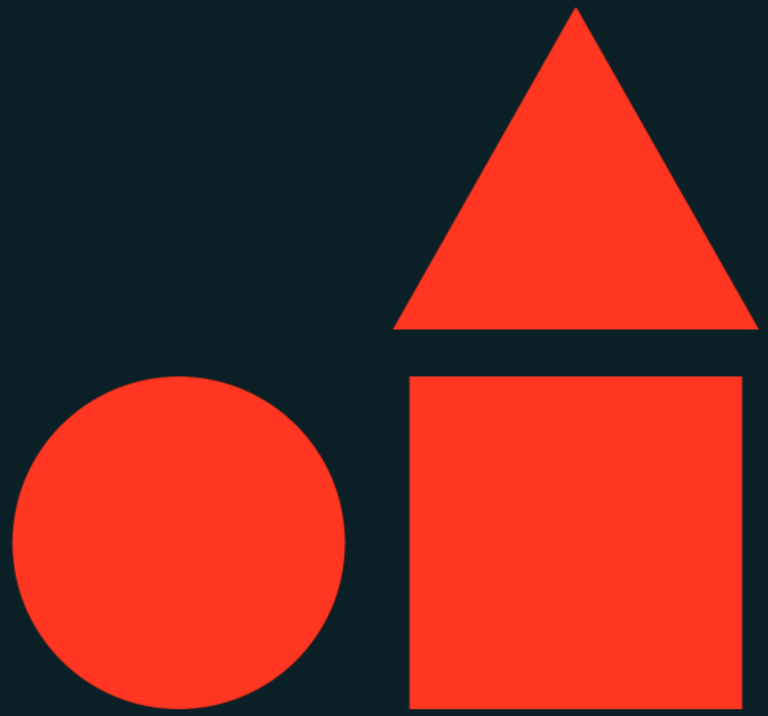
Get Started with Databricks for Data Warehousing



Databricks Academy
August 2025

Agenda

Databricks Overview	Time	Lecture	Demo	Lab
Databricks Data Intelligence Platform	8 mins	✓		
Databricks Workspace Walkthrough	15 mins		✓	

Agenda

Using Databricks for Data Warehousing	Time	Lecture	Demo	Lab
Introduction to Data Warehousing with Databricks	10 mins	✓		
Databricks SQL Warehouses	5 mins	✓		
Delta Lake Overview	8 mins	✓		
Using Delta Lake features with Databricks SQL	5 mins		✓	

Agenda

Data Ingestion and Transformation	Time	Lecture	Demo	Lab
Delta Lake UniForm	5 mins	✓		
Data Ingestion Techniques Overview	5 mins	✓		
Data Ingestion Techniques	10 mins		✓	
Data Transformation	8 mins	✓		
Exploring Data Transformation in Databricks	10 mins		✓	

Agenda

Data Orchestration and Querying Capabilities	Time	Lecture	Demo	Lab
Orchestration in Databricks	5 mins	✓		
Setting Up and Managing Serverless Lakeflow Jobs	15 mins		✓	
Databricks Querying Capabilities	5 mins	✓		

Agenda

Data Presentation	Time	Lecture	Demo	Lab
Introduction to AI/BI	5 mins	✓		
Creating a Dashboard in Databricks	5 mins		✓	
Creating Genie Spaces	5 mins		✓	
Data Warehousing Comprehensive Lab	30 mins			✓

Lab Exercise Environment

Technical Details

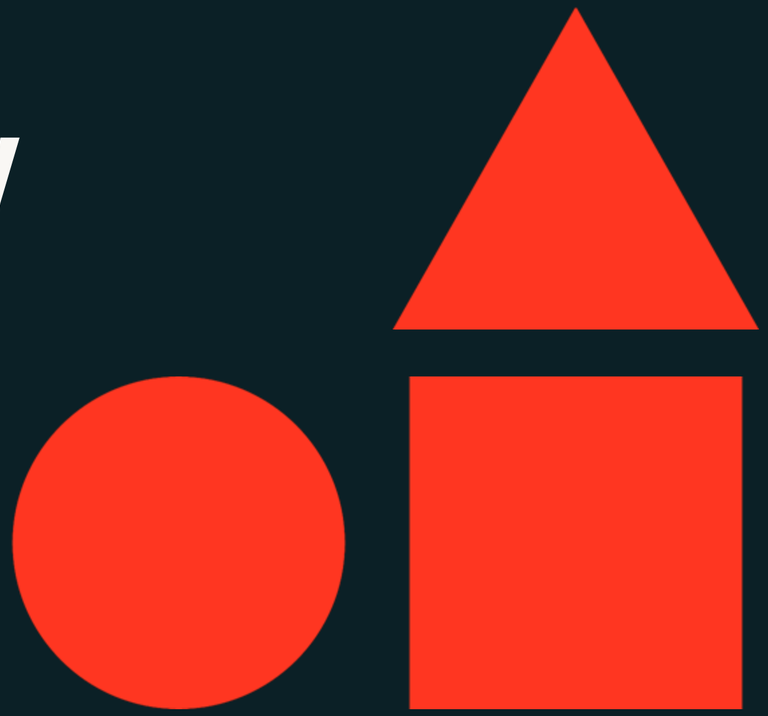
- Your lab environment is provided by Vocareum.
- It will open in a new tab.
- It has been configured with the permissions and resources required to accomplish the tasks outlined in the lab exercise.
- Third party cookies must be enabled in your browser for Vocareum's user experience to work properly.
- Make sure to enable pop ups!

Course Learning Objectives

- Describe the available compute options for workloads performed on the Databricks Data Intelligence Platform.
- List the products and features Databricks offers for different data-centric needs within the Databricks Platform.
- Navigate the Databricks Workspace UI.
- Use Databricks to complete common data warehousing tasks.
- Explain the purpose of Delta Lake for data warehousing purposes.
- Describe and apply various techniques for ingesting and transforming data in Databricks.
- Explain how Databricks support data orchestration needs within the platform.
- Describe the impact of data warehousing work on Databricks AI/BI needs.
- Identify the AI supported features in the Databricks platform.

databricks

Databricks Overview



Get Started with Databricks for Data Warehousing

Learning Objectives

- List the four key product lines available from Databricks.
- Describe how each product line serves its respective audience.
- Navigate the Databricks Workspace UI to locate key features and functionalities.

Agenda

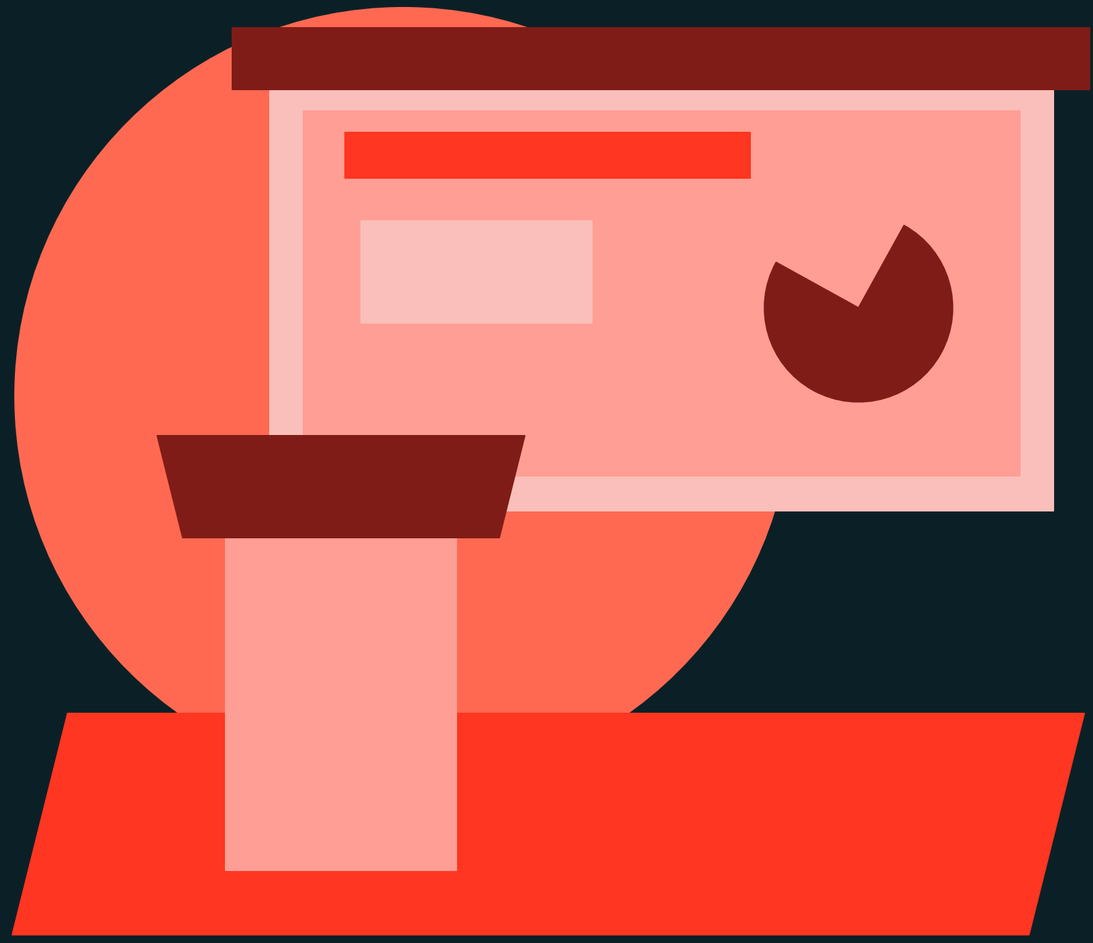
Databricks Overview	Time	Lecture	Demo	Lab
Databricks Data Intelligence Platform	8 mins	✓		
Databricks Workspace Walkthrough	15 mins		✓	

databricks

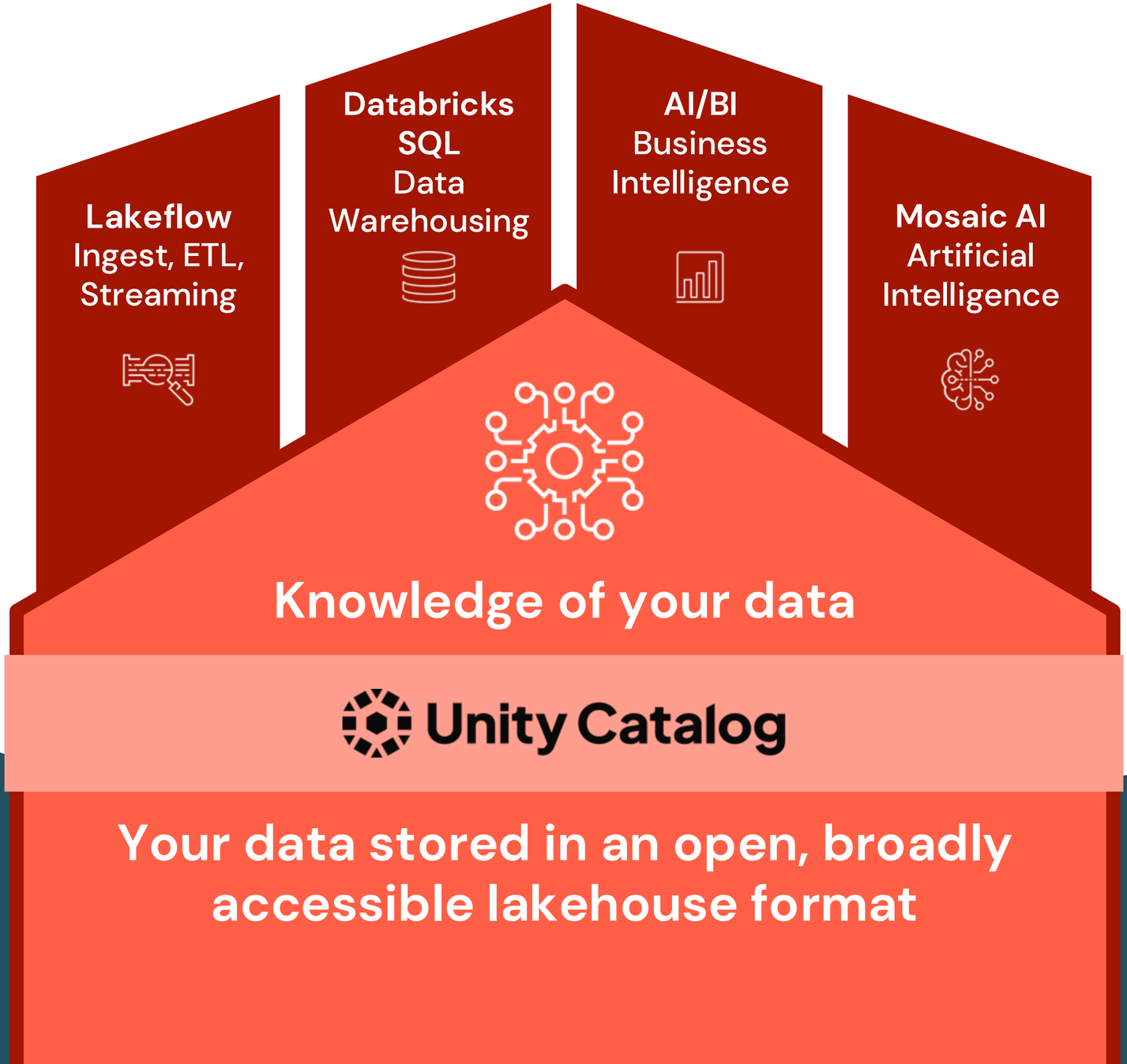
Databricks Overview

LECTURE

Databricks Data Intelligence Platform



databricks


Lakeflow
Ingest, ETL, Streaming
Databricks SQL
Data Warehousing
AI/BI
Business Intelligence
Mosaic AI
Artificial Intelligence
Knowledge of your data
Unity Catalog
Your data stored in an open, broadly accessible lakehouse format

Databricks simplifies the complexity in your data estate

Unify and own your data

Adopt an open source Lakehouse format of choice

Break free from lock-in with open formats

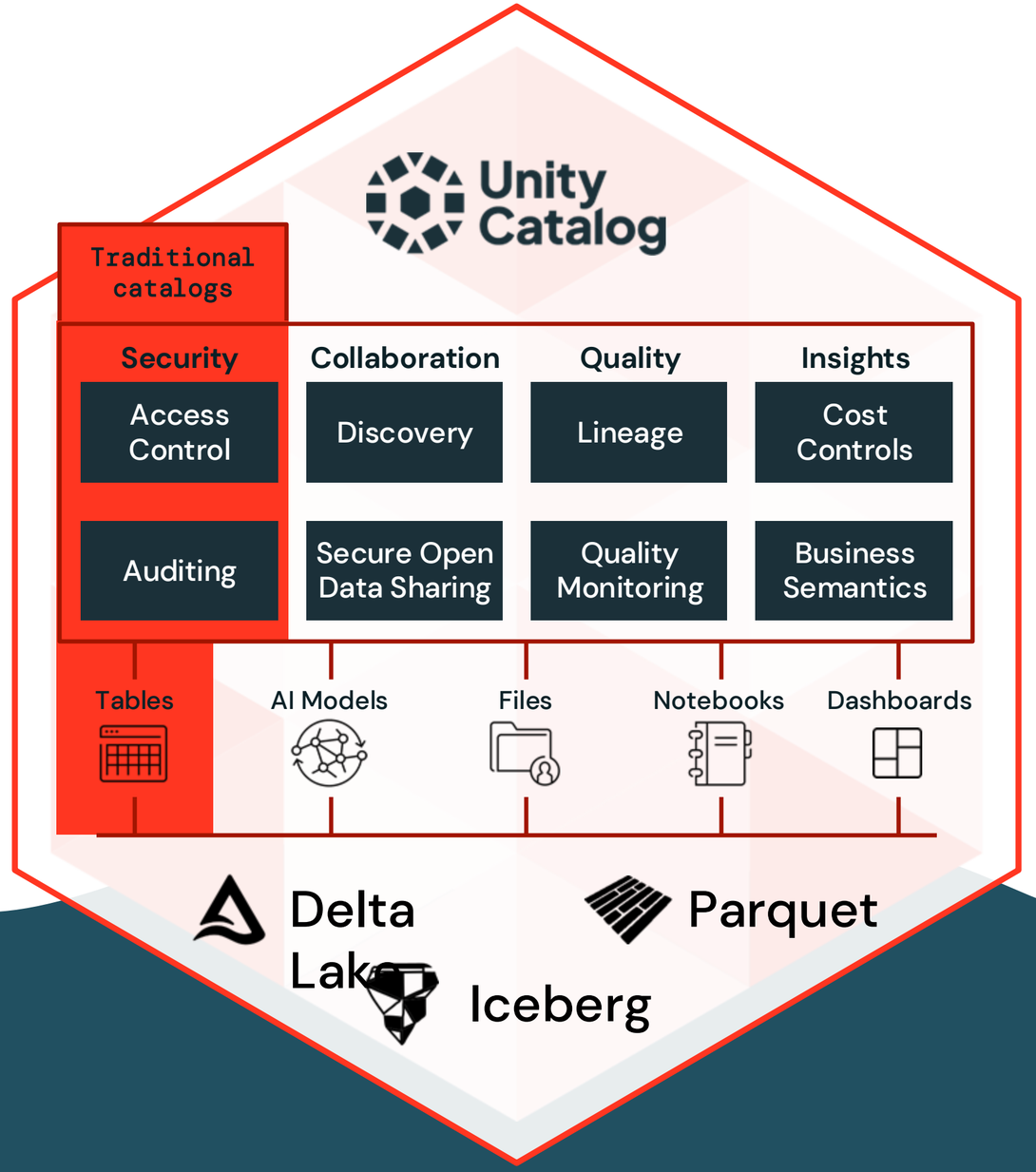
Reduce unnecessary costs from multiple copies of data

Iceberg Delta Lake Parquet

Protect and understand your data

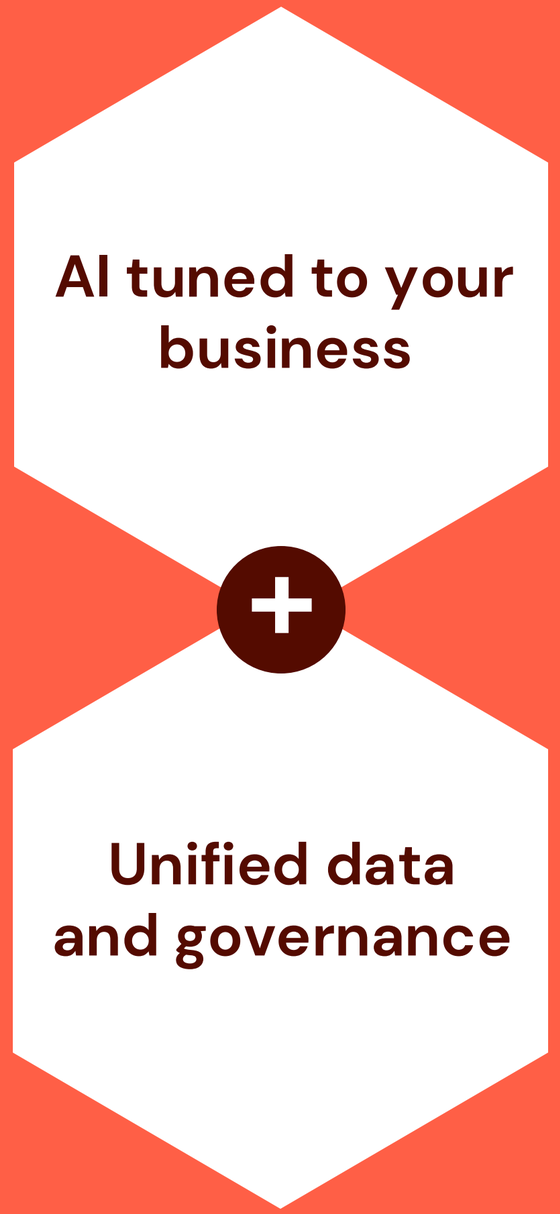
Unified capabilities for **every use case**

Unified governance for **all assets**

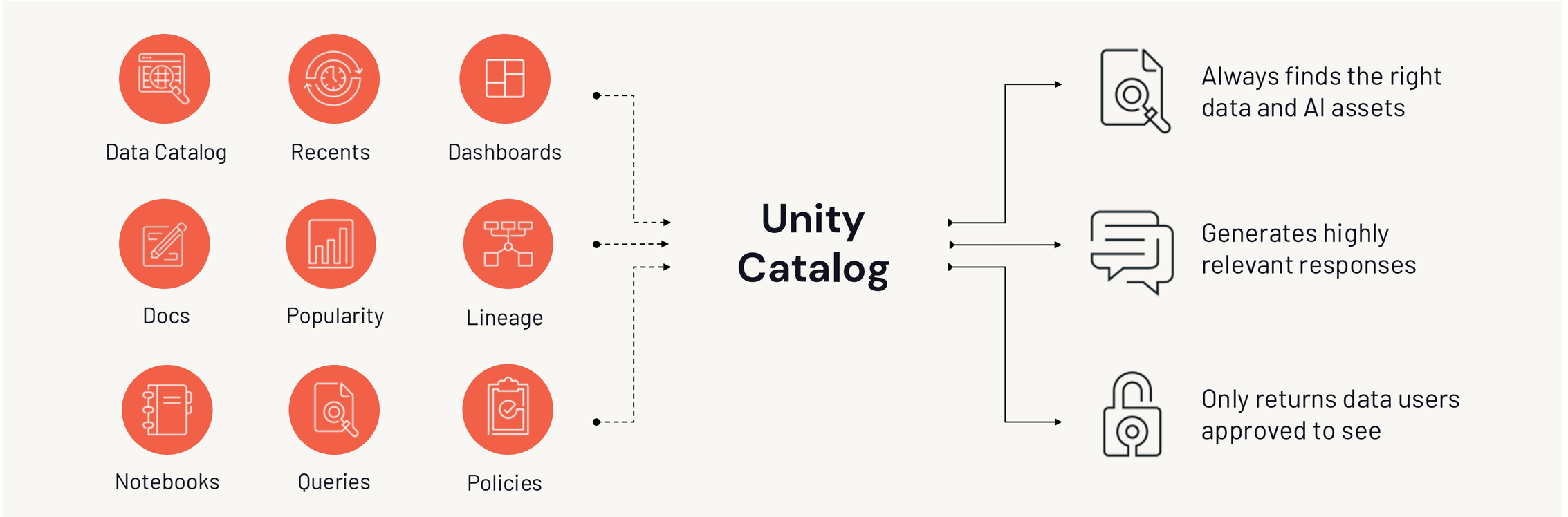


databricks

Data Intelligence Platform



Data Intelligence delivers the unique context of your business



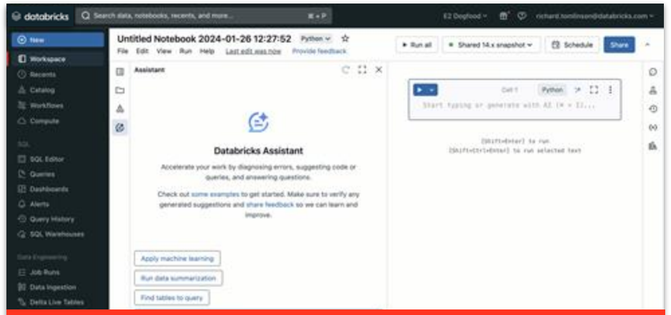
All of the signals flowing into Unity Catalog allow AI to learn

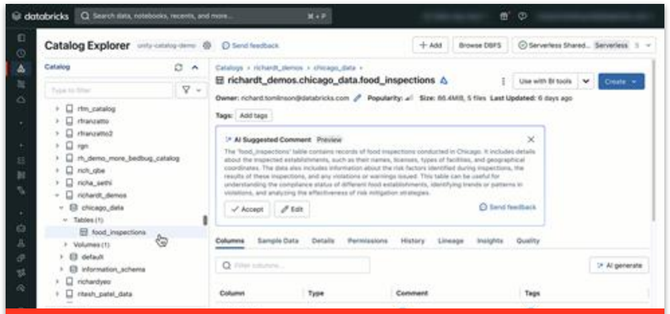
Databricks gains awareness of your unique data and business context

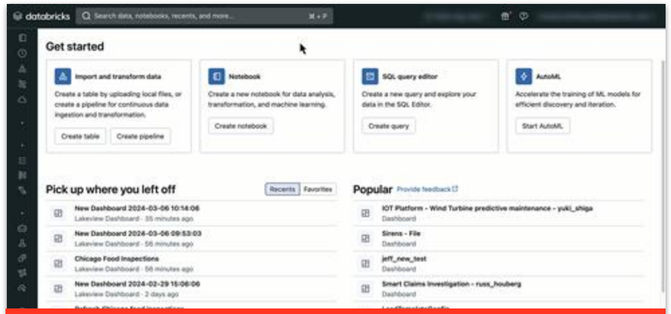
Boosts productivity for all users

Data Scientists | Data Engineers | SQL Analysts | Business Users

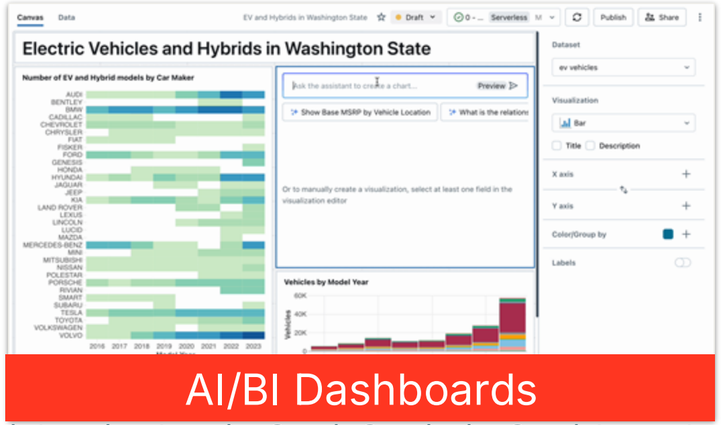
For Data Teams

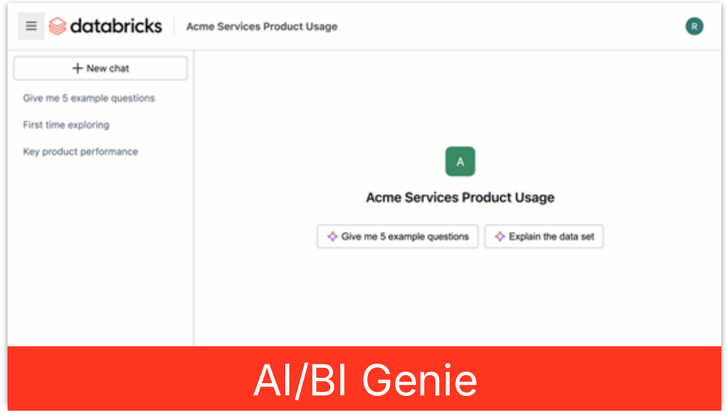




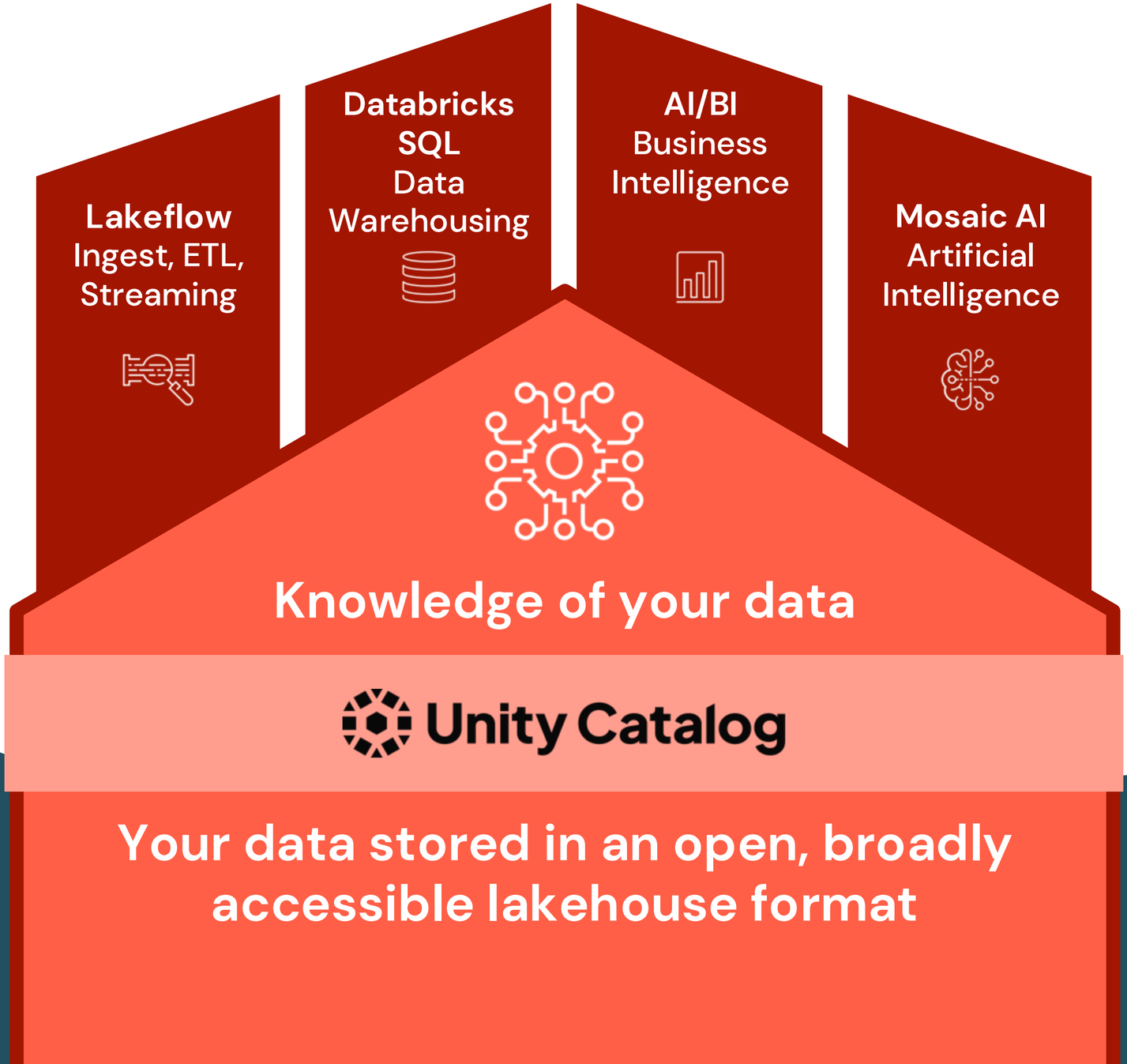


For Business Teams





databricks

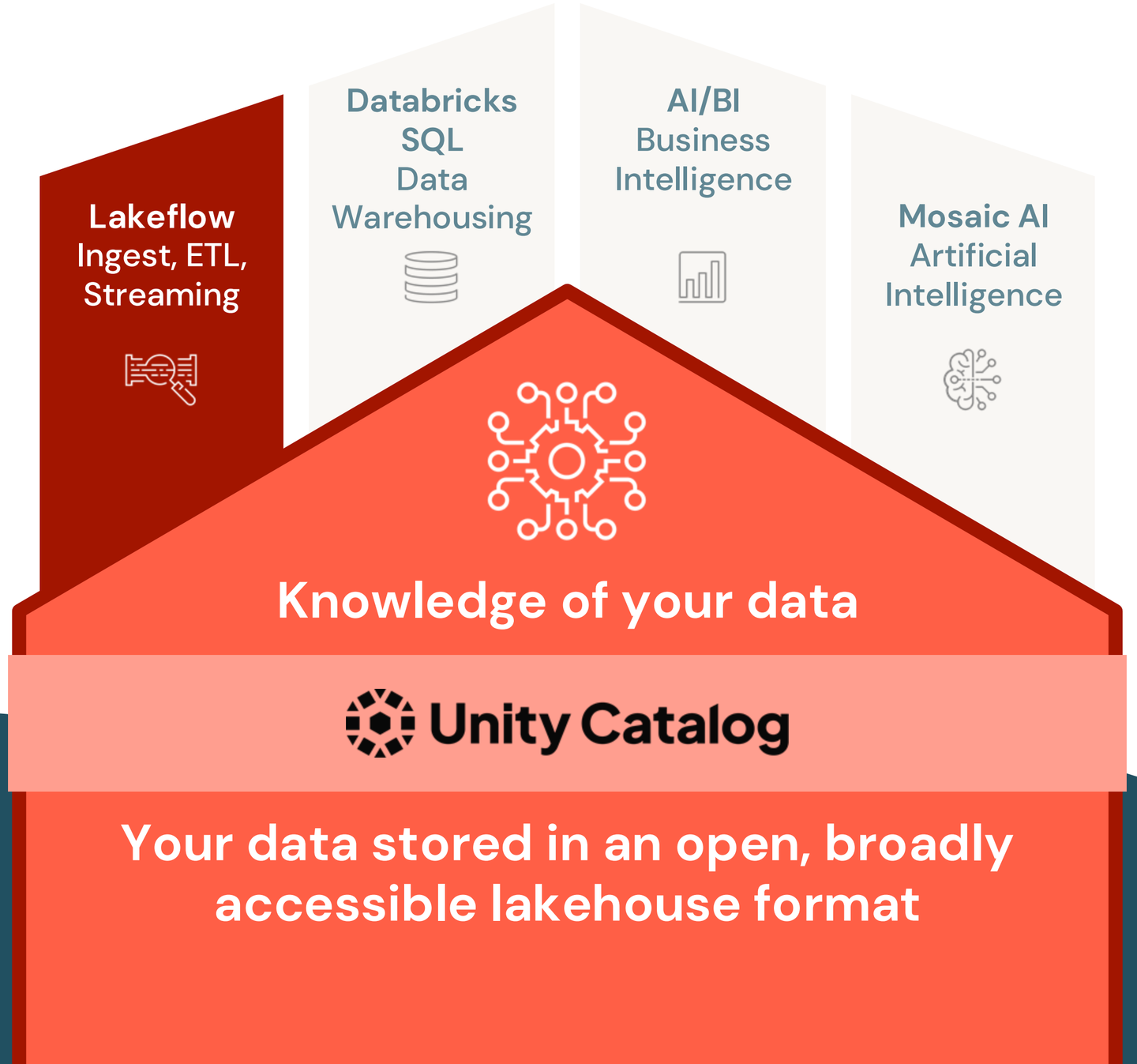

Lakeflow
Ingest, ETL, Streaming
Databricks SQL
Data Warehousing
AI/BI
Business Intelligence
Mosaic AI
Artificial Intelligence
Knowledge of your data
Unity Catalog
Your data stored in an open, broadly accessible lakehouse format

Achieving data intelligence begins with getting data in

databricks

Lakeflow

Reliable & automated dataflow from systems-of-record



Get the right data without the hassle

databricks Lakeflow

Connect

Easily connect key data sources

ORACLE DATABASE · PostgreSQL · MySQL · Microsoft SQL Server · salesforce · workday. · Google Analytics · servicenow · Google Ads · ORACLE NETSUITE · SharePoint · Dynamics 365

DLT

Reliable data pipelines made easy

Jobs

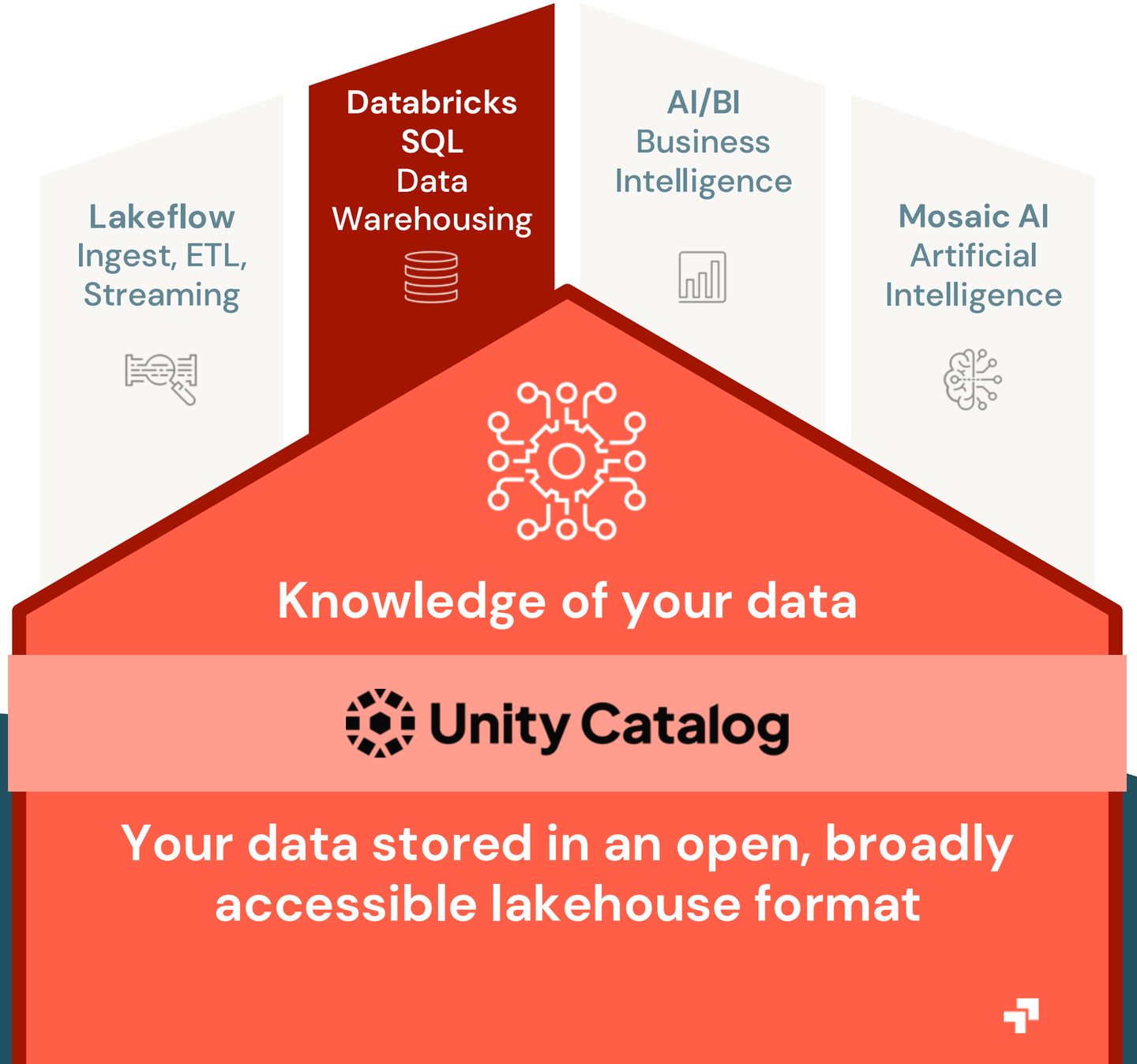
Unified orchestration for analytics and AI

What is data intelligence for analytics and BI?

databricks

Databricks SQL

The intelligent enterprise data warehouse



Complete data warehousing capabilities

Lower migration cost

Serverless Warehouses	Views/Temp Views	ANSI SQL	Data Quality Monitoring	Monitor Permissions
1P/3P Connectors	SQL Scripting	SQL UDFs	Table ACLs	Billing System Tables
CDC Support	H3 GeoSpatial	Session Vars	Lakehouse Federation	Internet On/Off
Streaming Tables	Spatial SQL (ST_)	SQL Alerts	HMS Federation	Queries System Table
Materialized Views	Foreign/Primary Keys	Dashboards	Table Lineage	Warehouses Sys Table
Autoloader for Ingest	Lateral Col Alias	Dashboard Sharing	Entity Relation Diagram	WH Events Sys Table
Scheduled Workflows	Named Arguments	Publish to PowerBI	Row/Col Security	Query Duration Limits
1p/3p Orchestration	HyperLogLog	Publish to Tableau	OAuth	Foreign/Primary Keys
SQL Editor	Array Functions	Cloud Fetch	ABAC	Query History
Python/Go Connectors	Identifier Clause	Python UDFs	Marketplace	Query Profiling
SQL Rest API	Variant Data Type	Row Level Concurrency	100K+ User Support	Warehouse Monitoring

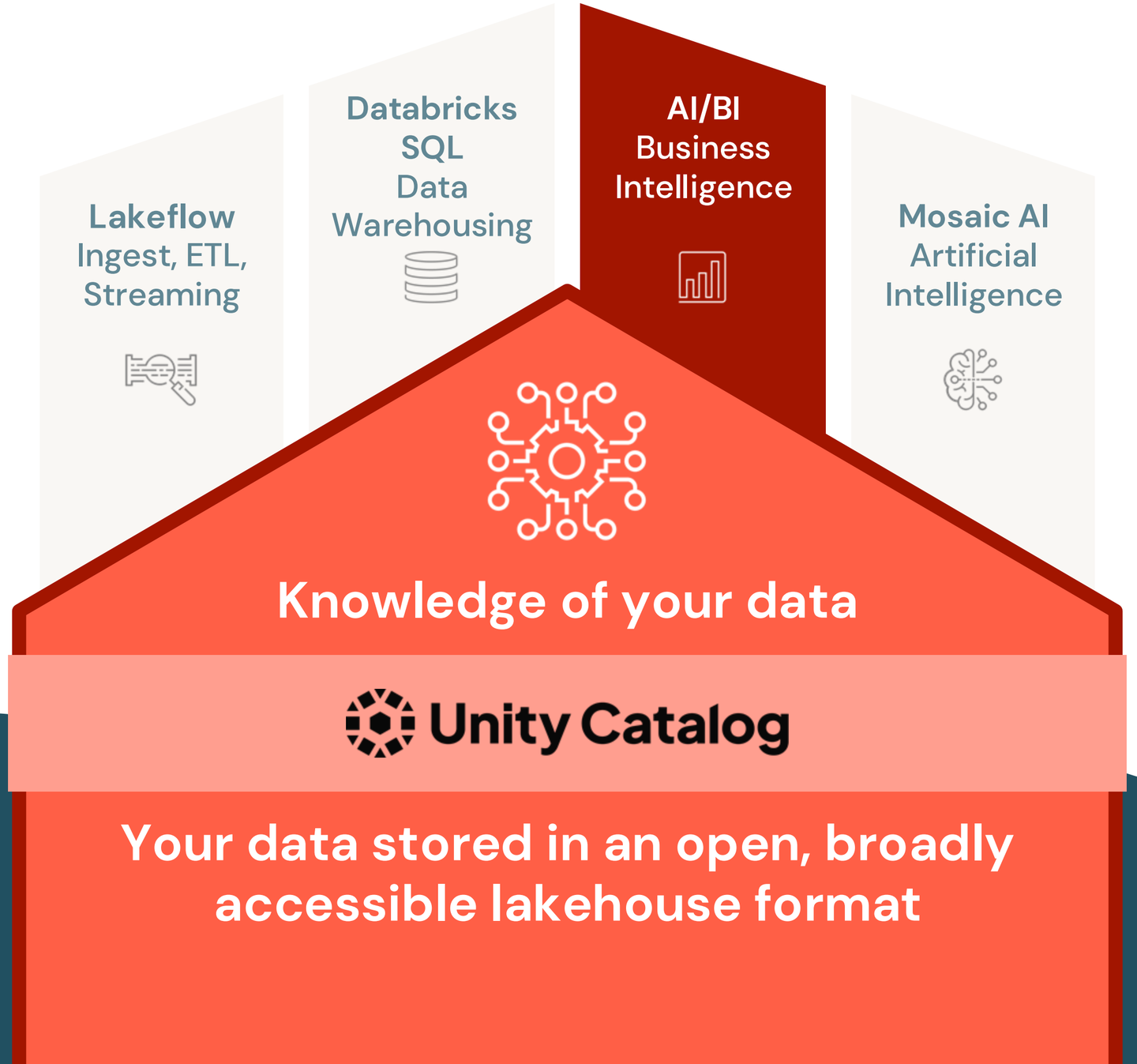
Foundational Functionality

Governance and Administration

databricks

AI/BI

Intelligent analytics
for real-world data



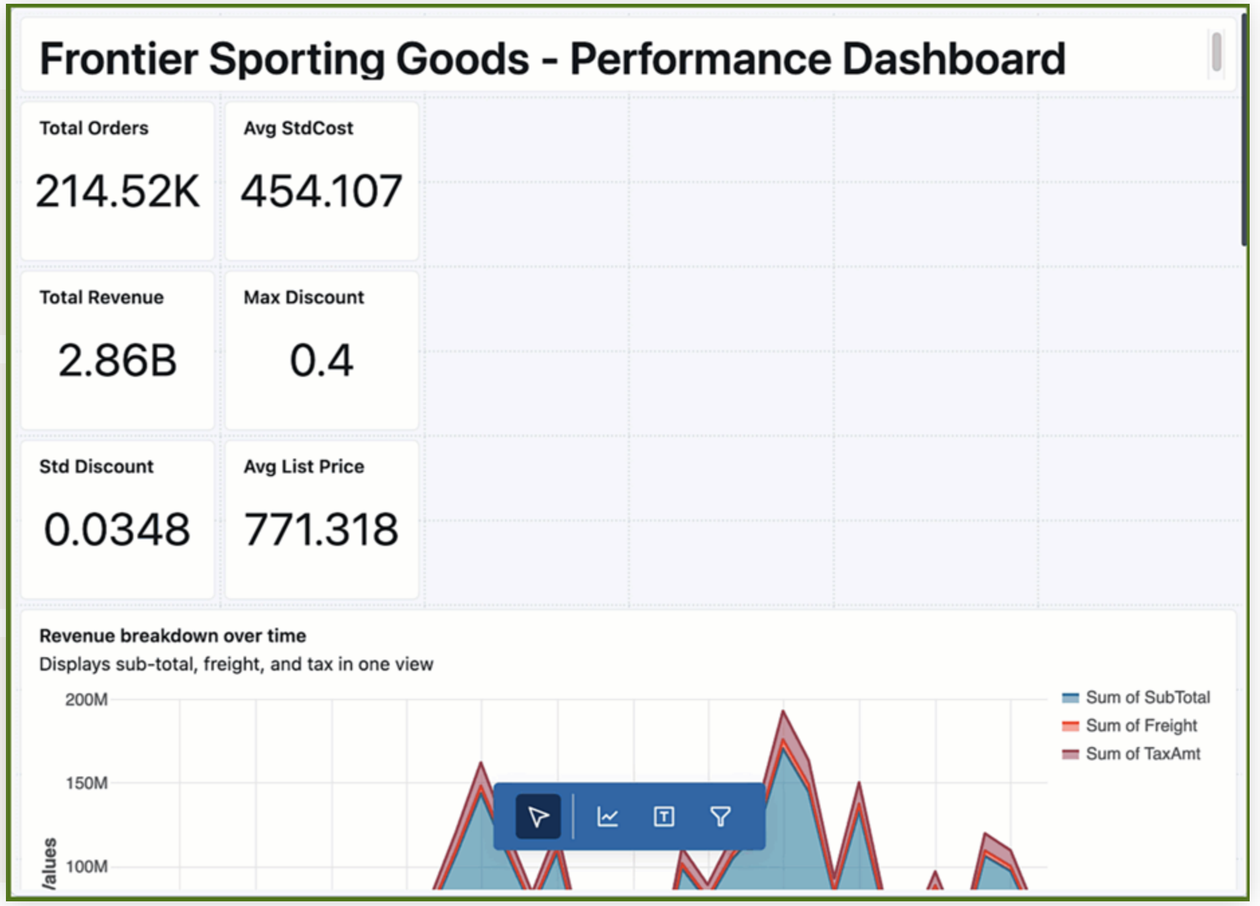
Dashboards to analyze and visualize data

Get useful results with AI-powered understanding

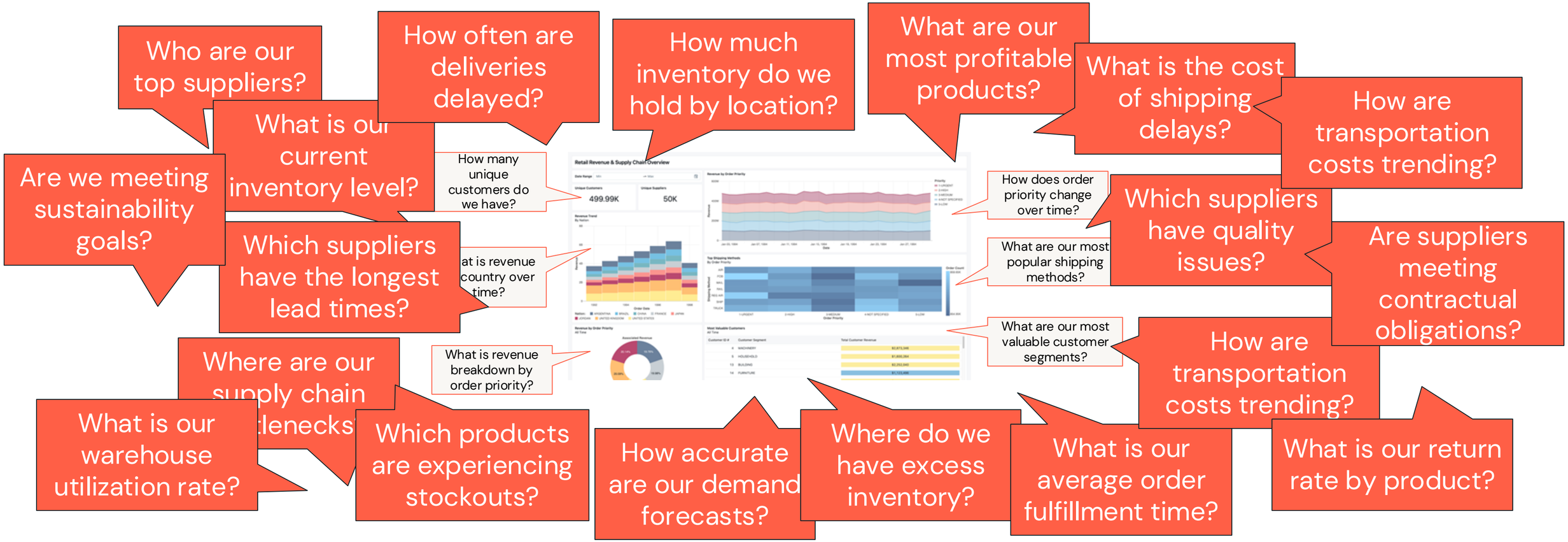
Scale BI to everyone with no additional cost

Blazing fast results

Secure by design with Unity

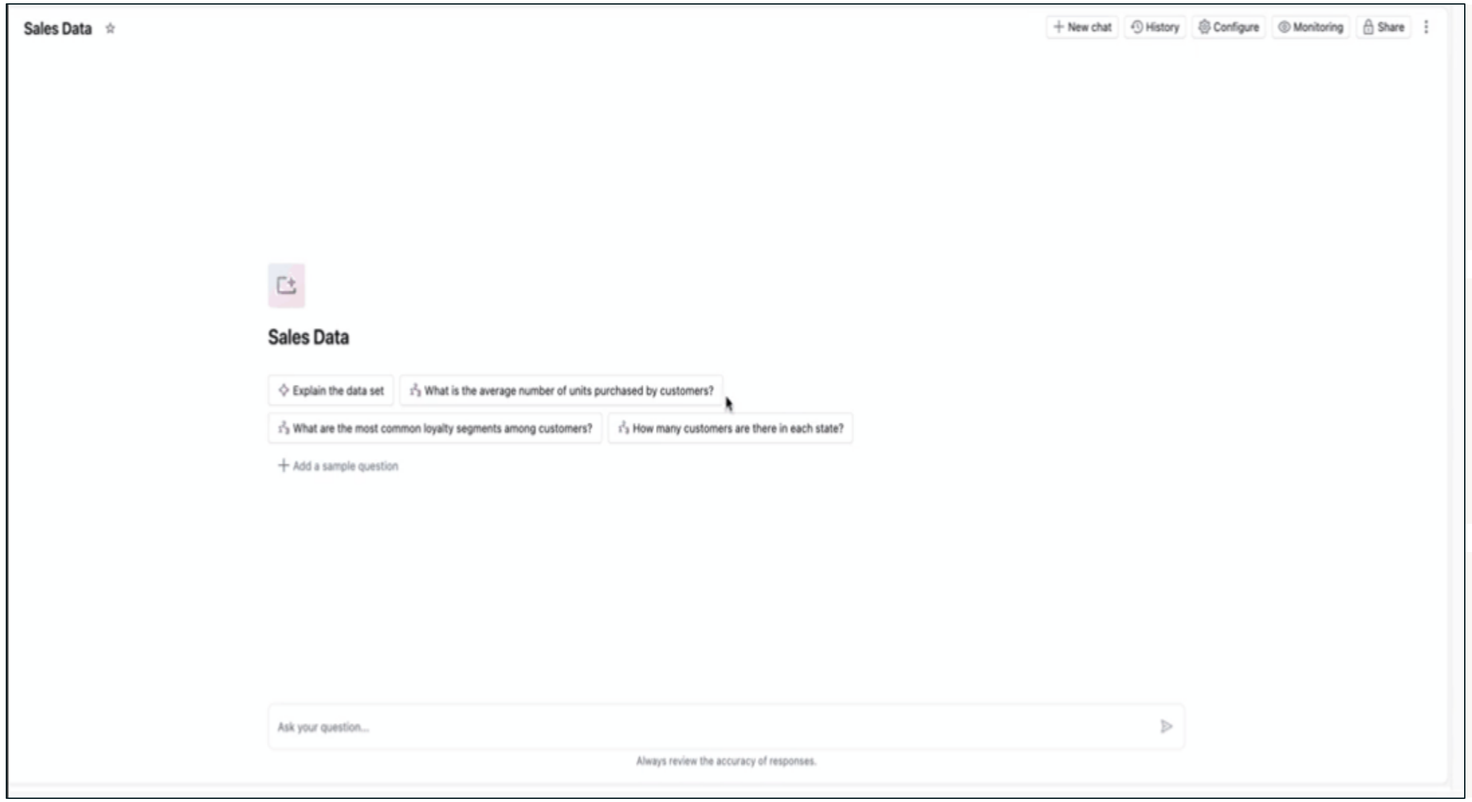


But, you can’t anticipate every question users will ask



Genie lets you go beyond dashboards

Talk to your data to explore any question



Query your enterprise data directly with open Q&A

Maintain context between sessions to follow-up on research

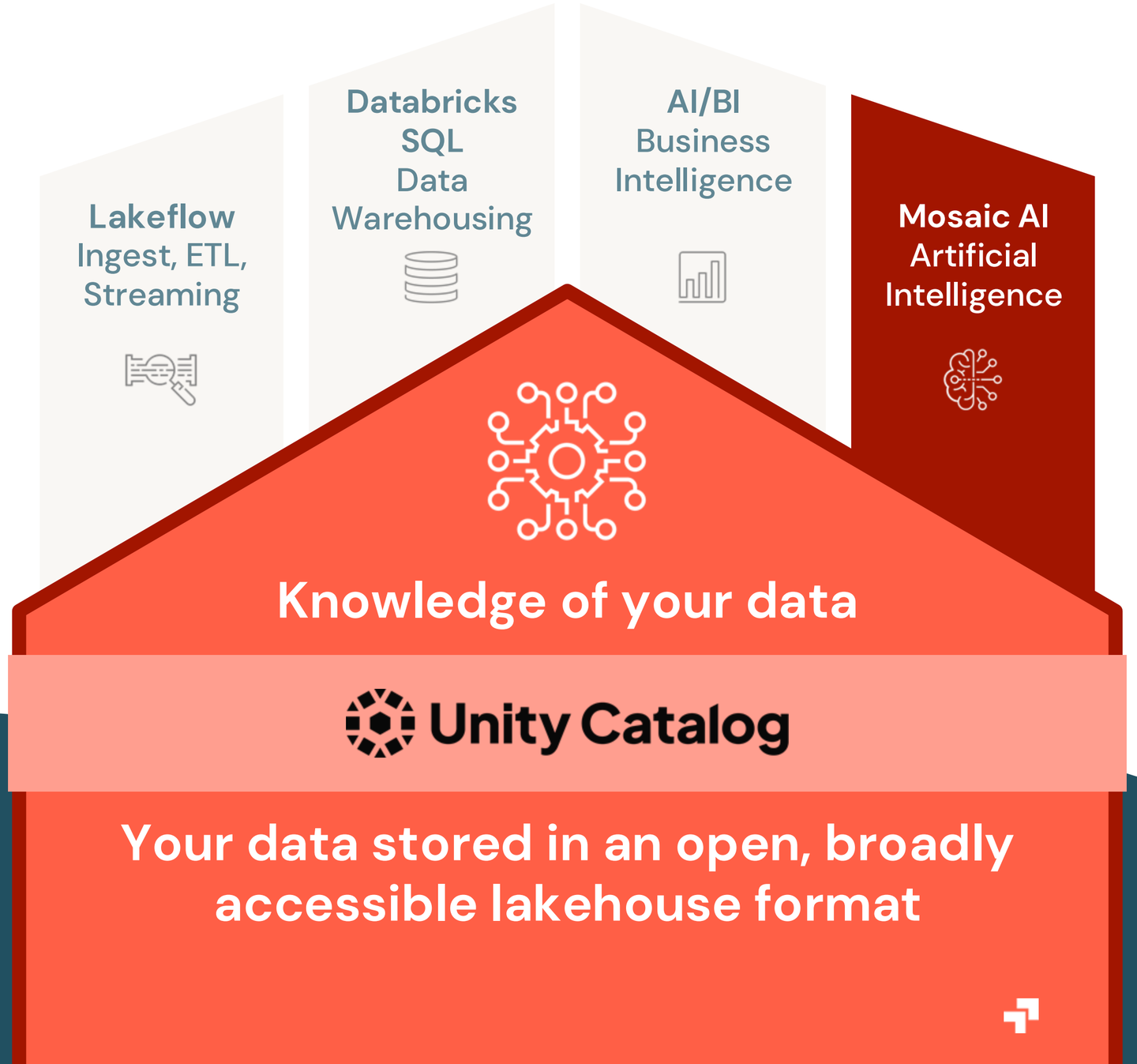
Get responses in natural language, data tables, and visualizations

What is data intelligence for AI applications?

databricks

Mosaic AI

Create domain-specific agentic applications



General Intelligence

Large models
trained on the entire web
leveraging scaling laws

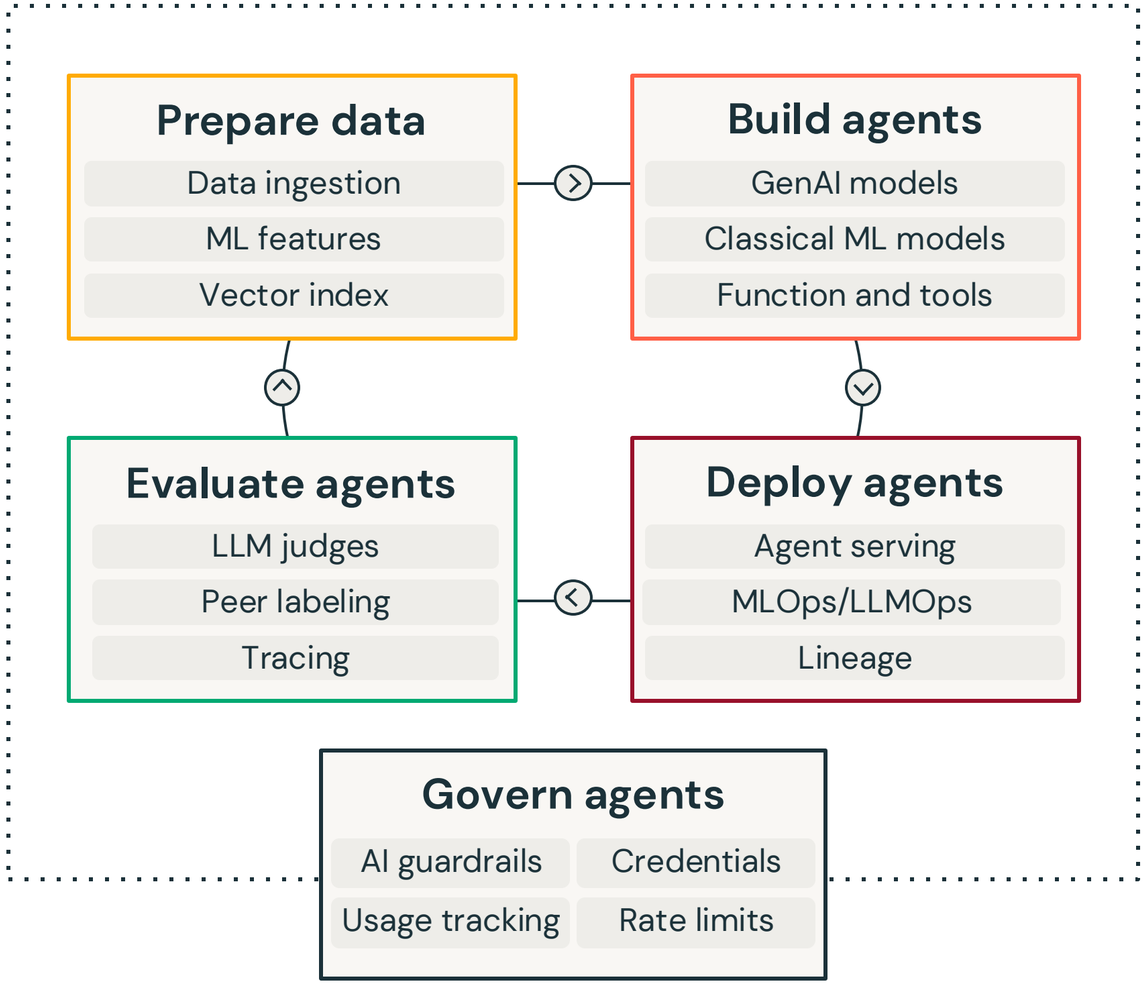
VS

Data Intelligence

AI agents that **reason**
on your data and solve
domain-specific problems

Get accurate, domain-specific results

Mosaic AI is the complete agent platform

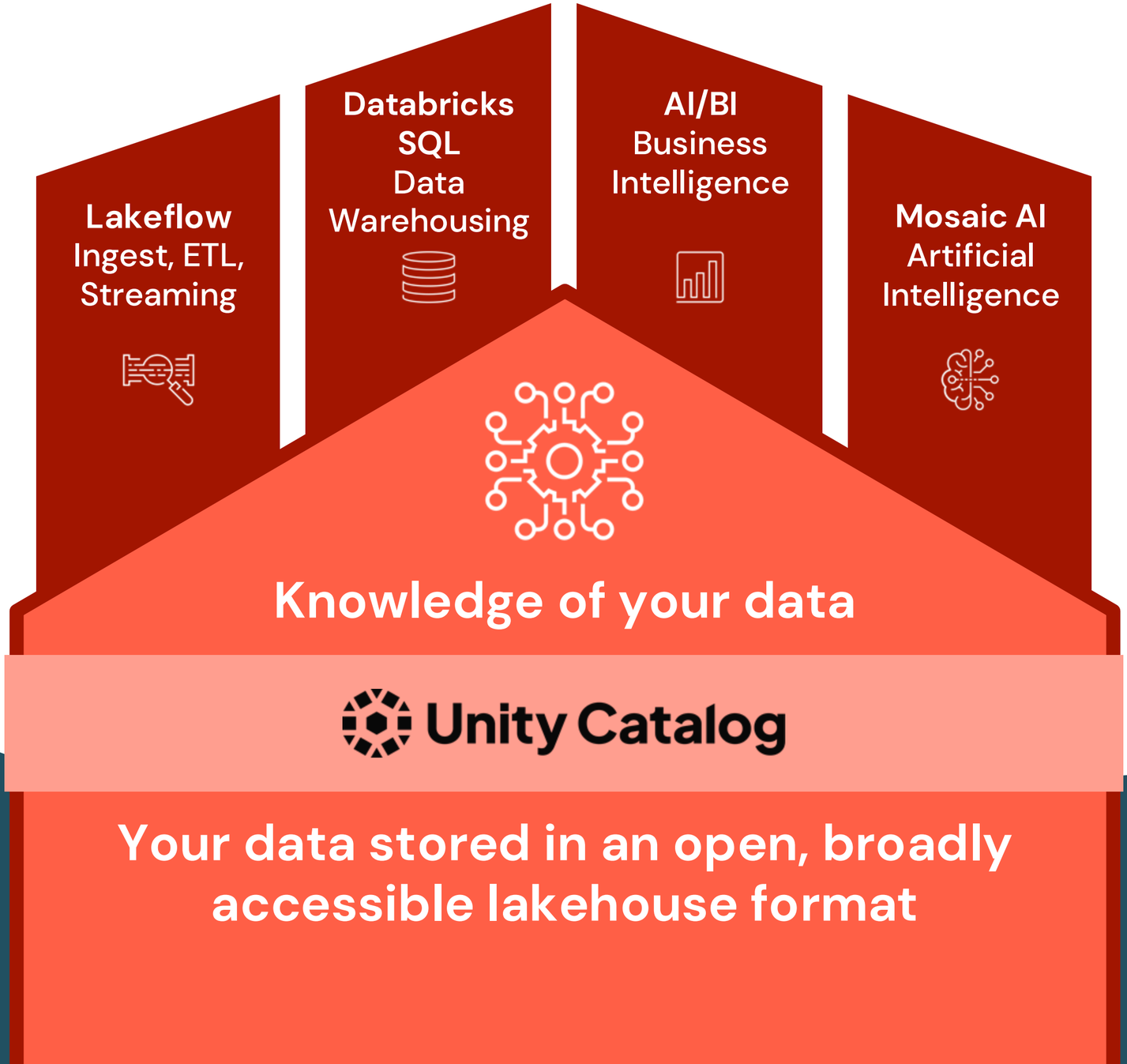


Agents that reason across every enterprise system

Support for all existing and future AI models

Build trust with guardrails, evaluation, and monitoring

databricks

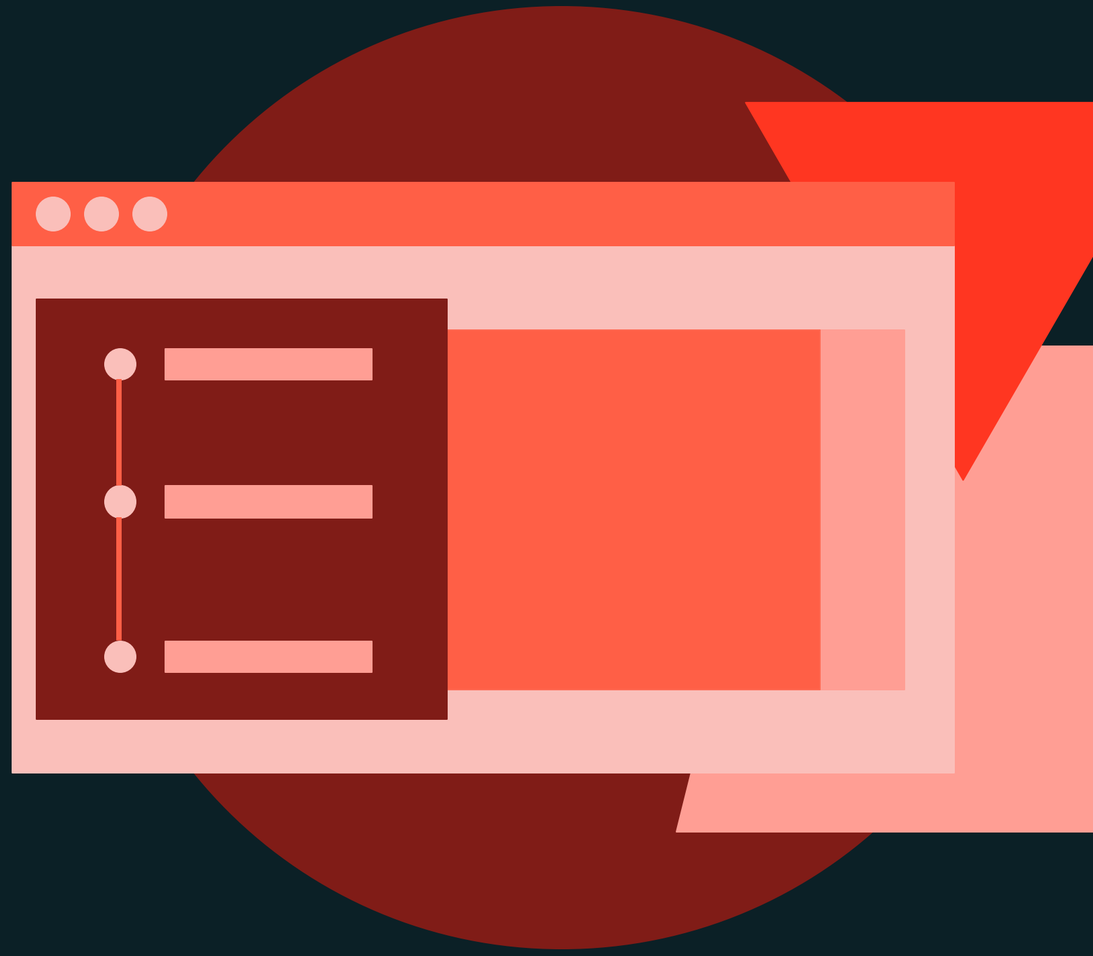

Lakeflow
Ingest, ETL, Streaming
Databricks SQL
Data Warehousing
AI/BI
Business Intelligence
Mosaic AI
Artificial Intelligence
Knowledge of your data
Unity Catalog
Your data stored in an open, broadly accessible lakehouse format

databricks

Databricks Overview

DEMONSTRATION

Databricks Workspace Walkthrough



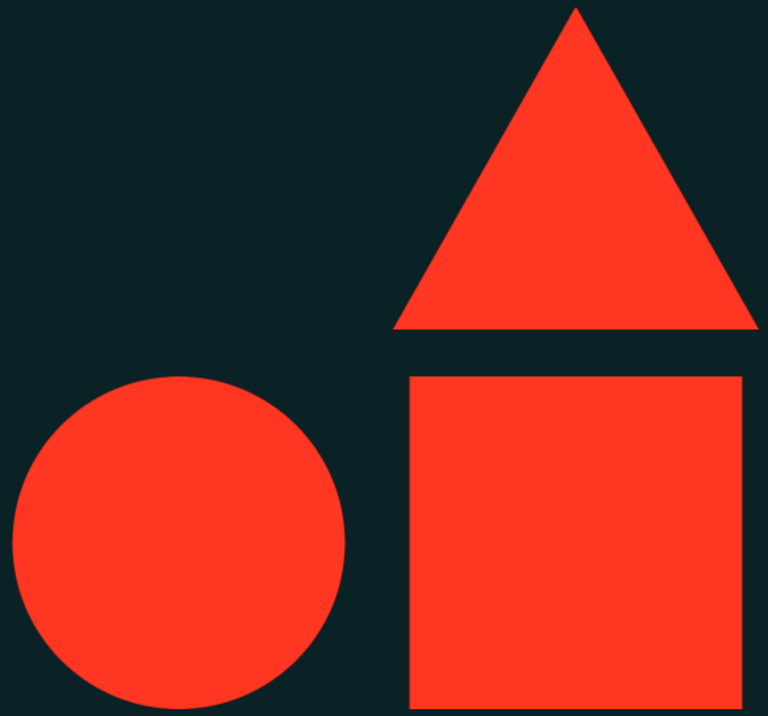
Follow-along Instructions

Estimated Time: 15 minutes

For this demonstration, the instructor will take you on a guided tour of the platform. If you have access to the Vocareum lab environment, feel free to follow-along to explore the workspace and get oriented with where future labs will be completed.

databricks

Using Databricks for Data Warehousing



Get Started with Databricks for Data Warehousing

Objectives

- Use Databricks to complete common data warehousing tasks.
- Explain the purpose of Delta Lake for data warehousing purposes.

Agenda

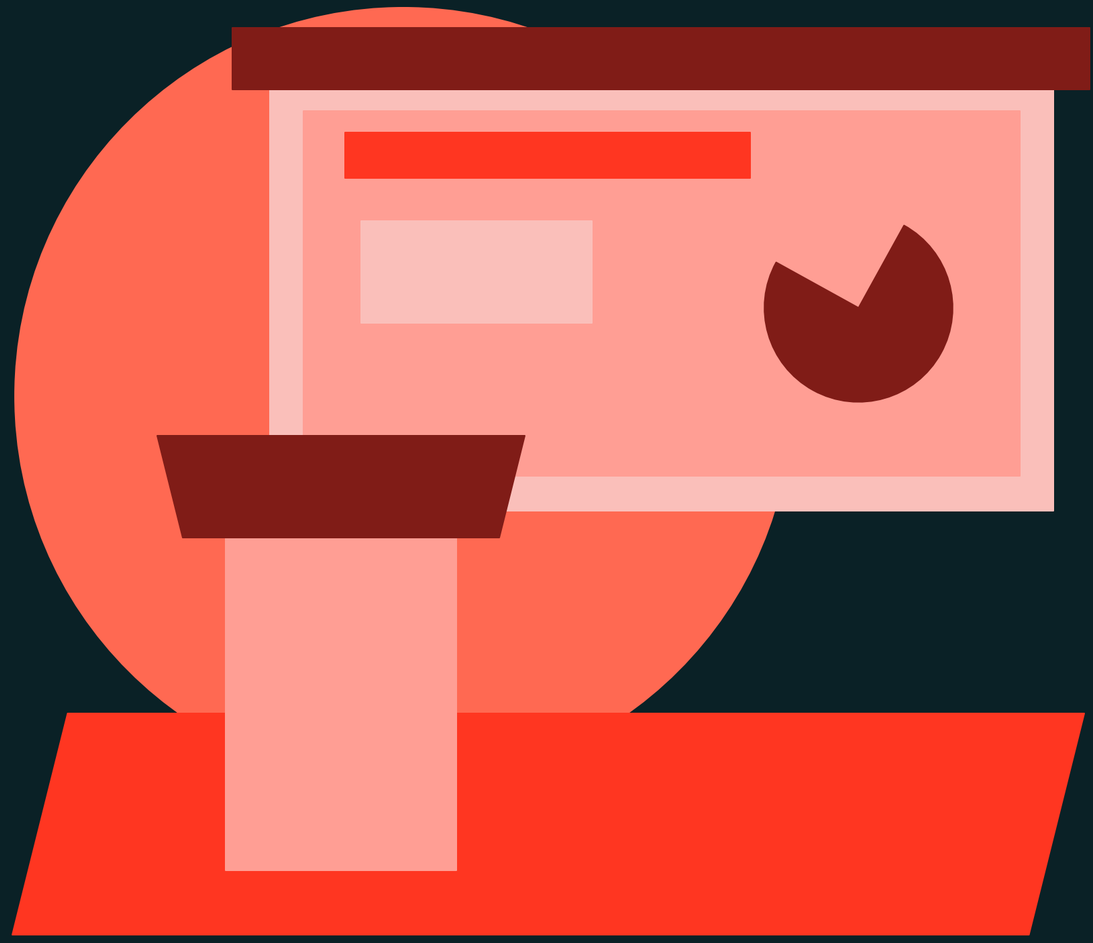
Using Databricks for Data Warehousing	Time	Lecture	Demo	Lab
Introduction to Data Warehousing with Databricks	10 mins	✓		
Databricks SQL Warehouses	5 mins	✓		
Delta Lake Overview	8 mins	✓		
Using Delta Lake features with Databricks SQL	5 mins		✓	

databricks

Using Databricks for Data Warehousing

LECTURE

Introduction to Data Warehousing with Databricks

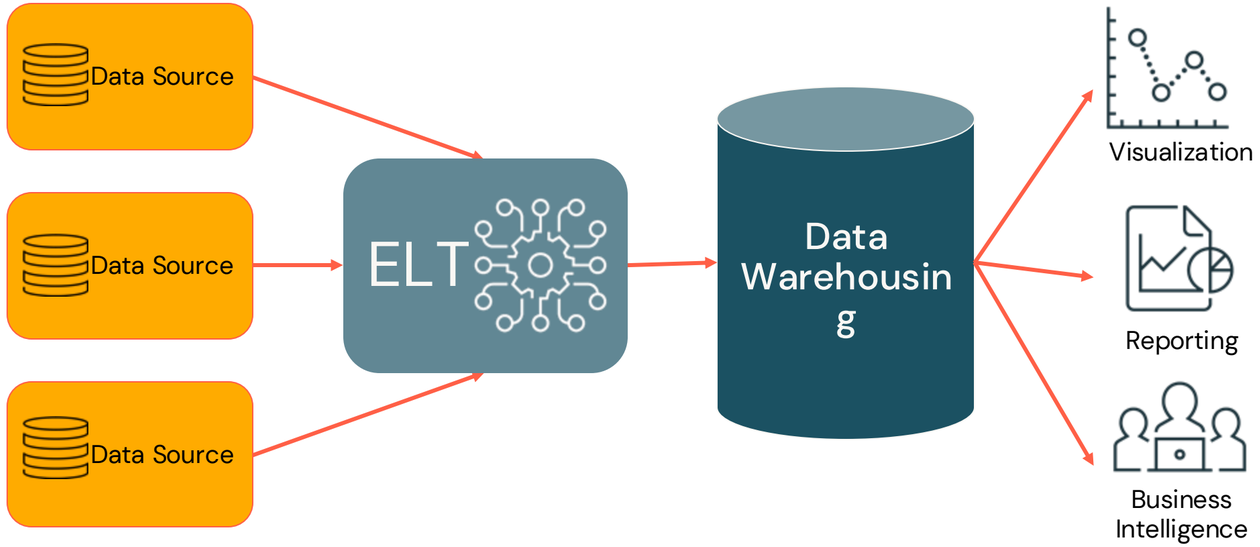


Introduction to Data Warehousing

A Centralized Solution for Reporting, Analytics, and Insights

A data warehouse serves as the foundation for reporting, analytics, and BI, helping organizations make informed decisions.

- **Centralized Storage:** Integrates data from multiple sources (databases, logs, etc.).
- **ETL Process:** Data is extracted, cleaned, and organized for analysis.
- **Optimized for Insights:** Powers visualization, reporting, and BI tools.

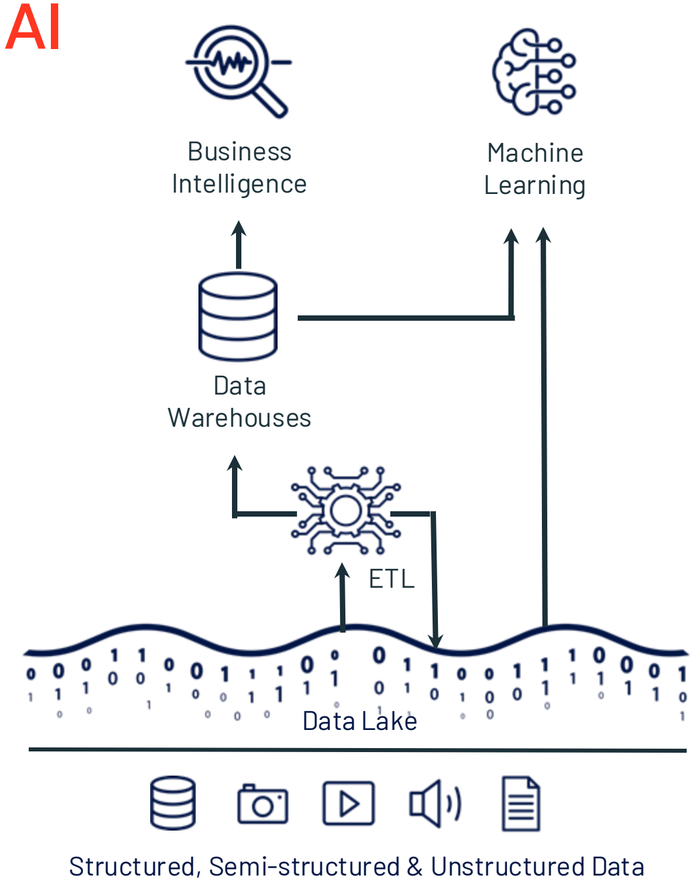


Data warehousing professionals ensure the ETL process runs smoothly and data is centralized, reliable, and optimized for reporting and insights.

Why Data Warehousing Matters

Unified data powers analytics, reporting, and AI

- Consolidates Data from Disparate Sources
- Improves Data Quality and Consistency
- Enables Reporting, Analytics, and BI
- Supports Better Decision-Making



Unified data enables faster insights and better decisions, driving organizational success.

Challenges in Data Warehousing

Data warehousing can be complex and expensive.

Users need specialized technical skills

Complex infrastructure

Data silos and fragmented systems

Issues with data governance and compliance

Slow performance at scale

Inconsistent data quality and access

- Users must have specialized knowledge to operate and manage traditional data warehouses.
- Talent is hard to obtain and retain as the industry is highly competitive.
- Expensive to train up these skills to have them lost to other companies.

Challenges in Data Warehousing

Data warehousing can be complex and expensive.

Users need specialized technical skills

Complex infrastructure

Data silos and fragmented systems

Issues with data governance and compliance

Slow performance at scale

Inconsistent data quality and access

- Redundant tooling within the data warehouse make it difficult to achieve consistency
- Systems and processes become inefficient over time and need to be frequently adjusted to maintain performance requirements
- Maintenance difficulties arise with easily outdated foundational components

Challenges in Data Warehousing

Data warehousing can be complex and expensive.

Users need specialized technical skills

Complex infrastructure

Data silos and fragmented systems

Issues with data governance and compliance

Slow performance at scale

Inconsistent data quality and access

- Data is spread across multiple systems in the ecosystem, fragmenting management
- Difficulty in consolidating insights and identifying accurate and clean data
- Delayed decision-making as processes are lengthy and heavily manual

Challenges in Data Warehousing

Data warehousing can be complex and expensive.

Users need specialized technical skills

Complex infrastructure

Data silos and fragmented systems

Issues with data governance and compliance

Slow performance at scale

Inconsistent data quality and access

- Lack of centralized access controls in traditional platforms
- Challenges in ensuring compliance with regulations (GDPR, HIPAA)
- Limited capabilities to track data lineage and challenges with tackling quality issues

Challenges in Data Warehousing

Data warehousing can be complex and expensive.

Users need specialized technical skills

Complex infrastructure

Data silos and fragmented systems

Issues with data governance and compliance

Slow performance at scale

Inconsistent data quality and access

- Runaway increase in cost for cloud compute as the data warehouses scale
- Degraded performance as the amount of data increases from small initial testing
- Lack of access in traditional on-premise data warehouses to the latest technologies that impact performance capabilities

Challenges in Data Warehousing

Data warehousing can be complex and expensive.

Users need specialized technical skills

Complex infrastructure

Data silos and fragmented systems

Issues with data governance and compliance

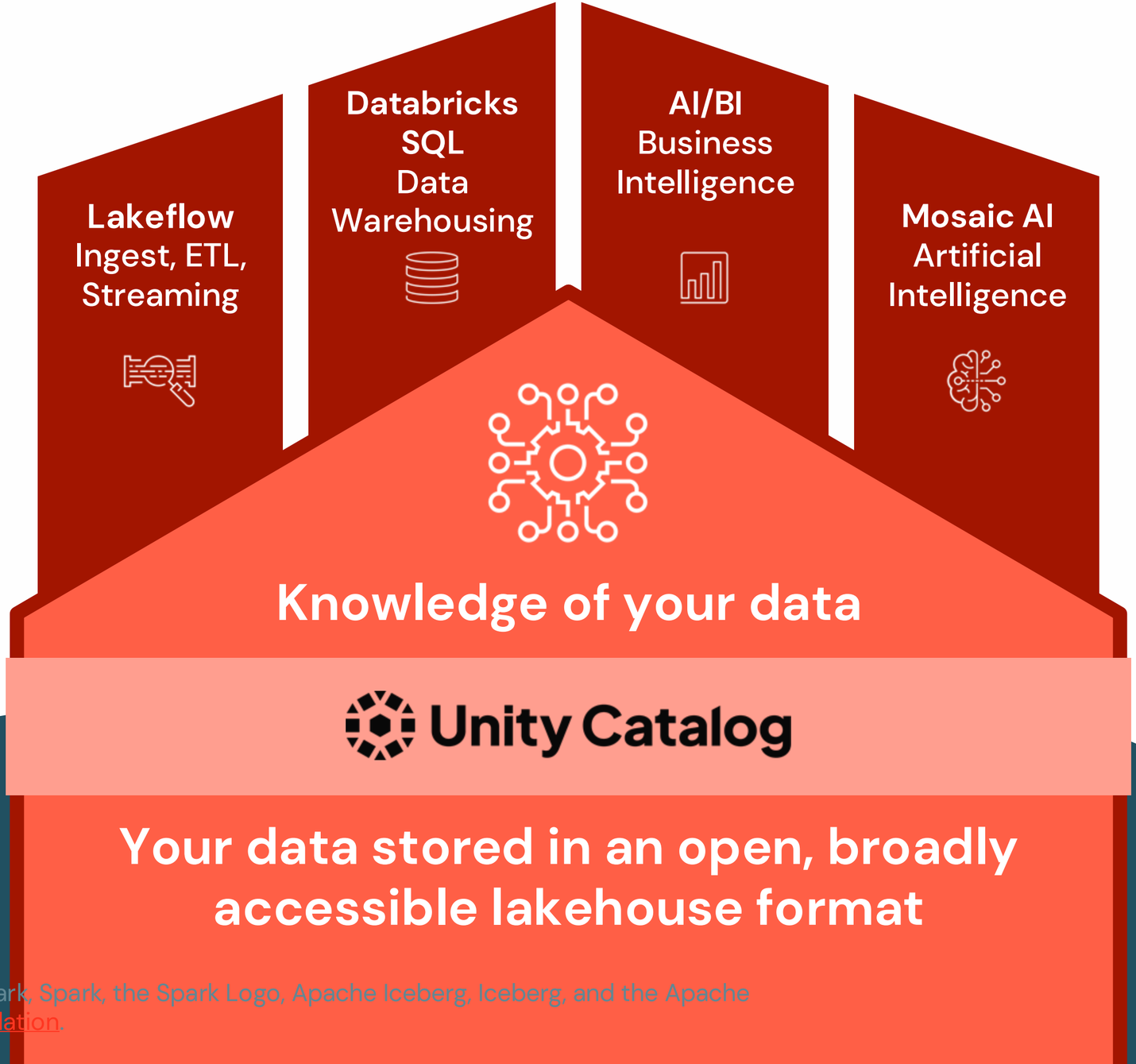
Slow performance at scale

Inconsistent data quality and access

- Inconsistent, incomplete, or outdated data persists within the environment
- Limited access for business users leads to manual overhead to process requests for access
- Complex processes requiring manual oversight to clean and validate data persist regardless of the volume of data in the warehouse

With Databricks, data warehousing becomes simpler, faster, and smarter – enabling you to overcome traditional challenges.

Databricks Data Intelligence Platform



What Makes Databricks Different?

Databricks Data Warehousing: Simplified, Centralized, and Powerful

Unified Data Management:

- Combines structured and semi-structured data using Delta Tables for centralized storage.

Transform with Delta:

- Leverages Databricks' compute for ETL and real-time transformations.

Powered by Unity Catalog:

- Ensures governance, metadata management, and seamless collaboration across teams.

Optimized Outputs:

- Enables data marts, machine learning models, and applications for actionable insights.

What Makes Databricks Different?

Databricks Data Warehousing: Simplified, Centralized, and Powerful

Unified Data Management:

- Combines structured and semi-structured data using Delta Tables for centralized storage.

Transform with Delta:

- Leverages Databricks' compute for ETL and real-time transformations.

Powered by Unity Catalog:

- Ensures governance, metadata management, and seamless collaboration across teams.

Optimized Outputs:

- Enables data marts, machine learning models, and applications for actionable insights.

What Makes Databricks Different?

Databricks Data Warehousing: Simplified, Centralized, and Powerful

Unified Data Management:

- Combines structured and semi-structured data using Delta Tables for centralized storage.

Transform with Delta:

- Leverages Databricks' compute for ETL and real-time transformations.

Powered by Unity Catalog:

- Ensures governance, metadata management, and seamless collaboration across teams.

Optimized Outputs:

- Enables data marts, machine learning models, and applications for actionable insights.

What Makes Databricks Different?

Databricks Data Warehousing: Simplified, Centralized, and Powerful

Unified Data Management:

- Combines structured and semi-structured data using Delta Tables for centralized storage.

Transform with Delta:

- Leverages Databricks' compute for ETL and real-time transformations.

Powered by Unity Catalog:

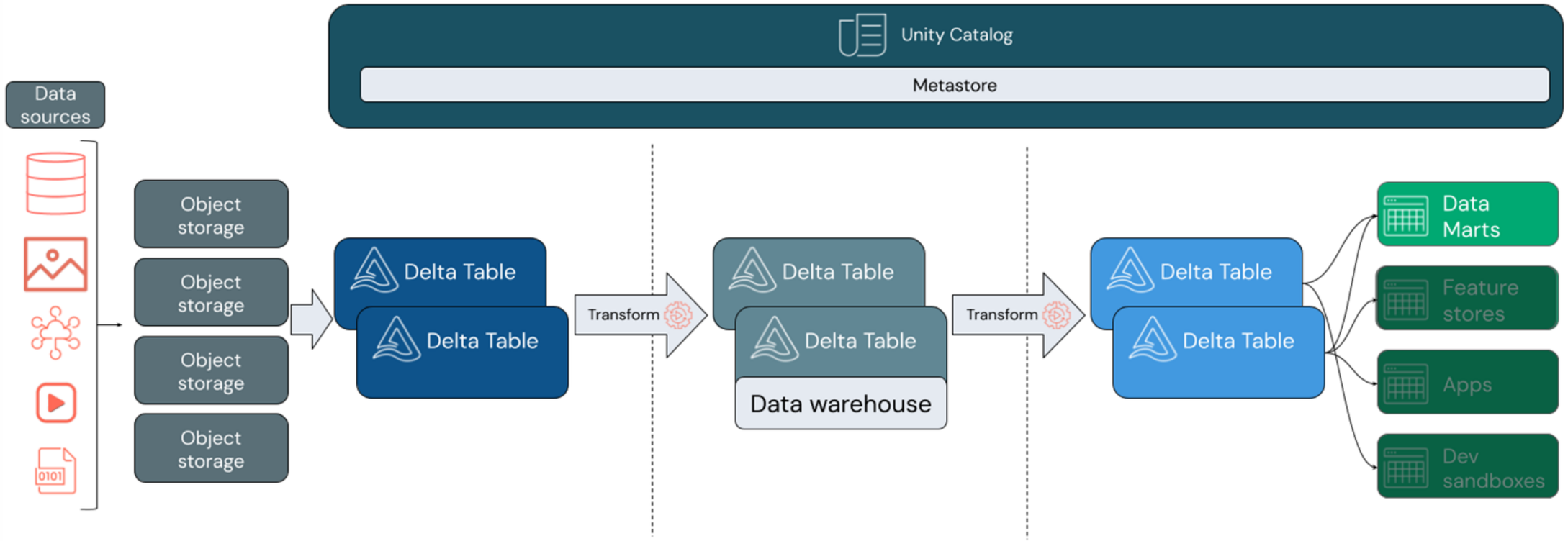
- Ensures governance, metadata management, and seamless collaboration across teams.

Optimized Outputs:

- Enables data marts, machine learning models, and applications for actionable insights.

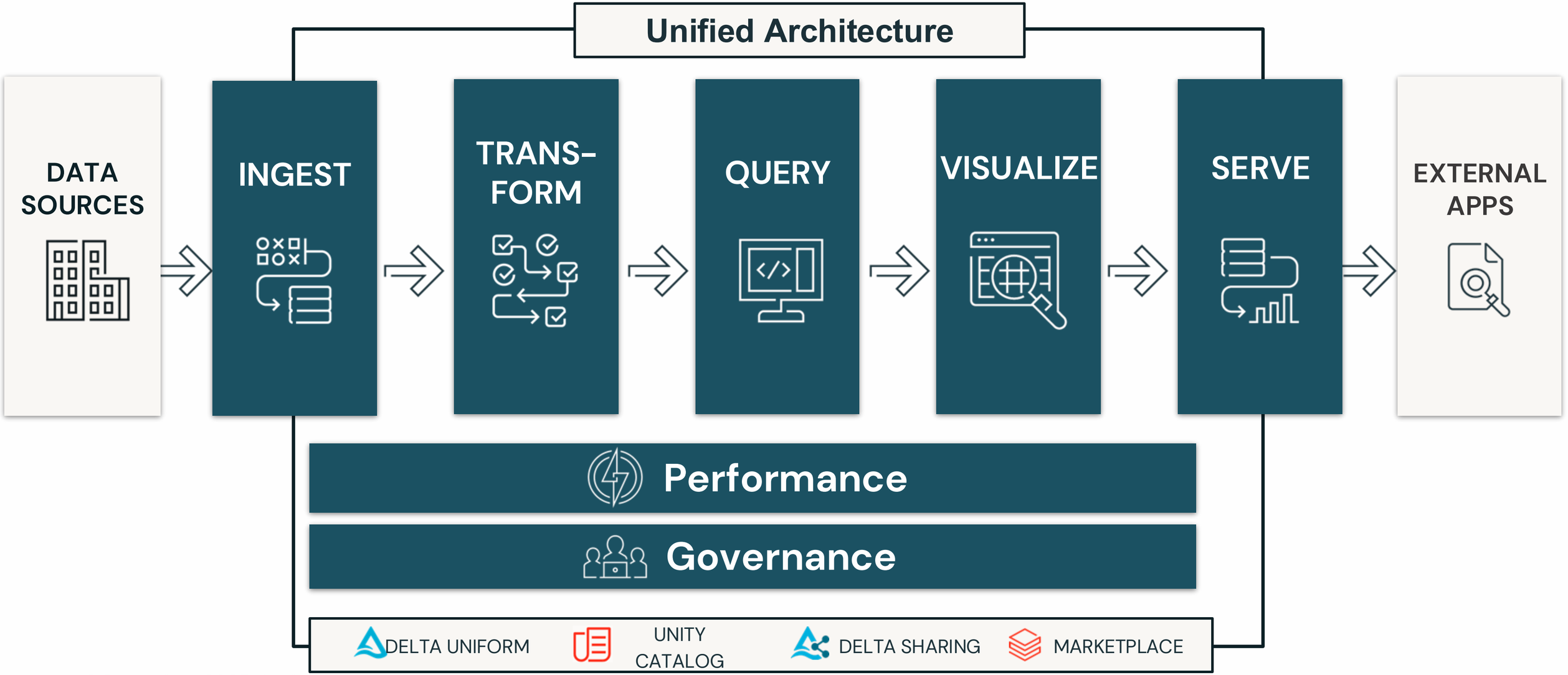
Databricks Approach to Data Warehousing

From raw data to insights: Leveraging Delta Tables and Unity Catalog.

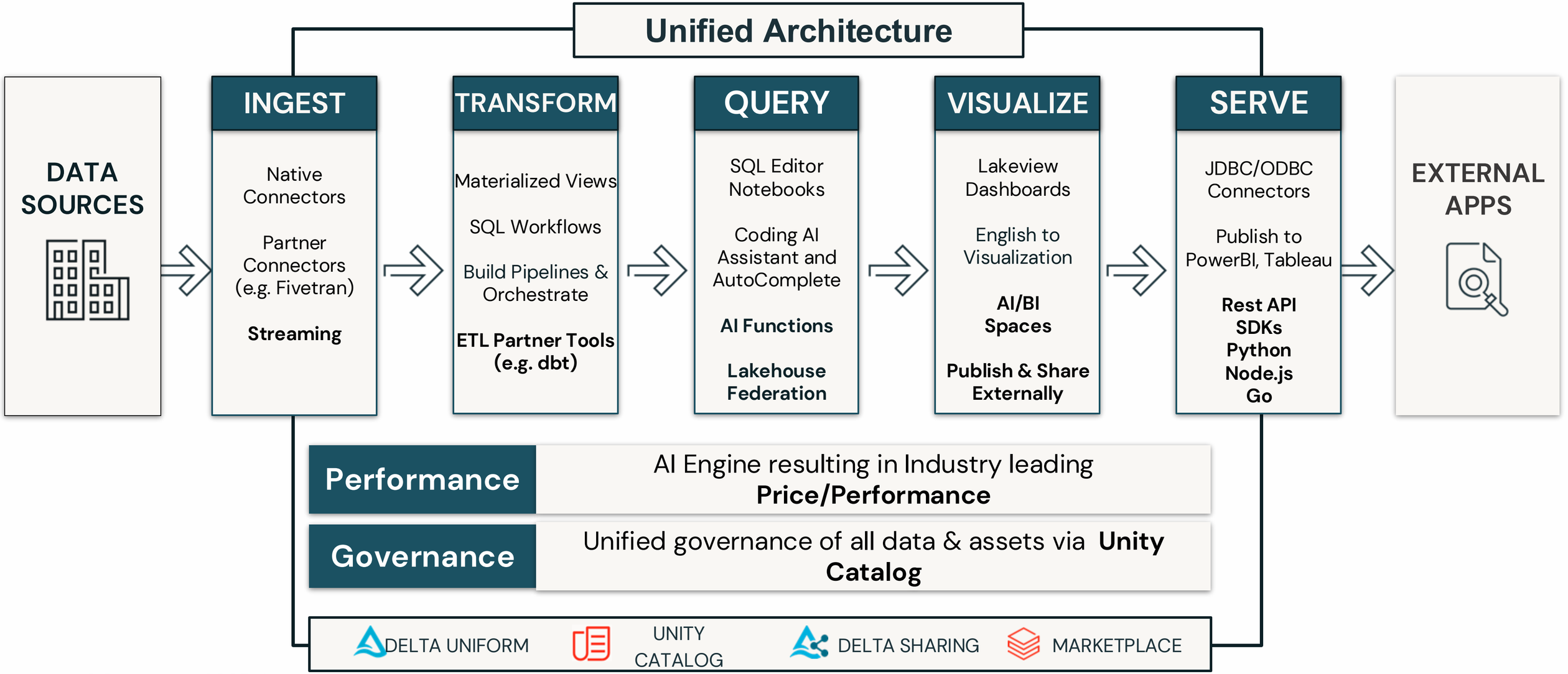


Databricks combines unified storage, real-time transformations, and governance to enable smarter data-driven insights.

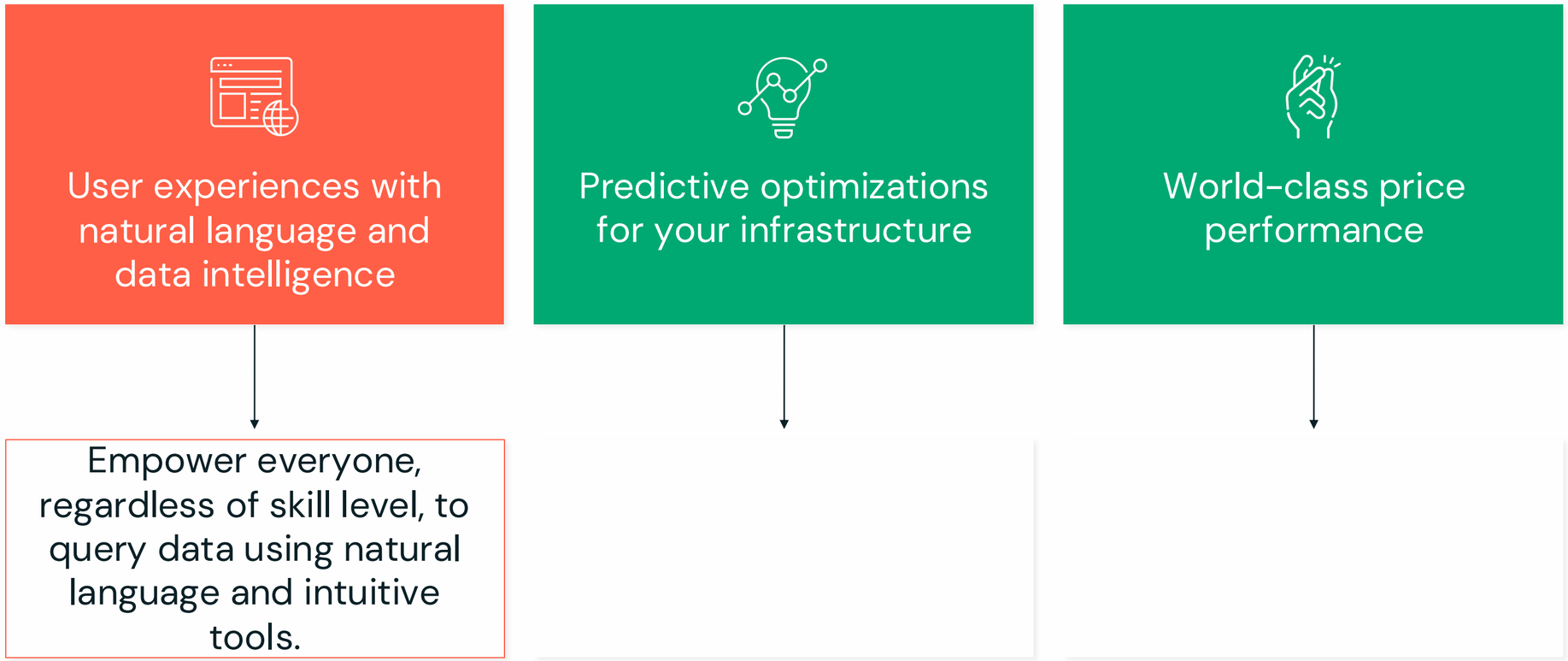
Complete Data Warehousing Solution



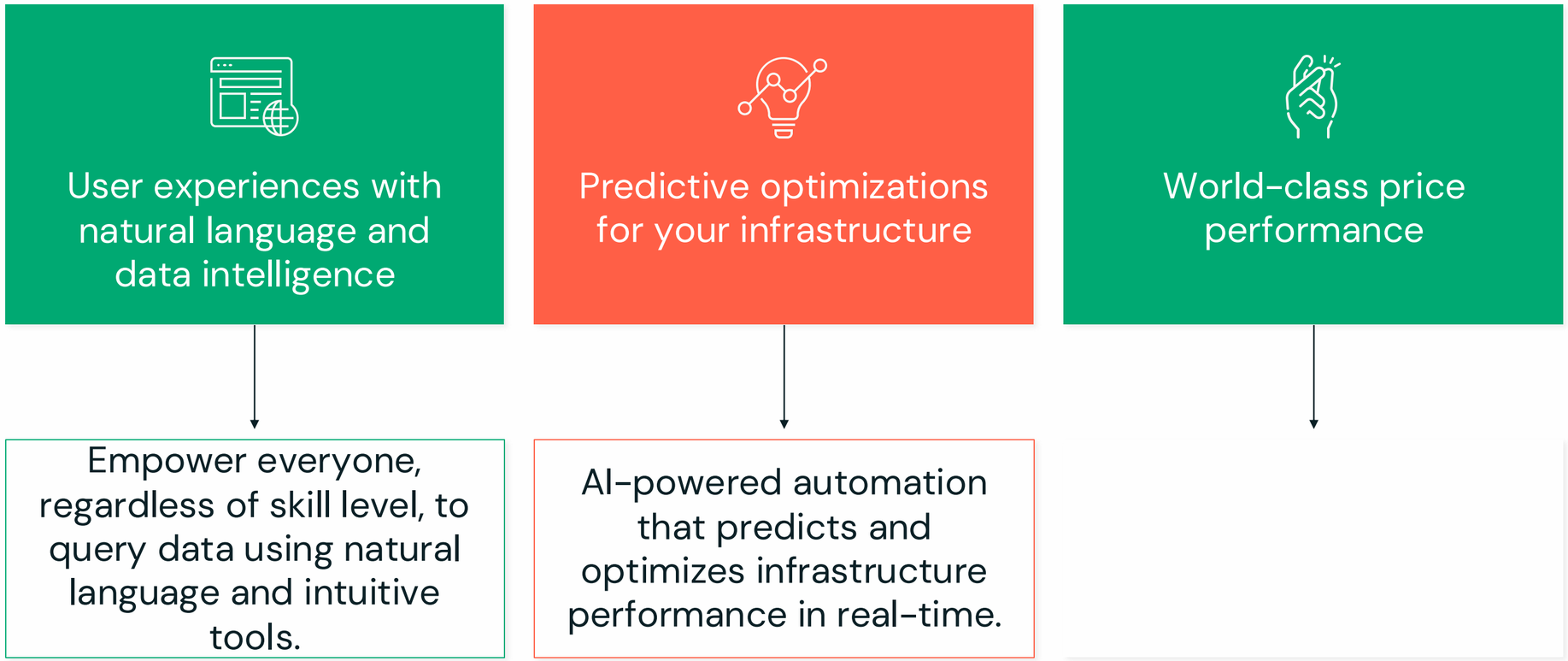
Complete Data Warehousing Solution



The Future of Data Warehousing: Smarter, Faster, and AI-Powered



The Future of Data Warehousing: Smarter, Faster, and AI-Powered



The Future of Data Warehousing: Smarter, Faster, and AI-Powered

User experiences with natural language and data intelligence

Empower everyone, regardless of skill level, to query data using natural language and intuitive tools.

Predictive optimizations for your infrastructure

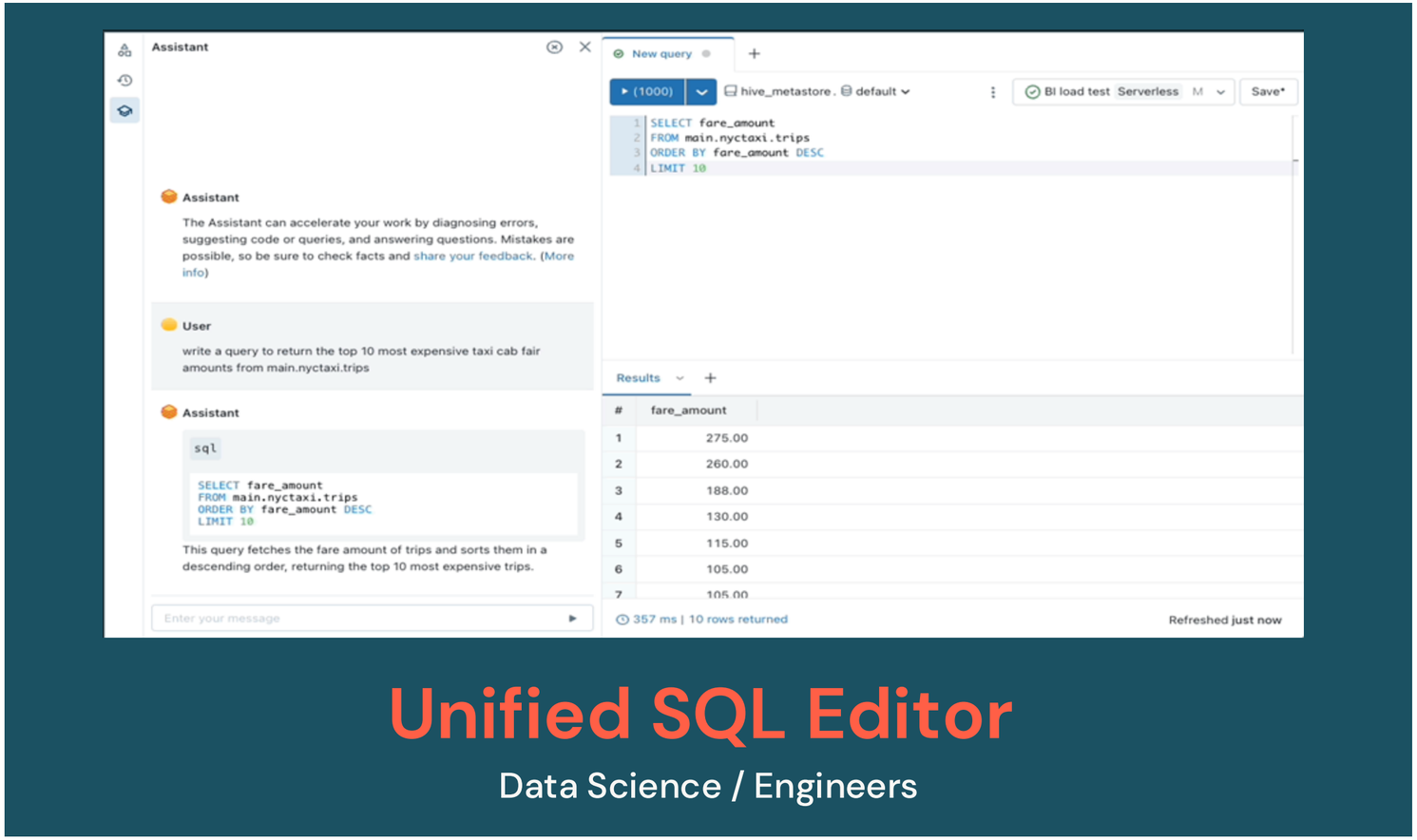
AI-powered automation that predicts and optimizes infrastructure performance in real-time.

World-class price performance

Achieve unmatched cost efficiency with best-in-class performance for every workload.

Tools for Intelligent Data Warehousing

Powered by Intelligence Engine



Unified SQL Editor

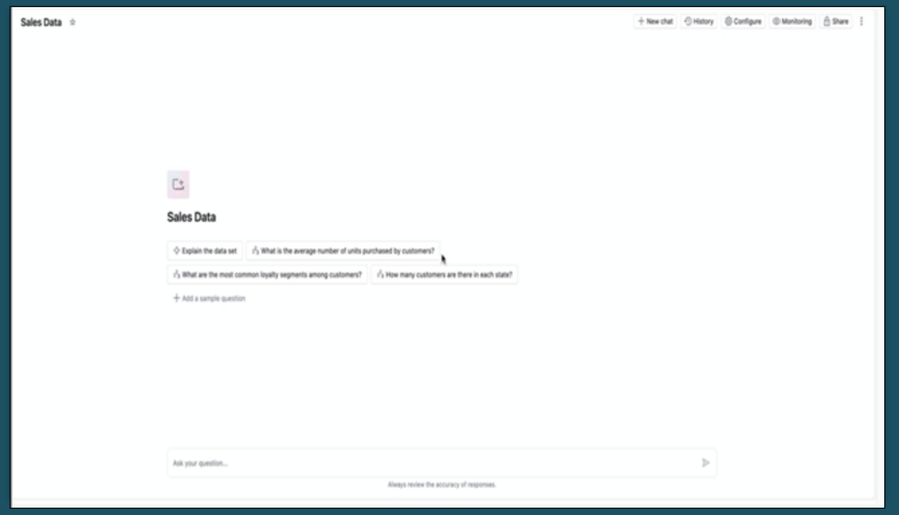
Data Science / Engineers

Tools for Intelligent Data Warehousing

Powered by Intelligence Engine

AI/BI Dashboards

Analysts

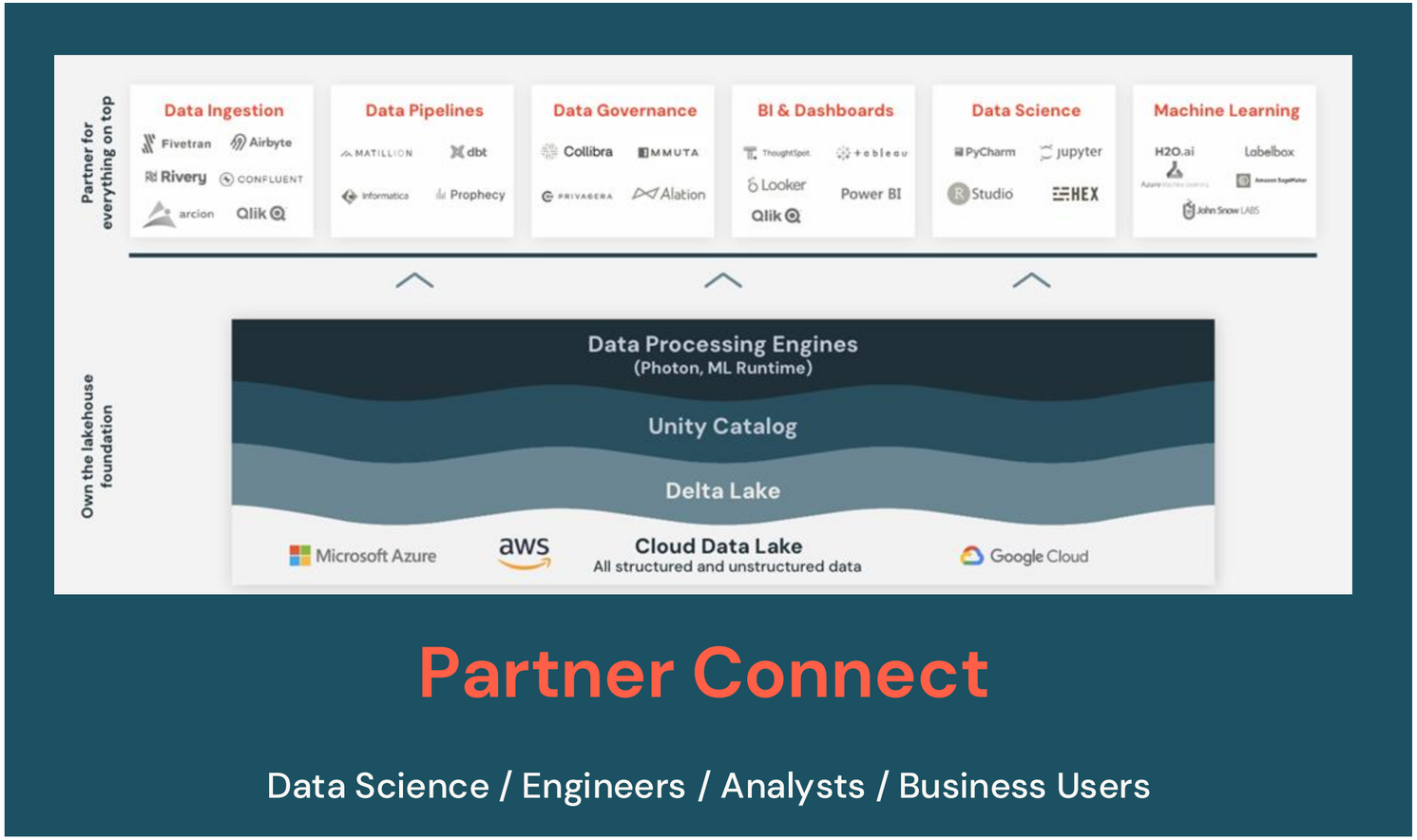


AI/BI Genie

Business Users

Tools for Intelligent Data Warehousing

Powered by Intelligence Engine

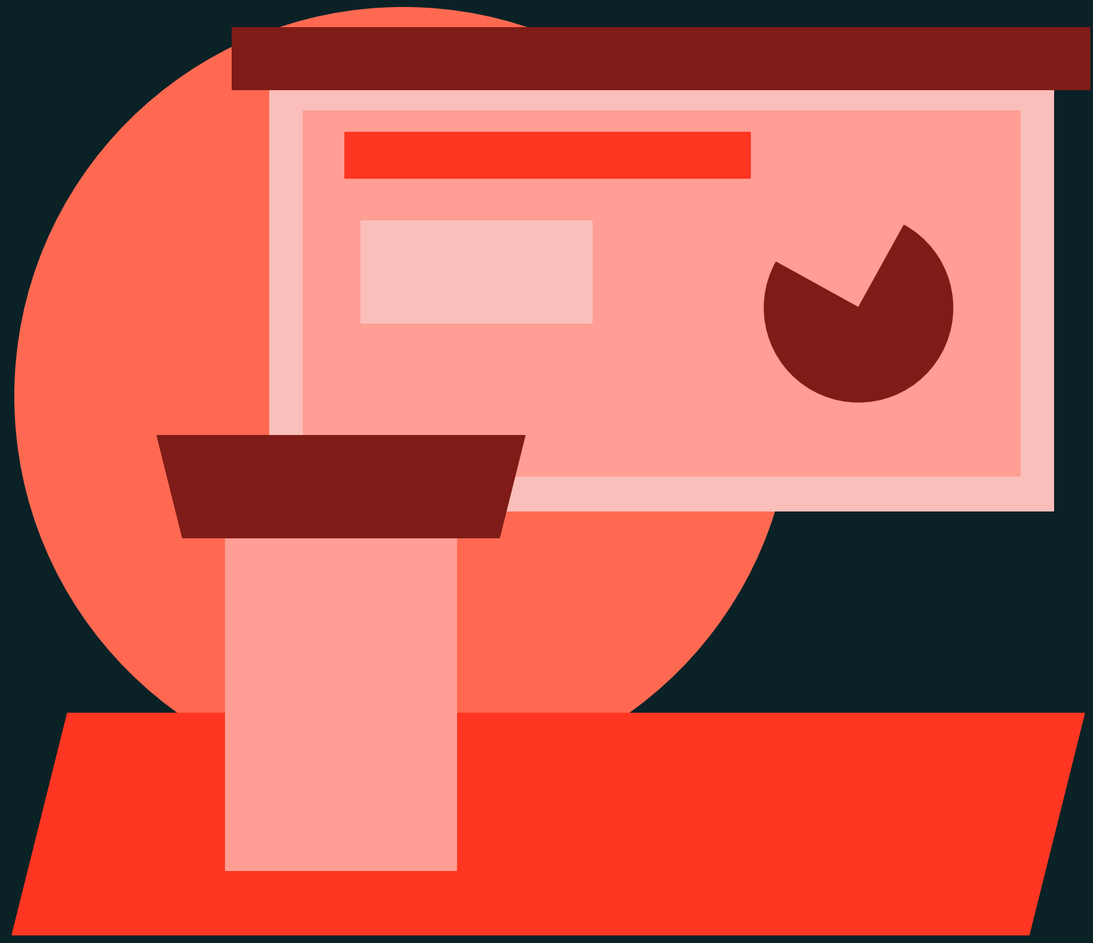


databricks

Using Databricks for Data Warehousing

LECTURE

Databricks SQL Warehouses



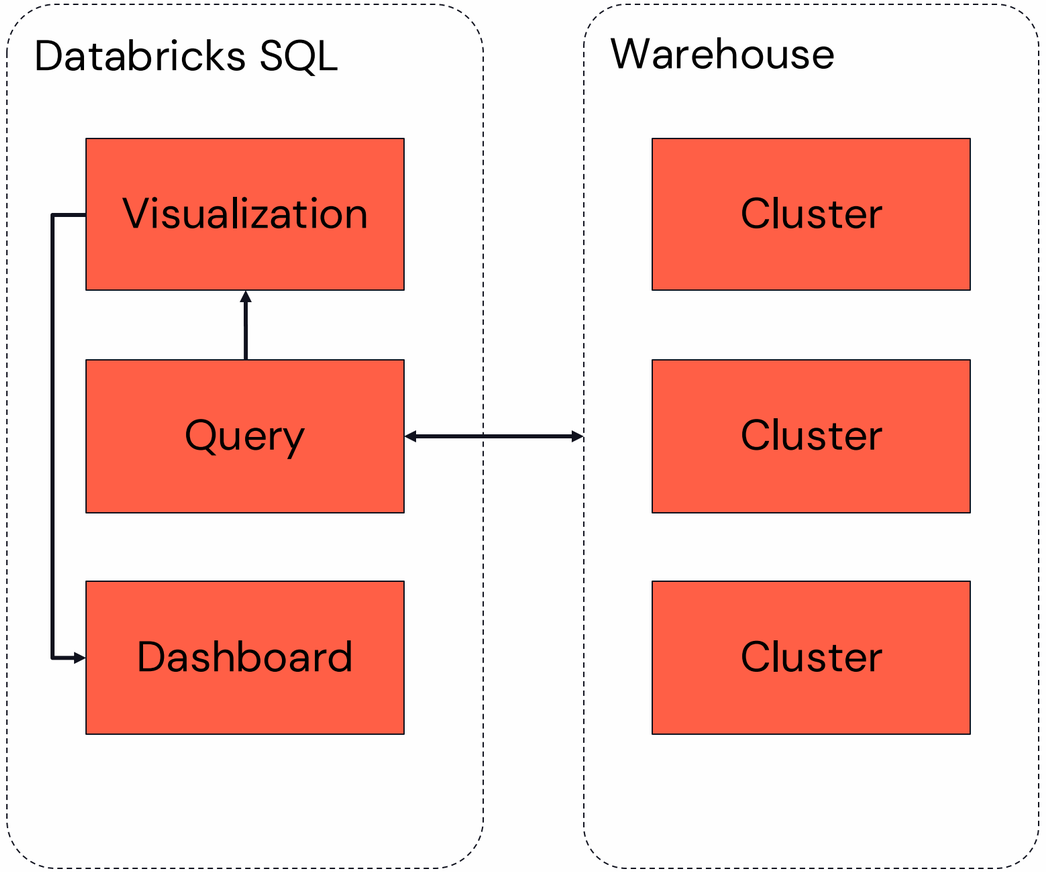
SQL Warehouses

Overview

SQL Warehouse in Databricks is a scalable compute resource optimized for executing SQL queries, analyzing data, creating visualizations, and supporting business intelligence (BI) workflows.

Key Features:

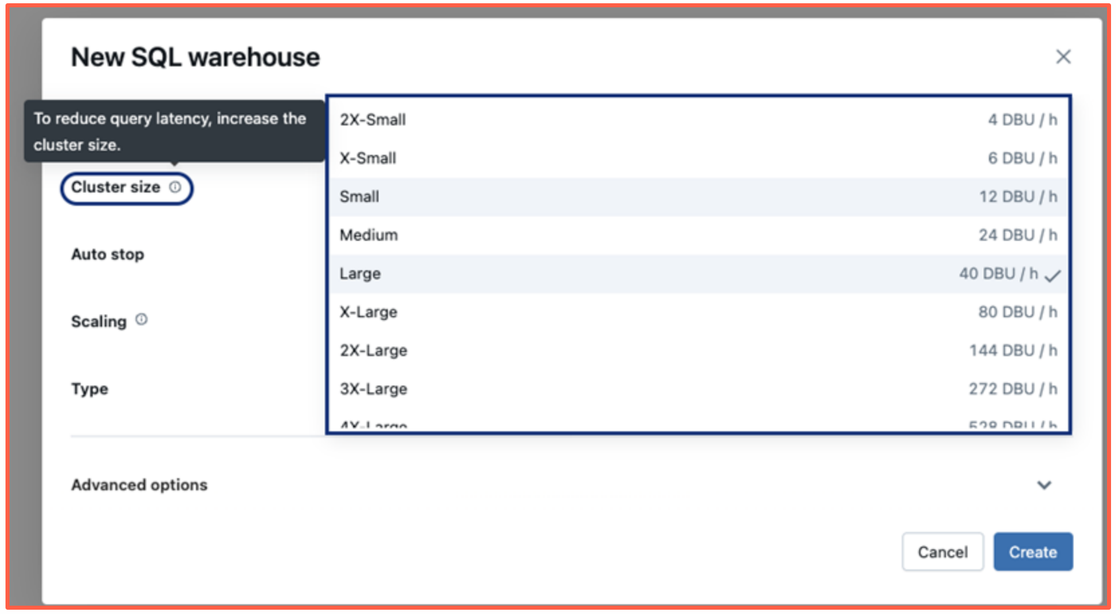
- Scalable compute for distributed SQL workloads.
- Optimized for data queries and BI workflows.
- Supports integration with dashboards and analytics tools.



SQL Warehouses

Cluster Size:

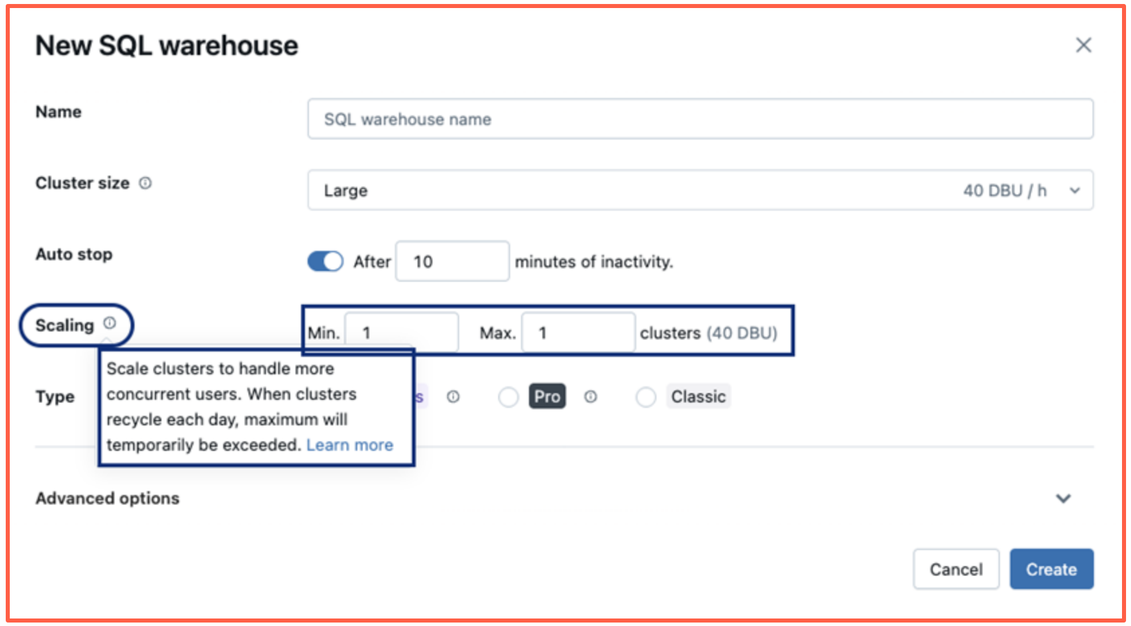
- Refers to the computational power allocated to a SQL Warehouse.
- Expressed in **“t-shirt sizes”** such as **Small**, **Medium**, **Large**, and **X-Large**.
- Each size defines the number of compute nodes and capacity for parallel processing.
- **Example:**
  - **Small:** Suitable for light queries.
  - **X-Large:** Ideal for heavy, high-concurrency workloads.



SQL Warehouses

Scaling:

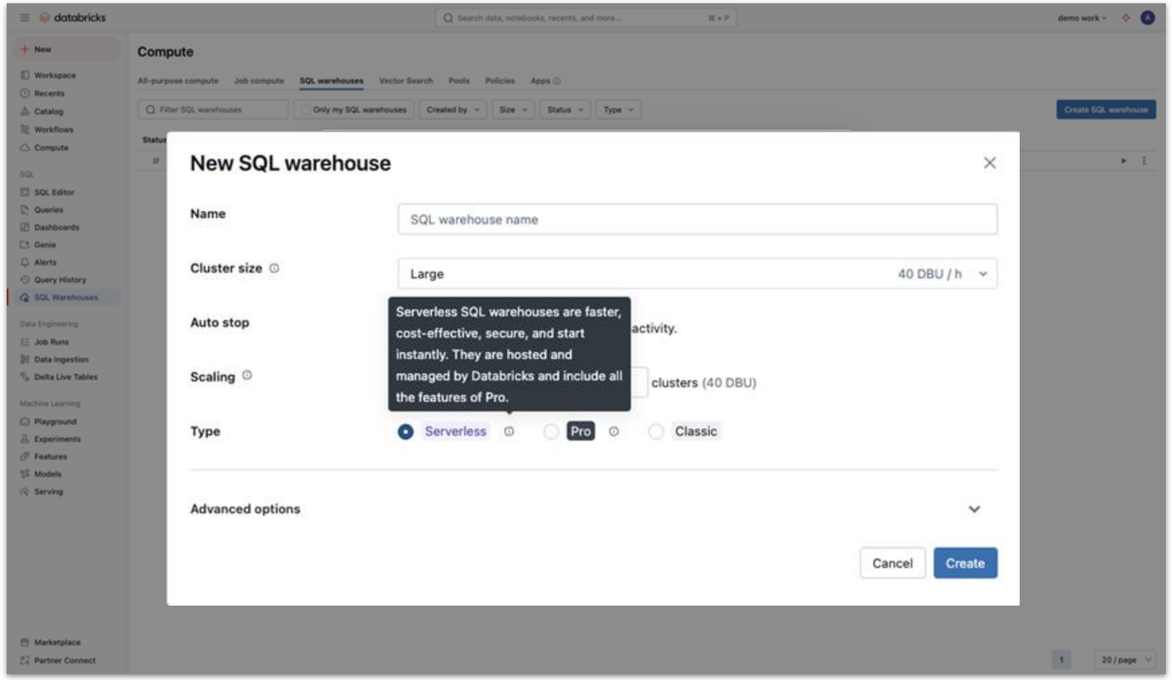
- Ensures SQL Warehouses adapt to varying workloads by dynamically allocating resources.
- **Users can configure:**
  - **Minimum clusters:** Base capacity for consistent performance.
  - **Maximum clusters:** Auto-scale limit to handle peak demand.
- **Benefits:**
  - **Cost Efficiency:** Only pay for resources used.
  - **Performance Optimization:** Prevents resource bottlenecks.



Serverless Compute for Databricks SQL

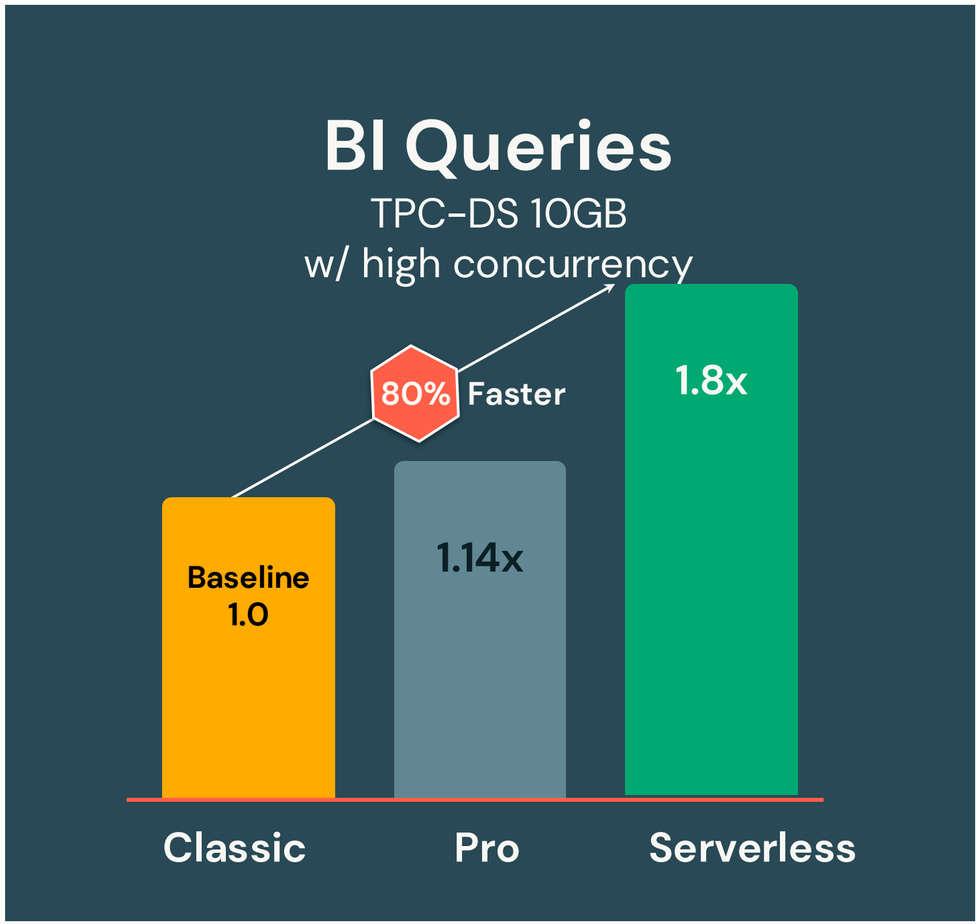
Instant, elastic SQL compute decoupled from storage

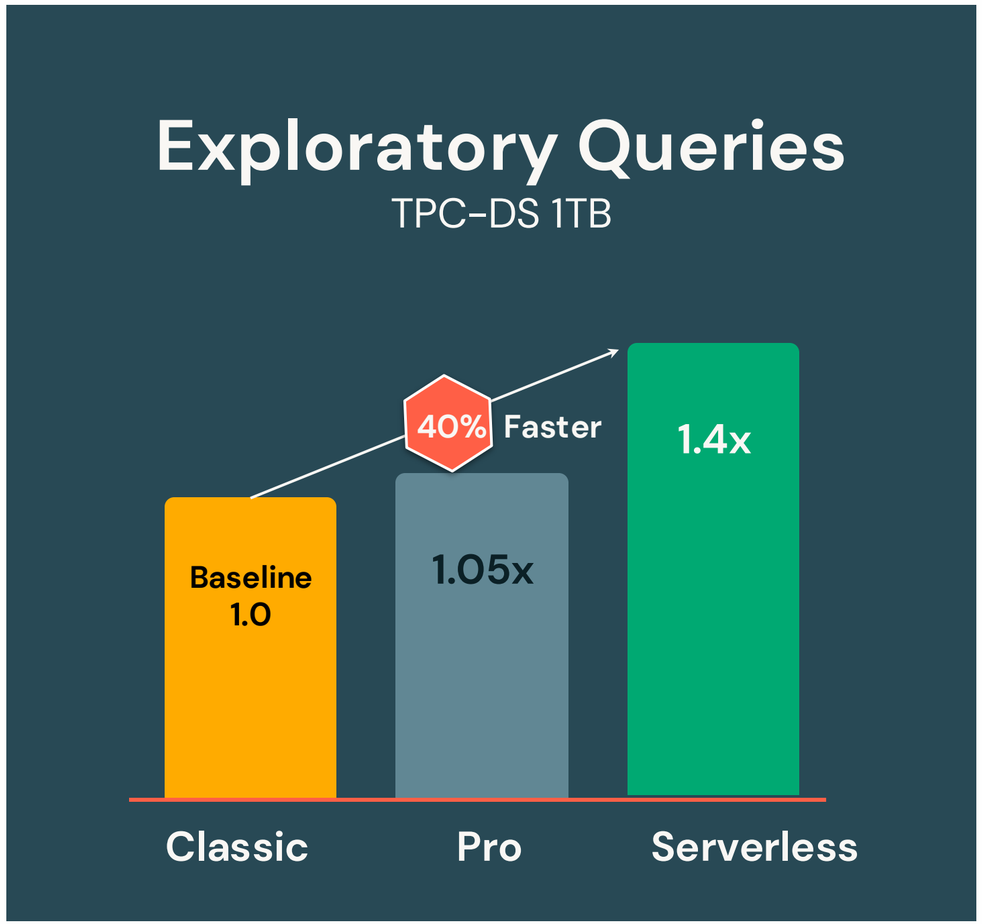
- Quickly setup **instant**, **elastic SQL warehouse** - decoupled from storage - Powered by Photon
- Automatically determines instance types and configuration for best price/perf (up to 12x)
- **High concurrency** built-in, automatic load balancing
- Intelligent **workload management** and faster reads from cloud storage
- Instant startup, greater availability, and **40% average lower overall costs with serverless**



Serverless SQL has the Best Performance!

Let's see how they perform in real-world scenarios:





databricks

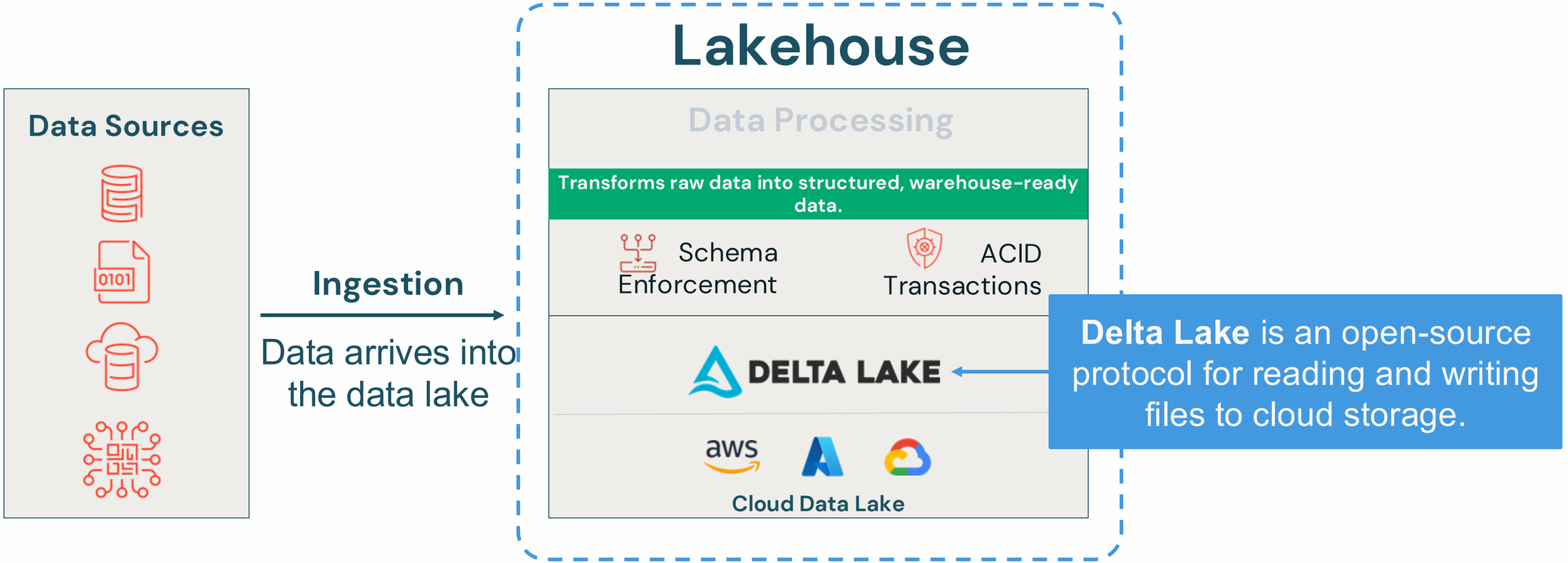
Using Databricks for Data Warehousing

LECTURE

Delta Lake Overview

Delta Lake Overview

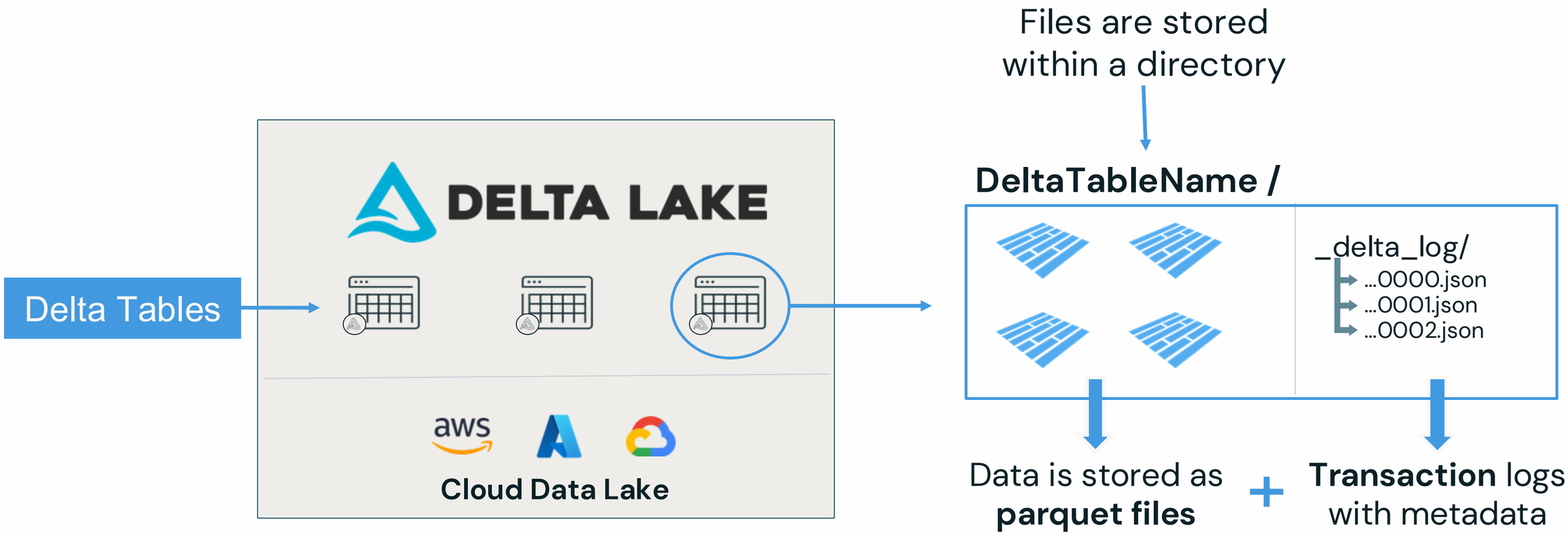
Delta Lake: Foundation for Modern Data Warehousing.



Delta Lake ensures reliability, performance, and scalability for modern data warehousing workflows.

Delta Lake Overview

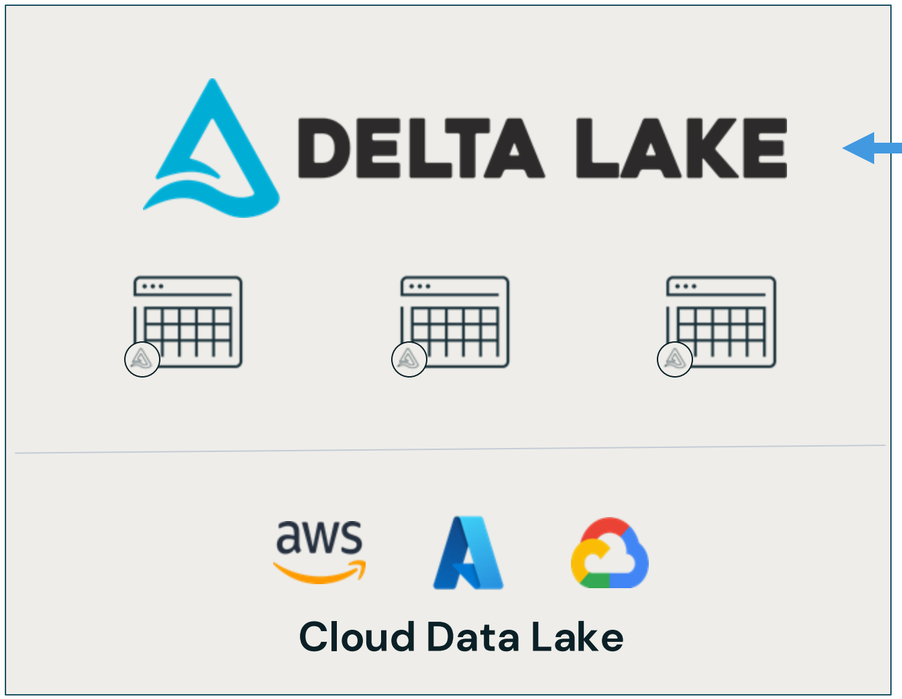
Delta Tables: The Backbone of Reliable Data Warehousing



Delta Tables support both batch and streaming data, making them flexible for real-time and historical analytics in a warehouse.

Delta Lake Overview

Delta Lake: Default Format for Databricks Warehousing



Delta Lake is the default format for tables created in Databricks

SQL

```
CREATE TABLE orders USING DELTA
```

Python

```
df.write.format("delta")
```

Delta Tables power Serverless SQL Warehouses, providing instant scalability and optimized performance for data warehousing workflows.

Delta Lake Overview

Key Features

ACID Transactions

DML Operations

Time Travel

Schema Evolution and Enforcement

Many more!

Delta Lake Overview

Key Features

ACID Transactions

DML Operations

Time Travel

Schema Evolution and Enforcement

Many more!

Delta Lake provides ACID transactional guarantees at the table level to data stored in **cloud-based object storage**.

Atomicity - Entire transaction completes

Consistency – Data follows rules or it will be rolled back

Isolation – One transaction completed before the start of another

Durability – Data is saved in a persistent state once completed

Delta Lake Overview

Key Features

ACID Transactions

DML Operations

Time Travel

Schema Evolution and Enforcement

Many more!

Data Manipulation Language (DML)

- **INSERT** - Add new rows
- **UPDATE** - Update existing values
- **DELETE** - Delete rows
- **MERGE** - Selects records from a source table and performs multiple DML operations on a target table

Delta Lake Overview

Key Features

ACID Transactions

DML Operations

Time Travel

Schema Evolution and Enforcement

Many more!

Delta Tables keep a **transaction log** for each **version (writes)** of the table.

- Historical Querying
- Snapshot Isolation
- Auditing

Delta Lake Overview

Key Features

ACID Transactions

DML Operations

Time Travel

Schema Evolution and Enforcement

Many more!

Schema Evolution

- Automatically **adjusts** the schema of your Delta table as your data changes

Delta Lake Overview

Key Features

ACID Transactions

DML Operations

Time Travel

Schema Evolution and Enforcement

Many more!

Schema Evolution

- Automatically **adjusts** the schema of your Delta table as your data changes

Schema Enforcement

- Ensures that any data written to the Delta table **matches** the table's defined schema

Delta Lake Overview

Key Features

ACID Transactions

DML Operations

Time Travel

Schema Evolution and Enforcement

Many more!

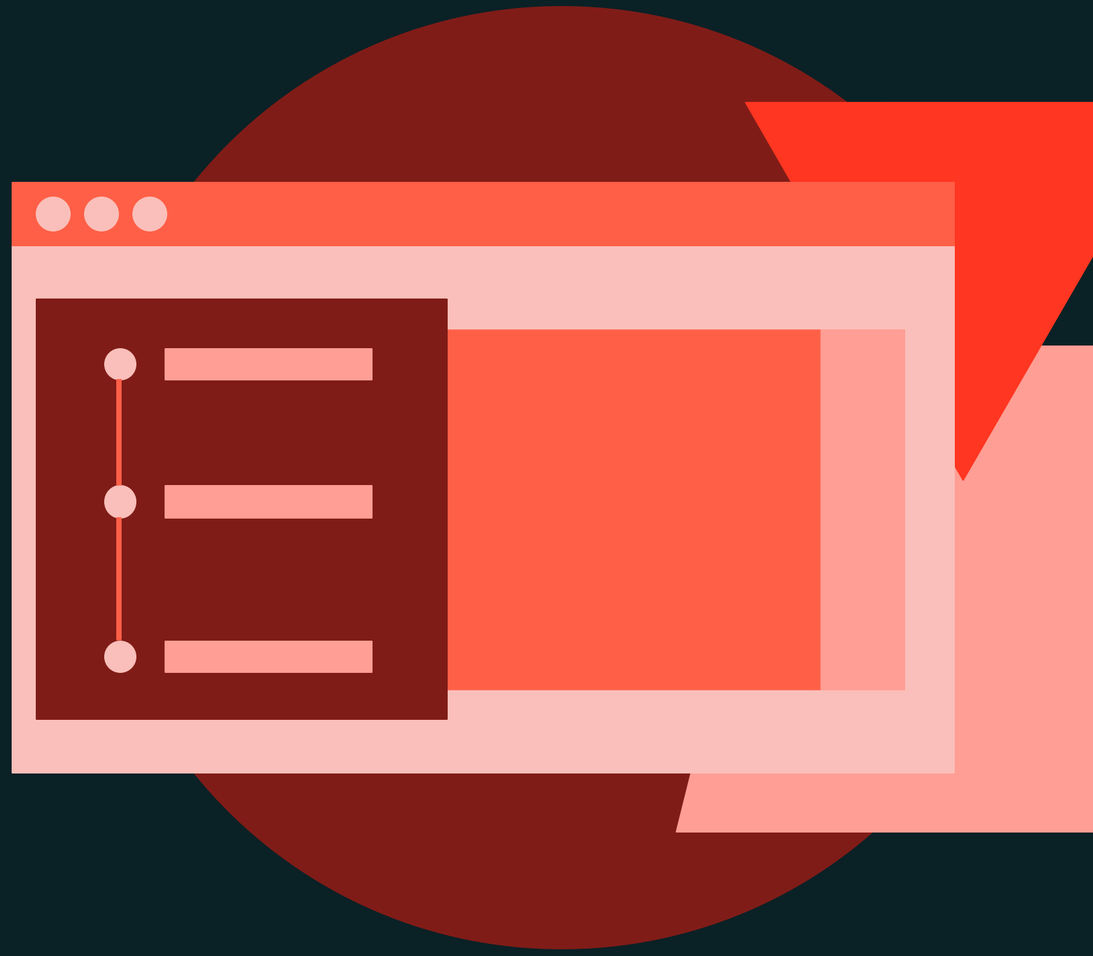
- Unified Batch and Streaming
- Performance
- Scalable Metadata
- Optimization
- **Delta Lake is Open Source!**

databricks

Using Databricks for Data Warehousing

DEMONSTRATION

Using Delta Lake features with Databricks SQL

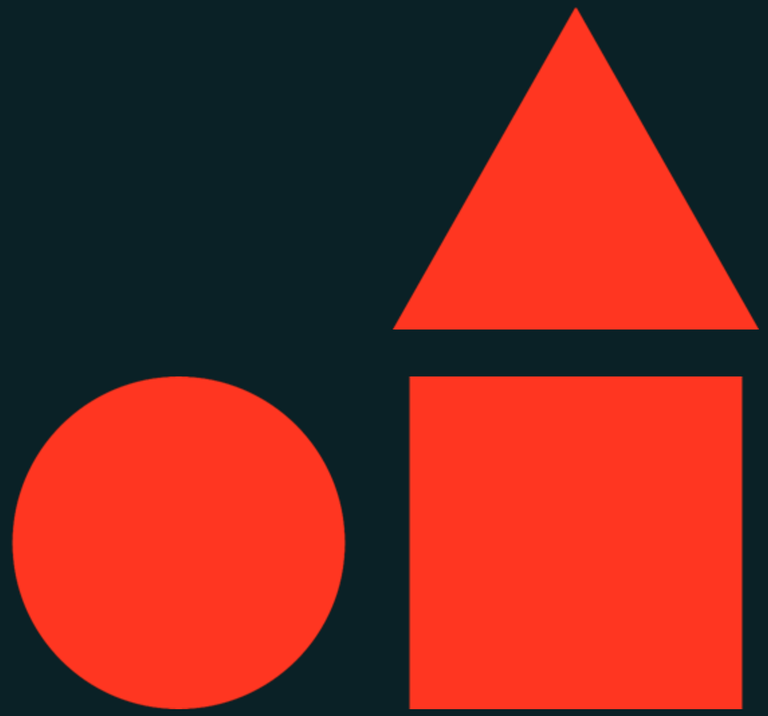


High Level Steps

- Explore key features of Delta Lake such as Time Travel, Version History, and metadata management.
- Use SQL commands like **DESCRIBE EXTENDED**, **DESCRIBE HISTORY**, **VERSION AS OF**, and **RESTORE TABLE**.
- Understand how Delta Lake's features are beneficial for managing data in data warehousing scenarios.
- Work with the **sales_orders** table in the <schema_name> schema.

databricks

Data Ingestion and Transformation



Get Started with Databricks for Data Warehousing

Objectives

- Describe and apply various techniques for ingesting and transforming data in Databricks.

Agenda

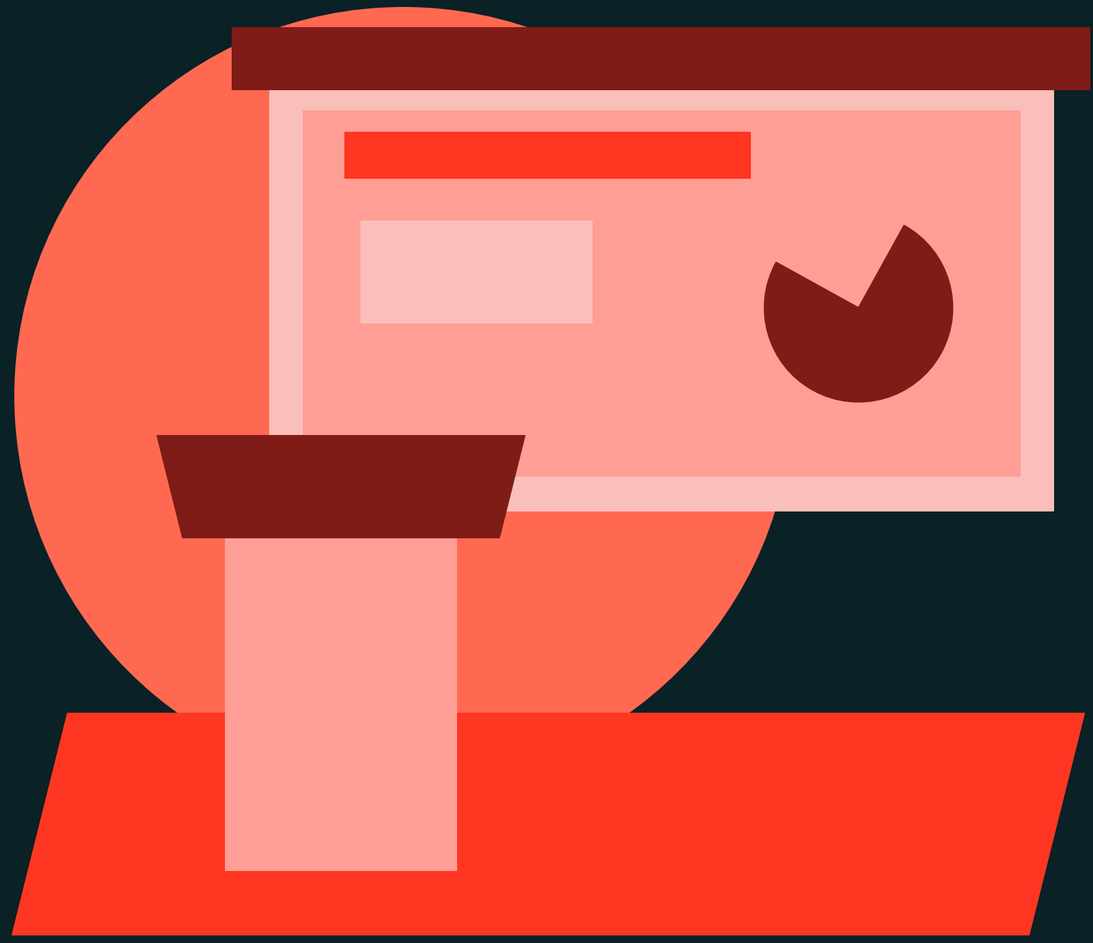
Data Ingestion and Transformation	Time	Lecture	Demo	Lab
Delta Lake UniForm	5 mins	✓		
Data Ingestion Techniques Overview	5 mins	✓		
Data Ingestion Techniques	10 mins		✓	
Data Transformation	8 mins	✓		
Exploring Data Transformation in Databricks	10 mins		✓	

databricks

Data Ingestion and Transformation

LECTURE

Delta Lake UniForm



Delta Lake with UniForm

Rise of Open Table Formats

Which open table format to pick?

Delta Lake

Hudi

Iceberg

Delta Lake with UniForm simplifies data management by bridging the gap between different table formats, empowering data warehouses with unified access and control.

Introducing Delta Lake with Uniform (Delta Universal Format)

Delta Lake with UniForm

Open Table Formats Overview

Apache Hudi

Delta Lake

Apache Iceberg

Metadata
Used for transactional source of truth, concurrency control, etc.

Apache Hudi	Delta Lake	Apache Iceberg
Metadata	Metadata	Metadata
Parquet	Parquet	Parquet

Connector Ecosystem

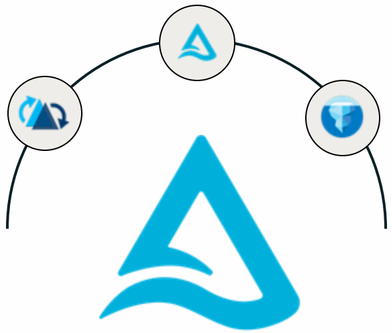
Apache Hudi: trino, Starburst, DORIS, ClickHouse, StarRocks

Delta Lake: Microsoft Fabric, kafka, Athena, amazon REDSHIFT, dbt

Apache Iceberg: presto, snowflake, HIVE, Google Big Query, dremio

Delta Lake with UniForm

How does it work?



Delta Lake With UniForm

Metadata

Metadata
Used for transactional source of truth, concurrency control, etc.

✓ **Metadata automatically generated to make Delta accessible as Iceberg/Hudi readers**

Delta Lake with UniForm

How does it work?

Delta Lake With UniForm

Metadata
Used for transactional source of truth, concurrency control, etc.

Data
All formats use Parquet!

- ✓ Metadata automatically generated to make Delta accessible as Iceberg/Hudi readers
- ✓ **Parquet files remain the same**

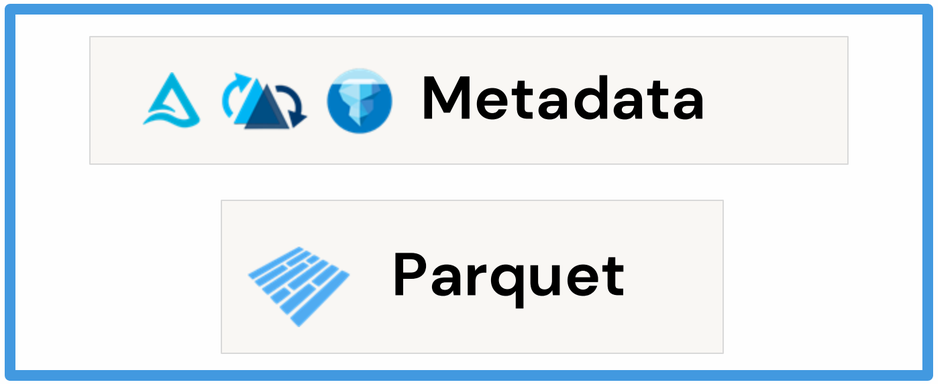
Delta Lake with UniForm

How does it work?

Delta Lake With UniForm

Metadata
Used for transactional source of truth, concurrency control, etc.

Data
All formats use Parquet!



- ✓ Metadata automatically generated to make Delta accessible as Iceberg/Hudi readers
- ✓ Parquet files remain the same
- ✓ **Metadata is co-located with data**

Delta Lake with UniForm

How does it work?

Delta Lake With UniForm

Metadata
Used for transactional source of truth, concurrency control, etc.

Metadata

Data
All formats use Parquet!

Parquet

- ✓ Metadata automatically generated to make Delta accessible as Iceberg/Hudi readers
- ✓ Parquet files remain the same
- ✓ Metadata is co-located with data
- ✓ **Data stored in Delta can be read as if it were Iceberg or Hudi**

Connector Ecosystem

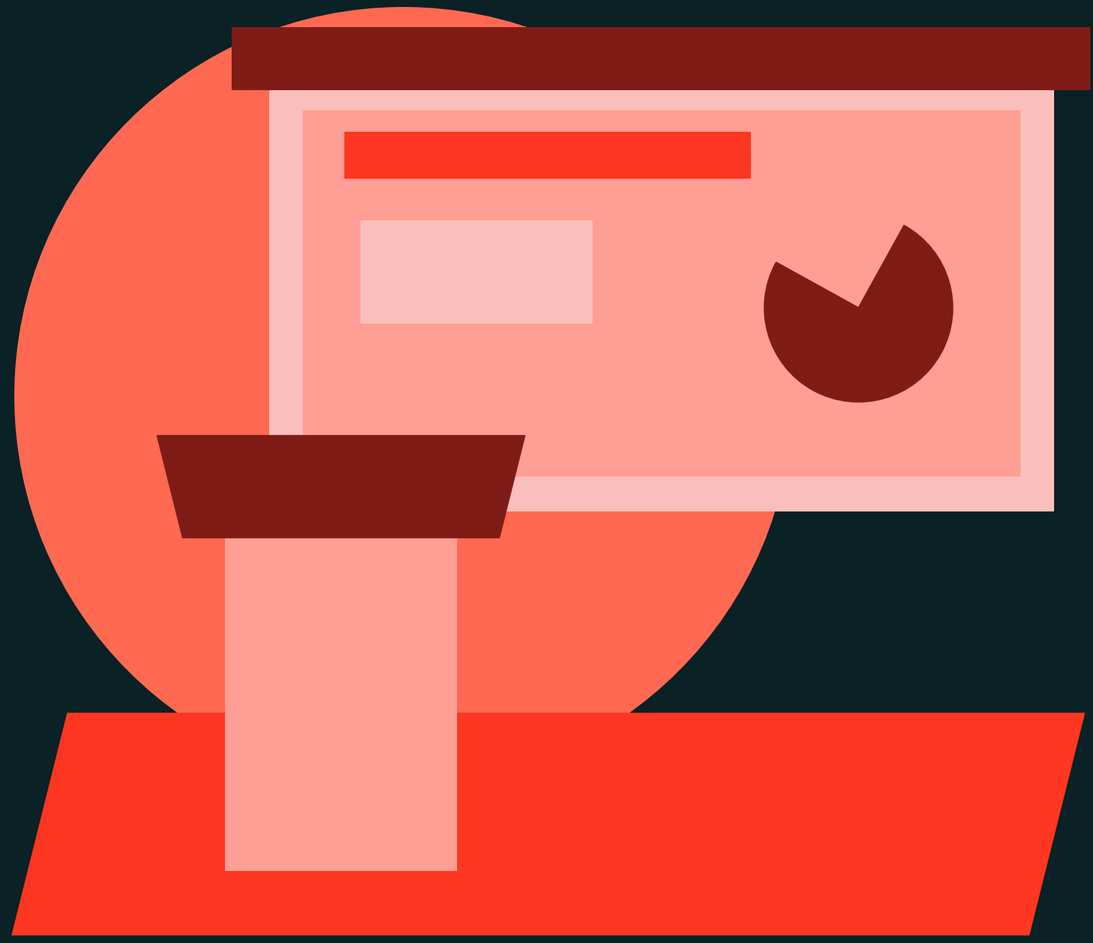
trino, Starburst, Microsoft Fabric, kafka, Athena, presto, snowflake, HIVE, DORIS, ClickHouse, StarRocks, amazon REDSHIFT, dbt, Google Big Query, dremio

databricks

Data Ingestion and Transformation

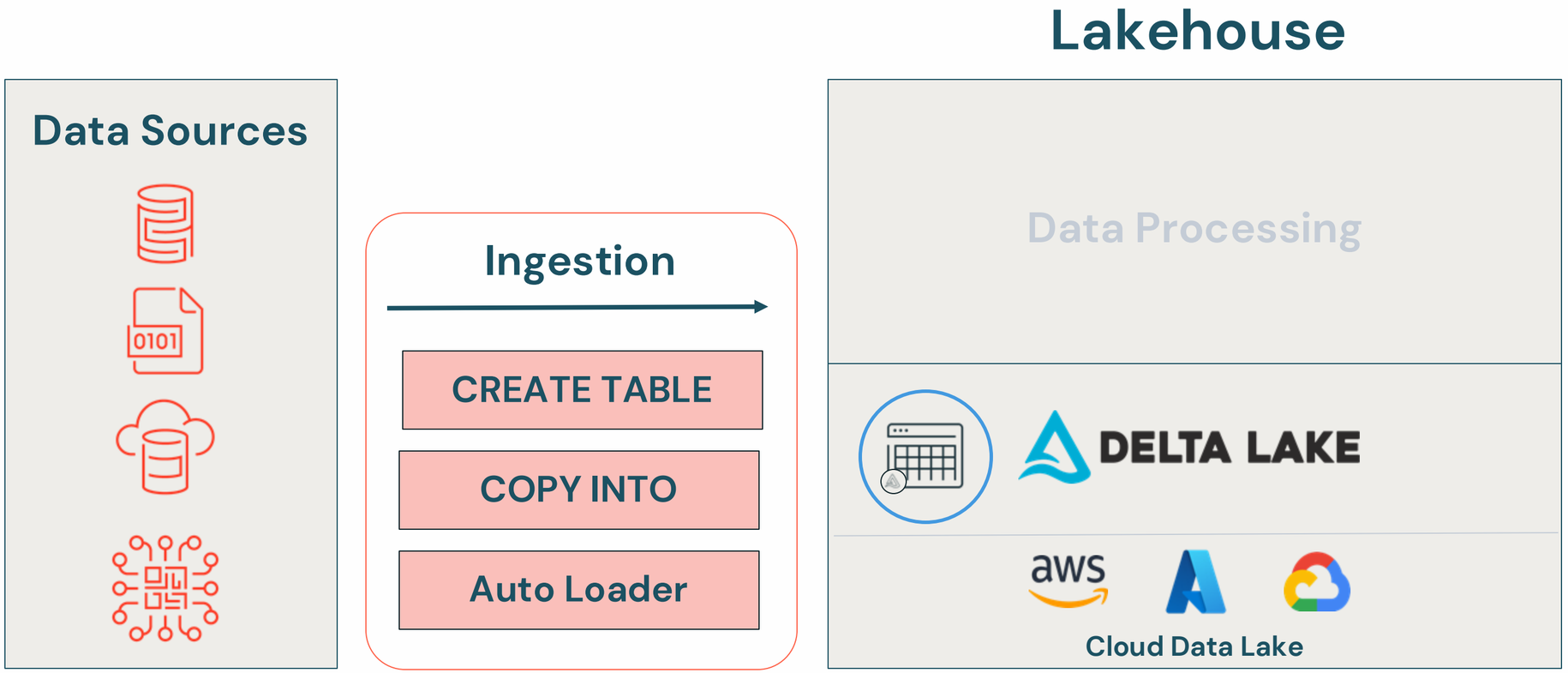
LECTURE

Data Ingestion Techniques Overview



Delta Lake Data Ingestion Techniques

Overview



Delta Lake Data Ingestion Techniques

Overview

CREATE TABLE

Upload UI

COPY INTO

Auto Loader

Delta Lake Data Ingestion Techniques

Overview

CREATE TABLE

Upload UI

COPY INTO

Auto Loader

```
CREATE TABLE mydeltatable
USING DELTA
AS
<query>
```

USING DELTA is optional.

- CREATE TABLE AS (CTAS) creates a table by selecting data from an existing table or data source
- Default table format is **Delta**

Delta Lake Data Ingestion Techniques

Overview

CREATE TABLE

Upload UI

COPY INTO

Auto Loader

- The UPLOAD UI provides a **point-and-click interface** to upload files and create tables
- Allows you to upload CSV, TSV, JSON, Avro, Parquet, or text files
- The UI also enables you to upload files directly to a Unity Catalog volume

Delta Lake Data Ingestion Techniques

Overview

CREATE TABLE

Upload UI

COPY INTO

Auto Loader

```
COPY INTO mydeltatable
FROM 'your-path'
FILE_FORMAT = 'format'
FILE_OPTIONS = ('format-options')
```

- Loads files from a **file location** into a Delta table
- Supports **various file formats** and cloud storage locations
- Automatically can handle **schema changes**.
- **Idempotent** operation—Already-loaded files in the source are skipped

Delta Lake Data Ingestion Techniques

Overview

CREATE TABLE

Upload UI

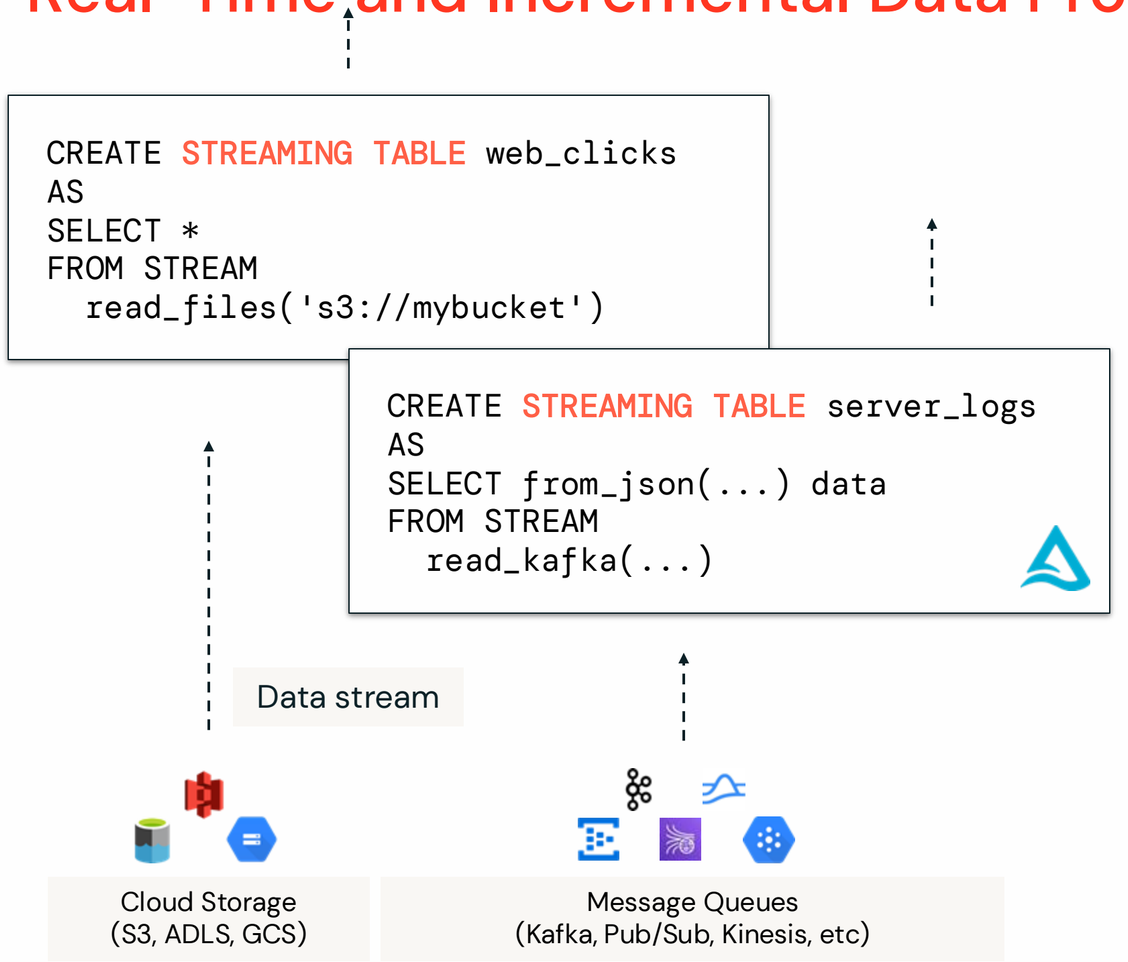
COPY INTO

Auto Loader

- Auto Loader **incrementally (streaming) processes** new data files as they arrive in cloud storage
- Automatically **infers schemas** and accommodates **schema changes**
- Includes a **rescue data** column for data that does not adhere to the schema

Streaming Tables in Delta Lake

Real-Time and Incremental Data Processing Made Simple



Benefits:

1. Unlock real-time use cases. Ability to support real-time analytics/BI, machine learning and operational use cases with streaming data.
2. Better scalability. More efficiently handle high volumes of data via incremental processing vs large batches.
3. Enable more practitioners. Simple SQL syntax makes data streaming accessible to all data engineers and analysts.

Databricks LakeFlow Connect

Native ingestion connectors for the Data Intelligence Platform

Efficient ingestion pipelines

Simple setup and maintenance

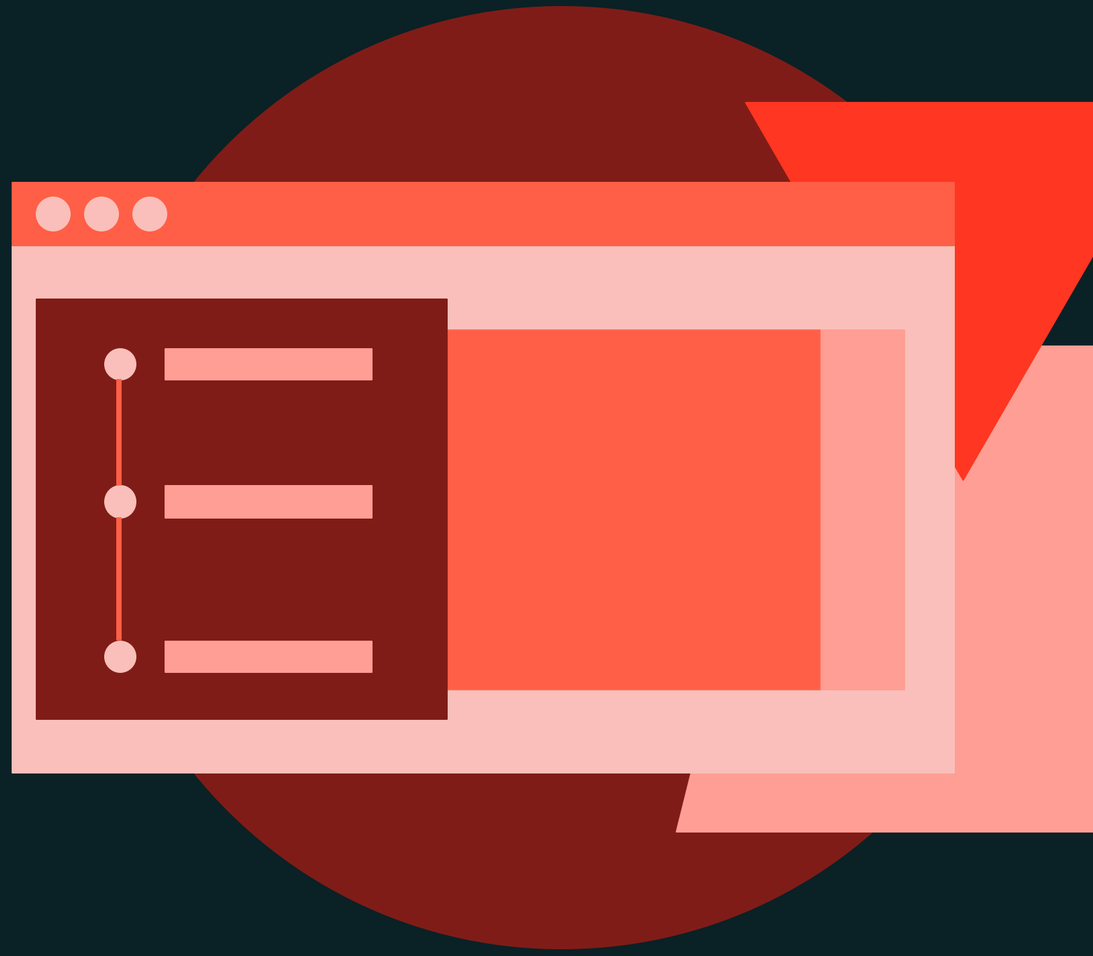
Unified orchestration, observability, and governance

databricks

Data Ingestion and Transformation

DEMONSTRATION

Data Ingestion Techniques



High Level Steps

- Ingest data into Delta Lake using CREATE TABLE AS SELECT.
- Incrementally load data using COPY INTO.
- Use the Databricks Upload UI for data ingestion.
- Set up real-time ingestion pipelines using Auto Loader.
- Understand the role of Lakeflow Connect in data ingestion workflows.

databricks

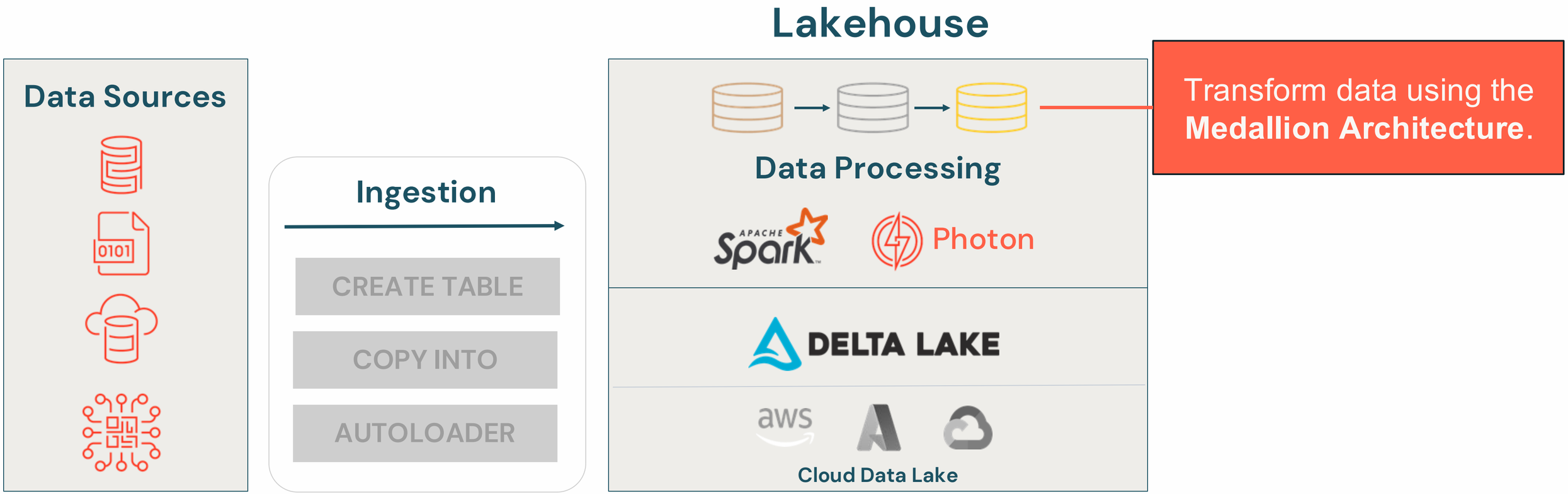
Data Ingestion and Transformation

LECTURE

Data Transformation

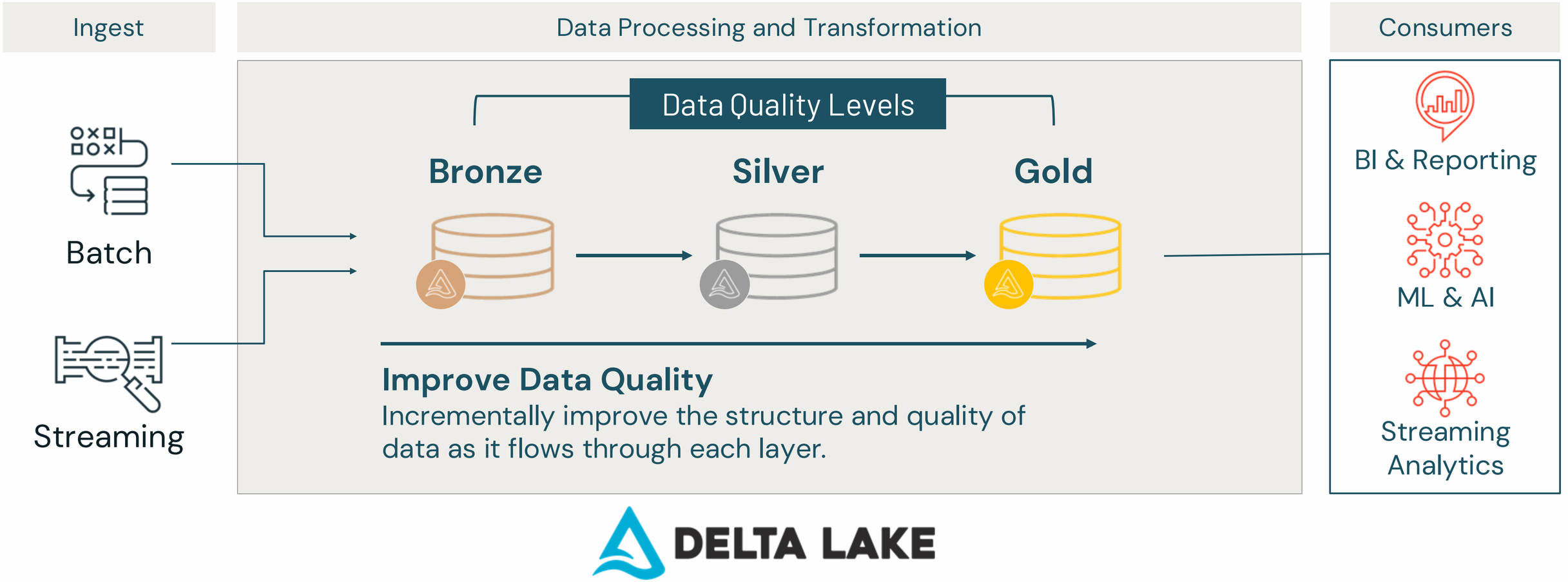
Data Transformation Overview

Medallion Architecture (Multi Hop)



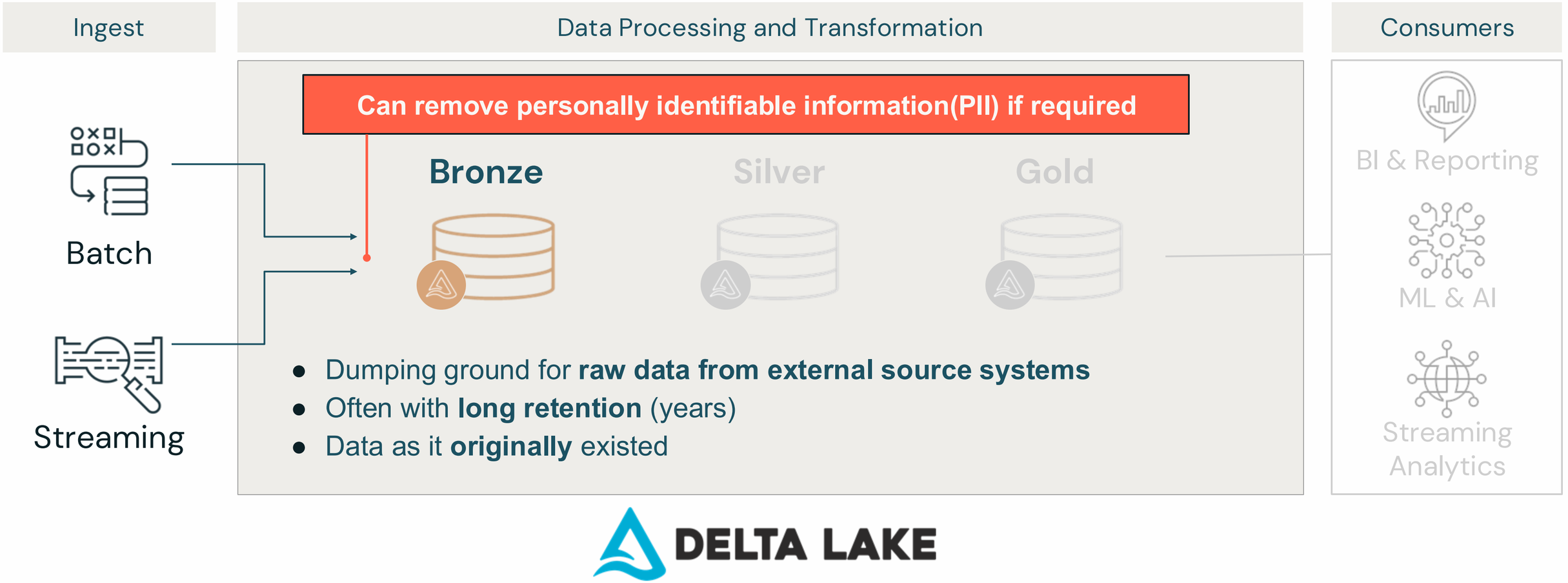
Data Transformation Overview

Medallion Architecture (Multi Hop)



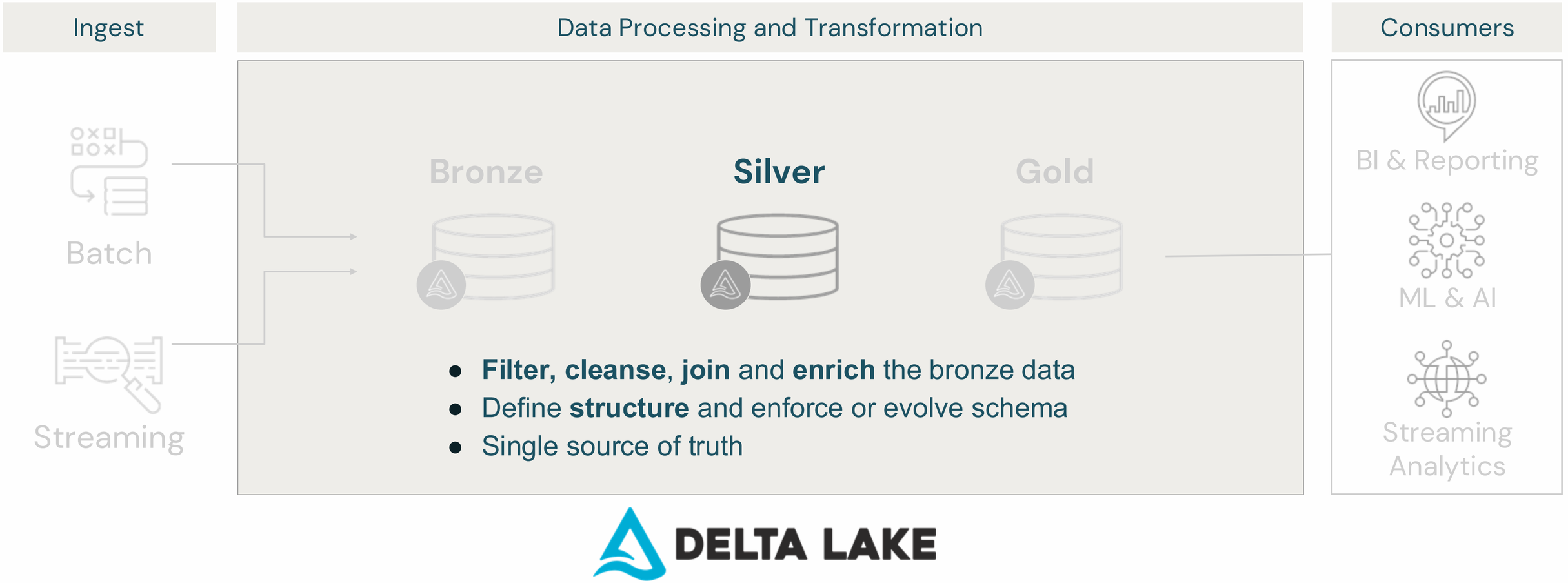
Data Transformation Overview

Medallion Architecture (Multi Hop)



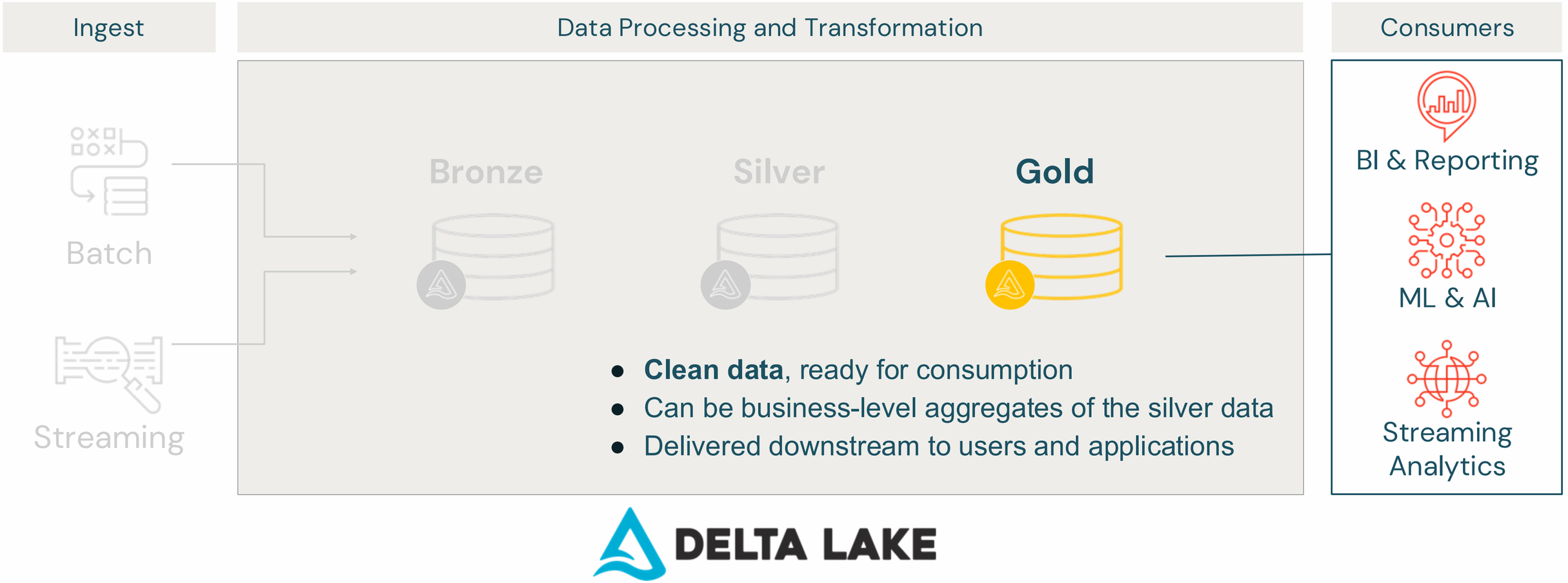
Data Transformation Overview

Medallion Architecture (Multi Hop)



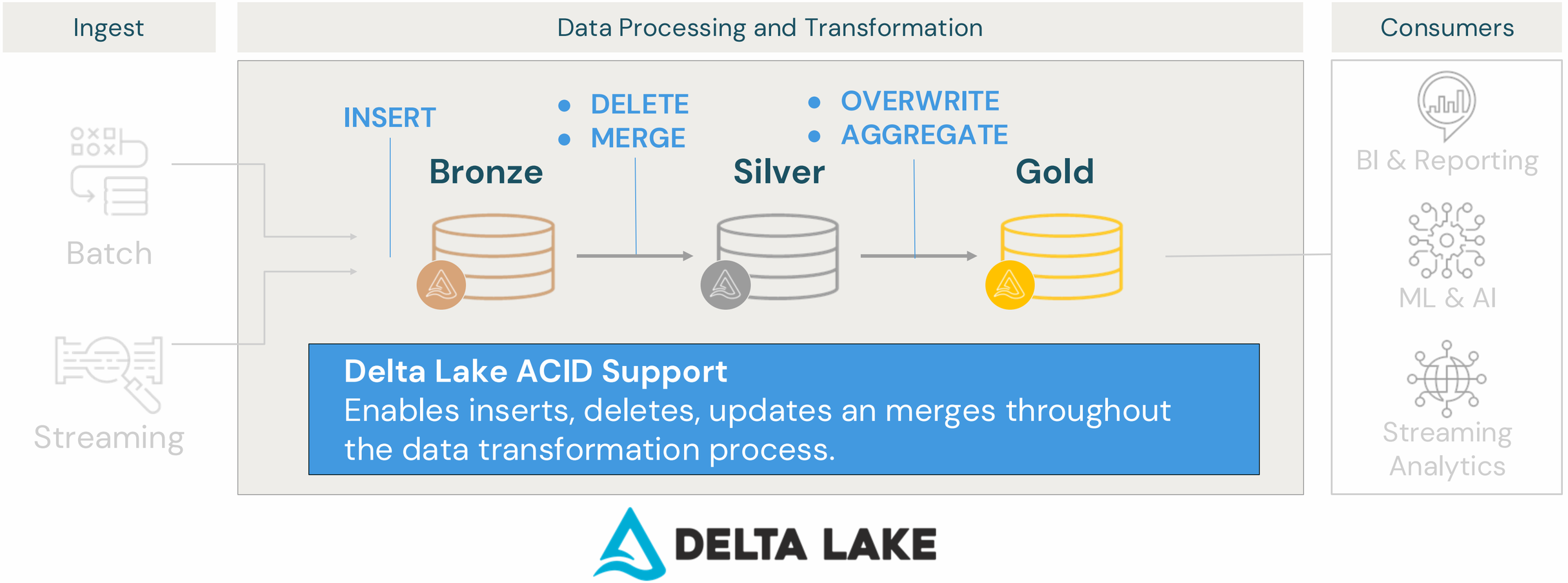
Data Transformation Overview

Medallion Architecture (Multi Hop)



Data Transformation Overview

Medallion Architecture (Multi Hop)

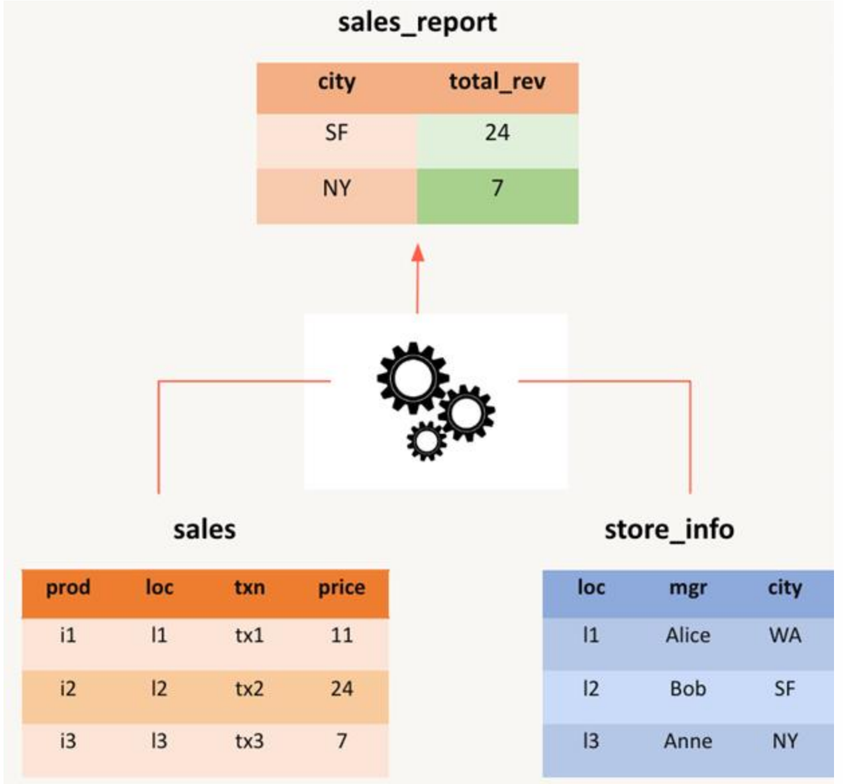


Materialized Views

Speed up queries with pre-computed results

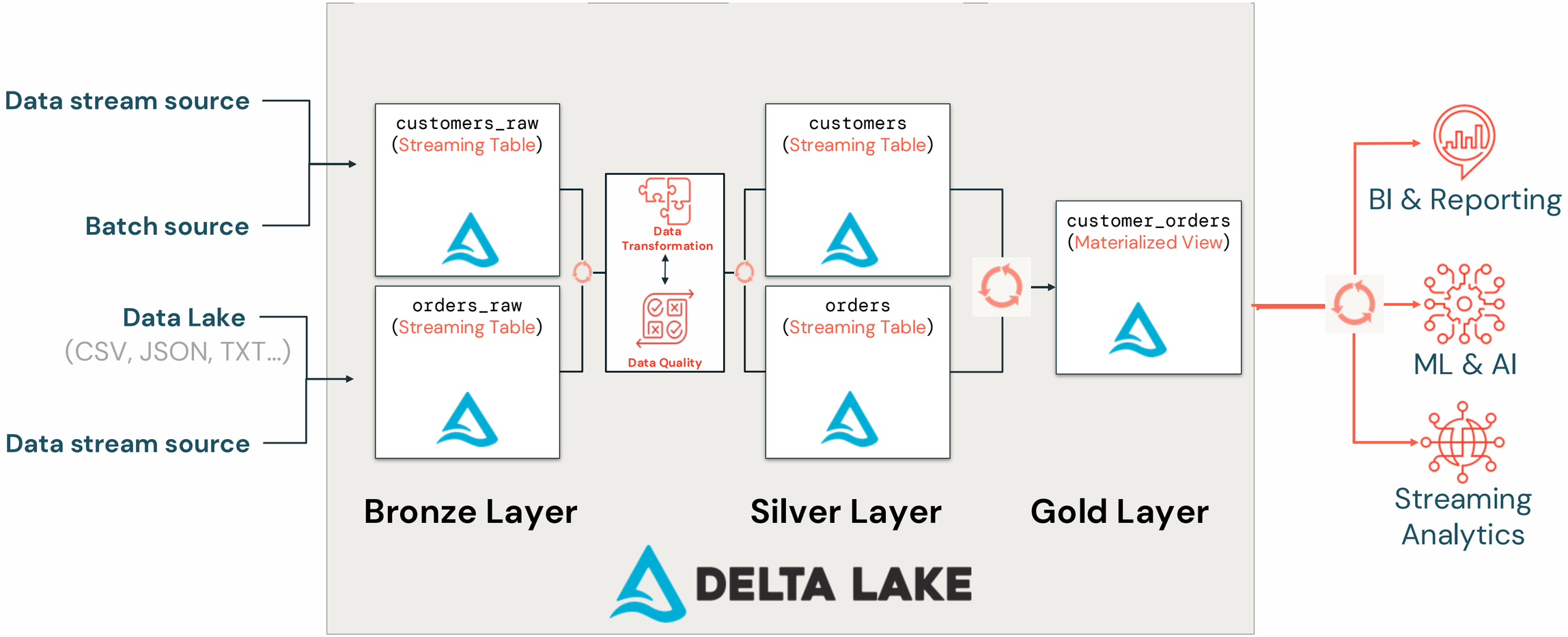
Accelerate end-user queries and reduce infrastructure costs with efficient, incremental computation

- Accelerate **BI dashboards** and **ETL queries**
- **Streaming**: build MVs on top of live tables
- **Easy ELT**: Simplify reporting by cleaning, enriching, denormalizing the base tables
- **Data Sharing & Access Control**: control what info can be seen by internal and external users and organizations



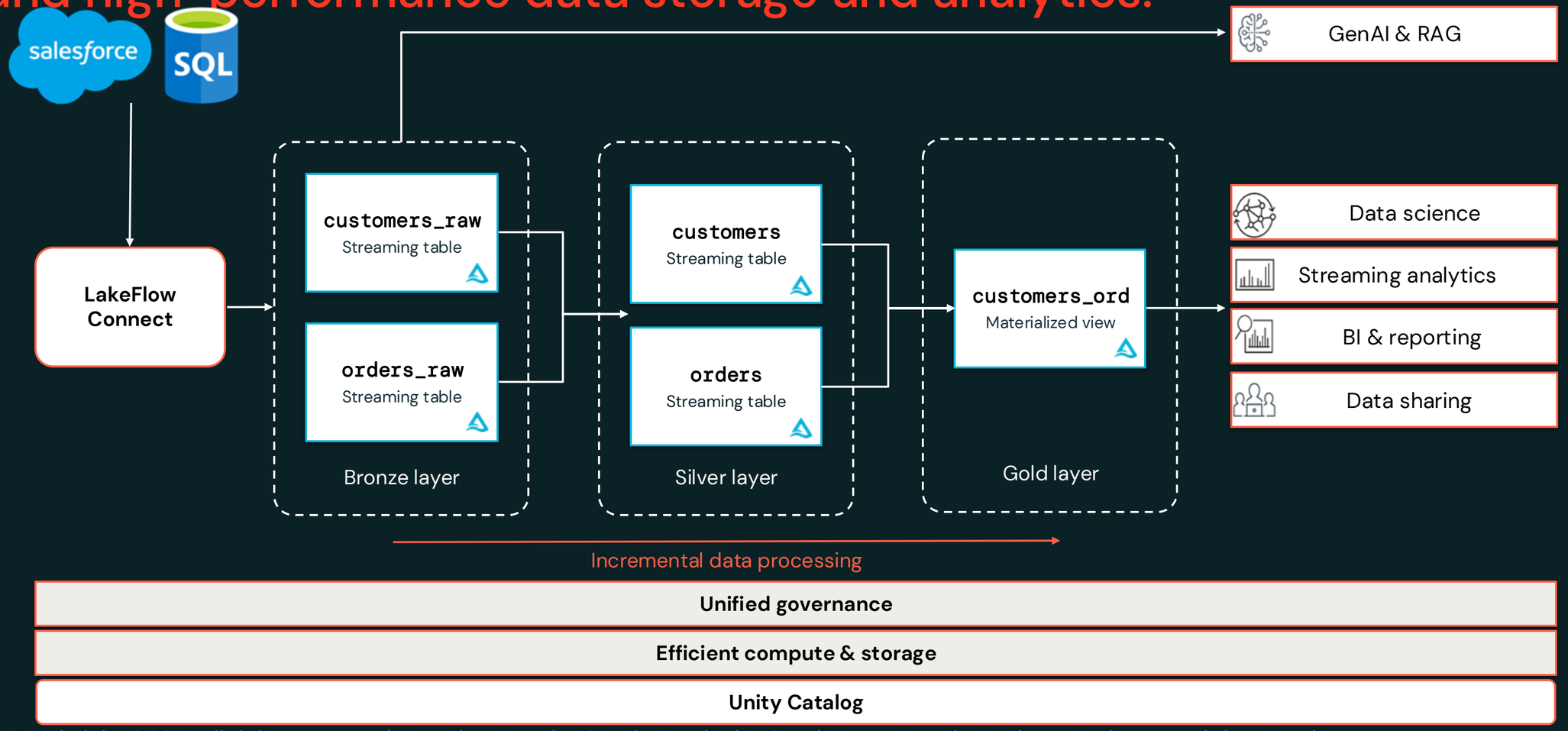
Constructing a Data Transformation Pipeline

Real World Architecture is Typically More Complex



LakeFlow Connect in Action

Medallion Architecture powered by Delta Lake ensures reliable, scalable, and high-performance data storage and analytics.

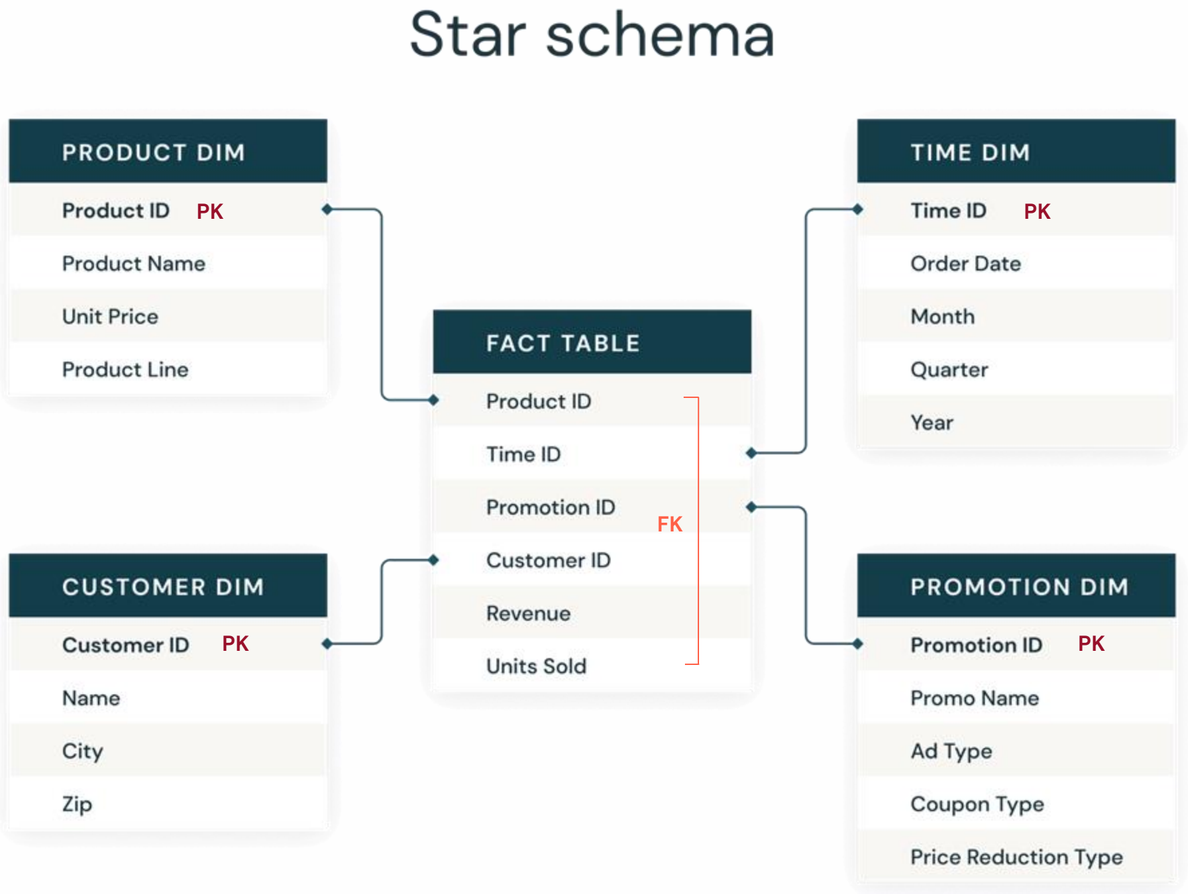


Overview of Data Modeling with Databricks

Introduction to Data Modeling in Databricks

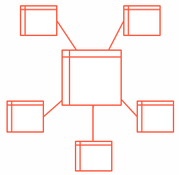
Data Modeling enables structured, high-quality, and consistent data to drive accurate analysis and informed decision-making.

- **Support for Various Modeling Approaches:**
  - Design scalable models using Star Schema, Snowflake Schema, or Data Vault.
- **Primary Key/Foreign Key Constraints:**
  - Ensure referential integrity in Delta tables for reliable relationships.
- **Surrogate Keys:**
  - Enable efficient data management with auto-incrementing IDs.



Modern Data Modeling Approaches

Key Modeling Approaches:



Top-Down Approach (Inmon):

- Relational Modeling with normalized data as the core of the Data Warehousing
- Data marts (often dimensional or denormalized models)

Bottom-Up Approach (Kimball):

- Dimensional Modeling
- Denormalized data model, built as a star or snowflake schema: central fact tables surrounded by dimension tables

Data Vault (Linstedt):

The Data Vault model is based on three basic entity types:

- **Hubs** separate core business concepts
- **Links** store relationships between business concepts
- **Satellites** store the attributes of a business concepts or relationships

Unique Databricks Features:

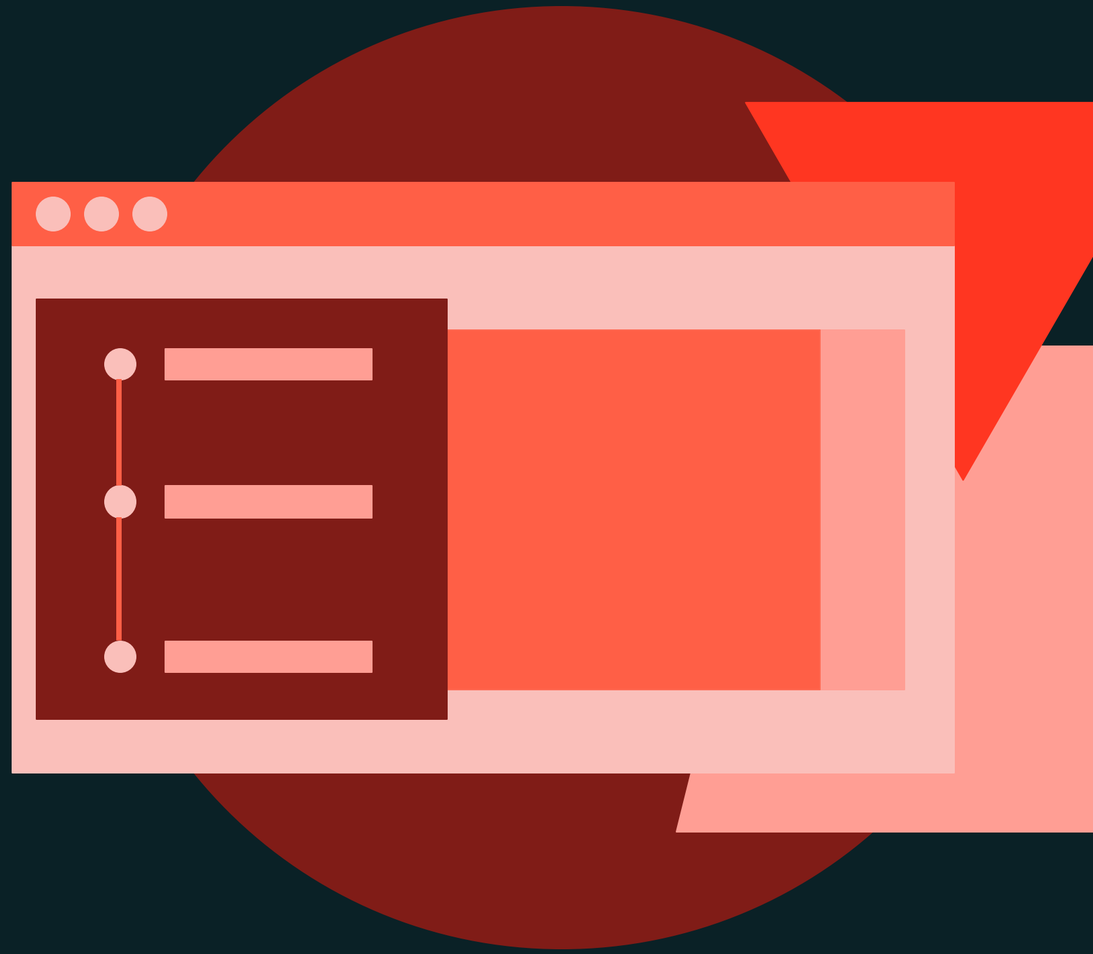
- Seamless integration with Delta Lake for schema management.
- Identity columns for generating Surrogate Keys.

databricks

Data Ingestion and Transformation

DEMONSTRATION

Exploring Data Transformation in Databricks

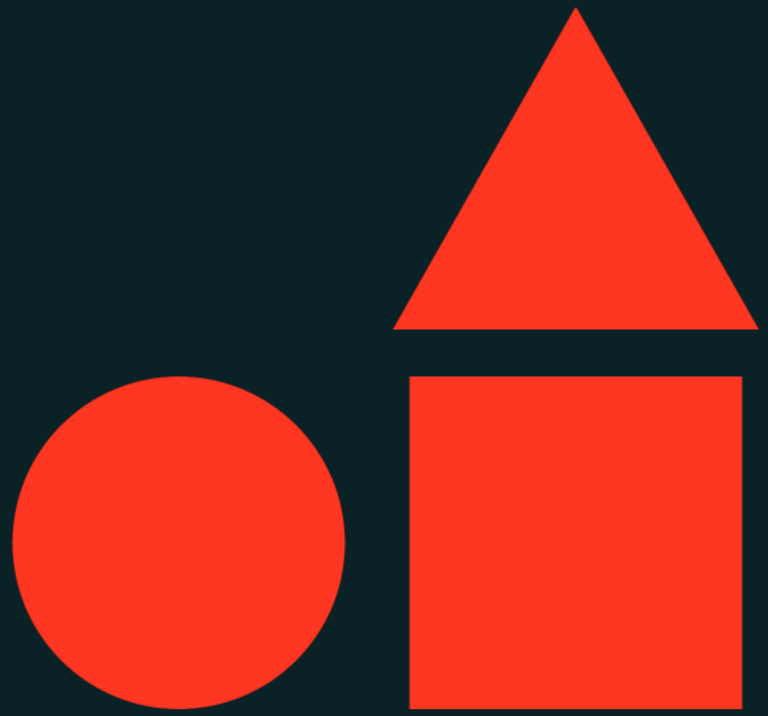


High Level Steps

- Create a multi-hop transformation pipeline using the Medallion Architecture.
- Use Streaming Tables and Materialized Views to transform raw data into meaningful insights.
- Demonstrate how to move data from the Bronze layer to the Silver and Gold layers while maintaining data quality.

databricks

Data Orchestration and Querying Capabilities



Get Started with Databricks for Data Warehousing

Objectives

- Explain how Databricks support data orchestration needs within the platform.

Agenda

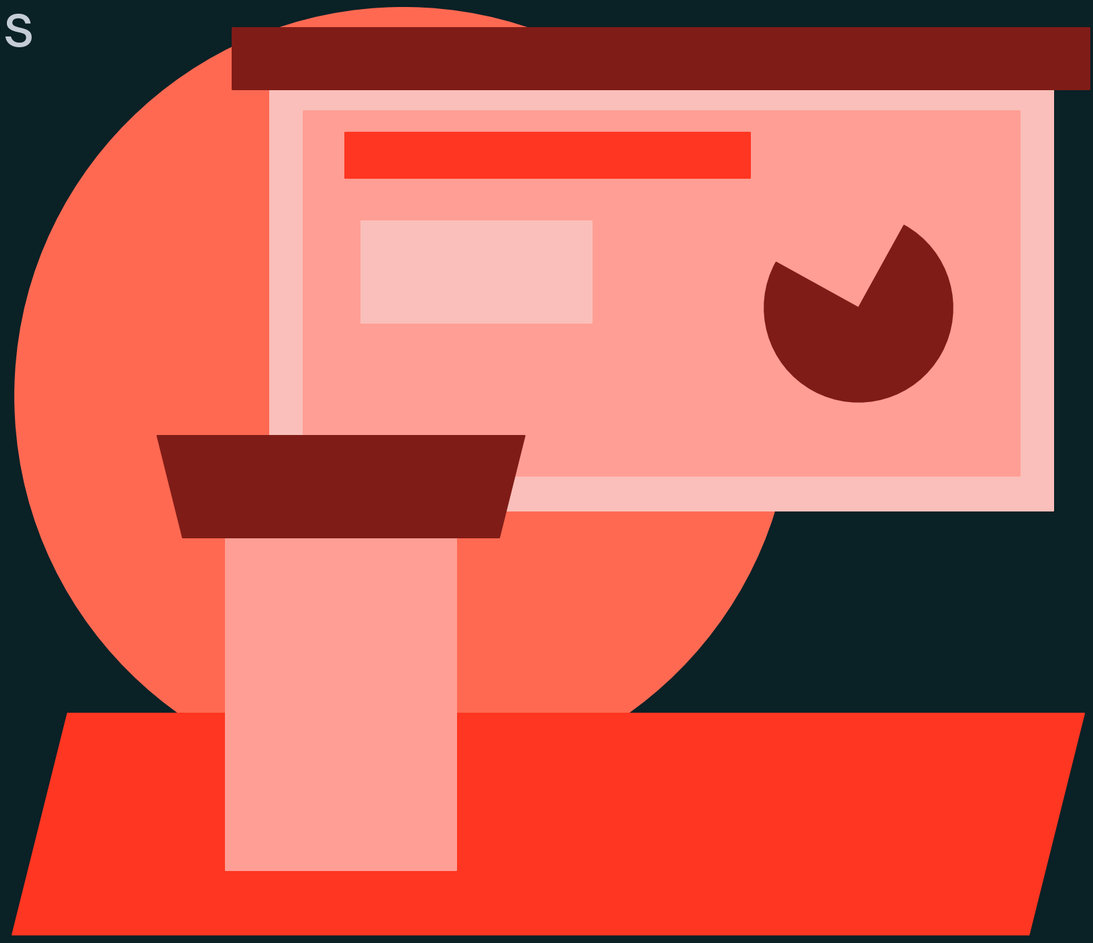
Data Orchestration and Querying Capabilities	Time	Lecture	Demo	Lab
Orchestration in Databricks	5 mins	✓		
Setting Up and Managing Serverless Lakeflow Jobs	15 mins		✓	
Databricks Querying Capabilities	5 mins	✓		

databricks

Data Orchestration and Querying Capabilities

LECTURE

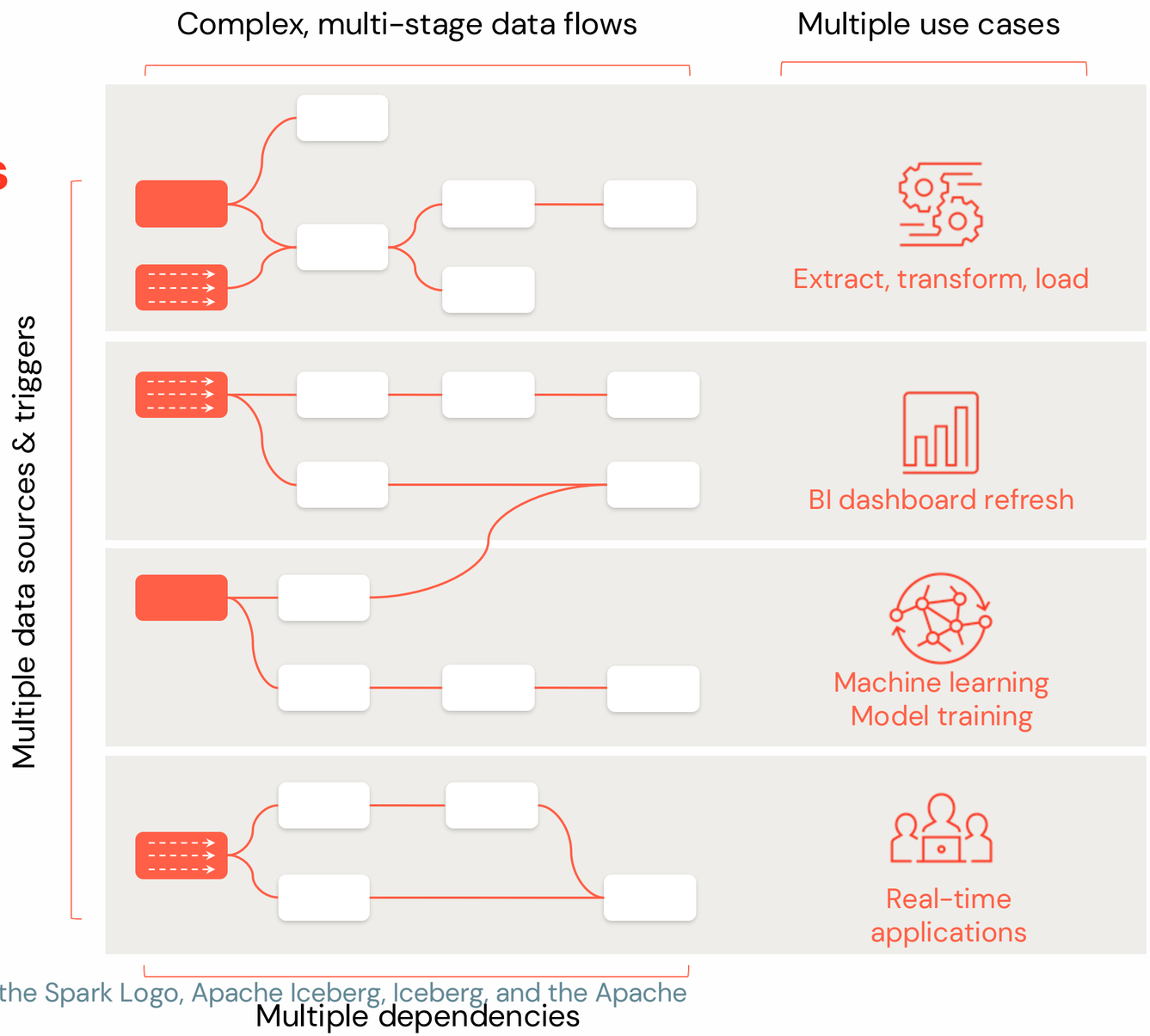
Orchestration in Databricks



Modern Data Warehousing Requires Modern Orchestration

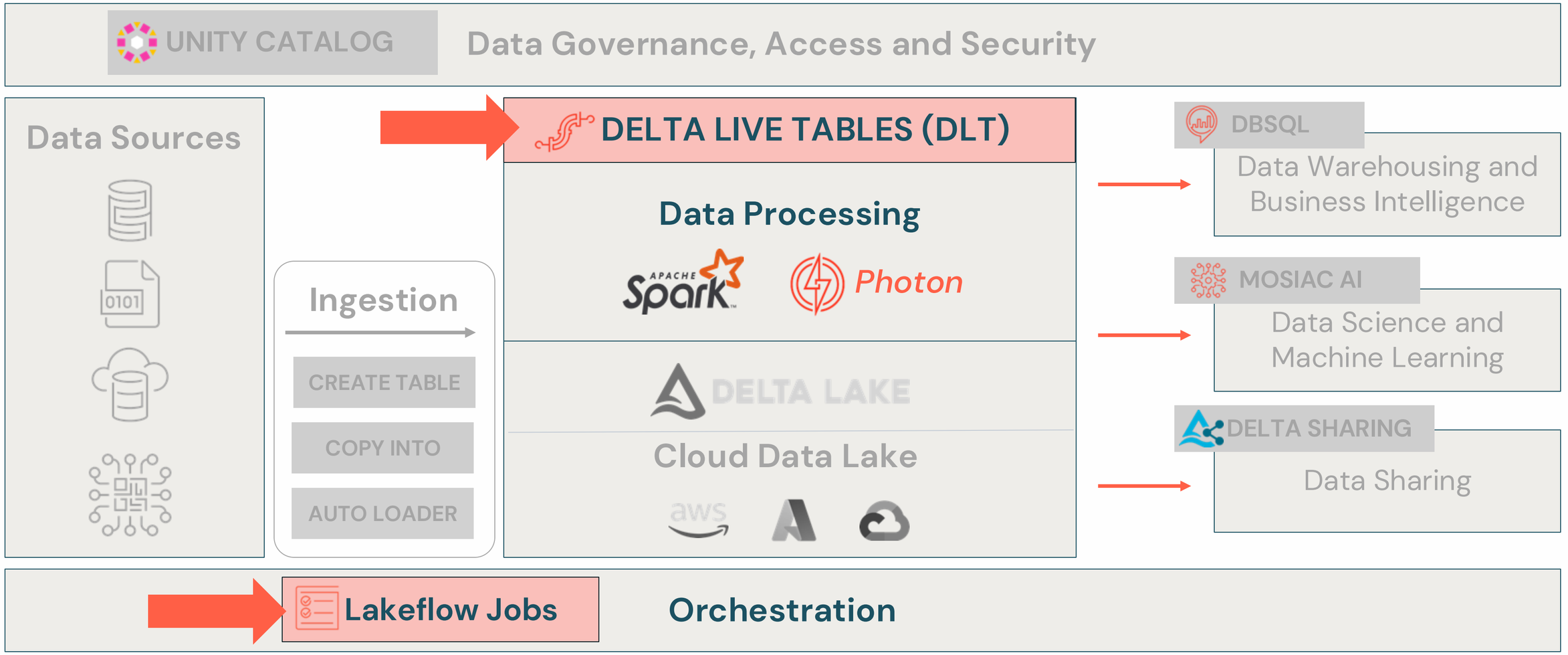
Orchestrating processes across all data, analytics and AI use cases is business critical

- Handles complex, multi-stage data flows.
- Manages multiple dependencies in Lakeflow Jobs.
- Supports diverse use cases, including:
  - Extract, transform, load (ETL).
  - BI dashboard refresh.
  - Machine learning model training.
  - Real-time applications.



Unified Orchestration Using Databricks

Overview



Databricks Lakeflow Jobs

Reliable orchestration for data, analytics, and AI

Orchestrate Anything Anywhere

Simple Lakeflow Jobs Authoring

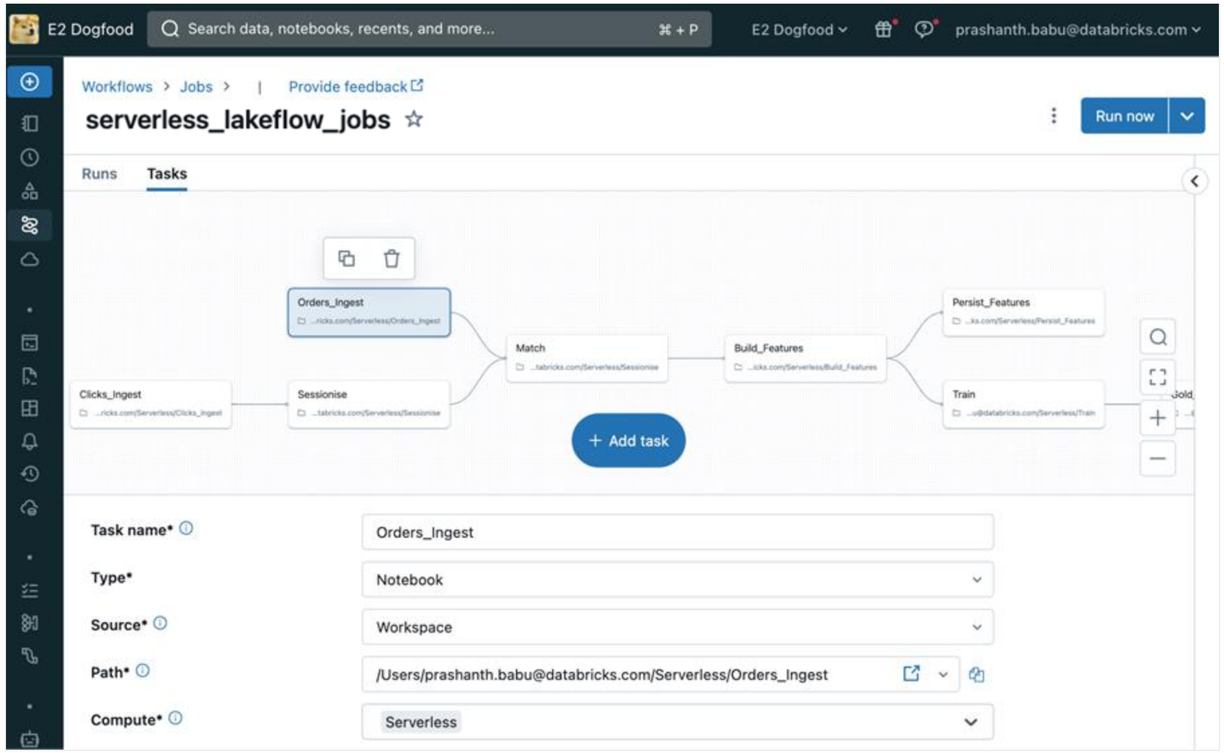
Deep Platform Integration

Proven Reliability

Fully Managed

Serverless Lakeflow Jobs

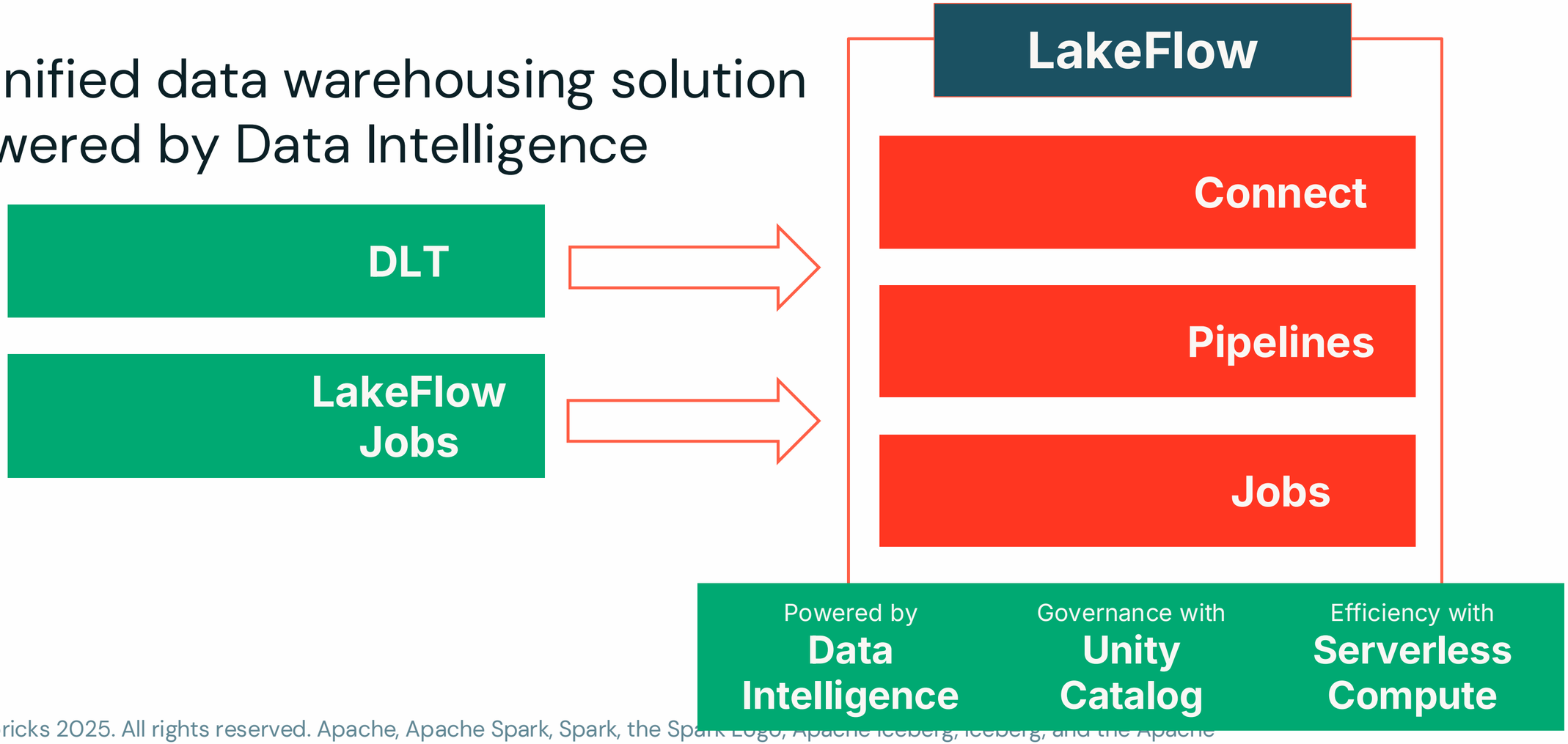
- Fully **Managed Infrastructure**, eliminates the human efforts to handle compute, providing fast analysis & cost savings.
- Advanced **Autoscaler** ensures efficient resource management
- **Elastic scaling** up and down is achieved using a warm pool of machines.



LakeFlow

The evolution of data warehousing on Databricks

A unified data warehousing solution powered by Data Intelligence

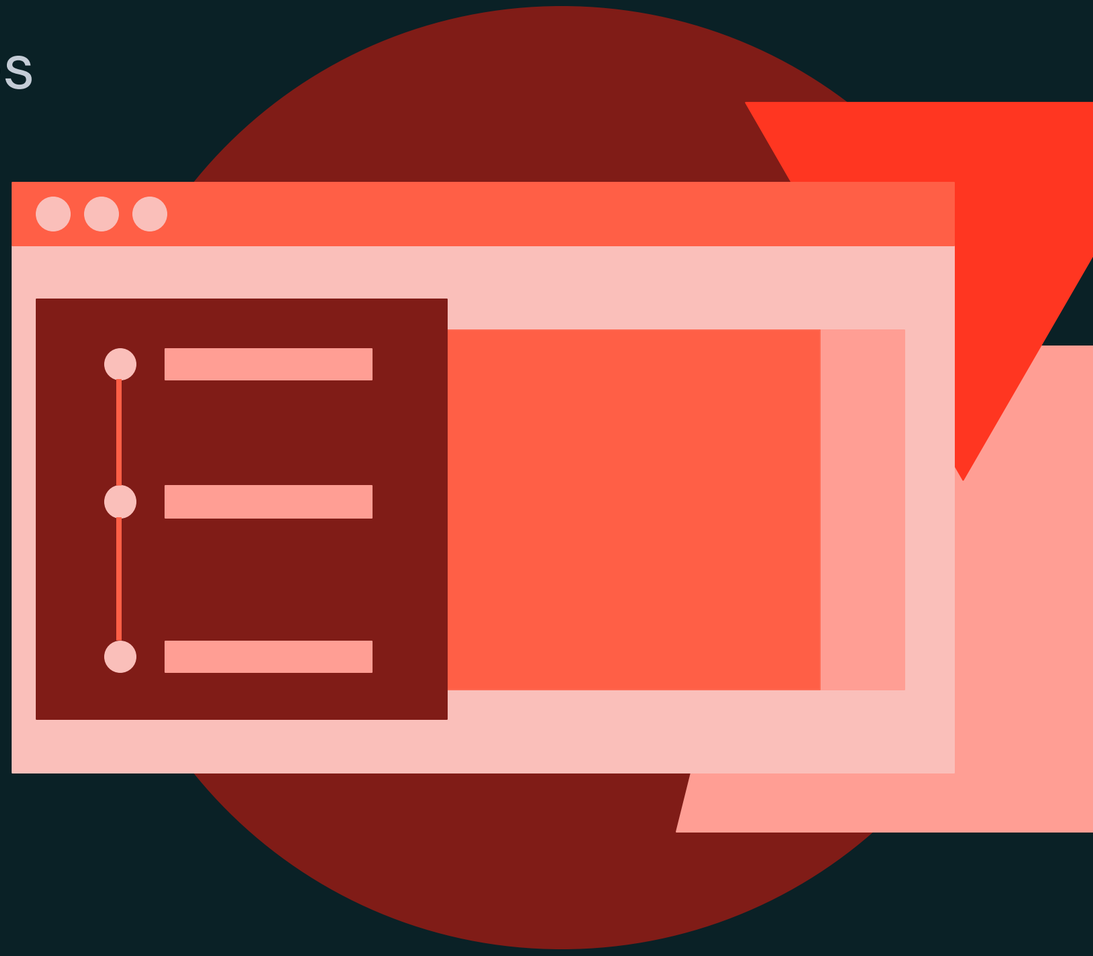


databricks

Data Orchestration and Querying Capabilities

DEMONSTRATION

Setting Up and Managing Serverless Lakeflow Jobs



High Level Steps

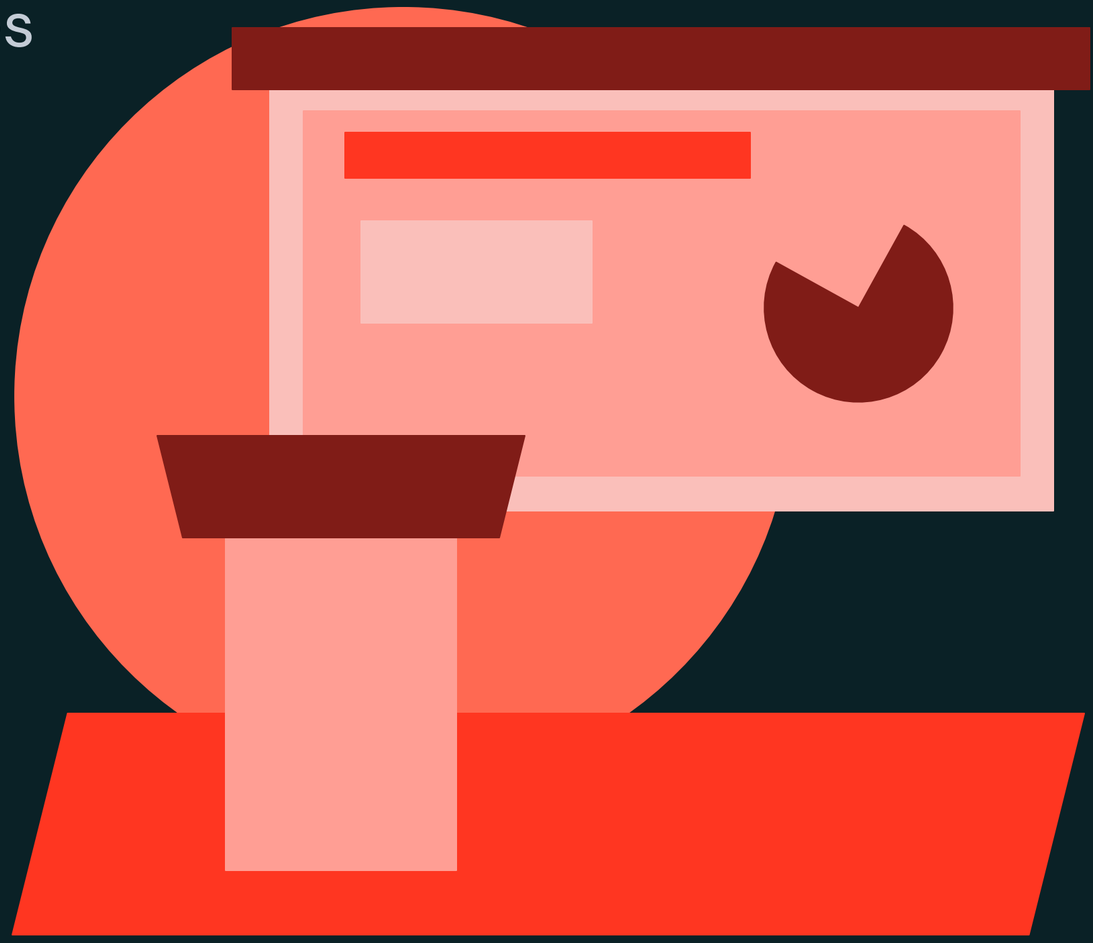
- Create a Lakeflow Job with SQL tasks for a Medallion Architecture pipeline.
- Set up task dependencies and implement conditional logic for lakeflow jobs control.
- Use the Lakeflow Job UI for monitoring and data lineage visualization to trace data transformations and dependencies.
- Configure error handling and retries for tasks.
- Schedule Lakeflow Job using manual triggers.
- Set up notifications for monitoring and analyze the execution history

databricks

Data Orchestration and Querying Capabilities

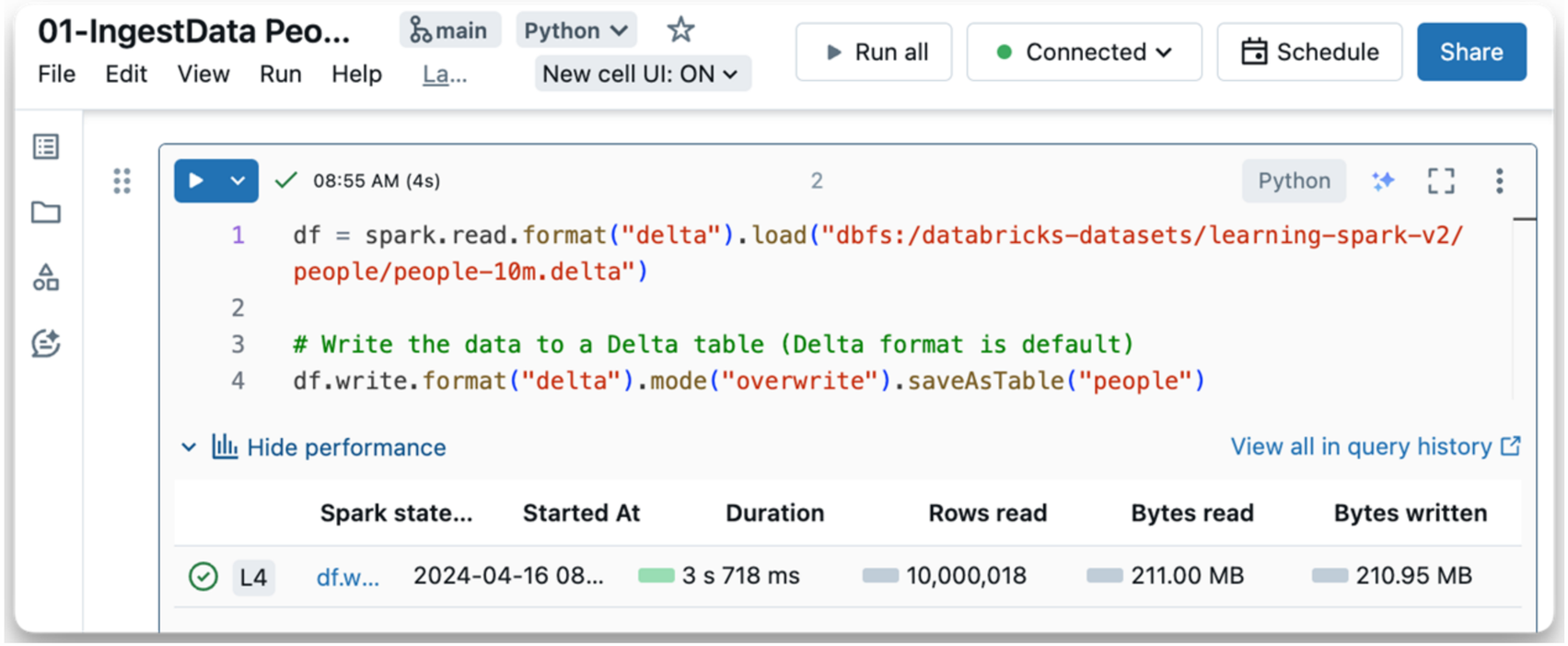
LECTURE

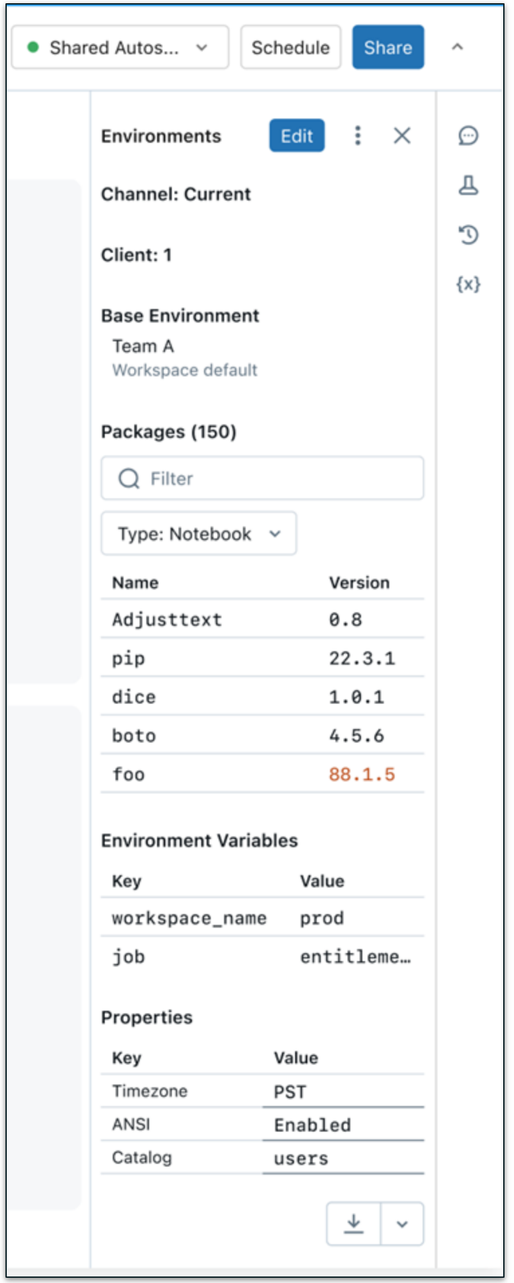
Databricks Querying Capabilities



Serverless Compute for Notebooks

With *Environments* for dependency management





Fast and easy to use

- Starts in seconds

Portable, reusable environments

Automatically cached

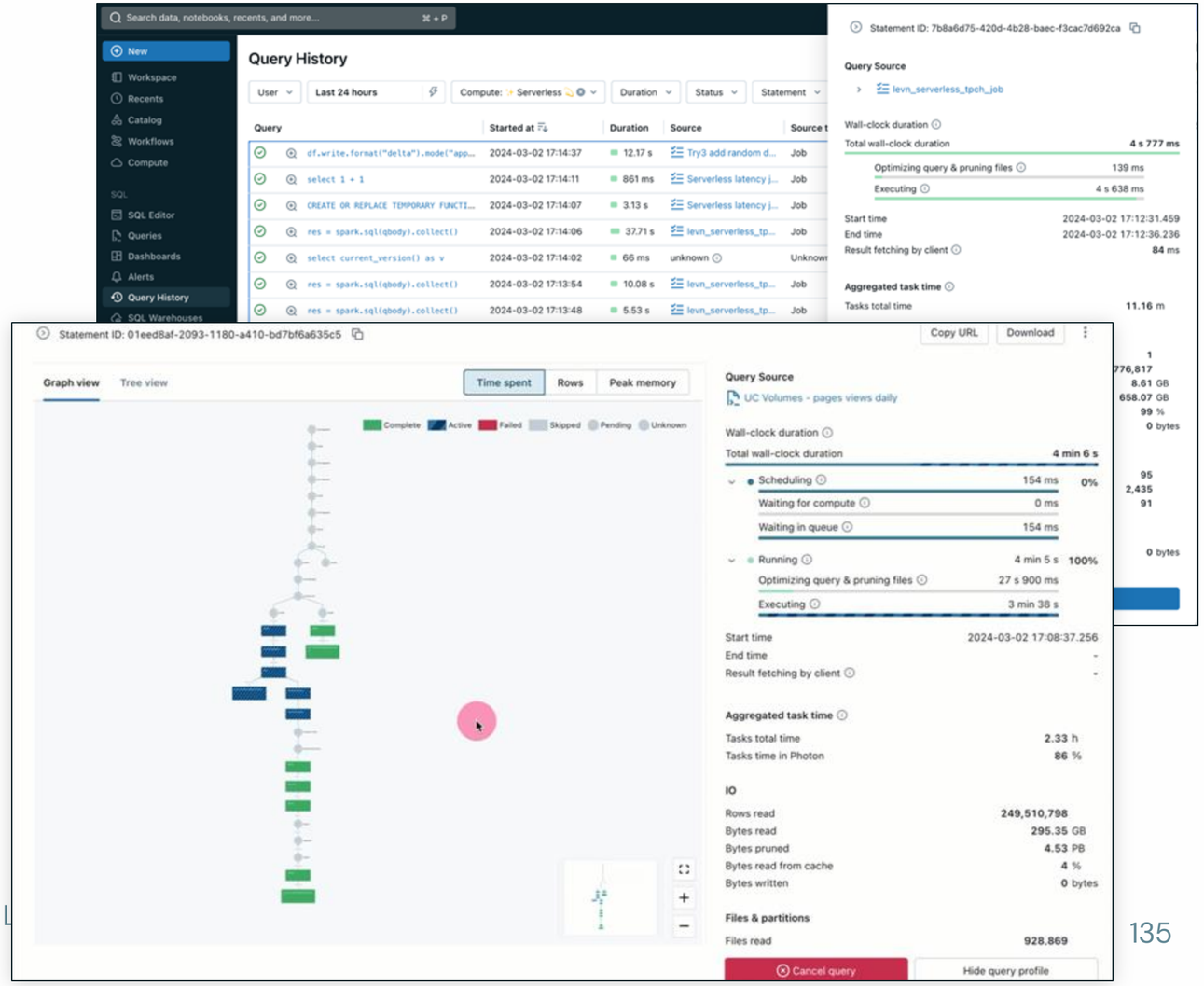
Base Python and package versions

Customizable by users or admins

Query Insights

Extended to SQL Queries & DataFrames on Serverless Compute

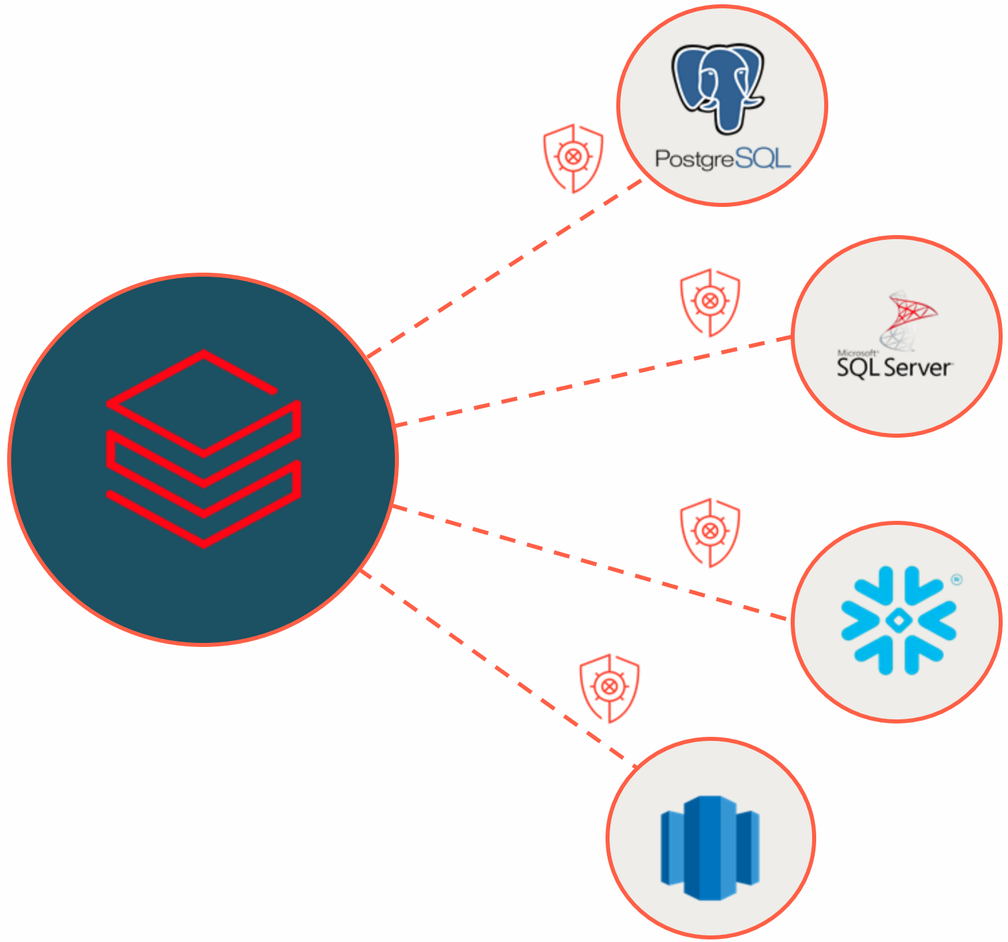
- **Query History**
  - Searchable and filterable overview of all current and past statements + cross-linking to source object that generated the query
- **Query Profile**
  - Deep-dive into execution details to identify performance bottlenecks and optimization opportunities
- **Live Query Profile** – *coming soon*
  - Performance insights updated live as the query executes to enable immediate action



Query Federation

Discover, query, and govern your data wherever it lives

- Connect data sources to Unity Catalog and set policies across them
  - Table, Row, Col, tags, etc.
- Efficient optimization & caching across sources



AI Functions for DBSQL

Access Proprietary and Open Source LLMs directly within DBSQL

- Enable analysts to classify data with LLMs
- Extract actionable insights using AI language models.
- Generate product descriptions, summaries, and more by simply adding the **ai_query** function.

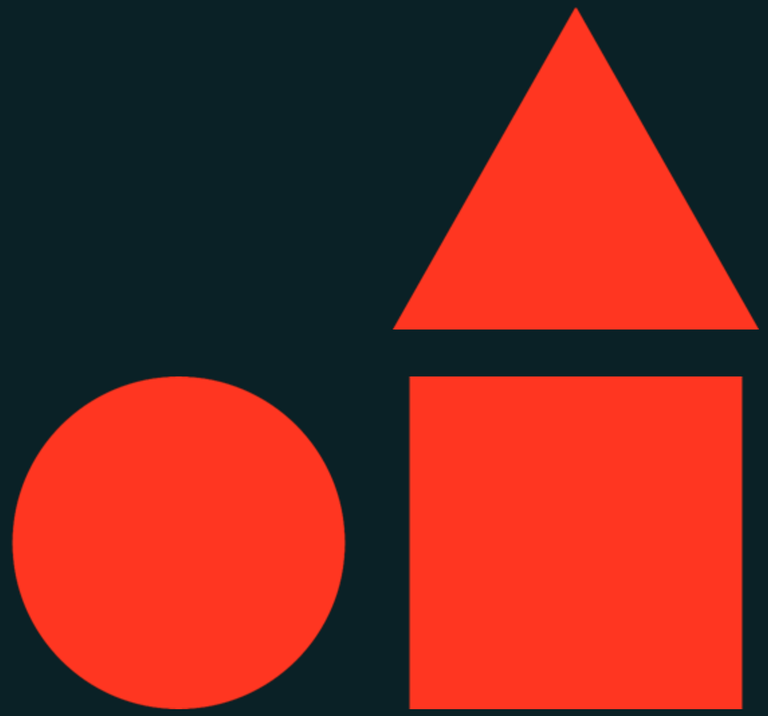
```
SELECT
 sku_id,
 product_name,
 ai_query(
   "my-external-openai-chat",
   "You are a marketing expert for a winter holiday promotion targeting GenZ. Generate a promotional text in 30 words mentioning a 50% discount for product: " || product_name
 )
FROM
 uc_catalog.schema.retail_products
WHERE
 inventory > 2 * forecasted_sales
```

Autocomplete Code or Queries

- Generate code from comments inside a notebook cell or SQL editor

databricks

Data Presentation



Get Started with Databricks for Data Warehousing

Objectives

- Describe the impact of data warehousing work on Databricks AI/BI needs.
- Identify the AI supported features in the Databricks platform.

Agenda

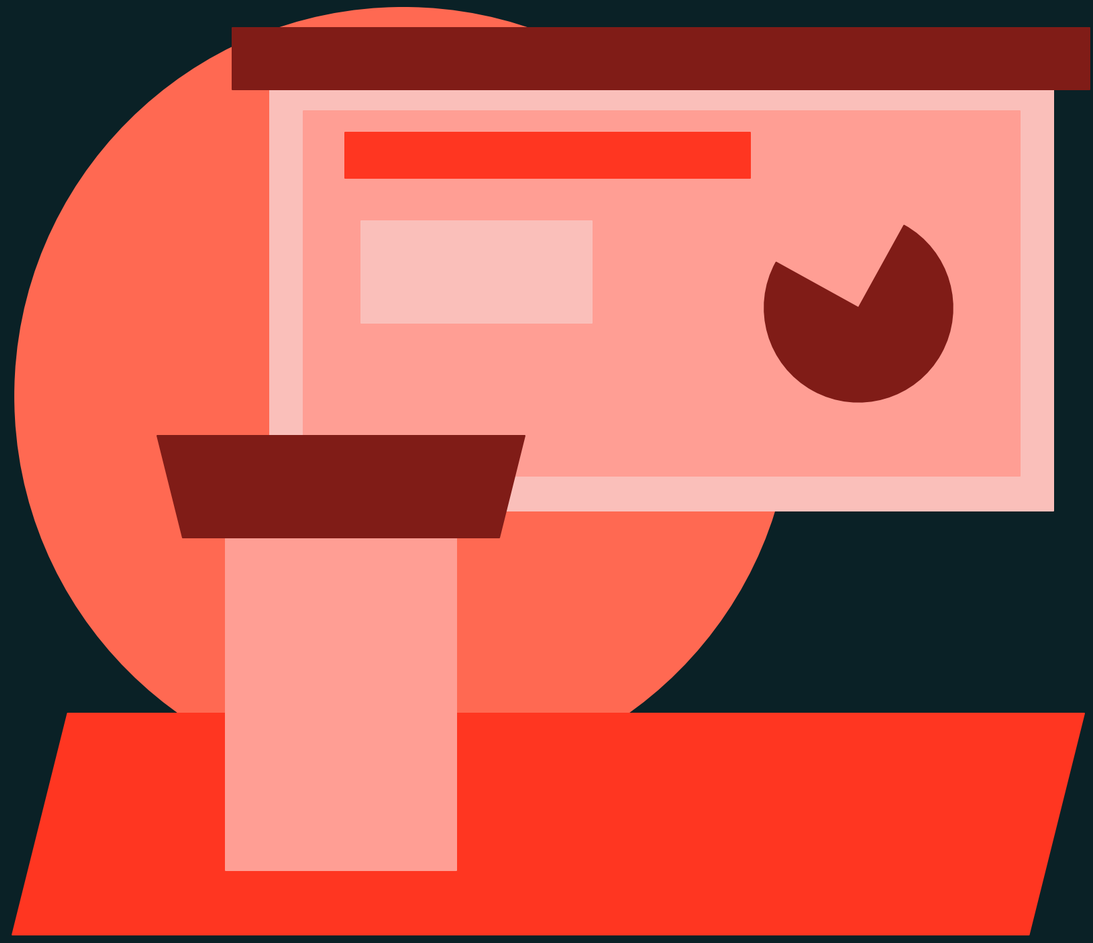
Data Presentation	Time	Lecture	Demo	Lab
Introduction to AI/BI	5 mins	✓		
Creating a Dashboard in Databricks	5 mins		✓	
Creating Genie Spaces	10 mins		✓	
Data Warehousing Comprehensive Lab	30 mins			✓

databricks

Data Presentation

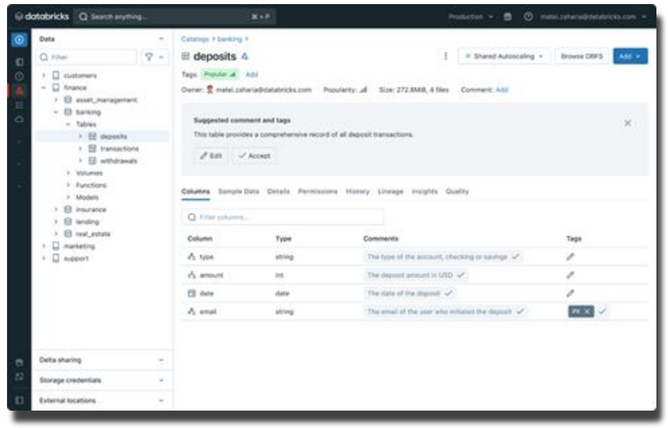
LECTURE

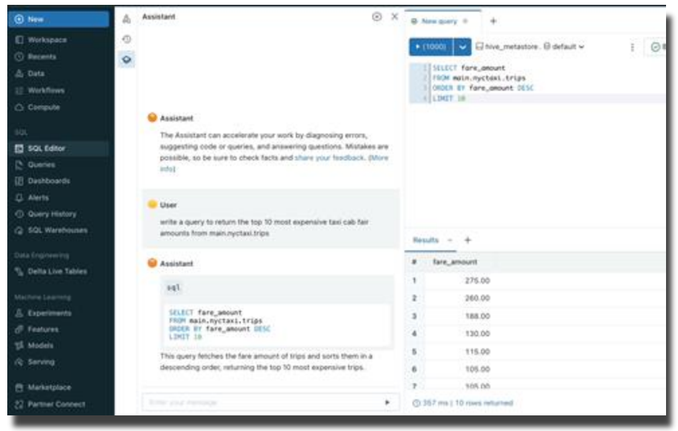
Introduction to AI/BI



Infusing AI in our user experience

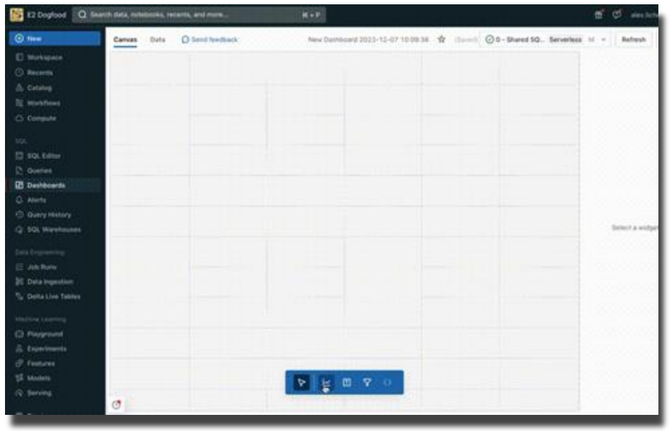
AI-generated docs for easier search and discoverability

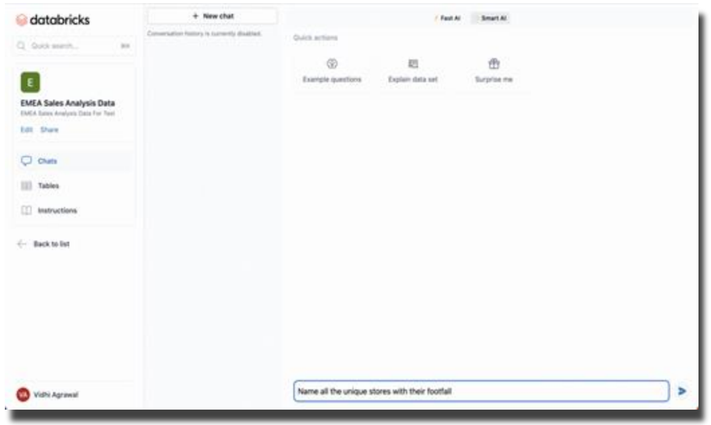




AI Assistant for quick code fixes, queries and descriptions

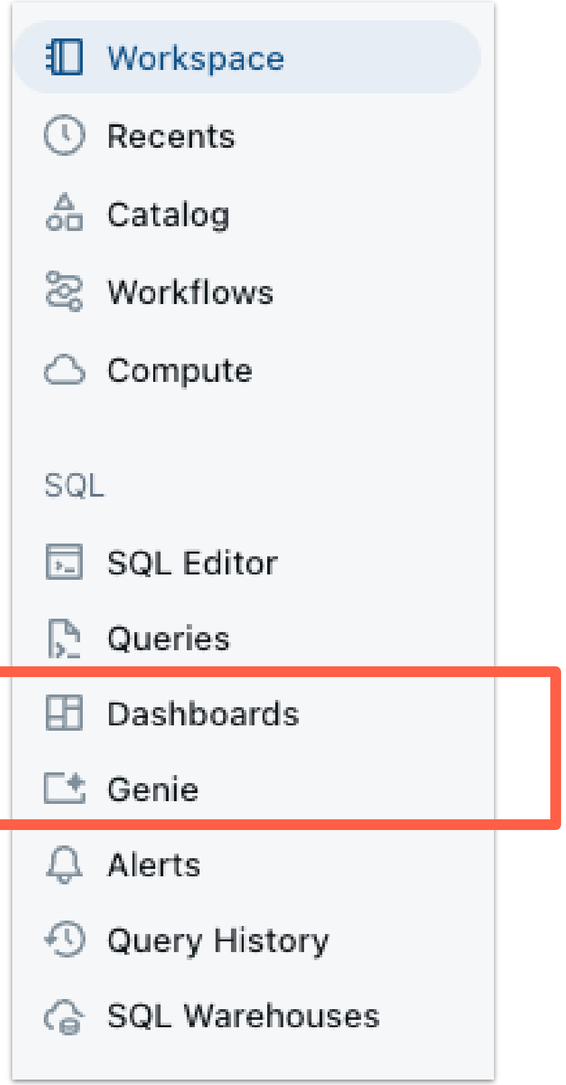
AI Assistant for quick creation of visualizations





AI-powered Genie Spaces to quickly understand your data using natural language

Databrick's AI/BI Offerings



Located under the SQL header in the navigation:

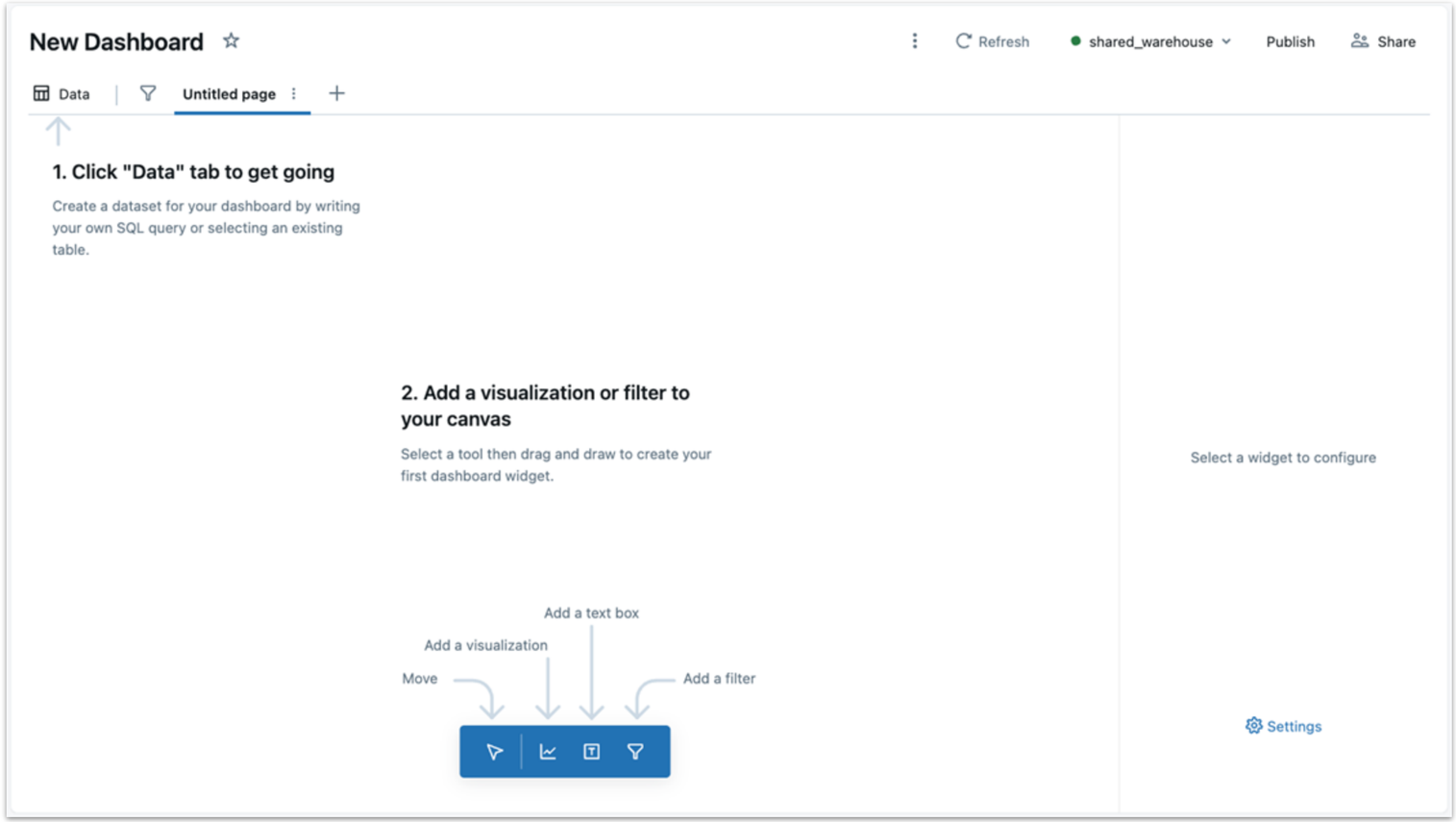
AI/BI Dashboards

- Presentation environment for visualizations, titles, filters, and other features traditionally found in presentation software.

AI/BI Genie

- Unique space associated to a specific curated data set for asking questions using natural language to gain further insight or alternative visualizations without needing a data analyst to oversee the development.

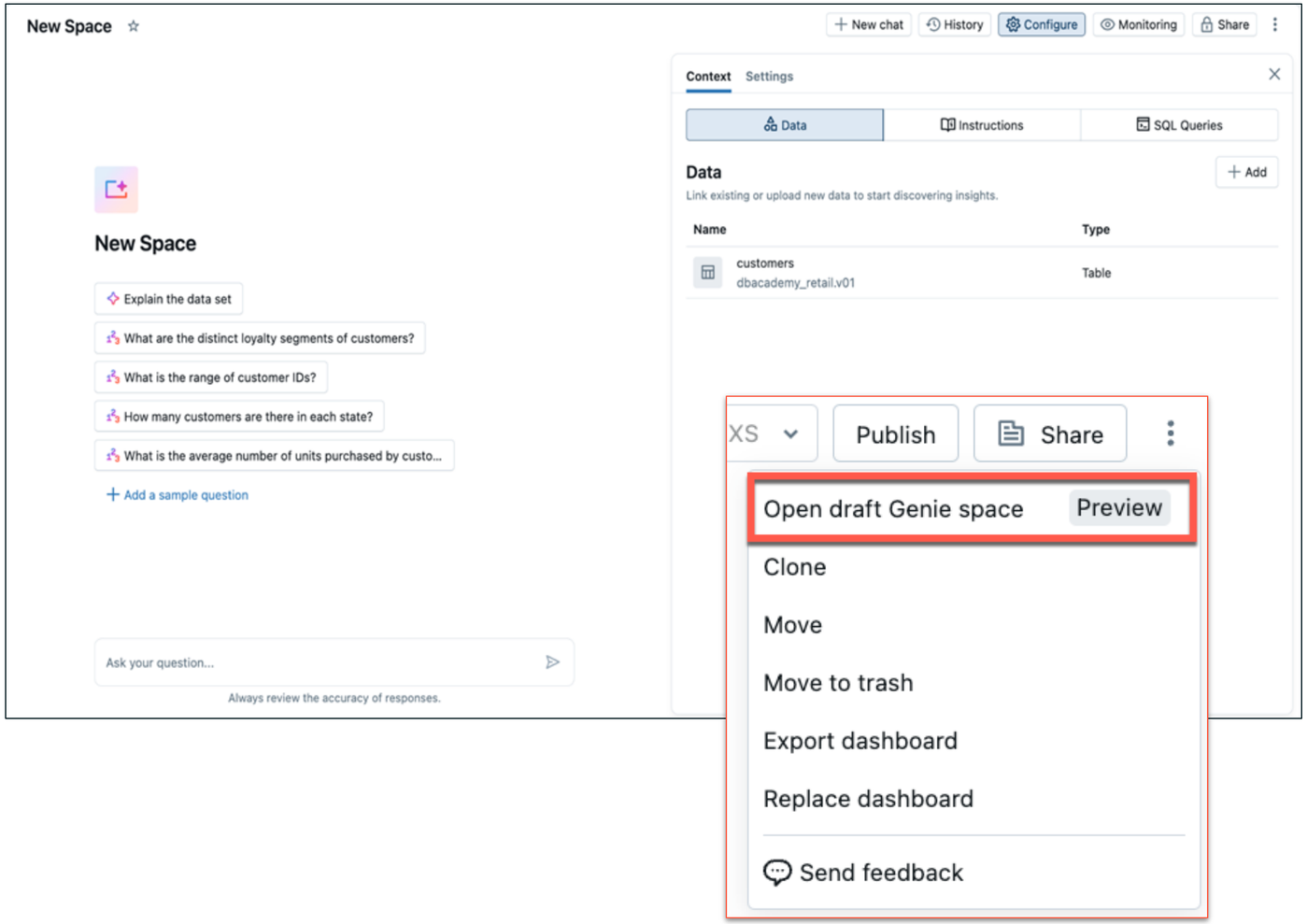
AI/BI Dashboards



- Support for technical and non-technical users.
- Simple, easy to user interface.
- All-in-one sharing manages the permissions for the data, queries, and dashboard all from one interface.

AI/BI Genie

- Easily created straight from a Dashboard.
- Simple UI for creation outside of Dashboards.
- Supports follow-up discussions directly with the data.



AI/BI Genie Spaces

Ask questions of your data in natural language using your terms

- Supports non technical users with a no-code solution to gaining data insights using natural language.

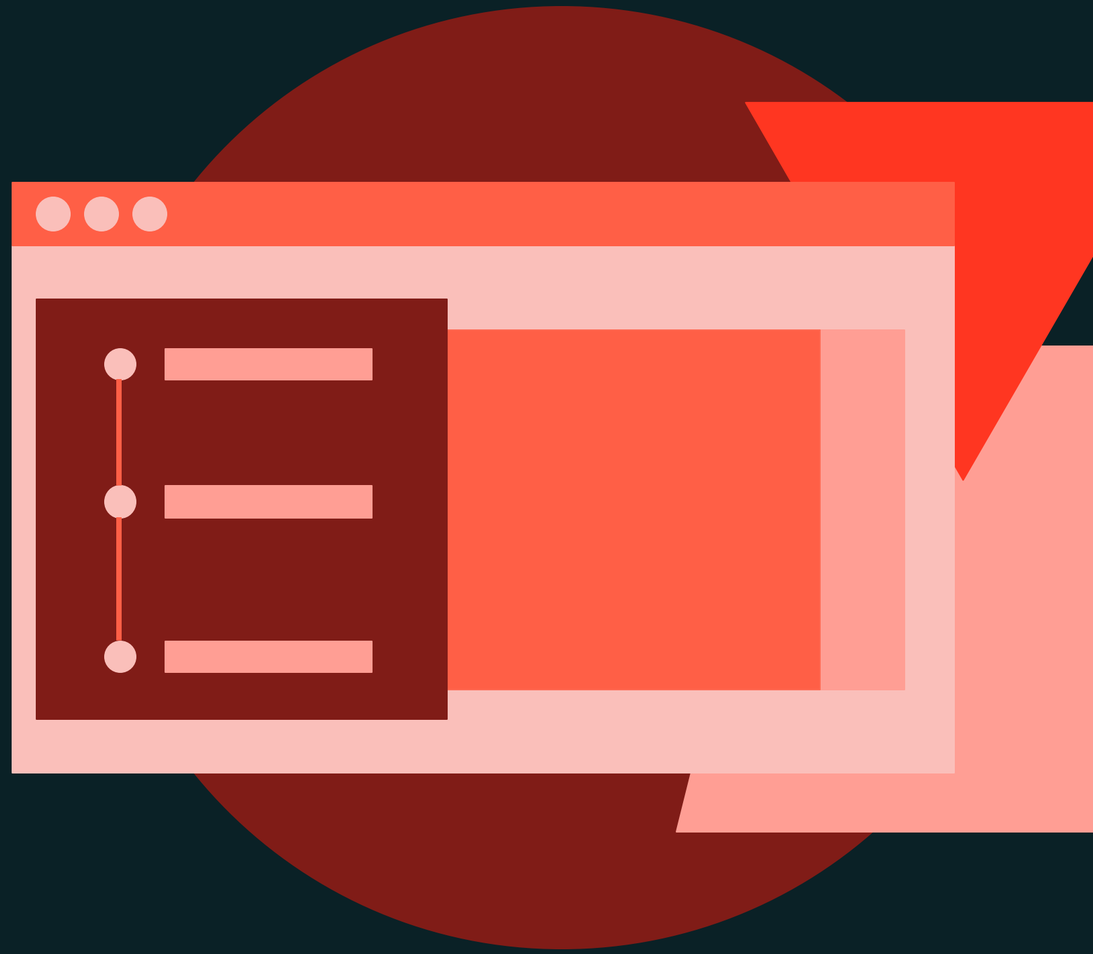
- Provides a self-service access point to the data for follow-up questions without needing to involve a data analytics specialist.

databricks

Data Presentation

DEMONSTRATION

Creating a Dashboard in Databricks



Follow-along Instructions

Estimated Time: 5 minutes

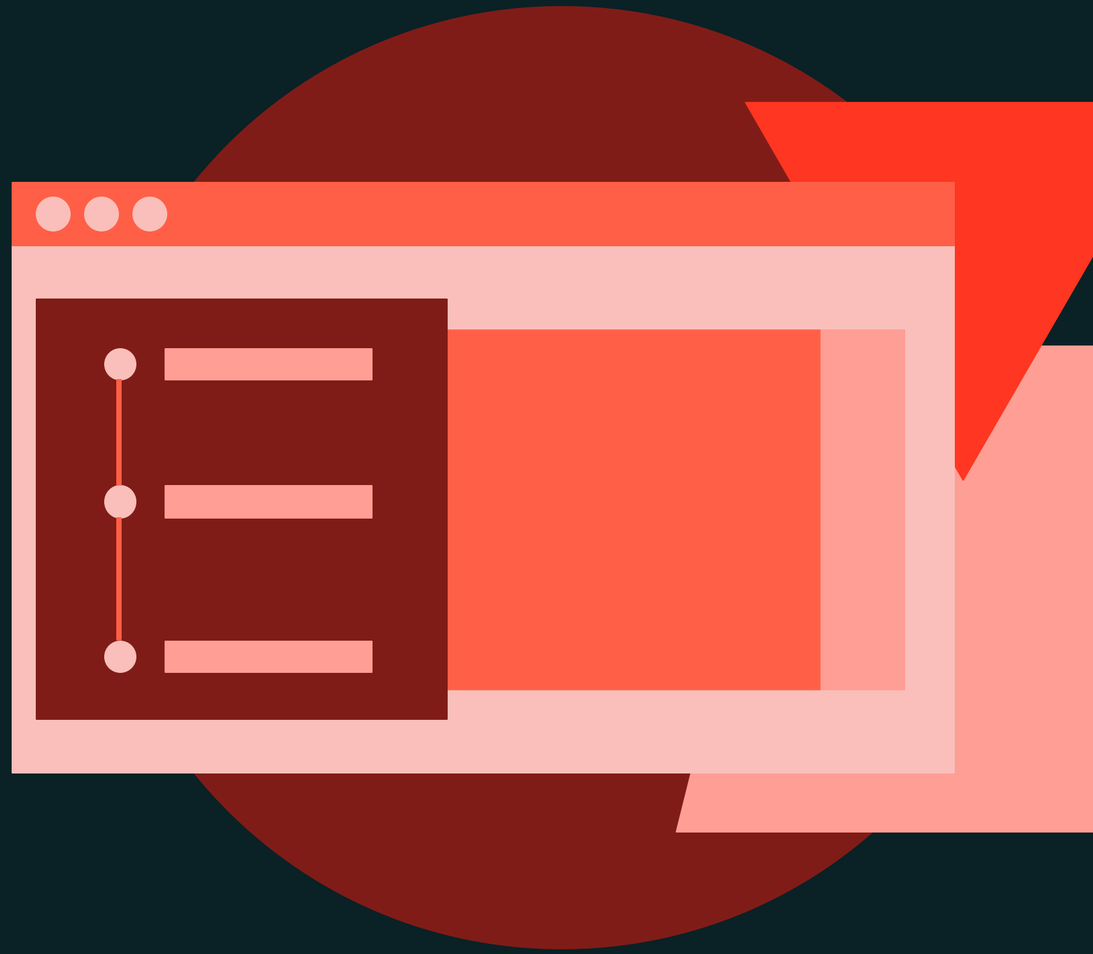
For this demonstration, the instructor will guide you through the Dashboards interface, including adding data sets both with the UI and through a query, creating visualizations, and adding features like filters to the dashboard. This demo closes with a brief introduction to sharing your dashboards with others in your business. If you have access to the Vocareum lab environment, feel free to follow-along to explore the workspace and get oriented with where future labs will be completed.

databricks

Data Presentation

DEMONSTRATION

Creating AI/BI Genie Spaces



Demonstration Details

Estimated Time: 5 minutes

For this demonstration, the instructor will walk you through creating Genie Spaces both from a dashboard but also from scratch and discuss the capabilities of this tool.

High Level Steps

- Understand Databricks AI/BI Genie Spaces:
- Create Genie Spaces:
- Leverage Genie Space Features:
- Create Genie Spaces from Dashboards:
- Understand Genie’s Monitoring and Preview Features:

databricks

Get Started with Databricks for Data Warehousing

LAB EXERCISE

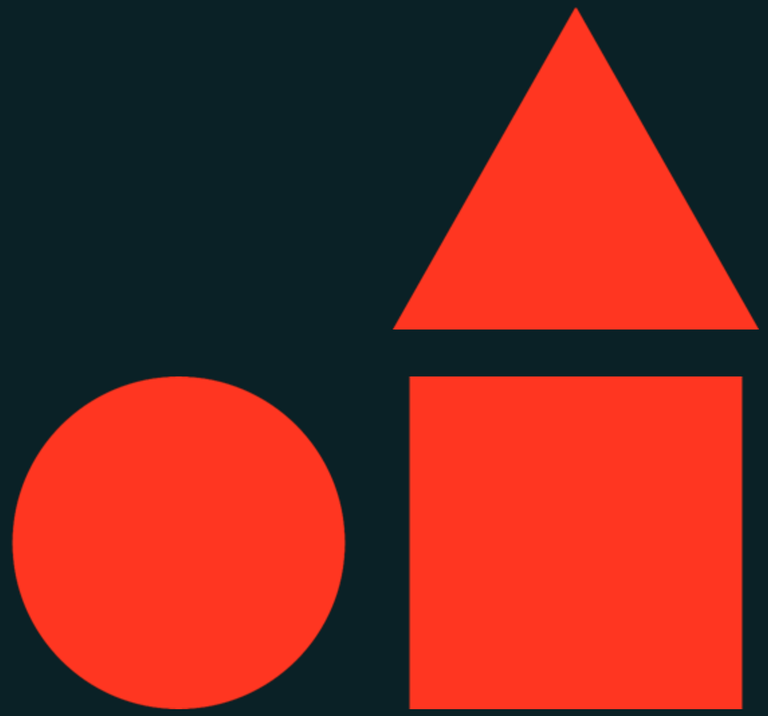
Data Warehousing Comprehensive Lab

Lab Exercise Goals

- What the learner is to accomplish within the lab environment as part of completing the lab.
- Explain how it relates to the lecture or demonstration preceding the exercise.
- Announce whether the lab will be completed with the instructor or if it’s an independent exercise to be completed in the following X minutes.

databricks

Summary and Next Steps



Get Started with Databricks for Data Warehousing

Course Learning Objective Recap

- Describe the available compute options for workloads performed on the Databricks Data Intelligence Platform.
- List the products and features Databricks offers for different data-centric needs within the Databricks Platform.
- Navigate the Databricks Workspace UI.
- Use Databricks to complete common data warehousing tasks.
- Explain the purpose of Delta Lake for data warehousing purposes.
- Describe and apply various techniques for ingesting and transforming data in Databricks.
- Explain how Databricks support data orchestration needs within the platform.
- Describe the impact of data warehousing work on Databricks AI/BI needs.
- Identify the AI supported features in the Databricks platform.