Cloudera DataFlow

SQL Server CDC to Kudu

Date published: 2021-04-06
Date modified: 2024-01-09

CLOUDERA

Legal Notice

Contents

ReadyFlow: SQL Server CDC to Kudu [Technical Preview]

You can use the SQL Server CDC to Kudu ReadyFlow to retrieve CDC events from a SQL Server source table and stream them to a Kudu destination table.

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

Note:

This ReadyFlow is considered Technical Preview and is not designed for production use. The flow does not support schema changes or primary key field updates. Make sure to assign the correct permissions for the Kudu destination table to the specified CDP Workload User.

SQL Server CDC to Kudu [Technical Preview] ReadyFlow details	
Source	SQL Server Table
Source Format	SQL Server Table
Destination	Kudu
Destination Format	Kudu Table

Prerequisites

Learn how to collect the information you need to deploy the SQL Server CDC to Kudu [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 SQL Server database server hostname and port.
- You have obtained the SQL Server 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 SQL Server table.
- You have performed the SQL Server setup tasks required to run Debezium.

For DataFlow

- You have enabled DataFlow for an environment.

  For information on how to enable DataFlow for an environment, see Enabling DataFlow for an Environment.
- You have created a Machine User to use as the CDP Workload User.

- You have given the CDP Workload User the EnvironmentUser role.

  1. From the Management Console, go to the environment for which DataFlow is enabled.
  2. From the Actions drop down, click Manage Access.
  3. Identify the user you want to use as a Workload User.

     > **Note:**
     >
     > The CDP Workload User can be a machine user or your own user name. It is best practice to create a dedicated Machine user for this.

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

  For information on how to synchronize your user to FreeIPA, see Performing User Sync.
- You have granted your CDP user the DFCatalogAdmin and DFFlowAdmin roles to enable your user to add the ReadyFlow to the Catalog and deploy the flow definition.

  1. Give a user permission to add the ReadyFlow to the Catalog.

     a. From the Management Console, click User Management.
     b. Enter the name of the user or group you wish to authorize in the Search field.
     c. Select the user or group from the list that displays.
     d. Click Roles Update Roles .
     e. 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.

  2. Give your user or group permission to deploy flow definitions.

     a. From the Management Console, click Environments to display the Environment List page.
     b. Select the environment to which you want your user or group to deploy flow definitions.
     c. Click Actions Manage Access to display the Environment Access page.
     d. Enter the name of your user or group you wish to authorize in the Search field.
     e. Select your user or group and click Update Roles.
     f. Select DFFlowAdmin from the list of roles.
     g. Click Update Roles.

  3. Give your user or group access to the Project where the ReadyFlow will be deployed.

     a. Go to DataFlow Projects .
     b. Select the project where you want to manage access rights and click ⋮ More Manage Access .

  4. Start typing the name of the user or group you want to add and select them from the list.
  5. Select the Resource Roles you want to grant.
  6. Click Update Roles.
  7. Click Synchronize Users.

For your data ingest target

- You have a Real-Time Data Mart cluster running Kudu, Impala, and Hue in the same environment for which DataFlow has been enabled.
- You have the Kudu Master hostnames.

  1. From Management Console, click Data Hub Clusters.
  2. Select the Real-Time Data Mart cluster to which you want to ingest data into.
  3. Click the Hardware tab.
  4. Copy the FQDN for each Kudu Master.
- You have the Kudu database name.

- You have created the Kudu table that you want to ingest data into.

  1. Navigate to your Real Time Data Mart cluster and click Hue from the Services pane.
  2. Click the Tables icon on the left pane.
  3. Select the default database, and click + New to create a new table.
  4. In the Type field, select Manually and click Next.
  5. Provide the table Name, Format, Primary keys, and any partitions.
  6. Click Submit. The newly created table displays in the default database Tables pane.
  7. Check the Kudu UI **Tables** tab for the name of the table you created. You will need this table name when you use the DataFlow Deployment wizard to deploy the ReadyFlow.

- You have assigned permissions via IDBroker or in Ranger to enable the CDP Workload User to access the Kudu table that you want to ingest data into.

  1. From the base cluster on CDP Public Cloud, select Ranger.
  2. Select your Real Time Data Mart cluster from the **Kudu** folder.
  3. Click Add New Policy policy.
  4. On the **Create Policy** page, enter the Kudu table name in the topic field.
  5. Add the CDP Workload User in the Select User field.
  6. Add the Insert and Select permissions in the Permissions field.
  7. Click Save.

Related Concepts

List of required configuration parameters for the SQL Server CDC to Kudu [Technical Preview] ReadyFlow

List of required configuration parameters for the SQL Server CDC to Kudu [Technical Preview] ReadyFlow

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

Table 1: SQL Server CDC to Kudu [Technical Preview] ReadyFlow configuration parameters

Parameter Name	Description
CDP Workload User	Specify the CDP machine user or workload username that you want to use to authenticate to Kudu. Ensure this user has the appropriate access rights to the Kudu table.
CDP Workload User Password	Specify the password of the CDP machine user or workload user you are using to authenticate to Kudu.
Destination Database Table Name	Specify the destination database table name in the form: [database_name].[table_name]
Kudu Master Hosts	Specify the Kudu Master hostnames in a comma separated list.
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 1433.
Source Database Table Name	Specify the source database table name in the format [schema_name].[table_name]
Source Database User	Specify the source database user.

Related Concepts

List of required configuration parameters for the SQL Server CDC to Kudu [Technical Preview] ReadyFlow

Related Information

Deploying a ReadyFlow