



Introduction to Databricks

Bring reliability, performance,
and security to your data lake



October 2022



Continue learning on Databricks Academy

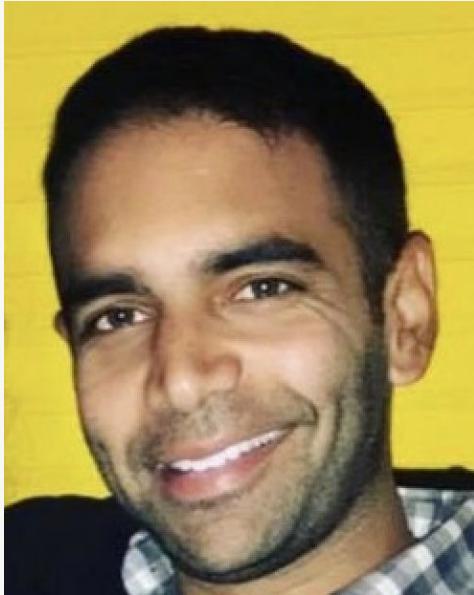
New link: [Databricks Academy](#) and check our [full list of courses](#)

The screenshot shows the Databricks Academy platform. At the top, there's a navigation bar with the Databricks logo, a search bar, and a menu icon. Below the header, a banner says "Welcome to Databricks training!". A navigation sidebar on the left includes a home icon and the text "Home User Home Page". The main content area features a large photo of a diverse group of people. Below the photo, there are two sections: "Enrolled Learning" and "My Task List". The "Enrolled Learning" section has a search bar and a grid of course cards. The "My Task List" section shows a summary of tasks: DEADLINES (1), NOT STARTED (8), IN PROGRESS (1), and ILT/WEBINAR.



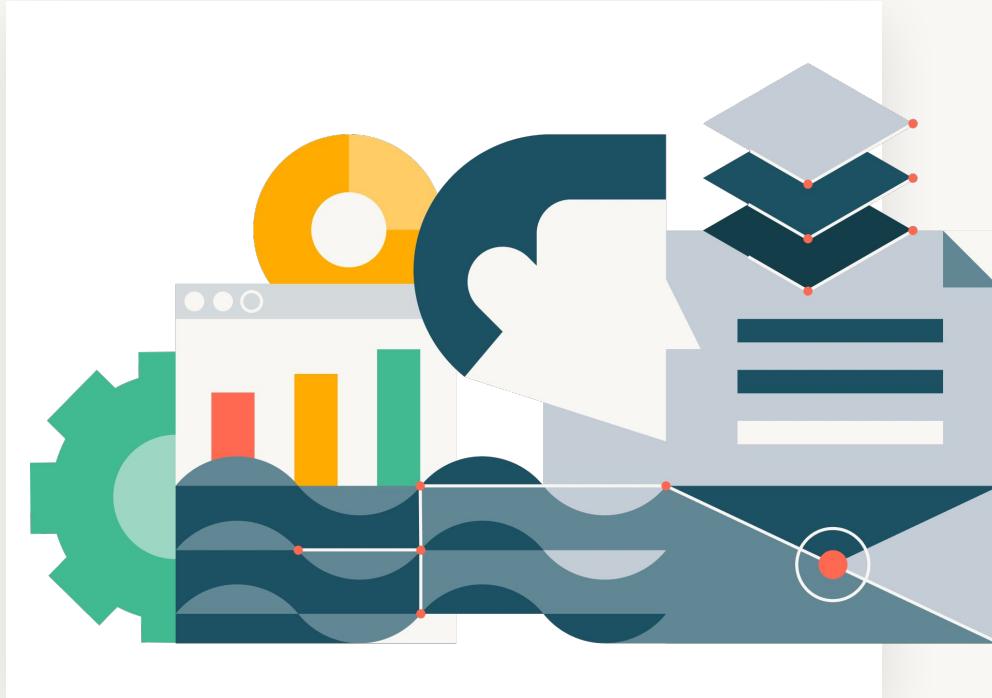
About Me

Jai Karve, Solutions Architect, Databricks



- Native Houstonian
- Shellfish enthusiast
- Likes to experiment with facial hair
- New Dad

Presentation Overview



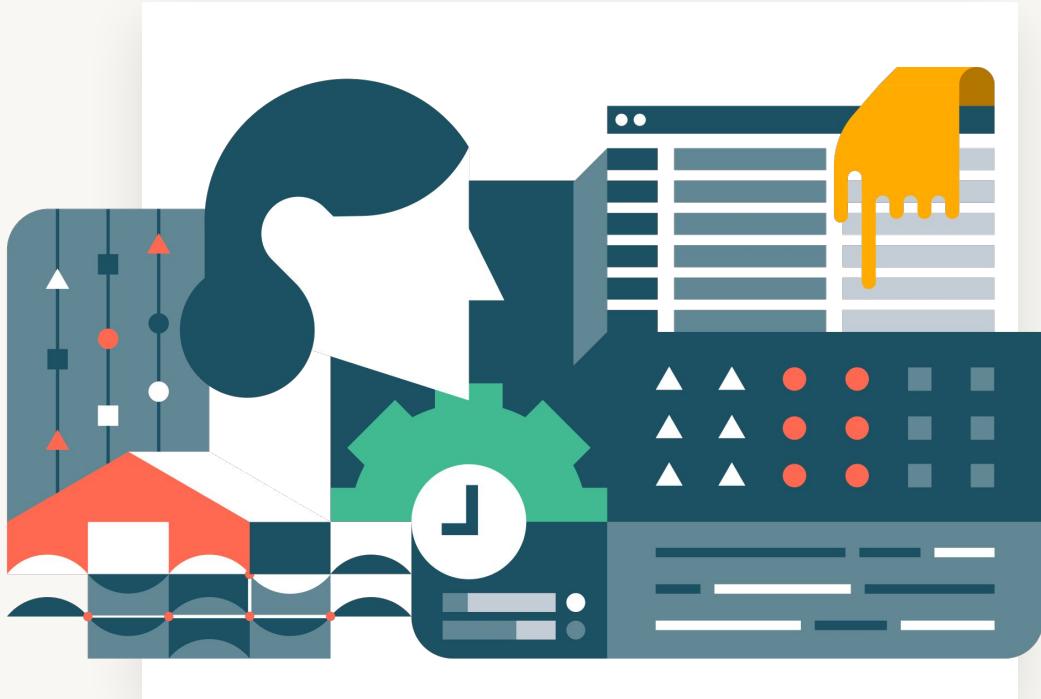
What we'll cover

- Introduction to the Lakehouse
- Integration with BI Tools
- Delta Live Tables
- Performance Optimizations
- How Oxy Currently Uses Databricks

Hands on Overview

What we'll do

- Dataframes
- Data Persistence
- Data Ingestion
- Schema Enforcement/Evolution
- Performance Optimizations



Let's get you set up



Step 1: Workspace Registration

The temporary workspace you'll be using in today's workshop is administered by a third party vendor.

Complete this short registration form for the hands-on portion of the workshop.

The image shows two side-by-side screenshots. On the left is a landing page for 'Databricks Cloud Workshops' titled 'Databricks on Azure'. It features a red logo icon, the title 'Databricks Cloud Workshops', the subtitle 'Databricks on Azure', the author 'By : Databricks', a note 'Please sign up to get access to the lab Environment', and a duration '7 hour(s) and 0 minute(s)' with an email link '_cloudlabs@databricks.com'. On the right is a registration form titled 'Register Now' with fields for 'First Name*', 'Last Name*', 'Email*', and a checkbox for agreeing to terms and privacy policy. A 'Submit' button is at the bottom.

Databricks Cloud Workshops

Databricks on Azure

By : Databricks

Please sign up to get access to the lab Environment

7 hour(s) and 0 minute(s)

_cloudlabs@databricks.com

Azure,Databricks

EN

First Name*

Last Name*

Email*

I agree to the Databricks [Terms of Service](#) and acknowledge the Databricks [Privacy Policy](#) (required).

Submit



Step 2: Launching the workspace

**Click “Launch Lab” and
allow 5-10 minutes for
the environment to load.**

The screenshot shows a Databricks workspace interface. At the top, there's a blue header bar with the Databricks logo and a menu icon. Below the header, the text "Databricks on Azure | Jan 13th | United States" is displayed. In the center, a message says "Please click on 'Launch Lab' button to activate your lab environment." To the right of this message is a prominent blue button labeled "LAUNCH LAB".





Delta Lake: The foundation of your lakehouse

Bring reliability, performance,
and security to your data lake



October 5, 2022

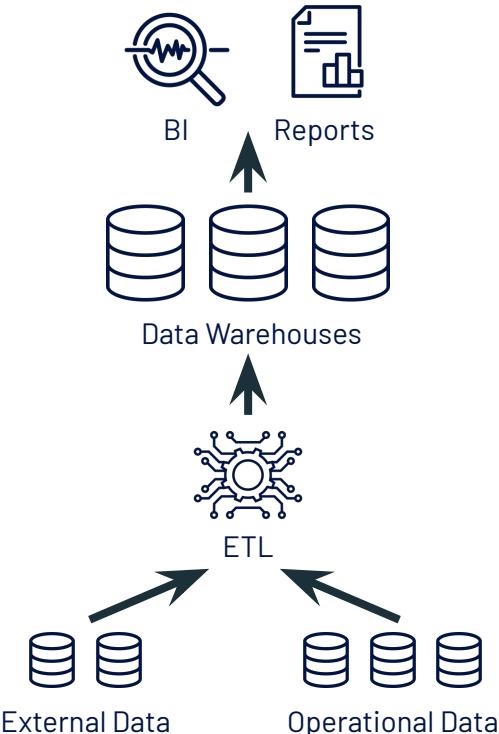


Data Warehouses

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

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

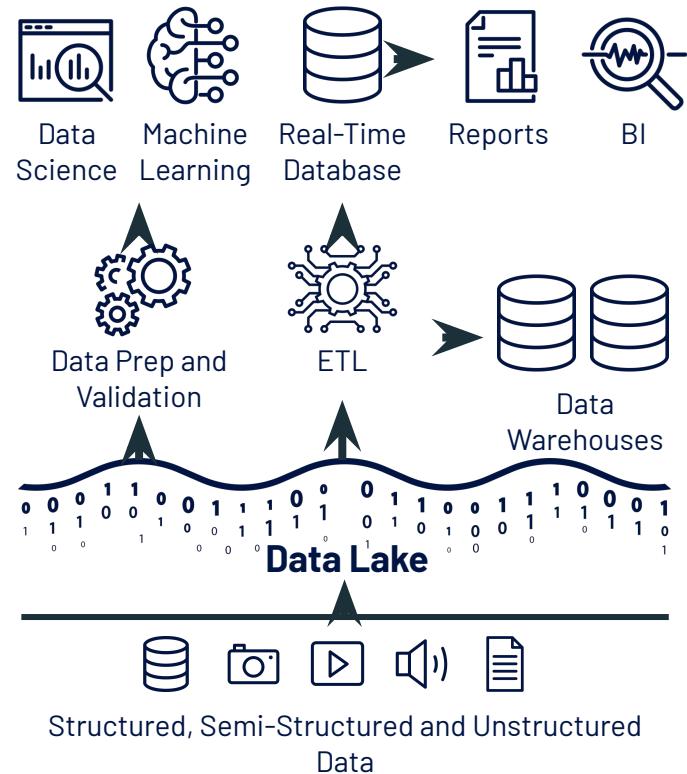
Therefore, most data is stored in
data lakes & blob stores



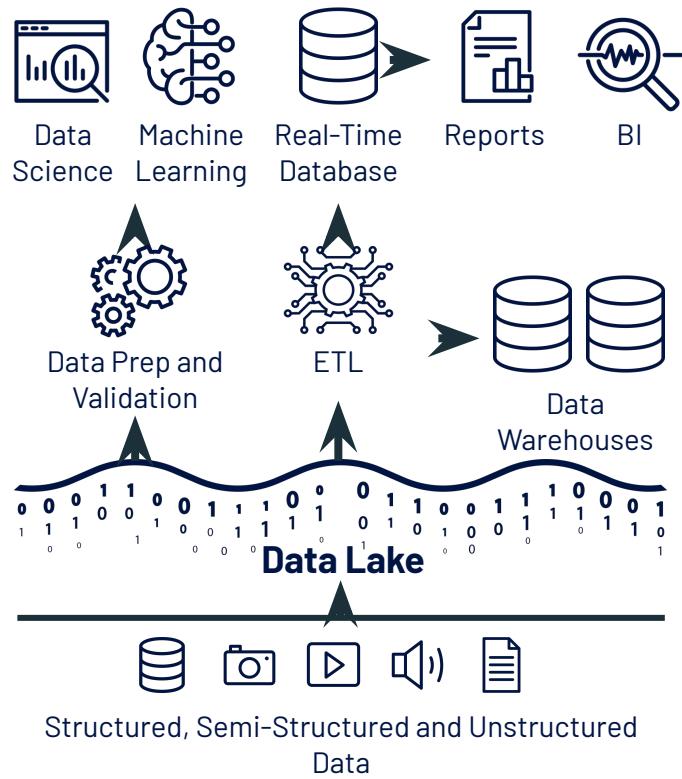
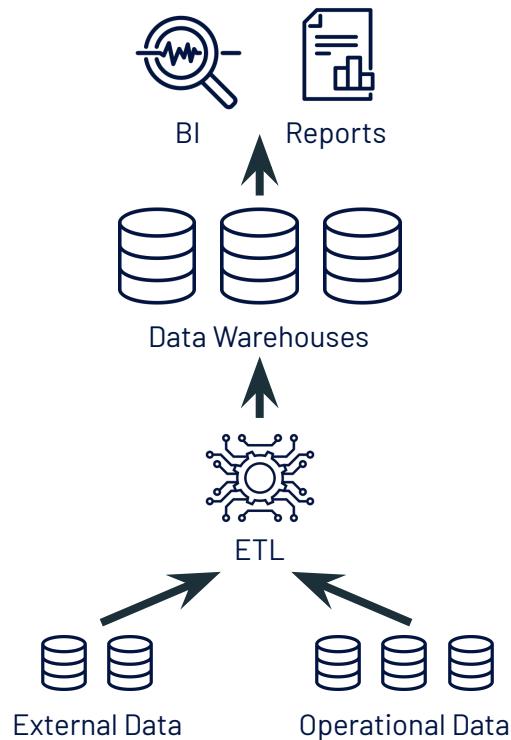
Data Lakes

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

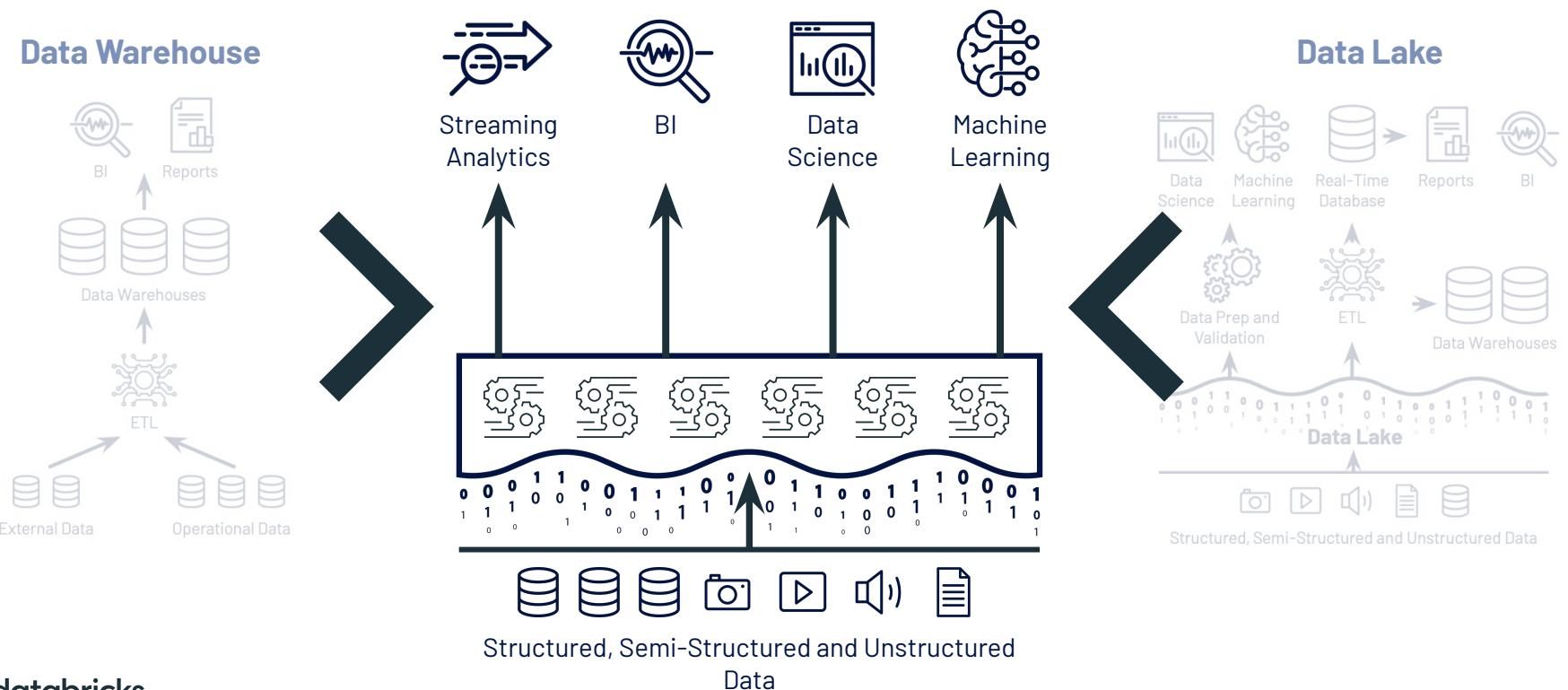
- Poor BI support
- Complex to set up
- Poor performance
- Unreliable data swamps



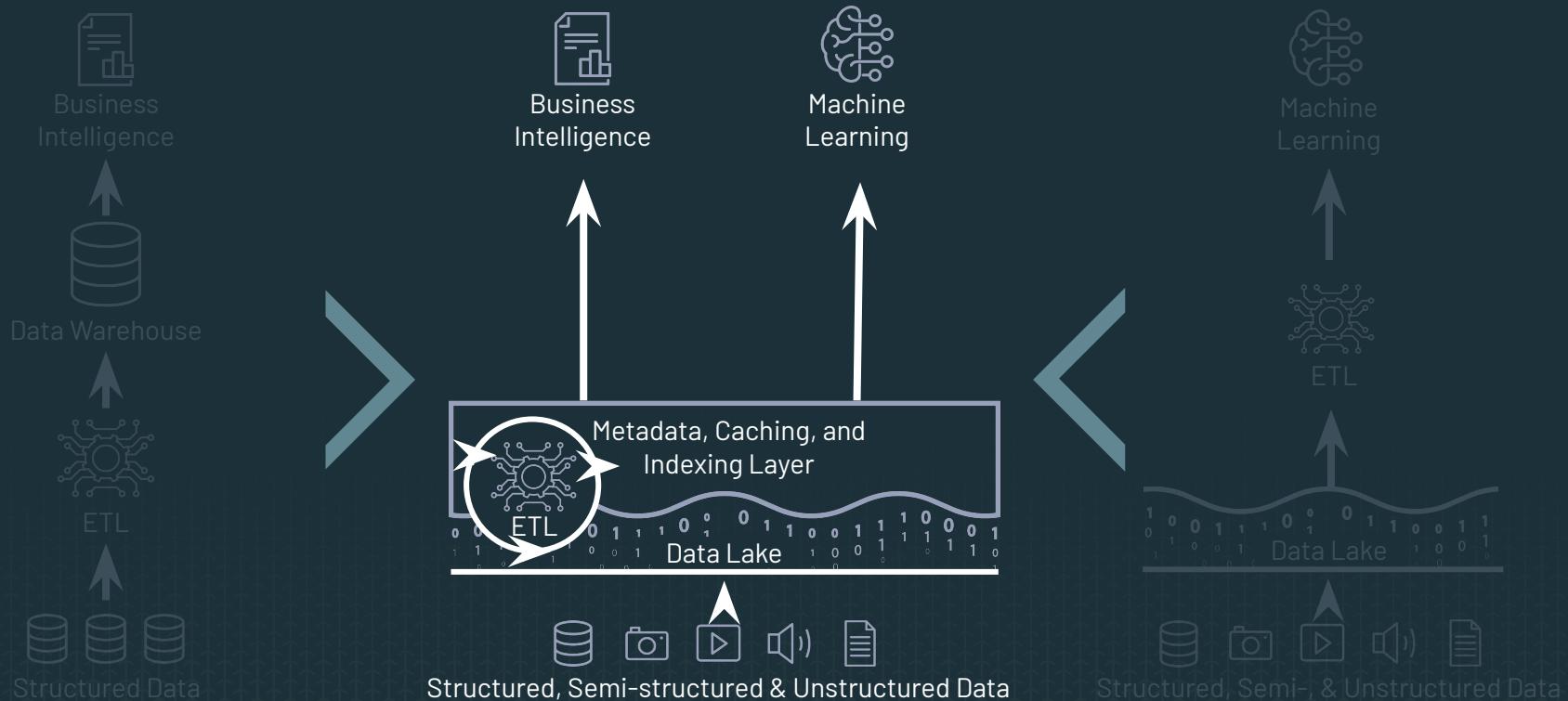
How do we get the best of both worlds?



Lakehouse



New Way Forward: lakehouse



Data Lake



An open approach to bringing
data management and governance
to data lakes

Better reliability with transactions

48x faster data processing with indexing

Data governance at scale with
fine-grained access control lists

Data Warehouse





Innovations lay the foundation for Lakehouse Architecture

Delta Lake Innovations

- Reliability → **ACID Transactions**
- Performance → **Indexing**
- Governance → **Table ACLs**
- Quality → **Expectations**

Positive Business Impact

- Quality data **accelerates innovation**
- Lower TCO with a simple architecture
- Automation **increases productivity**
- Reduces security risk

Open

- Open Source
- Data is in your cloud store (It's mostly parquet)
- No lock-in. Easy to convert to parquet in-place

Under the Covers - Mostly Parquet

Objects (13)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	_delta_log/	Folder	-	-	-
<input type="checkbox"/>	part-00000-08caa61b-05a9-4fbe-909e-2cb1bb3d2243-c000.snappy.parquet	parquet	January 14, 2021, 21:55:22 (UTC+01:00)	434.1 KB	Standard
<input type="checkbox"/>	part-00000-4a31568d-9c74-4e38-a158-710bf20951e9-c000.snappy.parquet	parquet	January 14, 2021, 21:43:36 (UTC+01:00)	434.1 KB	Standard
<input type="checkbox"/>	part-00000-8f4c5fd1-99e6-4366-8856-5ef8b16083d2-c000.snappy.parquet	parquet	January 14, 2021, 18:30:04 (UTC+01:00)	434.1 KB	Standard
<input type="checkbox"/>	part-00000-a7c53bf8-405f-47e4-b42b-4d7ad2471711-c000.snappy.parquet	parquet	January 14, 2021, 17:42:12 (UTC+01:00)	434.1 KB	Standard
<input type="checkbox"/>	part-00000-b3d33634-3884-43ed-a2e9-a6f528665342-c000.snappy.parquet	parquet	January 14, 2021, 17:39:12 (UTC+01:00)	434.1 KB	Standard
<input type="checkbox"/>	part-00000-ea7be554-e892-4238-a9f8-e464087e92c8-c000.snappy.parquet	parquet	July 30, 2021, 17:09:20 (UTC+02:00)	434.1 KB	Standard
<input type="checkbox"/>	part-00001-110bf228-b844-4295-807d-526a61b1a559-c000.snappy.parquet	parquet	January 14, 2021, 18:30:04 (UTC+01:00)	150.4 KB	Standard
<input type="checkbox"/>	part-00001-55f16a53-7031-41fe-834c-f35cf0ca85eb-c000.snappy.parquet	parquet	January 14, 2021, 17:39:12 (UTC+01:00)	150.4 KB	Standard
<input type="checkbox"/>	part-00001-6956200e-037b-4ada-9f3b-6fbf196d5c11-c000.snappy.parquet	parquet	January 14, 2021, 21:55:20 (UTC+01:00)	150.4 KB	Standard
<input type="checkbox"/>	part-00001-b75c47cf-3a80-4236-85dc-75f3da840d53-c000.snappy.parquet	parquet	January 14, 2021, 17:42:12 (UTC+01:00)	150.4 KB	Standard
<input type="checkbox"/>	part-00001-d7ab239e-1109-4b65-984c-9282ddd96cbd-c000.snappy.parquet	parquet	July 30, 2021, 17:09:20 (UTC+02:00)	150.4 KB	Standard
<input type="checkbox"/>	part-00001-ee86fdd5-4330-4438-abf6-a640a74182a9-c000.snappy.parquet	parquet	January 14, 2021, 21:43:35 (UTC+01:00)	150.4 KB	Standard

1. Hard to append data
2. Modification of existing data difficult
3. Jobs failing mid way
4. Real-time operations hard
5. Costly to keep historical data versions
6. Difficult to handle large metadata
7. “Too many files” problems
8. Poor performance
9. Data quality issues

ACID Transactions

Make every operation transactional

- It either fully succeeds - or it is fully aborted for later retries

/path/to/table/_delta_log

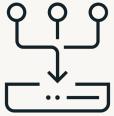
- 0000.json
- 0001.json
- 0002.json
- ...
- 0010.parquet
- 0010.json
- 0011.json

Delta Lake Key Features



ACID Transactions

Protect your data with serializability, the strongest level of isolation.



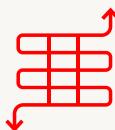
Unified Batch/Streaming

Exactly once semantics ingestion to backfill to interactive queries



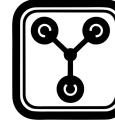
Scalable Metadata

Handle petabyte-scale tables with billions of partitions and files at ease



Schema Evolution / Enforcement

Prevent bad data from causing data corruption



Time Travel

Access/revert to earlier versions of data for audits, rollbacks, or reproduce



Audit History

Delta Lake log all change details providing a full audit trail



Open Source

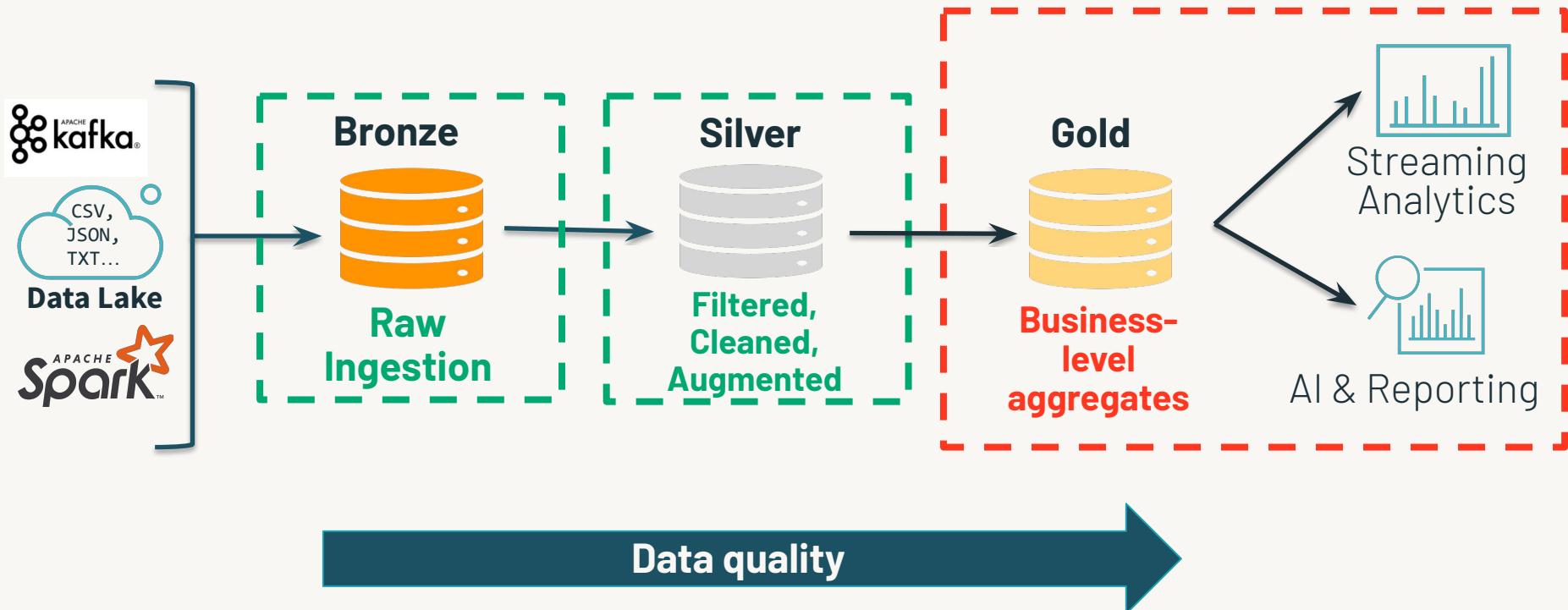
Community driven, open standards, open protocol, open discussions



DML Operations

SQL, Scala/Java and Python APIs to merge, update and delete datasets

The Delta Lake -- Multihop medallion architecture



Walkthrough of Databricks Workspace



Workspace Setup



Step 3: Accessing the Workspace

Copy-paste the provided Databricks Workspace URL into an incognito window.

Tip: Using an incognito window prevents you from logging into an existing workspace. If issues persists, try disconnecting from VPN and trying again.

Environment Details

Here are your credentials to login to [Microsoft Azure](#) and access the On Demand Lab.

Fields	Credentials	Action
Username	odl_user_512944@databrickslabs.com	
Password	wpsj11LAU*kn	
Resource Group : 20438		
Key	Value	Action
Databricks Workspace URL	https://adb-766847864694359.19.azure.databricks.net	



Step 4-5: Signing in on Azure

Sign in with the provided username and password from the credentials page

Select **Sign in with Azure AD**



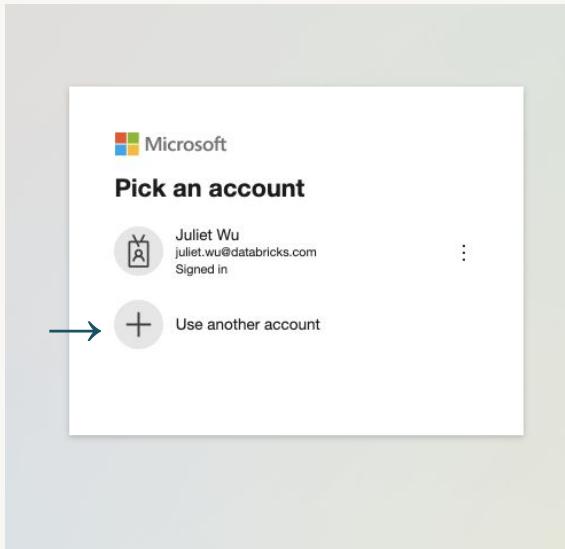
Sign In to Databricks

Sign in using Azure Active Directory Single Sign On. [Learn more](#)

→ **Sign in with Azure AD**

Contact your site administrator to request access.

Select **Use another account**



Copy-paste **provided credentials**



Sign in

Email, phone, or Skype

Can't access your account?

Back

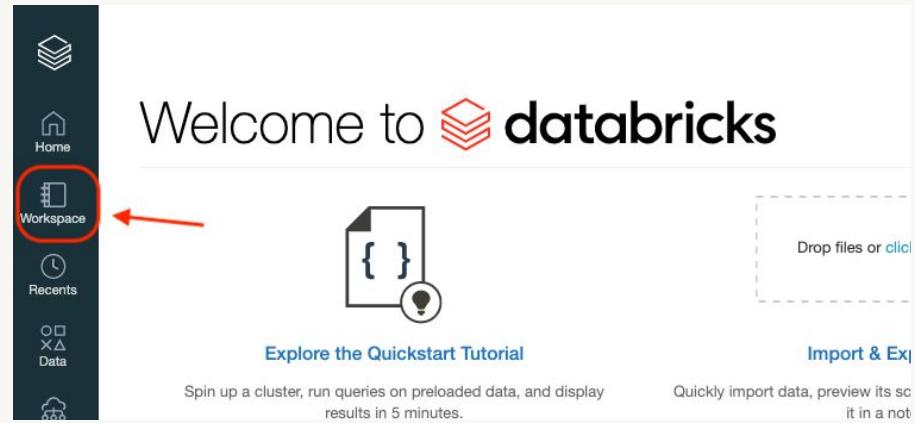
Next

Sign-in options

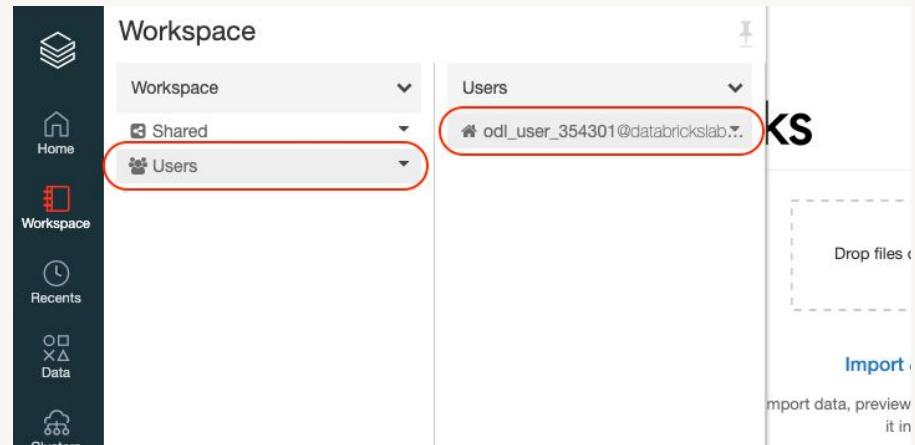
Arrive in your Databricks Workspace

Find the notebook under the Workspace tab and within your User folder.

1. From the navigation panel on the left side, select **Workspace**
2. Select **Users** to find your folder of course materials



The screenshot shows the Databricks homepage. On the left, a dark sidebar navigation bar includes icons for Home, Workspace (circled in red with an arrow pointing to it), Recents, Data, and Clusters. The main area features the "Welcome to databricks" header, a "Explore the Quickstart Tutorial" button, and a "Drop files or click" input field. To the right, there's an "Import & Export" section with a "Drop files or click" input field and a "Import" button.

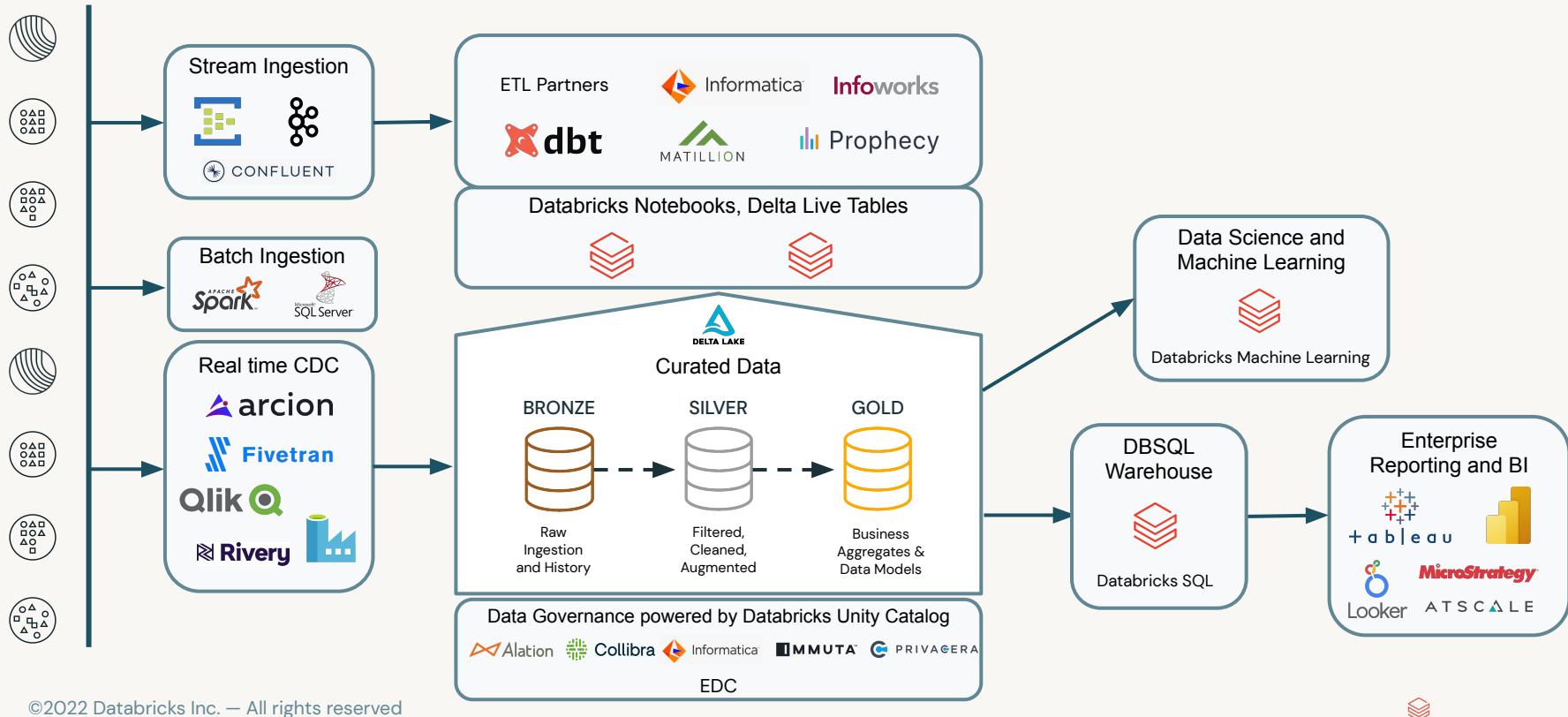


The screenshot shows the Databricks workspace interface. The left sidebar has the same navigation as the homepage. The main workspace area has two dropdown menus: "Workspace" and "Users". The "Users" dropdown is open, showing a list of users, with one user entry circled in red. The "Import & Export" section on the right is partially visible.

Hands On Exercises



Modern Data Warehousing on Databricks

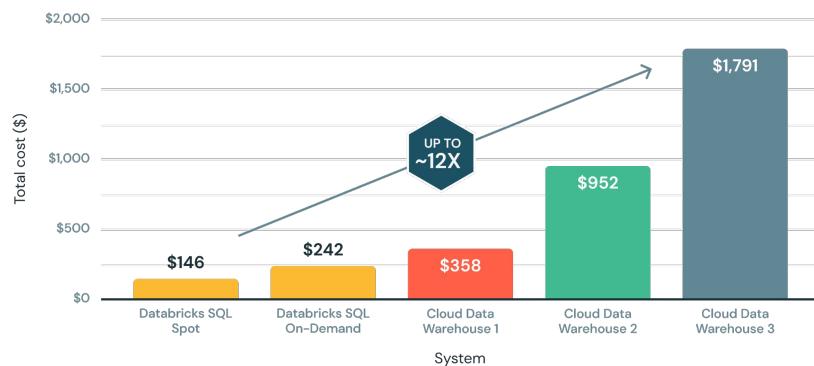


Built from the ground up for best performance

Lightning fast analytics for all queries

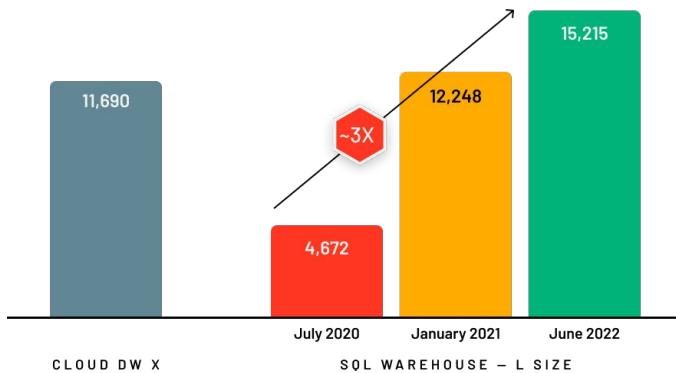
100TB TPC-DS Price/Performance

Lower is better



10GB TPC-DS @ 32 Concurrent Streams (Queries/Hr)

Higher is better



Databricks sets official data warehousing performance record
Learn more: <https://dbricks.co/benchmark>

The Journey to Photon

Next Generation Query Engine

MPP Execution model, written in native code and leveraging vectorized CPU primitives.

Exabytes of data processed

Billions of queries executed

3-8x

faster interactive workloads

1/5

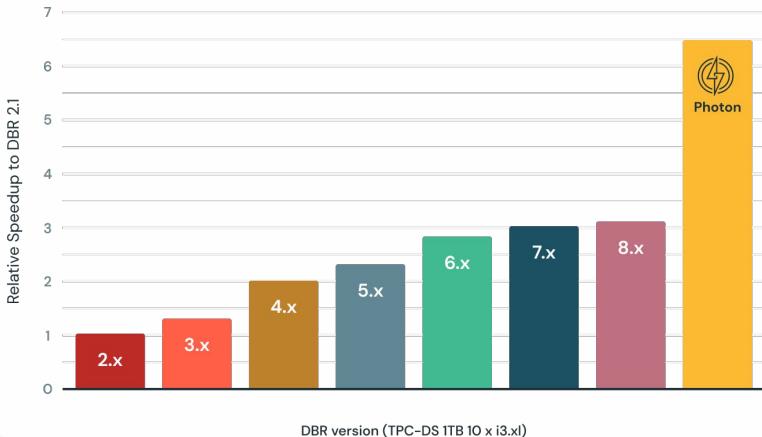
compute cost for ETL

30%

average TCO savings

Relative Speedup to DBR 2.1 by DBR version

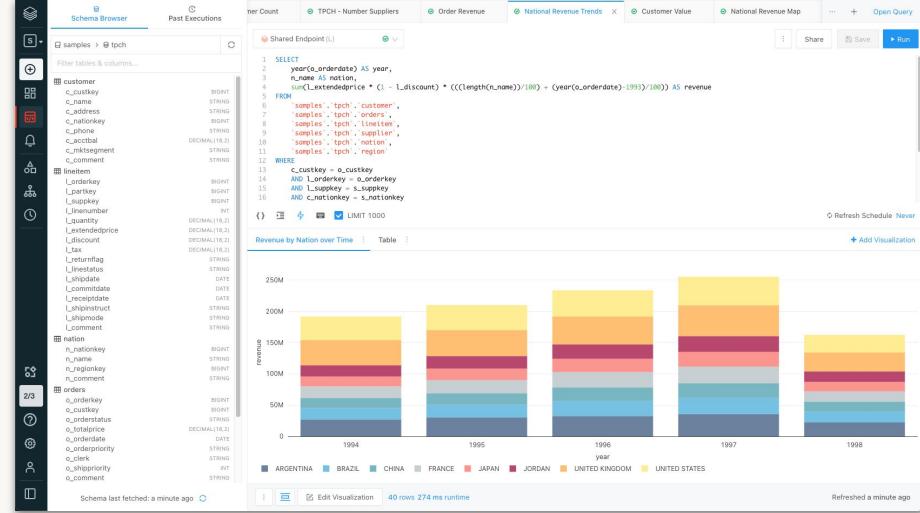
Higher is better



Databricks SQL Query Editor

Collaboratively query, explore, and transform data in-place

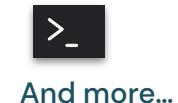
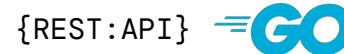
- Easily **ingest** data from cloud storage, local files, or business applications using Fivetran
- Discover data, explore database schema, and query data using **ANSI SQL**
- Save, share, and reuse queries across teams to get to results faster
- Orchestrate queries, alerts, and dashboards with automated **workflows**
- Stay up to date with alerts and automatic refresh schedules



Data Consumption

Query from any tool

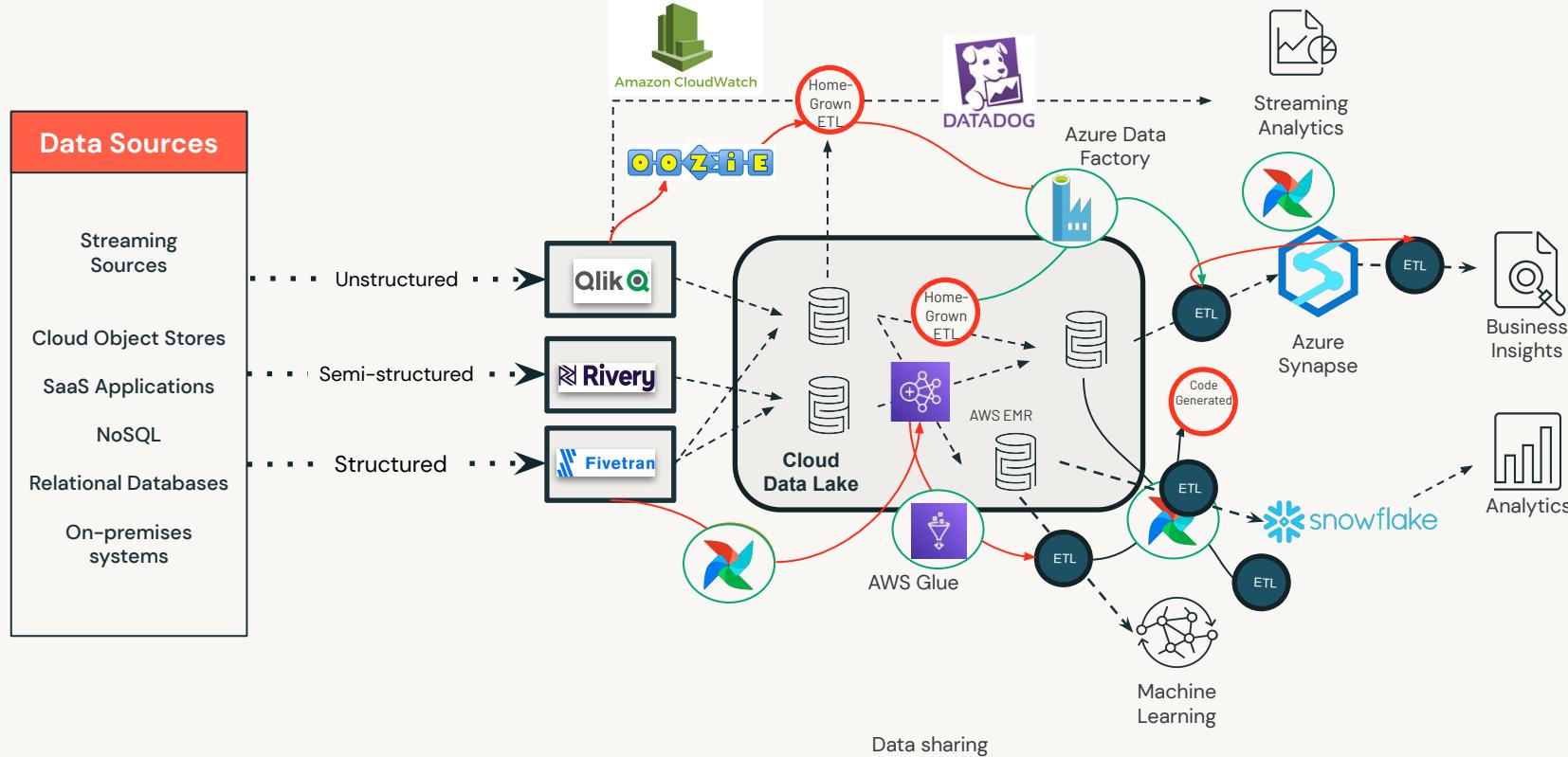
Connect existing dashboards or brand new ones to the latest and freshest data, leverage your favorite tools to find new insights, and build custom data apps powered by the lakehouse with tools and languages you already know.



Delta Live Tables



But there is complexity in the data delivery....



Delta Live Tables

The best way to do ETL on the lakehouse

```
CREATE STREAMING LIVE TABLE raw_data  
AS SELECT *  
FROM cloud_files ("/raw_data",  
"json")
```

```
CREATE LIVE TABLE clean_data  
AS SELECT ...  
FROM LIVE.raw_data
```



Accelerate ETL development

Declare **SQL or Python** and DLT automatically orchestrates the DAG, handles retries, changing data

Automatically manage your infrastructure

Automates complex tedious activities like **recovery, auto-scaling, and performance optimization**

Ensure high data quality

Deliver reliable data with built-in **quality controls, testing, monitoring, and enforcement**

3

5

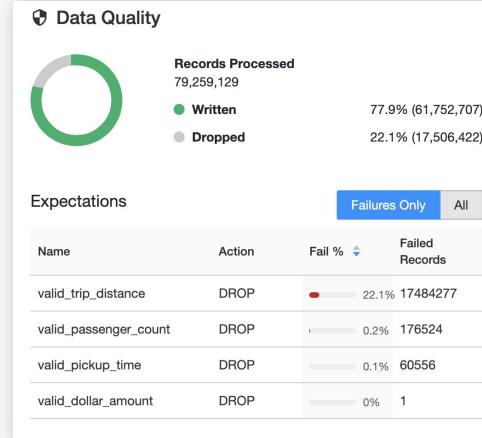
Unify batch and streaming

Get the simplicity of SQL with freshness of streaming with one **unified API**

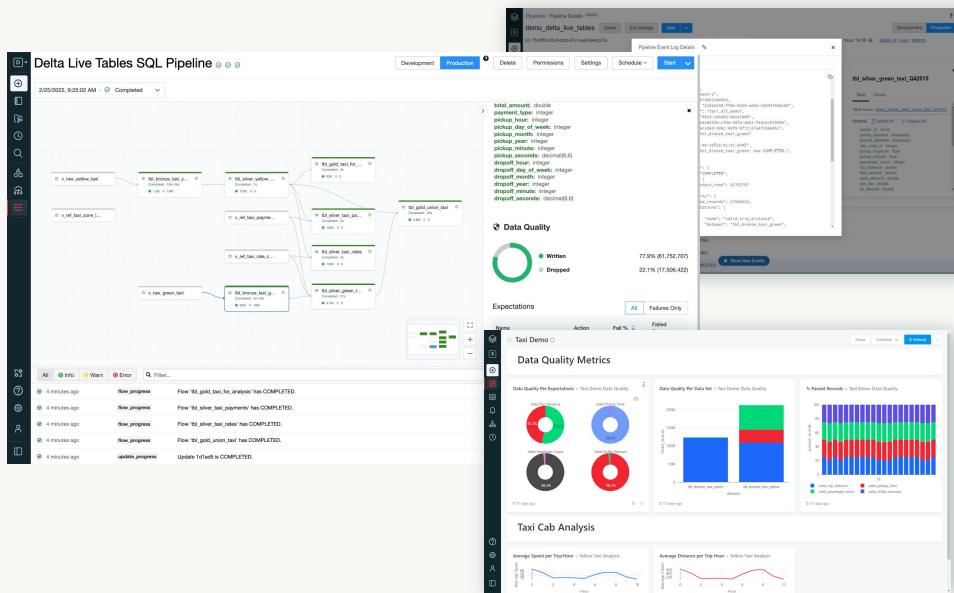
Data quality validation and monitoring

- Define data quality and integrity controls within the pipeline with data expectations
- Address data quality errors with flexible policies: fail, drop, alert, quarantine(future)
- All data pipeline runs and quality metrics are captured, tracked and reported

```
/* Stage 1: Bronze Table drop invalid rows */
CREATE STREAMING LIVE TABLE fire_account_bronze AS
( CONSTRAINT valid_account_open_dt EXPECT (account_open_dt is not null
and (account_close_dt > account_open_dt)) ON VIOLATION DROP ROW
COMMENT "Bronze table with valid account ids"
SELECT * FROM fire_account_raw ...
```



Data pipeline observability

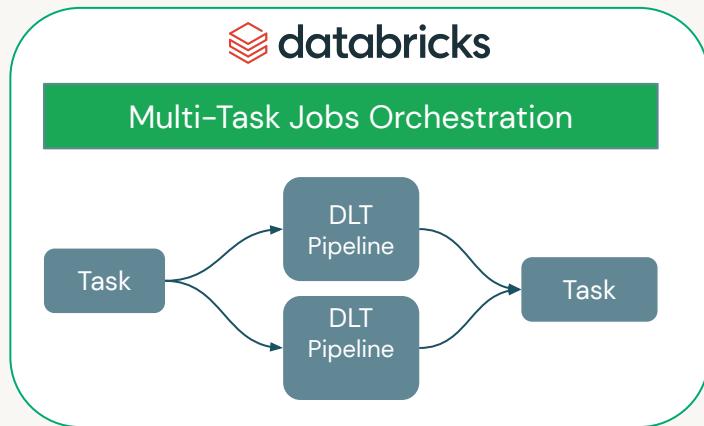


- High-quality, high-fidelity lineage diagram that provides visibility into how data flows for impact analysis
- Granular logging for operational, governance, quality and status of the data pipeline at a row level
- Continuously monitor data pipeline jobs to ensure continued operation
- Notifications using Databricks SQL



Workflow Orchestration

Simplify orchestration and management of data pipelines



- Easily **orchestrate** DLT Pipelines and tasks in the same DAG
- **Fully integrated** in Databricks platform, making inspecting results, debugging faster
- Orchestrate and manage workloads in **multi-cloud environments**
- You can run a Delta Live Tables pipeline as part of a data processing workflow with **Databricks jobs, Apache Airflow, or Azure Data Factory**.



Our Current Work at Occidental

Decline Curve Analysis (DCA) auto-forecasting: automated DCA forecast that updates the EUR forecasts on a monthly basis for all wells in the DJ basin (ca. 9,500) in under two minutes that is on a schedule using Azure Data Factory (ADF)

Econ Engine: Identify the most profitable/economic strategy for well completions

Data Pipeline: data pipelining that executes daily and curates and process DJ data (e.g., production, completion design, well characteristics, geology, operations, etc. all from disparate internal and external sources); critical for well completions strategies, general analysis



Thank you for attending!

