Agenda – 2 Days

Day 1

General

0900 – Welcome & Introduction
0910 – Databricks Overview
0930 – Lakehouse / Demo

Data Engineering

0950 – Data Engineering Overview
1020 – Lab Setup
1030 – Morning Tea
1045 – Data Engineering Lab
1300 – Lunch

Analytics/BI

1400 – Analytics/BI Overview
1420 – Databricks SQL Lab
1445 – Afternoon Tea
1500 – Databricks SQL Lab Cont'd
1530 – Databricks Roadmap

Day 2

Data Science

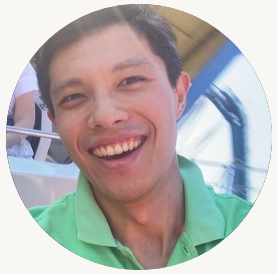
0950 – Databricks ML Overview
1020 – Lab Setup & Break
1030 – Databricks ML Lab

Wrap Up

1200 – Closing Comments
1210 – Q&A

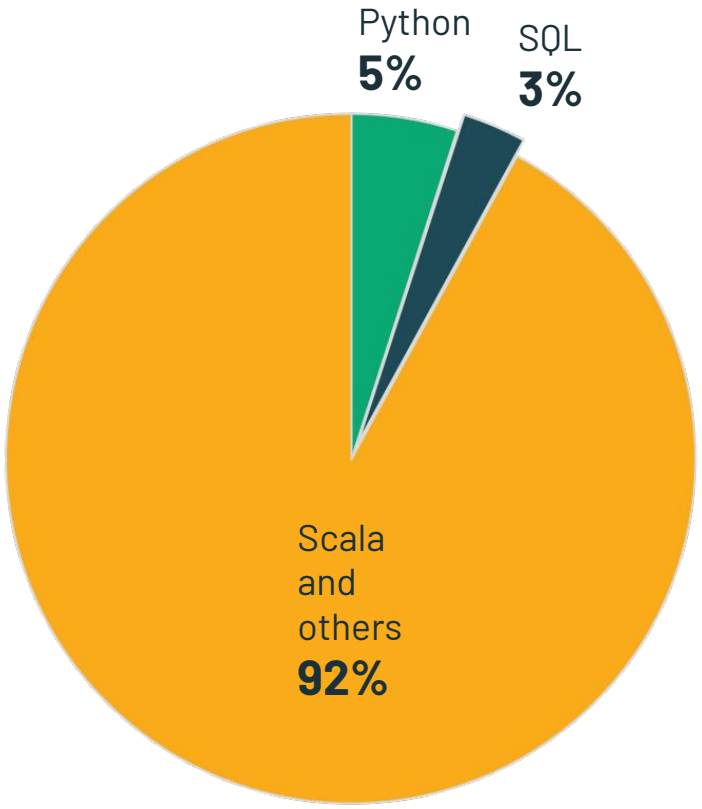
databricks

Databricks SQL Overview

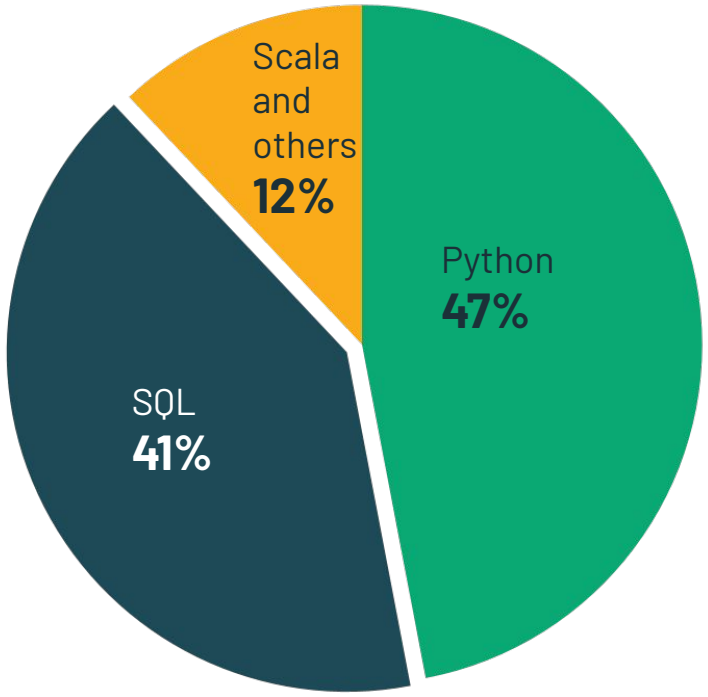


Spark usage among Databricks customers

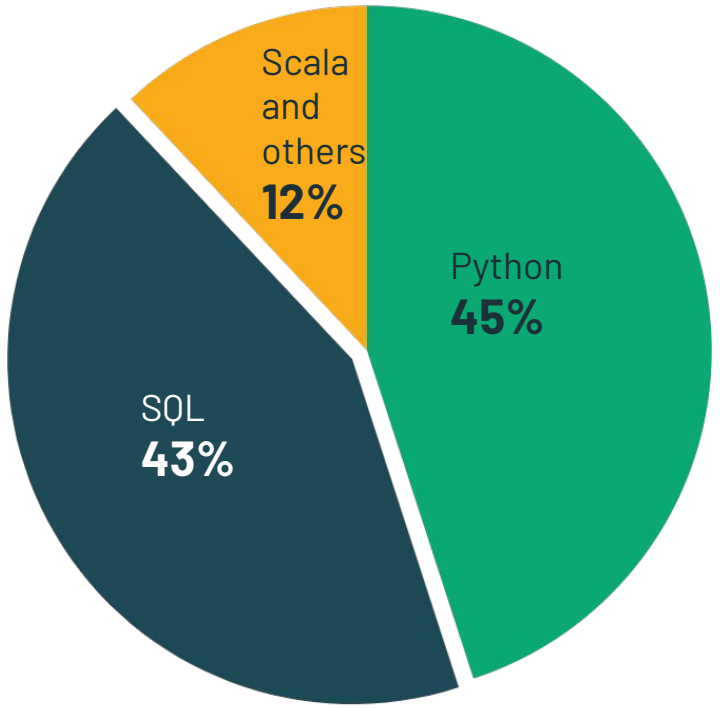
2013



2020



2021

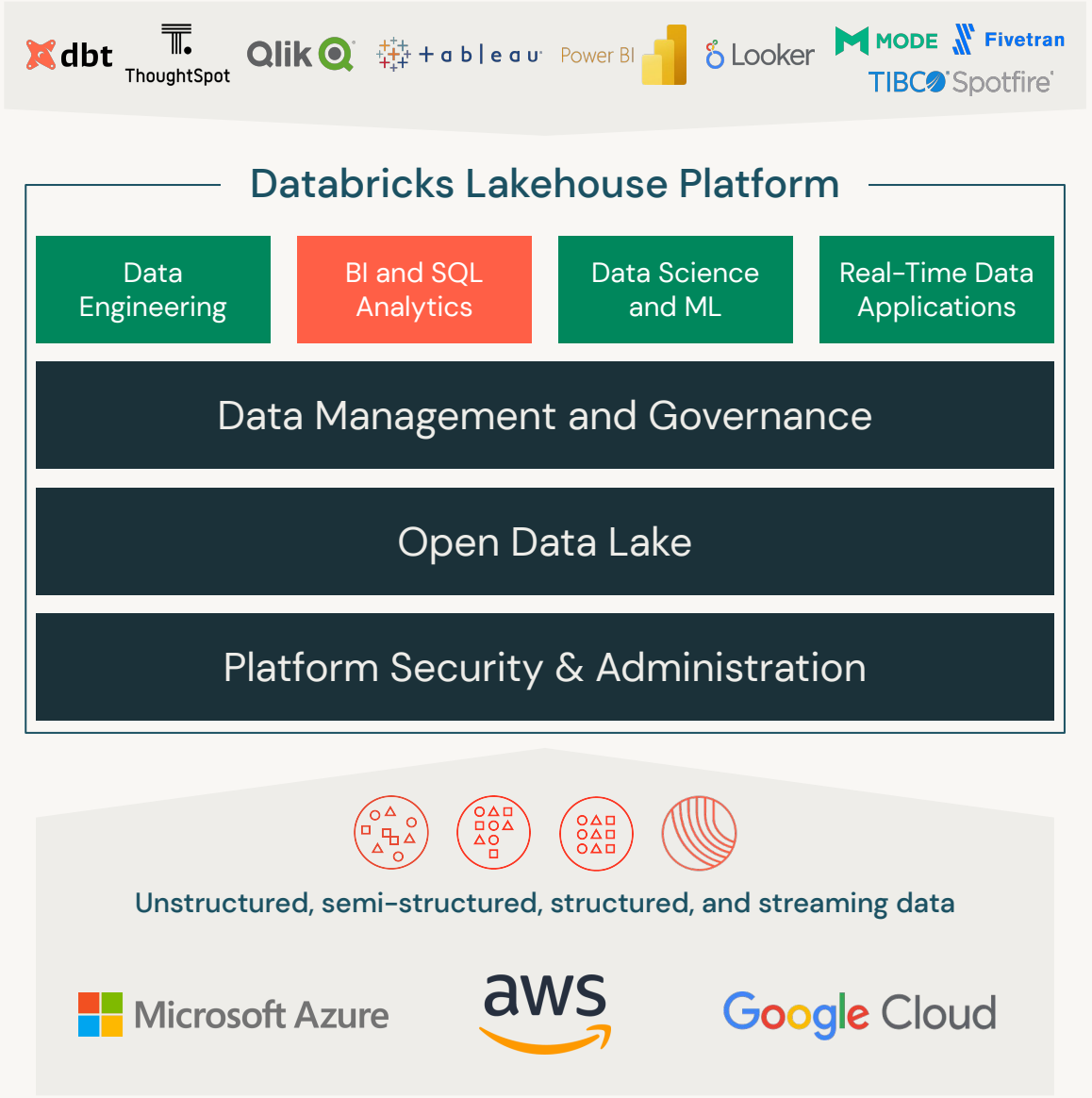


Databricks SQL

Analytics on the Lakehouse

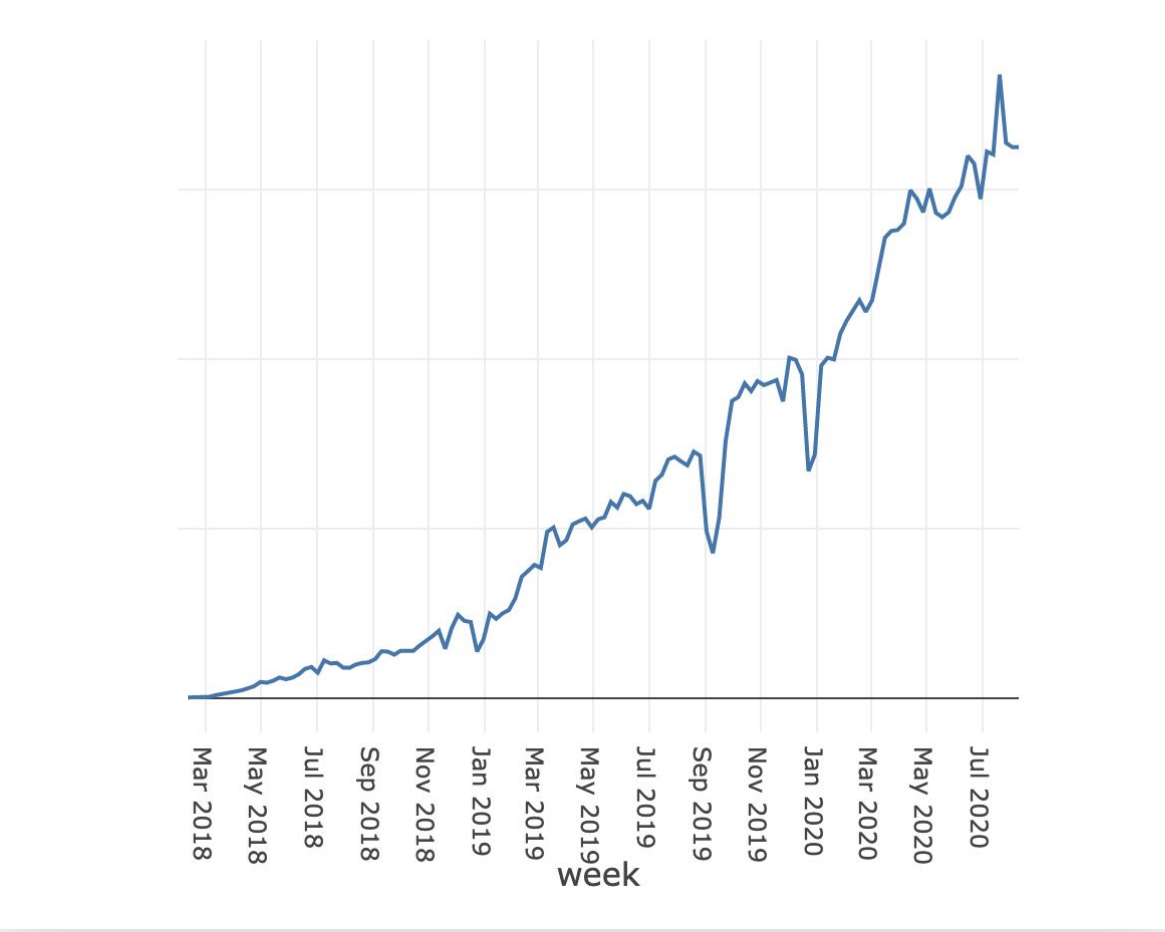
World-Class Performance at Data Lake Economics

- Fast and predictable performance for all queries
- Simplified administration and fine-grained governance
- Analytics on the freshest data with your tools of choice



Why building the Lakehouse ?

What we heard from our customers



SQL Usage Growth on Databricks (2018-2020)

1. Delta Lake enabled robust data management on the data lake.
2. Customers increasingly use SQL to directly query data lake data.
3. All the data is going to the lake. Only a portion gets into the data warehouse.

"We do not want another data warehouse"

– Our customers

Myth or Reality ?

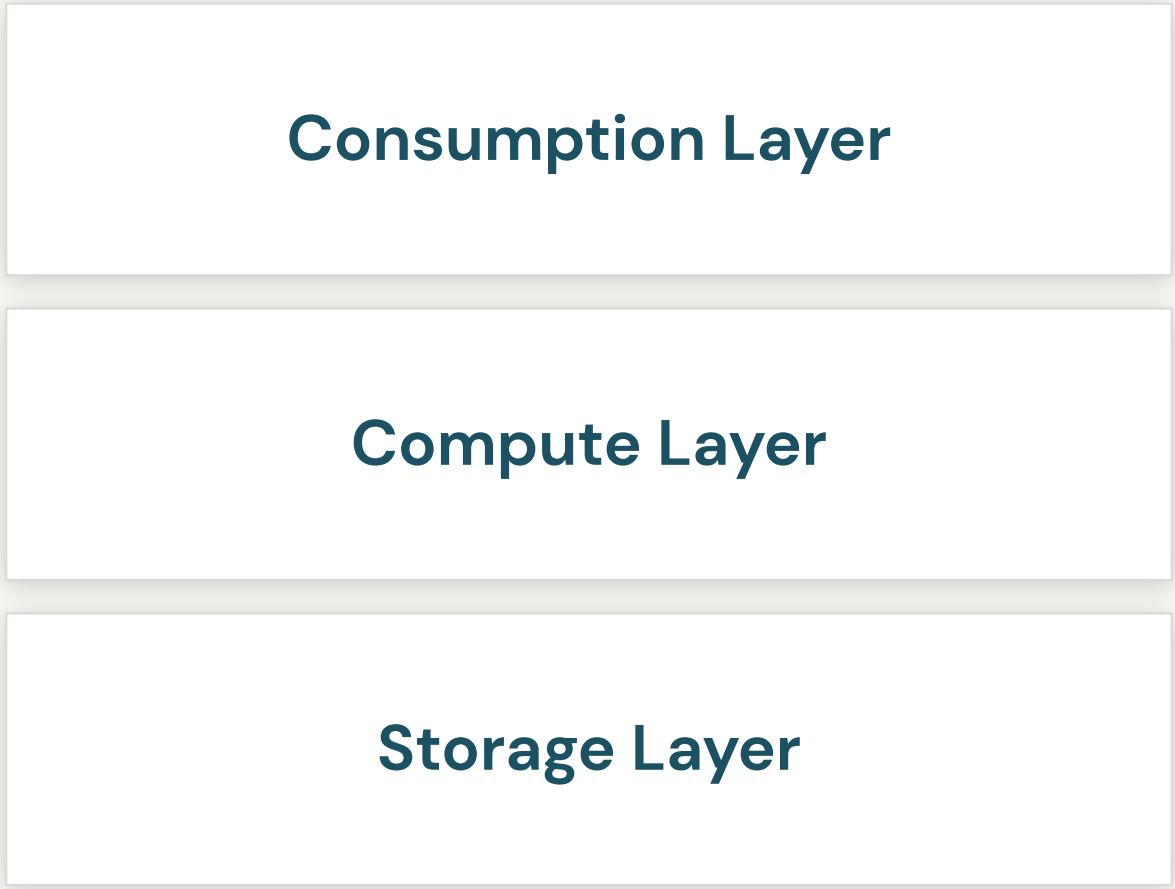
Databricks sets official data warehousing performance record

and outperforms cloud data warehouses with lakehouse

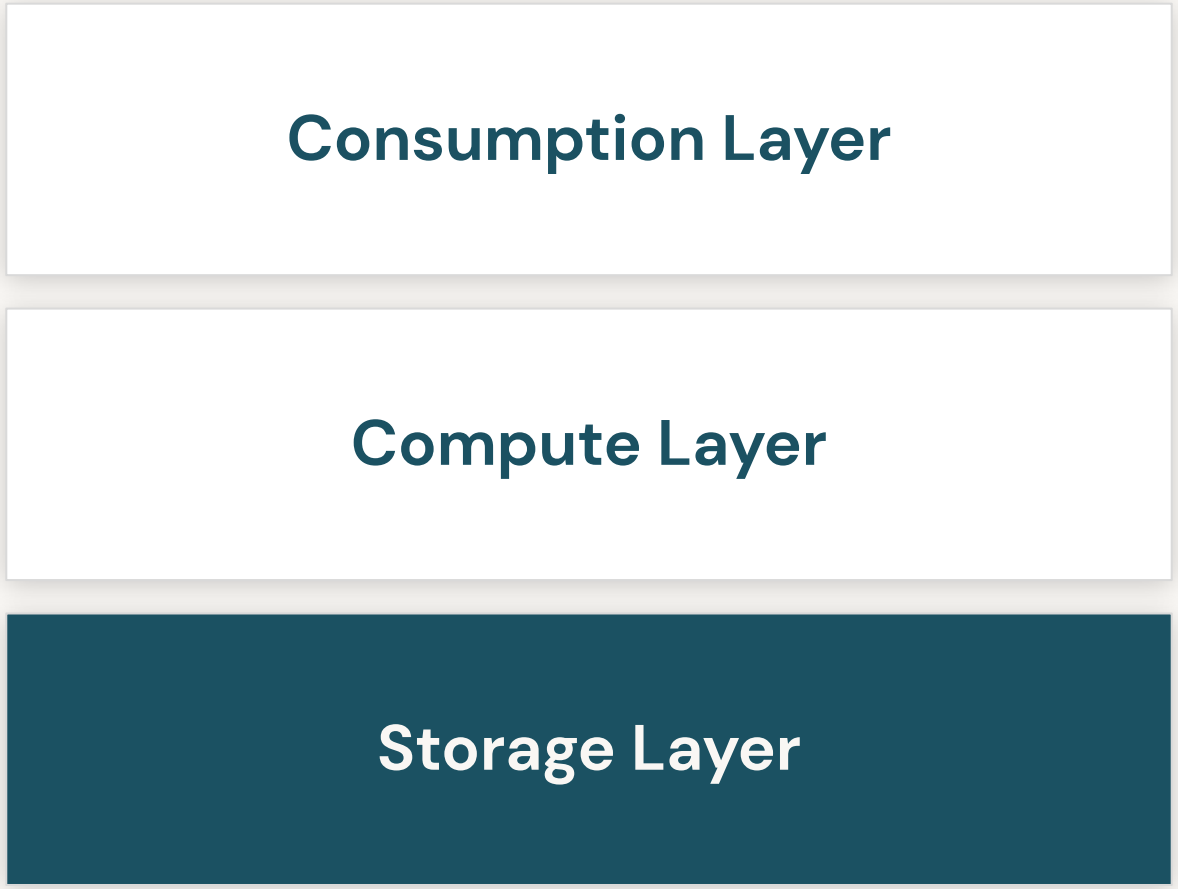
Learn more at https://dbricks.co/benchmark

How to build a Lakehouse?

Inner Workings of Lakehouse



Inner Workings of Lakehouse



Delta Lake is the foundation of the Lakehouse

An open format storage layer built for lake-first architecture

ACID Transactions, Time travel, Highly available

Advanced indexing, Caching, Auto-tuning

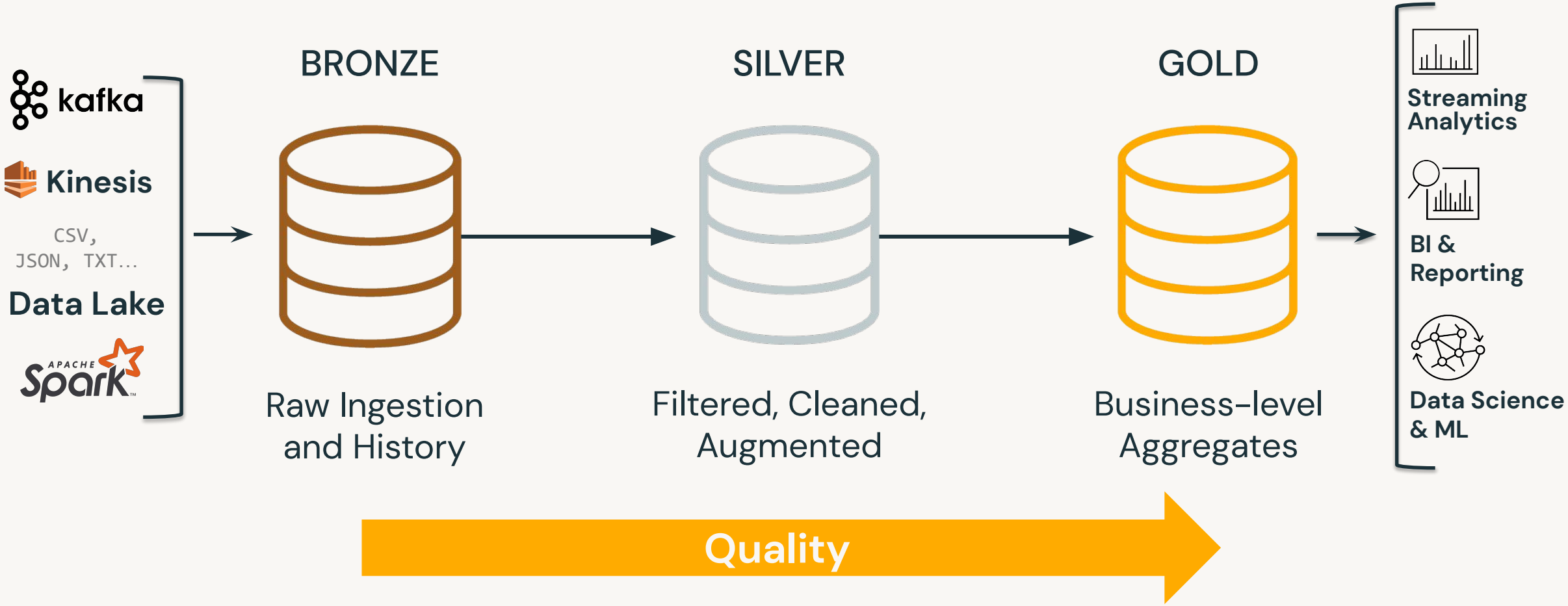
Fine-grained, role-based access controls

Streaming & batch, Analytics & ML

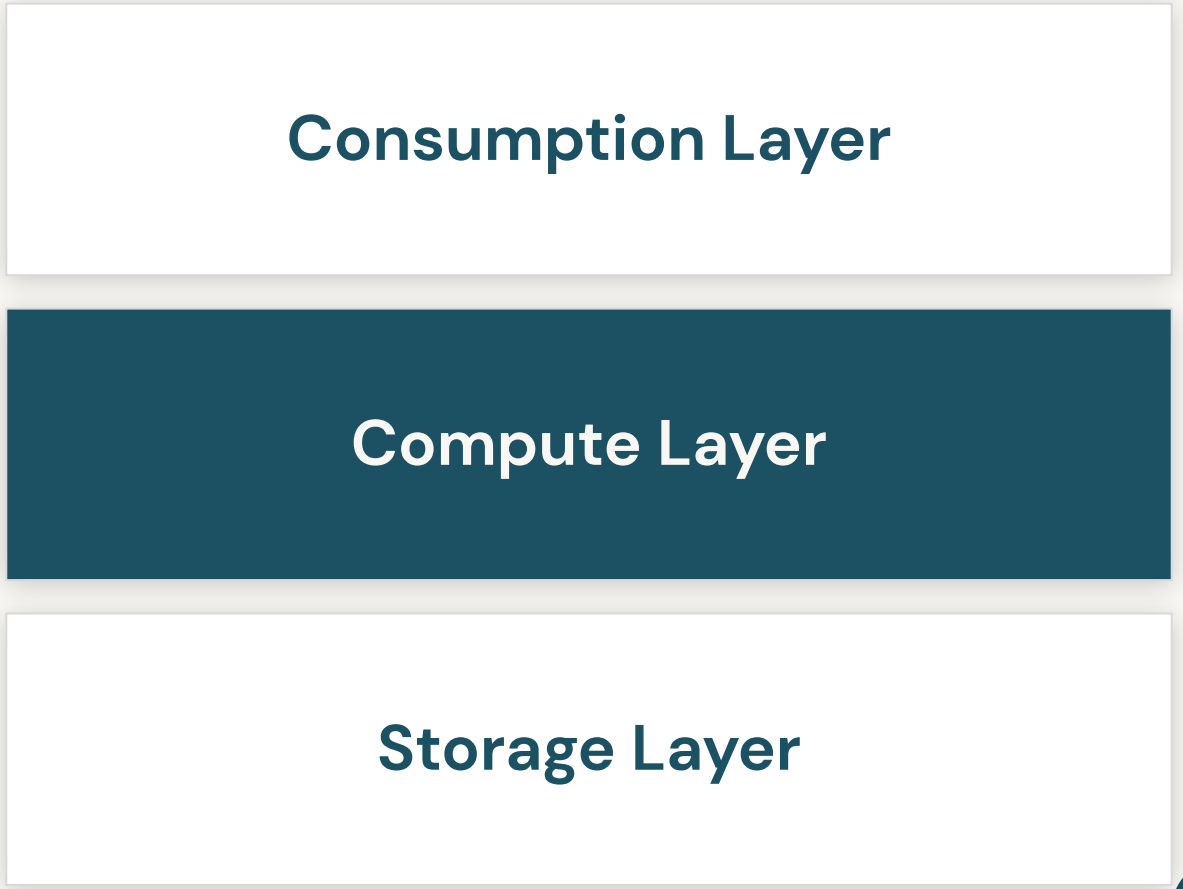
Python, SQL, R, Scala

Building the foundation of the Lakehouse

Greatly improve data quality for end users – on your existing data lake



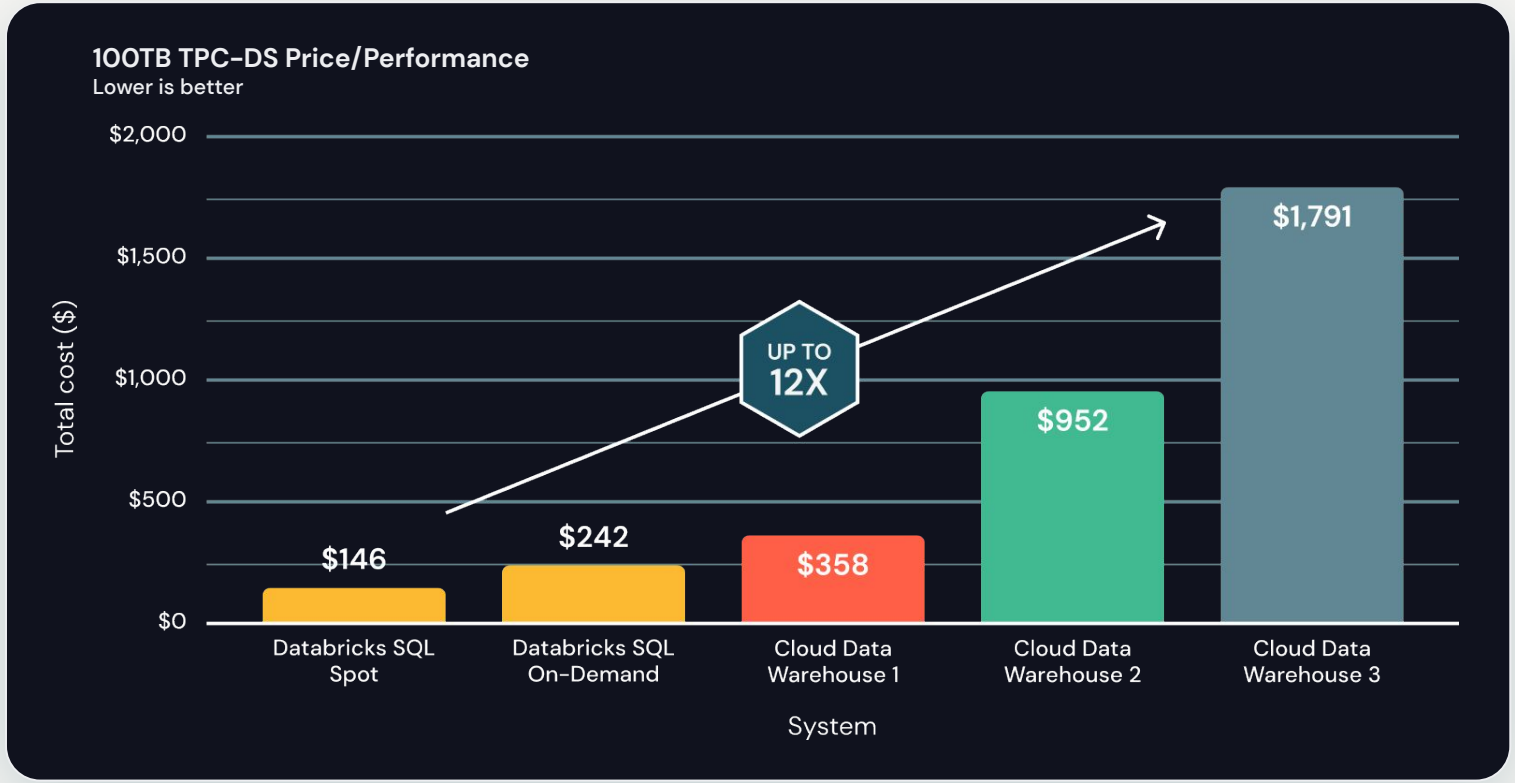
Inner Workings of Lakehouse



Best price / performance

Lightning fast analytics

Query and analyze your most complete and freshest data with
up to 12x better price/performance than legacy cloud data warehouses.

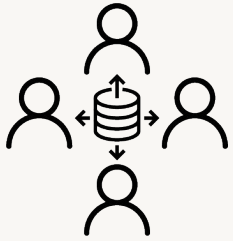


Source: Performance Benchmark with Barcelona Supercomputing Center

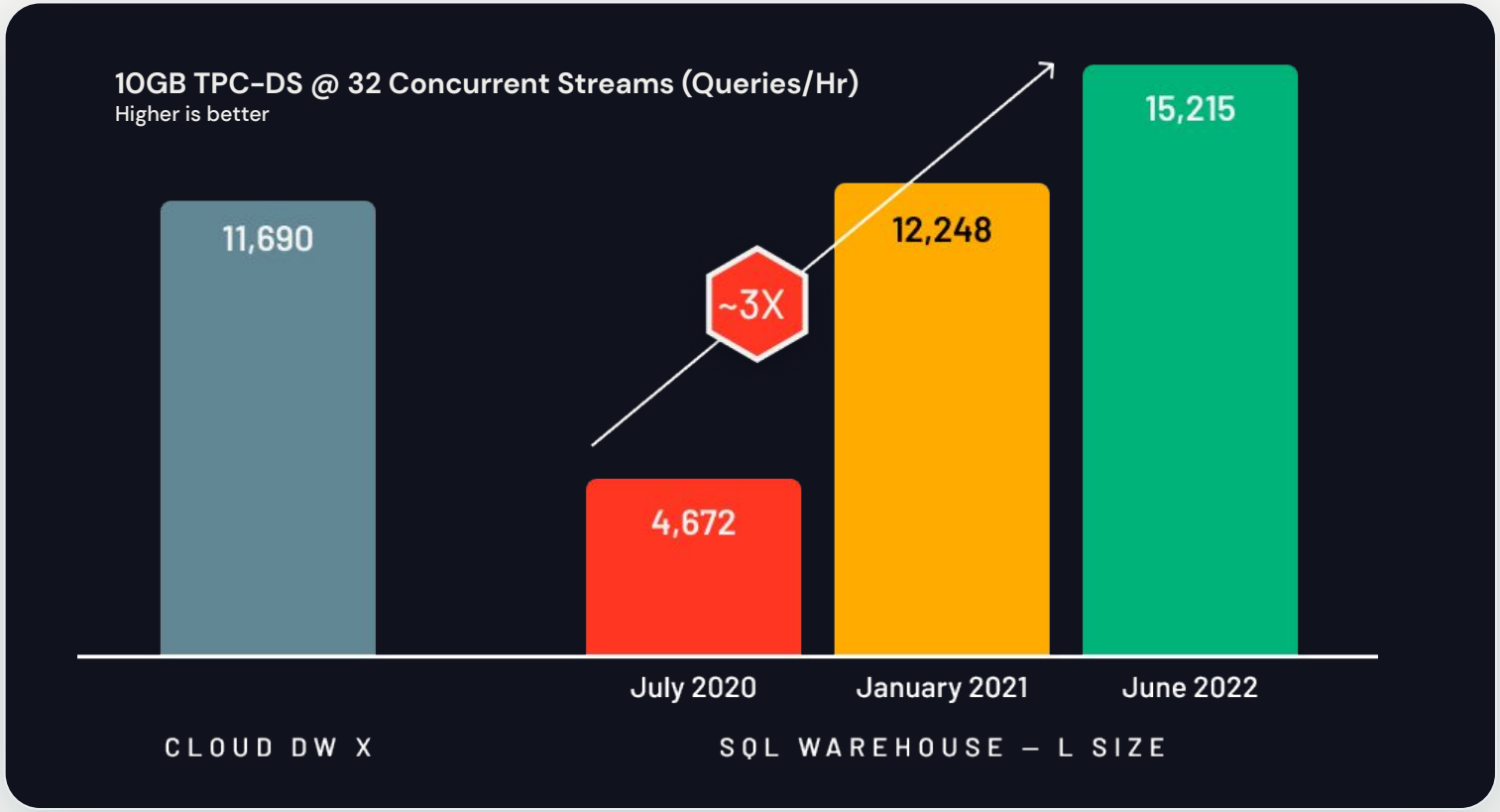
Fast and predictable performance for all queries

Beyond large query performance

High Concurrency



Shaving system overhead off and improving QpH by 3X

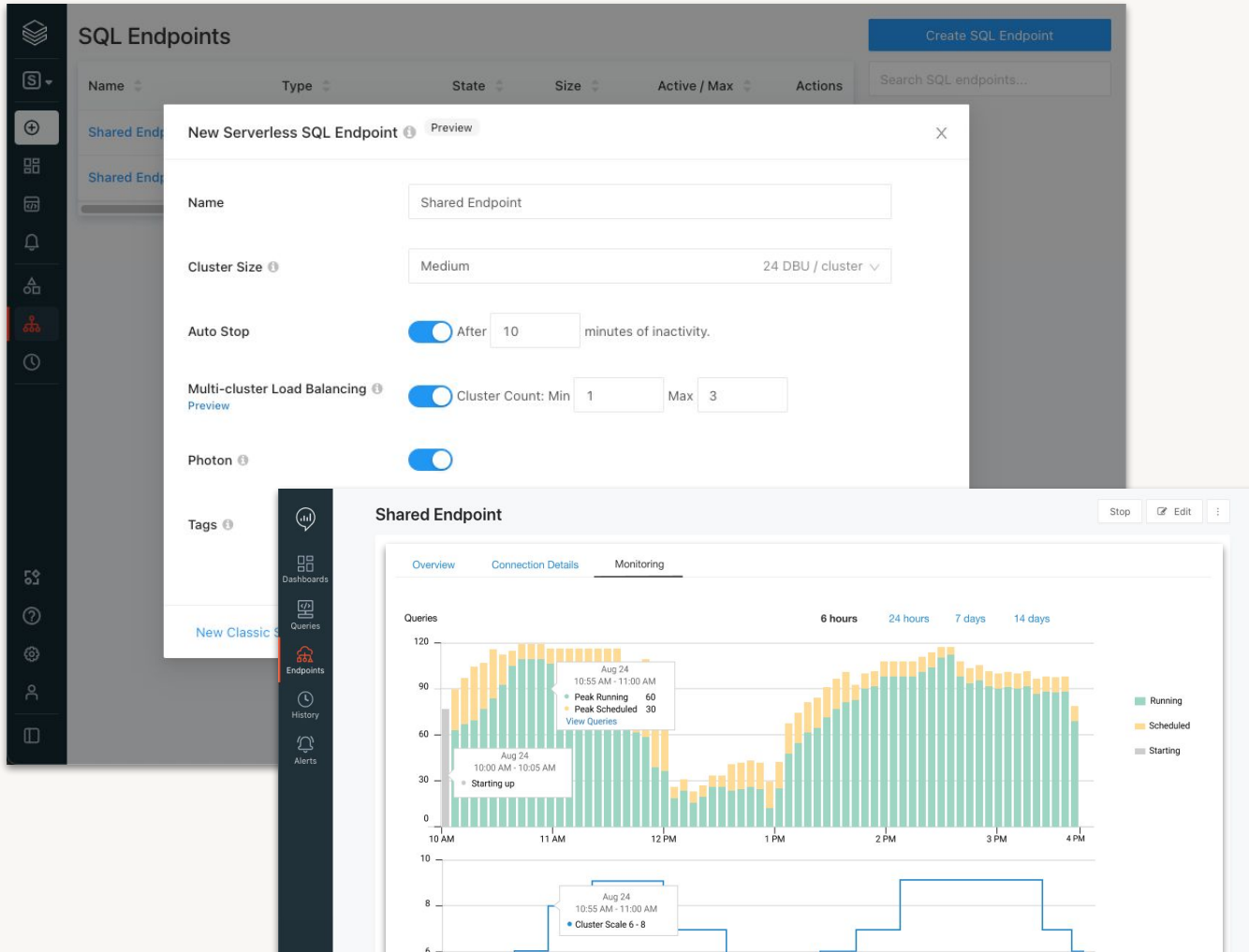


Simple administration – no management required

Elastic, instant compute decoupled from storage – now in Serverless

Quickly setup instant, elastic SQL compute decoupled from storage. Databricks automatically determines instance types and configuration for the best price/performance.

Then, easily manage users, data, and resources with endpoint monitoring, query history, and data explorer.

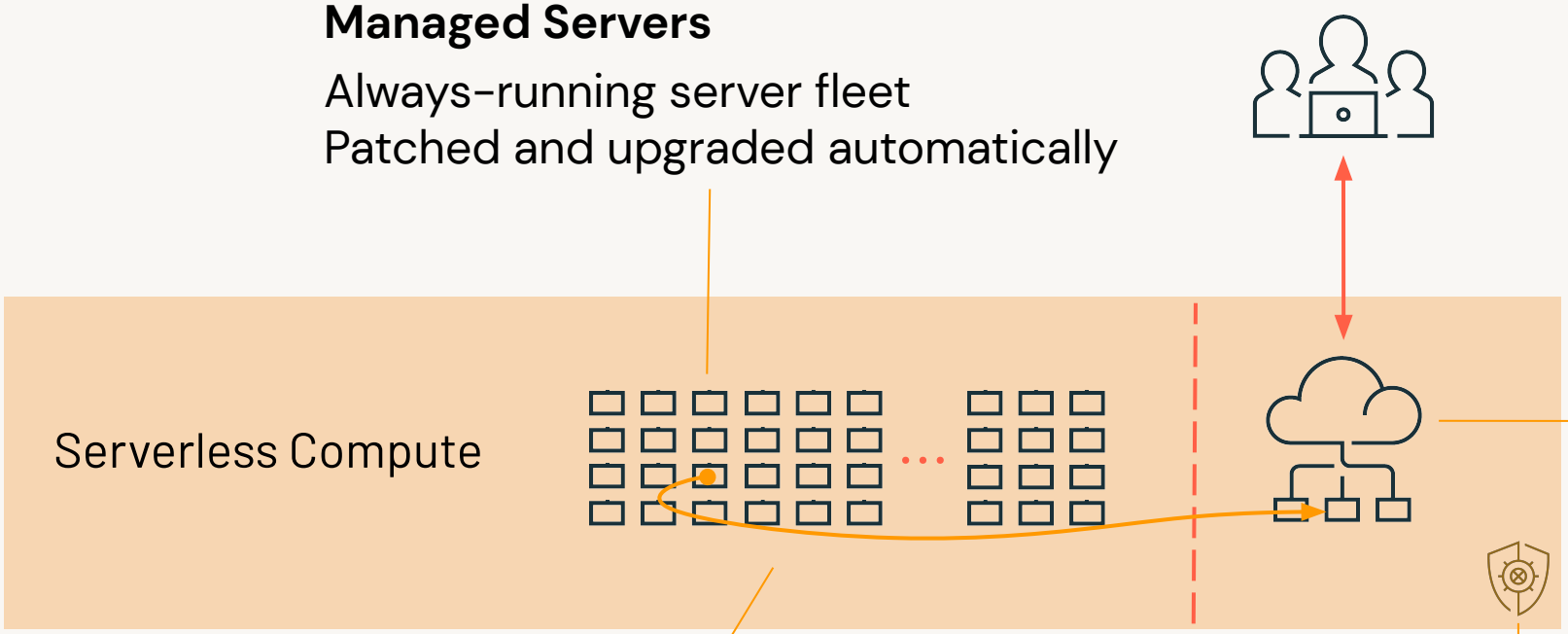


Databricks SQL Serverless

In Preview

Managed Servers

Always-running server fleet
Patched and upgraded automatically



Optimized Capacity

Idle clusters removed 10 minutes after last query

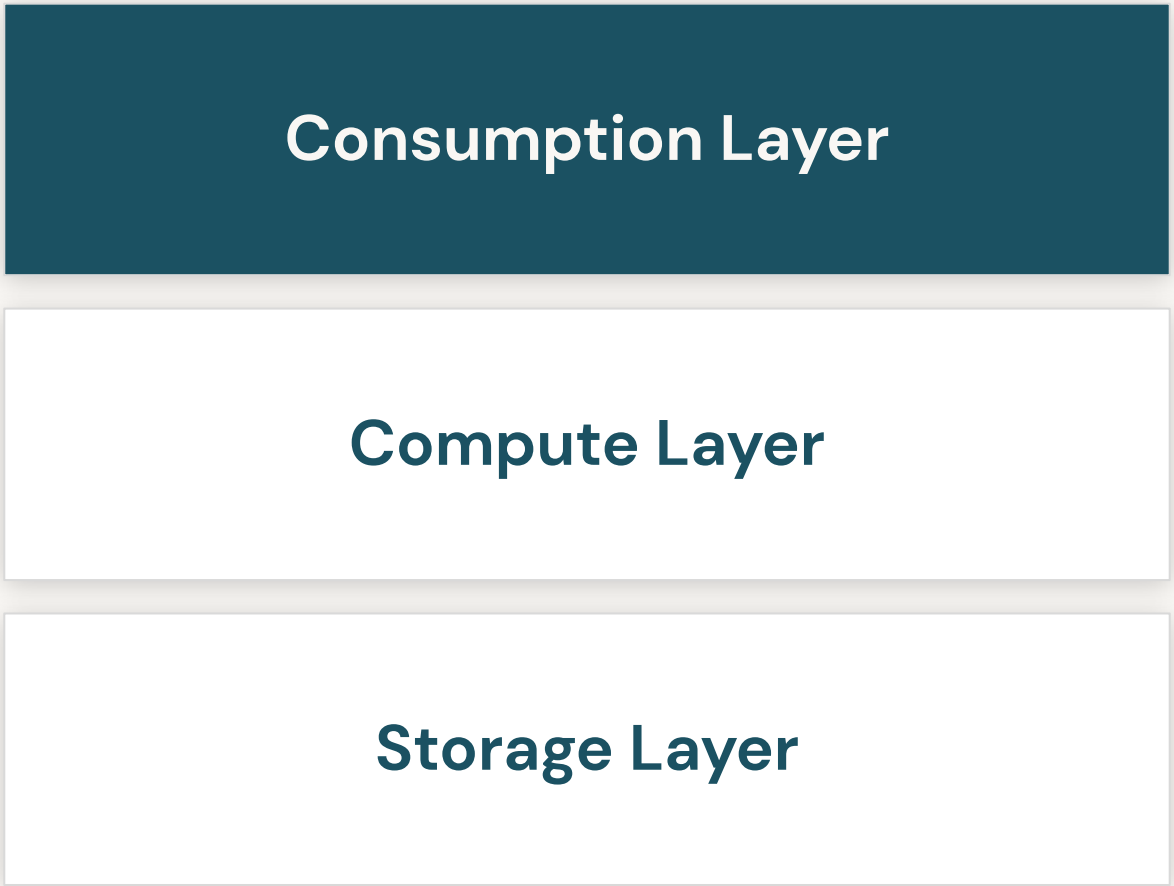
Instant Compute

~15 second allocation to users
No waiting for clusters to startup or scale-up

Secure

3-layer isolation with data encryption

Inner Workings of Lakehouse

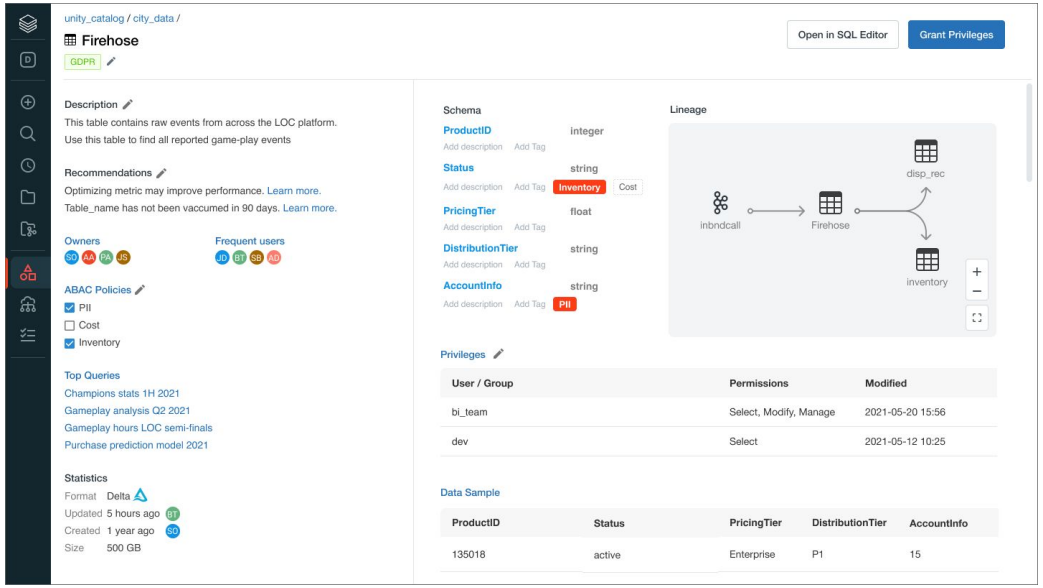


In Preview

Fine-grained governance on the Lakehouse

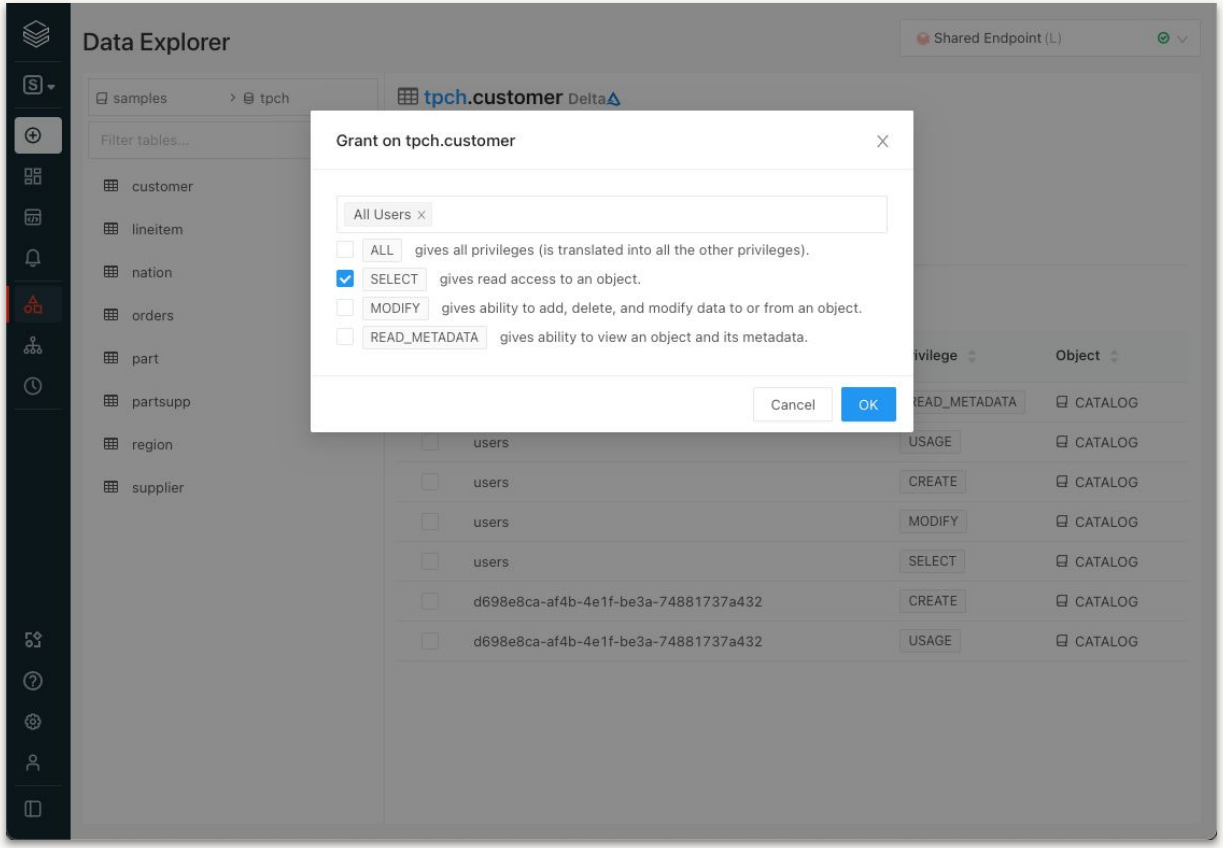
Introducing Databricks Unity Catalog

- **Centrally catalog, search, and discover** data and AI assets
- Simplify governance with a **unified cross-cloud governance model**
- **Easily integrate** with your existing Enterprise Data Catalogs
- **Securely share live data across platforms** with delta sharing



Easy to govern self-served analytics

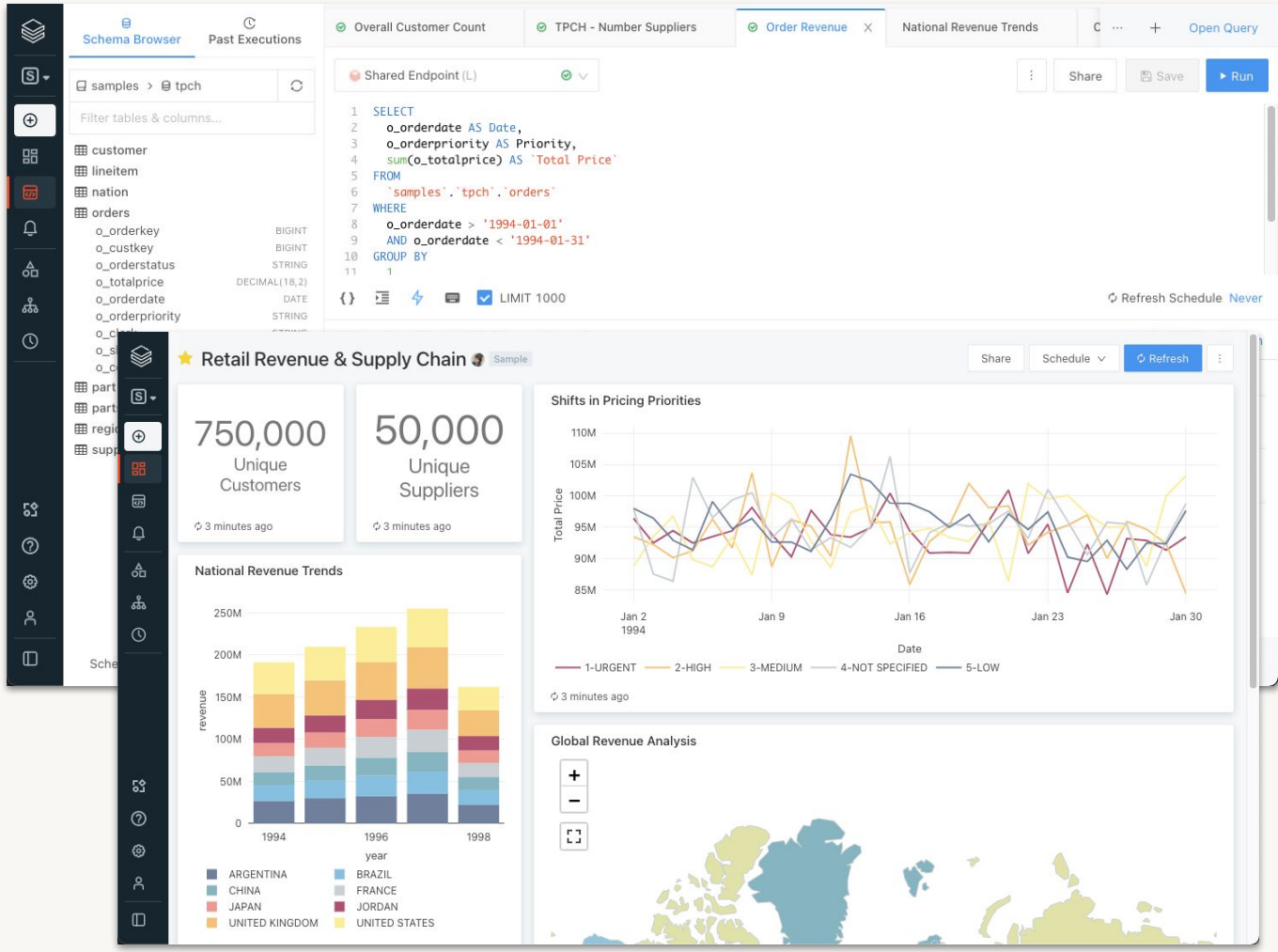
- Confidently onboard new users, discover, secure, and govern data
- Manage costs and usage effectively with endpoints monitoring and query history
- Meet compliance needs with built-in audit trail



First-class SQL development experience

Analytics & BI on the Lakehouse

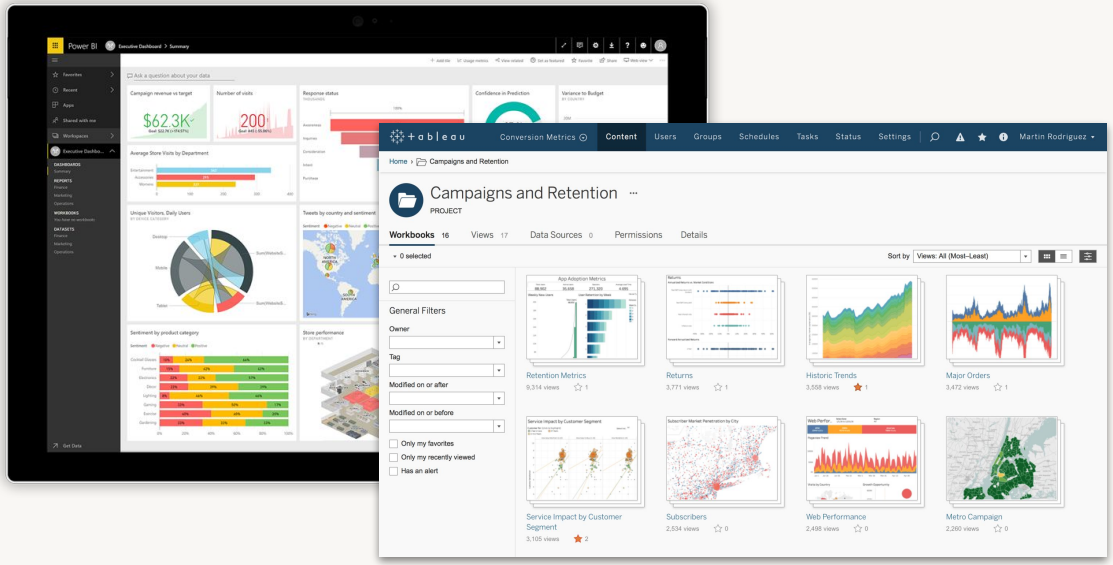
Query data lake data using familiar **ANSI SQL**, and find and share new insights faster with the built-in SQL query editor, alerts, visualizations, and interactive dashboards.



A platform for your tools of choice

Analytics & BI on the Lakehouse

Get critical business data in with one click integrations, and benefit from fast performance, low latency, and high user concurrency for **your existing BI tools**.



Ingest & ETL

dbt Fivetran

BI & Analytics

tableau Power BI

Qlik TIBCO Spotfire

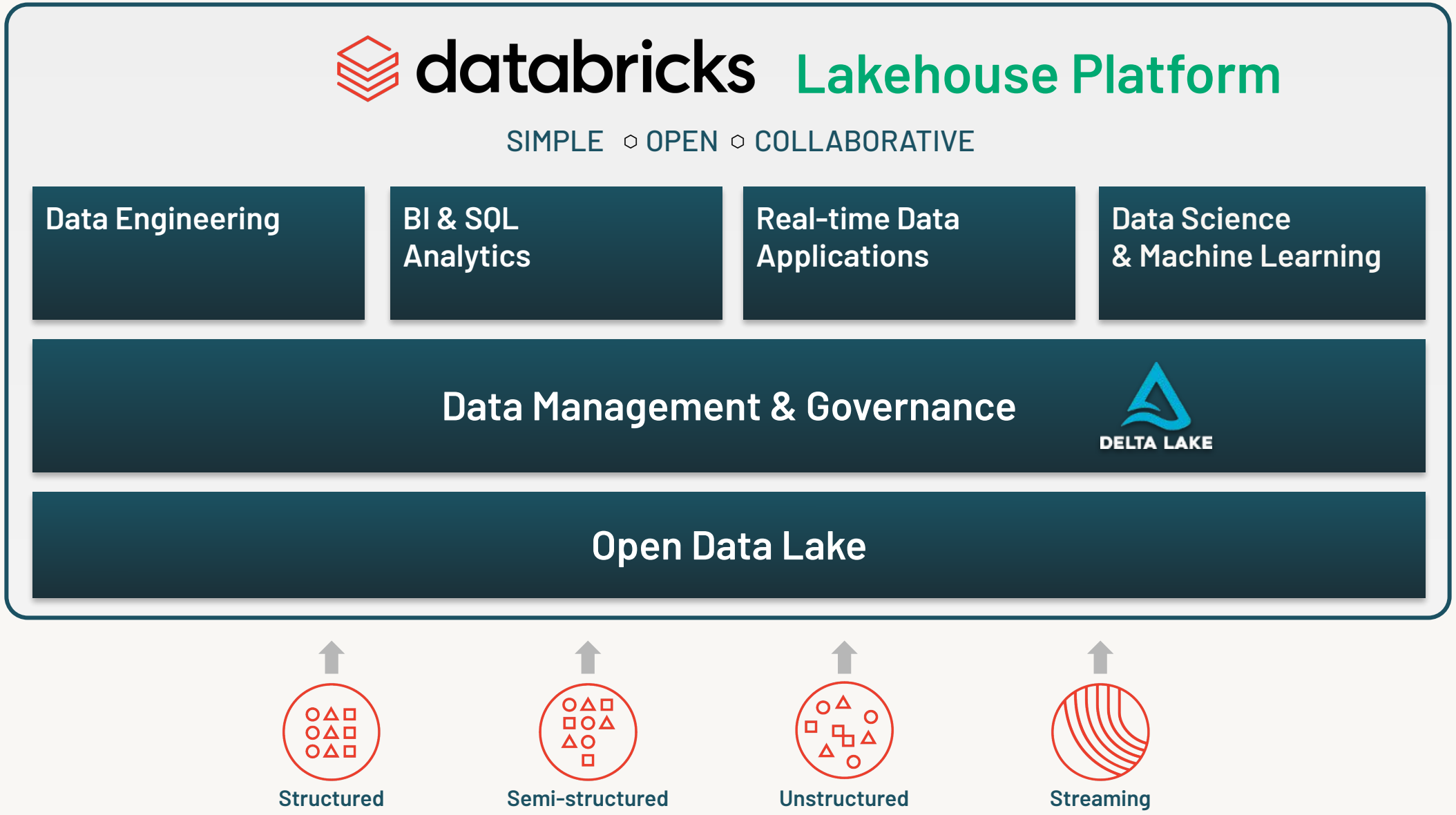
MicroStrategy Looker

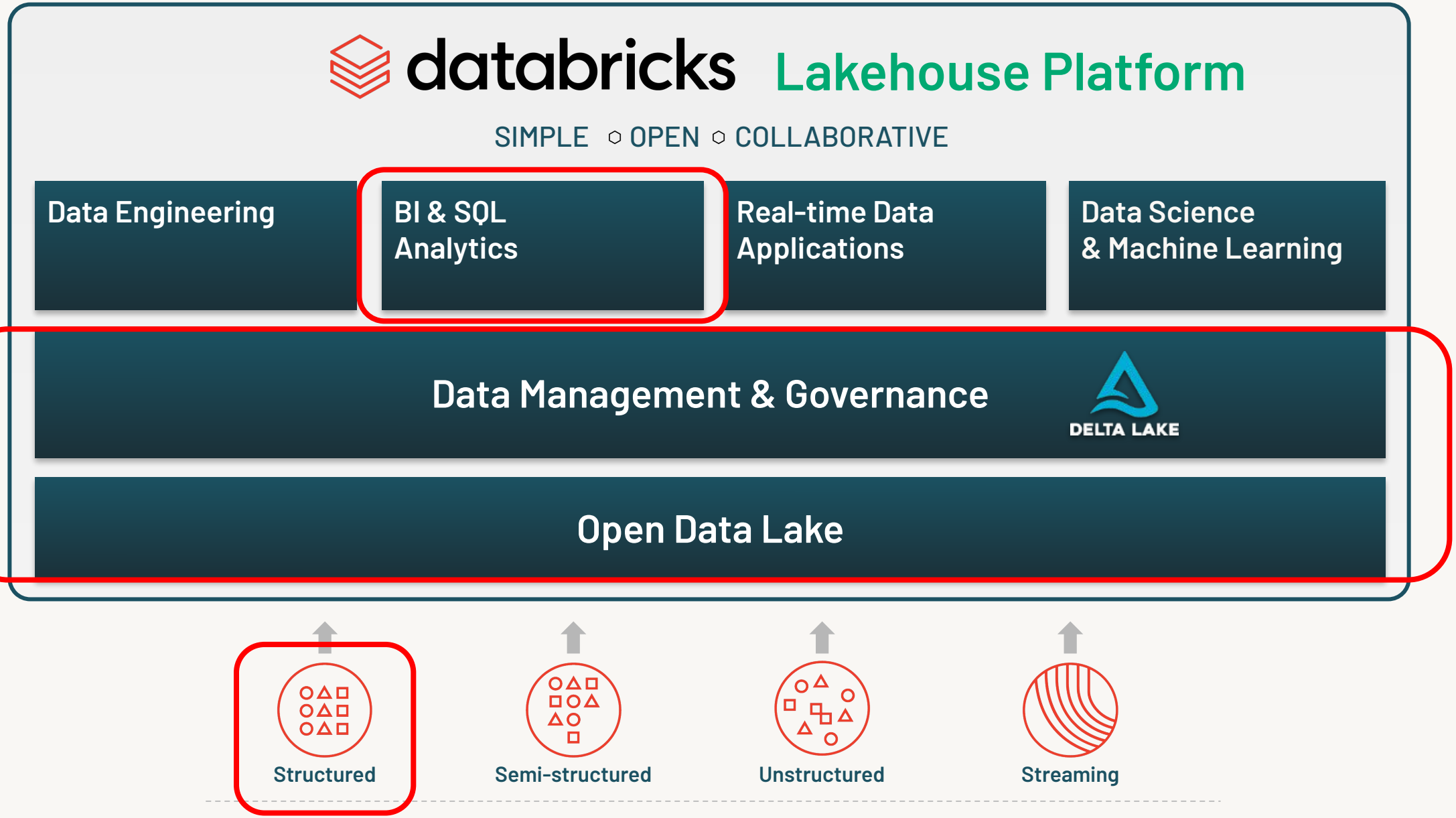
MODE ThoughtSpot

Databricks SQL Lab Overview

Lab Context

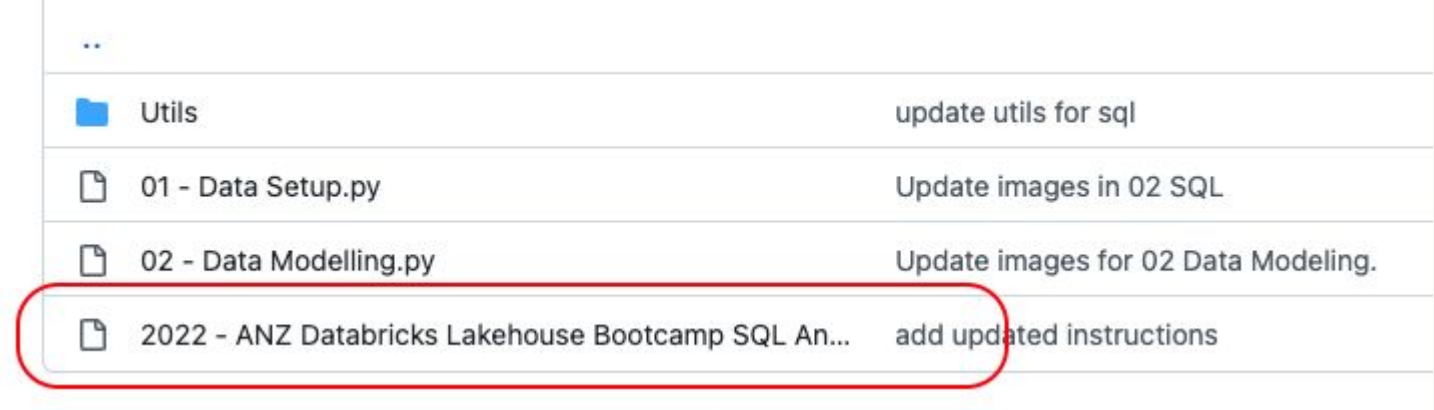
- As part of the lab today, we will be working with APJuice data from different juice outlets across APJ
- Dataset consists of sales, stores, customer and product info
- The lab will focus on analysing data that has been processed using batch to build a star schema data model in Delta Lake Gold Layer and then aggregating the data for dashboarding





Download The Workbook

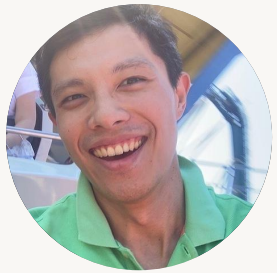
The workbook is in the lab git repo in



https://bit.ly/3NQiFLS

https://github.com/zivilenorkunaite/apjbootcamp2022/LAB 02-SQL/

Databricks SQL Lab



Afternoon Tea & Feedback