



Kristin Dahl, Senior Solution Engineer

September 15, 2021

# 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:

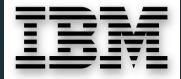
# Databricks Community ([community.databricks.com](https://community.databricks.com))

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**

# Meet the Presenter



Kristin Dahl  
Senior Solution Engineer  
[kristin@databricks.com](mailto:kristin@databricks.com)



# Agenda

1. Lakehouse Design

2. Delta Streaming Pipeline

3. SQL Analytics

# Let's summarize our data challenges

With traditional Data Warehouses and Data Lakes we still lack:

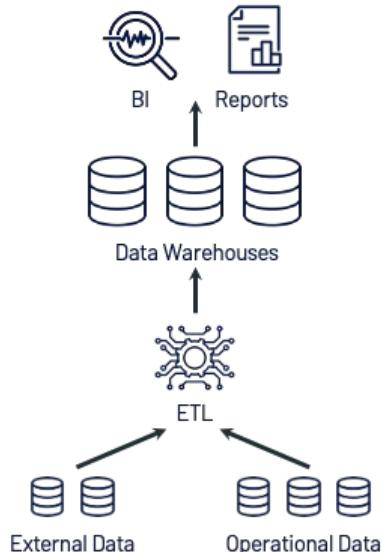
- Reliability (lack of ACID, pipeline fragmentation for batch + stream)
- Performance (small files problem, object store limitations, etc.)
- Support for Data Science, Machine Learning & BI under one roof
- Openness (vs proprietary software and tooling)

# 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

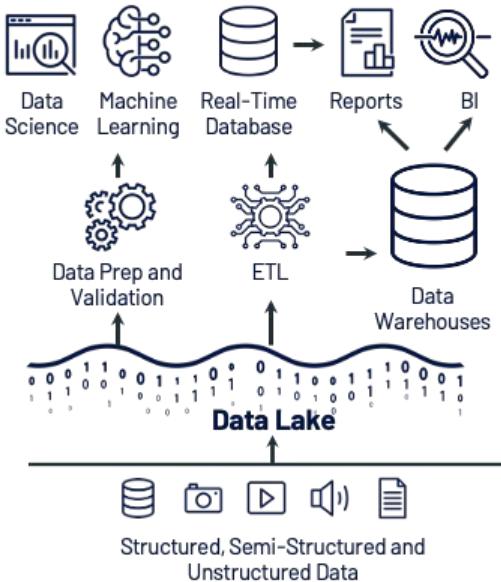
Therefore, most data is stored in data lakes & blob stores

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

## Data Warehouse



## Data Lakes



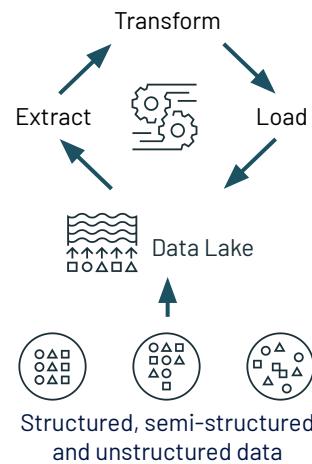
- Poor BI support
- Complex to set up
- Poor performance
- Unreliable data swamps

# Today, most enterprises struggle with data

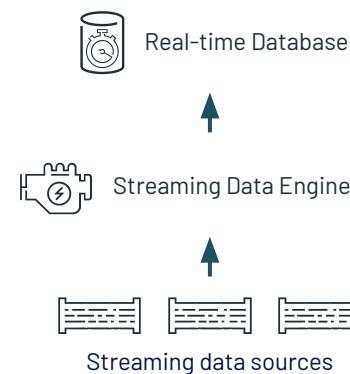
## Data Warehousing



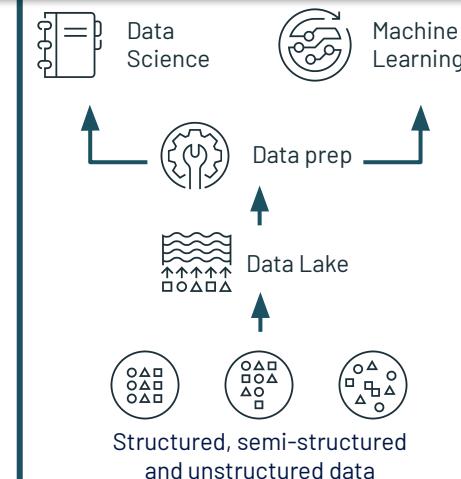
## Data Engineering



## Streaming



## Data Science & Machine Learning



Siloed stacks increase data architecture complexity

# Today, most enterprises struggle with data

## Data Warehousing

Amazon Redshift  
Azure Synapse  
Snowflake  
SAP  
Oracle Autonomous Data Warehouse

## Data Engineering

Teradata  
Google BigQuery  
IBM Db2  
Oracle Autonomous Data Warehouse  
Hadoop  
Amazon EMR  
Google Dataproc  
Apache Airflow  
Apache Spark  
Cloudera

## Streaming

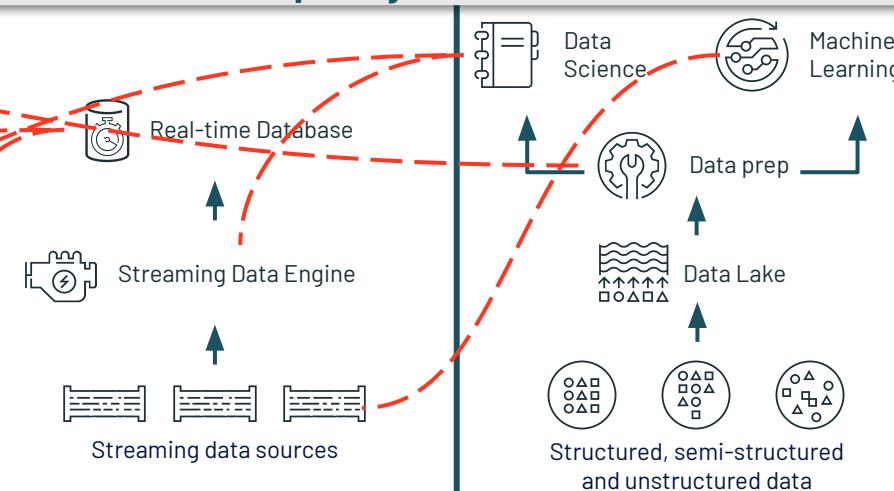
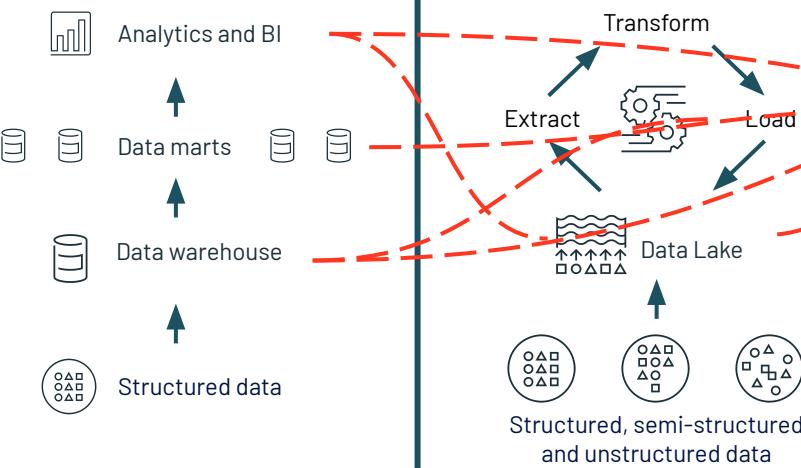
Apache Kafka  
Apache Flink  
Azure Stream Analytics  
Tibco Spotfire  
Apache Spark  
Amazon Kinesis  
Google Dataflow  
Confluent

## Data Science & Machine Learning

Jupyter  
Azure ML Studio  
Domino Data Labs  
TensorFlow  
Amazon SageMaker  
MatLAB  
SAS  
PyTorch

**Disconnected systems and proprietary data formats make integration difficult**

**Siloed stacks increase data architecture complexity**



# Today, most enterprises struggle with data

## Data Warehousing



Data Analysts



## Data Engineering



Data Engineers

## Streaming



Data Engineers

## Data Science & Machine Learning



Data Scientists

### Siloed data teams decrease productivity

Amazon Redshift

Azure Synapse

Snowflake

SAP

Teradata

Google BigQuery

IBM Db2

Oracle Autonomous Data Warehouse

Hadoop

Amazon EMR

Google Dataproc

Apache Airflow

Apache Spark

Cloudera

Apache Kafka

Apache Flink

Azure Stream Analytics

Tibco Spotfire

Apache Spark

Amazon Kinesis

Google Dataflow

Confluent

Jupyter

Azure ML Studio

Domino Data Labs

TensorFlow

Amazon SageMaker

MatLAB

SAS

PyTorch

### Disconnected systems and proprietary data formats make integration difficult

### Siloed stacks increase data architecture complexity



Analytics and BI



Data marts



Data warehouse



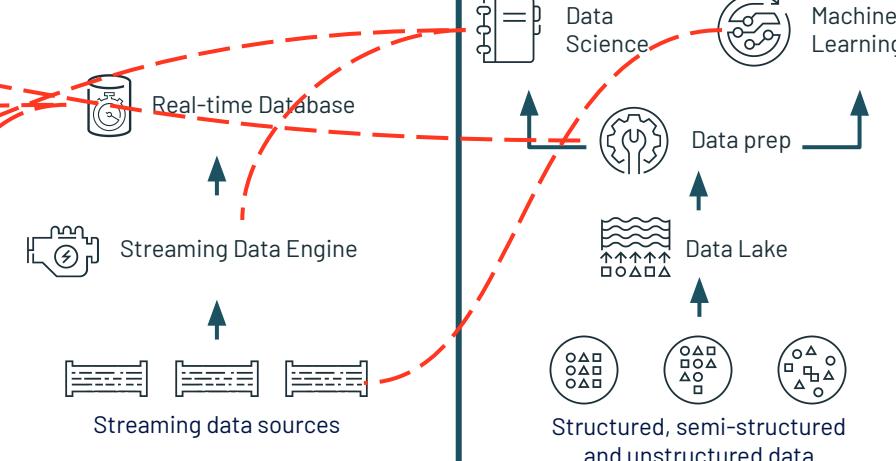
Structured data



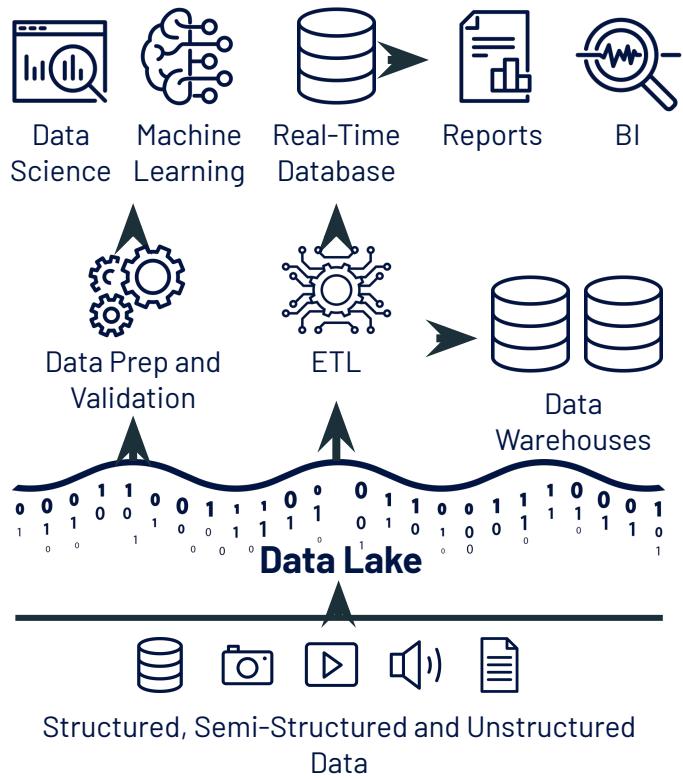
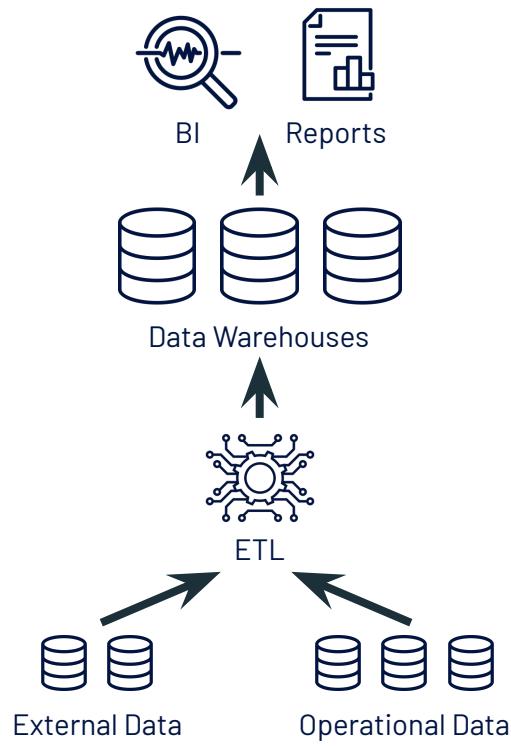
Data Lake



Structured, semi-structured and unstructured data



# Coexistence is not a desirable strategy



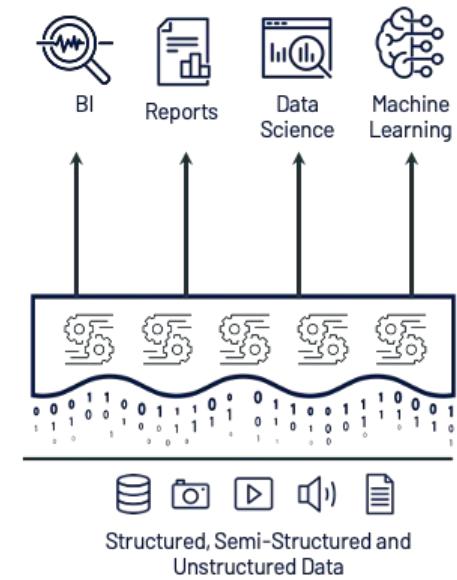
## Data Warehouse



## Data Lakes



## Lakehouse



# Lakehouse

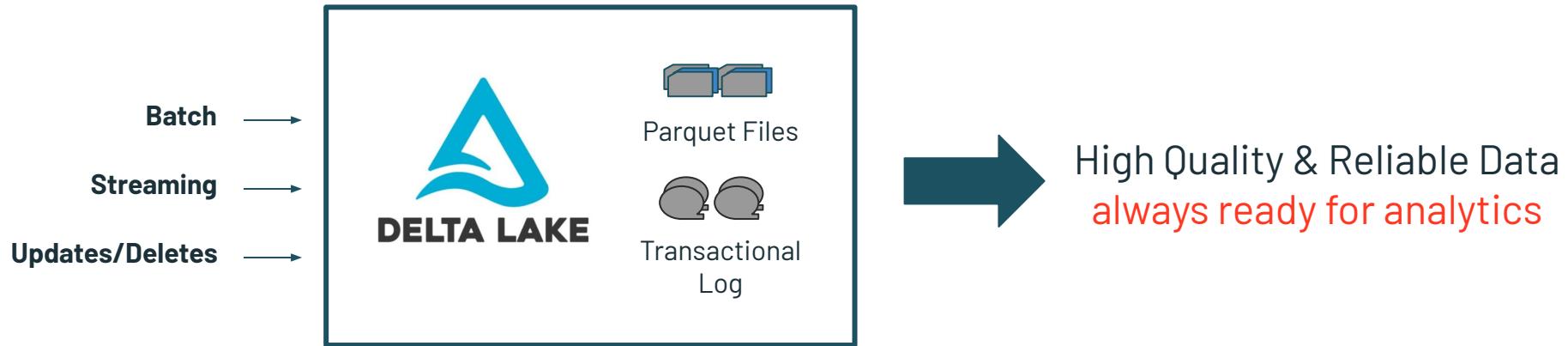


One platform for every use case

**Structured transactional layer**

Data Lake for all your data

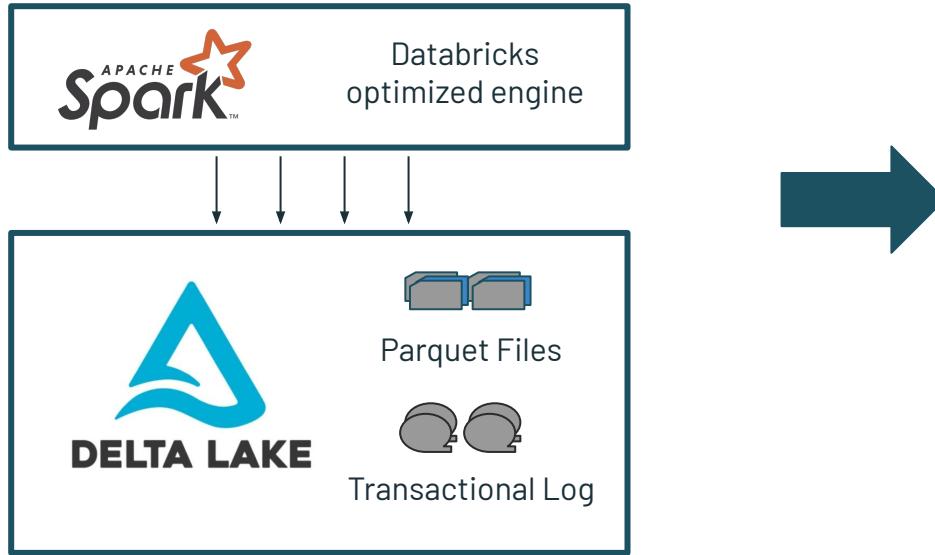
# Delta Lake ensures data reliability



## Key Features

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

# Delta Lake optimizes performance



Highly Performant  
queries at scale

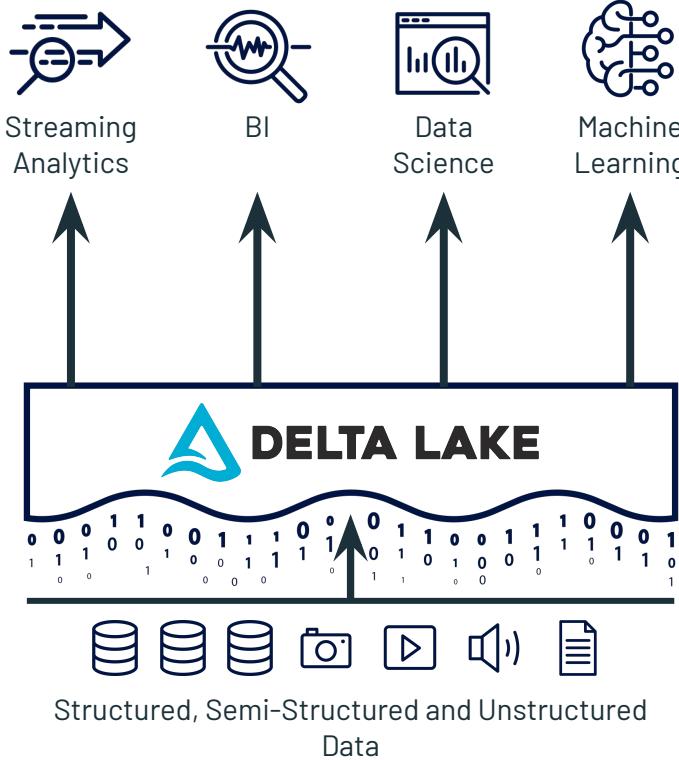
Automatically managed  
through



## Key Features

- Indexing
- Compaction
- Data skipping
- Caching

# Lakehouse



One platform for every use case

Structured transactional layer

Data Lake for all your data

# Lakehouse



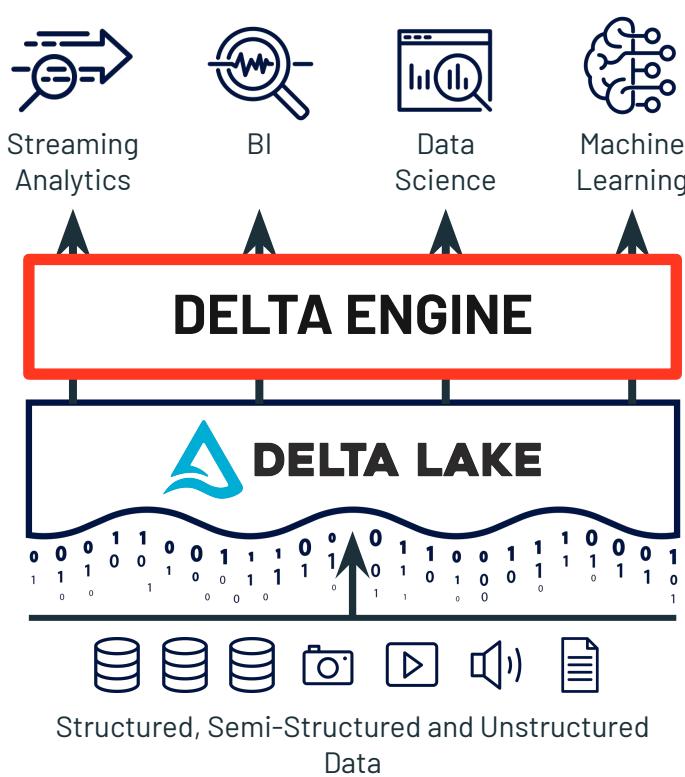
# One platform for every use case

# High performance query engine

## Structured transactional layer

## Data Lake for all your data

# Lakehouse with Databricks



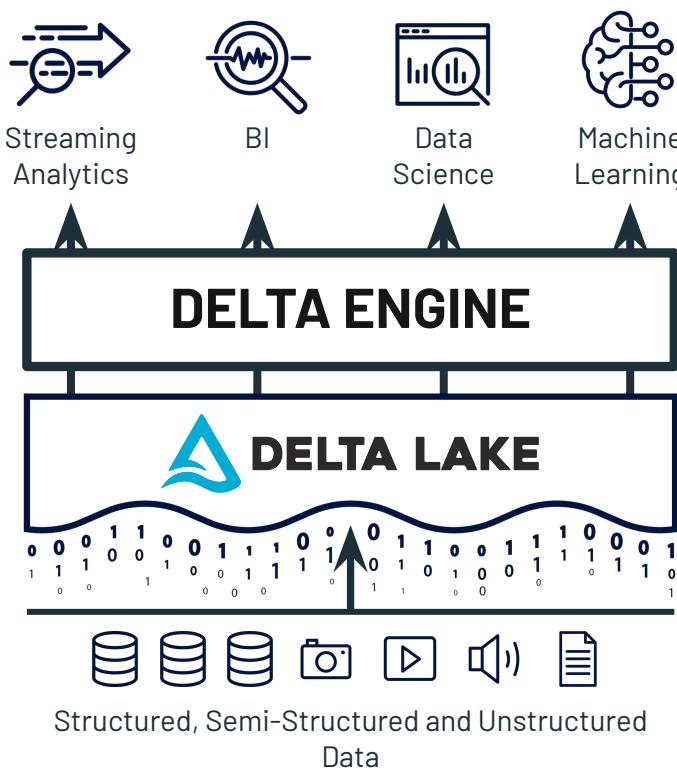
One platform for every use case

**High performance query engine**

Structured transactional layer

Data Lake for all your data

# Lakehouse with Databricks



One platform for every use case

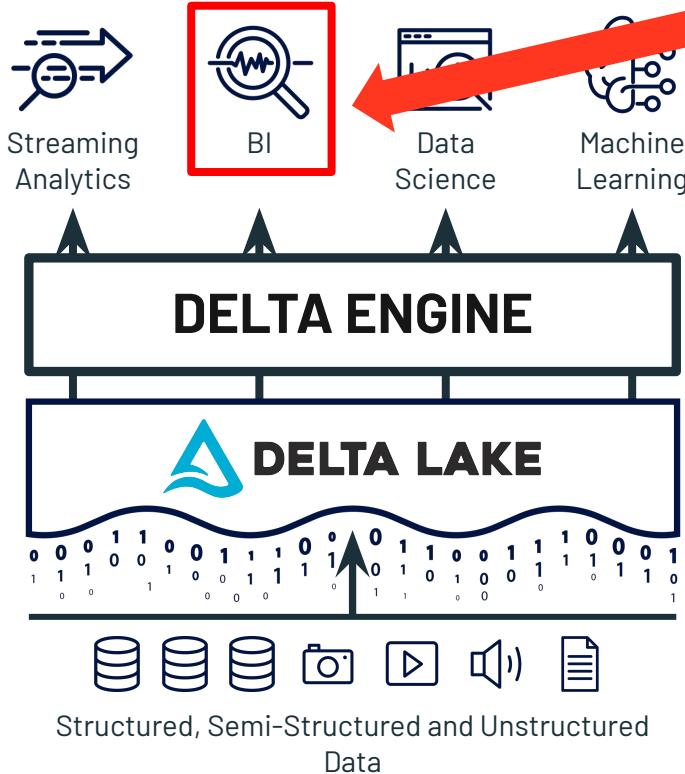
High performance query engine

Structured transactional layer

Data Lake for all your data

# Lakehouse with Databricks

In this session, we are going to focus on SQL Analytics.



One platform for every use case

High performance query engine

Structured transactional layer

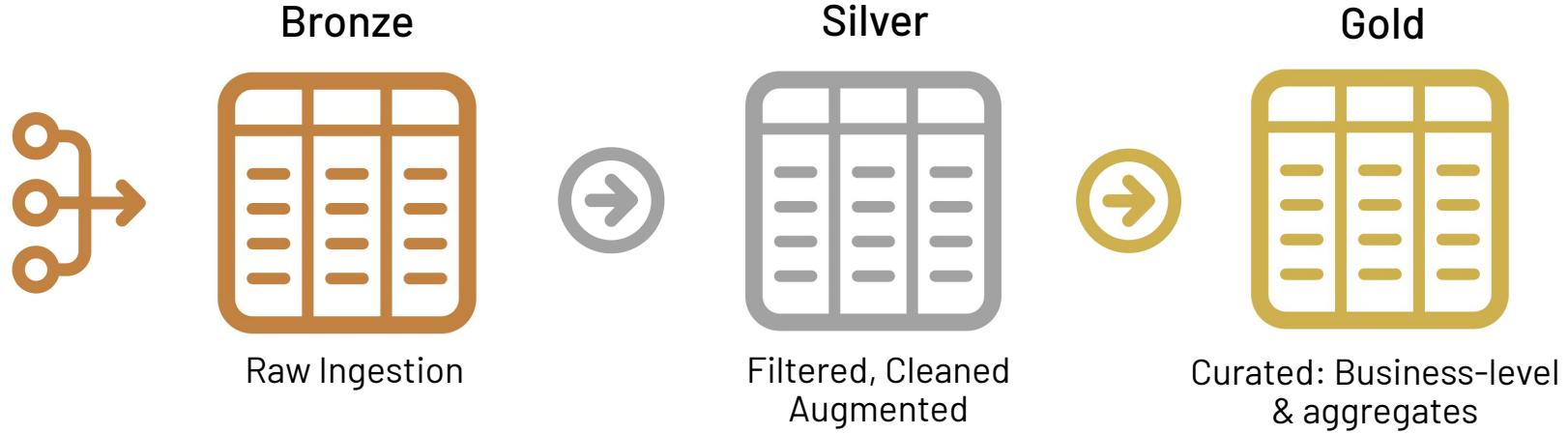
Data Lake for all your data

# What is your architecture today?

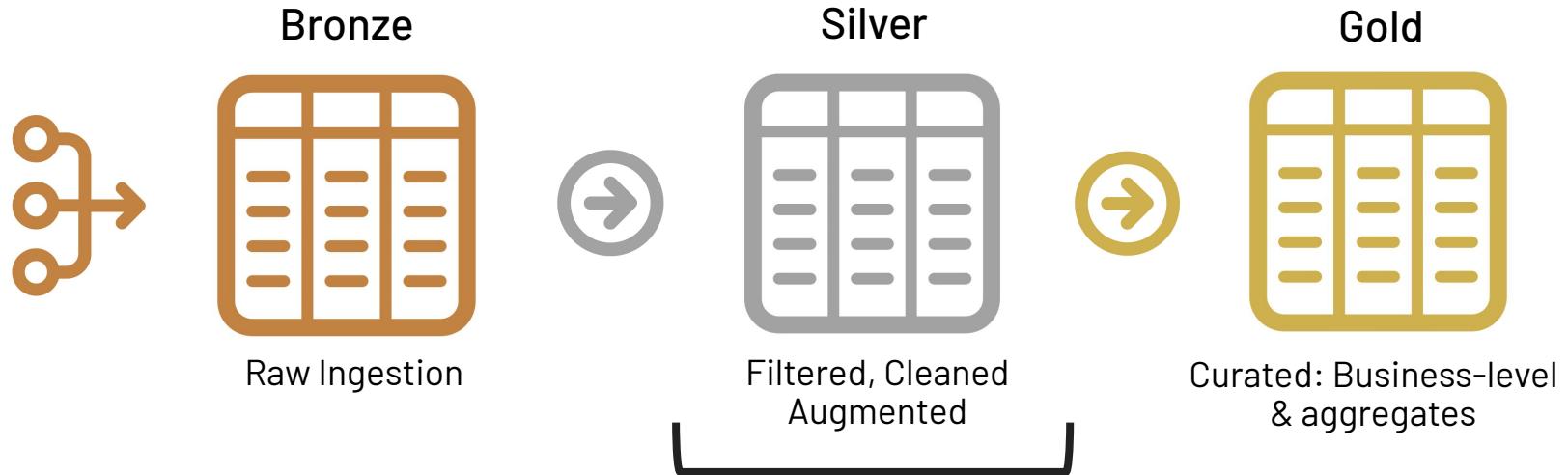
- Data warehouse centric
- Data lake centric
- Lakehouse centric
- N/A

*How do I get started?  
What do other folks do?*

# A few different approaches



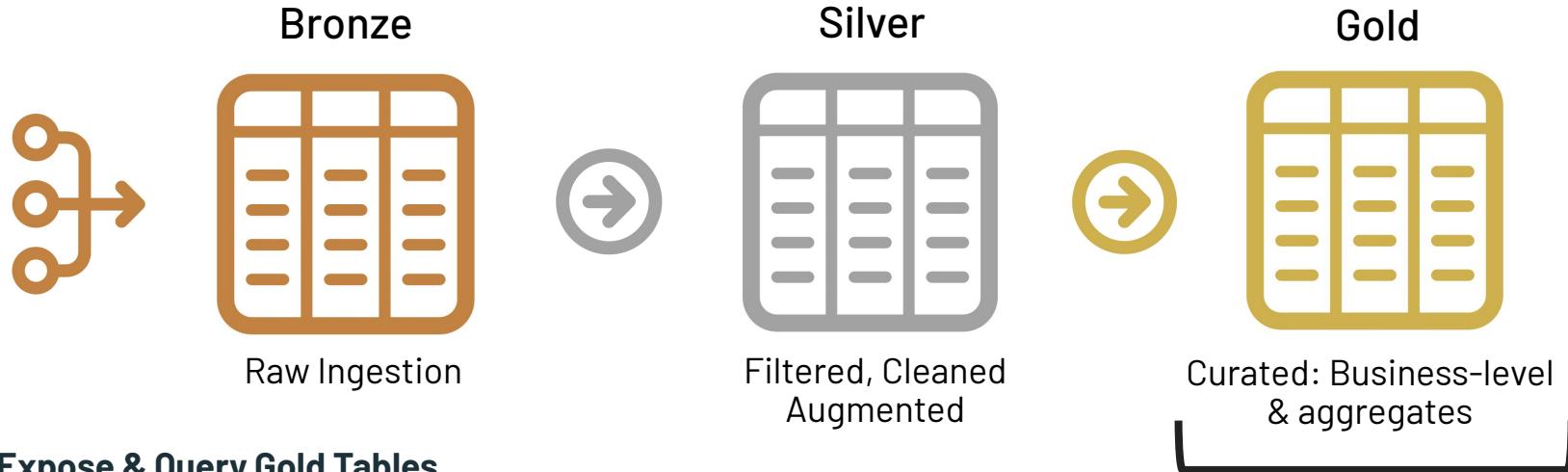
# A few different approaches



## Expose & Query Intermediate Tables

- Data engineering or “analytics engineering” does *\*some\** curation.
- Users are usually more advanced and are comfortable with SQL. So, they may create their own assets for reporting / BI, like aggregated or denormalized tables in a self-service manner.
- Provides high agility + opens up most of the data lake to users.

# A few different approaches



## Expose & Query Gold Tables

- Data engineering or “analytics engineering” curates & provides access to curated / gold level tables to the rest of the organization.
- Usually follows best-practices with proper modeling. (e.g. Kimball, denormalized reporting or mixed)
- Works well for less-technical users, as well as serving external users (e.g. companies selling data / insights)
- In this model, end-users typically do less self-service / “last-mile” ETL and rather rely on curated

# Product Lenses



A simpler, more contextual user experience for different types of users on Databricks via product



## Workspace

Data science and engineering



## SQL Analytics

SQL editor and dashboarding



Built on  
Redash!

# 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

# We're Hiring Solution Architects

If you or someone in your network is interested, we have East and West coast positions available.

Email [ericka.styles@databricks.com](mailto:ericka.styles@databricks.com) with your resume or LinkedIn. You can ask your presenters for more information.

[databricks.com/company/careers](https://databricks.com/company/careers)

Looking for applicants with backgrounds in any of these skills: data engineering, machine learning, **or** analytics.

Also hiring in Europe and Asia.



# SQL-native user interface for analysts

- Familiar SQL Editor
  - Auto Complete
  - Built in visualizations
  - Data Browser
- Automatic Alerts
  - Trigger based upon values
  - Email or Slack integration
- Dashboards
  - Simply convert queries to dashboards
  - Share with Access Control

