



Databricks 101

Getting Started Series

Crystal Chang



What is Databricks?

What is a Lakehouse?

How does it work?

What does it look like?

2

How can I get started?



10,000+
global customers

\$1.5B+
in revenue

\$4B
in investment

Inventor of the **lakehouse**
&
Pioneer of **generative AI**



databricks

The data and AI company

Gartner-recognized Leader
Database Management Systems
Data Science and Machine Learning Platforms

Creator of



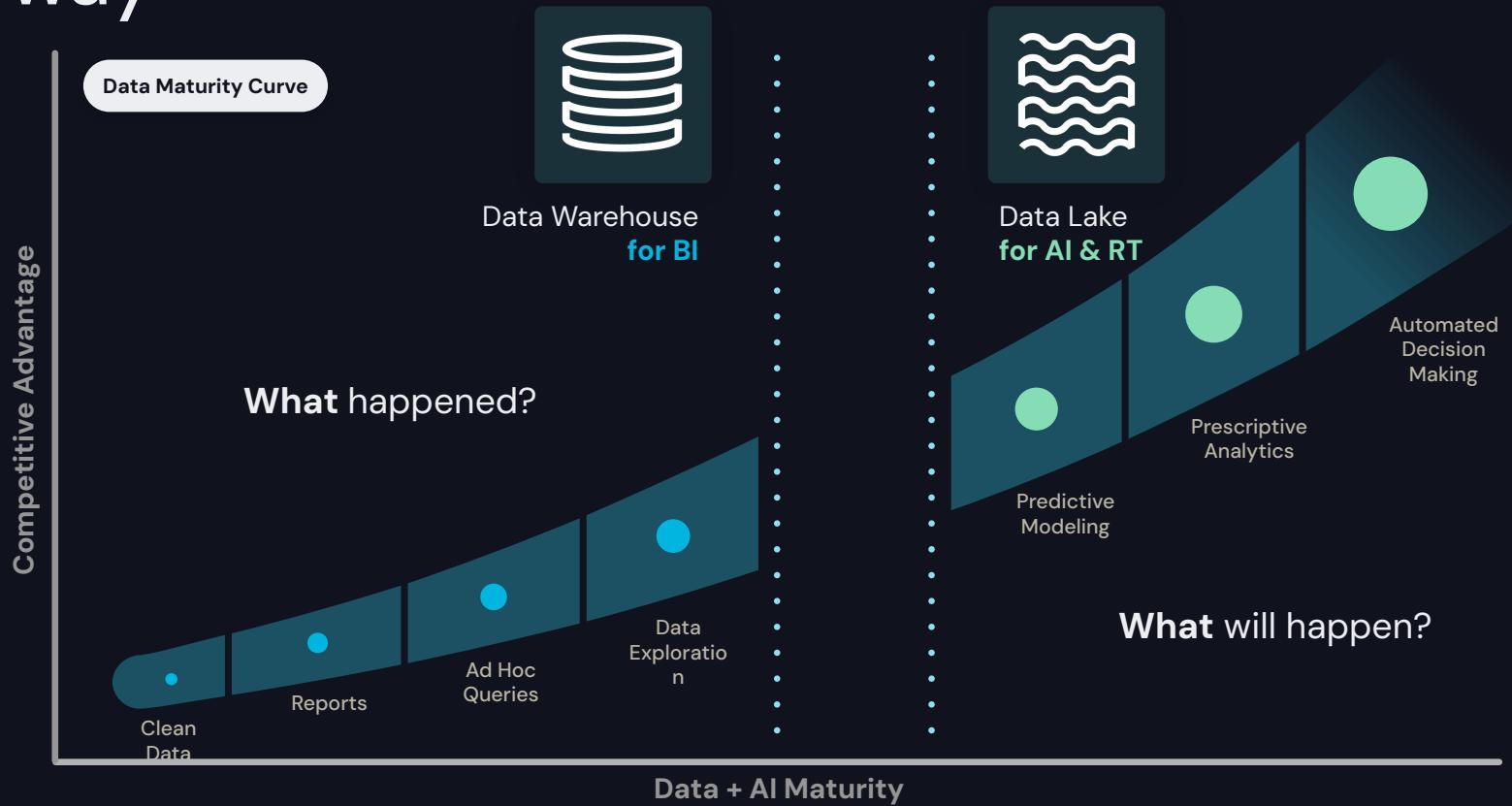
mlflow™



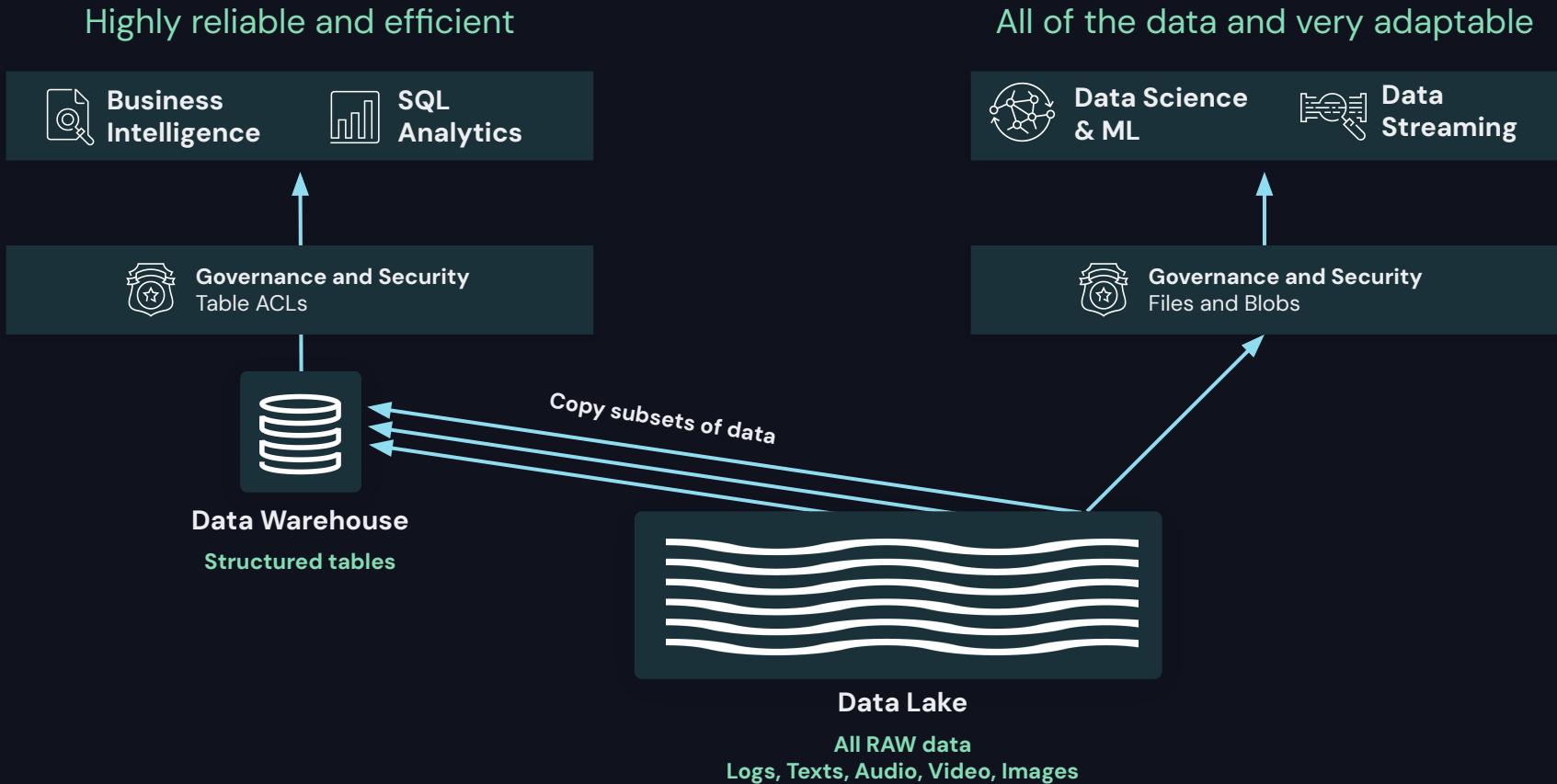
Lakehouse?



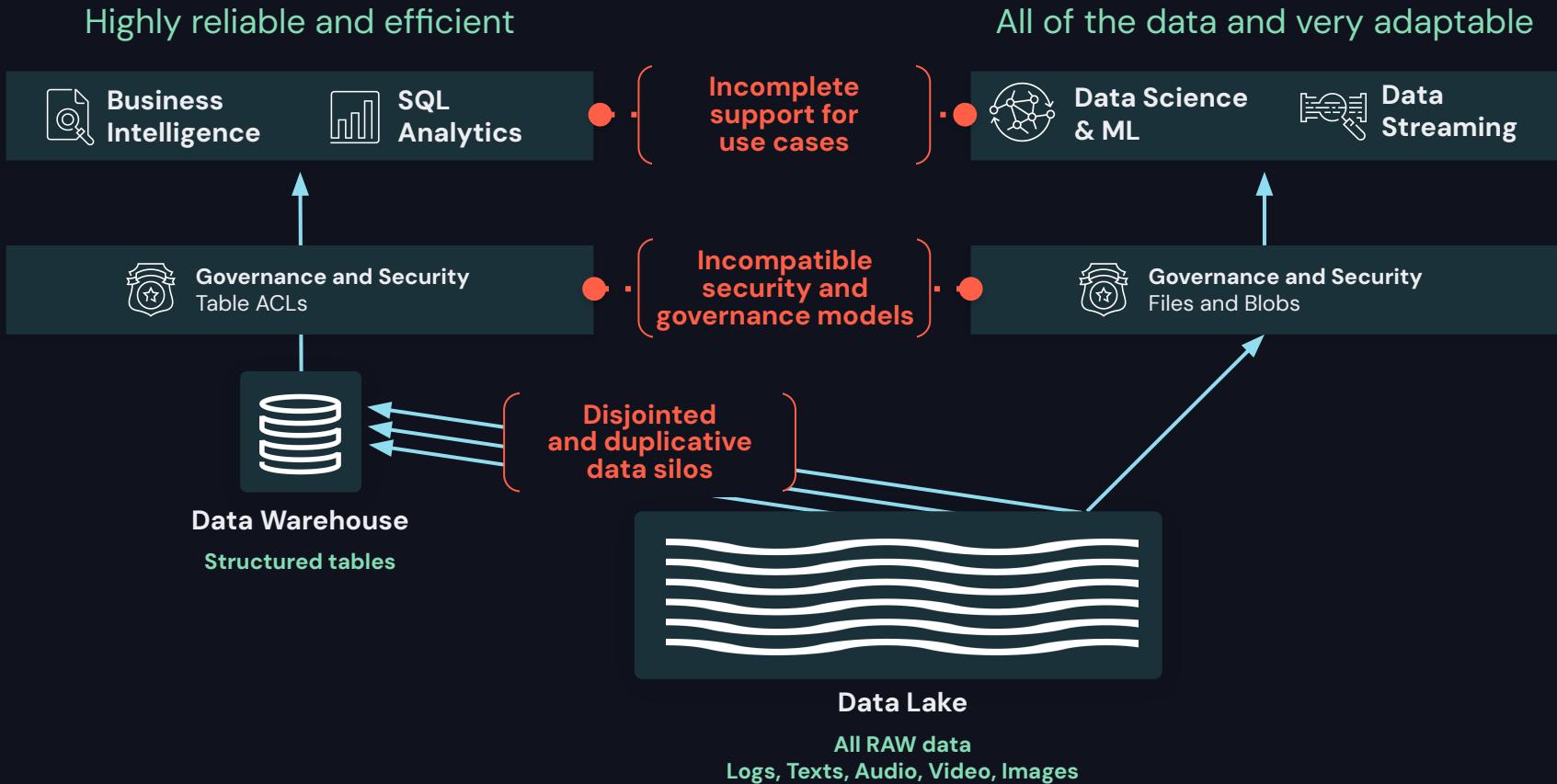
Two incompatible architectures get in the way



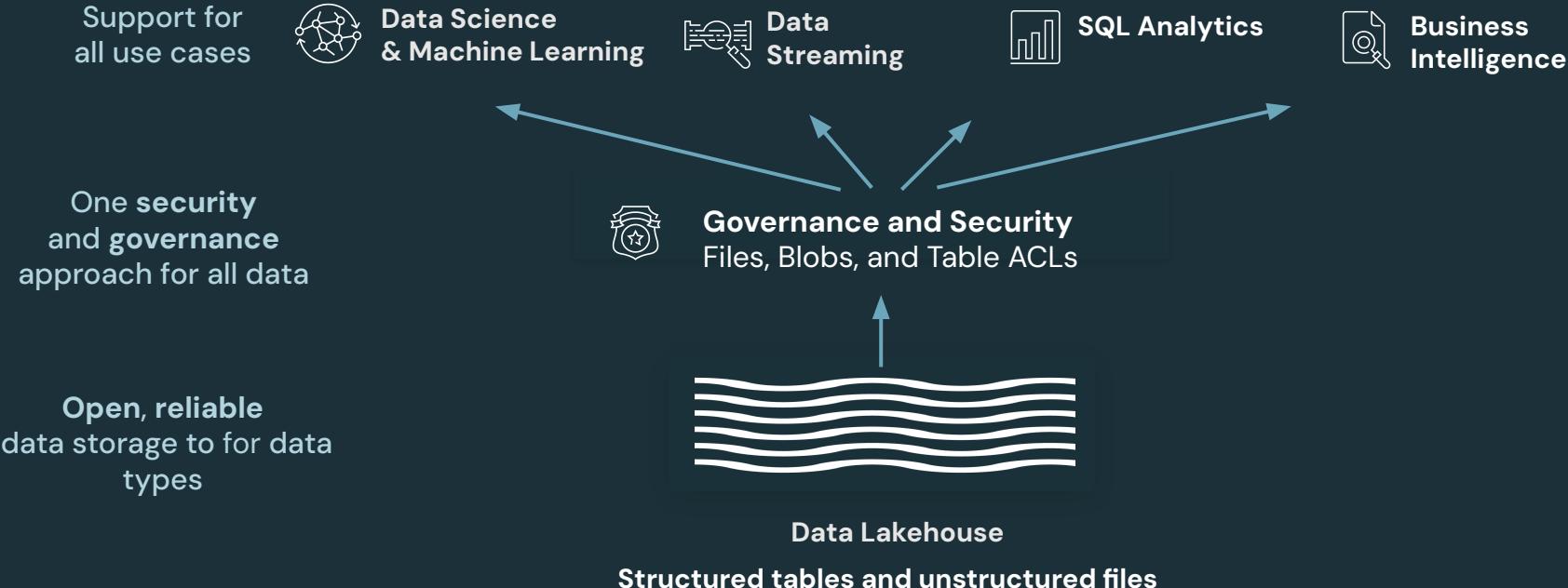
Almost all cloud deployments are two-tiered, Require 5+ Platforms



Complex duplicated architecture



Lakehouse simplifies data, analytics, and AI



The four functional areas and sub-disciplines of data, analytics, and AI

The 4Ds of Databricks



Databricks Data Intelligence Platform

An open, unified foundation for all your data

Data Science
& AI

Mosaic AI

ETL &
Real-time Analytics

Delta Live Tables

Orchestration

Workflows

Data
Warehousing

Databricks SQL

Unified security, governance, and cataloging

Unity Catalog

Unified data storage for reliability and sharing

Delta Lake

Open Data Lake

All structured, semi-structured, unstructured data
(Logs, texts, audio, video, images, etc.)

The Data Lakehouse

An open, unified foundation for all your data

Data Science
& AI

Mosaic AI

ETL &
Real-time Analytics

Delta Live Tables

Orchestration

Workflows

Data
Warehousing

Databricks SQL

Unified security, governance, and cataloging

Unity Catalog

Unified data storage for reliability and sharing

Delta Lake

Open Data Lake

All raw data
(Logs, texts, audio, video, images)

2020

Databricks pioneered
the lakehouse
architecture

TODAY

74% of global
enterprises have
adopted lakehouse

MIT Technology Review
Insights, 2023



Databricks Data Intelligence Platform

Data-centric AI

Gen AI

- Custom models
- Model serving
- RAG

End-to-end AI

- MLOps (MLflow)
- AutoML
- Monitoring
- Governance

Mosaic AI

Create, tune, and serve custom LLMs

Delta Live Tables

Automated data quality

Workflows

Job cost optimized based on past runs

Databricks SQL

Text-to-SQL

The AI powered data intelligence engine to understand the semantics of your data

DatabricksIQ

Unity Catalog

Securely get insights in natural language

Delta Lake

12

Data layout is automatically optimized based on usage patterns

Open Data Lake

All raw data
(Logs, texts, audio, video, images)



Databricks Data Intelligence Platform



"Project Genie"

Data and AI for all with natural language



How can we extend data and AI to everyone in the organization?

Mosaic AI

Create, tune, and serve custom LLMs

Delta Live Tables

Automated data quality

V

The AI powered data intelligence engine to understand

DatabricksSIG

Unity Catalog

Securely get insights in natural language

Delta Lake

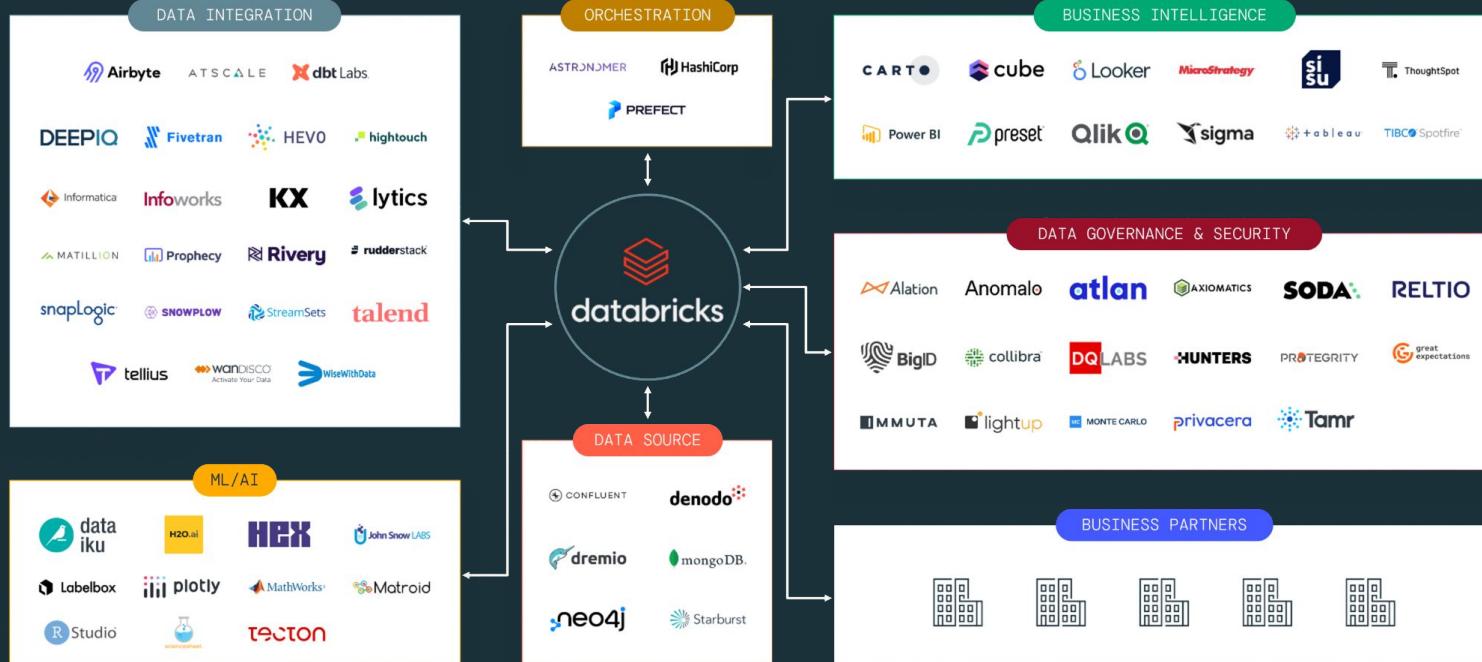
Data layout is automatically optimized for performance

Open Data Lake



Built on an open foundation

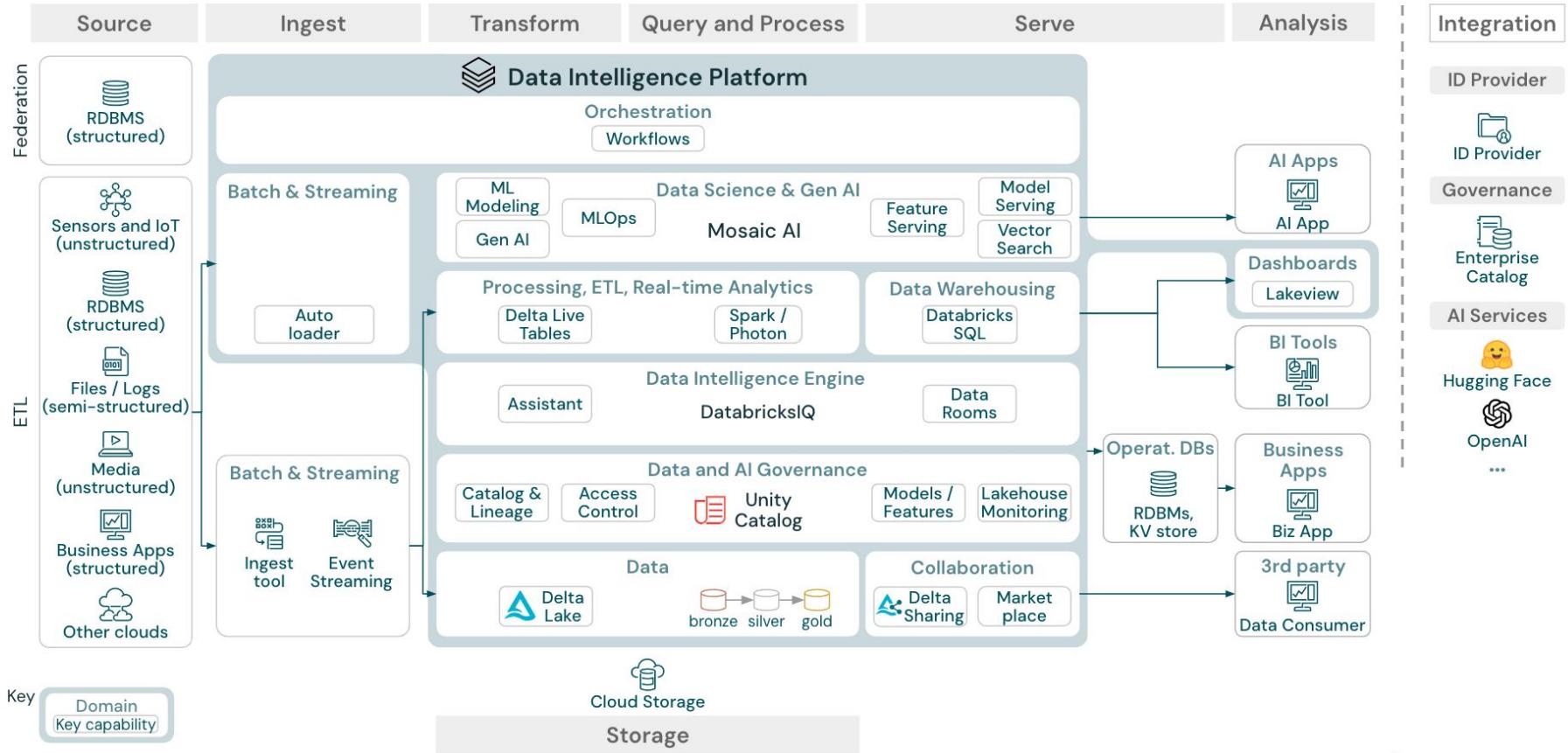
Easily integrate with the entire data and AI ecosystem



How does it fit in with my current architecture?



Databricks Data Intelligence Platform



Peeling the onion: How does it work?



Databricks Data Intelligence Platform

An open, unified foundation for all your data

Data Science
& AI

Mosaic AI

ETL &
Real-time Analytics

Delta Live Tables

Orchestration

Workflows

Data
Warehousing

Databricks SQL

Unified security, governance, and cataloging

Unity Catalog

Unified data storage for reliability and sharing

Delta Lake

Open Data Lake

All structured, semi-structured, unstructured data
(Logs, texts, audio, video, images, etc.)



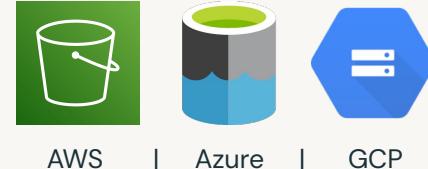


Delta Format

Batch
Streaming
Updates/Deletes



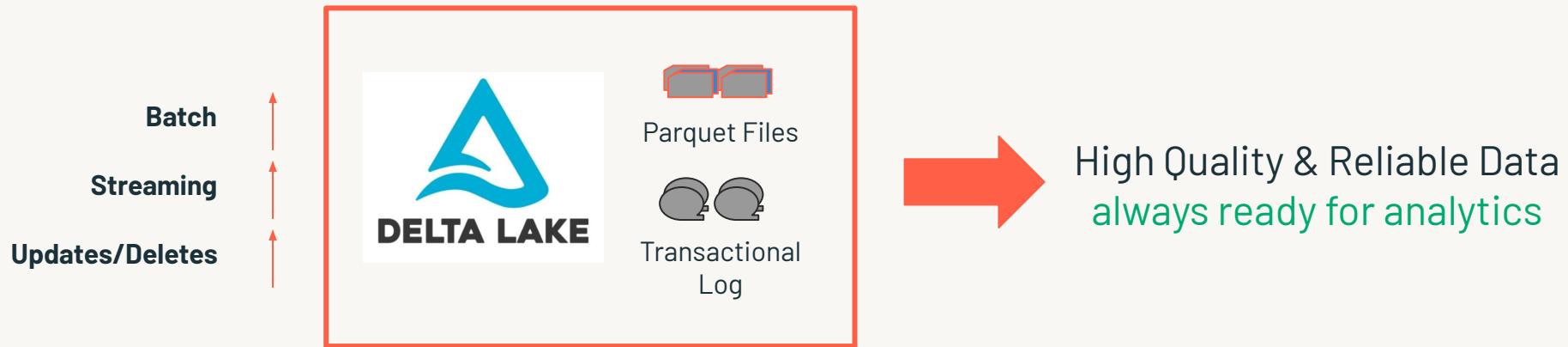
Transaction Log
Data Files



table_name1/
_delta_log/
0000.json
0001.json
....
file1.parquet
file2.parquet
....



Delta Lake ensures data reliability

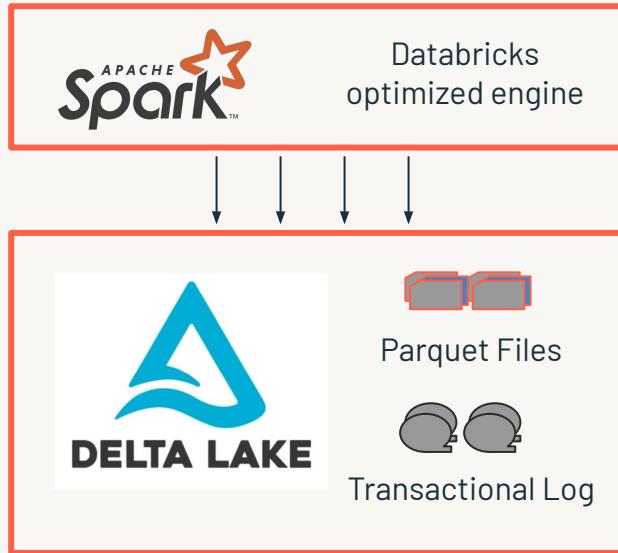


Key Features

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



Delta Lake optimizes performance



Highly Performant
queries at scale

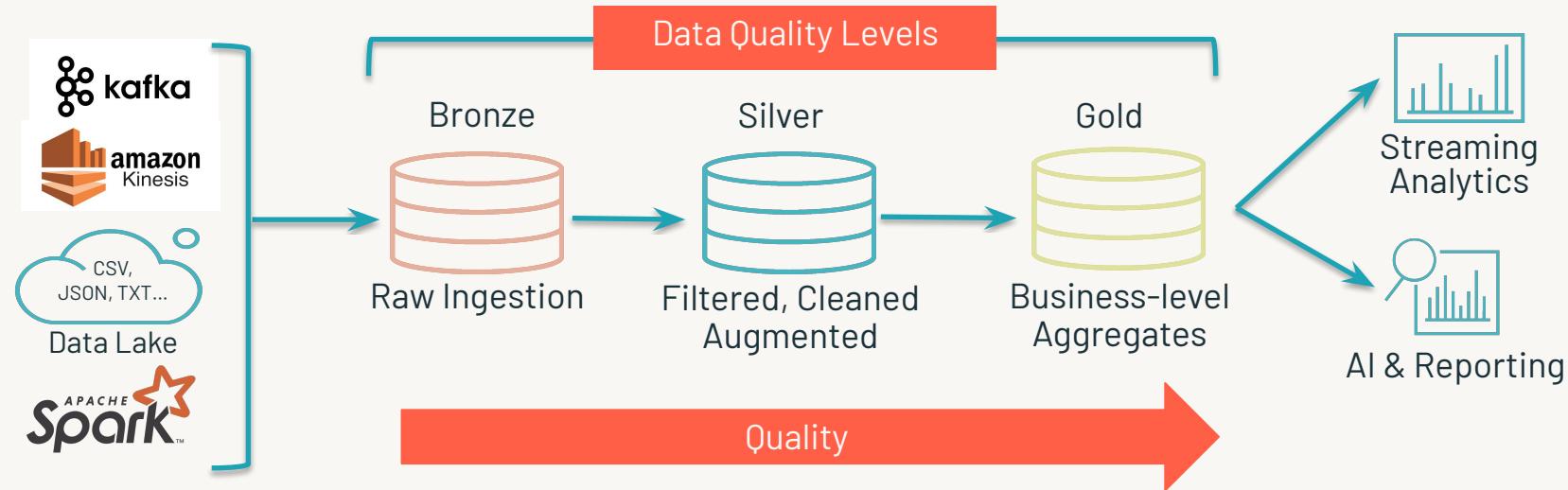
Automatically managed
through



Key Features

- Caching
- Compaction
- Data skipping
- Z-ordering

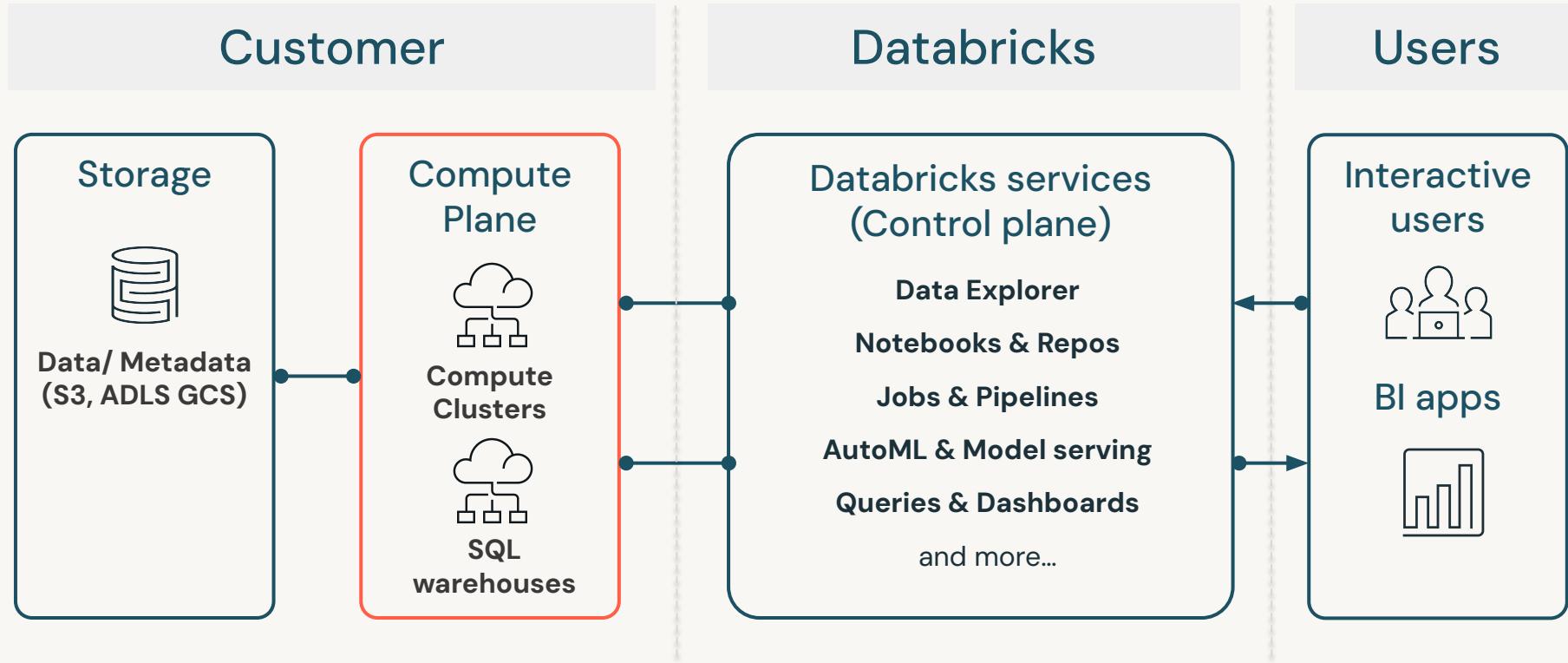
The Delta Lake



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

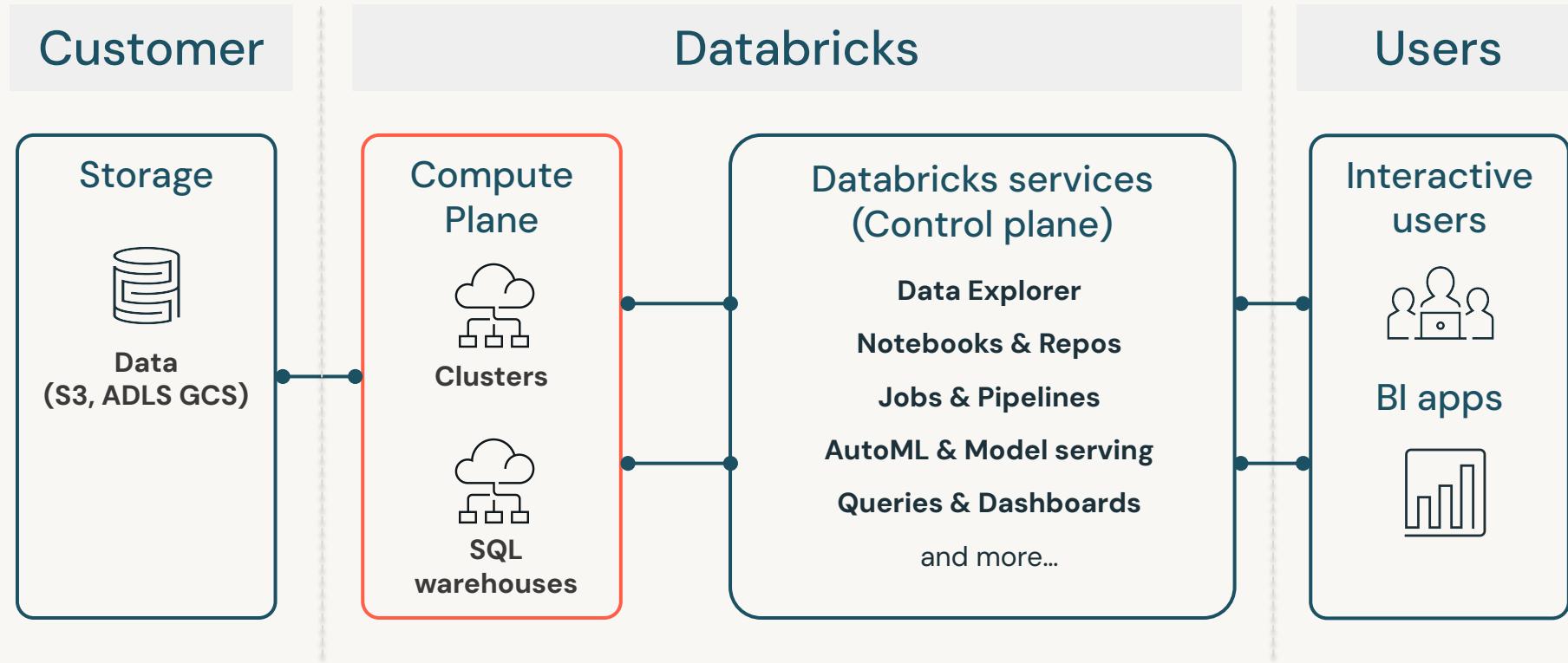
Compute on Databricks

Past and present architecture



How will Compute work in future?

Serverless Architecture



Databricks Data Intelligence Platform

An open, unified foundation for all your data

Data Science
& AI

Mosaic AI

ETL &
Real-time Analytics

Delta Live Tables

Orchestration

Workflows

Data
Warehousing

Databricks SQL

Unified security, governance, and cataloging

Unity Catalog

Unified data storage for reliability and sharing

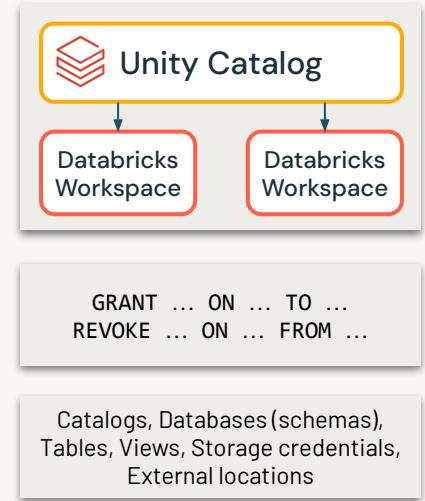
Delta Lake

Open Data Lake

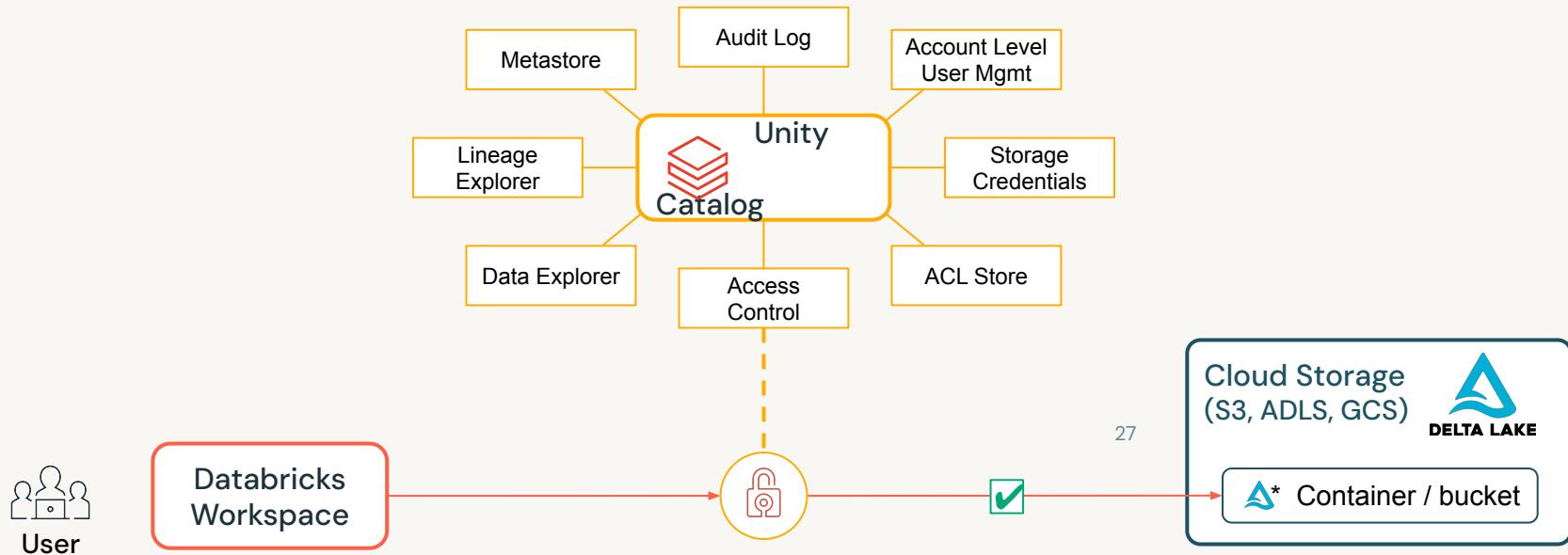
All structured, semi-structured, unstructured data
(Logs, texts, audio, video, images, etc.)

Unity Catalog – Key Capabilities

- Centralized metadata and user management
- Centralized data access controls
- Data lineage
- Data access auditing
- Data search and discovery
- Secure data sharing with Delta Sharing



Unity Catalog – Architecture



Centralized Access Controls

Centrally grant and manage access permissions across workloads

Using ANSI SQL DCL

```
GRANT <privilege> ON <securable_type>  
<securable_name> TO `<principal>`
```

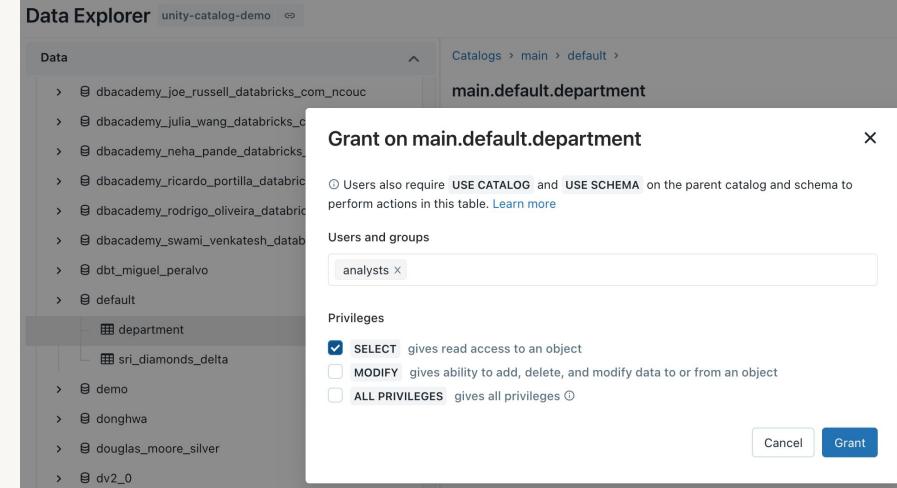
```
GRANT SELECT ON iot.events TO engineers
```

Choose permission level

'Table'= collection of files in S3/ADLS

Sync groups from your identity provider

Using UI



Row Level Security and Column Level Masking

Provide differential fine grained access to datasets

Only show specific rows

```
CREATE FUNCTION <name> (<parameter_name>  
<parameter_type> .. )  
RETURN {filter clause whose output must be a boolean}
```

```
CREATE FUNCTION us_filter(region STRING)  
RETURN IF(IS_MEMBER('admin'), true, region="US");
```

```
ALTER TABLE sales SET ROW FILTER us_filter ON region;
```

Test for group membership

Assign reusable filter to table

Specify filter predicates

Mask or redact sensitive columns

```
CREATE FUNCTION <name> (<parameter_name>,  
<parameter_type>, [, <column>...])  
RETURN {expression with the same type as the first  
parameter}
```

```
CREATE FUNCTION ssn_mask(ssn STRING)  
RETURN IF(IS_MEMBER('admin'), ssn, "****");
```

```
ALTER TABLE users ALTER COLUMN table_ssn SET MASK  
ssn_mask;
```

29

Test for group membership

Assign reusable mask to column

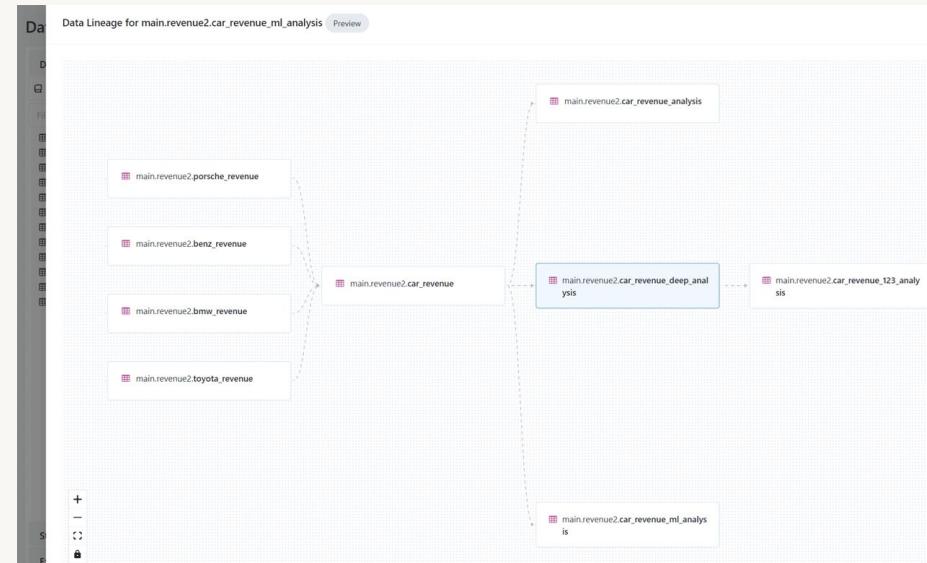
Specify mask or function to mask



Automated lineage for all workloads

End-to-end visibility into how data flows and consumed in your organization

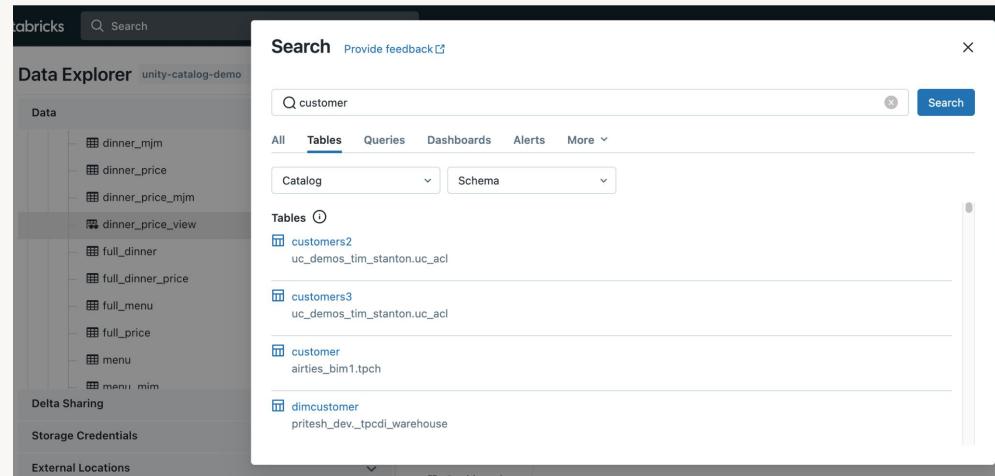
- Auto-capture runtime data lineage on a Databricks cluster or SQL warehouse
- Track lineage down to the table and column level
- Leverage common permission model from Unity Catalog
- Lineage across tables, dashboards, workflows, notebooks



Built-in search and discovery

Accelerate time to value with low latency data discovery

- UI to search for data assets stored in Unity Catalog
- Unified UI across DSML + DBSQL
- Leverage common permission model from Unity Catalog
- Apply semantic tags to data and search across tags



Let's take a test drive



Imagine I'm working for a family
restaurant chain in Singapore called
Just Eat Lah!



Craft my marketing campaign:

Send a *personalised* push notification to predicted churn users to get them to spend at least \$50 again



What if your AI could do more?

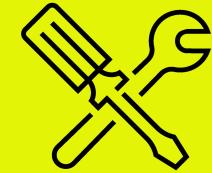
Look up information about a customer from your enterprise data

Create a customer support ticket

Return an order to your customer

File a JIRA issue

Execute a code snippet



Tools

ANNOUNCING

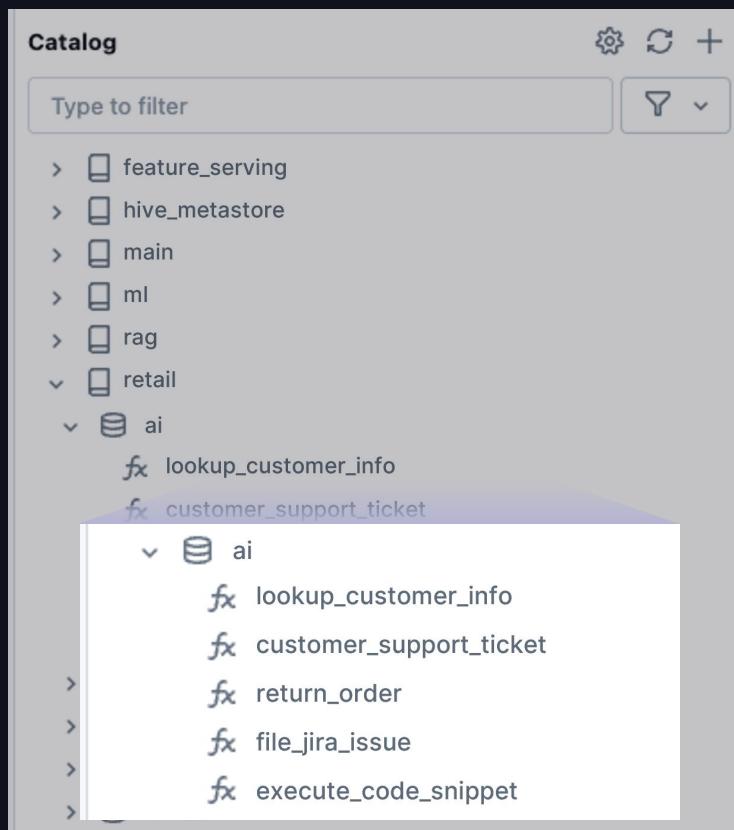
Mosaic AI Tool Catalog

Author, publish and
share enterprise AI tools

Leverage trusted compute
and credential management

Fully integrated
into Unity Catalog

Deploy 



The screenshot shows the Unity Catalog interface with a search bar and filter options at the top. Below is a tree view of tool categories:

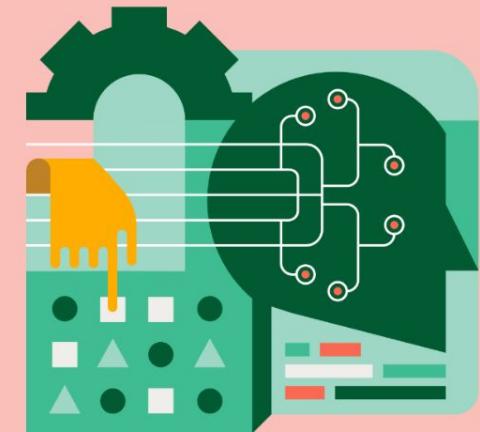
- feature_serving
- hive_metastore
- main
- ml
- rag
- retail
 - ai
 - lookup_customer_info
 - customer_support_ticket
- ai
 - lookup_customer_info
 - customer_support_ticket
 - return_order
 - file_jira_issue
 - execute_code_snippet

How do I get started?



Data + AI Professional Workshop Series

Build your data engineering, analytics, data science and AI skills



Join us at the upcoming sessions!

- Well Architected Lakehouse
- Data Engineering Optimization Best Practices
- Gen AI in Action: Build your first LLM App
- Security and Governance for Data + AI
- Simplify Your Streaming With Delta Live Tables
- Gen AI in Action: Accelerate LLM App to Production

Scan to register



Take a training course



Register for a free instructor-led
2h fundamentals course

- Data Engineering
- Data Analysis
- Machine Learning
- Platform Administration

<p>DATA ANALYST</p> <p>Get Started with Data Analysis on Databricks</p> <p>This content provides an introduction to Databricks SQL. Participants will learn about ingesting data, producing visualizations and dashboards, and...</p> <p> FREE @ 2H INSTRUCTOR-LED ONBOARDING</p>	<p>Get Started with Databricks for Business Leaders</p> <p>This course is designed to introduce Business Leaders to Databricks and the Databricks Lakehouse Platform.</p> <p> FREE @ 1H INSTRUCTOR-LED ONBOARDING</p>	<p>DATA ENGINEER</p> <p>Get Started with Databricks for Data Engineering</p> <p>In this course, you will learn basic skills that will allow you to use the Databricks Data Intelligence Platform to perform a simple data engineering workflow. You w...</p> <p> FREE @ 2H INSTRUCTOR-LED ONBOARDING</p>
<p>PLATFORM ADMINISTRATOR</p> <p>Get Started with Databricks Platform Administration</p> <p>This content serves as a starting point for platform administrators new to the Databricks Lakehouse Platform. It focuses on Unity Catalog and guides learners...</p> <p> FREE @ 2H INSTRUCTOR-LED ONBOARDING</p>	<p>MACHINE LEARNING PRACTITIONER</p> <p>Get Started with Machine Learning on Databricks</p> <p>Join us for this live, introductory session for data scientists and machine learning practitioners onboarding onto the Databricks Lakehouse Platform. ...</p> <p> FREE @ 2H INSTRUCTOR-LED ONBOARDING</p>	



Databricks Academy



Visit academy.databricks.com for:

- Foundational e-learning **free** for our customers
- Featuring curated self-paced curriculums for each persona
- For users who want to prove their knowledge of the platform, we have a number of certifications

Click [here](#) for registration instruction

The screenshot shows the Databricks Academy homepage. It features a dark header with the text "Databricks Academy". Below the header, there's a banner with the text "Earn your credentials" and "Select certification overview courses are now available free". A button says "Get started today". To the right, there's a section titled "RECOMMENDED" with a link to "Databricks Community".

This is a landing page titled "Welcome to Databricks Academy!". It includes a sub-headline: "To get started with your learning experience, please review the course 'Databricks Academy Guide' in the 'Enrolled Learning' section." The page is decorated with colorful geometric shapes.



Start Building!

Create your Databricks Free Trial Account

- How to Deploy your Databricks environment
 - [AWS Docs Video](#) | [Azure](#) | [GCP](#)

Connect with the Community

- Join the [Databricks Community](#) where you can discover latest features, collaborate with peers, get help from experts
- Customer-exclusive [Office Hours](#) connect you directly with experts through a live Q&A where you can ask all your Databricks questions
- Discover [events](#) in your area

The image features the Databricks logo at the top left. To its right is a screenshot of a code editor window titled "Predictive Maintenance" containing Python code for an MLflow experiment. Below the code editor is a "Dashboard" section titled "Raw Data" which displays a line chart with multiple colored lines representing different data series over time. A legend on the right side of the dashboard lists five stages: Pre-processing, Training, Evaluation, Inference, and Monitoring, each represented by a green checkmark icon.

databricks

**Databricks
Free Trial**

Predictive Maintenance

```
1 import mlflow
2 from hyperopt import SparkTrials
3
4 spark_trials = SparkTrials()
5 with mlflow.start_run():
6     argmin = fmin(
7         fn=objective,
8         space=search_space,
9         algo=algo,
10        max_evals=16,
11        trials=spark_trials)
```

Test drive the full Databricks platform free for 14-days

Databricks Solution Accelerators

Deliver data analytics and AI value faster

Fully functional set of resources for tackling the most common, high-impact use cases

Designed to help Databricks customers go from idea to proof-of-concept (PoC) in less than two weeks

Includes: Notebook, webpage, explainer video, blog, etc

Industry Solutions

Deliver the data and AI-driven outcomes that matter most — faster

[Start your free trial](#) [Schedule a demo](#)



Databricks Solution Accelerators

Save hours of discovery, design, development and testing with Databricks Solution Accelerators. Our purpose-built guides — fully functional notebooks and best practices — speed up results across your most common and high-impact use cases. Go from idea to proof of concept (PoC) in as little as two weeks.

Start using Solution Accelerators with your free Databricks trial or your existing account.

[Start your free trial today →](#)

80+ Solution Accelerators

 Financial Services	 Healthcare & Life Sciences	 Comm, Media & Entertainment	 Retail & Consumer Goods	 Manufacturing & Energy	 Cross-Industry & Public Sector
<ul style="list-style-type: none"> • Transaction Embeddings for Personalization • Model Risk Management • Risk Management (VaR) • Regulatory Reporting • Real-time Fraud Detection • Modern Investment Platform w/Time Series • ESG Performance Analytics • Smart Claims • NLP for Customer Service Analytics • ...and others 	<ul style="list-style-type: none"> • Abstracting RWE w/OMOP • Automated PHI Removal • Adverse Drug Event Detection • Genomic Pipelines (GWAS) • R&D Optimization w/Knowledge Graphs • Digital Pathology Image Analysis • FHIR & HL7 Interop. • Price Transparency • LLM for Biomedical Literature • LLM for Clinical Notes Summarization • ...and others 	<ul style="list-style-type: none"> • Multi-touch Attribution • Real-time Bidding Optimization • Media Mix Modeling • Sales Forecasting & Ad Attribution • Toxicity Detection in Gaming • Responsible Gaming • Video Quality of Experience • Subscriber Churn Prediction • Telco Network Analytics • ...and others 	<ul style="list-style-type: none"> • Customer Identify Resolution • Customer Segmentation • Propensity Scoring • Survival Analysis & LTV • Recommenders • Demand Forecasting • Real-time Point of Sale Analytics • On-shelf Availability • Safety Stock Analysis • Pricing Analytics with Redkite • LLM for Retail – Search, Review Summary, Image or Copy Generation • ...and others 	<ul style="list-style-type: none"> • Digital Twin • Predictive Maintenance (IoT) • OEE: Equipment Monitoring • Computer Vision Foundations • Part Forecasting • Quality Inspection w/Computer Vision • Supply Chain Optimization • Grid-Edge Analytics • Managing Recalls with Barcode Traceability • ...and others 	<p>LLM</p> <ul style="list-style-type: none"> • LLM – Customer Service & Knowledge Base • Product Search • Personalized Recommendations • Review Summarization • Visual Concept Design • Copy Generation <p>Cyber</p> <ul style="list-style-type: none"> • Splunk Connector • Threat Detection w/DNS • Incident Investigation using Graphistry • IOC Matching & Multi-cloud Federation <p>Public Sector</p> <ul style="list-style-type: none"> • Entity Resolution • Automated Record Linking

<https://www.databricks.com/solutions/accelerators>



80+ Solution Accelerators

 Financial Services	 Healthcare & Life Sciences	 Comm, Media & Entertainment	 Retail & Consumer Goods	 Manufacturing & Energy	 Cross-Industry & Public Sector
<ul style="list-style-type: none"> Merchant Classification for Personalization Transaction Embedding for Personalization Model Risk Management Risk Management (VaR) Regulatory Reporting Reputation Risk Geospatial Analytics to Identify Fraud Real-time Financial Fraud Protection Anti-Money Laundering Modern Investment Platform w/Time Series ESG Performance Analytics Smart Claims NLP for Customer Service Analytics LLM for Customer Service & Knowledge Base 	<ul style="list-style-type: none"> Abstracting RWE: OMOP, PSM, and Survival Analysis Real-World Data Extraction w/NLP (eg., for Oncology) Automated PHI Removal Adverse Drug Event Detection Genome-Wide Association Studies (Target Identification) Cohort Building w/Knowledge Graphs R&D Optimization w/Knowledge Graphs Digital Pathology Image Analysis Patient / Disease Risk Prediction Disease Profiling with TCGA FHIR Interoperability Ingest Streaming HL7 for Patient Analytics Medicare Risk Adjustment Price Transparency Social Determinants of Health HEDIS Engine with ApolloMed Scaling Geospatial Nearest Neighbor Searches LLM for Biomedical Literature LLM for Clinical Notes Summarization 	<ul style="list-style-type: none"> Multi-touch Attribution Real-time Bidding Optimization Media Mix Modeling Responsible Gaming Sales Forecasting & Ad Attribution Toxicity Detection in Gaming Video Quality of Experience Subscriber Churn Prediction Telco Customer Churn Prediction Telco Network Analytics Enhancing CDP (Amperity) with Personalized Recommendations LLM for Customer Service & Knowledge Base 	<ul style="list-style-type: none"> Customer Identify Resolution Customer Lifetime Value Customer Segmentation Recency, Frequency & Monetary (RFM) Segmentation Propensity Scoring Survival Analysis & LTV Recommendation Engines Retention Management Demand Forecasting Real-time Point of Sale Analytics On-shelf Availability Order Picking Optimization Scalable Route Generation Item / Product (Fuzzy) Matching Safety Stock Analysis Pricing Analytics with Redkite Enhancing CDP (Amperity) with Personalized Recommendations LLM for Customer Service & Knowledge Base LLM for Retail - Product Search, Recommendations, Product Review Summary 	<ul style="list-style-type: none"> Digital Twin Predictive Maintenance (IoT) OEE: Equipment Monitoring Computer Vision Foundations Part Forecasting Quality Inspection w/Computer Vision Supply Chain Optimization Grid-Edge Analytics Computer Vision for Power & Utility Inspection Managing Recalls with Barcode Traceability LLM for Manufacturing – Q&A over custom datasets 	<p>LLM</p> <ul style="list-style-type: none"> LLM for Customer Service & Knowledge Base Product Search Personalized Recommendations Product Review Summarization Visual Concept Design Product Copy Generation <p>Cyber</p> <ul style="list-style-type: none"> Cyber Analytics (Splunk Connector) Threat Detection with DNS Incident Investigation using Graphistry IOC Matching and Multicloud Query Federation <p>Public Sector</p> <ul style="list-style-type: none"> Entity Resolution Automated Record Liking

<https://www.databricks.com/solutions/accelerators>



Learning Resources

- [Databricks Academy](#) – See these instructions on how to access and view self-paced training videos.
 - [Course Catalog](#) Create an account using your company email address
- [Databricks Academy Github](#) – Get the code to follow along
- [Databricks Certifications](#) – Get certified using Databricks
- [Databricks Demo Hub](#) – Watch short demos of Databricks products!
- [DBDemos.AI](#) – Install demos in your workspaces with best practices, ready to go.
- [Instructor Led Training](#) (if purchased) – be sure to check out the [learning paths!](#)



Useful Resources

Developing on Databricks with Python		ETL & Streaming	
Developing with Python in Databricks	AWS , Azure	Delta Lake Official Documentation	
Uploading Python Libraries	AWS , Azure	Delta Lake Best Practices	
Visualizations in Python	AWS , Azure	Structured Streaming Guide	
Introduction to DataFrames	AWS , Azure	Simplify Streaming Stock Data Analysis Using Databricks Delta	
Pandas User-Defined Functions	AWS , Azure	Designing ETL Pipelines with Structured Streaming and Delta Lake	
Migrate Single Node Workloads to Databricks	AWS , Azure	Workflows and Jobs	
Databricks Connect	AWS , Azure	Databricks Workflows	
Pandas APIs on Apache Spark	Apache Spark Docs , AWS , Azure	Databricks Jobs	
Using BI tools with Databricks		Delta Live Tables	
Connecting Business Intelligence Applications	AWS , Azure		
SQL on Databricks	AWS , Azure		





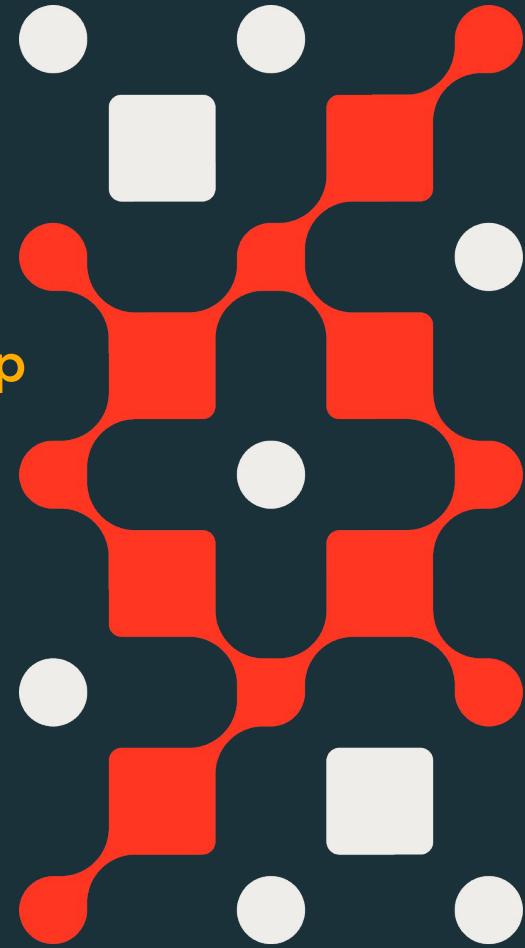
Generative AI World Cup

So you think you can hack

bit.ly/genai-world-cup-roadmap

Ready to launch your AI game to new heights? Join this virtual hackathon and stand to win amazing prizes, including \$50,000 in total cash prizes, featured blogs and potential speaking opportunity at events, and a world cup trophy.

SUBMISSIONS DUE **OCTOBER 18 2024**





Live Q&A





Appendix

Crystal Chang

28 March 2024



ANNOUNCING

Mosaic AI Model Training Fine-tuning

No-code fine-tuning
based on OSS models

Serve on Databricks
in one click

The screenshot shows the Mosaic AI Model Training interface. On the left, there are four sections: General, Training data, Model registration, and Advanced options >. The General section is currently active, displaying three task options: Chat Completion (selected), Continued Pre-training, and Instruction Finetuning. Below these is a "Select Foundation Model" dropdown set to "Llama 3 8B Instruct". A detailed view of this dropdown is shown in a modal window on the right, listing various models under categories: DBRX (DBRX-Base, DBRX-Instruct), Llama 3 (Llama 3 70B Instruct, Llama 3 70B, Llama 3 8B Instruct, Llama 3 8B). The "Llama 3 8B Instruct" option is selected.

General

Task

- Chat Completion
Finetune your model on chat logs between a user and an AI assistant
- Continued Pre-training
Train your model with additional text data to add new knowledge to a model
- Instruction Finetuning
Finetune your model on structured prompt-response data to adapt the model to a new task

Select Foundation Model
Models trained in Model Training may be subject to license requirements and/or use policies. [Learn more](#)

Llama 3 8B Instruct

Search

codellama/codellama-70b

Select Foundation Model
Models trained in Model Training may be subject to license requirements and/or use policies. [Learn more](#)

Llama 3 8B Instruct

Search

codellama/codellama-70b

DBRX

DBRX-Base
databricks/dbrx-base

DBRX-Instruct
databricks/dbrx-instruct

Llama 3

Llama 3 70B Instruct
meta-llama/Meta-Llama-3-70B-Instruct

Llama 3 70B
meta-llama/Meta-Llama-3-70B

✓ Llama 3 8B Instruct
meta-llama/Meta-Llama-3-8B-Instruct

Llama 3 8B
meta-llama/Meta-Llama-3-8B

Mosaic AI Model Training Pre-training

Build custom LLMs with
your enterprise data at
10x lower costs



Your
Data

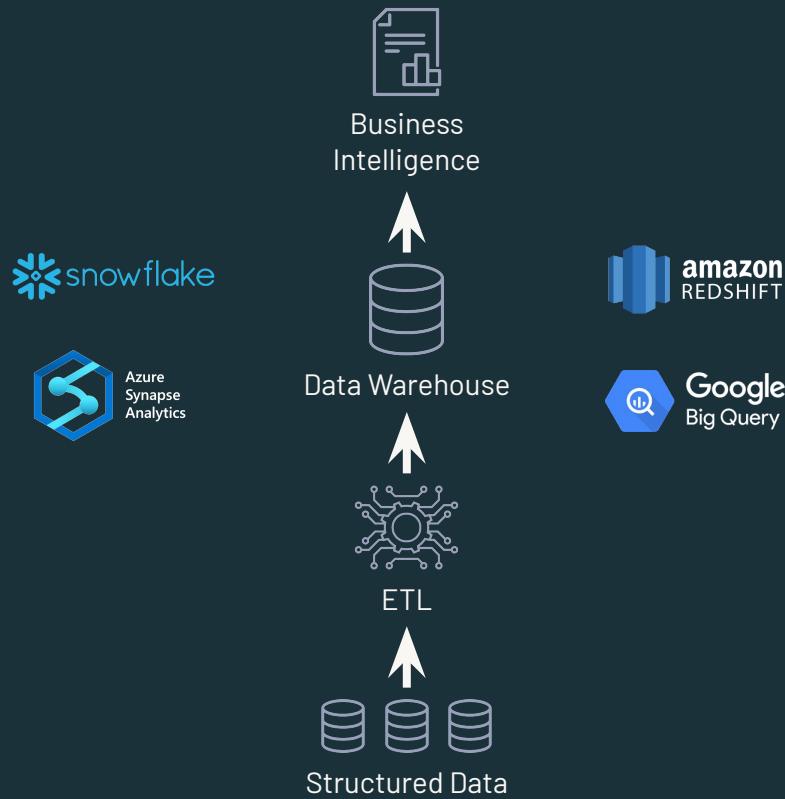


**Mosaic AI Model
Training**



Your
Model

Build



Data Warehouses

Pros

- Great for Business Intelligence (BI) applications

Cons

- Limited support for Machine Learning (ML) workloads
- Proprietary systems with only a SQL interface

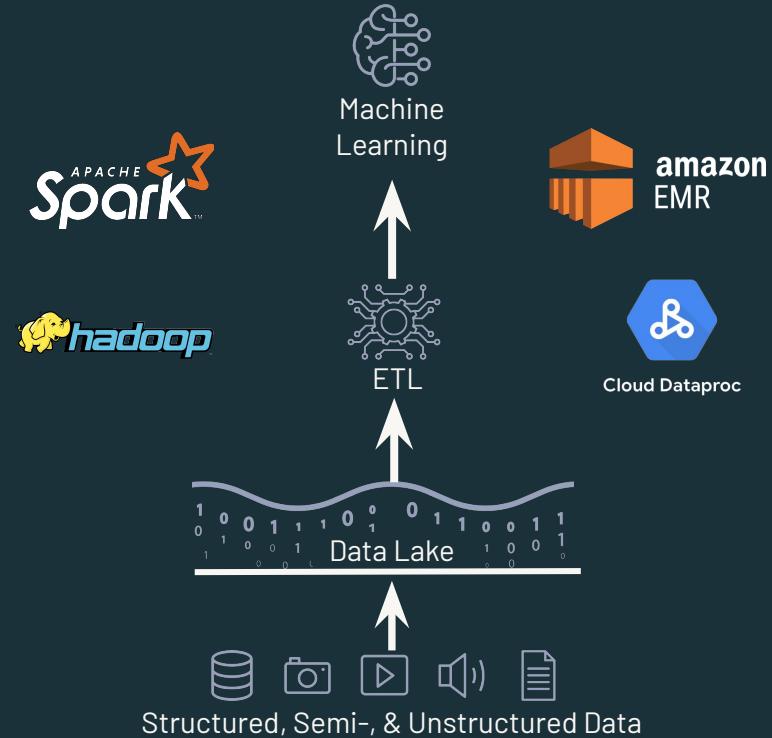
Data Lakes

Pros

- Supports ML
- Open formats and big ecosystem

Cons

- Poor support for BI
- Complex data quality problems





7-Eleven uses Databricks to build customer-centric solutions that drive revenue from personalization and the optimization of supply chain operations.

Challenge

- Data silos made cross-team collaboration difficult
- Analysts were working 16 hour days to respond to business demands
- Data infrastructure couldn't match the speed of the business

Solution

- Lakehouse democratizes access to data and expands DSML teams' capacity
- Analysts rapidly execute queries using Databricks SQL with reliable and complete real-time data
- The elimination of complexity improves data quality and reliability

Impact

\$109M

in accelerated revenue through Customer 360

\$3M

saved annually in cloud compute costs

35%

increase in data team productivity



Grab uses Databricks to collaborate, experiment, and develop more innovative features to continually enhance consumer-centric experiences.

Use Case

- Personalization and targeted marketing campaigns
- Inconsistent understanding of their customer: Disparate data teams, different products based on various customer segmentation

Why Databricks?

- Databricks powers Grab's in-house Consumer 360 platform, enabling them to develop more innovative features to continually enhance customer-centric experiences
- Delta Lake ingests and optimizes 1000s of user-generated signals and data sources from various applications, to enhance data integrity and security

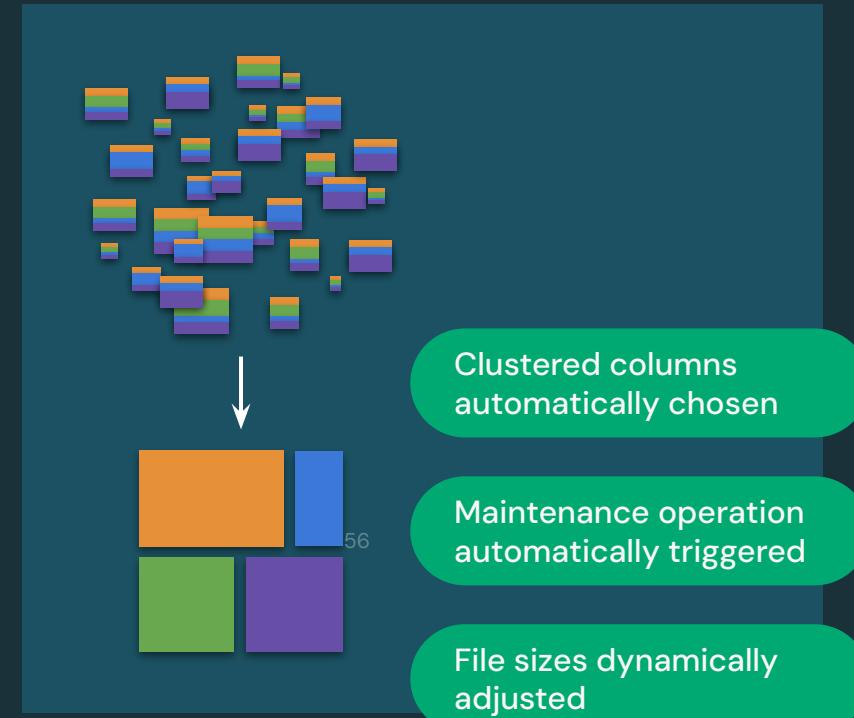
Impact

- Gained a unified understanding of customers to **personalize recommendations to millions of users everyday**
- New customer features can be developed faster, **lowering the costs of experimentation**, and **accelerating innovation**

Intelligence simplifies data management

Easy, automated data management with Delta Lake

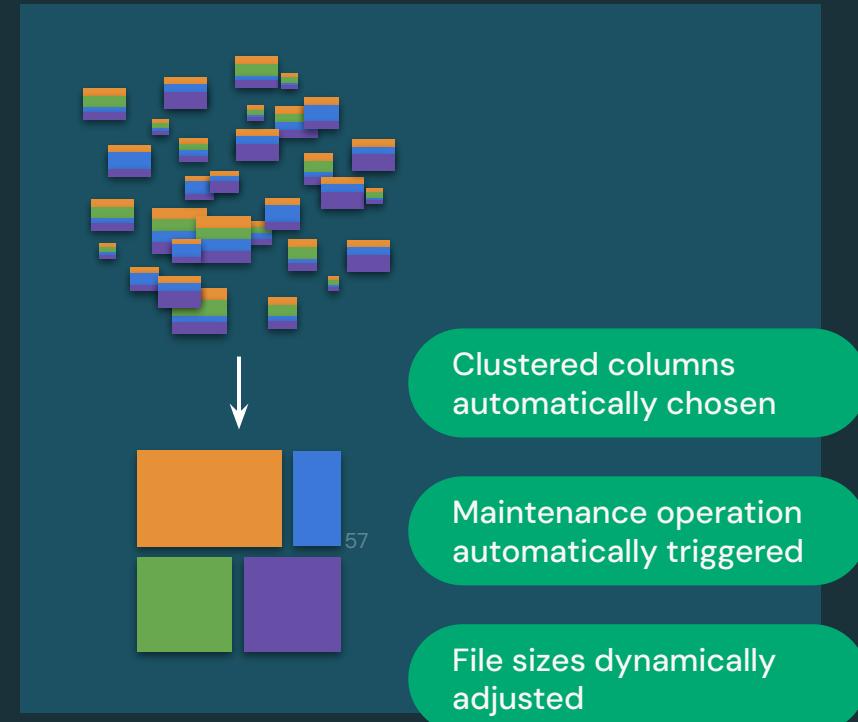
- Focus on your **data application**, the intelligence engine tunes your data
- AI-optimized data layout for **enhanced performance** in SQL warehouses and clusters
- New data clustered automatically and incrementally optimized to **reduce TCO**
- Adapts to new usage to **reduce risk over time**



Intelligence simplifies data management

Easy, automated data management with Delta Lake

- Focus on your **data application**, the intelligence engine tunes your data
- AI-optimized data layout for **enhanced performance** in SQL warehouses and clusters
- New data clustered automatically and incrementally optimized to **reduce TCO**
- Adapts to new usage to **reduce risk over time**



AI-powered Warehousing is fast and simple

Simple

Text-to-SQL

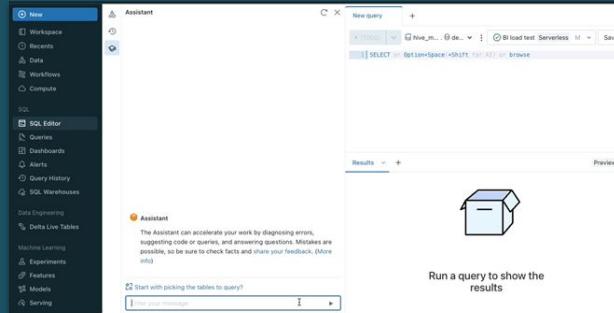
Transforms natural language into actionable insights

Powered by UC

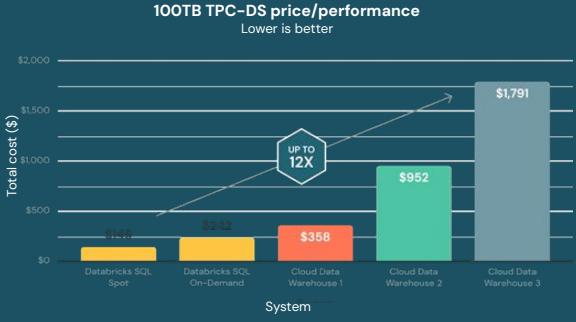
Understands your data, lineage, and dependencies

AI-powered assistant

Codes for you, Identifies issues and provides solutions



The screenshot shows the Databricks SQL Editor interface. On the left is a sidebar with options like Workspace, Data, Workflows, Compute, and SQL. The main area has tabs for New, Assistant, and New query. The New query tab shows a code editor with a query: `SELECT * FROM optionSpace`. Below the code editor are 'Results' and 'Preview' tabs. A large callout box labeled 'Assistant' provides information about the AI feature, including a link to learn more. At the bottom right, there's a button to 'Run a query to show the results'.



The chart compares the total cost for processing 100TB of data using different systems. The Y-axis represents 'Total cost (\$)' from \$0 to \$2,000. The X-axis lists the systems: Databricks SQL Spot, Databricks SQL On-Demand, Cloud Data Warehouse 1, Cloud Data Warehouse 2, and Cloud Data Warehouse 3. An arrow points from the Databricks SQL On-Demand bar to a callout box stating 'UP TO 12X'. The bars are color-coded: yellow for spot and on-demand, red for Cloud Data Warehouse 1, green for Cloud Data Warehouse 2, and blue for Cloud Data Warehouse 3.

System	Total cost (\$)
Databricks SQL Spot	\$140
Databricks SQL On-Demand	\$342
Cloud Data Warehouse 1	\$358
Cloud Data Warehouse 2	\$952
Cloud Data Warehouse 3	\$1,791

Fast and affordable

Instant queries

Auto-tuning and predictive IO, scale with data size

Intelligent, serverless workload management

Instant stop and start, route queries to optimize utilization and provide the best TCO

Orchestration with intelligence unlocks data power

Orchestrate and automate anything, anywhere, anytime



Serverless: instantly trigger ingestion, transform, refresh BI and retrain model altogether



Simple and accessible with **AI-based debugging**

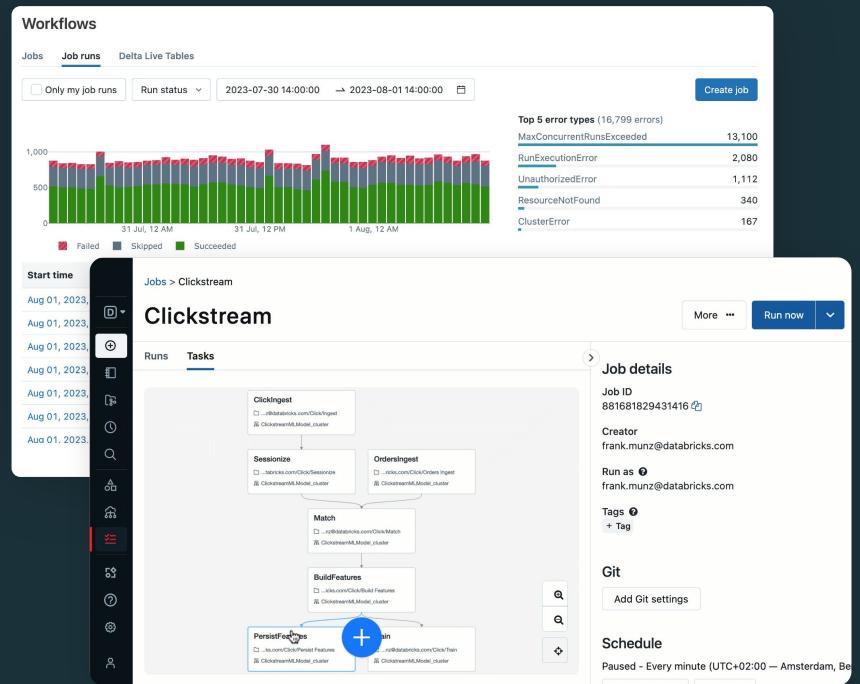


Reliable with automatic checkpoint and recovery

Automatic resource allocation to **reduce TCO**



Trust your execution with automatic monitoring and smart alerting (ex: job slower than usual)



Intelligence ensures robust and reliable ETL

Delta Live Table: declare your table with SQL / Python, let the engine be the expert

Robust pipeline for **Citizen Data Engineers**, code creation and debugging **assisted by AI**

