EBOOK

Why the Data Lakehouse Is Your Next Data Warehouse

2ND EDITION

databricks

Contents

Preface

Historically, data teams have had to resort to a bifurcated architecture to run traditional BI and analytics workloads, copying subsets of the data already stored in their data lake to a legacy data warehouse. Unfortunately, this led to the lock-in, high costs and complex governance inherent in proprietary architectures.

Our customers have asked us to simplify their data architecture. We decided to accelerate our investments to do just that.

We introduced Databricks SQL to simplify and provide data warehousing capabilities and first-class support for SQL on the Databricks Lakehouse Platform, for all your existing tools. We use the term "lakehouse" to reflect our customers' desire to combine the best of data warehouses and data lakes. With the lakehouse, you can now establish one source of truth for all data and enable all workloads from AI to BI on one platform. And we want to provide you with ease-of-use and state-of-the-art performance at the lowest cost.

This eBook covers how we went back to the drawing board to build Databricks SQL — the last mile of enabling data warehousing capabilities for your existing data lakes — as part of the Databricks Lakehouse Platform.

Reynold Xin
Original Creator of Apache Spark,™
Co-founder and Chief Architect,
Databricks

Introduction

Most organizations operate their business with a complex data architecture that combines data warehouses and data lakes. For one thing, data lakes are great for machine learning (ML). They support open formats and a large ecosystem. But data lakes have poor support for business intelligence (BI) and suffer complex data quality problems. Data warehouses, on the other hand, are great for BI applications. But they have limited support for ML workloads, can't handle natural language data, large-scale structured data, or raw, video, audio or image files, and are proprietary systems with only a SQL interface.

As a result, data is moved around the organization through data pipelines and systems that create a multitude of data silos. A large amount of time is spent maintaining these pipelines and systems rather than creating new value from data, and downstream consumers struggle to get a single source of truth of the data due to the inherent siloing of data that takes place. The situation becomes very expensive, and decision-making speed and quality are negatively affected.

Unifying these systems can be transformational in how we think about data.

The need for simplification

It is time for a new data architecture that can meet both today's and tomorrow's needs. Without any compromise. Advanced analytics and ML are one of the most strategic priorities for data-driven organizations today, and the amount of unstructured data is growing exponentially. So it makes sense to position the data lake as the center of the data infrastructure. However, for this to be achievable, the data lake needs to adopt the strengths of data warehouses.

The answer is the lakehouse, an open data architecture enabled by a new open and standardized system design: one that implements data structure and data management features similar to those in a data warehouse, directly on the low-cost storage used for data lakes.

DOWNLOAD NOW

Building the Data Lakehouse

Bill Immon, Father of the Data Warehouse

Our Approach: The Databricks Lakehouse Platform

Our customers have asked us for simplification. This is why we've embarked on this journey to deliver one simple, open and collaborative platform for all your data, AI and BI workloads on your existing data lakes.

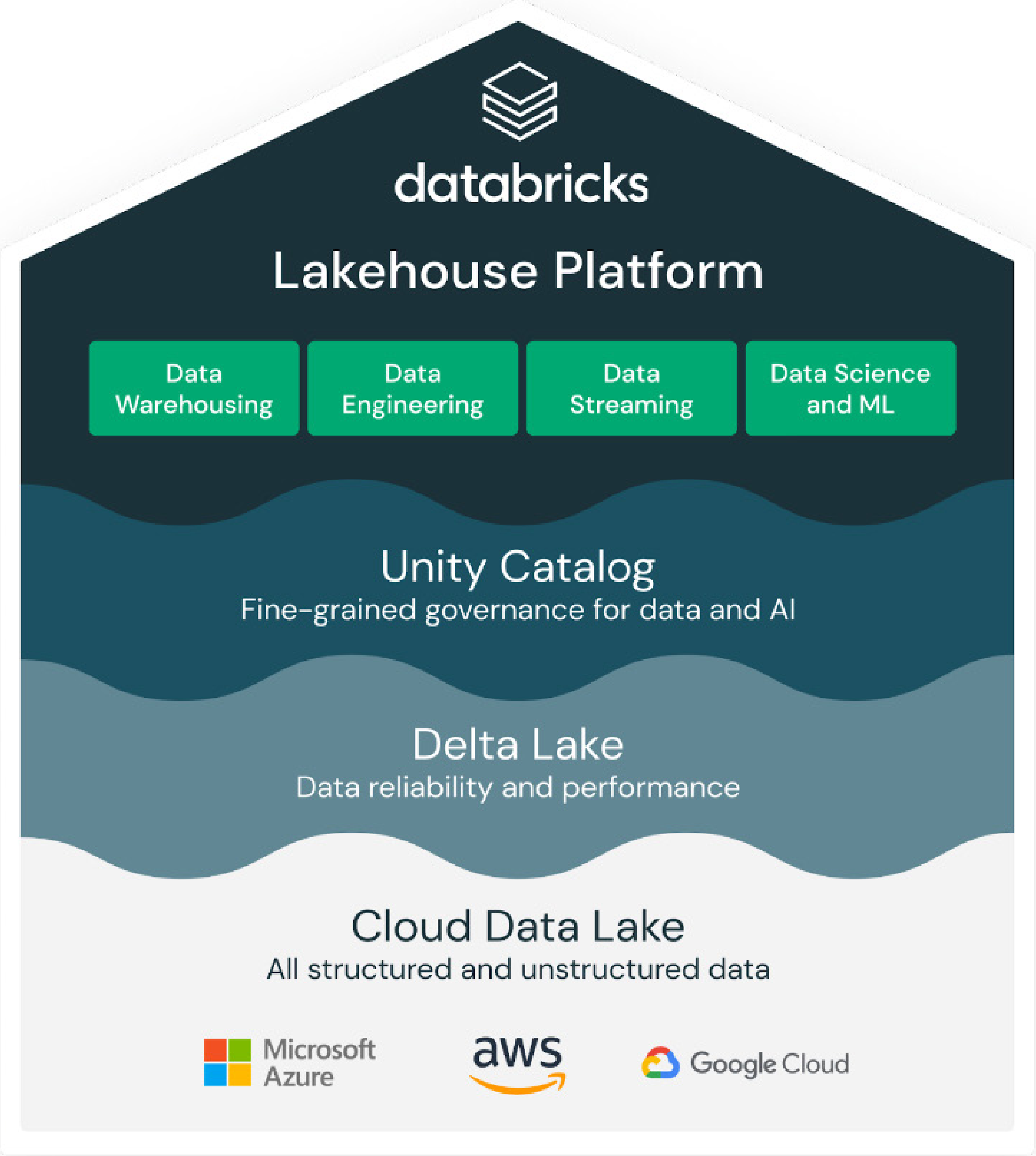
The Databricks Lakehouse Platform greatly simplifies data architectures by combining the data management and performance typically found in data warehouses with the low-cost, flexible object stores offered by data lakes.

It's built on open source and open standards to maximize flexibility, and lets you store all your data — structured, semi-structured and unstructured — in your existing data lake while still getting the data quality, performance, security and governance you'd expect from a data warehouse. Data only needs to exist once to support all of your data, AI and BI workloads on one common platform — establishing one source of truth.

Finally, the Lakehouse Platform provides tailored and collaborative experiences so data engineers, data scientists and analysts can work together on one common platform across the entire data lifecycle — from ingestion to consumption and the serving of data products — and innovate faster.

Let's look at how, with the right data structures and data management capabilities in place, we can now deliver data warehouse and analytics capabilities on your lakehouse. That's where Databricks SQL (DB SQL) comes in.

DISCOVER LAKEHOUSE ›



Introducing Databricks SQL: The Best Data Warehouse Is a Lakehouse

Databricks 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 vendor lock-in. Reduce resource management overhead with serverless compute, and easily ingest, transform and query all your data in place to deliver real-time business insights faster. In fact, DB SQL now holds the new world record in 100TB TPC-DS, the gold standard performance benchmark for data warehousing.

Built on open standards and APIs, the lakehouse provides an open, simplified and multicloud architecture that brings the best of data warehousing and data lakes together, and integrations with a rich ecosystem for maximum flexibility.

Why Databricks SQL?

Best Price/Performance

Lower costs, get world-class performance, and eliminate the need to manage, configure or scale cloud infrastructure with serverless.

Built-In Governance

Establish one single copy for all your data using open standards, and one unified governance layer across all data teams using standard SQL.

Rich Ecosystem

Use SQL and any tool like Fivetran, dbt, Power BI or Tableau along with Databricks to ingest, transform and query all your data in place.

Break Down Silos

Empower every analyst to access the latest data faster for downstream real-time analytics, and go effortlessly from BI to ML.

WATCH A DEMO ❯

Common use cases

Thousands of customers like Atlassian, SEGA and Punchh are using Databricks SQL to enable self-served analytics for hundreds of analysts across their organizations, and to build custom data applications to better serve their customers. Below are some examples of use cases for Databricks SQL.

Query data lake data with your BI tools of choice

Enable business analysts to directly query data lake data using their favorite BI tool and avoid data silos. Reengineered and optimized connectors ensure fast performance, low latency and high user concurrency to your data lake. Now analysts can use the best tool for the job on one single source of truth for your data.

Collaboratively explore the freshest data

Empower every analyst and SQL professional in your organization to quickly find and share new insights by providing them with a collaborative and self-served analytics experience. Confidently manage data permissions with fine-grained governance, share and reuse queries, and quickly analyze and share results using interactive visualizations and dashboards.

Build rich and custom data applications

Build more effective and tailored data applications for your own organization or your customers. Benefit from the ease of connectivity, management and better price/performance of DB SQL to simplify development of data-enhanced applications at scale, all served from your data lake.

> **At Atlassian, we have proven that there is no longer a need for two separate data things. Technology has advanced far enough for us to consider one single unified lakehouse architecture.**
>
> **Rohan Dhupelia**
> Data Platform Senior Manager,
> Atlassian

The Inner Workings of the Lakehouse

In the next chapter, we'll unpack the three foundational layers of the Databricks Lakehouse Platform and how we went back to the drawing board to build this experience. Specifically, we'll dive into how we built Databricks SQL to deliver analytics and data warehousing workloads on your lakehouse.

Those layers are:

1. The storage layer, or how we store and govern data
2. The compute layer, or how we process queries
3. The consumption layer, or the tools you can use to interface with the system

PART 1: STORAGE LAYER

In order to bring the best of data lakes and data warehouses, we needed to support the openness and flexibility of data lakes, as well as the quality, performance and governance you'd expect from a data warehouse.

Storage layer attributes — data lake vs. data warehouse vs. data lakehouse

Data Lake	Data Warehouse	Data Lakehouse
Open format	Closed, proprietary format	Open format
Low quality, "data swamp"	High-quality, reliable data	High-quality, reliable data
File-level access control	Fine-grained governance (tables row/columnar level)	Fine-grained governance (tables row/columnar level)
All data types	Structured only	All data types
Requires manually specifying how to lay out data	Automatically lays out data to query efficiently	Automatically lays out data to query efficiently

Transactional guarantees for your data lake

The open source format Delta Lake — based on Parquet — solves historical data lake challenges around data quality and reliability. It is the foundation for the lakehouse, and Databricks SQL stores and processes data using Delta Lake.

For example, it provides ACID transactions to ensure that every operation either fully succeeds or fully aborts for later retries — without requiring new data pipelines to be created. It unifies batch and streaming pipelines so you can easily merge existing and new data at the speed required for your business. With Time Travel, Delta Lake automatically records all past transactions, so it's easy to access and use previous versions of your data for compliance needs or for ML applications. Advanced indexing, caching and auto-tuning allow optimization of Delta tables for the best query performance. Delta Lake also acts as the foundation for fine-grained, role-based access controls on the lakehouse.

As a result, Delta Lake allows you to treat tables in Databricks SQL just like you treat tables in a database: updates, inserts and merges can take place with high performance at the row level. This is particularly useful if you are inserting new data rapidly (e.g., in IoT or e-commerce use cases), or if you are redacting data (e.g., for compliance laws such as GDPR). Furthermore, Delta Lake provides you with one open and standard format — not only for SQL but also for Python, Scala and other languages — so you can run all analytical and ML use cases on the same data.

DELTA LAKE

Delta Lake provides the key

An open format storage layer built for lake-first architecture

- ACID transactions, Time Travel, highly available
- Advanced indexing, caching, auto-tuning
- Fine-grained, role-based access controls
- Streaming & batch, analytics & ML
- Python, SQL, R, Scala

Delta Lake brings data quality, performance and governance to the lakehouse

DOWNLOAD NOW

Delta Lake: The Definitive Guide

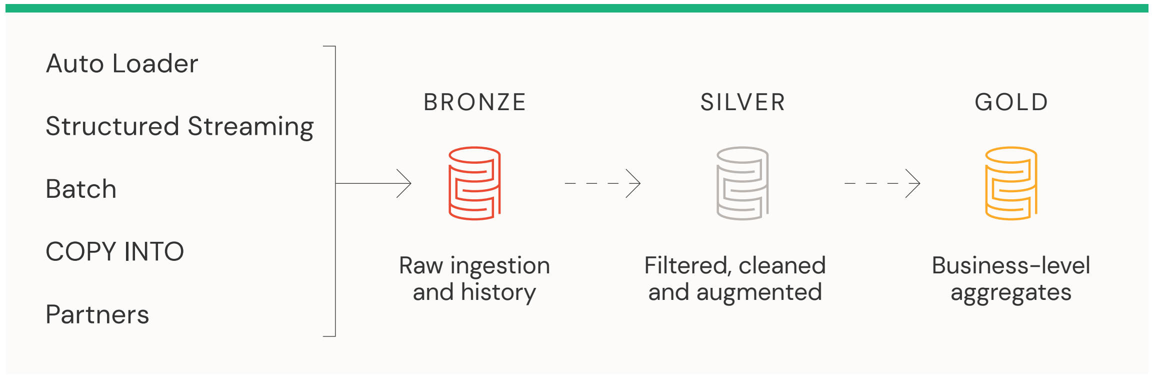
by O'Reilly

A framework for building a curated data lake

With the ability to ingest petabytes of data with auto-evolving schemas, Delta Lake helps turn raw data into actionable data by incrementally and efficiently processing data as it arrives from files or streaming sources like Kafka, Kinesis, Event Hubs, DBMS and NoSQL. It can also automatically and efficiently track data as it arrives with no manual intervention, as well as infer schema, detect column changes for structured and unstructured data formats, and prevent data loss by rescuing data columns that don't meet data quality specifications. And now with Partner Connect, it's never been easier to bring in critical business data from various sources.

As you refine the data, you can add more structure to it. Databricks recommends the Bronze, Silver and Gold pattern. It lets you easily merge and transform new and existing data — in batch or streaming — while benefiting from the low-cost, flexible object storage offered by data lakes. Bronze is the initial landing zone for the pipeline. We recommend copying data that's as close to its raw form as possible to easily replay the whole pipeline from the beginning, if needed. Silver is where the raw data gets cleansed (think data quality checks), transformed and potentially enriched with external data sets. Gold is the production-grade data that your entire company can rely on for business intelligence, descriptive statistics, and data science/machine learning.

By the time you get to Gold, the tables are high-value business-level metrics that have all the schema enforcement and constraints applied. This way, you can retain the flexibility of the data lake at the Bronze and Silver levels, and then use the Gold level for high-quality business data.


Auto Loader
Structured Streaming
Batch
COPY INTO
Partners
BRONZE
Raw ingestion and history
SILVER
Filtered, cleaned and augmented
GOLD
Business-level aggregates

LEARN MORE ›

An aside on batch and streaming data pipelines

The best way to set up and run data pipelines in the Bronze/Silver/Gold pattern recommended on the previous page is in Delta Live Tables (DLT). DLT makes it easy to build and manage reliable batch and streaming data pipelines that deliver high-quality data. It helps data engineering teams simplify ETL development and management with declarative pipeline development, automatic data testing, and deep visibility for monitoring and recovery.

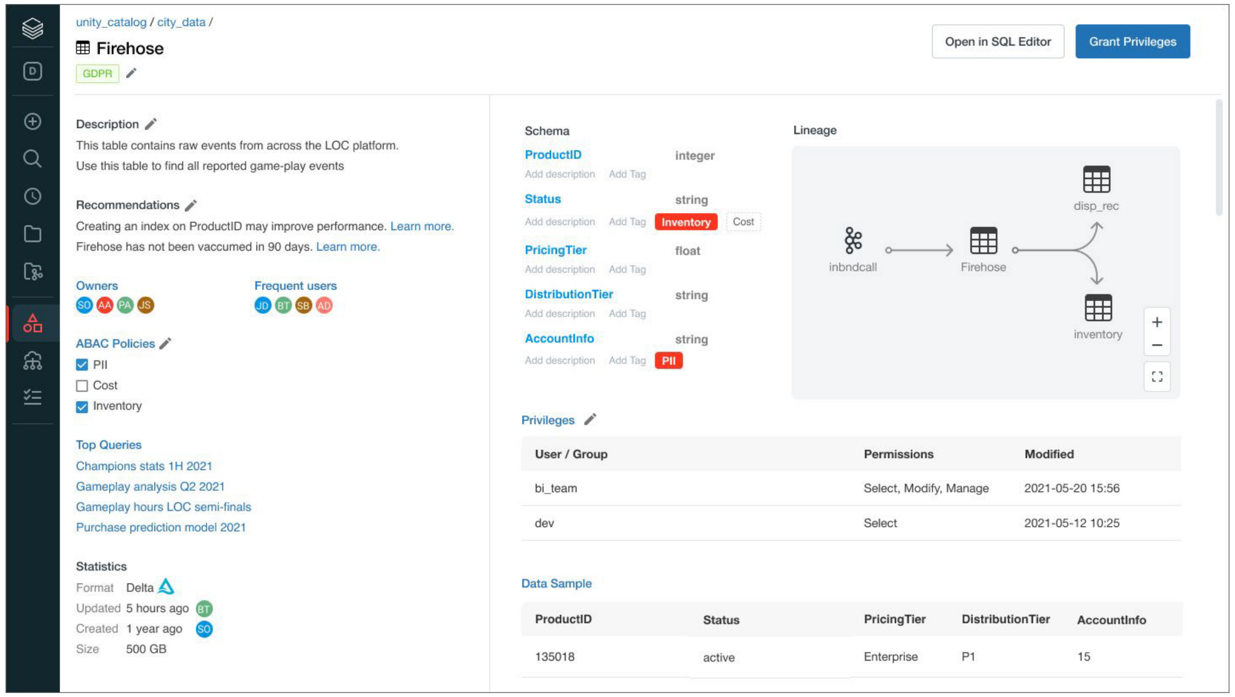
The fact that you can run all your batch and streaming pipelines together in one simple, declarative framework makes data engineering easy on the Databricks Lakehouse Platform. We regularly talk to customers who have been able to reduce pipeline development time from weeks — or months — to mere minutes with Delta Live Tables. And by the way, even data analysts can easily interrogate DLT pipelines for the queries they need to run, without knowing any sort of specialized programming language or niche skills.

One of the top benefits of DLT, and Delta Lake in general, is that it is built with streaming pipelines in mind. Today, the world operates in real time, and businesses are increasingly expected to analyze and respond to their data in real time. With streaming data pipelines built on DLT, analysts can easily access, query and analyze data with greater accuracy and actionability than with conventional batch processing. Delta Live Tables makes real-time analytics a reality for our customers.

Fine-grained governance on the lakehouse

Delta Lake is the foundation for open and secure data sharing and governance on the lakehouse. It underpins the Databricks Unity Catalog (in preview), which provides fine-grained governance across clouds, data and ML assets. Among the benefits of the Unity Catalog, it allows you to:

- **Discover, audit and govern data assets in one place:** A user-friendly interface, automated data lineage across tables, columns, notebooks, workflows and dashboards, role-based security policies, table or column-level tags, and central auditing capabilities make it easy for data stewards to discover, manage and secure data access to meet compliance and privacy needs directly on the lakehouse.
- **Grant and manage permissions using SQL:** Unity Catalog brings fine-grained centralized governance to data assets across clouds through the open standard SQL DCL. This means database administrators can easily grant permission to arbitrary, user-specific views, or set permissions on all columns tagged together, using familiar SQL.
- **Centrally manage and audit shared data across organizations:** Every organization needs to share data with customers, partners and suppliers to better collaborate and to unlock value from their data. Unity Catalog builds on open source Delta Sharing to centrally manage and govern shared assets within and across organizations.



unity_catalog / city_data /

Firehose

GDPR

Description

This table contains raw events from across the LOC platform.
Use this table to find all reported game-play events

Recommendations

Creating an index on ProductID may improve performance. Learn more.
Firehose has not been vaccumed in 90 days. Learn more.

Owners

SO AA PA JS

Frequent users

JD BT SB AD

ABAC Policies

☑ PII
☐ Cost
☑ Inventory

Top Queries

Champions stats 1H 2021
Gameplay analysis Q2 2021
Gameplay hours LOC semi-finals
Purchase prediction model 2021

Statistics

Format Delta
Updated 5 hours ago BT
Created 1 year ago SO
Size 500 GB

Schema

ProductID	integer	
Add description	Add Tag	
Status	string	
Add description	Add Tag	Inventory Cost
PricingTier	float	
Add description	Add Tag	
DistributionTier	string	
Add description	Add Tag	
AccountInfo	string	
Add description	Add Tag	PII

Lineage

Privileges

User / Group	Permissions	Modified
bi_team	Select, Modify, Manage	2021-05-20 15:56
dev	Select	2021-05-12 10:25

Data Sample

ProductID	Status	PricingTier	DistributionTier	AccountInfo
135018	active	Enterprise	P1	15

The Unity Catalog makes it easy for data stewards to discover, manage and secure data access to meet compliance and privacy needs on the lakehouse.

LEARN MORE ›

PART 2: COMPUTE LAYER

The next layer to look at is the compute layer, or how we process queries.

Apache Spark™ has been the de facto standard for data lake compute. It's great for processing terabytes and petabytes of data cheaply, but historically Spark SQL uses a nonstandard syntax and can be difficult to configure.

Data warehouses, on the other hand, tend to support short running queries really well, especially when you have a lot of users issuing queries concurrently. They tend to be easier to set up, but don't necessarily scale or they become too costly.

Compute layer attributes — data lake vs. data warehouse vs. data lakehouse

Data Lake	Data Warehouse	Data Lakehouse
High performance for large jobs (TBs to PBs)	High concurrency	High performance for large jobs (TBs to PBs)
Economical	Scaling is exponentially more expensive	Economical
High operational complexity	Ease of use	Ease of use

A popular belief is that large workloads require a drastically different system than low latency, high concurrency workloads. For example, there's the classic trade-off in computer systems between latency and throughput.

But after spending a lot of time analyzing these systems, we found that it was possible to simultaneously improve large query performance and concurrency and latency. Although the classic trade-offs definitely existed, they were only explicit when we optimized the system to the very theoretical optimal. It turned out the vast majority of software — and this includes all data warehouse systems and Databricks — were far away from optimal.

Simplified administration and instant, elastic SQL compute — decoupled from storage

To achieve world-class performance for analytics on the lakehouse, we chose to completely rebuild the compute layer. But performance isn't everything. We also want it to be simple to administer and cheaper to use. Databricks SQL leverages serverless SQL warehouses that let you get started in seconds, and it's powered by a new native MPP vectorized engine: Photon.

Databricks SQL warehouses are optimized and elastic SQL compute resources. Just pick the cluster size and Databricks automatically determines the best instance types and VMs configuration for the best price/performance. This means you don't have to worry about estimating peak demand or paying too much by overprovisioning. You just need to click a few buttons to operate. To further streamline the experience, simply use Databricks SQL Serverless. With the serverless capability, queries start rapidly with zero infrastructure management or configuration overhead. This lowers your total cost, as you pay only for what you consume without idle time or overprovisioned resources.

Since CPU clock speeds have plateaued, we also wanted to find new ways to process data faster, beyond raw compute power. One of the most impactful methods has been to improve the amount of data that can be processed in parallel. However, data processing engines need to be specifically architected to take advantage of this parallelism. So, from the ground up, we built Photon, a new C++ based vectorized query processing engine that dramatically improves query performance while remaining fully compatible with open Spark APIs. Databricks SQL warehouses are powered by Photon, which seamlessly coordinates work and resources and transparently accelerates portions of your SQL queries directly on your data lake. No need to move the data to a data warehouse.

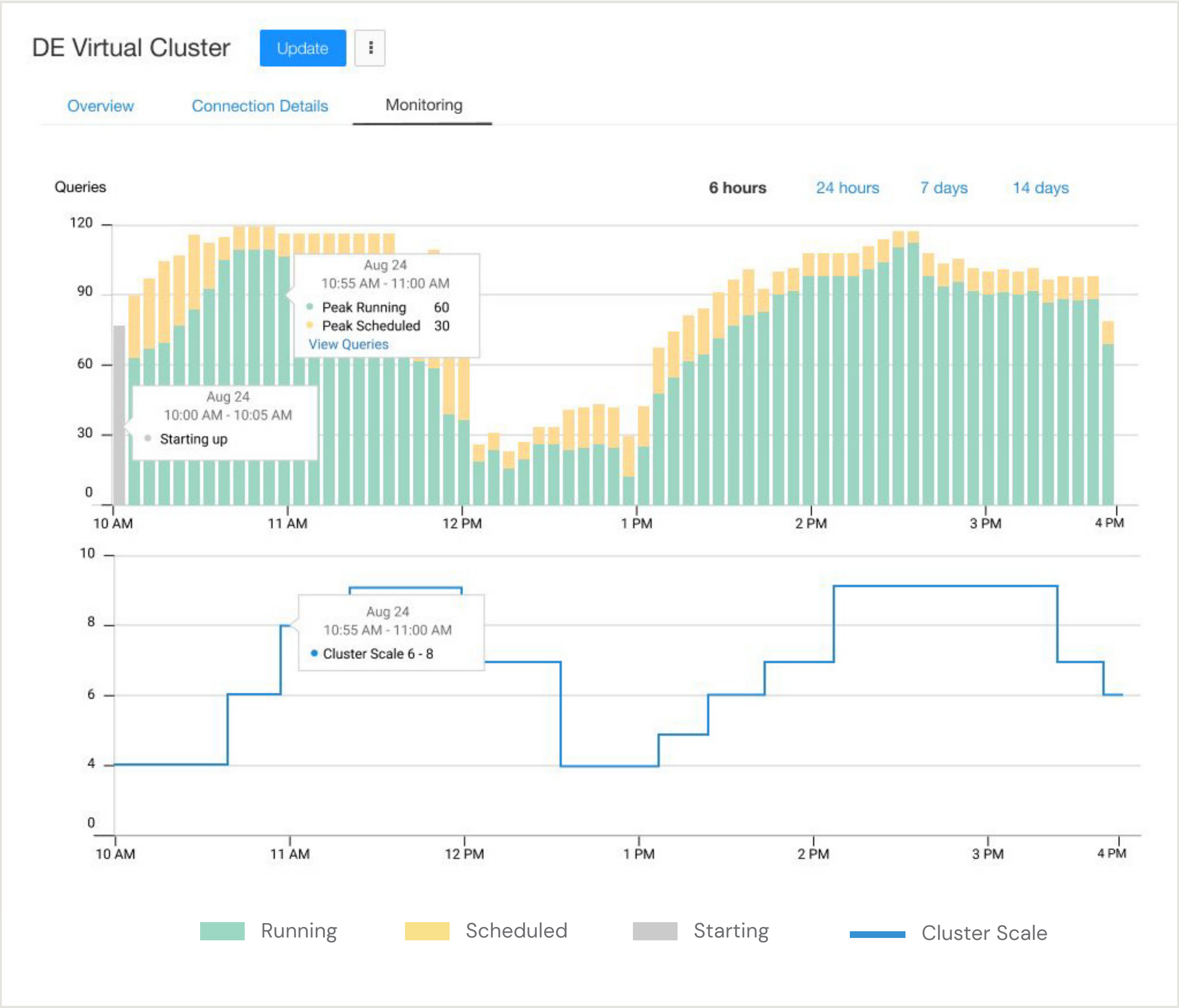
READ NOW

Photon: A Fast Query Engine for Lakehouse Systems

SIGMOD 2022 Best Industry Paper Award

Did you know?

Databricks SQL warehouses scale automatically throughout the day to better suit your business needs. Administration is simplified by identifying how many clusters can scale out with min and max, and Databricks SQL will auto-scale as needed. This ensures that you have ample compute to serve your needs, without overprovisioning. Administrators appreciate the ability to have better control over consumption costs, while users appreciate that their queries process as fast and efficiently as possible. For most BI and analytics use cases, using medium-size warehouses with scaling is a great balance of price/performance that fits most business needs.

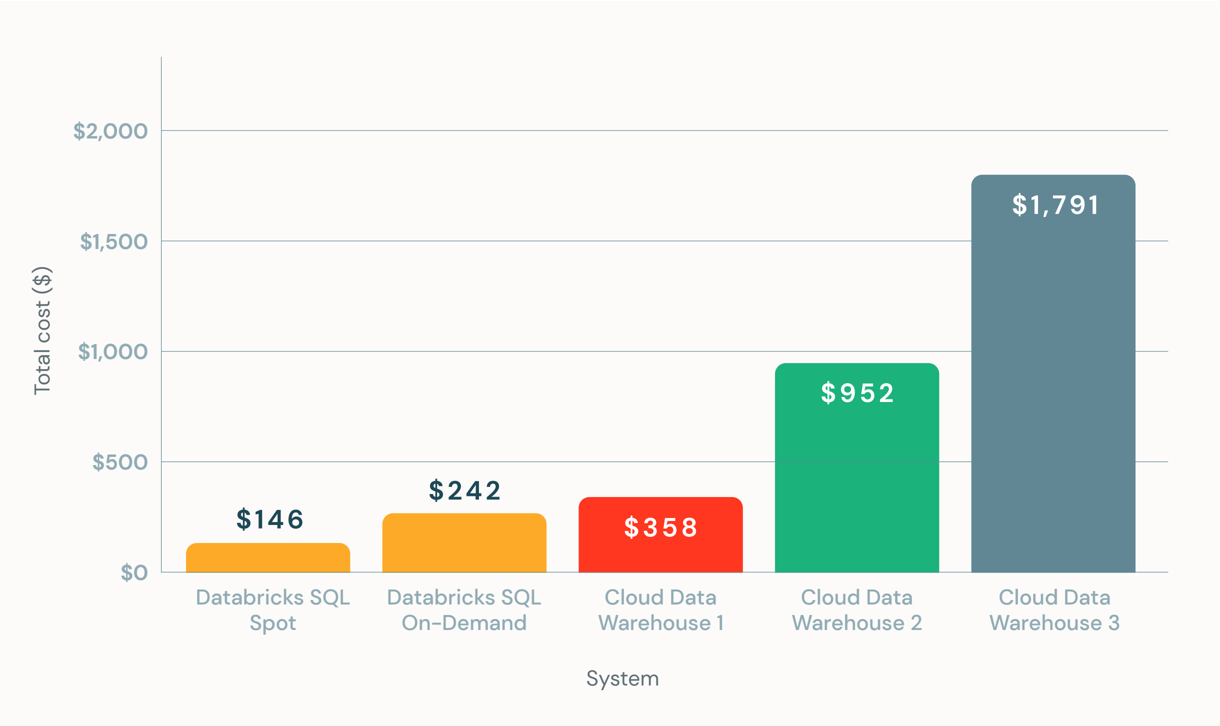


In the next section, we will discuss examples of Databricks SQL performance results on large-scale analytic workloads as well as highly concurrent workloads.

Large query performance: the fastest data warehouse

The industry standard benchmark used by data warehouses is TPC-DS. It includes 100 queries that range from very simple to very sophisticated to simulate decision support workloads. This benchmark was created by a committee formed by data warehousing vendors. The chart at right shows price/performance results running the 100TB version of TPC-DS, since for large workloads the numbers that ultimately matter pertain to the performance cost. As you can see, Databricks SQL outperforms all cloud data warehouses we have measured.

LEARN MORE ›



100TB TPC-DS price/performance benchmark (lower is better).

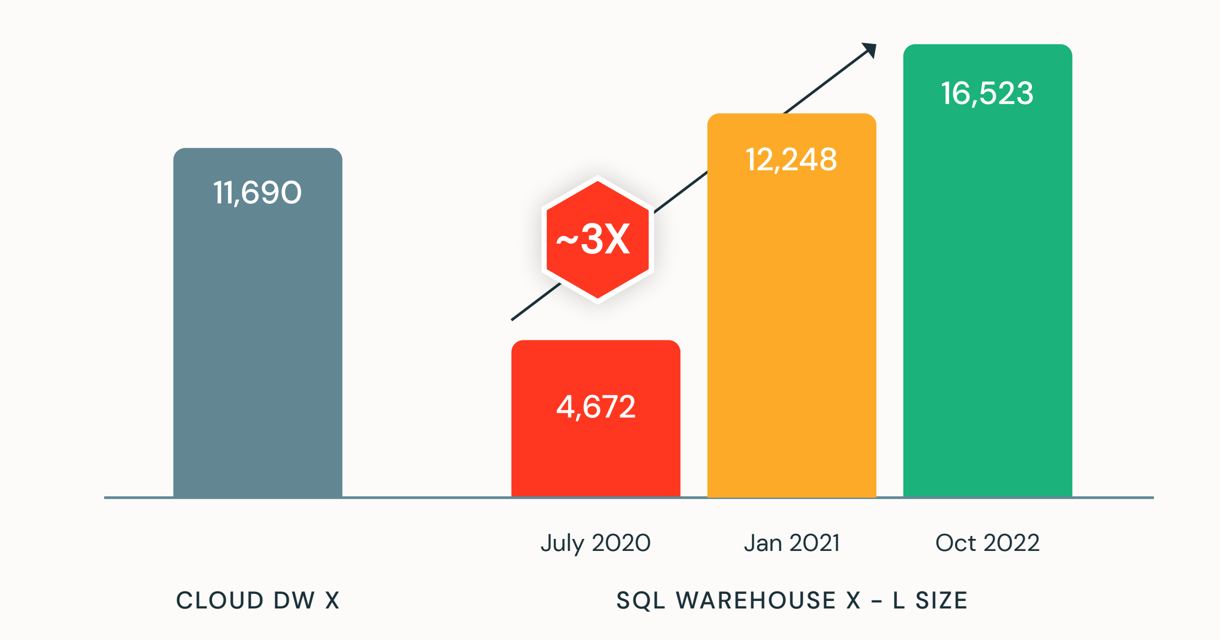
Did you know?

Databricks SQL has set a new world record in 100TB TPC-DS, the gold standard performance benchmark for data warehousing. Databricks SQL outperformed the previous record by 2.2x. And this result has been formally audited and reviewed by the TPC council.

Highly concurrent analytics workloads

Beyond large queries, it is also common for highly concurrent analytics workloads to execute over small data sets. To optimize concurrency, we used the same TPC-DS benchmark, but on a much smaller scale (10GB) and with 32 concurrent streams. We analyzed the results to identify and remove bottlenecks, and built hundreds of optimizations to improve concurrency. Databricks SQL now outperforms some of the best cloud data warehouses for both large queries and small queries with lots of users.

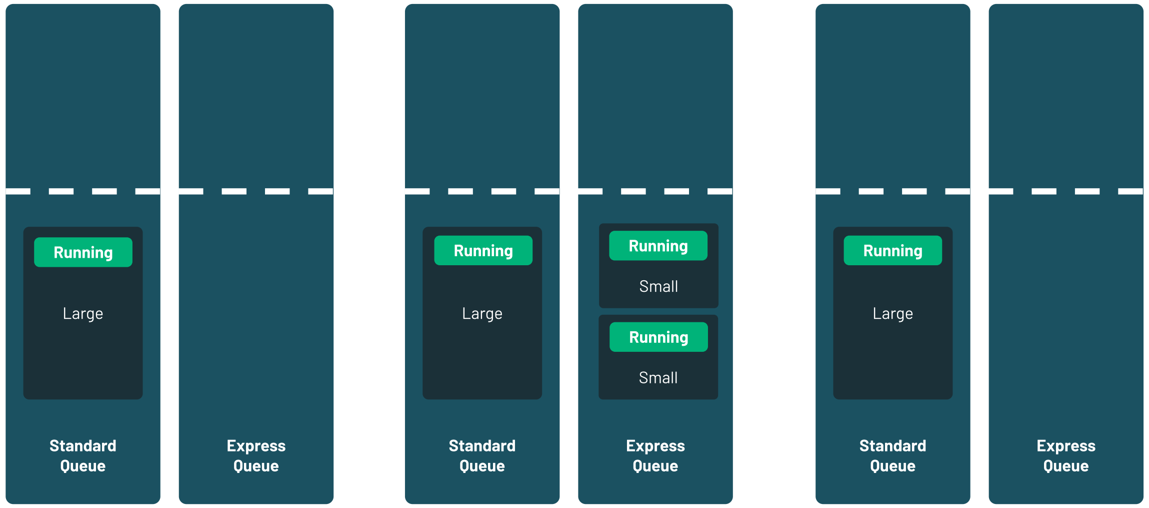
Real-world workloads, however, are not just about either large or small queries. Databricks SQL also provides intelligent workload management with a dual queuing system and highly parallel reads.



10GB TPC-DS queries/hr at 32 concurrent streams (higher is better).

Intelligent workload management with smart queuing system

Real-world workloads typically include a mix of small and large queries. Therefore the smart queuing and load balancing capabilities of Databricks SQL need to account for that too. Databrick SQL uses a smart dual queuing system (in preview) that prioritizes small queries over large, as analysts typically care more about the latency of short queries than large ones.



5:00 PM

Large query in progress

5:05 PM

Large query in progress

New small queries immediately in progress

5:05:10 PM

Large query in progress

Small queries complete! ✓

Highly parallel reads with improved I/O performance

It is common for some tables in a lakehouse to be composed of many files — for example, in streaming scenarios such as IoT ingest when data arrives continuously. In legacy systems, the execution engine can spend far more time listing these files than actually executing the query. Our customers told us they do not want to sacrifice performance for data freshness. With async and highly parallel I/O, when executing a query, Databricks SQL now automatically reads the next blocks of data from cloud storage while the current block is being processed. This considerably increases overall query performance on small files (by 12x for 1MB files) and "cold data" (data that is not cached) use cases as well.

LEARN MORE ›

PART 3: CONSUMPTION LAYER

The third layer of the Databricks Lakehouse Platform would similarly have to bridge the best of both data lakes and data warehouses. In the lakehouse, you would have to be able to work seamlessly with your tools of choice — whether you are a business analyst, data scientist, or ML or data engineer.

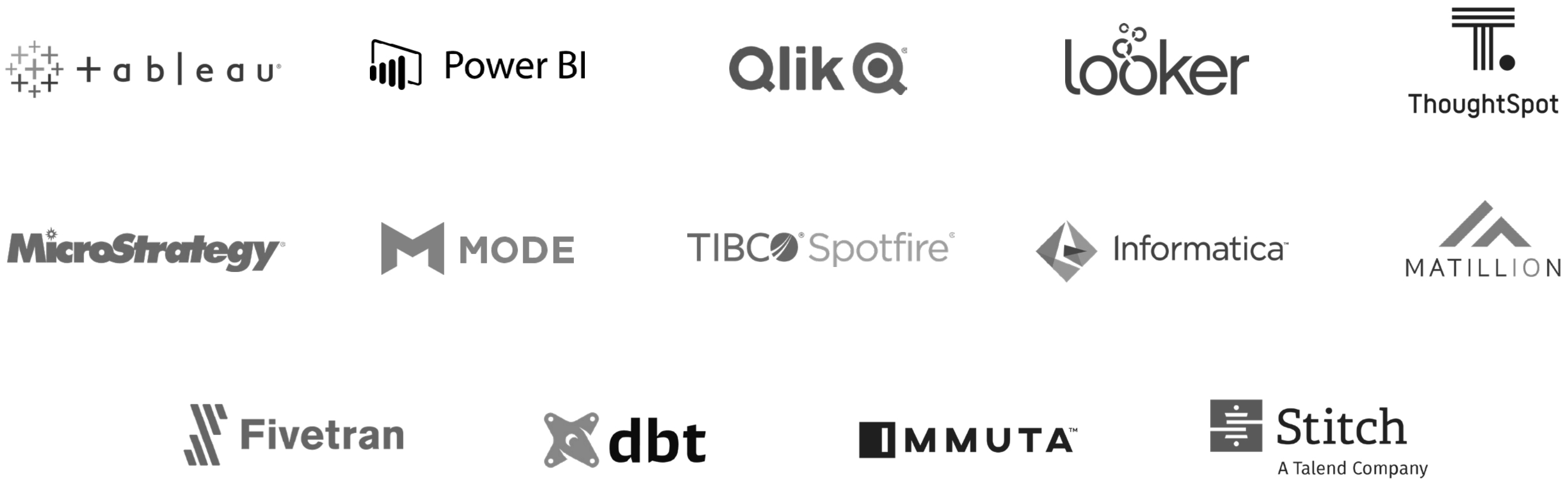
The lakehouse must treat Python, Scala, R and SQL programming languages and ecosystems as first-class citizens to truly unify data engineering, ML and BI workloads in one place.

Consumption layer attributes — data lake vs. data warehouse vs. data lakehouse

Data Lake	Data Warehouse	Data Lakehouse
Notebooks (great for data scientists)	Lack of support for data science/ML	Notebooks (great for data scientists)
Openness with rich ecosystem (Python, R, Scala)	Limited to SQL only	Openness with rich ecosystem (Python, R, Scala)
BI/SQL not 1st-class citizen	BI/SQL 1st-class citizen	BI/SQL 1st-class citizen

A platform for your tools of choice

At Databricks we believe strongly in open platforms and meeting our customers where they are. We work very closely with a large number of software vendors to make sure you can easily use your tools of choice on Databricks, like Tableau, Power BI or dbt. With Partner Connect, it's easier than ever to connect with your favorite tools, easier to get data in, easier to authenticate using single sign-on, and of course, with all the concurrency and performance improvements, we make sure that the direct and live query experience is great.



+ Any other Apache Spark-compatible client

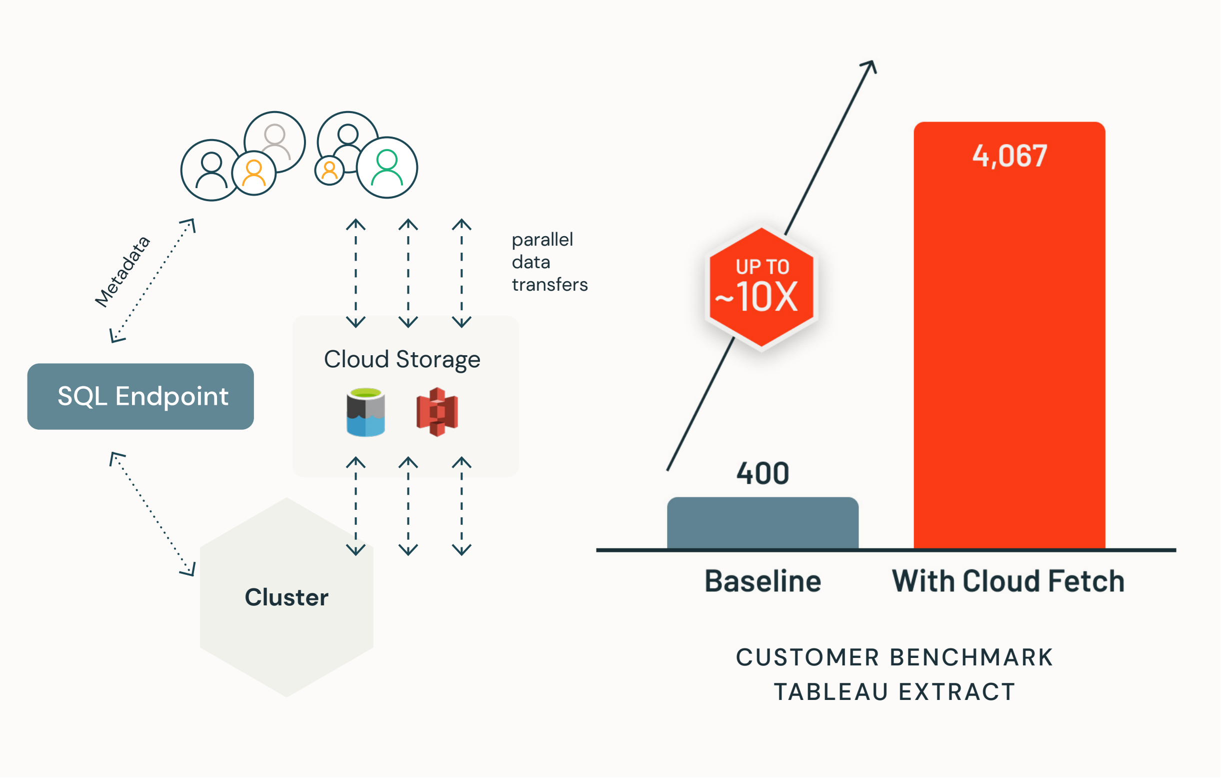
> Now more than ever, organizations need a data strategy that enables speed and agility to be adaptable. As organizations are rapidly moving their data to the cloud, we're seeing growing interest in doing analytics on the data lake. The introduction of Databricks SQL delivers an entirely new experience for customers to tap into insights from massive volumes of data with the performance, reliability and scale they need. We're proud to partner with Databricks to bring that opportunity to life.
>
> **Francois Ajenstat**
> Chief Product Officer, Tableau

Faster BI results retrieval with Cloud Fetch

Once query results are computed, cloud data warehouses often collect and stream back results to BI clients on a single thread. This can create a bottleneck and greatly slows down the experience if you are fetching anything more than a few megabytes of results in size. To provide analysts with the best experience from their favorite BI tools, we also needed to speed up how the system delivers results to BI tools like Power BI or Tableau once computed.

That's why we've reimagined this approach with a new architecture called Cloud Fetch. For large results, Databricks SQL now writes results in parallel across all of the compute nodes to cloud storage, and then sends the list of files using pre-signed URLs back to the client. The client then can download in parallel all the data from cloud storage. This approach provides up to 10x performance improvement in real-world scenarios.

LEARN MORE ›

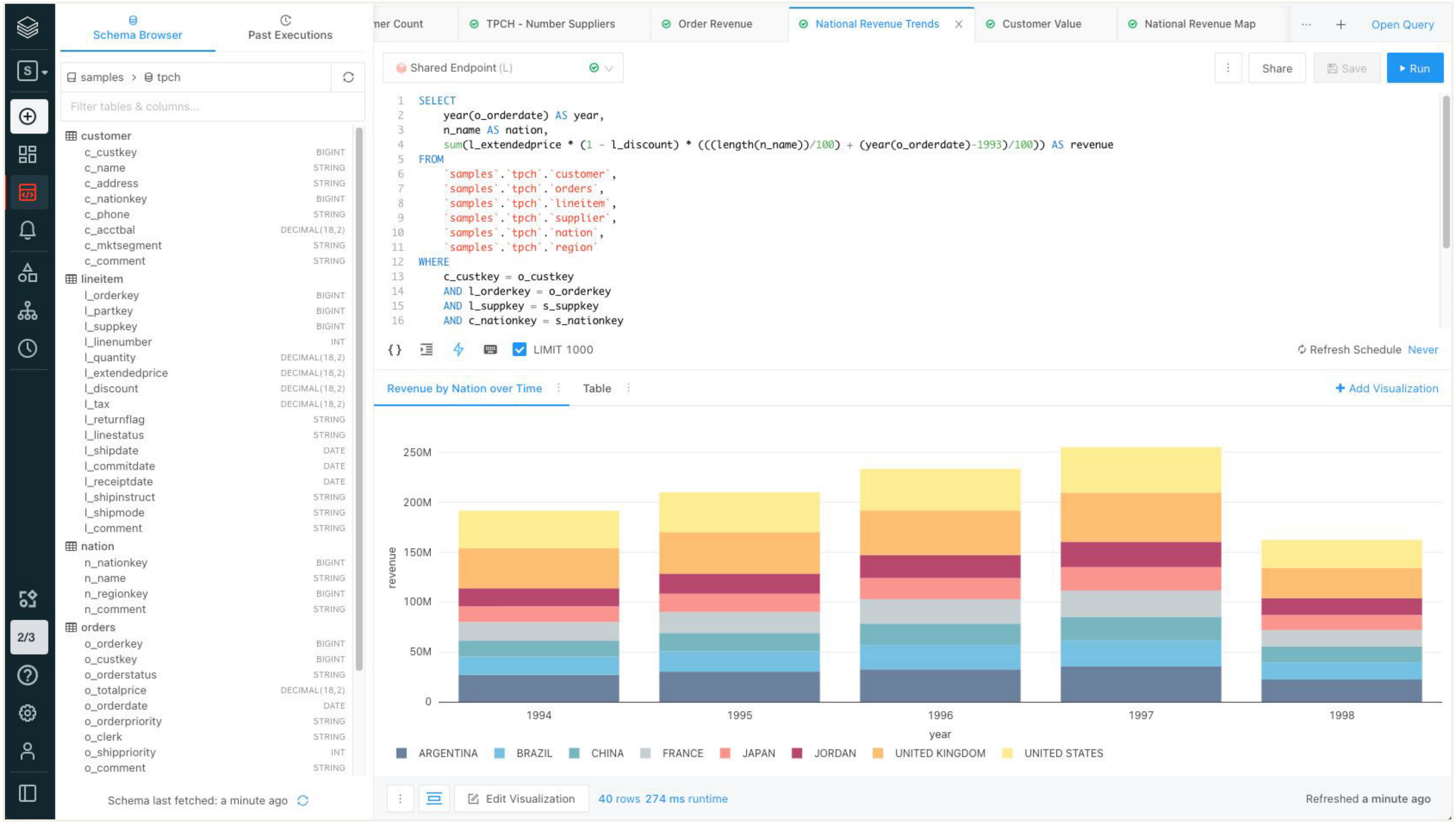


Cloud Fetch enables faster, higher-bandwidth connectivity to and from your BI tools.

A first-class SQL development experience

In addition to supporting your favorite tools, we are also focused on providing a native first-class SQL development experience. We've talked to hundreds of analysts using various SQL editors like SQL Workbench every day, and worked with them to provide the dream set of capabilities for SQL development.

For example, Databricks SQL now supports standard ANSI SQL, so you don't need to learn a special SQL dialect. Query tabs allow you to work on multiple queries at once, autosave gives you peace of mind so you never have to worry about losing your drafts, integrated history lets you easily look at what you have run in the past, and intelligent auto-complete understands subqueries and aliases for a delightful experience.



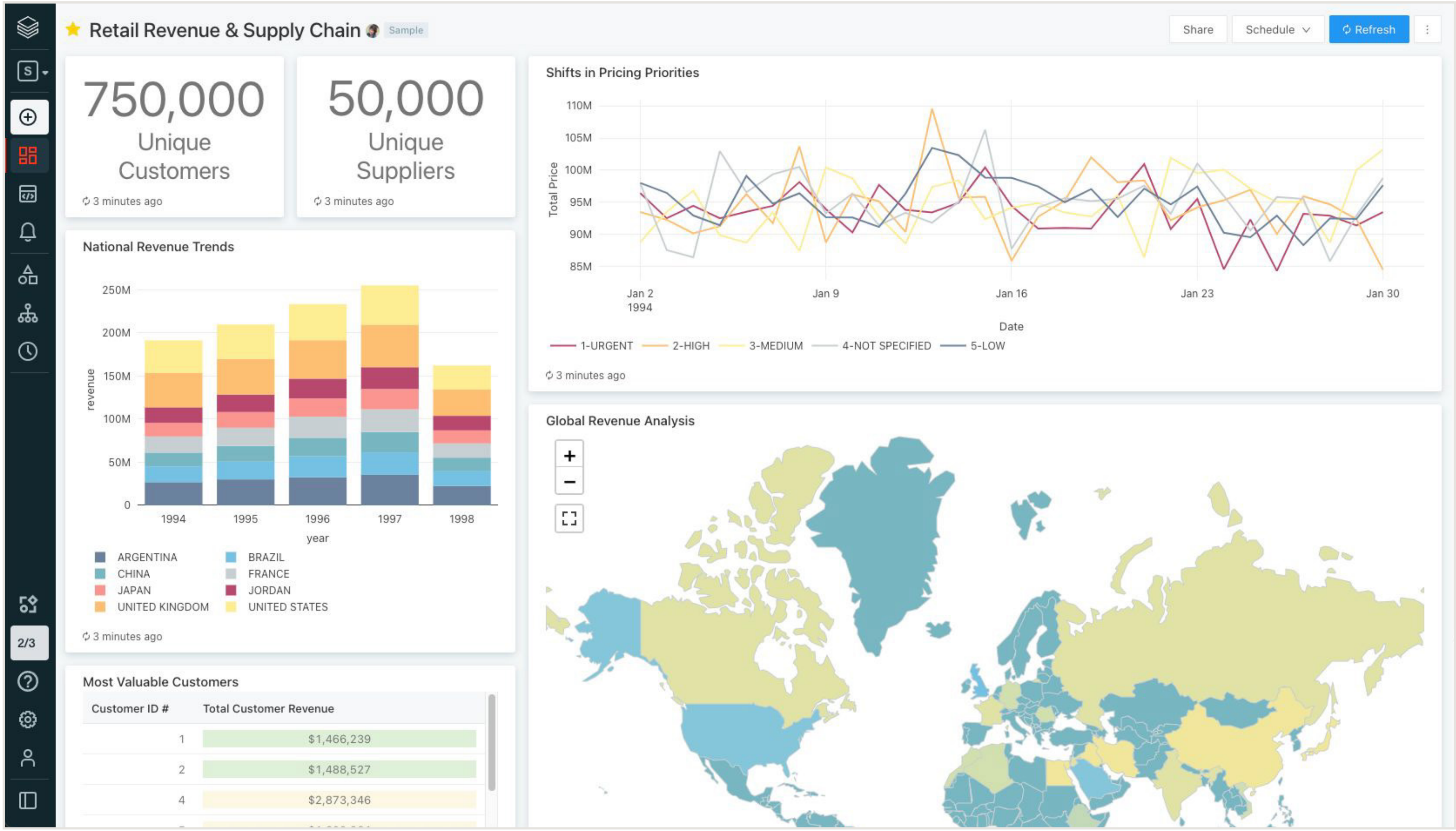
customer	
c_custkey	BIGINT
c_name	STRING
c_address	STRING
c_nationkey	BIGINT
c_phone	STRING
c_acctbal	DECIMAL(18,2)
c_mktsegment	STRING
c_comment	STRING
lineitem	
l_orderkey	BIGINT
l_partkey	BIGINT
l_suppkey	BIGINT
l_linenumber	INT
l_quantity	DECIMAL(18,2)
l_extendedprice	DECIMAL(18,2)
l_discount	DECIMAL(18,2)
l_tax	DECIMAL(18,2)
l_returnflag	STRING
l_linestatus	STRING
l_shipdate	DATE
l_commitdate	DATE
l_receiptdate	DATE
l_shipinstruct	STRING
l_shipmode	STRING
l_comment	STRING
nation	
n_nationkey	BIGINT
n_name	STRING
n_regionkey	BIGINT
n_comment	STRING
orders	
o_orderkey	BIGINT
o_custkey	BIGINT
o_orderstatus	STRING
o_totalprice	DECIMAL(18,2)
o_orderdate	DATE
o_orderpriority	STRING
o_clerk	STRING
o_shippriority	INT
o_comment	STRING

Schema last fetched: a minute ago

```sql
SELECT
    year(o_orderdate) AS year,
    n_name AS nation,
    sum(l_extendedprice * (1 - l_discount) * (((length(n_name))/100) + (year(o_orderdate)-1993)/100)) AS revenue
FROM
    `samples`.`tpch`.`customer`,
    `samples`.`tpch`.`orders`,
    `samples`.`tpch`.`lineitem`,
    `samples`.`tpch`.`supplier`,
    `samples`.`tpch`.`nation`,
    `samples`.`tpch`.`region`
WHERE
    c_custkey = o_custkey
    AND l_orderkey = o_orderkey
    AND l_suppkey = s_suppkey
    AND c_nationkey = s_nationkey
```

The built-in SQL query editor allows you to quickly explore available databases, query and visualize results.

Finally, with Databricks SQL, analysts can easily make sense of query results through a wide variety of rich visualizations and quickly build dashboards with an intuitive drag-and-drop interface. To keep everyone current, dashboards can be shared and configured to automatically refresh, as well as to alert the team to meaningful changes in the data.



Retail Revenue & Supply Chain

750,000 Unique Customers

50,000 Unique Suppliers

Most Valuable Customers

Customer ID #	Total Customer Revenue
1	$1,466,239
2	$1,488,527
4	$2,873,346

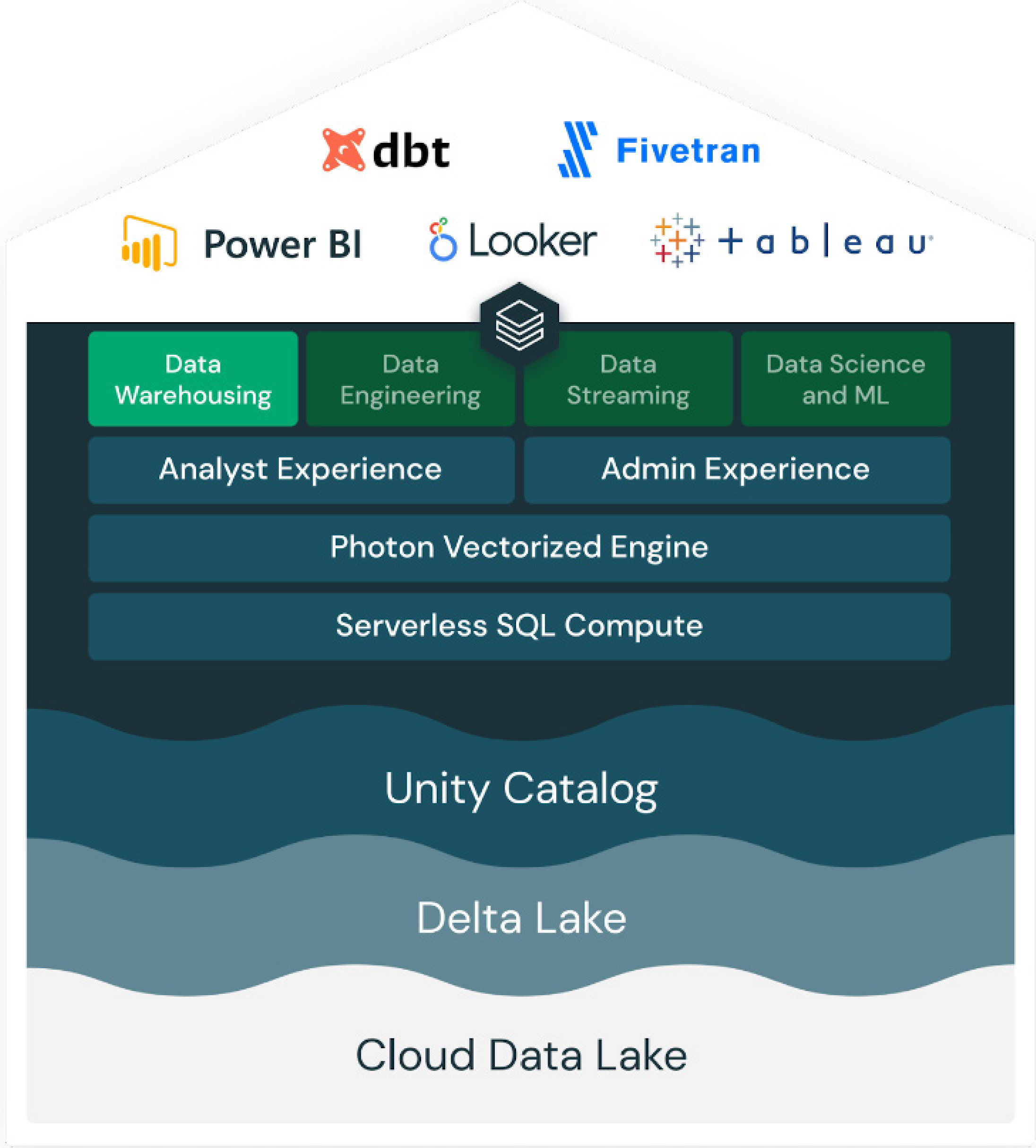
Easily combine visualizations to build rich dashboards that can be shared with stakeholders.

Conclusion

Databricks SQL leverages open source standard Delta Lake to turn raw data into actionable data, combining the flexibility and openness of data lakes with the reliability and performance of data warehouses. The Unity Catalog provides fine-grained governance on the lakehouse across all clouds using one friendly interface and standard SQL.

Databricks SQL also holds the new world record in 100TB TPC-DS, the gold standard performance benchmark for data warehousing. It is powered by Photon, the new vectorized query engine for the lakehouse, and by SQL warehouses for instant, elastic compute decoupled from storage.

Finally, Databricks SQL offers a native first-class SQL development experience, with a built-in SQL editor, rich visualizations and dashboards, and integrates seamlessly with your favorite BI- and SQL-based tools for maximum productivity.



Databricks SQL under the hood.

Atlassian

Atlassian is a leading provider of collaboration, development and issue-tracking software for teams. With over 150,000 global customers (including 85 of the Fortune 100), Atlassian is advancing the power of collaboration with products including Jira, Confluence, Bitbucket, Trello and more.

USE CASE

Atlassian uses the Databricks Lakehouse Platform to democratize data across the enterprise and drive down operational costs. Atlassian currently has a number of use cases focused on putting the customer experience at the forefront.

Customer support and service experience
With the majority of their customers being server-based (using products like Jira and Confluence), Atlassian set out to move those customers into the cloud to leverage deeper insights that enrich the customer support experience.

Marketing personalization
The same insights could also be used to deliver personalized marketing emails to drive engagement with new features and products.

Anti-abuse and fraud detection
They can predict license abuse and fraudulent behavior through anomaly detection and predictive analytics.

> **At Atlassian, we need to ensure teams can collaborate well across functions to achieve constantly evolving goals. A simplified lakehouse architecture would empower us to ingest high volumes of user data and run the analytics necessary to better predict customer needs and improve the experience of our customers. A single, easy-to-use cloud analytics platform allows us to rapidly improve and build new collaboration tools based on actionable insights.**
>
> **Rohan Dhupelia**
> Data Platform Senior Manager, Atlassian

SOLUTION AND BENEFITS

Atlassian is using the Databricks Lakehouse Platform to enable data democratization at scale, both internally and externally. They have moved from a data warehousing paradigm to standardization on Databricks, enabling the company to become more data driven across the organization. Over 3,000 internal users in areas ranging from HR and marketing to finance and R&D — more than half the organization — are accessing insights from the platform on a monthly basis via open technologies like Databricks SQL. Atlassian is also using the platform to drive more personalized support and service experiences to their customers.

- Delta Lake underpins a single lakehouse for PBs of data accessed by 3,000+ users across HR, marketing, finance, sales, support and R&D
- BI workloads powered by Databricks SQL enable dashboard reporting for more users
- MLflow streamlines MLOps for faster delivery
- Data platform unification eases governance, and self-managed clusters enable autonomy

With cloud-scale architecture, improved productivity through cross-team collaboration, and the ability to access all of their customer data for analytics and ML, the impact on Atlassian is projected to be immense. Already the company has:

- Reduced the cost of IT operations (specifically compute costs) by 60% through moving 50,000+ Spark jobs from EMR to Databricks with minimal effort and low-code change
- Decreased delivery time by 30% with shorter dev cycles
- Reduced data team dependencies by 70% with more self-service enabled throughout the organization

LEARN MORE ❯

ABN AMRO

As an established bank, ABN AMRO wanted to modernize their business but were hamstrung by legacy infrastructure and data warehouses that complicated access to data across various sources and created inefficient data processes and workflows. Today, Azure Databricks empowers ABN AMRO to democratize data and AI for a team of 500+ empowered engineers, scientists and analysts who work collaboratively on improving business operations and introducing new go-to-market capabilities across the company.

USE CASE

ABN AMRO uses the Databricks Lakehouse Platform to deliver financial services transformation on a global scale, providing automation and insight across operations.

Personalized finance

ABN AMRO leverages real-time data and customer insights to provide products and services tailored to customers' needs. For example, they use machine learning to power targeted messaging within their automated marketing campaigns to help drive engagement and conversion.

Risk management

Using data-driven decision-making, they are focused on mitigating risk for both the company and their customers. For example, they generate reports and dashboards that internal decision makers and leaders use to better understand risk and keep it from impacting ABN AMRO's business.

Fraud detection

With the goal of preventing malicious activity, they're using predictive analytics to identify fraud before it impacts their customers. Among the activities they're trying to address are money laundering and fake credit card applications.

"

> **Databricks has changed the way we do business. It has put us in a better position to succeed in our data and AI transformation as a company by enabling data professionals with advanced data capabilities in a controlled and scalable way.**

Stefan Groot
Head of Analytics Engineering,
ABN AMRO

SOLUTION AND BENEFITS

Today, Azure Databricks empowers ABN AMRO to democratize data and AI for a team of 500+ engineers, scientists and analysts who work collaboratively on improving business operations and introducing new go-to-market capabilities across the company.

- Delta Lake enables fast and reliable data pipelines to feed accurate and complete data for downstream analytics
- Integration with Power BI enables easy SQL analytics and feeds insights to 500+ business users through reports and dashboards
- MLflow speeds deployment of new models that improve the customer experience — with new use cases delivered in under two months

10x faster
time to market — use cases deployed in two months

100+
use cases to be delivered over the coming year

500+
empowered business and IT users

LEARN MORE ›

SEGA Europe

SEGA® Europe, the worldwide leader in interactive entertainment, is using the Databricks Lakehouse Platform to personalize the player experience and build its own machine learning algorithm to help target and tailor games for over 30 million of its customers.

> **Improving the player experience is at the heart of everything we do, and we very much see Databricks as a key partner, supporting us to drive forward the next generation of community gaming.**
>
> **Felix Baker**
> Data Services Manager, SEGA Europe

As housebound gamers looked to pass the time during the first lockdowns of 2020, some SEGA Europe titles, including Football Manager,™ saw over double the number of sales during the first lockdown compared to the year before. Furthermore, a number of SEGA titles experienced a more than 50% increase in players over the course of the COVID-19 pandemic. With more anonymized data being collected through an analytics pipeline than ever before, the team needed a dedicated computing resource to handle the sheer volume of data, extract meaningful insights from it and enable the data science team to improve general workflow.

LEARN MORE ›

About Databricks

Databricks is the lakehouse company. More than 7,000 organizations worldwide — including Comcast, Condé Nast and over 50% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark,™ Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.

START YOUR FREE TRIAL

Contact us for a personalized demo
databricks.com/contact

DISCOVER LAKEHOUSE

databricks