Cloudera Data Flow

Kafka Filter to Kafka

Date published: 2021-04-06
Date modified: 2025-09-30

CLOUDERA

Legal Notice

Contents

ReadyFlow: Kafka filter to Kafka

You can use the Kafka filter to Kafka ReadyFlow to move your data between two Kafka topics, while applying a schema to the data in Cloudera Data Flow.

Kafka filter to Kafka ReadyFlow description

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 Cloudera Schema Registry. You can filter events by specifying a SQL query in the `Filter Rule` parameter. The filtered events are then converted to the specified output data format and written to the destination Kafka topic. Failed Kafka write operations are retried automatically to handle transient issues. Define a KPI on the `failure_WriteToKafka` connection to monitor failed write operations.

Kafka filter to Kafka ReadyFlow details

ReadyFlow details	
Source	Kafka Topic
Source Format	JSON, CSV, or Avro
Destination	Kafka Topic
Destination Format	JSON, CSV, or Avro

Moving data with a Kafka filter to Kafka flow

You can use a Kafka filter to Kafka data flow when you want to filter specific events from a Kafka topic and write the filtered stream to another Kafka topic. For example, you could use this flow to filter out erroneous records that contain "null" for an important field. In order to filter the data using a SQL query, you need to provide a schema for the events that you are processing in the topic.

Your data flow can consume JSON, CSV, or Avro data from the source Kafka topic and write to the destination Kafka topic in any of these formats. The data flow parses the schema by looking up the schema name in the Cloudera Schema Registry. You can filter the events by specifying a SQL query. The filtered events are converted to the specified output data format, and they are written to the destination Kafka topic.

Prerequisites

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

For your data source and target

* You have created a Streams Messaging cluster in Cloudera on 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 two Kafka topics for your ingest data flow, according to your source and target requirements.

    1. Navigate to Management Console > Environments and select your environment.
    2. Select your Streams Messaging cluster.
    3. Click on the Streams Messaging Manager icon.
    4. Navigate to the Topics page.
    5. 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.

    6. Click Save.
    7. Repeat the previous steps to add another topic.

- 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 Cloudera Data Flow.

- You have assigned the Cloudera 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 Cloudera 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 Cloudera Data Flow

- You have enabled Cloudera Data Flow for an environment.

    For information on how to enable Cloudera Data Flow for an environment, see Enabling Cloudera Data Flow 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.

  1. From the Management Console, go to the environment for which Cloudera Data Flow 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 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.
  4. Give that user EnvironmentUser role.
- You have synchronized your user to the Cloudera on cloud environment that you enabled for Cloudera Data Flow.

  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.

  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 Data Flow 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.

Related Concepts
Required parameters
Related Information
Deploying a ReadyFlow

Required parameters

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

Table 1: Kafka Filter to Kafka 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 Kafka. Ensure this user has the appropriate access rights in Ranger for the source and target Kafka topics.
CDP Workload User Password	Specify the Cloudera machine user or workload user name that you want to use to authenticate to Kafka.
Data Input Format	Specify the format of your input data. If your data input is CSV, define a CSV delimiter for the data in the CSV Delimiter text box. If you use AVRO or JSON format, the delimiter is ignored. • CSV • JSON • AVRO
Data Output Format	Specify the format of your output data. If your data input is CSV, define a CSV delimiter for the data in the CSV Delimiter text box. If you use AVRO or JSON format, the delimiter is ignored. • CSV • JSON • AVRO
CSV Delimiter	If your source data is CSV, specify the delimiter here.
Filter Rule	If you want to filter your data for the destination topic, enter a filter rule expressed in SQL. Records matching the filter are written to the destination topic in Kafka. If you do not provide a specific filter rule, the default rule forwards all records. Default rule:SELECT * FROM FLOWFILE
Kafka Broker Endpoint	Specify the Kafka bootstrap servers string as a comma separated list in the format <host>:<port>.
Kafka Consumer Group ID	Add the name of the consumer group used for the source topic you are consuming from. Make sure to use the Consumer Group ID that the selected Cloudera Workload User is allowed to use.
Kafka Source Topic	Specify a topic name that you want to read from.
Kafka Destination Topic	Specify a topic name that you want to write to.
Kafka Producer ID	Use it to identify your data flow in SMM.
Schema Name	Specify the schema that you want to use in your data flow. Cloudera Data Flow looks up this schema in the Schema Registry you define with the Schema Registry Hostname. See the *Appendix* for an example schema.
Schema Registry Hostname	Specify the hostname of the Schema Registry running on the master node in the Streams Messaging cluster that you want to connect to. This must be the direct hostname of the Schema Registry itself, not the Knox Endpoint.

Related Concepts

Prerequisites