

Kafka to Snowflake

Date published: 2021-04-06

Date modified: 2024-01-09

CLOUdera

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 Overview: Kafka to Snowflake.....	4
Prerequisites.....	4
List of required configuration parameters for the Kafka to Snowflake ReadyFlow.....	8

ReadyFlow Overview: Kafka to Snowflake

You can use the Kafka to Snowflake ReadyFlow to move your data from a Kafka topic to a Snowflake table.

This ReadyFlow consumes JSON, CSV or Avro data from a source Kafka topic and parses the schema by looking up the schema name in the CDP Schema Registry. You can filter events based on a user provided SQL query. The filtered events are then merged into CSV files, compressed into gzip format and written to the destination Snowflake DB table.

Kafka to Snowflake ReadyFlow details	
Source	Kafka Topic
Source Format	JSON, CSV, Avro
Destination	Snowflake DB
Destination Format	Snowflake DB

Prerequisites

Learn how to collect the information you need to deploy the Kafka to Snowflake ReadyFlow, and meet other prerequisites.

For your data ingest source

- You have created a Streams Messaging cluster in CDP Public Cloud to host your Schema Registry.
For information on how to create a Streams Messaging cluster, see [Setting up your Streams Messaging Cluster](#).
- You have created at least one Kafka topic.

- Navigate to Management Console > Environments and select your environment.
- Select your Streams Messaging cluster.
- Click on the Streams Messaging Manager icon.
- Navigate to the Topics page.
- Click Add New and provide the following information:

- Topic name
- Number of partitions
- Level of availability
- Cleanup policy



Tip:

SMM has automatically set Kafka topic configuration parameters. To manually adjust them, click Advanced.

- Click Save.
- You have created a schema for your data and have uploaded it to the Schema Registry in the Streams Messaging cluster.

For information on how to create a new schema, see [Creating a new schema](#). For example:

```
{
  "type": "record",
  "name": "SensorReading",
  "namespace": "com.cloudera.example",
```

```

"doc":"This is a sample sensor reading",
"fields":[
  {
    "name":"sensor_id",
    "doc":"Sensor identification number.",
    "type":"int"
  },
  {
    "name":"sensor_ts",
    "doc":"Timestamp of the collected readings.",
    "type":"long"
  },
  {
    "name":"sensor_0",
    "doc":"Reading #0.",
    "type":"int"
  },
  {
    "name":"sensor_1",
    "doc":"Reading #1.",
    "type":"int"
  },
  {
    "name":"sensor_2",
    "doc":"Reading #2.",
    "type":"int"
  },
  {
    "name":"sensor_3",
    "doc":"Reading #3.",
    "type":"int"
  }
]
}

```

- You have the Schema Registry Host Name.
 1. From the Management Console, go to Data Hub Clusters and select the Streams Messaging cluster you are using.
 2. Navigate to the **Hardware** tab to locate the Master Node FQDN. Schema Registry is always running on the Master node, so copy the Master node FQDN.
- You have the Kafka broker end points.
 1. From the Management Console, click Data Hub Clusters.
 2. Select the Streams Messaging cluster from which you want to ingest data.
 3. Click the Hardware tab.
 4. Note the Kafka Broker FQDNs for each node in your cluster.
 5. Construct your Kafka Broker Endpoints by using the FQDN and Port number 9093 separated by a colon. Separate endpoints by a comma. For example:

```
broker1.fqdn:9093,broker2.fqdn:9093,broker3.fqdn:9093
```

Kafka broker FQDNs are listed under the **Core_broker** section.

- You have the Kafka Consumer Group ID.


This ID is defined by the user. Pick an ID and then create a Ranger policy for it. Use the ID when deploying the flow in DataFlow.

- You have assigned the CDP Workload User policies to access the consumer group ID and topic.
 1. Navigate to Management Console > Environments, and select the environment where you have created your cluster.
 2. Select Ranger. You are redirected to the Ranger **Service Manager** page.
 3. Select your Streams Messaging cluster under the **Kafka** folder.
 4. Create a policy to enable your Workload User to access the Kafka source topic.
 5. On the **Create Policy** page, give the policy a name, select topic from the drop-down list, add the user, and assign the Consume permission.
 6. Create another policy to give your Workload User access to the consumer group ID.
 7. On the **Create Policy** page, give the policy a name, select consumergroup from the drop-down list, add the user, and assign the Consume permission.
- You have assigned the CDP Workload User read-access to the schema.
 1. Navigate to Management Console > Environments, and select the environment where you have created your cluster.
 2. Select Ranger. You are redirected to the Ranger **Service Manager** page.
 3. Select your Streams Messaging cluster under the **Schema Registry** folder.
 4. Click Add New Policy.
 5. On the **Create Policy** page, give the policy a name, specify the schema details, add the user, and assign the Read permission.

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.

For your data ingest target

- You have created a Snowflake account.
- Ensure that the Snowflake database and table you are ingesting data to exist. If not, create them.
- You have set up your destination Snowflake table with the right columns and columns type.
- Your Snowflake user account has write access to the Snowflake database table.
- You have the schema of the Snowflake database. The default value is PUBLIC.
- You have created Snowflake Private and Public Keys and assigned the Public Key to your Snowflake account user.

For more information, see [Key Pair Authentication](#) in Snowflake documentation.

- You have created your Snowflake Pipe and configured the file format for CSV input.

For more information, see [CREATE FILE FORMAT](#) in Snowflake documentation.

Related Concepts

[List of required configuration parameters for the Kafka to Snowflake ReadyFlow](#)

List of required configuration parameters for the Kafka to Snowflake ReadyFlow

When deploying the Kafka to Snowflake ReadyFlow, you have to provide the following parameters. Use the information you collected in *Prerequisites*.

Table 1: Kafka to Snowflake ReadyFlow configuration parameters

Parameter Name	Description
CDP Workload User	Specify the CDP machine user or workload user name that you want to use to authenticate to Kafka. Ensure this user has the appropriate access rights to the source Kafka topic.
CDP Workload User Password	Specify the password of the CDP machine user or workload user you are using to authenticate against Kafka.
CSV Delimiter	If your source data is CSV, specify the delimiter here.
Data Input Format	Specify the format of your input data. You can use "CSV", "JSON" or "AVRO" with this ReadyFlow. <ul style="list-style-type: none"> • CSV • JSON • AVRO
Filter Rule	Specify the filter rule expressed in SQL to filter streaming events for the destination database. Records matching the filter will be written to the destination database. The default value forwards all records.
Kafka Broker Endpoint	Specify the Kafka bootstrap server.
Kafka Consumer Group ID	The name of the consumer group used for the source topic you are consuming from.
Kafka Source Topic	Specify a topic name that you want to read from.
Kafka Schema Name	Specify the schema name to be looked up in the CDP Schema Registry for the source Kafka topic.
Schema Registry Hostname	Specify the hostname of the Schema Registry you want to connect to. This must be the direct hostname of the Schema Registry itself, not the Knox Endpoint.
Snowflake Account Name	Specify the Snowflake account name.
Snowflake Organization	Specify the Snowflake organization name.
Snowflake Pipe	Specify the Snowflake pipe.
Snowflake Private Key File	Upload the Snowflake private key file.
Snowflake Private Key Password	Specify the password for the Snowflake private key.
Snowflake User Name	Specify your Snowflake user name.
Snowflake User Password	Specify your Snowflake user password.
Snowflake Warehouse	Specify your Snowflake warehouse. The default value is COMPUTE_WH.
Table Name	Specify the Snowflake table name you want to write to.
Target Database Name	Specify the Snowflake database you want to write to.
Target Schema Name	Specify the schema of the Snowflake database. The default value is PUBLIC.

Related Concepts

[Prerequisites](#)

[Related Information](#)

[Deploying a ReadyFlow](#)