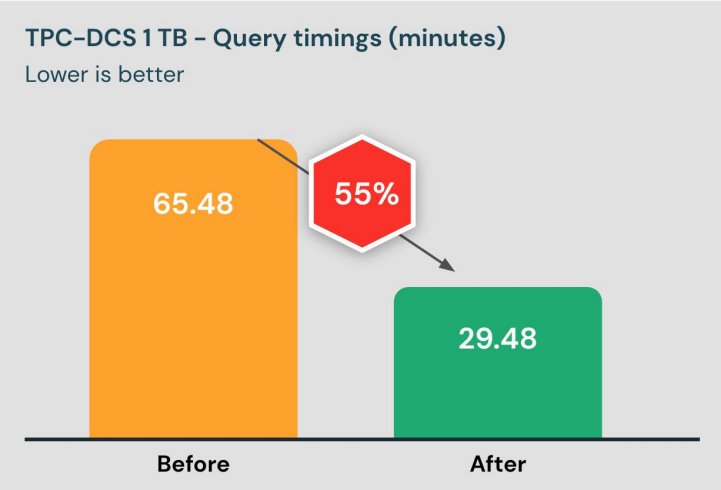
Lifecycle Archival Support

Auto Tuning of file sizes for all Unity Catalog tables

50% increased query performance through product improvement

Auto select target file size	Auto tune optimize writes	Background auto compaction

GA in Q2



Above: performance improvements from Auto Tune capabilities

Auto Maintenance

Automatically identify and trigger operations to maintain Delta Tables at best price-performance

Runs OPTIMIZE, VACUUM, ANALYZE, CLUSTERING

Prioritizes tables based on ROI

Out-of-box observability with system tables

Public Preview in Q2

1 **Phase 1 (Q2)**
OPTIMIZE, VACUUM
UC managed tables

2 **Phase 2**
+ ANALYZE
+ UC external tables

3 **Phase 3**
+ Sorting and clustering optimizations

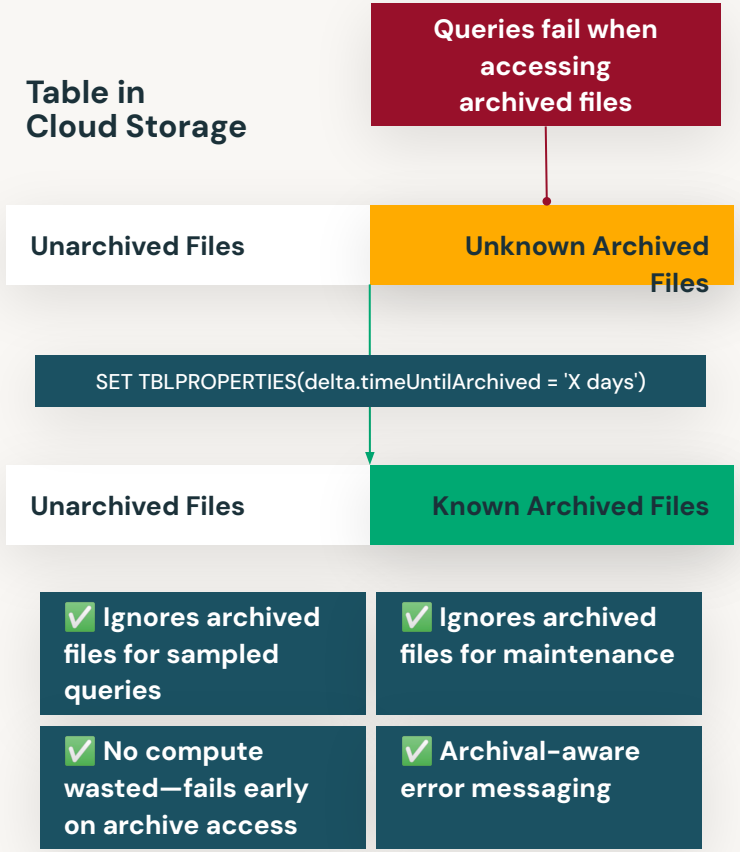
Data archival support

Save on storage costs by storing older data in your Delta table in archival storage (AWS Glacier Deep Archive/Azure Archive).

Ignores archived files for SELECT*, sampled queries, and maintenance operations

No wasted compute—query fails early if it knows it will access a file in archive

Public Preview in Q2



What’s coming in Photon this quarter?

Predictive I/O for updates

Additional instance types supported

Predictive I/O for updates

Up to 10x faster MERGE, UPDATE, and DELETE

Predictive I/O intelligently applies Deletion Vectors to significantly reduce rewrites while not compromising on read performance

MERGE

2-6x

faster vs. DBR
on real-world workloads

DELETE

2-10x

faster vs. DBR
on real-world workloads

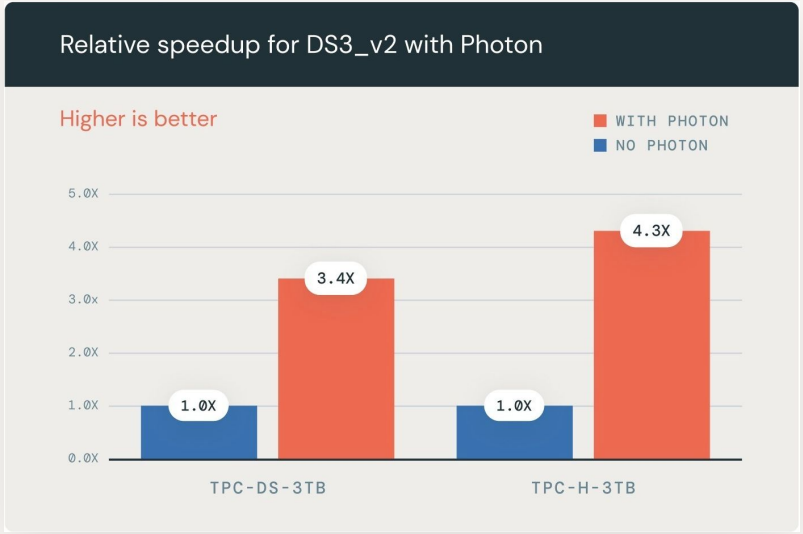
GA in Q2

Additional Instance types supported

Leverage Photon against a broader set of VMs and improve price performance

Offer maximal flexibility for choosing compute for all workloads

Support for general purpose VMs including DS3_v2, DS4_v2, DS5_v2, Standard: D8_v3, D16_v3, D32_v3, D64_v3 on Azure



New Instances Added Each Quarter

DATA+AI SUMMIT

World's largest data, analytics and AI conference

Explore sessions and register at
databricks.com/dataaisummit

IN-PERSON | JUNE 26-29 | SAN FRANCISCO

- Registration NOW OPEN!
- 10k people onsite
- 250+ breakout sessions
- 20+ hands-on-training sessions
- 2 keynotes featuring Databricks founders and guest speakers from DuckDB Labs, LangChain, PyTorch and more
- More meetups, parties and and fun than ever before

Hear from data and AI thought leaders about latest trends and innovations, including Apache Spark, Delta Lake, MLflow, Presto, dbt and more

Learn how others are applying the data lakehouse paradigm to unify data, analytics, and AI on one platform