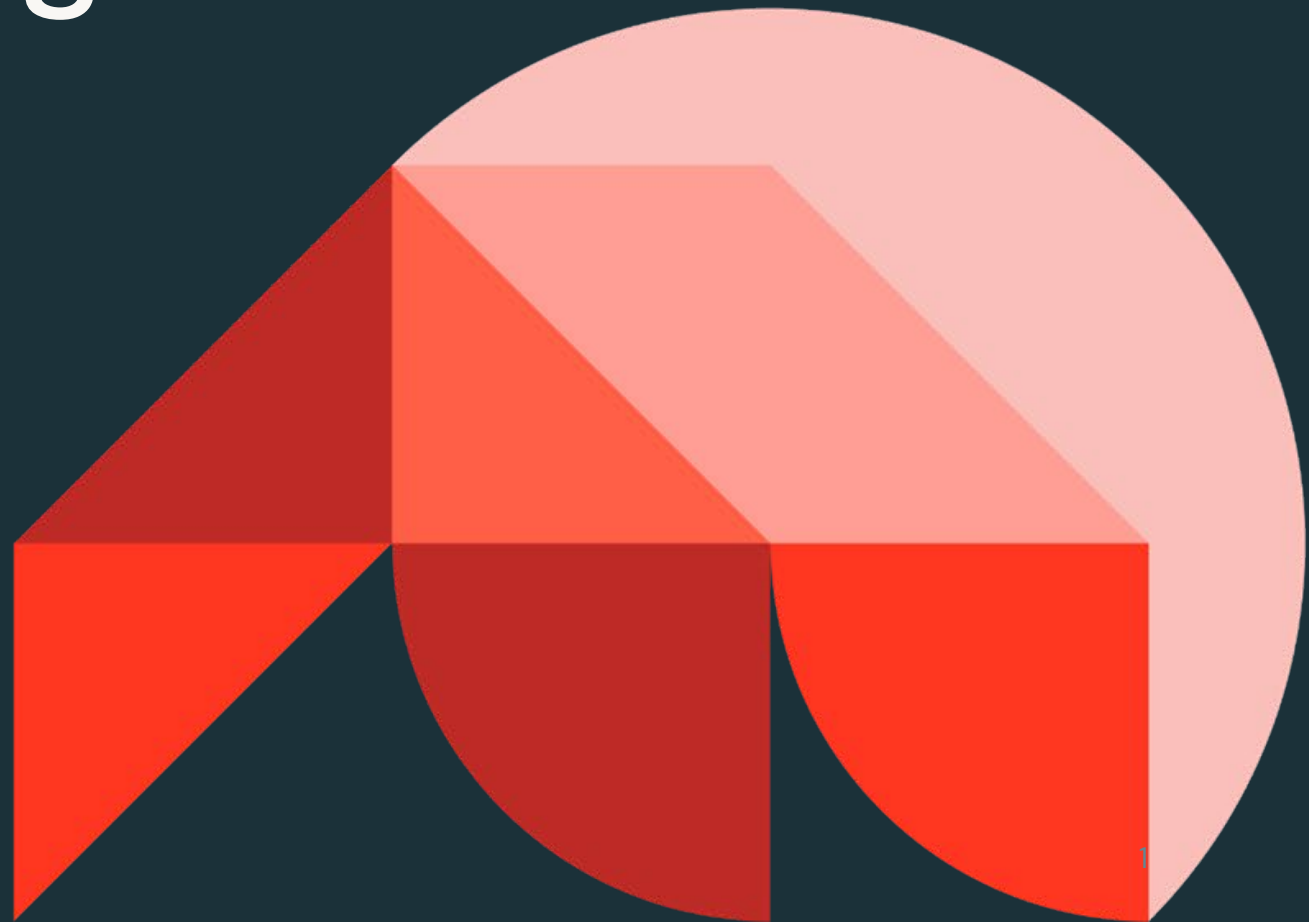
databricks

Your Guide to Data Engineering on the Data Intelligence Platform



Matt Jones, Sr. Product Marketing Manager, Databricks

Raju Mudunuri, Data Solutions & Delivery Sr. Leader, Lexmark

Agenda

I. Strategic Vision
II. Ingestion, Transformation, Orchestration
III. Streaming and Real-Time Processing
IV. Delta Lake, Unity Catalog, and Serverless
V. Embedded GenAI and Data Intelligence
VI. Lexmark's DE Journey with Databricks

Matt Jones

/ Senior Product Marketing Manager
Databricks

Raju Mudunuri

/ Data Solutions & Delivery Senior Leader
Lexmark

Product safe harbor statement

This information is provided to outline Databricks' general product direction and is for **informational purposes only**. Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks discretion and may not be delivered as planned or at all

AI initiatives are top of mind...

By 2026, **over 80%** of enterprises will be using GenAI in production environments, up from **less than 5%** in 2023

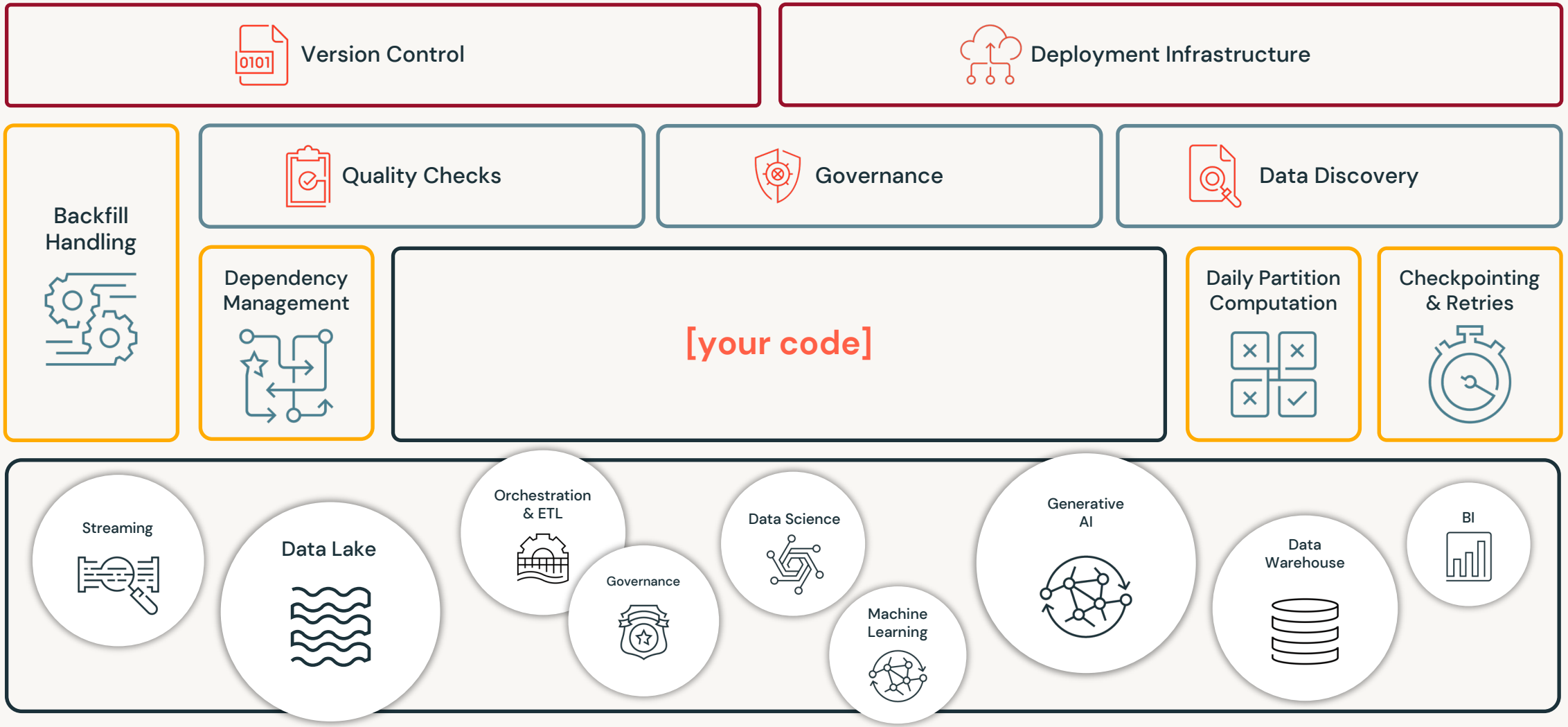
—2023 Gartner Hype Cycle for Generative AI

...but they will fail without reliable data

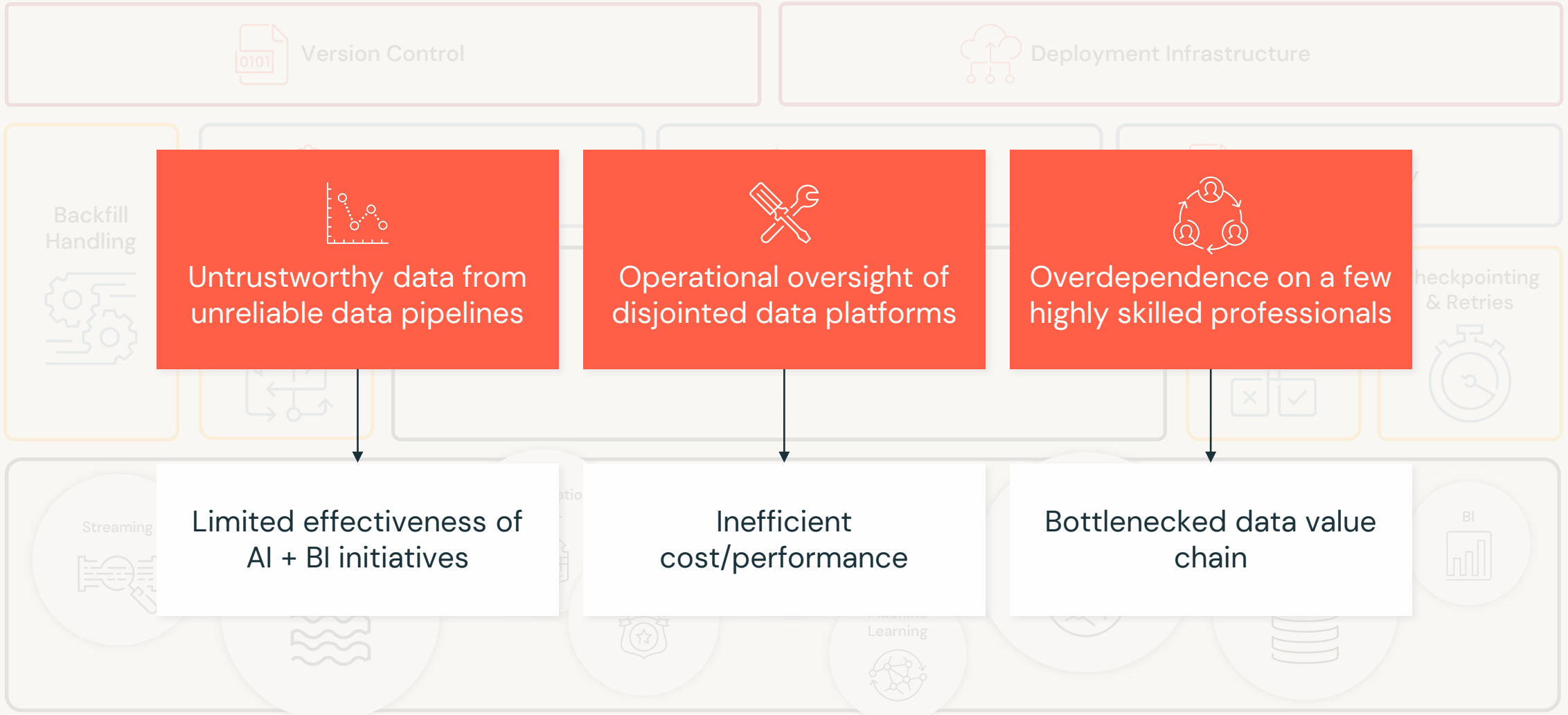
"Data problems are **the most likely factor** to jeopardize our AI/ML goals"

—MIT Technology Review Insights survey, 2023

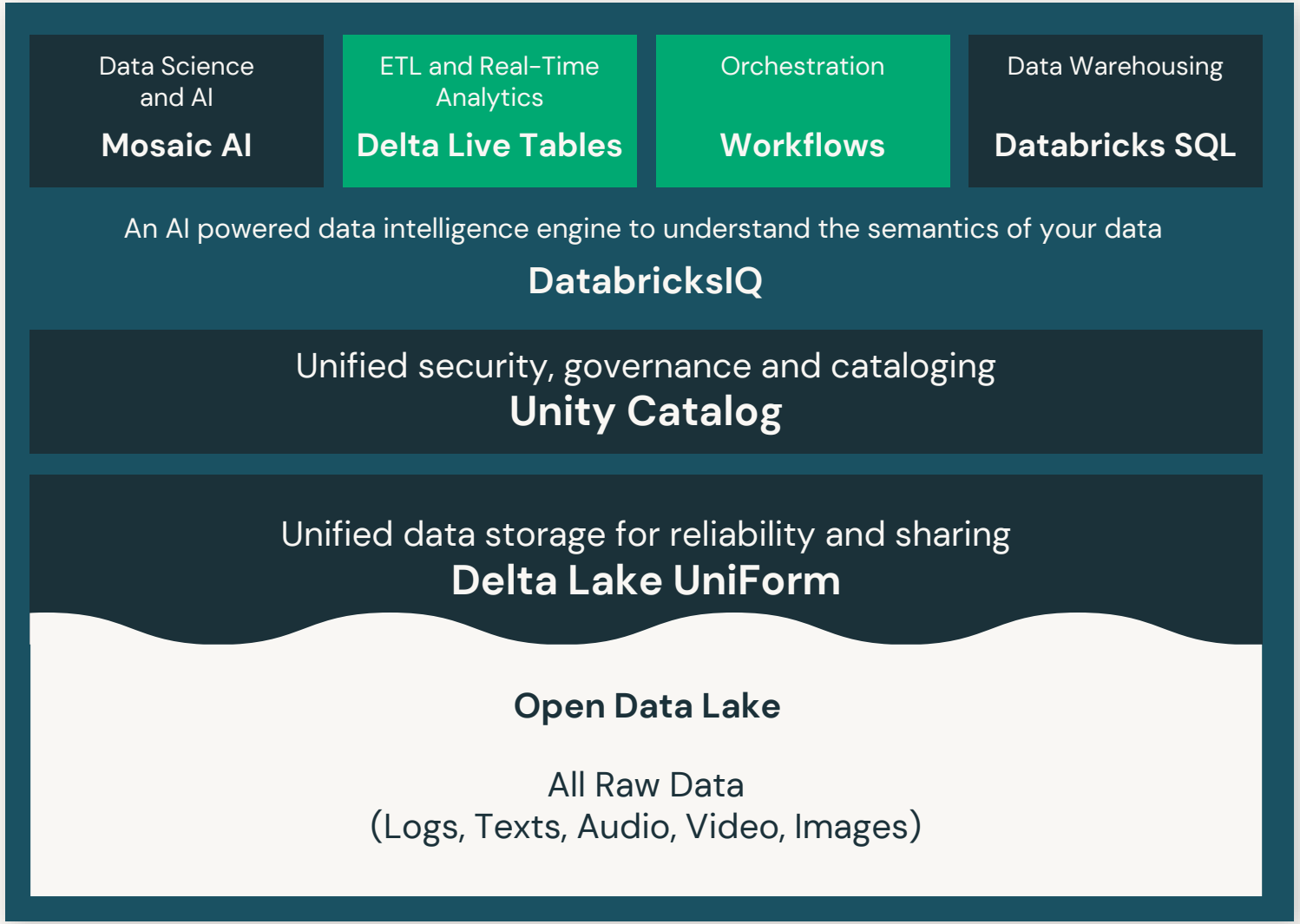
It’s difficult to build and operate data pipelines



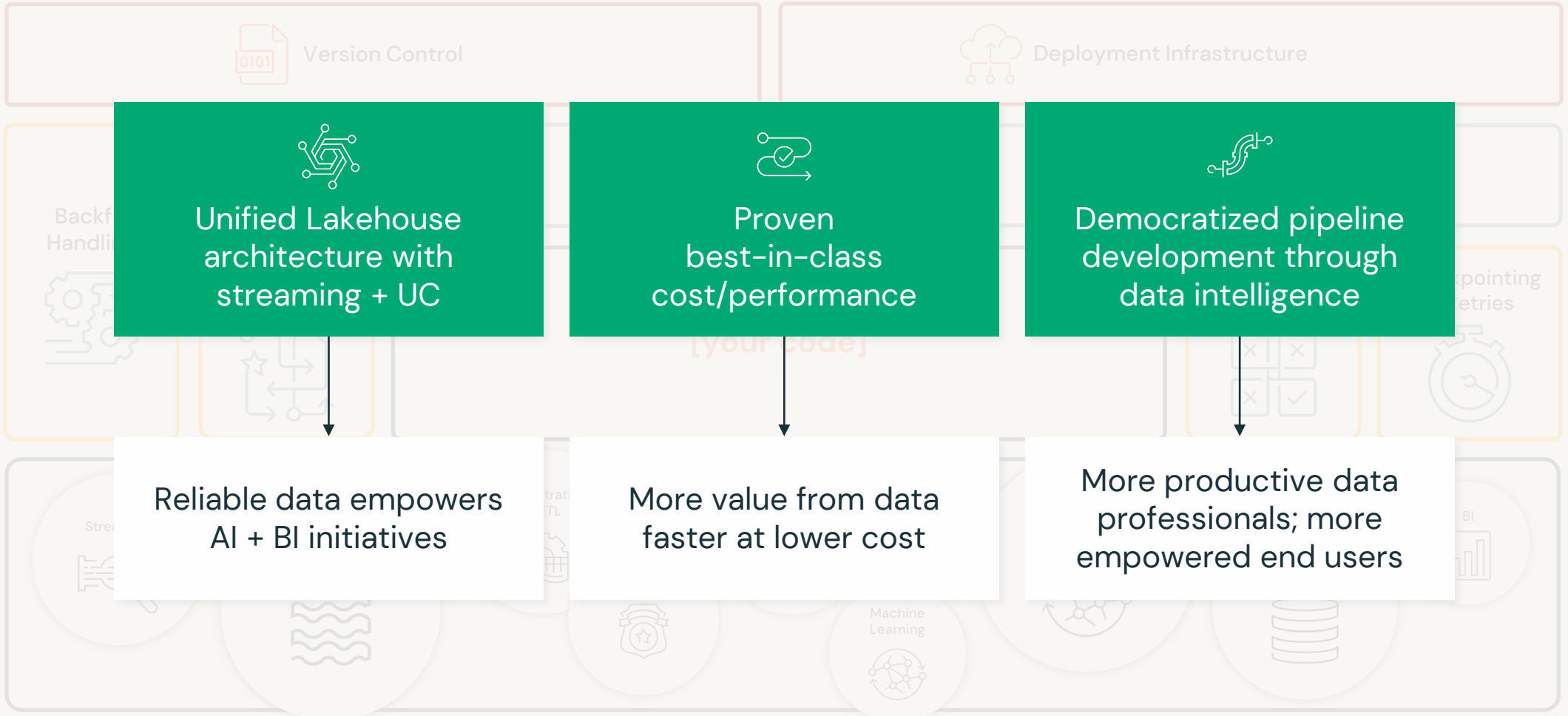
It’s difficult to build and operate data pipelines



Reliable data pipelines require a unified platform with data intelligence



The Databricks Data Intelligence Platform is the best place to do Data Engineering



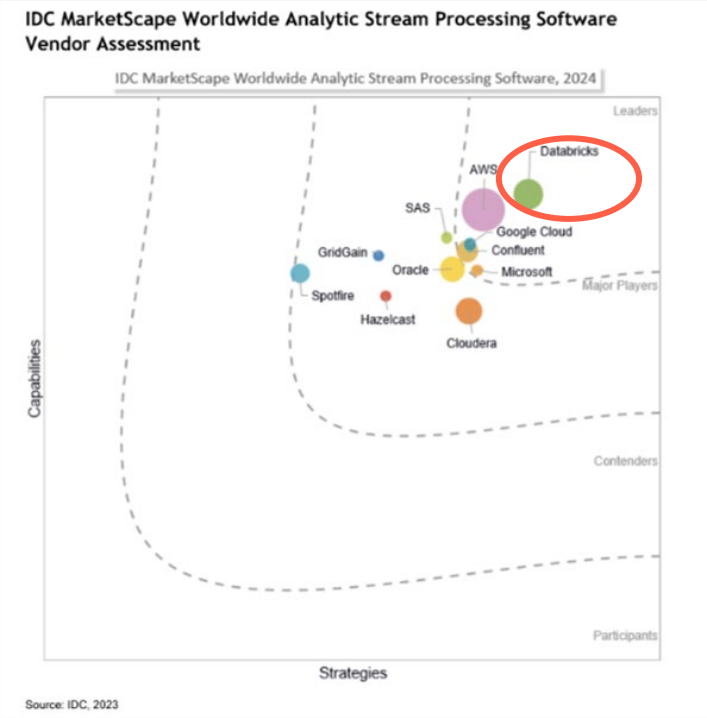
Trusted by organizations of all types

6,500+ Data Engineering customers across industries

A Recognized Leader in Streaming and Data Pipelines

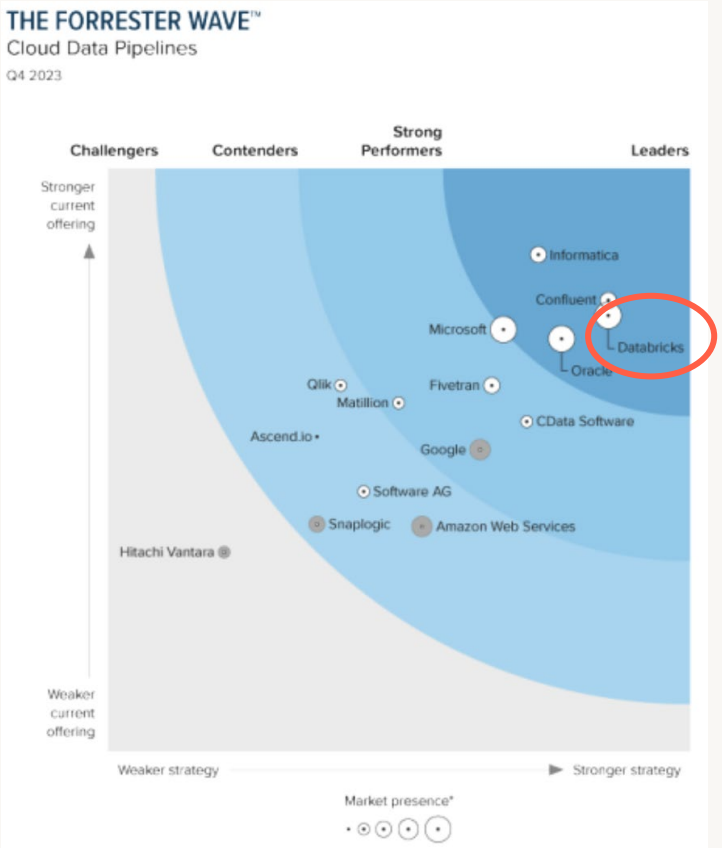
IDC MarketScape Leader

Analytic Stream Processing

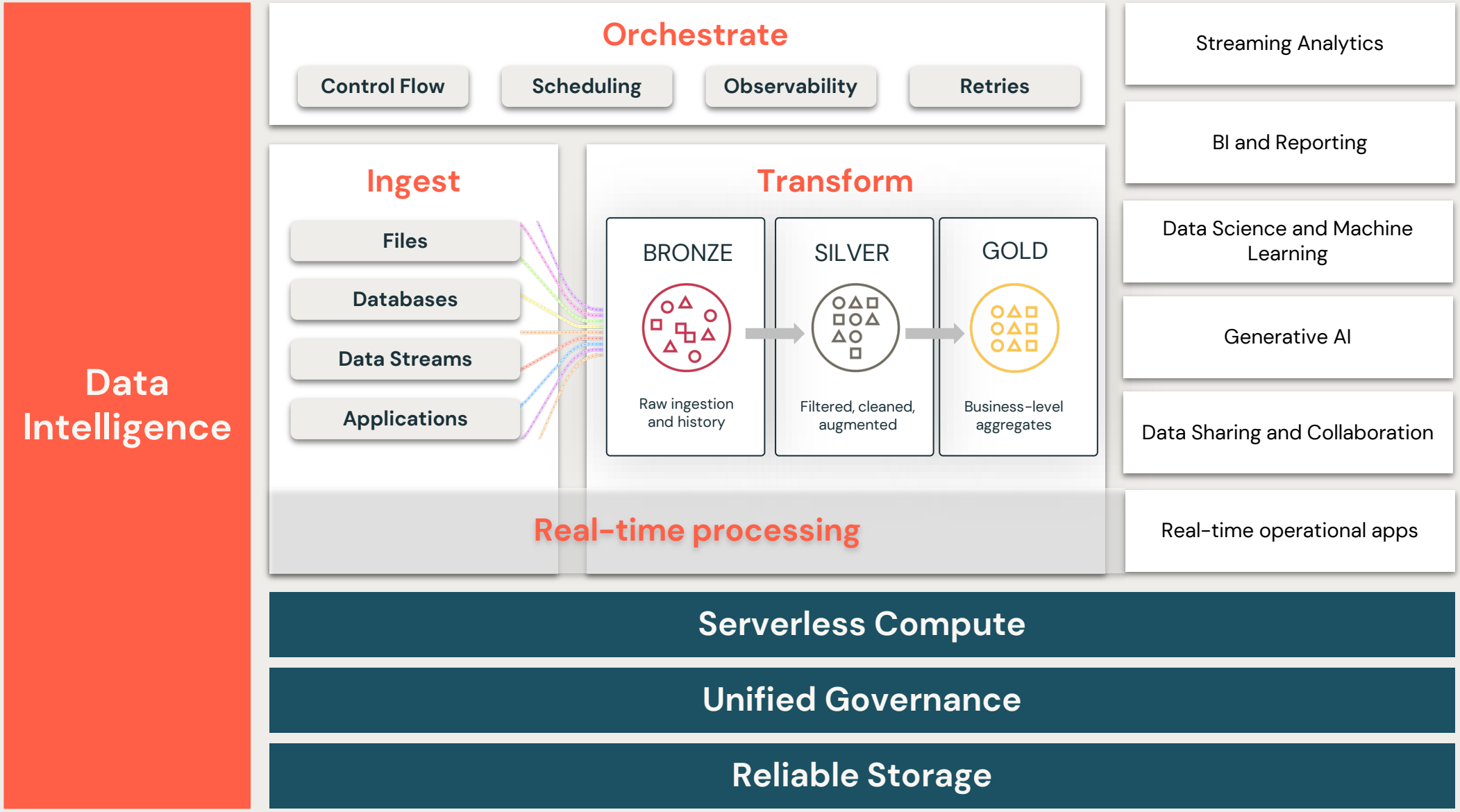


Forrester Wave Leader

Cloud Data Pipelines



databricks Data Intelligence Platform



databricks Data Intelligence Platform

Data Intelligence

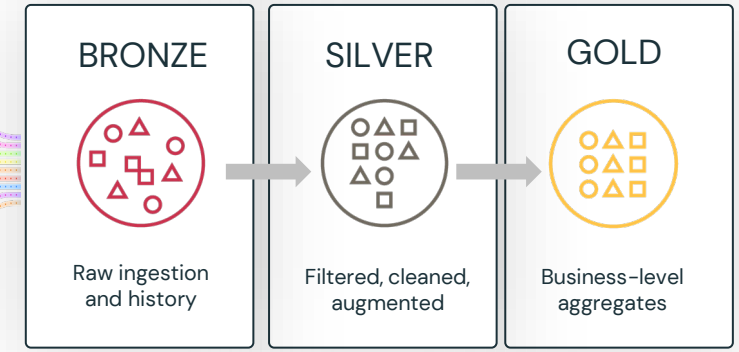
Orchestrate

Control Flow | Scheduling | Observability | Retries

Ingest

- Files
- Databases
- Data Streams
- Applications

Transform



Real-time processing

Streaming Analytics

BI and Reporting

Data Science and Machine Learning

Generative AI

Data Sharing and Collaboration

Real-time operational apps

Serverless Compute

Unified Governance

Reliable Storage

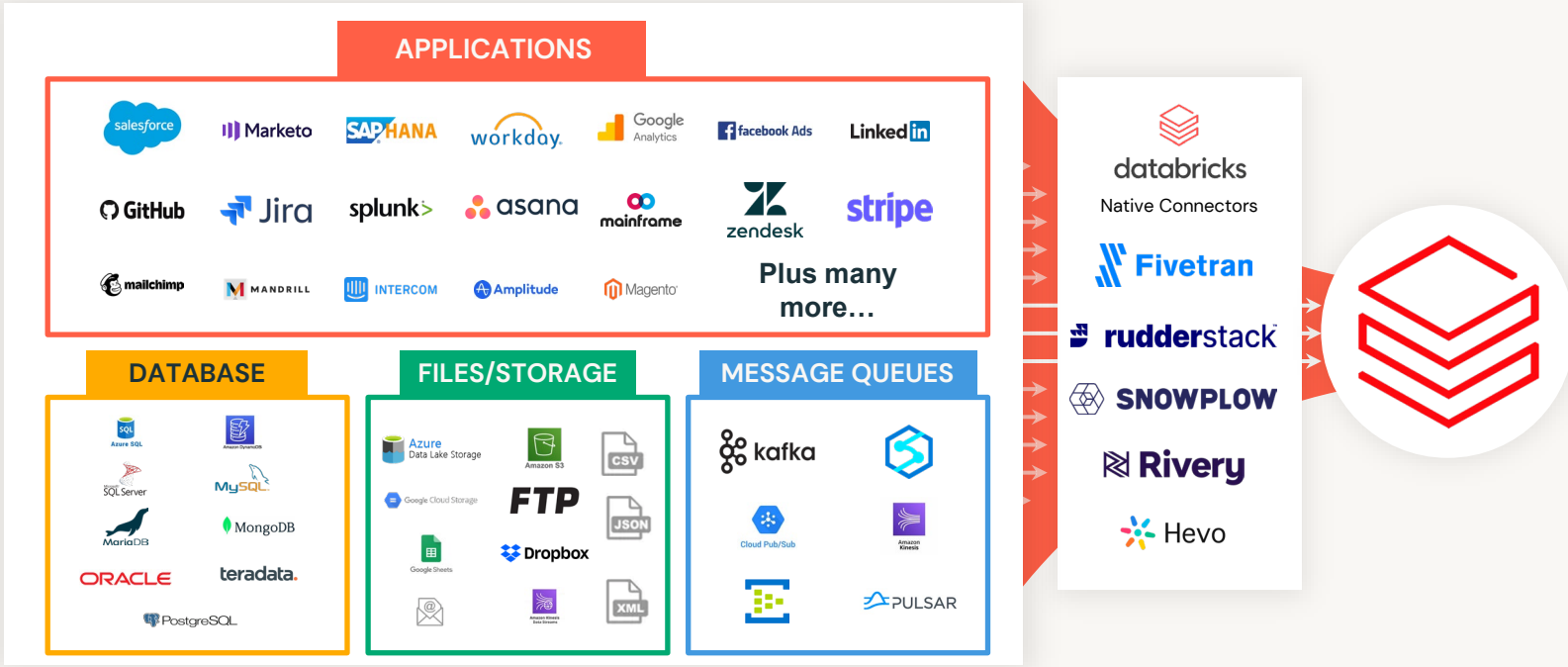
Ingest

Simple ingestion for every practitioner

Cost efficient, highly performant data ingestion

Native ingestion connectors

Large eco-system of partner solutions



Ingestion at any scale made simple

Landing data in the bronze layer

Source	Type	Tool	Environment	Method
Hundreds of sources	Structured	Databricks native connectors	Any notebook	Streaming
	Unstructured	Partner connectors	Databricks ingestion UI	Batch
			DLT Pipeline	
			DBSQL	

New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Genie Spaces
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
Models
Serving

Add data

Data sources

Search

From local files (2)

Create or modify table
Upload tabular data files to create a new table or replace an existing one

Upload files to volume
Upload files in any format, including structured, semi-structured, and unstructured data

Native integrations (12)

Azure Data Lak... | Azure Blob Stor... | Azure Data Lak... | Cassandra | Snowflake | JDBC | Kafka

Elasticsearch | MongoDB | Postgres | MySQL | DBFS

Free sample data from Databricks Marketplace (2,016)

Discover the open marketplace for data, analytics, and AI. View more on Databricks Marketplace

Property Chara... | CPG Retail and ... | B2B Commercia... | Places - Free N... | Atlas US Health... | See all...

Fivetran data sources (176) See all available ingest partners in Partner Connect

Google Drive | Google Ads | Google Analytics | Google Search ... | Salesforce | Facebook Ads | Facebook Pages

OneDrive | Hubspot | Zendesk Support | Shopify | Stripe | Jira | Microsoft Adver...

LinkedIn Ad Ana... | Webhooks | Postgres RDS | GitHub | NetSuite SuiteA... | Workday | Oracle

Data Pipelines Made Simple with DLT

Simple

- **Simple development**: Declarative programming for batch and streaming pipelines including ingestion, transformation, CDC/SCD and data quality expectations
- **Simple operations**: Serverless infrastructure for vertical/horizontal autoscaling, automated orchestration and fast startup & retries

Performant

- Rapid infrastructure scale-up
- Continuous mode for streaming
- Stream pipelining for fast ingestion and task parallelization
- Fast incremental transformation with Enzyme

Low TCO

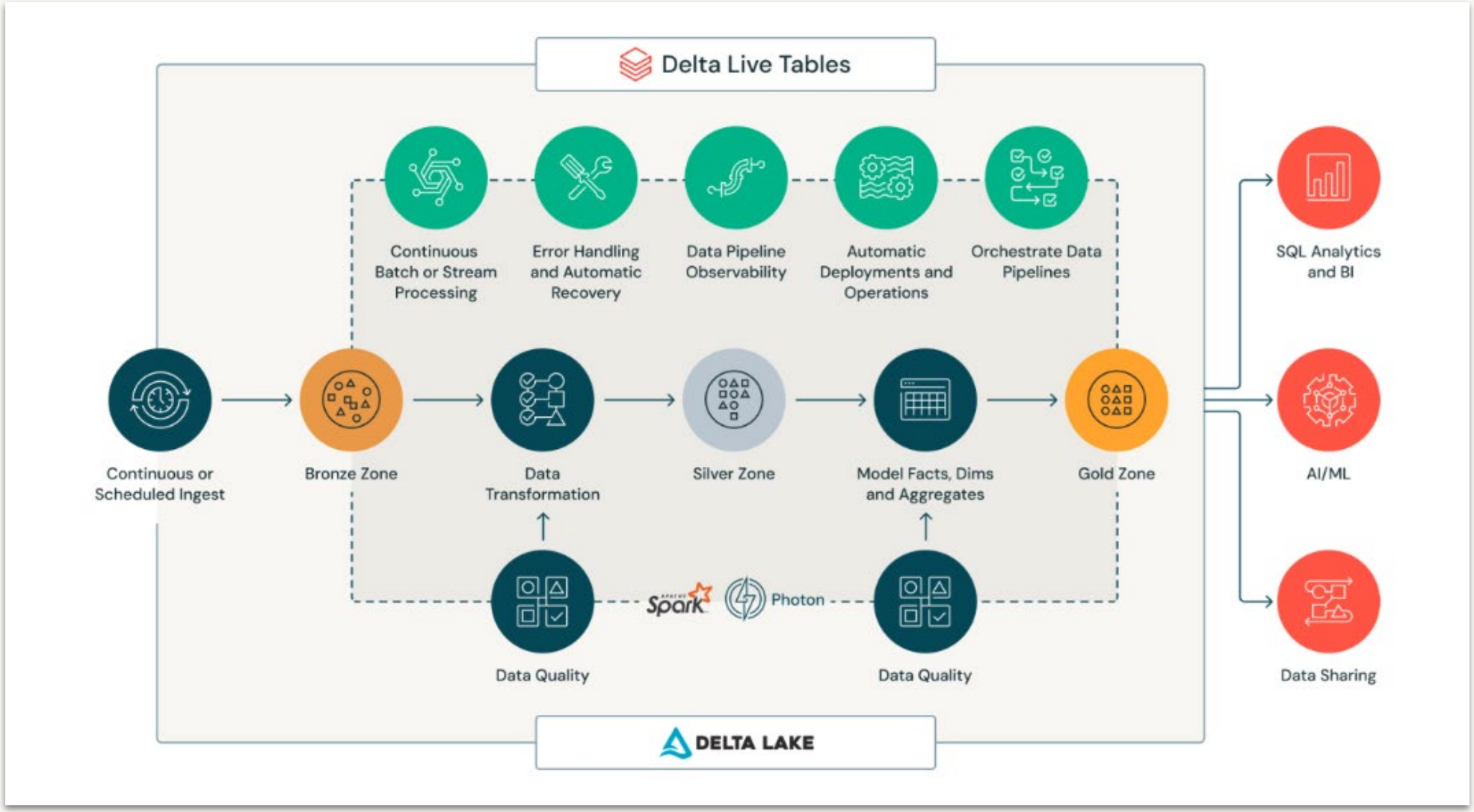
- **Efficient data processing:** Incremental ingestion and transformation
- **Efficient billing:** Only pay for what you use

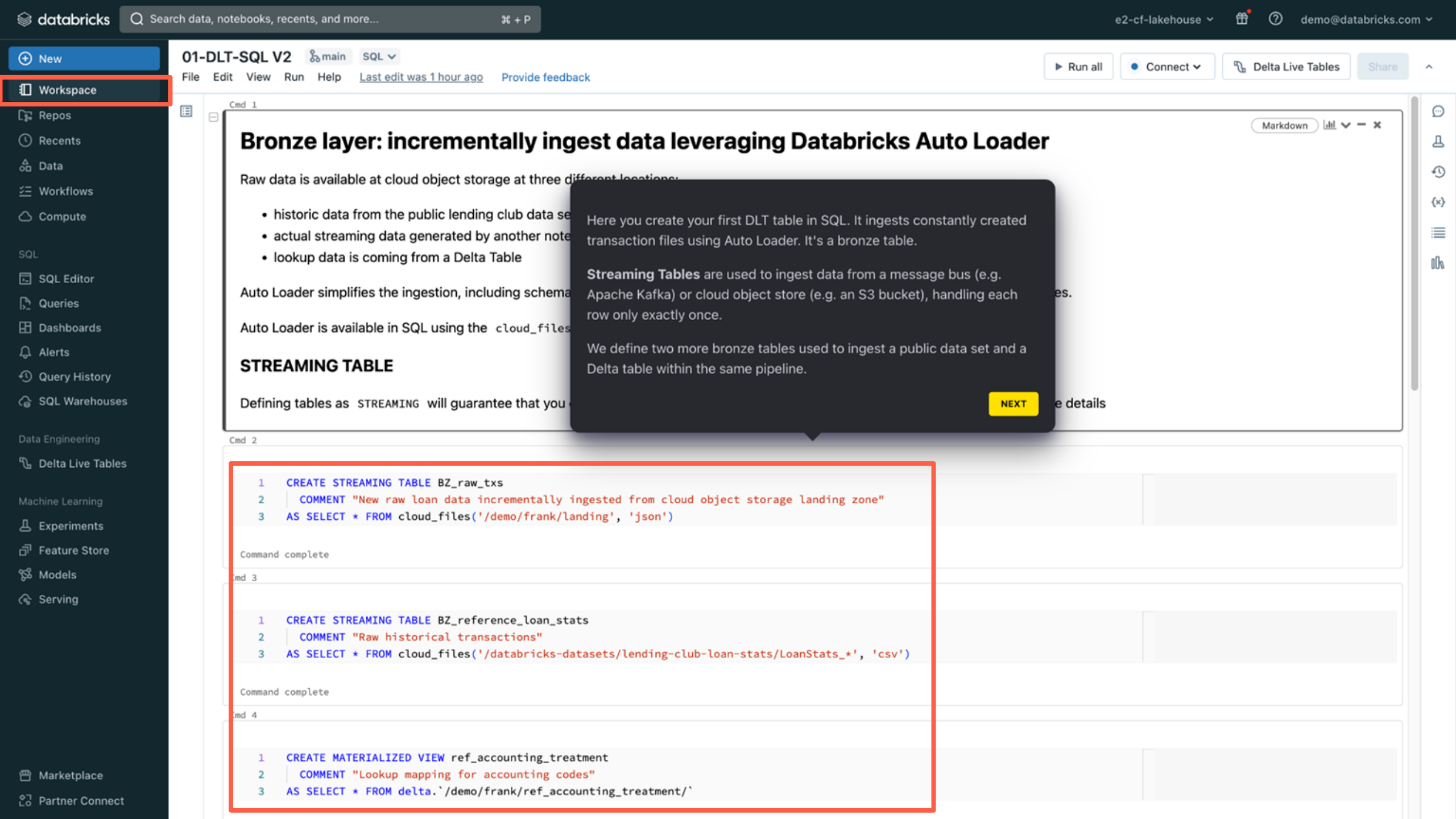
Delta Live Tables

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

CREATE MATERIALIZED VIEW clean_data
AS SELECT …
FROM raw_data
```

Data Pipelines Made Simple with DLT




databricks
Search data, notebooks, recents, and more... ⌘ + P
e2-cf-lakehouse
demo@databricks.com
New
Workspace
Repos
Recents
Data
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Delta Live Tables
Machine Learning
Experiments
Feature Store
Models
Serving
Marketplace
Partner Connect
01-DLT-SQL V2
main
SQL
File Edit View Run Help
Last edit was 1 hour ago
Provide feedback
Run all
Connect
Delta Live Tables
Share
Cmd 1
Markdown
Bronze layer: incrementally ingest data leveraging Databricks Auto Loader
Raw data is available at cloud object storage at three different locations:
historic data from the public lending club data se
actual streaming data generated by another note
lookup data is coming from a Delta Table
Auto Loader simplifies the ingestion, including schema
es.
Auto Loader is available in SQL using the cloud_files
STREAMING TABLE
Defining tables as STREAMING will guarantee that you
e details
Here you create your first DLT table in SQL. It ingests constantly created transaction files using Auto Loader. It's a bronze table.
Streaming Tables are used to ingest data from a message bus (e.g. Apache Kafka) or cloud object store (e.g. an S3 bucket), handling each row only exactly once.
We define two more bronze tables used to ingest a public data set and a Delta table within the same pipeline.
NEXT
Cmd 2
1 CREATE STREAMING TABLE BZ_raw_txs
2 COMMENT "New raw loan data incrementally ingested from cloud object storage landing zone"
3 AS SELECT * FROM cloud_files('/demo/frank/landing', 'json')
Command complete
Cmd 3
1 CREATE STREAMING TABLE BZ_reference_loan_stats
2 COMMENT "Raw historical transactions"
3 AS SELECT * FROM cloud_files('/databricks-datasets/lending-club-loan-stats/LoanStats_*', 'csv')
Command complete
Cmd 4
1 CREATE MATERIALIZED VIEW ref_accounting_treatment
2 COMMENT "Lookup mapping for accounting codes"
3 AS SELECT * FROM delta.`/demo/frank/ref_accounting_treatment/`

New
Workspace
Repos
Recents
Data
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Delta Live Tables
Machine Learning
Experiments
Feature Store
Models
Serving
Marketplace
Partner Connect

01-DLT-SQL V2

main | SQL

File Edit View Run Help | Last edit was 1 hour ago | Provide feedback

Run all | Connect | Delta Live Tables | Share

Show result

Cmd 5

Silver layer: joining tables while ensuring data quality

Once the bronze layer is defined, we'll create the sliver

To consume only incremental changes from the Bronz

DLT handles housekeeping such as compaction for us

Expectations

Expectations (`CONSTRAINT EXPECT`) enforce and trac

This silver table is used to join the previously defined bronze tables. It uses **Expectations** for data quality.

Expectations allow you to guarantee data arriving in tables meets certain data quality criteria. They also provide visual insights into data quality for each pipeline update.

NEXT

Cmd 6

```
CREATE STREAMING TABLE SV_cleaned_new_txs (
  CONSTRAINT `Payments should be this year`  EXPECT (next_payment_date > date('2020-12-31')),
  CONSTRAINT `Balance should be positive`    EXPECT (balance > 0 AND arrears_balance > 0) ON VIOLATION DROP ROW,
  CONSTRAINT `Cost center must be specified` EXPECT (cost_center_code IS NOT NULL) ON VIOLATION FAIL UPDATE

)
  COMMENT "Livestream of new transactions, cleaned and compliant"
AS SELECT txs.*, rat.id as accounting_treatment FROM stream(LIVE.BZ_raw_txs) txs
  INNER JOIN live.ref_accounting_treatment rat ON txs.accounting_treatment_id = rat.id
```

Command complete

Cmd 7

```
CREATE MATERIALIZED VIEW SV_historical_txs
  COMMENT "Historical loan transactions"
AS SELECT a.* FROM LIVE.BZ_reference_loan_stats a
  INNER JOIN LIVE.ref_accounting_treatment b USING (id)
```

Command complete

Cmd 8

New
Workspace
Repos
Recents
Data
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Delta Live Tables
Machine Learning
Experiments
Feature Store
Models
Serving
Marketplace
Partner Connect

01-DLT-SQL V2

main | SQL

File Edit View Run Help | Last edit was 1 hour ago | Provide feedback

Run all | Connect | Delta Live Tables | Share

Command complete

Cmd 8

Gold layer

Our last step is to implement the Gold Layer.

These tables will be requested at scale using a DWH (s eries using `pipelines.autoOptimize.zOrderCols` , and DLT will handle it.

The gold layer curates data for downstream systems aggregating data in silver tables.

We use a **Materialized View** (MV) here where the results have been precomputed. With serverless compute enabled these MVs are computed incrementally.

NEXT

Cmd 9

```
1  CREATE MATERIALIZED VIEW GL_total_loan_balances_1
2    COMMENT "Combines historical and new loan data for unified rollup of loan balances"
3    TBLPROPERTIES ("pipelines.autoOptimize.zOrderCols" = "location_code")
4  AS SELECT sum(revol_bal)  AS bal, addr_state   AS location_code FROM live.SV_historical_txs  GROUP BY addr_state
5    UNION SELECT sum(balance) AS bal, country_code AS location_code FROM live.SV_cleaned_new_txs GROUP BY country_code
```

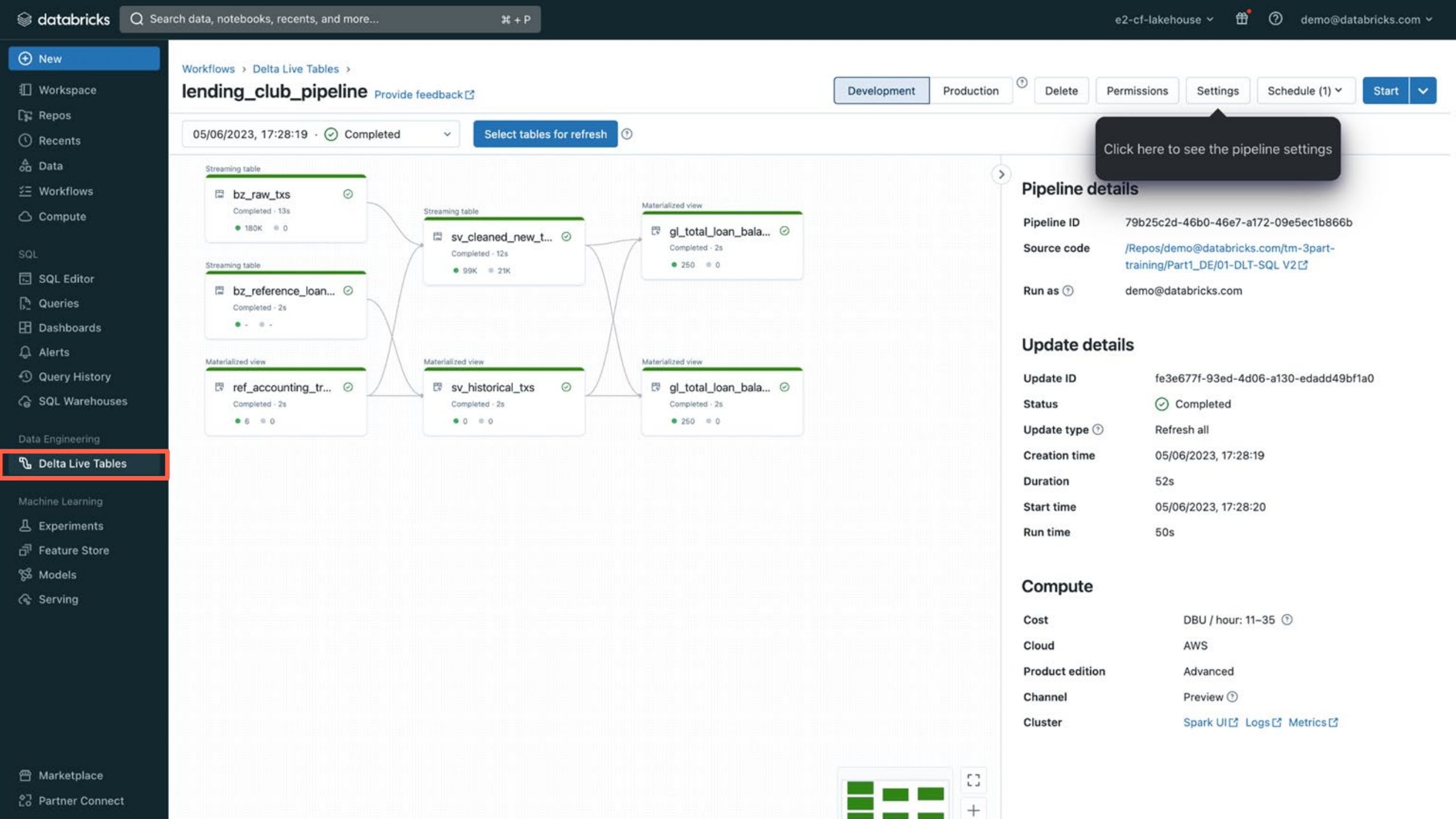
Command complete

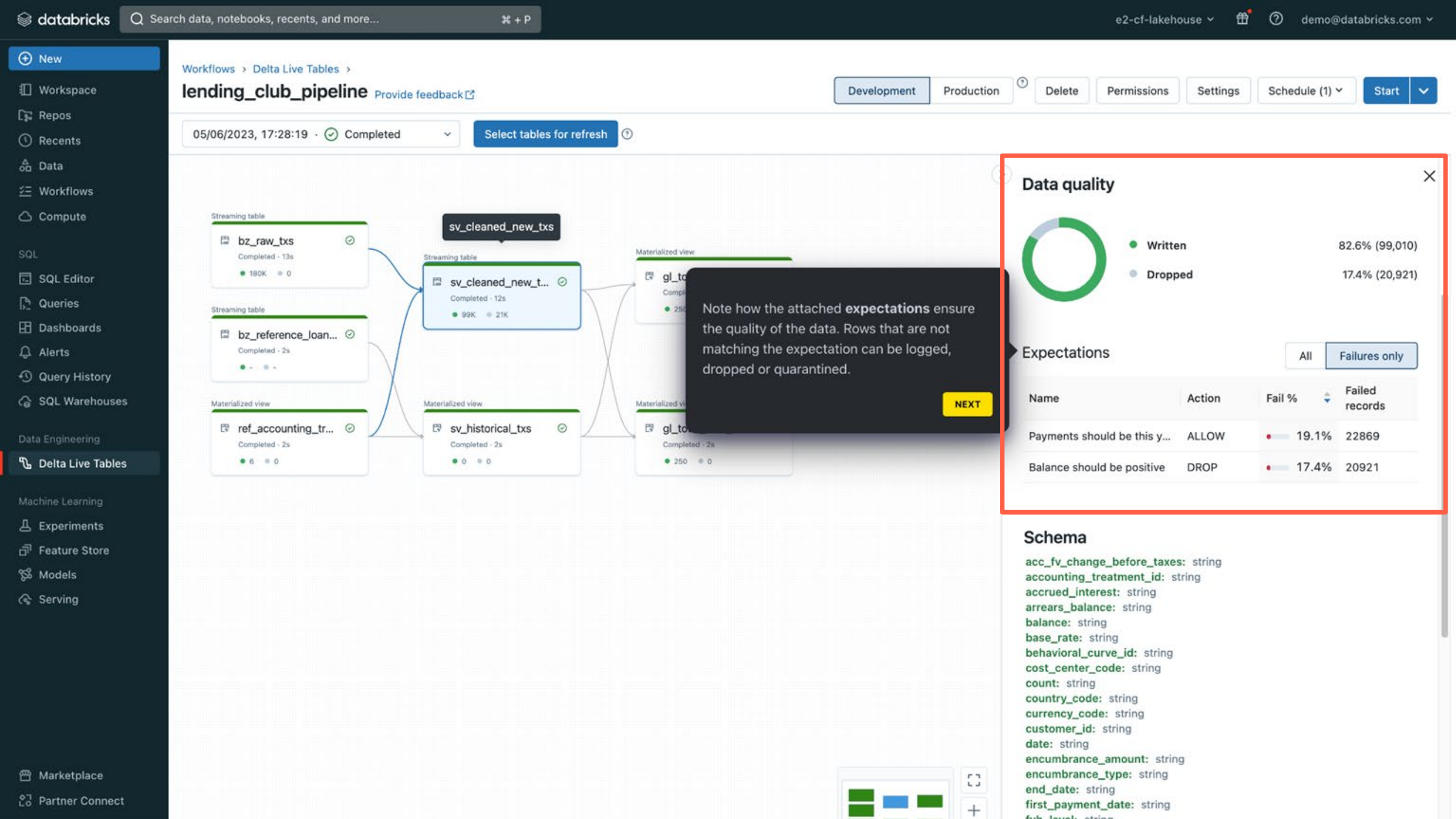
Cmd 10

```
1  CREATE MATERIALIZED VIEW GL_total_loan_balances_2
2    COMMENT "Combines historical and new loan data for unified rollup of loan balances"
3  AS SELECT sum(revol_bal)  AS bal, addr_state   AS location_code FROM live.SV_historical_txs  GROUP BY addr_state
4    UNION SELECT sum(balance) AS bal, country_code AS location_code FROM live.SV_cleaned_new_txs GROUP BY country_code
```

Command complete

Shift+Enter to run
Shift+Ctrl+Enter to run selected text


databricks
Search data, notebooks, recents, and more... ⌘ + P
e2-cf-lakehouse
demo@databricks.com
New
Workspace
Repos
Recents
Data
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Delta Live Tables
Machine Learning
Experiments
Feature Store
Models
Serving
Marketplace
Partner Connect
Workflows › Delta Live Tables ›
lending_club_pipeline
Provide feedback
Development
Production
Delete
Permissions
Settings
Schedule (1)
Start
Click here to see the pipeline settings
05/06/2023, 17:28:19 · Completed
Select tables for refresh
Streaming table
bz_raw_txs
Completed · 13s
180K 0
Streaming table
bz_reference_loan...
Completed · 2s
Materialized view
ref_accounting_tr...
Completed · 2s
6 0
Streaming table
sv_cleaned_new_t...
Completed · 12s
99K 21K
Materialized view
sv_historical_txs
Completed · 2s
0 0
Materialized view
gl_total_loan_bala...
Completed · 2s
250 0
Materialized view
gl_total_loan_bala...
Completed · 2s
250 0
Pipeline details
Pipeline ID 79b25c2d-46b0-46e7-a172-09e5ec1b866b
Source code /Repos/demo@databricks.com/tm-3part-training/Part1_DE/01-DLT-SQL V2
Run as demo@databricks.com
Update details
Update ID fe3e677f-93ed-4d06-a130-edadd49bf1a0
Status Completed
Update type Refresh all
Creation time 05/06/2023, 17:28:19
Duration 52s
Start time 05/06/2023, 17:28:20
Run time 50s
Compute
Cost DBU / hour: 11–35
Cloud AWS
Product edition Advanced
Channel Preview
Cluster Spark UI Logs Metrics


databricks
Search data, notebooks, recents, and more... ⌘ + P
e2-cf-lakehouse
demo@databricks.com
New
Workspace
Repos
Recents
Data
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Delta Live Tables
Machine Learning
Experiments
Feature Store
Models
Serving
Marketplace
Partner Connect
Workflows › Delta Live Tables ›
lending_club_pipeline
Provide feedback
Development
Production
Delete
Permissions
Settings
Schedule (1)
Start
05/06/2023, 17:28:19 · Completed
Select tables for refresh
Streaming table
bz_raw_txs
Completed · 13s
180K 0
Streaming table
bz_reference_loan...
Completed · 2s
Materialized view
ref_accounting_tr...
Completed · 2s
6 0
sv_cleaned_new_txs
Streaming table
sv_cleaned_new_t...
Completed · 12s
99K 21K
Materialized view
sv_historical_txs
Completed · 2s
0 0
Materialized view
Completed · 2s
250 0
Note how the attached expectations ensure the quality of the data. Rows that are not matching the expectation can be logged, dropped or quarantined.
NEXT
Data quality
Written 82.6% (99,010)
Dropped 17.4% (20,921)
Expectations
All
Failures only
Name
Action
Fail %
Failed records
Payments should be this y...
ALLOW
19.1%
22869
Balance should be positive
DROP
17.4%
20921
Schema
acc_fv_change_before_taxes: string
accounting_treatment_id: string
accrued_interest: string
arrears_balance: string
balance: string
base_rate: string
behavioral_curve_id: string
cost_center_code: string
count: string
country_code: string
currency_code: string
customer_id: string
date: string
encumbrance_amount: string
encumbrance_type: string
end_date: string
first_payment_date: string

New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Genie Spaces
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
Models
Serving

Pipeline settings

UI | JSON

General

* Pipeline name

Serverless

Pipeline mode

Triggered | Continuous

Source code

Paths to notebooks or files that contain pipeline source code.

* Paths

Add source code

Destination

Storage options

Hive Metastore | Unity Catalog Preview

Catalog

main

Target schema

Notifications

Summary

- Serverless enabled
- Photon enabled

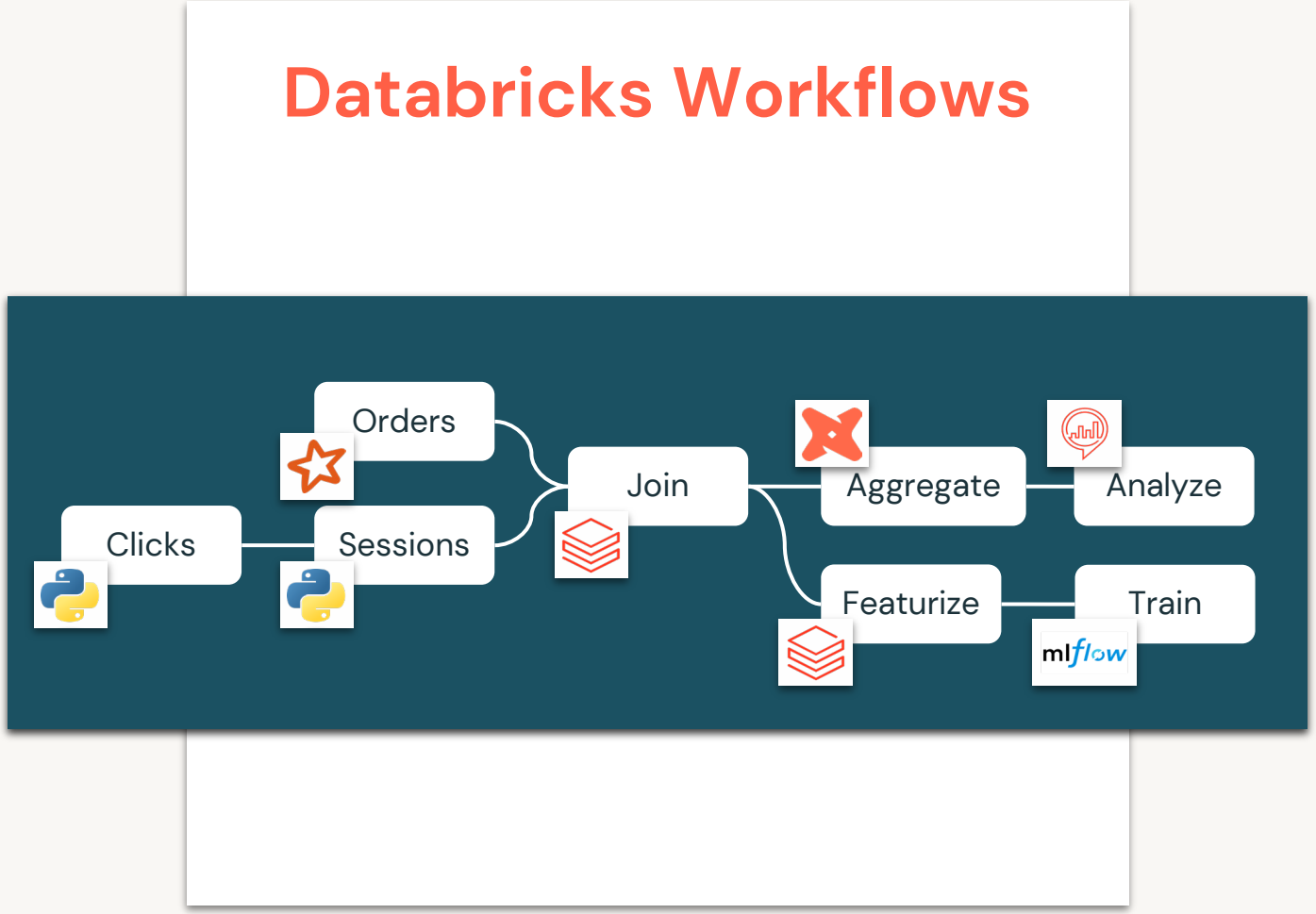
Orchestrate

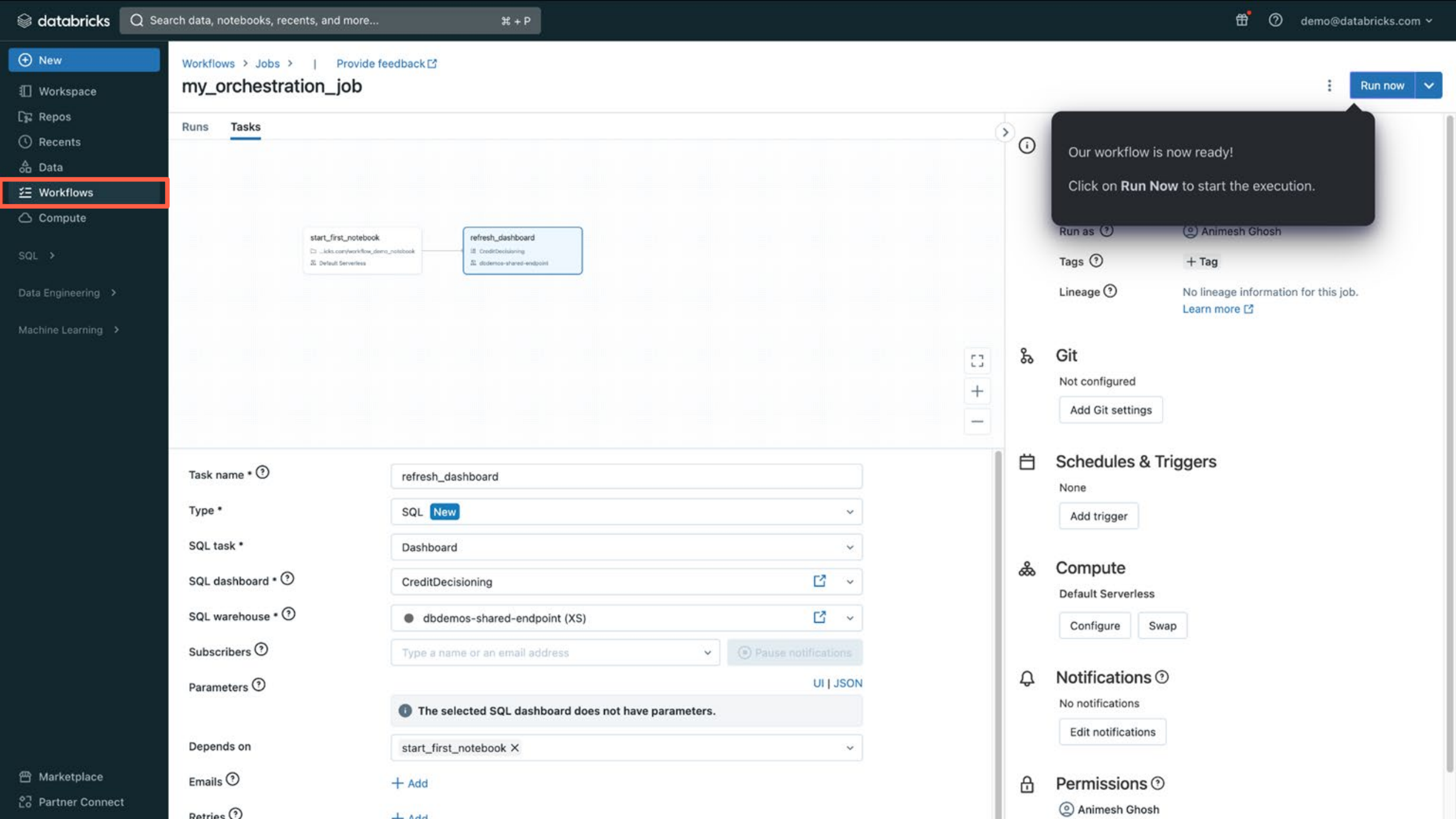
Unified orchestration for Data and AI with Databricks Workflows

Simple workflow authoring for any practitioner

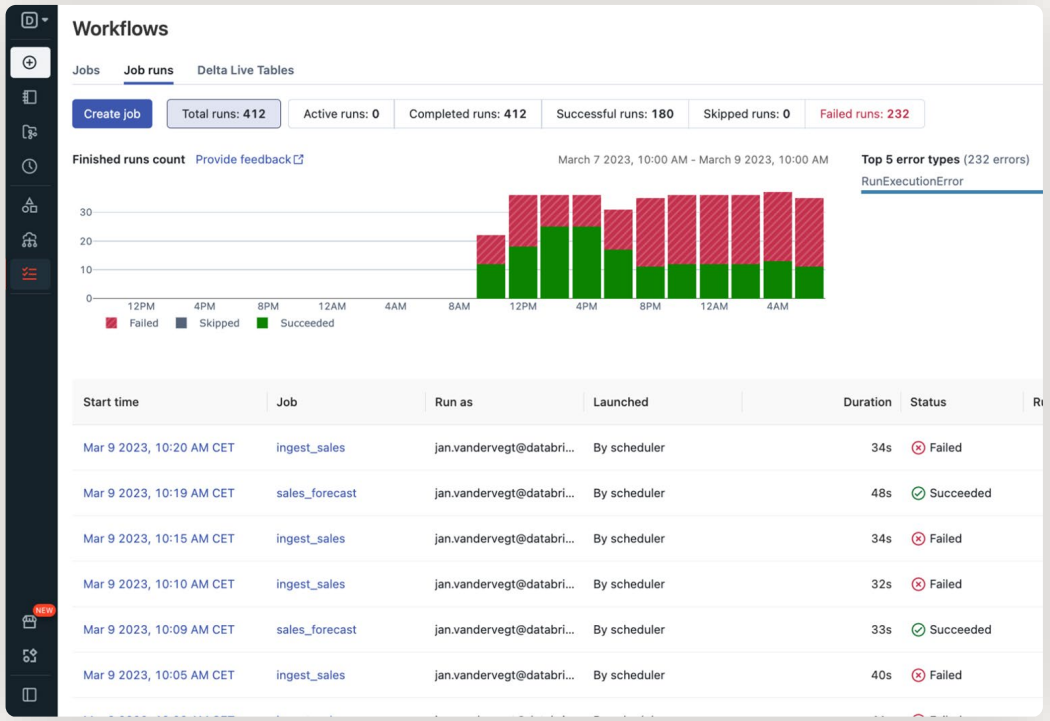
Actionable insights with end-to-end observability

Reliability at scale in production

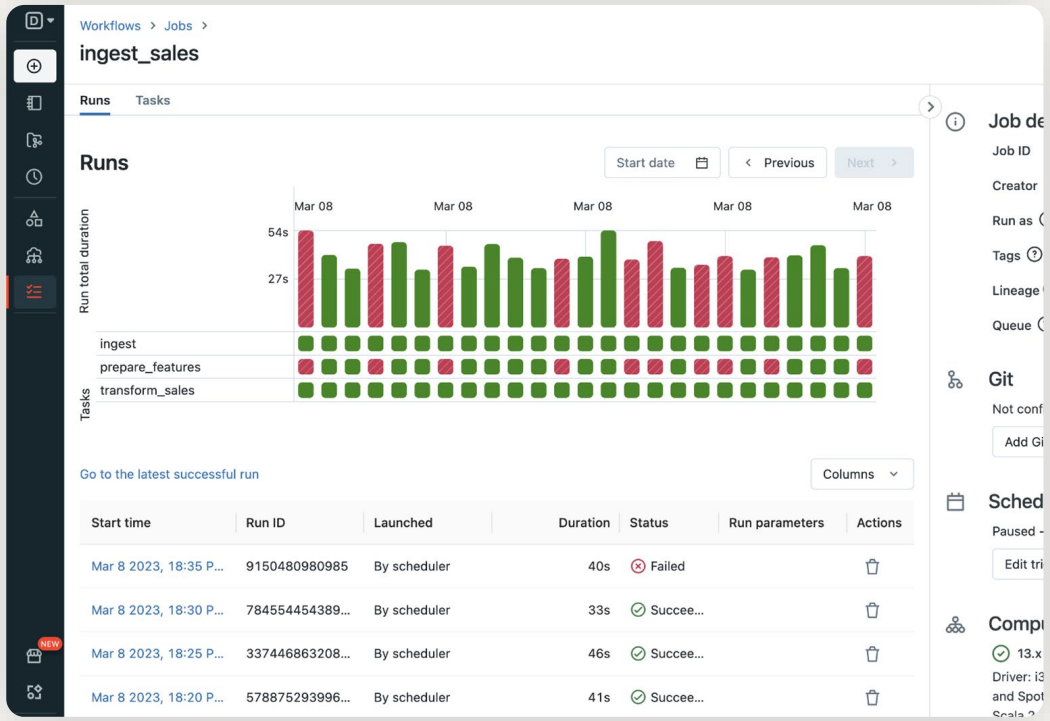



databricks
Search data, notebooks, recents, and more... ⌘ + P
demo@databricks.com
New
Workspace
Repos
Recents
Data
Workflows
Compute
SQL
Data Engineering
Machine Learning
Marketplace
Partner Connect
Workflows > Jobs > | Provide feedback
my_orchestration_job
Run now
Runs
Tasks
start_first_notebook
refresh_dashboard
Our workflow is now ready!
Click on Run Now to start the execution.
Run as
Animesh Ghosh
Tags
+ Tag
Lineage
No lineage information for this job.
Learn more
Git
Not configured
Add Git settings
Schedules & Triggers
None
Add trigger
Compute
Default Serverless
Configure
Swap
Notifications
No notifications
Edit notifications
Permissions
Animesh Ghosh
Task name *
refresh_dashboard
Type *
SQL New
SQL task *
Dashboard
SQL dashboard *
CreditDecisioning
SQL warehouse *
dbdemos-shared-endpoint (XS)
Subscribers
Type a name or an email address
Pause notifications
Parameters
UI | JSON
The selected SQL dashboard does not have parameters.
Depends on
start_first_notebook
Emails
+ Add
Retries
+ Add

Actionable insights from real-time monitoring



A simple and intuitive monitoring UI provides real-time metrics and detailed analytics for every workflow run



Drill down to understand which tasks are failing and why. Troubleshoot issues before your customers are impacted

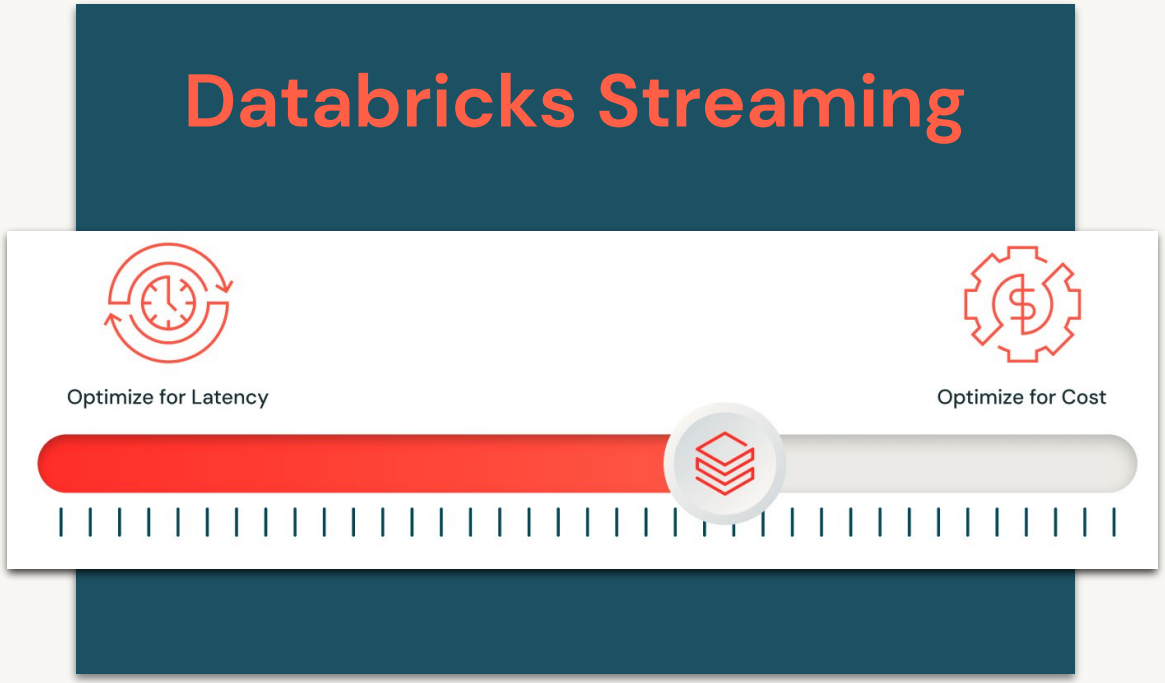
Real-time processing

Tunable streaming architecture built on the world's most popular streaming engine

Simplified development with unified batch and streaming APIs

Reliable operations through automatic checkpointing and failure recovery

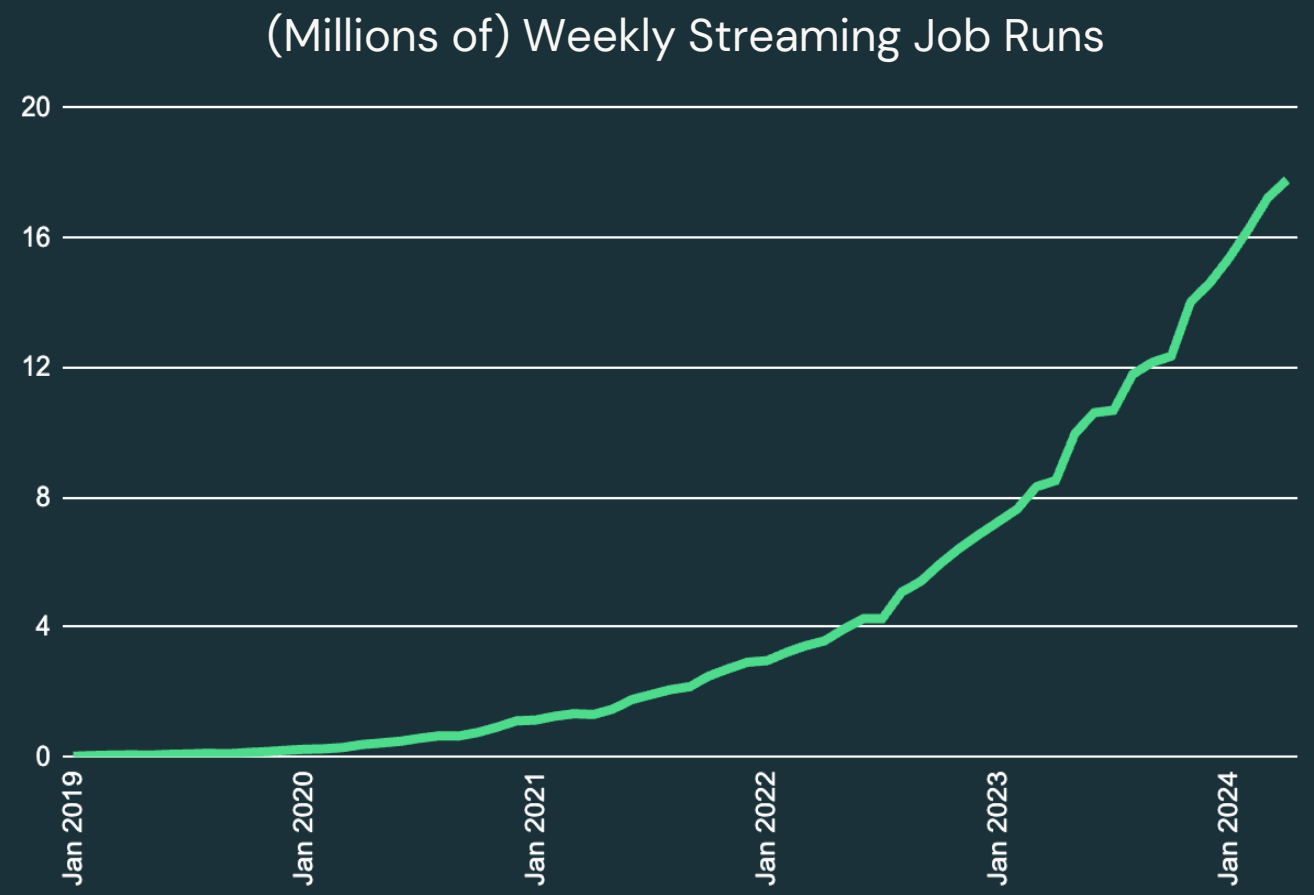
Easily adjustable throughput and latency for enterprise workload requirements



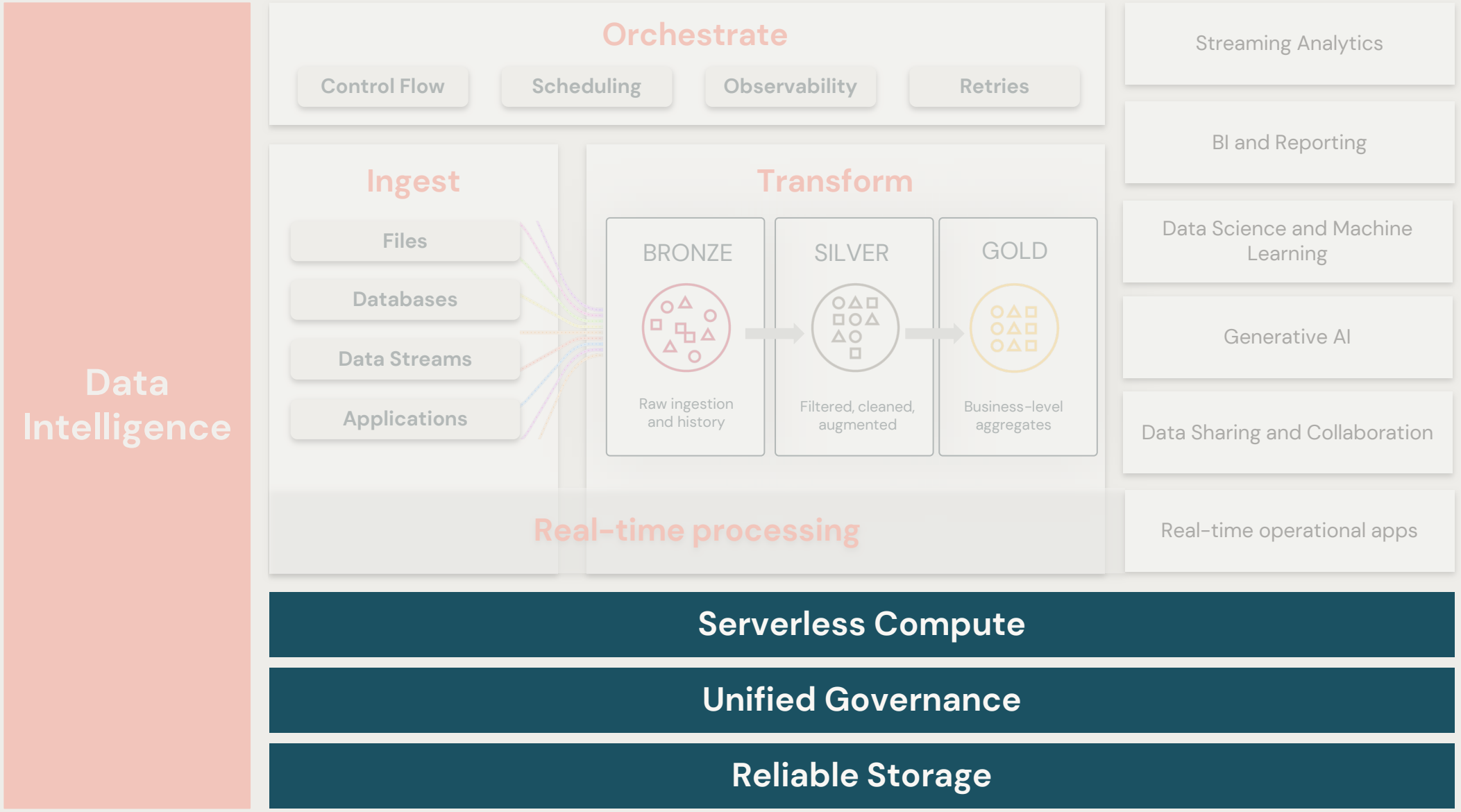
Streaming workloads continue to grow on the Databricks Data Intelligence Platform

>100%

YoY job growth



databricks Data Intelligence Platform



Built on a solid, open foundation

Reliable storage, unified governance and fully managed serverless compute

Delta Lake

The open format storage layer

High reliability and performance

Automatic instant translation across open formats with UniForm

Open source

Unity Catalog

Unified governance across all workloads

Single permission model for data and AI

AI powered monitoring, observability and lineage

Secure open data sharing

Serverless Compute

Hands off, fully managed compute

Fast startups

Cost efficient with smart autoscaling

Stable and secured by default

Serverless Compute

SIMPLE

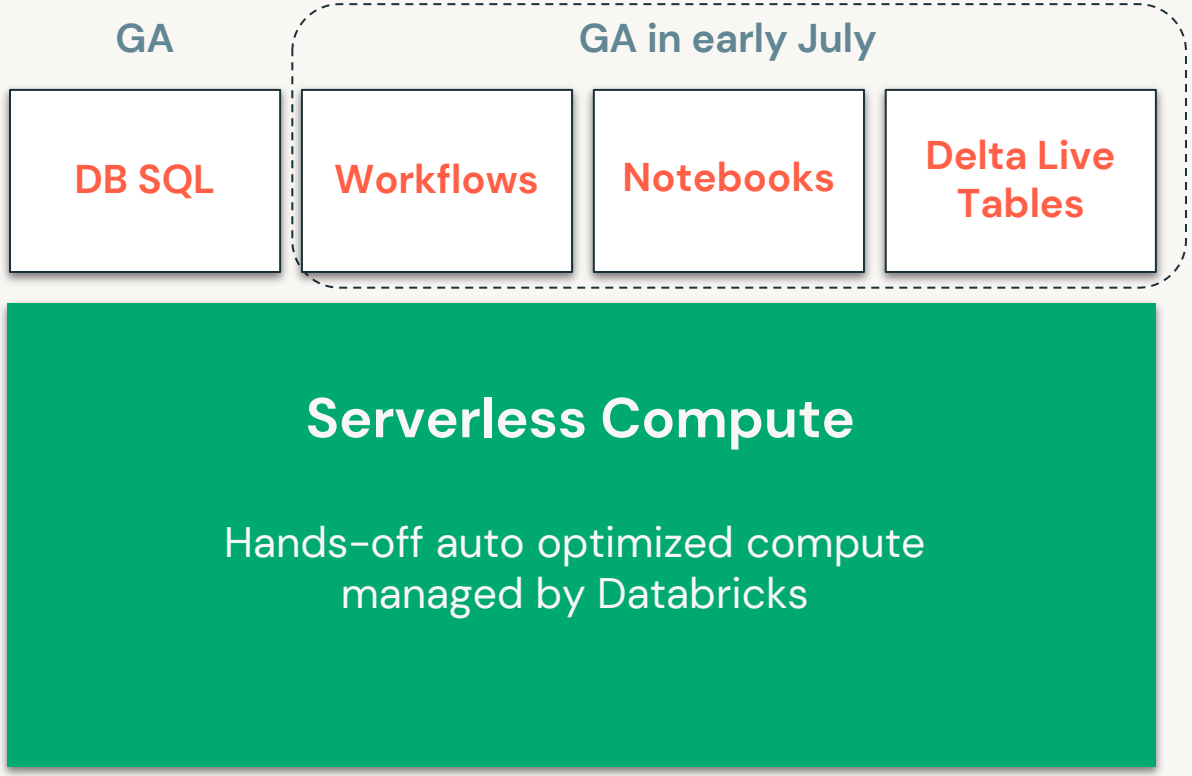
No knobs
No version upgrade
For any practitioner

COST EFFICIENT

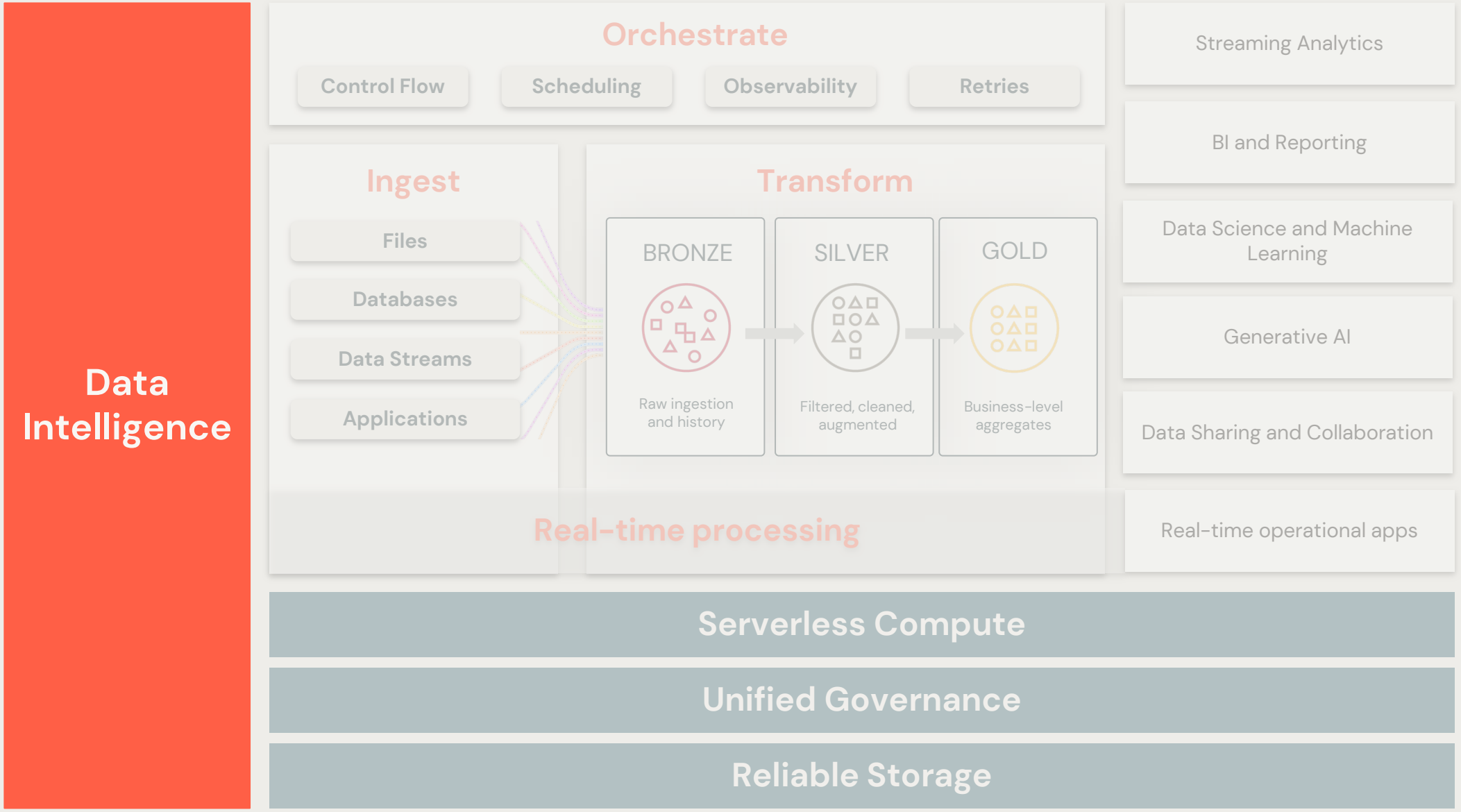
Fast startup times
Lower engineering overhead
Pay for what you use

RELIABLE

Reduced downtime
Secure by default



databricks Data Intelligence Platform



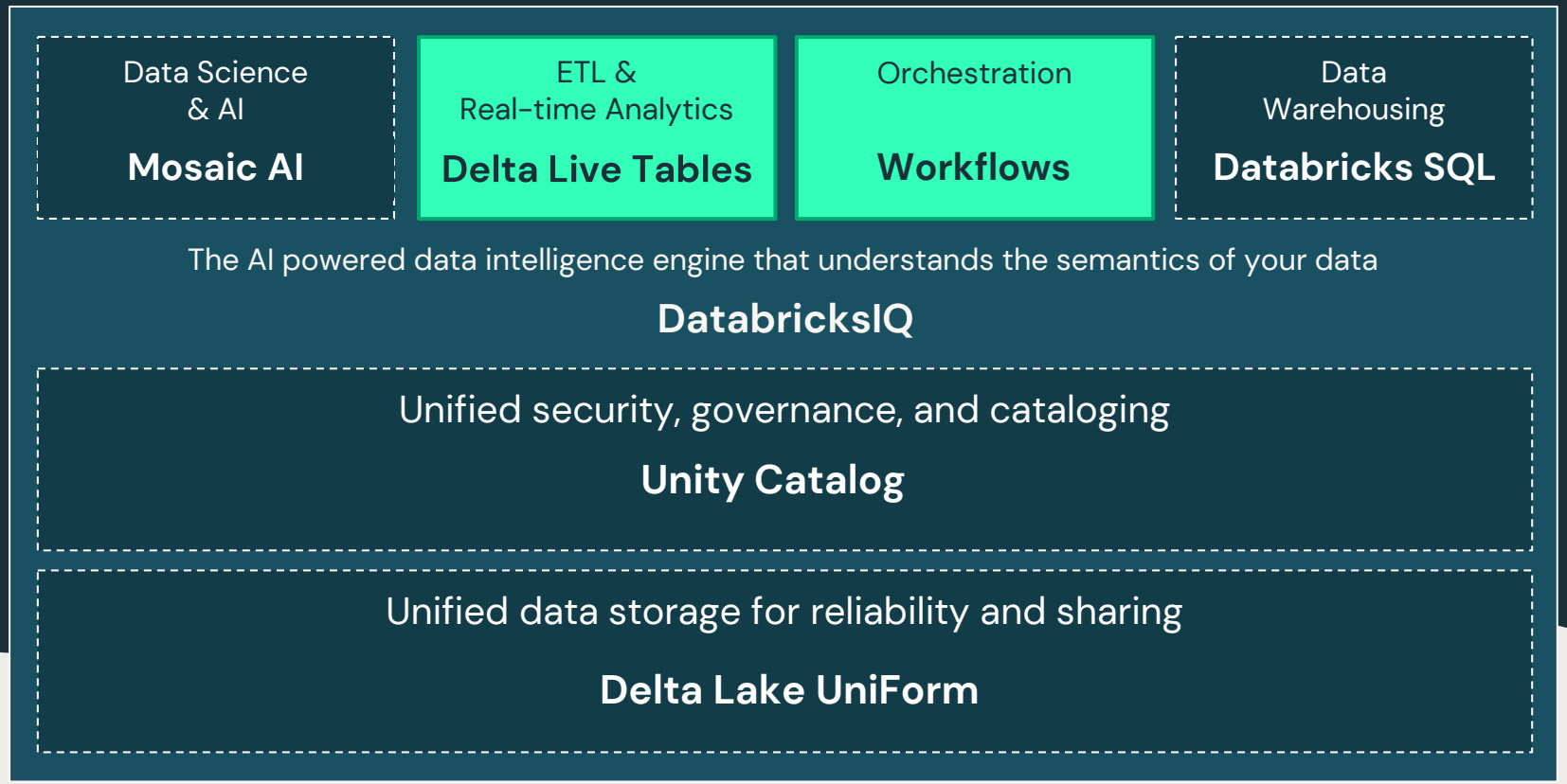
DatabricksIQ powers intelligent data engineering

AI to boost user productivity

- Intelligent Search
- Assistant in Notebooks, Jobs, & SQL Editor

AI to auto-optimize your workloads

- Intelligent Workload Management (Serverless)
- Intelligent orchestration and flow optimization

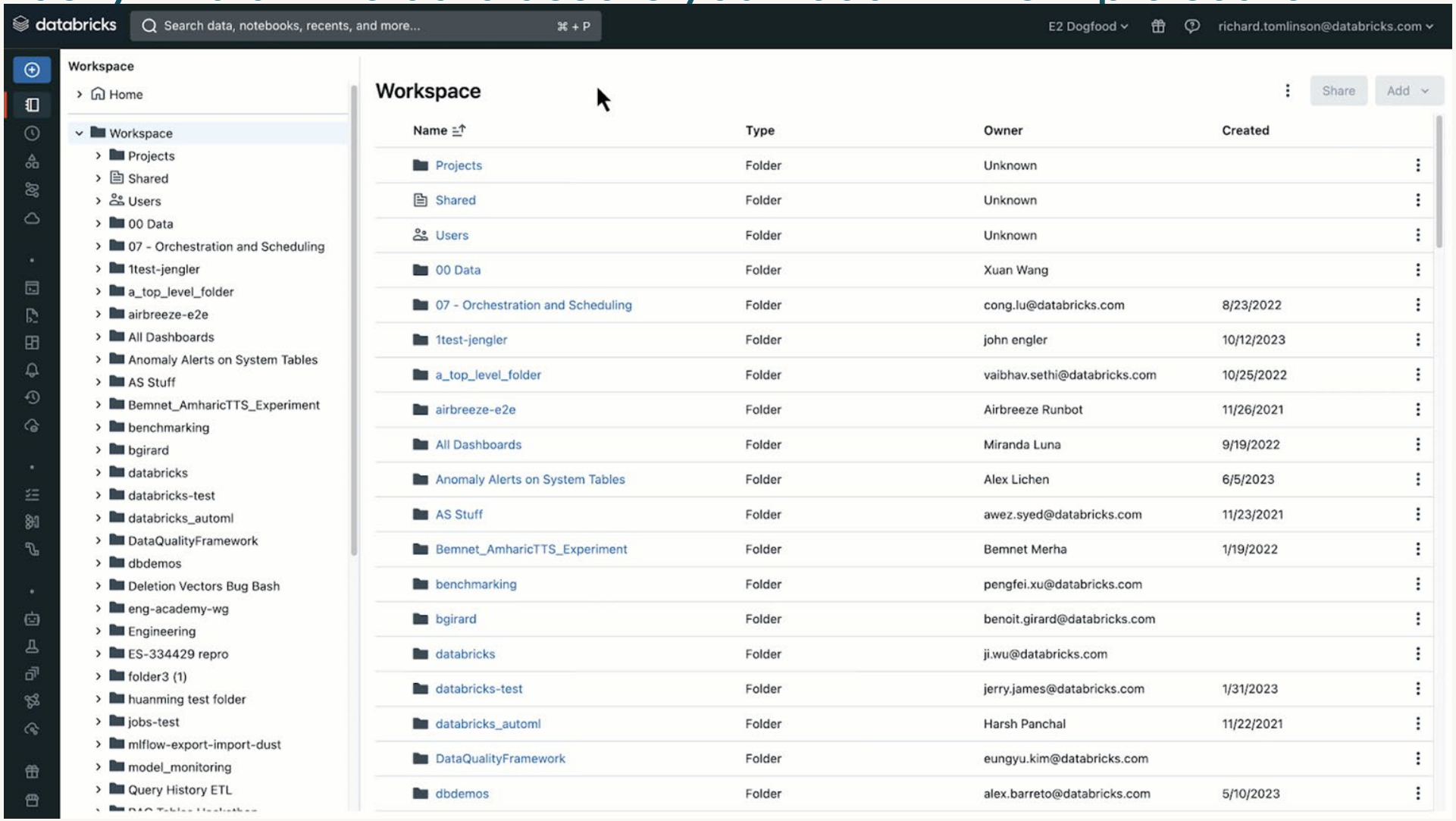


Open Data Lake

All Raw Data
(Logs, Texts, Audio, Video, Images)

Intelligent Search

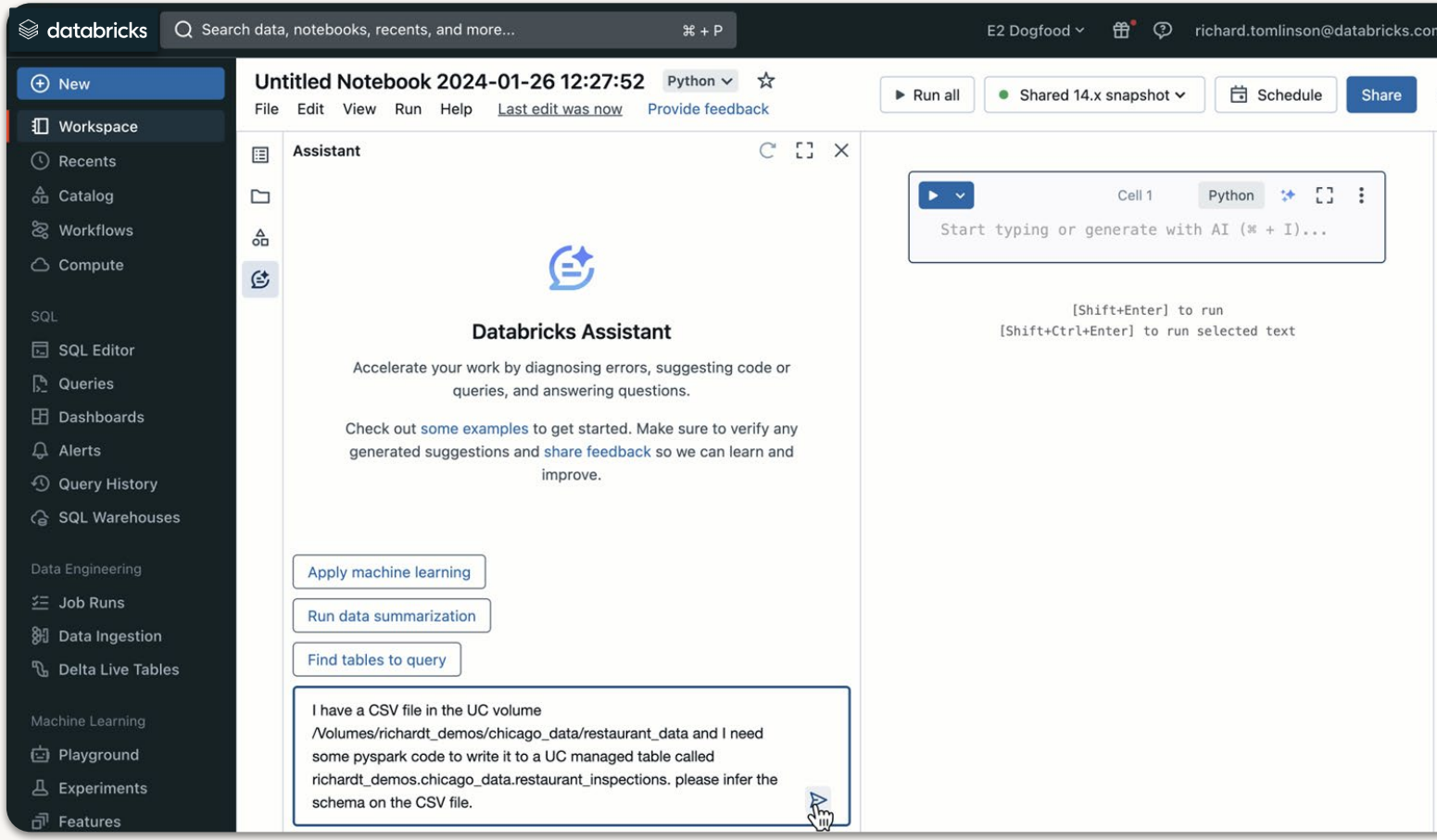
Easily find all the data assets you need with simple search



Databricks Assistant

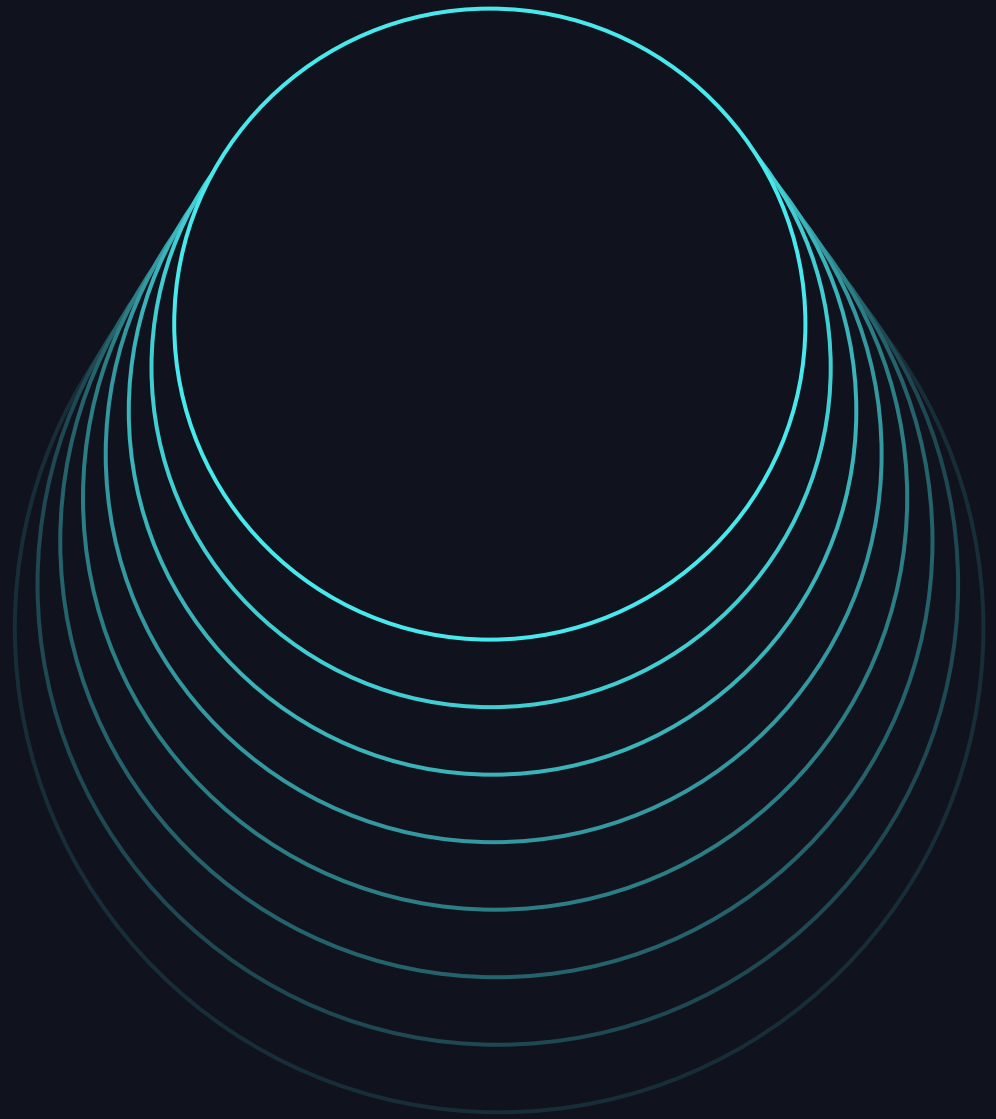
Boost productivity with your context-aware AI assistant

- Generate, explain, and fix code with natural language
- AI assistance in every user experience
- Powered by DatabricksIQ for highly relevant answers



DATA+AI SUMMIT
BY databricks

LEXMARK'S INTELLIGENT DATA ENGINEERING JOURNEY WITH DATABRICKS



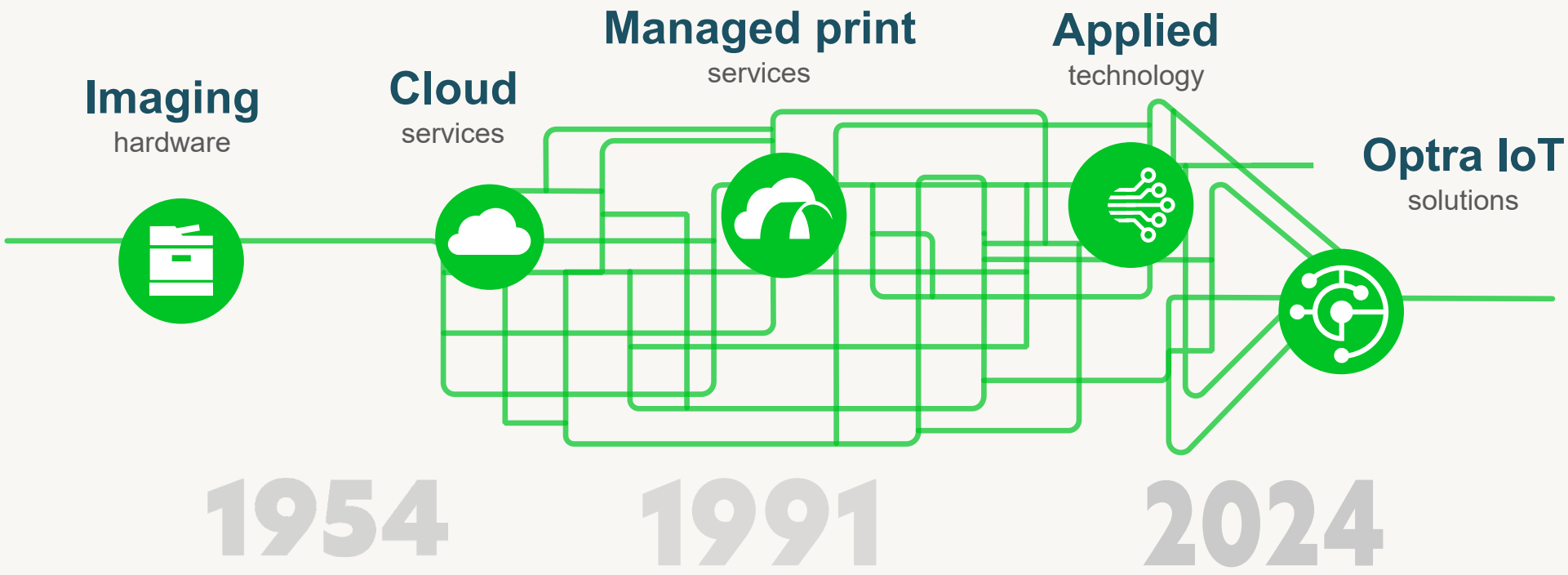
Raju Mudunuri
Lexmark Data Solutions & Strategy

Agenda

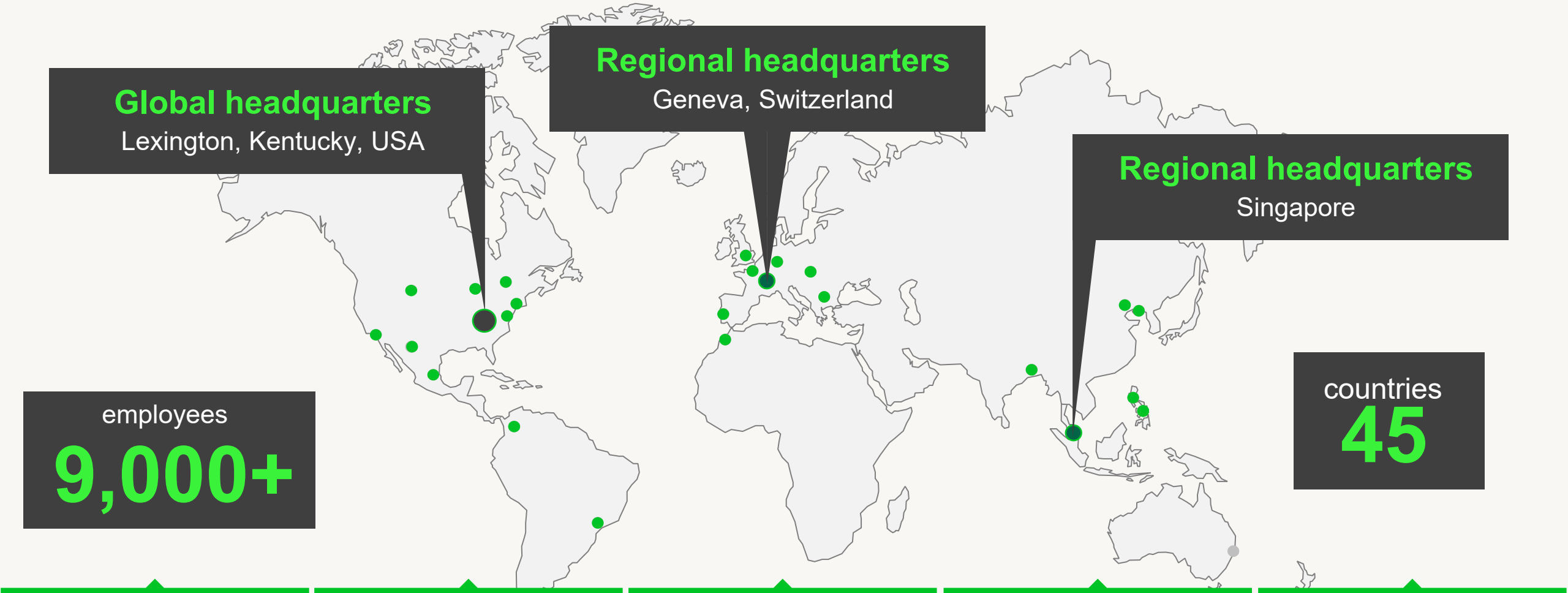
- Lexmark's corporate identity and mission
- Business Use Case - challenge and solution
- Technical and business outcomes
- Demo/screenshots

LEXMARK – Global Imaging & Technology Company

Creates cloud-enabled Imaging & IOT solutions powered by advance data analytics and cutting edge technologies of AI/ML to help worldwide customers quickly realize business outcomes



A Global Leader

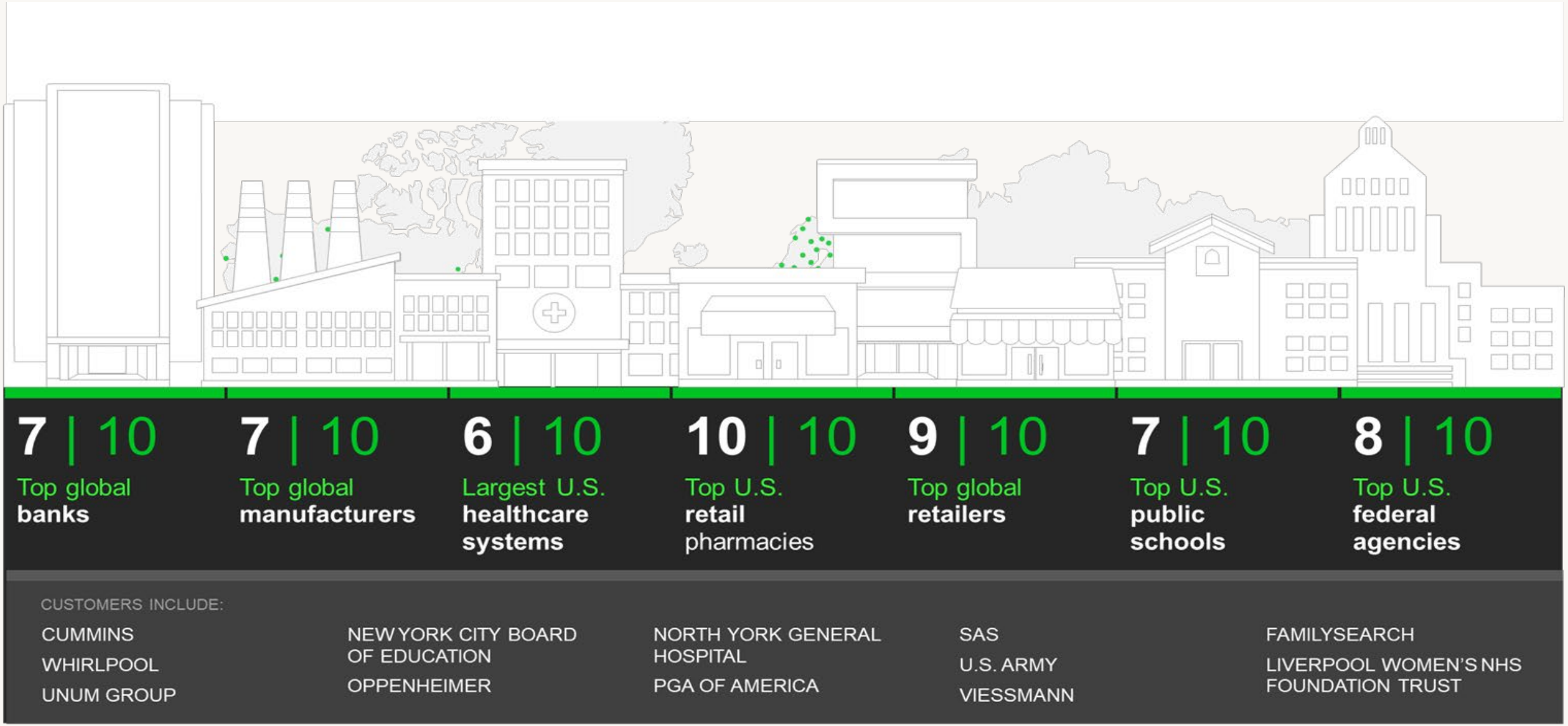


employees
9,000+

countries
45

Primary development centers	Manufacturing facilities	Shared services centers	Support centers	Major field sales presence

Enable customers make a lasting impression on *their* World



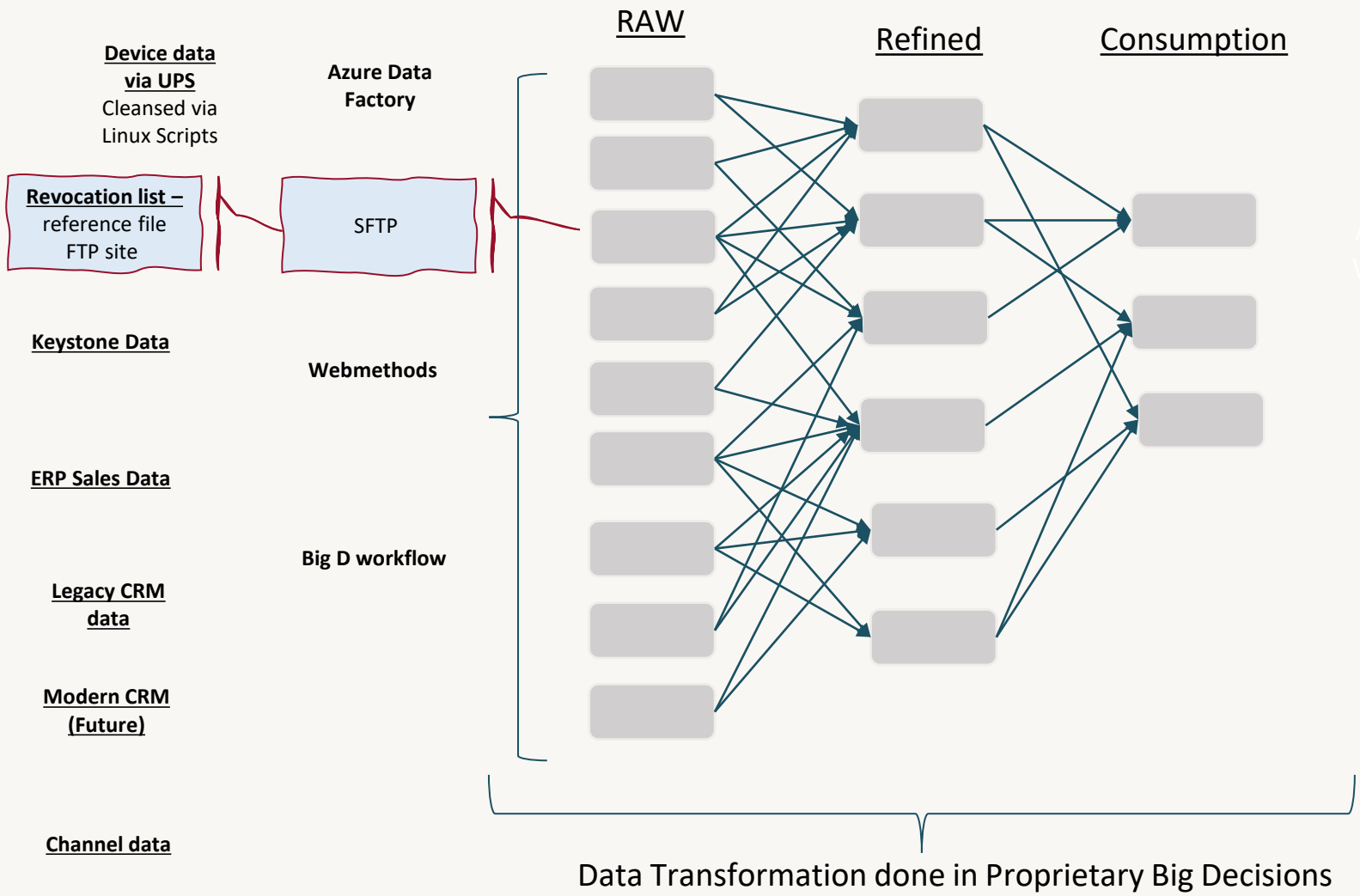
Use Case – Hardware Placement Optimization

Lexmark core business depends on productive hardware placements. However today, there are **less available pages** in the market due to:

- Post-COVID work environment
- Accelerated digitization efforts
- Global recession
- Increased aftermarket competition
- Trending flat imaging industry

Therefore, Lexmark must **optimize page usage** within each placement as we work to grow our overall install base

Legacy Solution – Hardware Placement Optimization



- No uniformity – varied tech stack
- Silos – data ingestion & processing
- Redundant data across
- Proprietary data refinement
- No Automation – CI/CD
- Lack of automated data Lineage
- Missing centralized governance and compliance stature
- Operational Overhead
- Difficulty applying security policies

Objectives – Hardware Placement Optimization

Business Problem:

Determine where to place our devices to provide the best ROI for the business

Business Objectives:

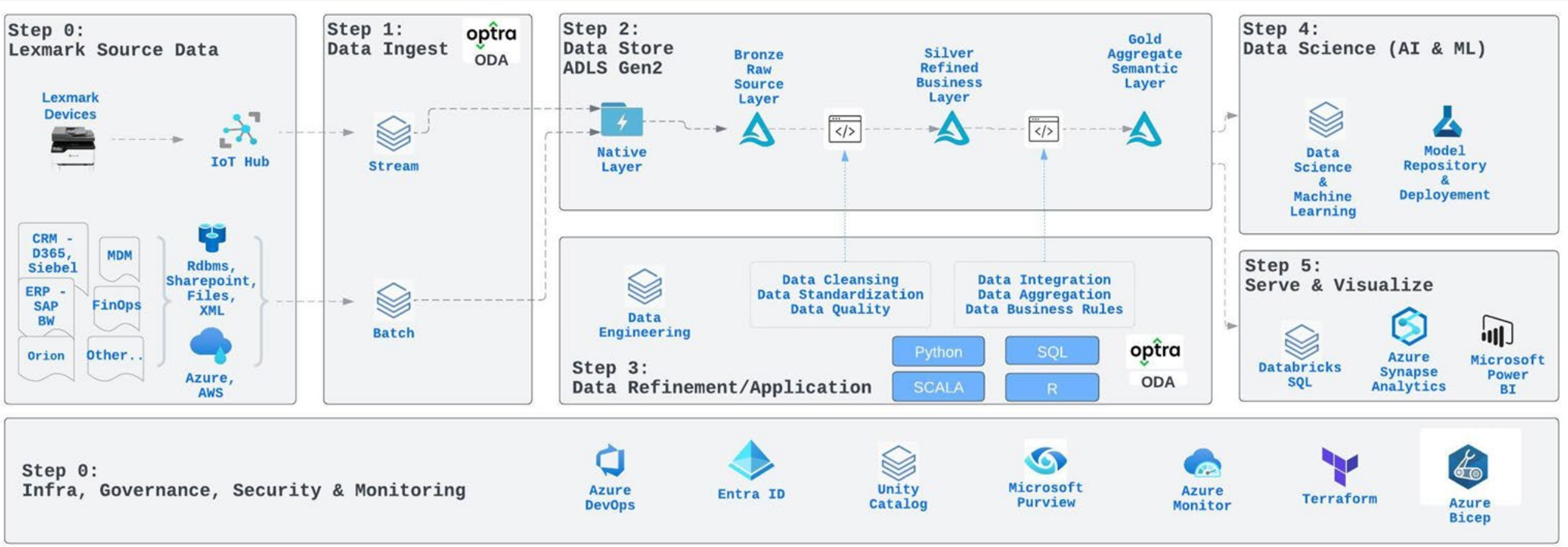
- Connect multiple data sources to printers → **Better understand device utilization**
- Leverage device utilization info → **Determine optimal product mix and paths-to-market**
- Discover page usage issues → **Take corrective actions with LoBs**

Technical Objectives:

- Mature from “data-aware” to “data-driven”
- Treat data as an asset
- Optimize time to insight
- Boost returns on data investment
- Build foundation for data to generate revenue

Business Solution: Reference Data Architecture

Moving from Strategy to Implementation – Simplification & Focus



- Reduce complexity and accelerate time to results
- Unified and consistent semantic layer for BI & Data Science
- Reduce Total Cost Of Ownership
- Improve Governance, Security, Compliance & Reliability

Technical Outcomes

Achievable due to the unified data architecture of Databricks Data Intelligence Platform

IT Data Operations

- ODA – sped up our ingestion process by **65%**
- Databricks Workflows reduced man hours from **months to weeks**

Metadata & Enhanced Data Governance

- Improved data governance and compliance stature
- Empower stakeholders to combine technical datasets and business glossary for advanced BI and data science

Community of Learning

- Leverage business teams' resources to enable others through education & training
- Help IT Data Team learn the basic tenets of business

Business Outcomes

Achievable due to the unified data architecture of Databricks Data Intelligence Platform

Business Continuity

- Recovery time SLA cut **95%+** from 24-48 hours to 0.5-2 hours

Accelerated Time-to-Market

- Reduce complexity and accelerate time to results
- Deliver additional **~12 to 15** data projects year over year to meet business demands

Revenue

- **$1-2M** annual net revenue benefit per each 1% improvement in page optimization

Wednesday

DATA ENGINEERING AND STREAMING

Getting Started with DLT Pipelines

Zoe Durand, Senior Product Manager, Databricks
Maarten de Haas, Product Architect, The HEINEKEN Company

Delta Live Tables (DLT) is a data pipelining framework for the Databricks Data Intelligence Platform that helps data teams simplify streaming and batch ETL cost-effectively, automating things like task orchestration, cluster management,...

Wednesday, Jun 12 | 5:10 PM - 5:50 PM PDT

RESERVE MY SEAT

DATA ENGINEERING AND STREAMING

What's New in Databricks Workflows - with Live Demos!

Saad Ansari, Product Management, Databricks
Roland Fäustlin, Product Manager, Databricks

Join the Databricks Workflows product team to learn about the latest capabilities of Databricks Workflows, the unified orchestrator for data, analytics, and AI on the Data Intelligence Platform. In this session, we'll show how you can easily...

/ Almost Full */*

Wednesday, Jun 12 | 11:20 AM - 12:00 PM PDT

RESERVE MY SEAT

DATA ENGINEERING AND STREAMING

Delta Live Tables in Depth: Best Practices for Intelligent Data Pipelines

Michael Armbrust, Distinguished Engineer, Databricks
Paul Lappas, Sr Staff Product Manager, Databricks

Join Databricks Distinguished Engineer Michael Armbrust and Product Lead Paul Lappas for the authoritative guide to Delta Live Tables (DLT). Learn what's new; what's coming (spoiler alert - some BIG news); and how to easily master the...

Wednesday, Jun 12 | 2:50 PM - 4:20 PM PDT

FULL

DATA ENGINEERING AND STREAMING

Databricks Streaming: Project Lightspeed Goes Hyperspeed

Karthik Ramasamy, Head of Streaming, Databricks
Ryan Nienhuis, Sr. Staff Product Manager, Databricks

For two years, Project Lightspeed has focused on next-gen streaming data processing with Spark Structured Streaming. This year, we've made some incredible strides in ultra low-latency processing.

Join us for a deep dive into...

Wednesday, Jun 12 | 4:00 PM - 4:40 PM PDT

RESERVE MY SEAT

Thursday

Data + AI Summit Keynote, Thursday

Alexander Booth, Asst Director of Research & Development, Texas Rangers
Ali Ghodsi, Co-Founder and CEO, Databricks
Matei Zaharia, Chief Technology Officer and Co-Founder, Databricks
Michael Armbrust, Distinguished Engineer, Databricks
Yejin Choi, Sr Research Director Commonsense AI, AI2, University of Washington
Zaheera Valani, VP, Engineering, Databricks

As the pioneer of lakehouse architecture, Databricks continues to set the vision for where this category is headed. This keynote will explore the foundations of the lakehouse with a deep dive into what's next for the Delta Lake community and...

Thursday, Jun 13 | 9:00 AM - 10:50 AM PDT

✓ RESERVED

DATA ENGINEERING AND STREAMING

Simplified Ingestion for the Data Intelligence Platform

Elise Georis, Staff Product Manager, Databricks
Peter Pogorski, Senior Product Manager, Databricks

To get value from data, practitioners must ingest data from different source databases and applications. Often, data movement from different sources often requires specialized knowledge of each source system. The Databricks Data Intelligence...

Thursday, Jun 13 | 11:20 AM - 12:00 PM PDT

RESERVE MY SEAT

DATA+AI SUMMIT

Thank you for joining us today.

Please complete your session survey!