databricks

Ingesting Enterprise Data with LakeFlow Connect

LECTURE

Ingesting Enterprise Data Overview
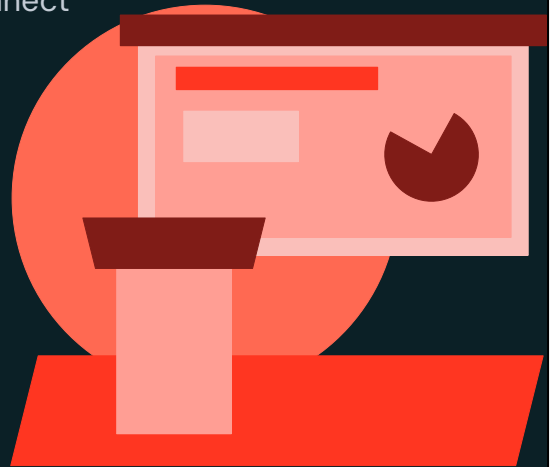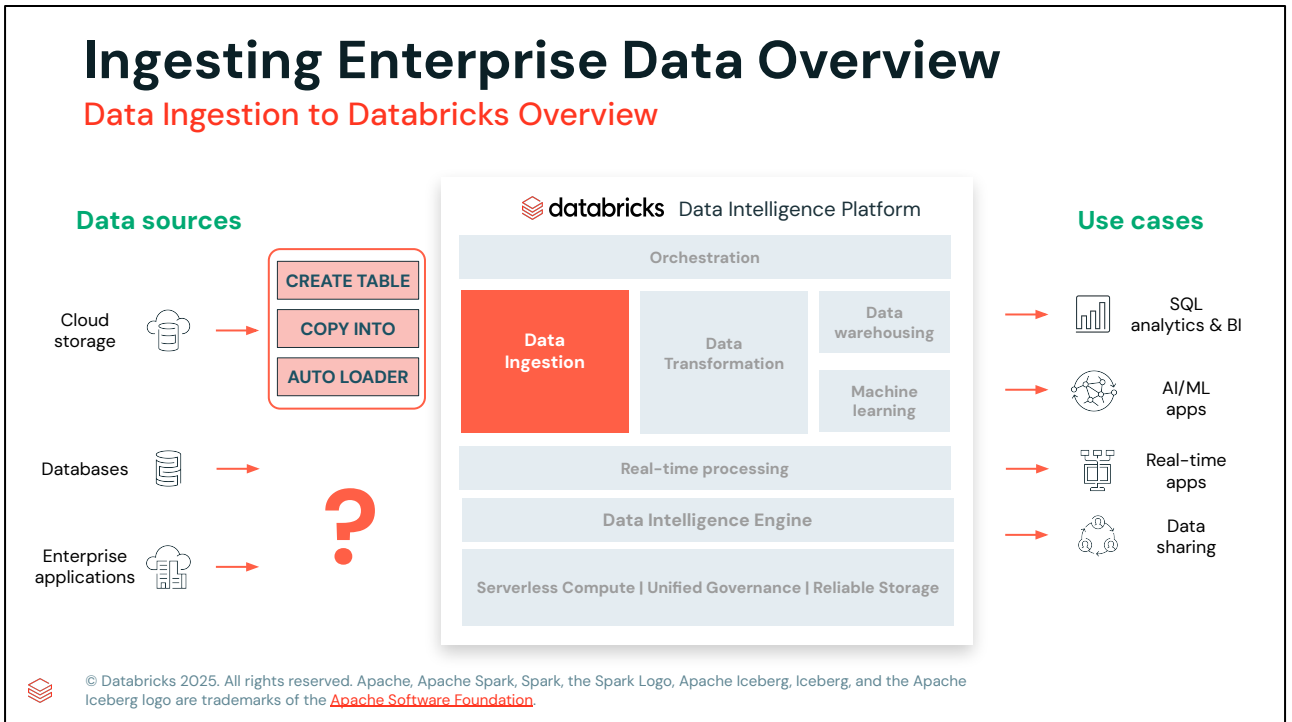
Ingesting enterprise data is streamlined with Lakeflow Connect Managed Connectors and Partner Connect, enabling fast, reliable integration from databases and applications into the Databricks Lakehouse—with flexible, fully managed, and partner-supported options.

Ingesting Enterprise Data Overview

Data Ingestion to Databricks Overview

Data sources

Cloud storage

Databases

Enterprise applications

?

CREATE TABLE

COPY INTO

AUTO LOADER

databricks Data Intelligence Platform

Orchestration

Data Ingestion

Data Transformation

Data warehousing

Machine learning

Real-time processing

Data Intelligence Engine

Serverless Compute | Unified Governance | Reliable Storage

Use cases

SQL analytics & BI

AI/ML apps

Real-time apps

Data sharing

So far, we've discussed ingesting data from cloud object storage into Databricks using techniques like CREATE TABLE, COPY INTO, and Auto Loader.

But what about ingesting data from databases or enterprise applications?

What techniques can we use to handle those sources?

Ingesting Enterprise Data Overview

Lakeflow Connect Managed Connectors

Lakeflow Connect
Managed Connectors

databricks

BENEFITS

- **Simplify** the process of ingestion data from a wide variety of enterprise databases and applications
- Provide a easy to use **user interface (UI)** (or you can use the API)
- **Fully managed by Databricks**, reducing the need for manual configuration or custom code

The first method for ingesting enterprise data is by using Lakeflow Connect Managed Connectors.

These connectors are built into Databricks and are designed to simplify the process of ingesting data from a wide variety of enterprise databases and applications.

They provide a low-code, fully managed experience, reducing the need for manual configuration or custom integration code.

Benefits summary:

- Simplify the process of ingestion data from a wide variety of enterprise databases and applications

- Provide a easy to use user interface (UI) (or you can use the API)

- Fully managed by Databricks, reducing the need for manual configuration or custom code

Ingesting Enterprise Data Overview
Data Ingestion with Lakeflow Connect Managed Connectors

Lakeflow Connect

Managed Connectors

"Point and click" or API

databricks
Data Intelligence Platform
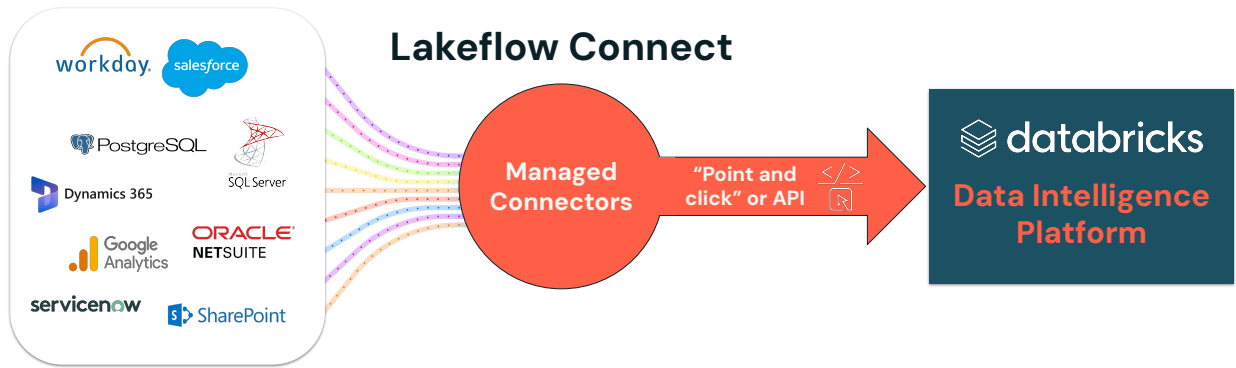
Efficient Databricks Managed Connectors for Ingestion into your Lakehouse
Managed connectors in Lakeflow Connect are in various release states

With Lakeflow Connect managed connectors, you can easily begin ingesting enterprise data from sources like Workday, Salesforce, PostgreSQL, SQL Server, and more.

These are highly efficient, Databricks-managed connectors designed specifically for fast, reliable ingestion into your Lakehouse.

NOTE: Managed connectors in Lakeflow Connect are in various release states. Some may be in public preview, others in GA. Be sure to check the official documentation for the latest details: https://docs.databricks.com/aws/en/release-notes/release-types

Ingesting Enterprise Data Overview

Lakeflow Connect Managed Connectors: SaaS Ingestion

Lakeflow Connect collects data from **external sources to Streaming Delta Tables** using a **serverless compute Declarative Pipelines pipeline**:

1. A Lakeflow Serverless Declarative Pipelines job **collects credentials** from Unity Catalog.
2. The job **reaches out** to the publicly accessible data source (e.g., API, open OLAP port, etc.).
3. The service transforms the data and stores it to a **Streaming Delta Table**.
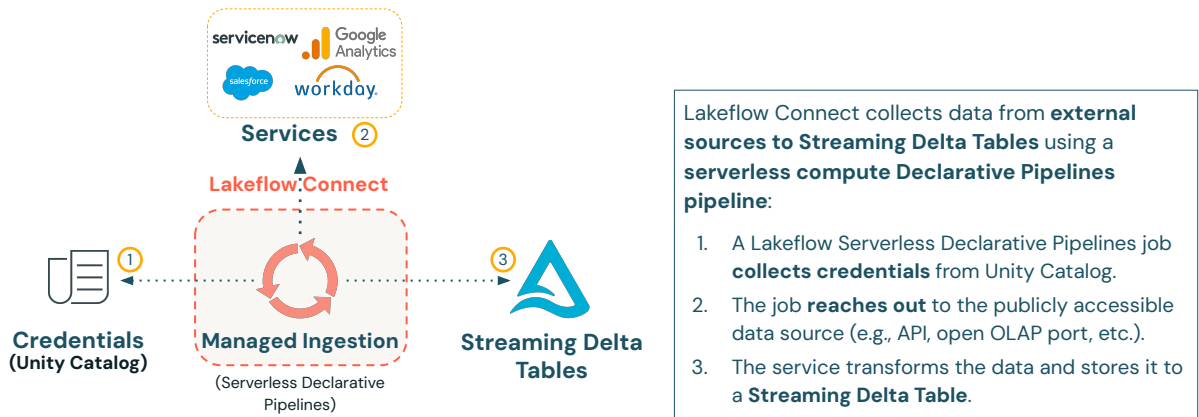
We're starting with an overview of the Lakeflow Connect managed connector architecture for SaaS applications.

Lakeflow Connect enables data ingestion from external, publicly accessible sources such as APIs or OLAP endpoints into Streaming Delta Tables, using serverless, declarative pipelines. You can setup these pipelines using the user interface (UI) or the API. Managed connectors leverage efficient incremental reads and writes to make data ingestion faster, scalable, and more cost-efficient, while your data remains fresh for downstream consumption

Here's how it works:

1. A Lakeflow Serverless Declarative Pipelines job collects credentials from Unity Catalog.

2. The job reaches out to the publicly accessible data source (e.g., API, open OLAP port, etc.).

3. The service transforms the data and stores it to a Streaming Delta Table.

Ingesting Enterprise Data Overview

Lakeflow Connect Managed Connectors: SaaS Ingestion

Services ②

Lakeflow Connect

Credentials
(Unity Catalog) ①

Managed Ingestion ←→
(Serverless Declarative Pipelines)

Streaming Delta Tables ③

Lakeflow Connect collects data from **external sources to Streaming Delta Tables** using a **serverless compute Declarative Pipelines pipeline**:

1. A Lakeflow Serverless Declarative Pipelines job **collects credentials** from Unity Catalog.
2. The job **reaches out** to the publicly accessible data source (e.g., API, open OLAP port, etc.).
3. The service transforms the data and stores it to a **Streaming Delta Table**.
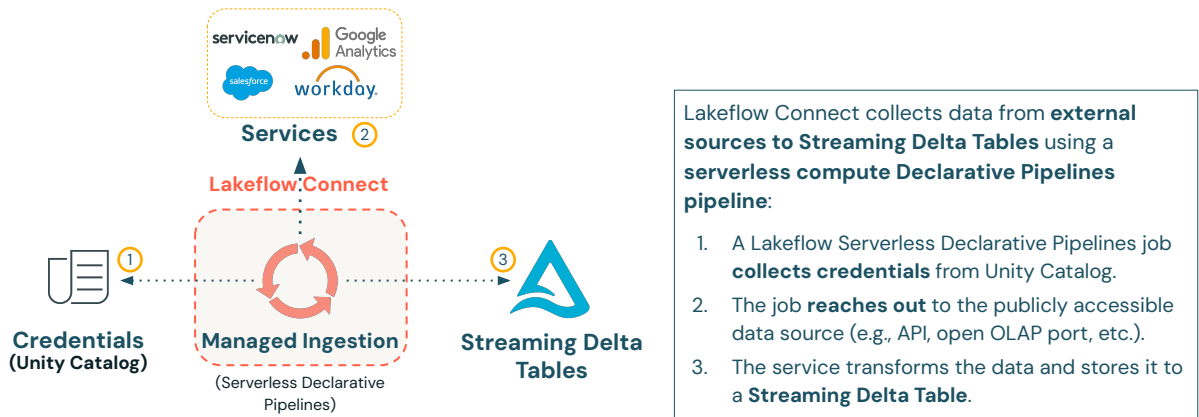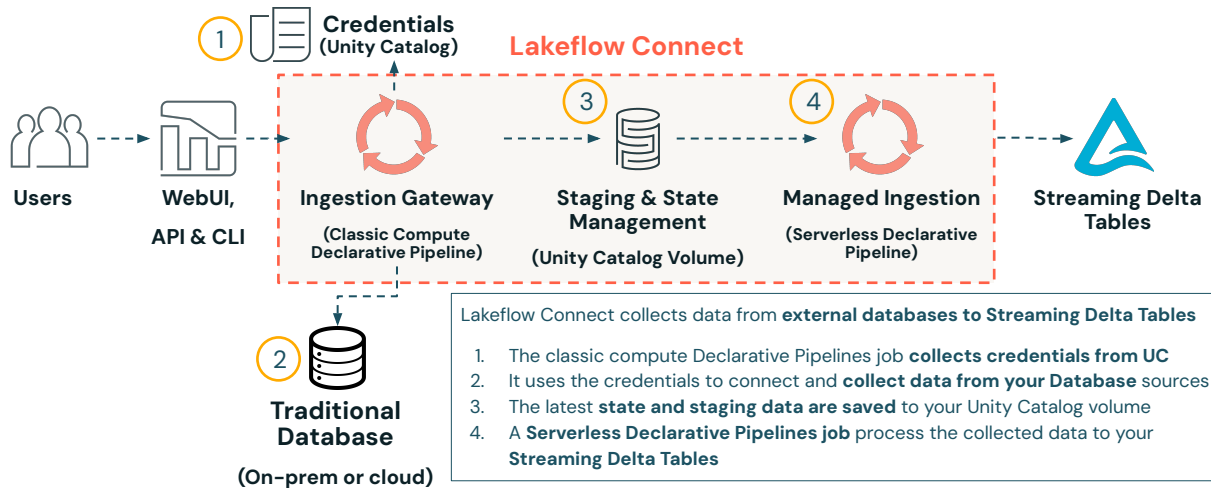
To support this flow, we've introduced a new pipeline type: the managed ingestion pipeline. Its primary role is to connect to public SaaS sources (like Salesforce or Workday), extract the data, and ingest it directly into a streaming table. These pipelines are largely predefined and managed by Databricks, enabling seamless handling of source-specific complexities.

Finally, for SaaS connectors, all data movement happens in the data plane. The control plane is only used for pipeline setup, monitoring (e.g., reading event logs), and management.

For more information view the SaaS managed connector components documentation: https://docs.databricks.com/aws/en/ingestion/lakeflow-connect/#saas-connector-components

Ingesting Enterprise Data Overview

Lakeflow Connect Managed Connectors: Database Ingestion

Credentials (Unity Catalog)

Lakeflow Connect

Users → WebUI, API & CLI → Ingestion Gateway (Classic Compute Declarative Pipeline) → Staging & State Management (Unity Catalog Volume) → Managed Ingestion (Serverless Declarative Pipeline) → Streaming Delta Tables

Traditional Database (On-prem or cloud)

Lakeflow Connect collects data from **external databases to Streaming Delta Tables**
1. The classic compute Declarative Pipelines job **collects credentials from UC**
2. It uses the credentials to connect and **collect data from your Database** sources
3. The latest **state and staging data are saved** to your Unity Catalog volume
4. A **Serverless Declarative Pipelines job** process the collected data to your **Streaming Delta Tables**

Now let's look at how the architecture changes when using a Lakeflow Connect managed database connector.

Like with SaaS connectors, this architecture is designed to move data into Streaming Delta Tables—but this time from external databases rather than public APIs.

How it works:

1. A classic compute Declarative Pipelines job retrieves credentials securely from Unity Catalog.

2. It uses those credentials to connect to your external database sources.

3. The job collects the latest state and change logs, storing staged data in a Unity Catalog volume.

4. A Serverless Declarative Pipelines job then processes that staged data and loads it into a Streaming Delta Table.

Ingesting Enterprise Data Overview
Lakeflow Connect Managed Connectors: Database Ingestion

Lakeflow Connect collects data from **external databases to Streaming Delta Tables**
1. The classic compute Declarative Pipelines job **collects credentials from UC**
2. It uses the credentials to connect and **collect data from your Database** sources
3. The latest **state and staging data are saved** to your Unity Catalog volume
4. A **Serverless Declarative Pipelines job** process the collected data to your **Streaming Delta Tables**
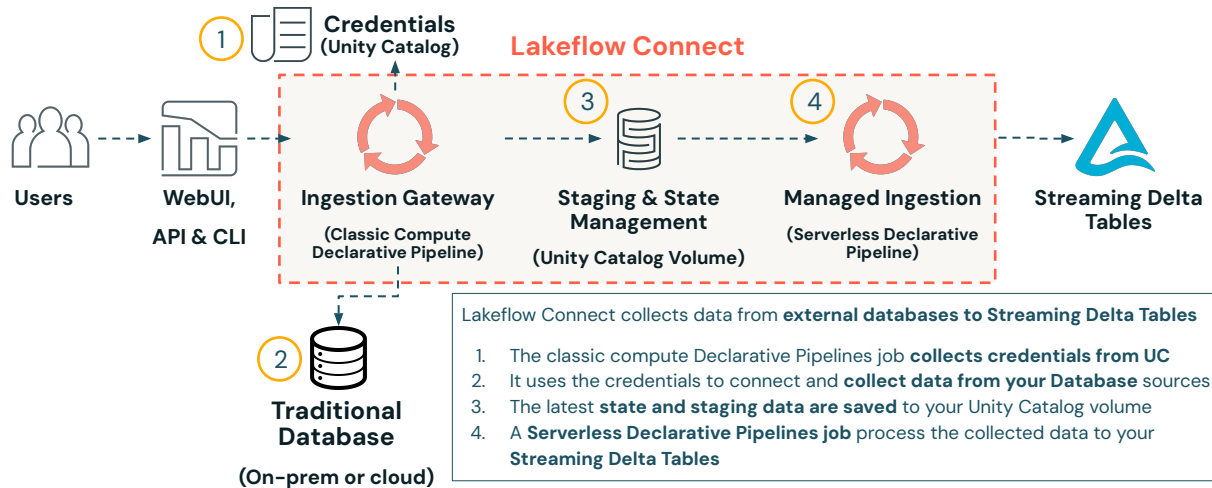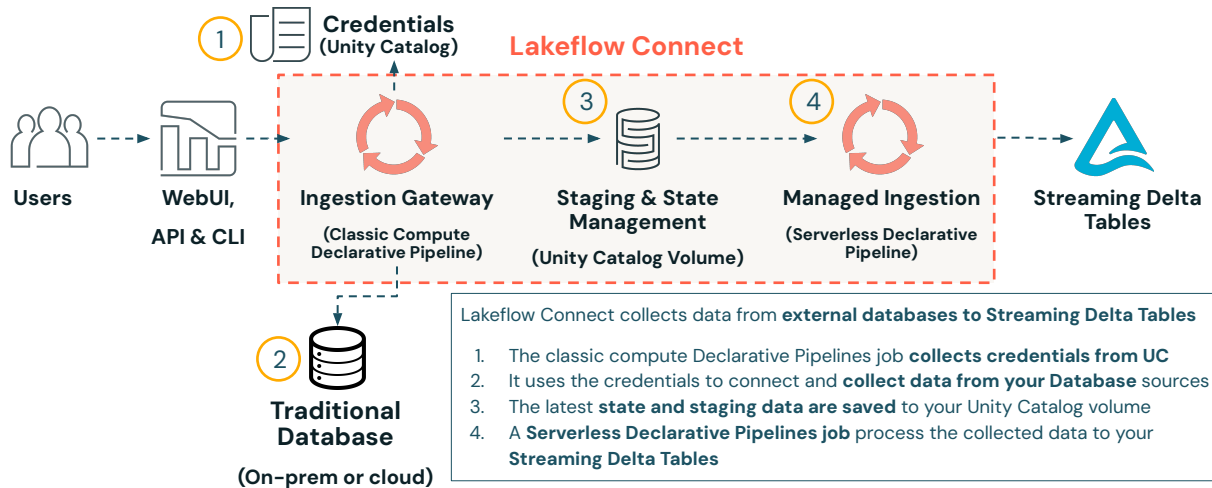
Key Components Introduced:

We're introducing two new architectural elements here:

- Ingestion Gateway:

  - A dedicated pipeline that connects to the database to extract:

    - Metadata
    - Snapshots
    - Change logs - It stages all of this in the Unity Catalog (UC) volume.

- Unity Catalog Volume:

  - This acts as the intermediate staging layer, enabling the next pipeline to pick up and stream data.

  - It's secured using standard UC mechanics, and by default, access is limited to the user running the pipeline.

Ingesting Enterprise Data Overview

Lakeflow Connect Managed Connectors: Database Ingestion

① Credentials (Unity Catalog)

Lakeflow Connect

Users

WebUI, API & CLI

Ingestion Gateway (Classic Compute Declarative Pipeline)

③ Staging & State Management (Unity Catalog Volume)

④ Managed Ingestion (Serverless Declarative Pipeline)

Streaming Delta Tables

② Traditional Database (On-prem or cloud)

Lakeflow Connect collects data from **external databases to Streaming Delta Tables**

1. The classic compute Declarative Pipelines job **collects credentials from UC**
2. It uses the credentials to connect and **collect data from your Database** sources
3. The latest **state and staging data are saved** to your Unity Catalog volume
4. A **Serverless Declarative Pipelines job** process the collected data to your **Streaming Delta Tables**

So why did we even add this gateway step?

For starters, it helps with networking. Many customers have databases that are [1] sitting behind a firewall, [2] that aren't publicly accessible to the internet, and so on. If you don't have an option like Private Link, we can deploy the gateway inside of your network.

The gateway also helps us limit the load that we're placing on the database. After all, you probably want to limit the number of direct connections to your data source. By splitting out the gateway from the pipeline, we'll be able to have one gateway speak to the database — and then fan out to N pipelines for that scalability. (For context, this isn't a problem in a SaaS connector because the load is typically controlled with API limits. But even there, we're using those limits as efficiently as we can.)

For more information view the database managed connector components documentation:
https://docs.databricks.com/aws/en/ingestion/lakeflow-connect/#database-connector-components

Ingesting Enterprise Data Overview

Data Ingestion with Partner Connect

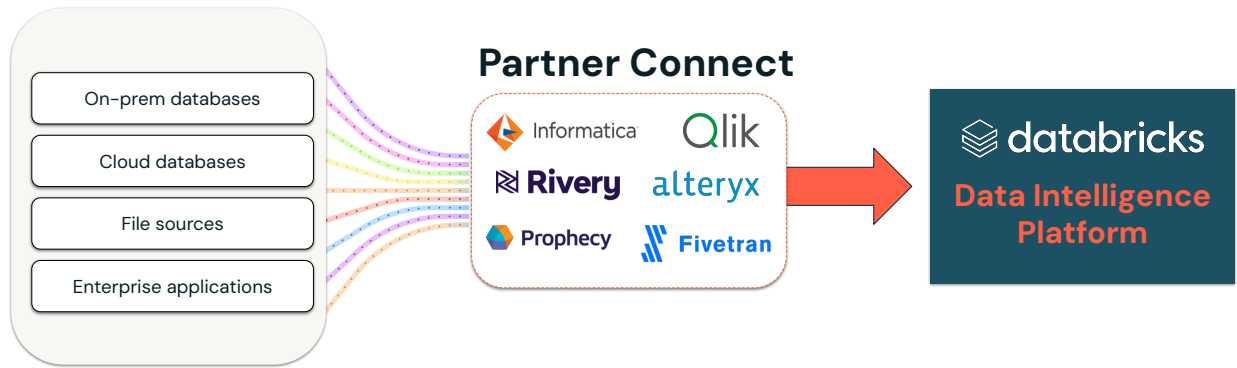Lakeflow Connect
Managed Connectors

Partner Connect

If there is not managed connector available for your specific data source, you can also use Partner Connect. Partner Connect lets you create trial accounts with select Databricks technology partners and connect your Databricks workspace to partner solutions from the Databricks UI. This allows you to try partner solutions using your data in the Databricks lakehouse, then adopt the solutions that best meet your business needs.

Ingesting Enterprise Data Overview
Data Ingestion with Partner Connect

Partner Connect

On-prem databases

Cloud databases

File sources

Enterprise applications

Informatica Qlik Rivery alteryx Prophecy Fivetran

databricks
Data Intelligence Platform

Utilize a rich ecosystem of partner solutions for data ingestion

Our ingestion partners remain vital. They offer a really wide range of connectors, and those connectors tend to have deep functionality that's maintained by dedicated teams of experienced engineers.

And we want you to continue having that choice, so we will continue to work with those partners to offer as many high-quality connectors as possible. Even when there is a native connector for a given source, we'll continue to support that optionality.

We're just now adding the critical connectors that customers have requested. It's about customer choice.

What is Databricks Partner Connect documentation

Partner connect

Thank you for completing this lesson and continuing your journey to develop your skills with us.