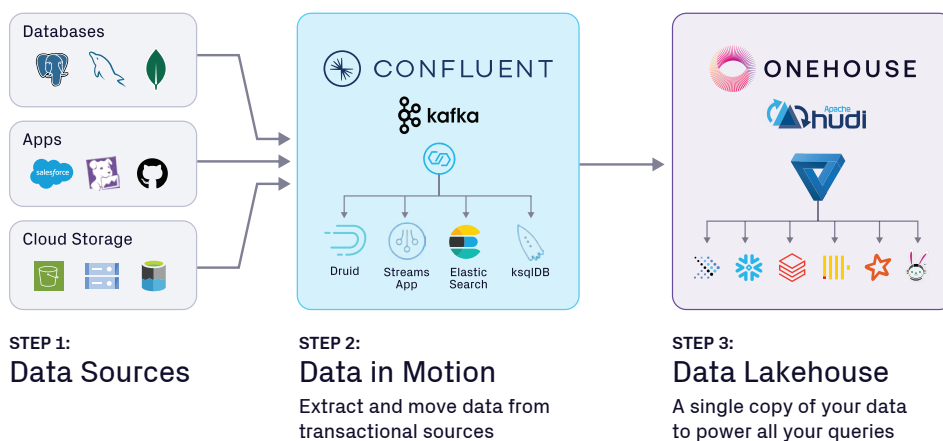# ONEHOUSE + CONFLUENT

# Universal Real-Time Workflows

Consolidate all your streaming and batch data workloads into unified, real-time incremental workflows in minutes with zero infrastructure management.

The universal data lakehouse from Onehouse is based on the industry's only independent, open source data lakehouse platform, providing a trusted single source of truth for all your workloads, with industry-leading native support for streaming data and incremental processing.

Confluent created the foundational platform for data-in-motion. With Confluent Cloud and Onehouse, you can build universal real-time workflows in minutes to power use cases across your entire data estate, including change data capture (CDC), analytics, AI and ML, and more.

## With the universal data lakehouse from Onehouse, you get:

- Near real-time data ingestion and incremental data preparation
- Support for all use cases including vector search and AI, machine learning, BI, and interactive analytics
- Managed reliability and performance optimizations



**STEP 1:**
### Data Sources

**STEP 2:**
### Data in Motion
Extract and move data from transactional sources

**STEP 3:**
### Data Lakehouse
A single copy of your data to power all your queries

It's simple to build real-time pipelines that ingest Apache Kafka topics from Confluent Cloud and then apply transformations with an analytics-ready data lakehouse from Onehouse. Thanks to industry-leading interoperability with Onehouse, your trusted data is optimized and ready to query from a number of popular services, including Snowflake for BI and analytics and Databricks for AI and ML.
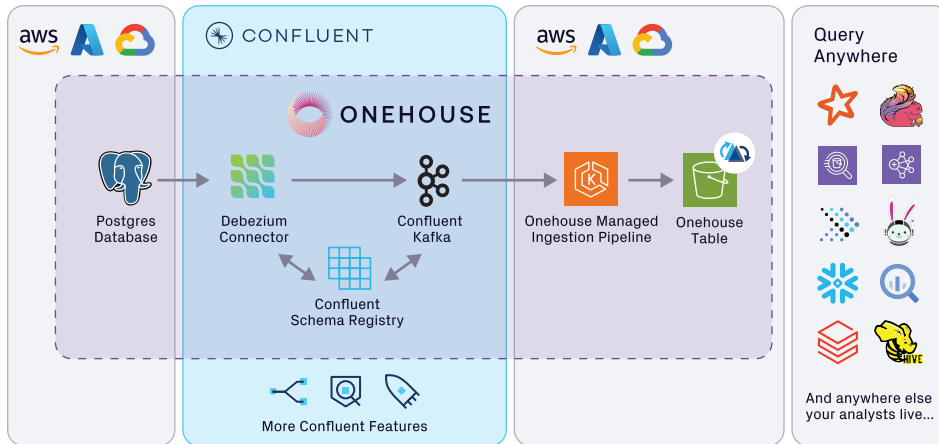
## Seamless Kafka ingestion into the data lakehouse

Onehouse integrates with Confluent Cloud to ingest your messages into a persistent data lakehouse for analytics. With Onehouse, you can:

- Efficiently merge Kafka messages into tables
- Easily handle late-arriving and out-of-order data
- Store data in open, analytics-optimized columnar formats
- Auto-optimize data for efficient queries
- Enable incremental ETL pipelines to deliver maximum data freshness

# Real-time operational database replication via change data capture — open and easy

Organizations often need to analyze data from their operational databases to build dashboards and uncover new insights about their customers and operations. However, relational databases such as PostgreSQL and MySQL are not a good fit for analytics at scale because of challenges with costs, storage limitations, and analytical workload performance.



*In the diagram above, we are replicating a PostgreSQL database to the Onehouse data lakehouse by capturing change logs with Confluent Cloud.*

Onehouse and Confluent Cloud are both designed for streaming data, forming a powerful combination for ingesting CDC data in realtime. As changes happen in your database, you can perform analytics on them live in the data lakehouse.

Onehouse automatically provisions and configures all necessary Confluent platform components, including the Debezium Confluent Connector, schema registry and Kafka topics. Onehouse then manages the pipeline lifecycle, including bootstrapping data, restarting connectors, and updating configurations.

## CDC for a Large-Scale Job Marketplace

A Onehouse customer with large deployments of MySQL has many transactional datasets. With Kafka and Onehouse, they extract changelogs and create low-latency CDC pipelines to build analytics-ready tables on S3.

### Sources

 MySQL      Debezium      Apache Kafka

### Analytics

 AWS Athena