ONEHOUSE

DATA SHEET

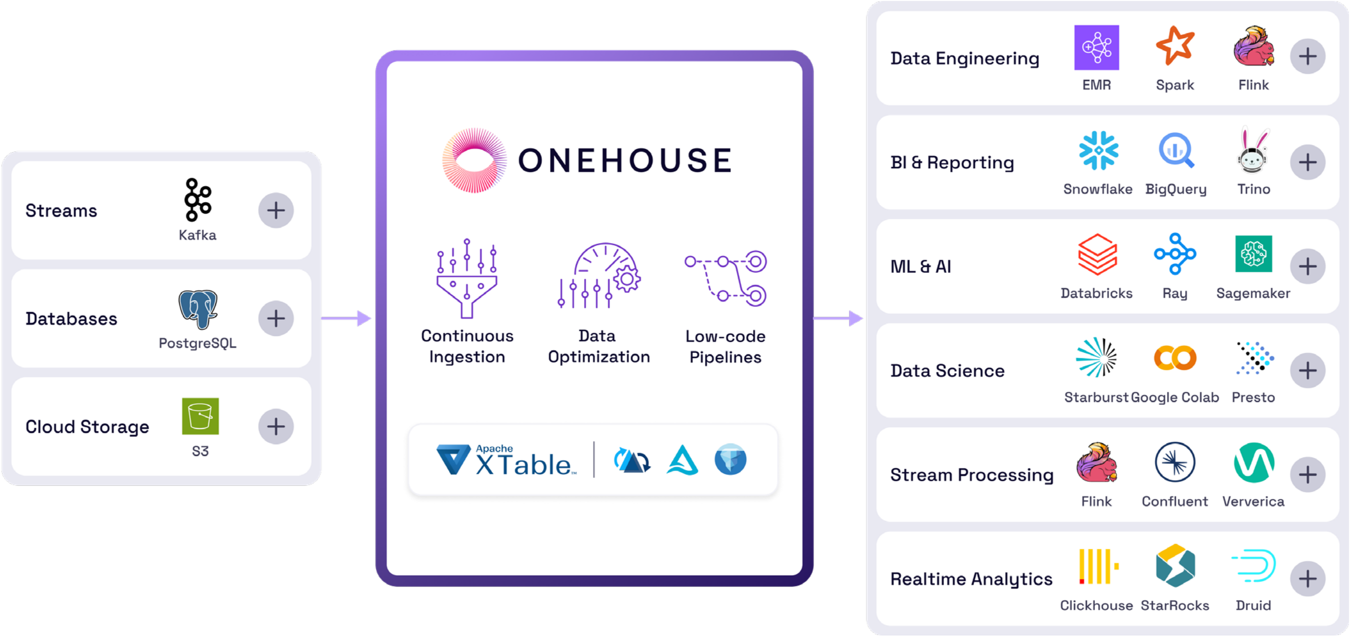
Data Integration Options

Contents

Introduction

Onehouse simplifies data ingestion by merging distinct pipelines for batch and streaming data sources into a unified, incremental pipeline that delivers data continuously. This approach eliminates the complexities of managing separate systems, offering a streamlined solution for real-time data integration.

Onehouse offers a broad range of connectors and integrates with partner platforms and open-source technologies, allowing you to connect to virtually any data source. Whether leveraging data from Amazon S3, Google Cloud Storage, Apache Kafka, or Confluent Cloud Kafka, Onehouse lets you choose from a diverse set of options to tailor data ingestion to your specific needs.



Onehouse Native Connectors

AWS S3

With the Onehouse Amazon S3 connector, ingest data from any S3 bucket. S3 files can be in a number of formats including Avro, JSON, JSONL, CSV, ORC, Parquet, and XML. You can either provide a schema that describes the incoming records, or Onehouse will infer the source schema by reading a sample of the files to be ingested. See below for a variety of options for collecting data into S3.

Google Cloud Storage

The Onehouse Google Cloud Storage connector allows you to continuously and incrementally stream data directly from any Google Cloud Storage (GCS) bucket into your Onehouse-managed lakehouse. Onehouse will infer the source schema by reading a sample of the files to be ingested. This connector also enables access to any data source accessible by a tool that can write data to GCS.

Apache Kafka

The Onehouse Apache Kafka connector allows you to stream data directly from any Kafka topic into your Onehouse-managed lakehouse. Messages can be serialized using Avro, JSON, JSON_SR (JSON Schema), and Protobuf. This connector also enables access to any data source accessible through the Apache Kafka Connect framework.

AWS MSK Kafka

The Onehouse Amazon MSK Kafka connector allows you to continuously stream data directly from AWS Managed Apache Kafka (MSK) into your Onehouse managed lakehouse. Messages can be serialized using Avro, JSON, JSON_SR (JSON Schema), and Protobuf.

Confluent Cloud Kafka

The Onehouse Confluent Cloud Kafka connector allows you to continuously stream data directly from Confluent Kafka into your Onehouse-managed lakehouse. Messages can be serialized using Avro, JSON, JSON_SR (JSON Schema), and Protobuf. Confluent Cloud provides over 120 pre-built connectors, custom connectors, and general-purpose connectors such as SFTP and HTTP.

Confluent Cloud CDC

The Onehouse Confluent Cloud CDC connector allows you to use your Confluent cluster to stream CDC data from relational databases into your managed lakehouse while additionally provisioning and managing the necessary resources within your Confluent Cloud account. Simply enter the details for your relational database and Confluent Cloud cluster, and then Onehouse will facilitate ingestion by automatically provisioning and managing resources in the Confluent cluster you provide.

MySQL CDC

The Onehouse MySQL CDC connector allows you to continuously stream data directly from a Postgres database into your Onehouse-managed lakehouse.

Connections continued on following page

Onehouse Native Connectors

Postgres CDC	The Onehouse Postgres CDC connector allows you to continuously stream data directly from a Postgres database into your Onehouse-managed lakehouse.
Existing Onehouse Tables	The Onehouse Existing Tables connector allows you to use your existing Onehouse tables to build a multi-stage pipeline.

Connections through Apache Kafka

Open-source Apache Kafka, supports a variety of connectors through the Kafka Connect framework. These connectors enable integration with various data systems and technologies. Here are some of the connectors available with open-source Apache Kafka:

File Stream Source Connector	Used for reading files from a filesystem into Kafka.
JDBC Source and Sink Connector	Allows integration with any relational database that has a JDBC driver.
JMS Source Connector	Connects Kafka with JMS-compliant message brokers, enabling the import of messages into Kafka.
Twitter Source Connector	Streams data from Twitter into Kafka, useful for real-time social media analytics.
RabbitMQ Connector	Allows consuming from RabbitMQ into Kafka.

These connectors are part of the broader Kafka Connect ecosystem, which is designed to be extensible. It allows for the development of custom connectors if the existing ones do not meet specific requirements. The open-source community around Kafka often contributes additional connectors, which can be found in various repositories and community hubs.

Connections through Confluent Cloud Kafka

Confluent Cloud provides a comprehensive range of connectors that support various data formats and integration scenarios including over 120 pre-built connectors. Here's a sampling of the available connectors. Check the Confluent Web site for more information.

Apache ActiveMQ	Reads messages from an ActiveMQ cluster and writes them to a Kafka topic.
AWS CloudWatch Logs	Imports data from Amazon CloudWatch Logs into Kafka topics.
AWS Kinesis	Pulls data from Amazon Kinesis and persists it to Kafka topics.
AWS Simple Queue Service (SQS)	Moves messages from an Amazon SQS Queue into Apache Kafka.
GitHub	Writes metadata from GitHub to Apache Kafka.
Google Cloud Storage (GCS)	Reads data from any type of file listed under a GCS bucket.
JDBC	Imports data from any relational database with a JDBC driver into Kafka.
JMS Source	Moves messages from any JMS-compliant broker into Kafka.
Microsoft Azure Cosmos	Reads records from an Azure Cosmos database and writes data to Apache Kafka.
Microsoft Azure Event Hubs	Polls data from Azure Event Hubs and persists it to an Apache Kafka topic.
Microsoft SQL Server CDC	Obtains a snapshot of the existing data in a Microsoft SQL Server database and then monitors and records all subsequent row-level changes to that data.
MongoDB Atlas	Moves data from a MongoDB replica set into an Apache Kafka cluster.
MySQL Change Data Capture (CDC)	Obtains a snapshot of the existing data in a MySQL database and then monitors and records all subsequent row-level changes to that data.
Oracle	CDC, Application API, and Ad-hoc queries into Kafka.

Connections continued on following page

Connections through Confluent Cloud Kafka

Salesforce CDC	Provides a way to monitor Salesforce records.
SFTP	Watches an SFTP directory for files and reads the data as new files are added to it.
Zendesk	Moves customer ticket information including updates into a Kafka topic.

More Ways to Get Data into Confluent Kafka Topics

Custom Connectors	Users can upload their custom connectors based on Kafka Connect plugins to Confluent Cloud. These can include connectors built from scratch, modified open-source connectors, or third-party connectors.
HTTP Source Connector	Periodically polls data from HTTP APIs and produces records from that source into Apache Kafka. It supports output data formats like Avro, JSON Schema (JSON-SR), Protobuf, and JSON (schemaless).

Connections through Amazon MSK Kafka

Amazon MSK (Managed Streaming for Apache Kafka) supports a variety of data sources through its integration capabilities, particularly using Kafka Connect and MSK Connect. Here is a representative list of data sources that can be integrated with Onehouse via Amazon MSK:

AWS Aurora	Compatible with MSK using the Debezium connector for MySQL and PostgreSQL, which captures row-level database changes.
AWS DynamoDB	Streams data changes from DynamoDB tables using connectors like the DynamoDB Streams Kinesis Adapter to capture and process data changes.
AWS Kinesis	Allows integration with Kinesis Data Streams, enabling real-time data processing and analytics.
AWS Lambda	Can be used to process data and then publish it to an MSK topic, enabling complex processing workflows.
AWS RDS (Relational Database Service)	Supports streaming data from various relational databases hosted on Amazon RDS, including MySQL, PostgreSQL, and Oracle.
AWS S3 (Simple Storage Service)	Often used as a data sink but can also be a source when combined with other AWS services like AWS Lambda to process and send data to MSK.
Change Data Capture (CDC) from Various Databases	Using connectors like Debezium, changes from databases such as Oracle, SQL Server, and others can be streamed into MSK.
Enterprise Messaging Systems (e.g., RabbitMQ, ActiveMQ)	These systems can be connected to MSK to stream message queue data into Kafka topics.
File Systems (e.g., Amazon EFS, Amazon FSx)	File systems can be sources of data when log files or other data are ingested into MSK for processing.
IoT Devices (via AWS IoT Core)	IoT device data can be streamed to MSK through AWS IoT Core, which acts as a bridge between IoT devices and powerful data processing capabilities.
Log Aggregation Systems (e.g., Fluentd, Logstash)	Used for collecting and streaming logs from various sources into MSK for centralized log analysis.

Connections continued on following page

Connections through Amazon MSK Kafka

Social Media APIs

Data from social media platforms can be ingested into MSK for real-time analytics and processing, typically using custom connectors or third-party services.

Web and Mobile Applications

Data generated by user interactions can be streamed directly into MSK using client libraries or indirectly via intermediate services like Amazon Kinesis or AWS Lambda.

Connections via Tools that Write to S3

For non-real-time architectures, other tools can connect to SaaS apps and store the data in S3, which Onehouse can then ingest.

Airbyte	An open-source data integration platform that syncs data from databases, APIs, and SaaS tools to warehouses, lakes, and databases, including Amazon S3.
AWS AppFlow	Automates bi-directional data flows between SaaS applications and AWS services, including Amazon S3, with no code required.
AWS Database Migration Service (DMS)	Supports continuous data replication with high availability and consolidates databases into a petabyte-scale data warehouse by streaming data to Amazon S3.
Fivetran	A fully managed automated data integration platform that loads data into cloud warehouses, databases, and other storage solutions like Amazon S3.
Hevo Data	A no-code data pipeline platform that automates the process of loading data from various SaaS applications and databases into Amazon S3 and other destinations.
Logstash	Part of the Elastic Stack, it dynamically ingests data from various sources, transforms it, and ships it to destinations like Amazon S3.
Matillion	Provides data transformation solutions for cloud data warehouses, including loading data into Amazon S3 as part of the ETL process.
Segment	Collects, cleans, and controls customer data and streams it into Amazon S3 and other data warehouses for advanced analysis.
Stitch	A simple, powerful ETL service built for developers that connects to today's most popular business tools and consolidates that data into Amazon S3.
Talend	Offers data integration and data integrity solutions, enabling users to transform, access, and manage data across multiple cloud and on-premises repositories, including Amazon S3.

Databases Supported via Debezium

Onehouse supports change data capture (CDC) via Debezium, and Debezium, in-turn, supports a variety of databases:

Cassandra	Debezium supports both Cassandra 3.x and 4.x, allowing you to track changes in this NoSQL database.
Db2	Debezium provides a connector for Db2, extending its change data capture capabilities to this platform.
Informix (Incubating)	An Informix connector is in development, signaling Debezium's intent to include this database in its supported list.
JDBC (Incubating)	Debezium is developing a JDBC connector, potentially allowing it to connect to a wider range of databases via JDBC drivers.
MongoDB	Debezium captures changes from MongoDB replica sets or sharded clusters, enabling real-time data streaming.
MySQL	Debezium reads MySQL's binary log to capture changes, providing a reliable way to track updates.
Oracle	Debezium supports Oracle using LogMiner and XStream, enabling change data capture for this database.
PostgreSQL	Debezium leverages PostgreSQL's logical decoding feature to stream changes, requiring a plugin installation.
Spanner	Debezium's connector for Spanner is also incubating, showing its commitment to expanding its support to Google's distributed database.
SQL Server	Debezium's connector for SQL Server captures changes, making it suitable for data integration tasks.
Vitess (Incubating)	Debezium has a connector for Vitess under development, indicating future support for this database clustering system.

Supported Data Streams via Spark

Onehouse natively supports Kafka, Confluent Kafka, and Amazon MSK Kafka as well as the following streams via Spark:

Apache Flink	A powerful stream processing framework that provides high-throughput, low-latency, and accurate results for stateful computations over data streams.
Redpanda	A Kafka-compatible streaming platform that offers improved performance and resource efficiency compared to Apache Kafka.
StreamNative	A cloud-native Kafka platform that simplifies the deployment and management of Kafka clusters while offering additional enterprise features.

Supported Cloud Storage Options via Spark

Onehouse will natively support the cloud storage option for the cloud provider that the Onehouse project is linked to. In addition, these storage options are available via Spark:

AWS S3	Amazon S3 is a scalable object storage service offering high availability and durability for storing Spark data in the AWS cloud.
Azure Storage	Microsoft's cloud storage solution, Azure Storage offers various storage options like blobs, files, and queues, accessible by Spark for data processing.
GCP Cloud Storage	Google Cloud Storage provides a similar object storage service to S3, but within the Google Cloud Platform ecosystem for seamless integration.
Hadoop	The foundation for Spark, Hadoop Distributed File System (HDFS) provides distributed storage across a cluster for large datasets.
MinIO	MinIO is a high-performance object storage server, compatible with the S3 API, often used for building private cloud storage solutions accessible by Spark.

Supported File Formats

Here's a closer look at the file formats supported, along with their key features:

Apache Avro	• Columnar storage format, often used with Hadoop and big data systems. • Schema-based (the schema defines the data's structure). • Efficient for both storage and processing, especially with complex data.
CSV (Comma-Separated Values)	• A simple, text-based format where each line represents a row and values within the row are separated by commas. • Widely used for tabular data, easily importable into spreadsheets.
JSON	• Text-based, human-readable format representing data as key-value pairs. • Hierarchical structure for nested data. • Extremely common in web applications and data exchange.
JSONL (JSON Lines)	• Variation of JSON where each line in the file is a valid JSON object. • Good for streaming data, as each line can be processed independently.
JSON_SR (JSON Schema)	• JSON_SR is a serialization format used in conjunction with the Confluent Schema Registry. JSON_SR messages directly embed the schema identifier within the data payload, ensuring the receiver knows how to interpret the data correctly.
ORC (Optimized Row Columnar)	• Columnar storage format designed for efficiency in Hadoop ecosystems. • Improves query performance and data compression.
Parquet	• Another columnar storage format, popular in big data environments. • Optimizes for fast querying and compression.
Protobuf	• Protobuf (Protocol Buffers) is a platform-neutral and language-neutral method for serializing structured data, designed by Google. It's known for being smaller, faster, and more efficient than alternatives like XML and JSON.
XML (Extensible Markup Language)	• Flexible, text-based format using tags to define data structure. • Can represent complex, hierarchical data.

Database Connectivity to Onehouse

Apache Doris	An MPP analytical database that accesses Onehouse data via HMS or AWS Glue with the Hudi extension.
AWS Athena	A serverless query service using the AWS Glue catalog and Hudi extension to query Onehouse data stored in S3.
AWS Redshift	A cloud data warehouse that can query Onehouse data using Iceberg external tables.
AWS Redshift Spectrum	Google Cloud's serverless data warehouse utilizing the Hudi external table feature to query Onehouse data.
Azure Synapse Analytics	Microsoft's cloud analytics service, supports connection to Onehouse via external tables.
Databricks Spark	The Spark engine in Databricks, connecting to Onehouse through the Unity Catalog's Delta Lake extension.
Databricks SQL	Part of Databricks Lakehouse Platform, it connects to Onehouse data through the Unity Catalog's Delta Lake extension.
DuckDB	A fast, in-process analytical database that can connect to Onehouse data through the Unity Catalog's Iceberg or Delta Lake extension.
GCP BigQuery	Google Cloud's serverless data warehouse utilizing the Hudi external table feature to query Onehouse data.
PrestoDB	Similar to TrinoDB, it's a distributed SQL query engine connecting to Onehouse using HMS or AWS Glue and the Hudi extension.
Snowflake	A cloud data warehouse that seamlessly integrates with Onehouse using its Snowflake Catalog and Iceberg external tables.
StarRocks	A high-performance MPP database that uses the Hive Metastore (HMS) or AWS Glue with the Hudi extension to access Onehouse data.

Try Onehouse for free at onehouse.ai/schedule-a-test-drive.