



Kristin Dahl, Solutions Architect

Roberto Salcido, Solutions Architect

October 13, 2021

Housekeeping

- Questions encouraged!
- This is meant to be a discussion
- If we do not answer your question during the call, we will follow-up with you afterwards to get you the information you need!
- Participate in the polls!

Agenda

- 11:00 AM PT | **Introduction & Databricks Delta SQL Overview**
- 11:30 AM PT | **Delta Streaming Pipeline**
- 12:00 AM PT | **SQL Analytics**
- 12:30 AM PT | **Event concludes**

Meet the Presenters



Kristin Dahl
Senior Solution Engineer
kristin@databricks.com



Roberto Salcido
Solutions Architect
roberto.salcido@databricks.com



Databricks SQL: Public Preview Available!

We are pleased to announce that [Databricks SQL](#) is now available to all Databricks customers using **premium** and **enterprise** workspaces without the need to request access.

Databricks SQL allows users to operate a multi-cloud lakehouse architecture that provides data warehousing performance at data lake economics with up to 6x better price/performance than traditional cloud data warehouses. This enables you to not only simplify your architecture, but also perform BI directly on the freshest and most complete data.

For more information, please visit <https://databricks.com/product/sql-analytics>

New Blog Series: Databricks SQL

Databricks is launching a new blog series on performance improvements and enhancements in Databricks SQL

<https://databricks.com/blog/2021/09/08/new-performance-improvements-in-databricks-sql.html>

New Performance Improvements in Databricks SQL



by Reynold Xin, Can Efeoglu, Cyrielle Simeone and Bilal Aslam
September 8, 2021

Originally **announced** at Data + AI Summit 2020 Europe, **Databricks SQL** lets you operate a multi-cloud **lakehouse** architecture that provides data warehousing performance at data lake economics. Our vision is to give data analysts a simple yet delightful tool for obtaining and sharing insights from their lakehouse using a purpose-built SQL UI and world-class support for popular BI tools.

This blog is the first of a series on Databricks SQL that aims at covering the innovations we constantly bring to achieve this vision: performance, ease of use and governance. This blog will cover recent performance optimizations as part of Databricks SQL for:

Databricks Community (community.databricks.com)

The screenshot shows the Databricks Community homepage. At the top, there's a navigation bar with links for Support Case, Help Center, Documentation, Knowledge Base, and Training. Below the navigation is a search bar and a login button. The main content area features a "Welcome to the Databricks Community" section with a message: "Get answers, network with peers, and let's solve the world's toughest problems, together." It includes a "How can we help you today?" dropdown and a "Ask a question" button. To the right is a decorative graphic of hexagonal portraits of community members. Below this is a "Browse Topics" section with categories like Administration, Data Engineering, Machine Learning, Data Science, Business Analytics, General, and Recent Discussions. Each category has a list of sub-topics and their counts. To the right of the browse topics is a "Top Contributors" section showing a list of users with their names, profile pictures, and post counts. Below that is a "Popular topics" section with a grid of tags and their counts.

Now live!

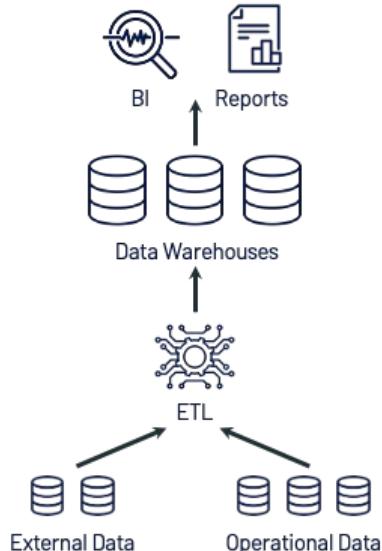
- Strong **cross-functional participation** across Databricks means expert advice available at scale
- **Federated search** across Documentation, Knowledge Base, Community posts + Academy (coming on roadmap!)
- Seamless account provisioning from workspaces, starting for AWS
- **Organized Topics** by use cases and features with participation from Databricks SMEs, partners and savvy practitioners
- Gamified Top Contributors to **celebrate and motivate** participation with exportable badges and recognitions (coming on roadmap!)
- Public and private **Groups with peers** for related conversations and deeper relationships
- Other than in private groups, all content will be public and indexed in search engines to **improvement discoverable for users**

Evolution of Data Management



purpose-built for BI and reporting, however...

Data Warehouse



- No support for video, audio, text
- No support for data science, ML
- Limited support for streaming
- Closed & proprietary formats

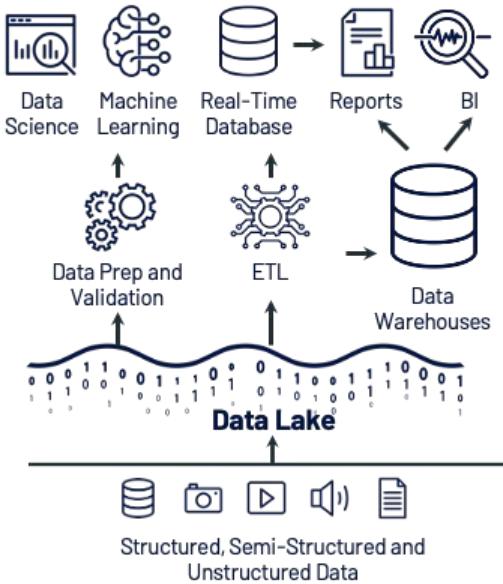
Therefore, most data is stored in data lakes & blob stores

could handle all your data for data science and ML, however...

Data Warehouse



Data Lakes



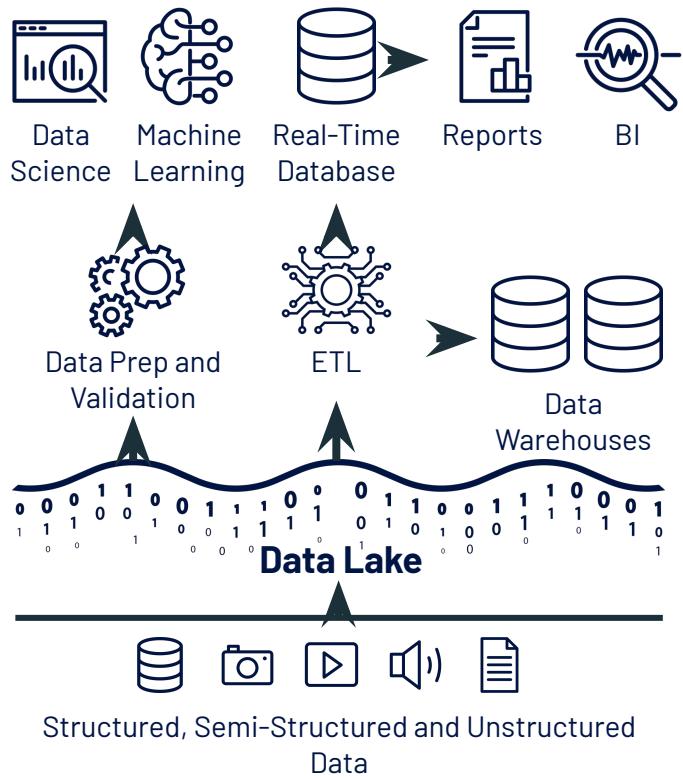
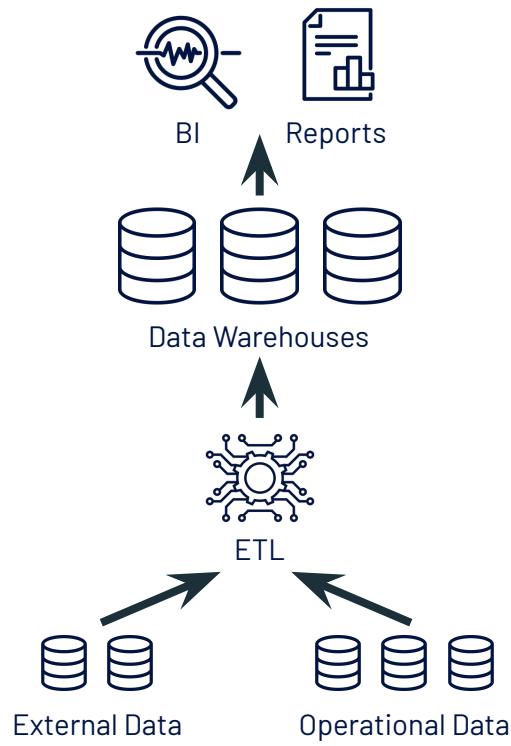
- Poor BI support
- Complex to set up
- Poor performance
- Unreliable data swamps

Let's summarize our data challenges

With traditional Data Warehouses and Data Lakes we still lack:

- Reliability (lack of ACID, pipeline fragmentation for batch + stream)
- Performance (small files problem, object store limitations, etc.)
- Support for Data Science, Machine Learning & BI under one roof
- Openness (vs proprietary software and tooling)

Coexistence is not a desirable strategy



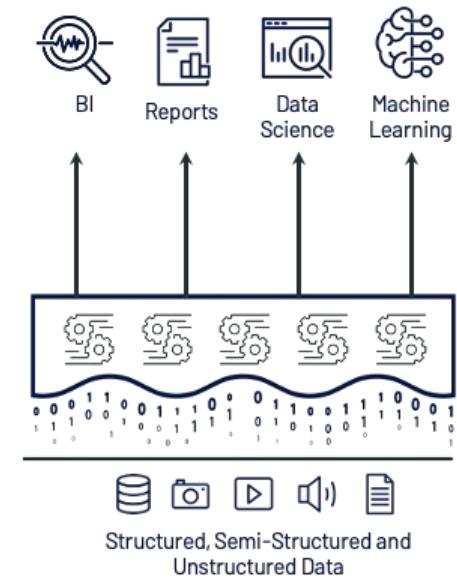
Data Warehouse



Data Lakes



Lakehouse



Data
Lake



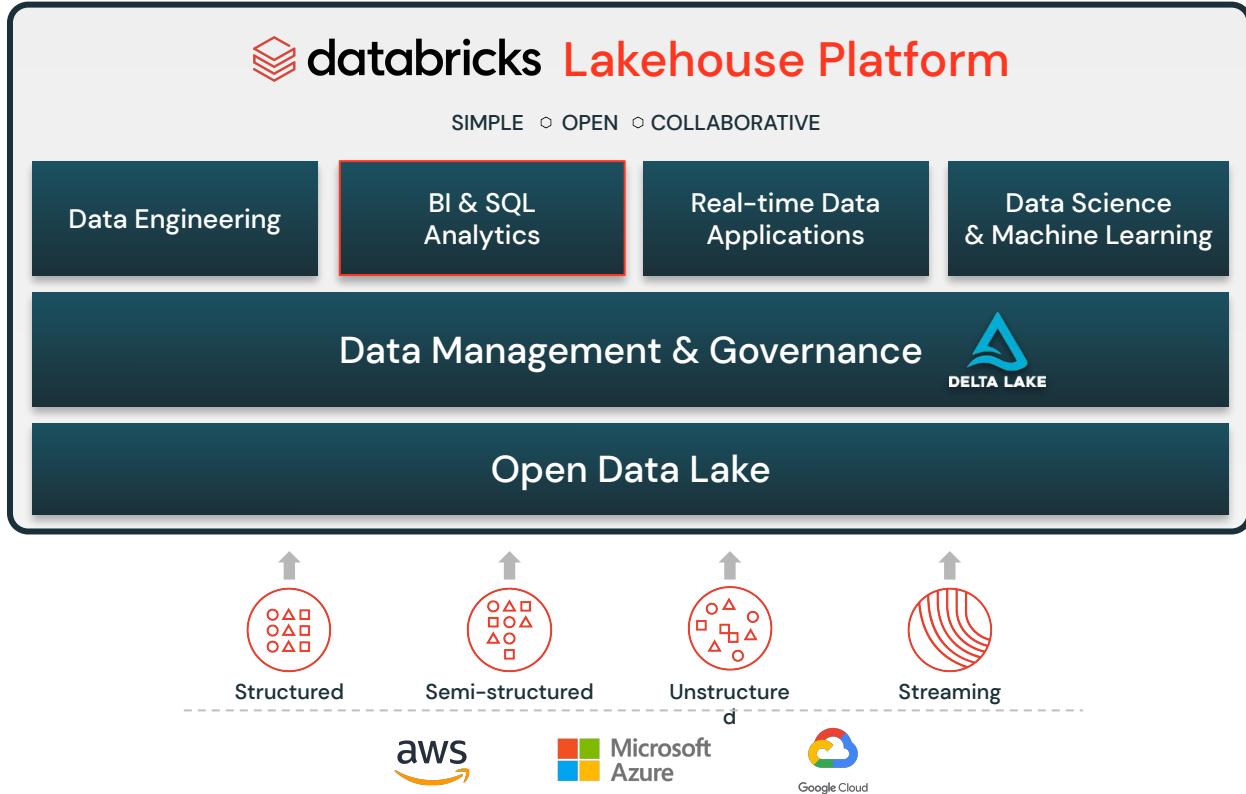
Lakehouse

One platform to unify all of
your data, analytics, and AI workloads

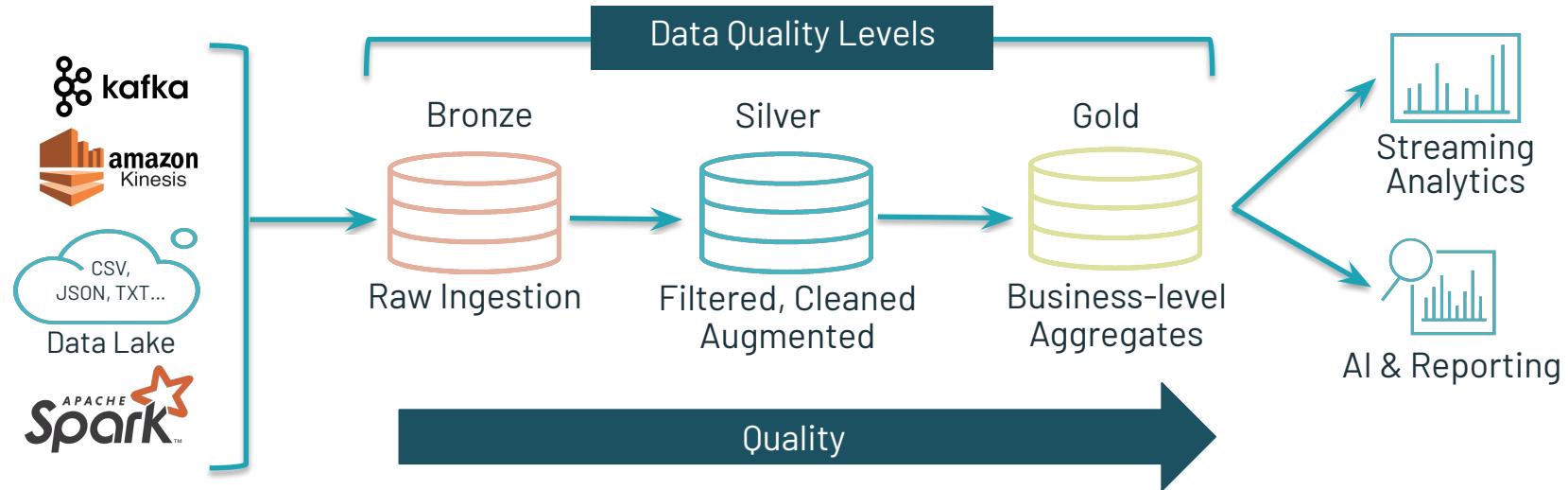
Data
Warehouse



Data Lakehouse is the foundation for BI & SQL Analytics

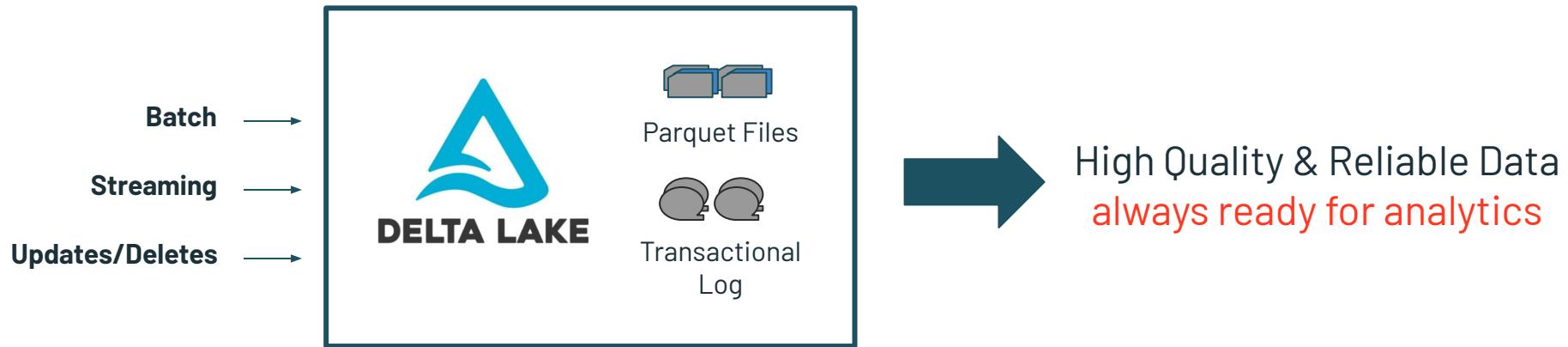


The Delta Lake



Delta Lake allows you to *incrementally* improve the quality of your data until it is ready for consumption.

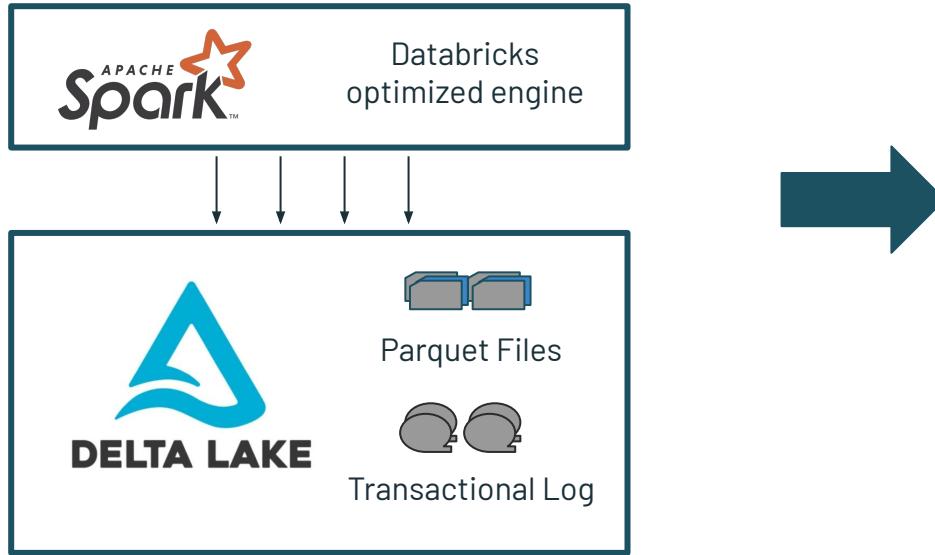
Delta Lake ensures data reliability



Key Features

- ACID Transactions
- Schema Enforcement
- Unified Batch & Streaming
- Time Travel/Data Snapshots

Delta Lake optimizes performance



Highly Performant
queries at scale

Automatically managed
through



Key Features

- Indexing
- Compaction
- Data skipping
- Caching

What is your architecture today?

- Data warehouse centric
- Data lake centric
- Lakehouse centric
- N/A

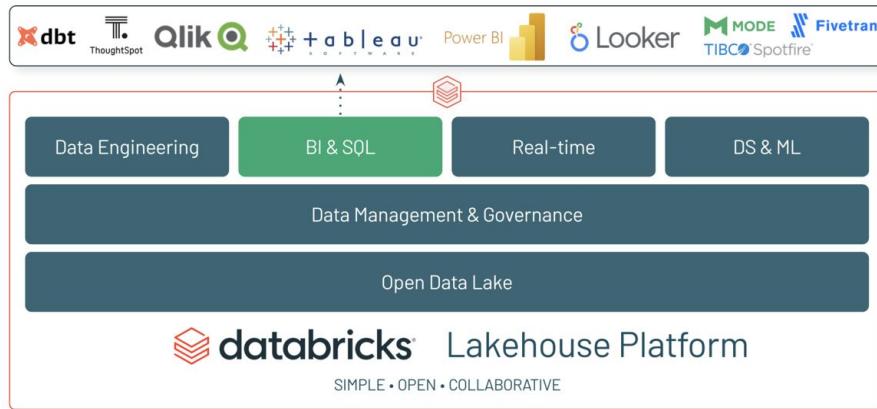
Databricks SQL Overview



Databricks SQL

Analytics on the latest and most complete data with data warehouse performance and data lake economics

- Fast and reliable performance for all queries
- Simplified administration and governance
- Analytics on all your data with your tools of choice



Use Cases

**Query data lake data
with your existing BI tools**



"Databricks is core to our business because its lakehouse architecture provides us a unified way to access, store and share actionable data. It's going to be the core of our analytics strategy going forward. No question about that."

— Jagan Mangalampalli, Director of Big Data,
Punchh

**Enable governed self-served
analytics for everyone**



"As a company focused on providing data-driven research to our customers, the massive amount of data in our data lake is our lifeblood. By leveraging Databricks and Delta Lake, we have already been able to democratize data at scale, while lowering the cost of running production workloads by 60%, saving us millions of dollars. We're excited to build on this momentum by leveraging the Databricks lakehouse architecture that will further empower everyone across our organization – from research analysts to data scientists – to interchangeably use the same data, helping us to provide innovative insights to our customers faster than ever before."

— Steve Pulec, Chief Technology Officer, YipitData

**Build data-enhanced
applications**



"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



LEARN MORE:

<https://databricks.com/product/databricks-sql>

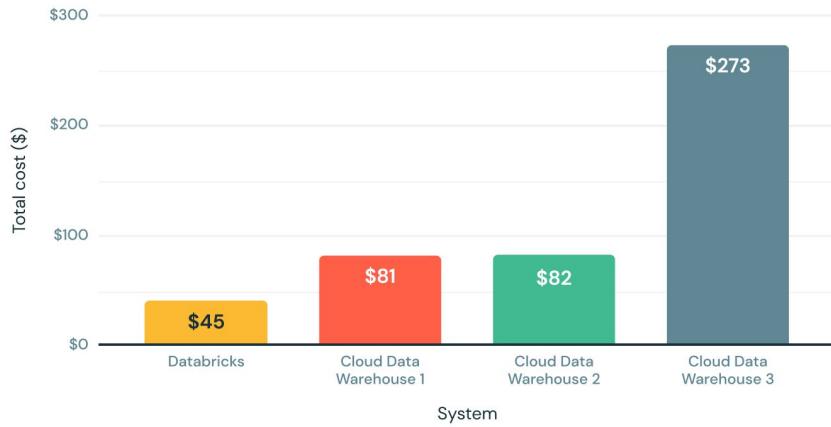
Fast and predictable
performance for all queries.

Better price / performance

Query and analyze your most complete and freshest data with **up to 6x better price/performance** than legacy cloud data warehouses.

30TB TPC-DS Price/Performance

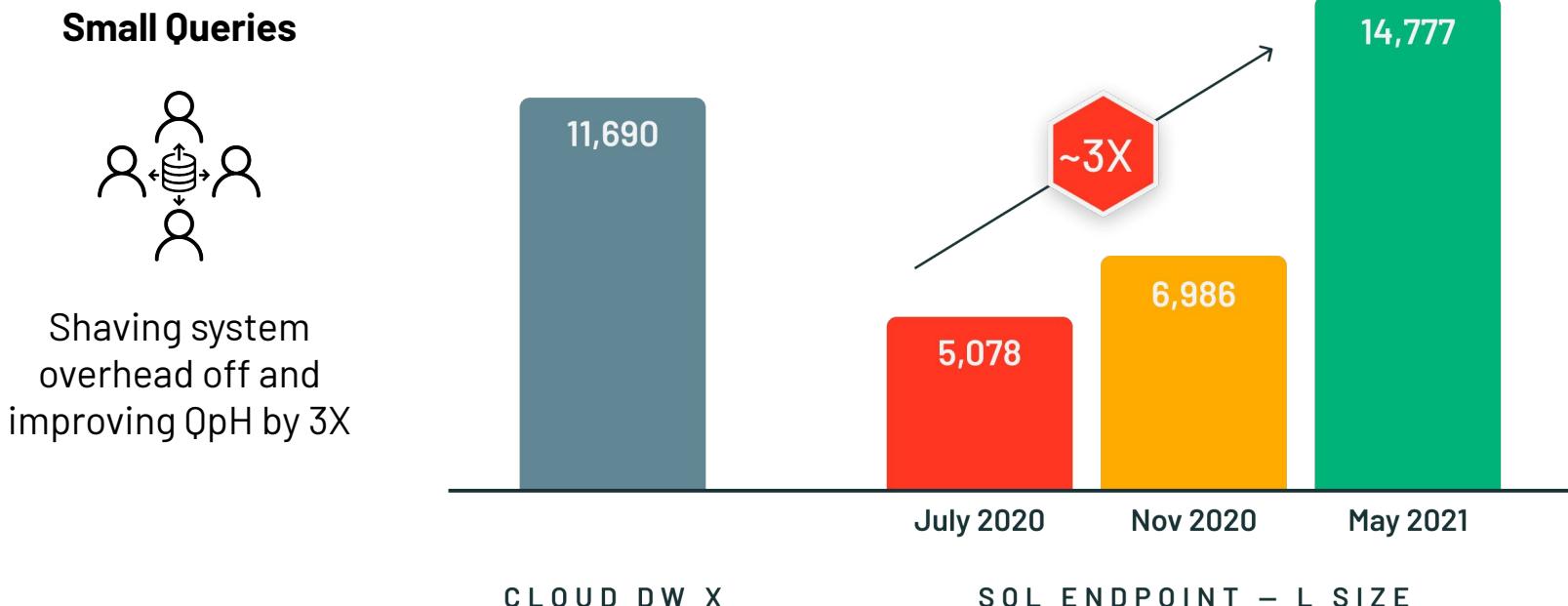
Lower is better



Source: Performance Benchmark with Barcelona Supercomputing Center

Fast and predictable performance for all queries

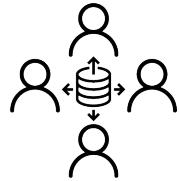
Beyond large query performance



Fast and predictable performance for all queries

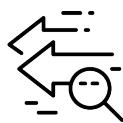
Beyond large query performance

Small Queries



Shaving system overhead off and improving QpH by 3X

Mixed Workloads

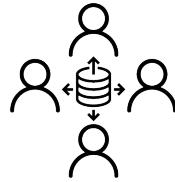


Dual queue to **avoid large queries blocking small ones**

Fast and predictable performance for all queries

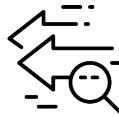
Beyond large query performance

Small Queries



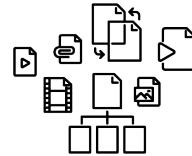
Shaving system overhead off and improving QpH by 3X

Mixed Workloads



Dual queue to **avoid large queries blocking small ones**

I/O Performance

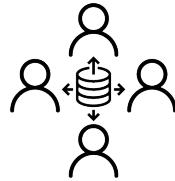


Async I/O for faster **small files** and **cold read** cases (~12x)

Fast and predictable performance for all queries

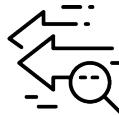
Beyond large query performance

Small Queries



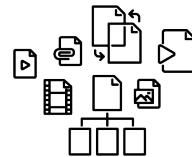
Shaving system overhead off and improving QpH by 3X

Mixed Workloads



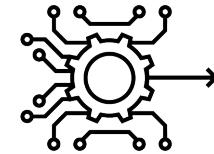
Dual queue to **avoid large queries blocking small ones**

I/O Performance



Async I/O for faster **small files** and **cold read** cases (~12x)

BI Results Retrieval



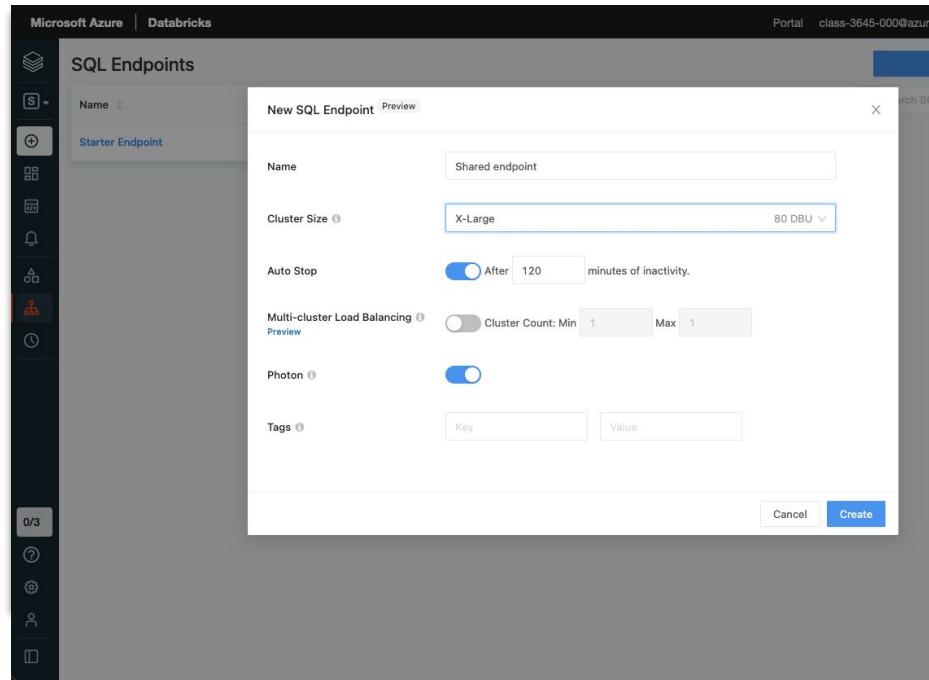
Cloud fetch to improve fetch throughput (10x)

Simplified administration and governance



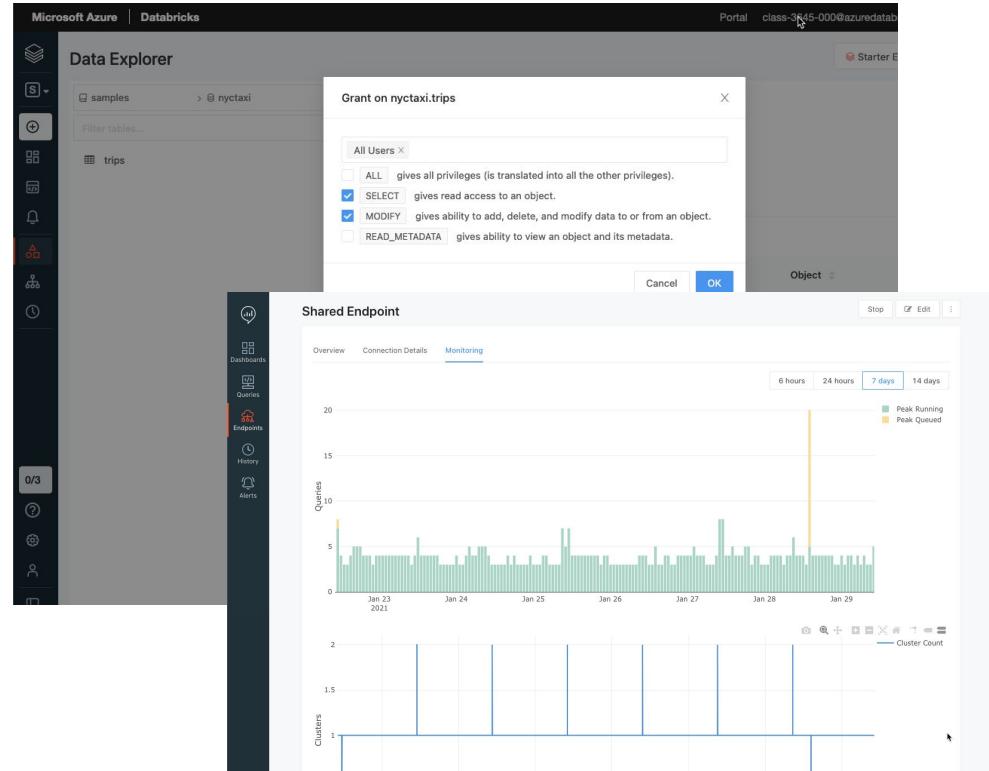
Get started quick with "No resource" management

- Quickly setup optimized compute resources with Databricks SQL endpoints (Powered by Photon)
- Easily and automatically scale to your needs with elasticity and concurrency scaling built-in
- Coming soon in Serverless !



Easy to govern self-served analytics

- Confidently onboard new users, discover, secure, and govern data in one place
- Manage costs and usage effectively with endpoints monitoring and query history
- Meet compliance needs with built-in audit trail



Easily troubleshoot and optimize queries

- Get full visibility into query execution with in-depth breakdown at operation levels
- Identify bottlenecks and expensive operations to optimize queries
- Analyze by query duration, data volumes & resource usage

The screenshot shows the Databricks Query History interface. On the left is a sidebar with icons for Create, Queries, Dashboards, Alerts, Endpoints, and History. The main area is titled "Query History" and displays a list of completed queries. One query is expanded to show its execution details. The "Execution Summary" tab is selected, showing a tree view of the query plan with metrics for each step. The plan includes operations like TakeOrderedAndProject, HashAggregate, CustomShuffleReader, Exchange, HashAggregate, Union, Project, BroadcastHashJoin, Project, Filter, ColumnarToRow, DataSourceScan, and Project. A tooltip over the DataSourceScan step provides specific performance metrics: duration 2.1min, volume: 200 rows | 10MB, and memory used: 15MB.

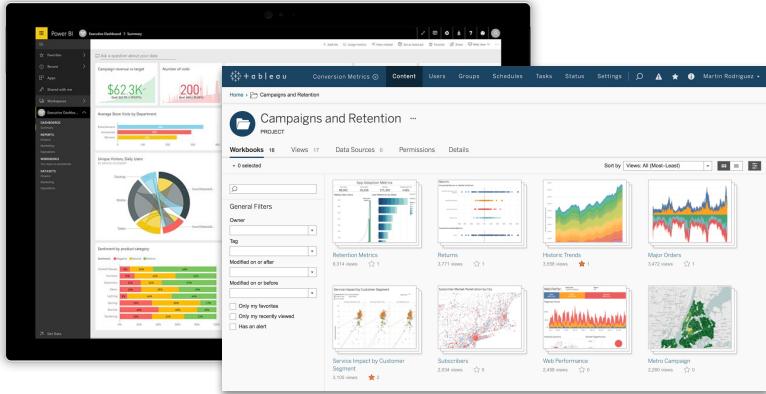
Step	Operation	Metric	Value
0	TakeOrderedAndProject	-	
1	HashAggregate	-	
2	CustomShuffleReader	-	
3	Exchange	-	
4	HashAggregate	-	
5	Union	-	
6	Project	-	
7	BroadcastHashJoin	-	
8	Project	-	
9	Filter	-	
10	ColumnarToRow	-	
11	DataSourceScan	duration	2.1min
11	DataSourceScan	volume	200 rows 10MB
11	DataSourceScan	memory used	15MB
12	Project	-	
13	ColumnarToRow	-	
14	LocalTableScan	-	
15	Project	-	
16	Filter	-	
17	ColumnarToRow	-	
18	RDDScan	-	

Analytics on all your data with your tools of choice



A platform for your BI tools of choice

Tap into one source of truth for all your data. Databricks SQL also provides fast performance, low latency, and high user concurrency for **your existing BI tools.**



Coming soon:



First-class SQL development experience

- Quickly **perform exploratory analysis** with the built-in SQL query editor, visualizations and dashboards.
- Trigger alerts for critical changes, and respond to business needs faster.

The screenshot shows the Databricks SQL interface. On the left is a dark sidebar with navigation icons for Create, Queries, Dashboards, Alerts, Endpoints, and History. The main area has a title bar "New Query" with tabs for "SQL", "Data", and "Notebook". A dropdown menu shows a selected item "19mar21-test2". Below it is a dropdown for "default". A "Filter tables & columns..." input field is present. To the right is a code editor with the following SQL query:

```
1 SELECT |
```

Below the code editor are buttons for "{{}}", "SQL", "LIMIT 1000", and a "Revert" button. The results section shows a "Table" icon and a message "Query has no result". It includes a "Refresh Now" button and a note "Execute/Refresh the query to show results." At the bottom, a status message says "Schema last fetched: 11 hours ago".

Delta & SQL Analytics - Lab Prep

Start your clusters now!

[How to Import a Notebook](#)

[Notebook Link](#)

[SQLA Documentation](#)

[New JSON SQL](#)

[Operators](#)

[Persona QuickStarts](#)

[Billable Usage Delivery](#)

[Audit Log Delivery](#)

[OverWatch](#)

If you have any questions or need assistance getting features turned on please contact the SA's giving this workshop

