

LAKEFLOW CONNECT

Introducing Databricks' Native Ingestion Connectors

Elise Georis & Peter Pogorski
June 2024



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.

ANNOUNCING

LakeFlow

One data engineering solution
powered by data intelligence

Ingest

Transform

Orchestrate

ANNOUNCING

LakeFlow

One data engineering solution
powered by data intelligence

Ingest

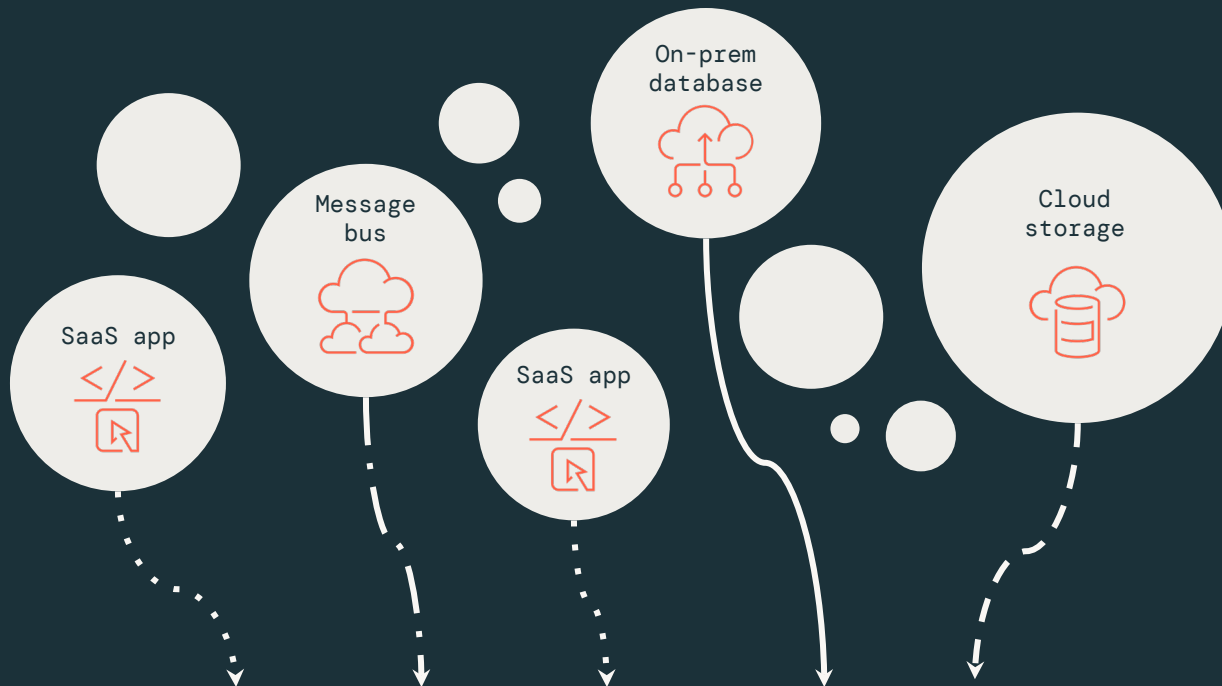
Transform

Orchestrate

AGENDA

1. State of the union
2. Overview of LakeFlow Connect
3. Demo
4. Deep-dive
5. FAQ

STATE OF THE UNION: today's problems



Data platform



TODAY'S PROBLEMS



Inefficiencies in data ingestion

➔ High costs; slow time to value



Dependencies on specialized teams

➔ Low productivity; siloed ownership



Patchwork solutions with limited governance

➔ Underutilized data; security risks

TODAY'S PROBLEMS



Inefficiencies in data ingestion

➔ High costs; slow time to value



Dependencies on specialized teams

➔ **Low productivity; siloed ownership**



Patchwork solutions with limited governance

➔ Underutilized data; security risks

TODAY'S PROBLEMS



Inefficiencies in data ingestion

➔ High costs; slow time to value



Dependencies on specialized teams

➔ Low productivity; siloed ownership

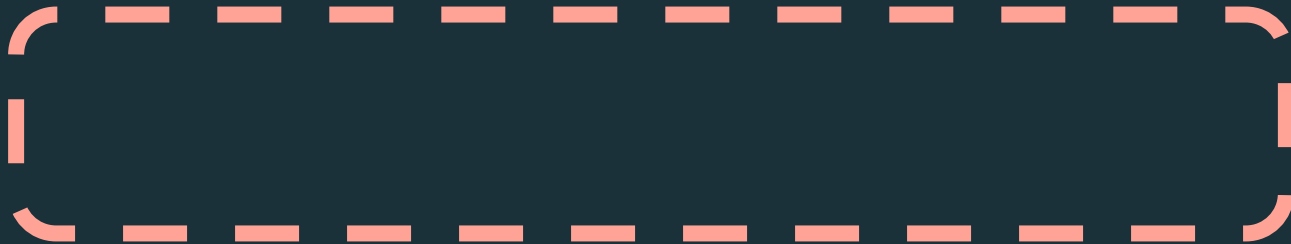


Patchwork solutions with limited governance

➔ **Underutilized data; security risks**

STATE OF THE UNION: today's solutions

Structured Streaming



Structured Streaming

Delta Live Tables

Structured Streaming



Delta Live Tables

Structured Streaming

LakeFlow Connect

Delta Live Tables

Structured Streaming

INTRODUCING LAKEFLOW CONNECT:

efficient data ingestion for everyone

LAKEFLOW CONNECT



Simple and low-maintenance



Unified with the lakehouse



Efficient end-to-end

LAKEFLOW CONNECT



Simple and low-maintenance → Fewer headaches, quicker time to value, democratized data



Unified with the lakehouse



Efficient end-to-end

LAKEFLOW CONNECT



Simple and low-maintenance →

- Schema evolution
- Observability and alerts
- Retries and error handling
- Schema mapping
- Data sampling
- SCD type 2
- Simple UI and API
- ...



Unified with the lakehouse



Efficient end-to-end

LAKEFLOW CONNECT



Simple and low-maintenance



Unified with the lakehouse



Secure and healthy pipelines that live where you do your work



Efficient end-to-end

LAKEFLOW CONNECT



Simple and low-maintenance



Unified with the lakehouse



- Unity Catalog
- Workflows
- RAG Studio
- Single interface for pipelines
- Single account for ingestion
- ...



Efficient end-to-end

LAKEFLOW CONNECT



Simple and low-maintenance



Unified with the lakehouse



Efficient end-to-end

➔ Lower costs, better performance,
better scalability

LAKEFLOW CONNECT



Simple and low-maintenance



Unified with the lakehouse



Efficient end-to-end



- Incremental reads
- Incremental writes
- Incremental transformations
- ...

Enzyme incrementally refreshes materialized views.



Delta-tracked
changes



Query plan
analysis



Monotonic append

Partition recompute

MERGE updates

Full recompute

Cost model



Optimal update
technique



ROADMAP

SUBJECT TO CHANGE, BASED ON YOUR FEEDBACK

Applications





A grid of application logos with their release status. The first row contains Salesforce (Private preview), Workday (Private preview), and Oracle NetSuite (Coming soon). The second row contains ServiceNow (Coming soon), Google Analytics (Coming soon), and SharePoint (Coming soon). The third row contains Meta, Google Ads, and Dynamics 365, followed by an ellipsis.

 Private preview	 Private preview	 Coming soon
 Coming soon	 Coming soon	 Coming soon
		 ...

Databases



A grid of database logos with their release status. The first row contains Microsoft SQL Server (Private preview) and PostgreSQL (Coming soon). The second row contains Oracle Database, MySQL, and IBM DB2. The third row contains MongoDB and Amazon DynamoDB, followed by an ellipsis.

 Private preview	 Coming soon	
		
	 ...	

DEMO



SETTING THE SCENE



I'm a data engineer at a car company.



My data lives in several places, including Salesforce and SQL Server.



My data quality varies.

GOALS



High-quality data to help my organization make informed decisions

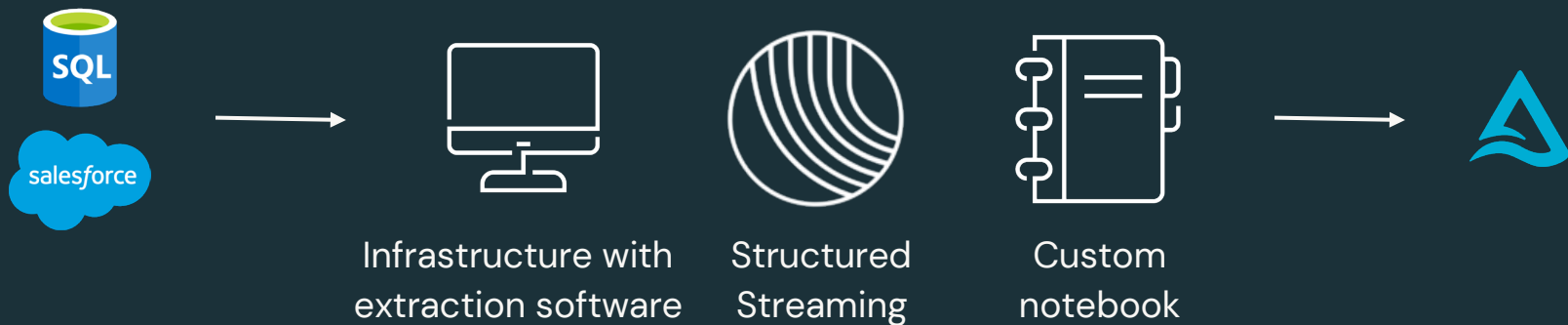


Dashboards to share insights with my stakeholders

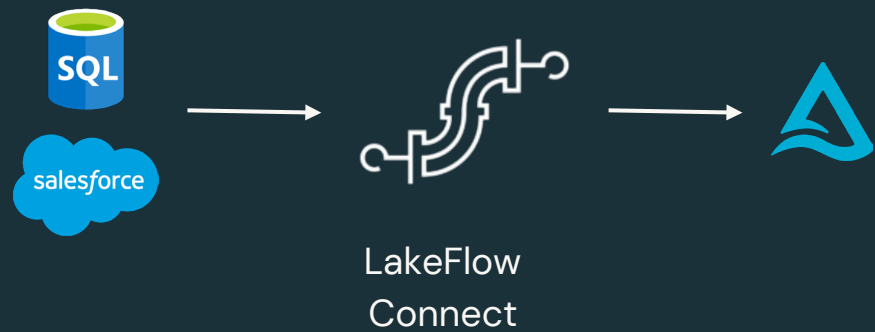


AI to help the sales team generate sales plans

BEFORE



AFTER



+ New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Genie

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Pipelines

Machine Learning

Playground

Experiments

Features

Models

Serving

Marketplace

Partner Connect

Add data

Get started by connecting to a data source or uploading a local file.

Databricks connectors



Salesforce Sales Cloud



Workday Reports



Azure SQL Database



AWS RDS for SQL Server



Amazon S3

File upload



Create or modify table

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



Upload files to a volume

Add files in any format to a non-tabular dataset managed in Unity Catalog

Fivetran connectors [See all available ingest partners in Partner Connect](#)



OneDrive



Google Drive



Jira



GitHub



Webhooks

[See all >](#)

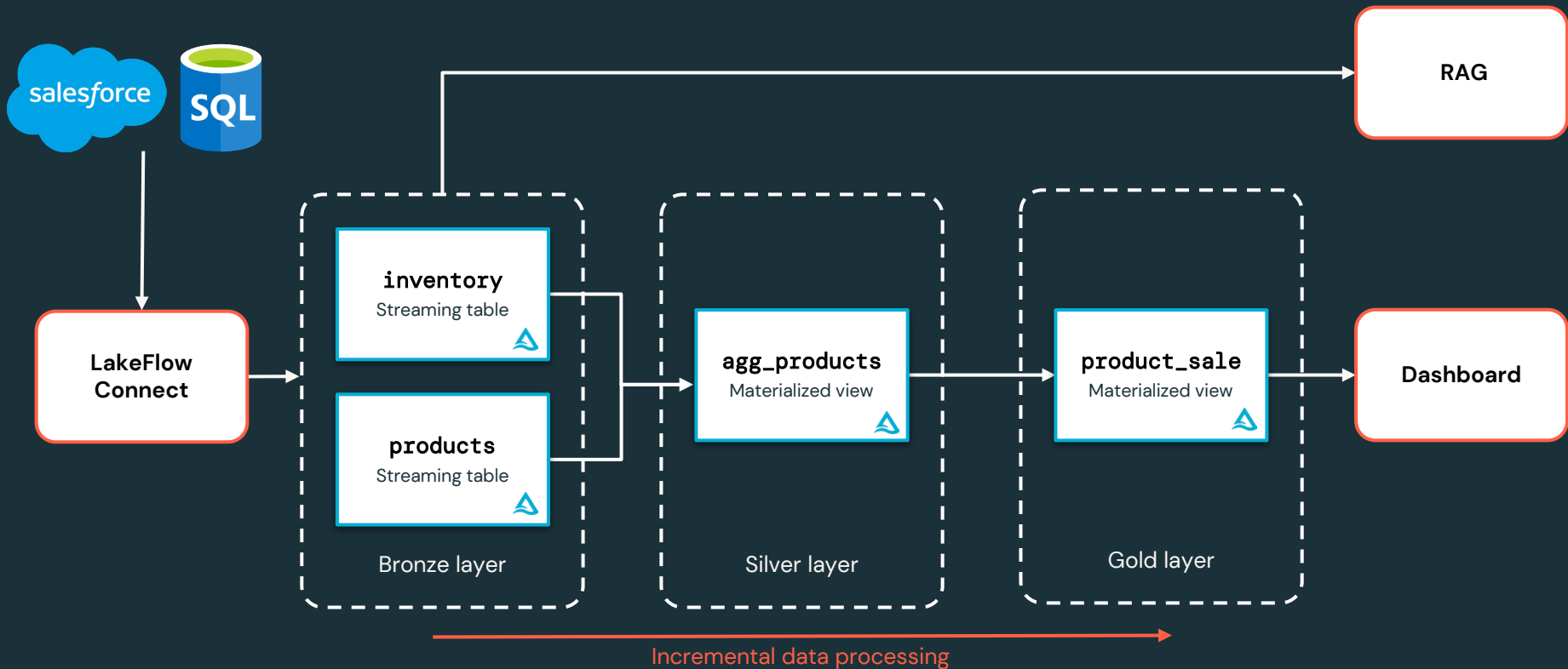
Legacy file upload



Upload files to DBFS

This page is maintained for backwards compatibility. We recommend uploading files to a volume.





Unity Catalog

+ New

- Workspace
- Recents
- Catalog
- Workflows
- Compute
- SQL
- SQL Editor
- Queries
- Dashboards
- Alerts
- Query History
- SQL Warehouses
- Genie Spaces

- Data Engineering
- Job Runs
- Pipelines
- Machine Learning
- Playground
- Experiments
- Features
- Models
- Serving

Workflows > Delta Live Tables >

speedy_motors_etl

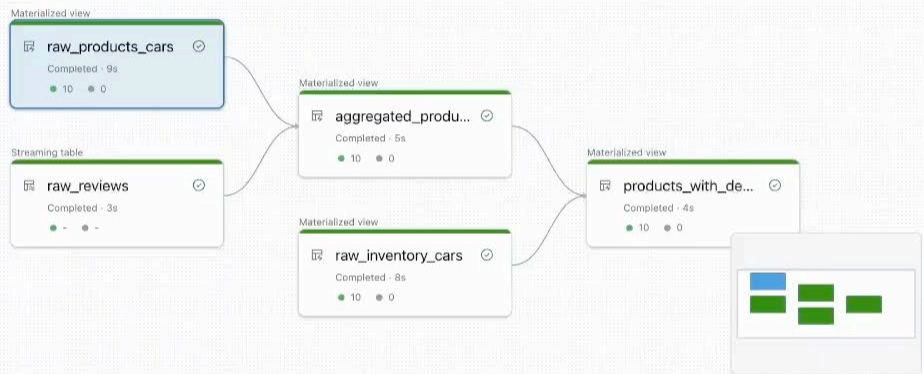
Provide feedback

Development Production Settings Schedule (1) Start

6/2/2024, 1:09:55 PM · Completed

Select tables for refresh

Graph List



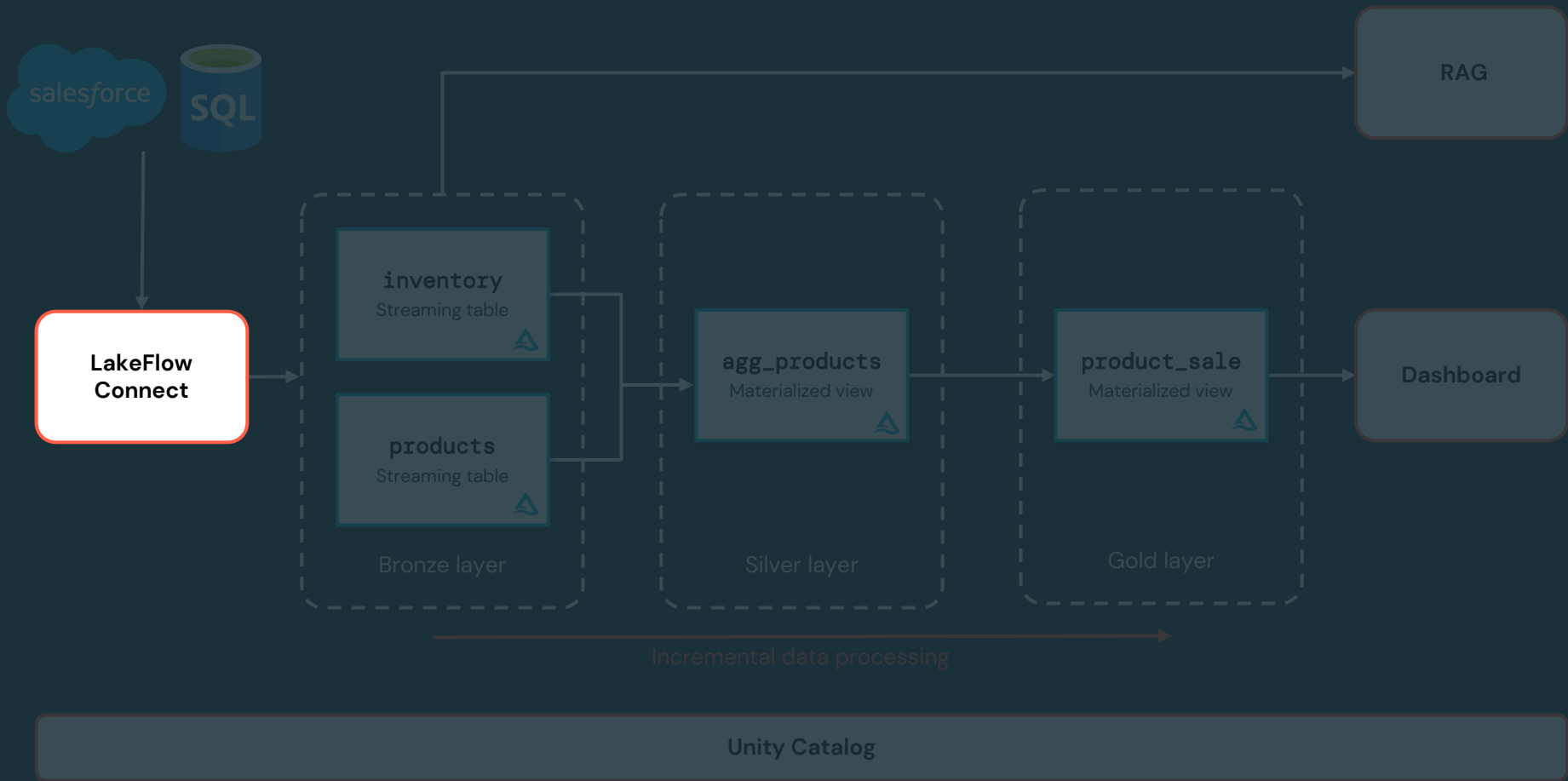
raw_products_cars

Details Data quality Schema Flows

Id: string
Name: string
ProductCode: string
Description: string
IsActive: boolean
CreatedDate: timestamp
CreatedById: string
LastModifiedDate: timestamp
LastModifiedById: string
SystemModstamp: timestamp
Family: string
ExternalDataSourceId: string
ExternalId: string

All Info Warning Error Filter...

3 hours ago	flow_progress	Flow 'products_with_dealership_info' is STARTING.
3 hours ago	flow_progress	Flow 'products_with_dealership_info' is RUNNING.
3 hours ago	flow_progress	Flow 'products_with_dealership_info' has COMPLETED.
3 hours ago	memory_utilization	Collected memory utilization on the cluster during termination
3 hours ago	update_progress	Update cc3da0 is COMPLETED.



WHAT IS A CONNECTOR?

UC connection

to store credentials securely

DLT pipeline

to ingest data efficiently

Workflows DAG

to orchestrate your ETL

Unity Catalog

for unified security, governance, cataloging, and lineage

Delta Lake

for reliable data storage that's externally accessible

WHAT IS A CONNECTOR?

UC connection
to store credentials securely

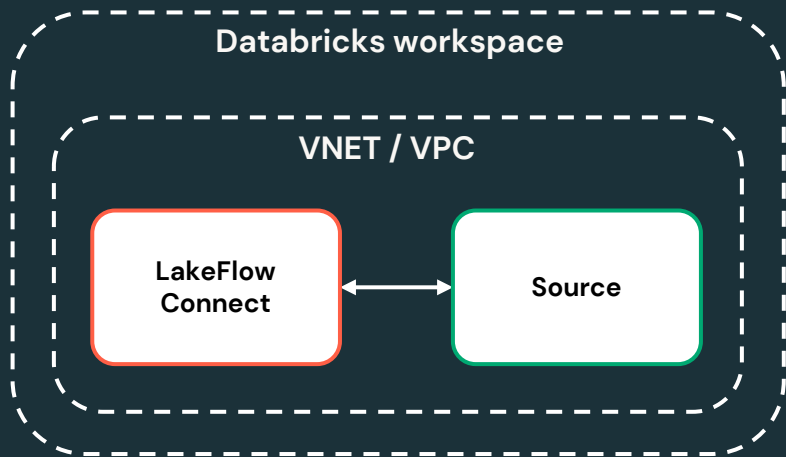
DLT pipeline
to ingest data efficiently

Workflows DAG
to orchestrate your ETL

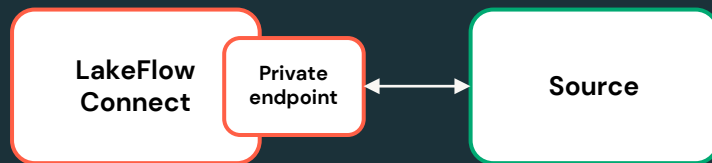
Unity Catalog
for unified security, governance, cataloging, and lineage

Delta Lake
for reliable data storage that's externally accessible

Workspace deployed with a virtual network that can access the database



Private Link



WHAT IS A CONNECTOR?

UC connection
to store credentials securely

DLT pipeline
to ingest data efficiently

Workflows DAG
to orchestrate your ETL

Unity Catalog
for unified security, governance, cataloging, and lineage

Delta Lake
for reliable data storage that's externally accessible

Adding a pipeline schedule

Production | Search | ajay.alfred@databricks.com

Ingest data from Salesforce Sales Cloud [Ingestion pipeline]

STEP 5

Settings
Configure the pipeline

Pipeline schedule

On-demand

Schedule

Refresh every: 24 Hours

Timezone: (UTC-07:00) Pacific Time (US and Canada)

Show cron syntax

Notifications: + Add notification

Cancel Previous Save pipeline Save and run pipeline



creates a job with a pipeline task.

Workflows > Jobs > | Provide feedback

Salesforce pipeline ☆

Runs Tasks

run_salesforce_ingestion_pipeline
Salesforce accounts -> prod.sfdc.accounts...

Task name* run_salesforce_ingestion_pipeline

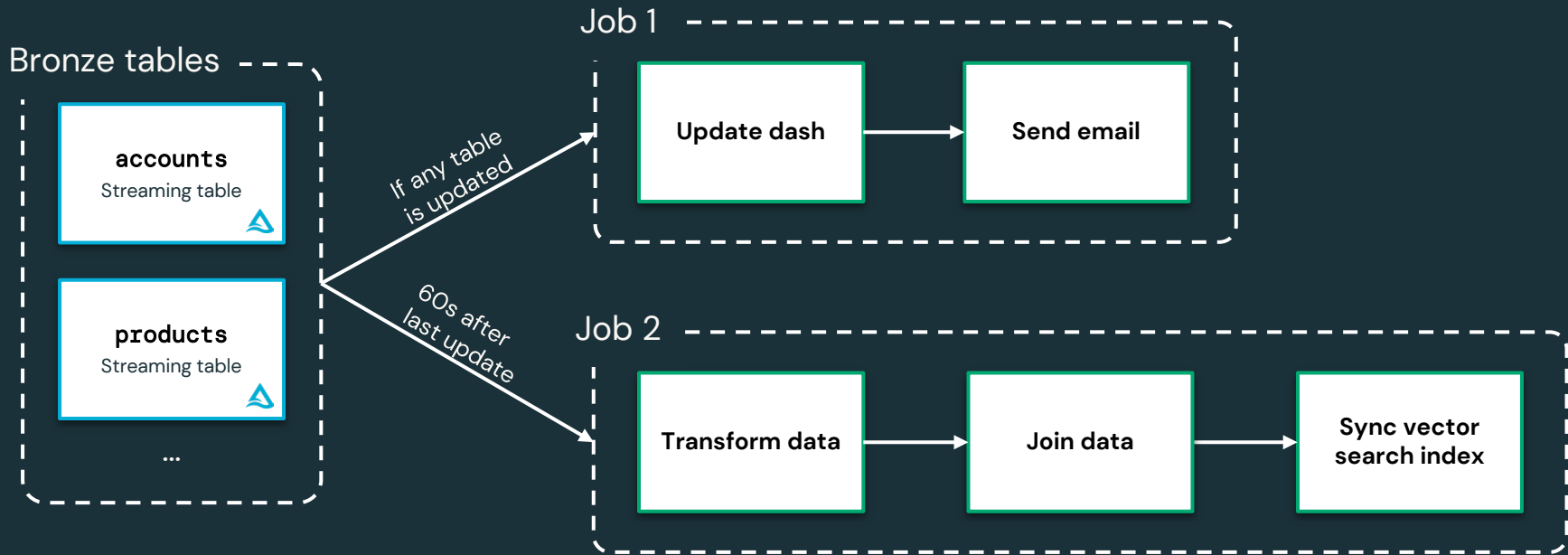
Type* Ingestion pipeline

Pipeline* Salesforce accounts -> prod.sfdc.accounts 2024-01-05 09:17:01

Notifications elise.georis@databricks.com
On failure

Edit notifications

Cancel Create task



BENCHMARKING

# objects	Total # rows	Observed latency (min)
100	14M	~14
250	21M	~35
100	194M	~100

* Not a formal benchmark or guarantee. Performance varies based on setup.

FREQUENTLY ASKED QUESTIONS

FAQ



What's the relationship with Arcion?

FAQ



What's the relationship with Arcion?



...with Auto Loader?

LakeFlow Connect

Roadmap

Delta Live Tables

✓ Available

Structured Streaming

✓ Available



FAQ



What's the relationship with Arcion?



...with Auto Loader?



...with Databricks' ingestion partners?

FAQ



What's the relationship with Arcion?



...with Auto Loader?



...with Databricks' ingestion partners?



...with Lakehouse Federation and Delta Sharing?

LAKEFLOW CONNECT: efficient data ingestion for everyone



1. Simple
2. Unified
3. Efficient

RECOMMENDED SESSIONS

Attend today or view online

Session	Date, Time
Your Guide to Data Engineering on the Data Intelligence Platform	Tues, 6/11, 9:00 AM
Delta Live Tables in Depth: Best Practices for Intelligent Data Pipelines	Wed, 6/12, 2:50 PM
Databricks Streaming: Project Lightspeed Goes Hyperspeed	Wed, 6/12, 4:00 PM
Getting Started with DLT Pipelines	Wed, 6/12, 5:10 PM
Streaming Data Pipelines: From Supernovas to LLMs	Thurs, 6/13, 12:30 PM
Introducing the New Python Data Source API for Apache Spark	Thurs, 6/13, 2:50 PM

Join waitlist



[tinyurl.com/
connector-waitlist](https://tinyurl.com/connector-waitlist)

Vote for sources



[tinyurl.com/
connector-survey](https://tinyurl.com/connector-survey)



Learn more at the summit!



Databricks
Events App



Tells us what you think

- We kindly request your valuable feedback on this session.
- Please take a moment to rate and share your thoughts about it.
- You can conveniently provide your feedback and rating through the **Mobile App**.



What to do next?

- Discover more related sessions in the mobile app!
- Visit the Demo Booth: Experience innovation firsthand!
- More Activities: Engage and connect further at the Databricks Zone!



Get trained and certified

- Visit the Learning Hub Experience at **Moscone West, 2nd Floor!**
- Take complimentary certification at the event; come by the Certified Lounge
- Visit our Databricks Learning website for more training, courses and workshops! databricks.com/learn



