

PostgreSQL CDC to Iceberg

Date published: 2021-04-06

Date modified: 2024-06-03



Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

ReadyFlow: PostgreSQL CDC to Iceberg [Technical Preview].....	4
Prerequisites.....	4
List of required configuration parameters for the PostgreSQL CDC to Iceberg [Technical Preview] ReadyFlow.....	6

ReadyFlow: PostgreSQL CDC to Iceberg [Technical Preview]

You can use the PostgreSQL CDC to Iceberg [Technical Preview] ReadyFlow to retrieve CDC events from a PostgreSQL source table and stream them into Iceberg.

This ReadyFlow uses Debezium to retrieve CDC events (INSERT, UPDATE, DELETE) from a PostgreSQL source table to stream the events to an Iceberg destination table. Failed Iceberg write operations are retried automatically to handle transient issues. Define a KPI on the failure_WriteIcebergTable connection to monitor failed write operations.



Note: This ReadyFlow is considered Technical Preview and is not designed for production use.



Note:

This ReadyFlow does not support schema changes or primary key field updates. The destination table must be Iceberg format v2. Make sure to assign the correct permissions for the Iceberg destination table to the specified CDP Workload User.

PostgreSQL CDC to Iceberg [Technical Preview] ReadyFlow details	
Source	PostgreSQL Table
Source Format	PostgreSQL Table
Destination	Iceberg
Destination Format	Parquet

Prerequisites

Learn how to collect the information you need to deploy the PostgreSQL CDC to Iceberg [Technical Preview] ReadyFlow, and meet other prerequisites.

For your data ingest source



Note: Do not change primary key field values in your source table after configuring the ReadyFlow. Doing so will cause the ReadyFlow to reject updates.



Note: You need to take care of field case sensitivity when defining source and destination table structure.

- You have obtained the PostgreSQL database server hostname and port.
- You have obtained the PostgreSQL schema name and table name. Take note of the table structure, specifically field case sensitivity.
- You have obtained a username and password to access the PostgreSQL table.
- You have performed the [PostgreSQL setup tasks required to run Debezium](#).

For Cloudera DataFlow

- You have enabled Cloudera DataFlow for an environment.

For information on how to enable Cloudera DataFlow for an environment, see [Enabling Cloudera DataFlow for an Environment](#).

- You have created a Machine User to use as the Cloudera Workload User.

- You have given the Cloudera Workload User the EnvironmentUser role.
 - From the Management Console, go to the environment for which Cloudera DataFlow is enabled.
 - From the Actions drop down, click Manage Access.
 - Identify the user you want to use as a Workload User.

**Note:**

The Cloudera Workload User can be a machine user or your own user name. It is best practice to create a dedicated Machine user for this.


- Give that user EnvironmentUser role.
- You have synchronized your user to the Cloudera Public Cloud environment that you enabled for Cloudera DataFlow.

For information on how to synchronize your user to FreeIPA, see [Performing User Sync](#).

- You have granted your Cloudera user the DFCatalogAdmin and DFFlowAdmin roles to enable your user to add the ReadyFlow to the Catalog and deploy the flow definition.
 - Give a user permission to add the ReadyFlow to the Catalog.
 - From the Management Console, click User Management.
 - Enter the name of the user or group you wish to authorize in the Search field.
 - Select the user or group from the list that displays.
 - Click Roles Update Roles .
 - From Update Roles, select DFCatalogAdmin and click Update.



Note: If the ReadyFlow is already in the Catalog, then you can give your user just the DFCatalogViewer role.

- Give your user or group permission to deploy flow definitions.
 - From the Management Console, click Environments to display the Environment List page.
 - Select the environment to which you want your user or group to deploy flow definitions.
 - Click Actions Manage Access to display the Environment Access page.
 - Enter the name of your user or group you wish to authorize in the Search field.
 - Select your user or group and click Update Roles.
 - Select DFFlowAdmin from the list of roles.
 - Click Update Roles.
- Give your user or group access to the Project where the ReadyFlow will be deployed.
 - Go to DataFlow Projects .
 - Select the project where you want to manage access rights and click  More Manage Access .
- Start typing the name of the user or group you want to add and select them from the list.
- Select the Resource Roles you want to grant.
- Click Update Roles.
- Click Synchronize Users.

For your data ingest target

- In Cloudera Data Warehouse, you have activated the same environment for which Cloudera DataFlow has been enabled. This will create a default database catalog. For more information, see [Activating an AWS environment from Cloudera Data Warehouse](#) or [Activating Azure environments](#).
- You have created a Hive Virtual Warehouse referencing the default database catalog. For more information, see [Creating your first Virtual Warehouse](#).
- You have created the Iceberg table that you want to ingest data into, running in your Hive Virtual Warehouse. For more information, see [Iceberg table creation from Hive](#).

Related Concepts[List of required configuration parameters for the PostgreSQL CDC to Iceberg \[Technical Preview\] ReadyFlow](#)

List of required configuration parameters for the PostgreSQL CDC to Iceberg [Technical Preview] ReadyFlow

When deploying the PostgreSQL CDC to Iceberg [Technical Preview] ReadyFlow, you have to provide the following parameters. Use the information you collected in *Prerequisites*.

Table 1: PostgreSQL CDC to Iceberg [Technical Preview] ReadyFlow configuration parameters

Parameter Name	Description
CDP Workload User	Specify the Cloudera machine user or workload user name that you want to use to authenticate to Hive. Ensure this user has the appropriate access rights to the Iceberg table.
CDP Workload User Password	Specify the password of the Cloudera machine user or workload user you are using to authenticate to Hive.
CDPEnvironment	The Cloudera Environment configuration resources.
Hive Catalog Namespace	Specify the Hive catalog namespace.
Iceberg Table Name	Specify the Iceberg table name you want to write to. Table must be Iceberg format v2.
Source Database Name	Specify the source database name.
Source Database Password	Specify the source database password.
Source Database Server Host Name	Specify the source database server host name.
Source Database Server Port	Specify the source database server port. The default value is 5432.
Source Database Table Name	Specify the source database table name in the form:[schema_name].[table_name].
Source Database User	Specify the source database user.

Related Concepts[Prerequisites](#)**Related Information**[Deploying a ReadyFlow](#)