Cloudera DataFlow

Kafka to S3 Avro

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

https://docs.cloudera.com/

Legal Notice

Contents

ReadyFlow: Kafka to S3 Avro

You can use the Kafka to S3 Avro ReadyFlow to move your data from a Kafka topic to an Amazon S3 bucket.

This ReadyFlow consumes JSON, CSV or Avro data from a source Kafka topic and merges the events into Avro files before writing the data to S3. The flow writes out a file every time its size has either reached 100MB or five minutes have passed. Files can reach a maximum size of 1GB. You can specify the topic you want to read from as well as the target S3 bucket and path. Failed S3 write operations are retried automatically to handle transient issues. Define a KPI on the `failure_WriteToS3` connection to monitor failed write operations.

ReadyFlow details	
Source	Kafka topic
Source Format	JSON, CSV, Avro
Destination	Amazon S3
Destination Format	Avro

Moving data to object stores

Cloud environments offer numerous deployment options and services. There are many ways to store data in the cloud, but the easiest option is to use object stores. Object stores are extremely robust and cost-effective storage solutions with multiple levels of durability and availability. You can include them in your data pipeline, both as an intermediate step and as an end state. Object stores are accessible to many tools and connecting systems, and you have a variety of options to control access.

Prerequisites

Learn how to collect the information you need to deploy the Kafka to S3 Avro 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.

  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.

- 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.

  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 your source S3 path and bucket.

- Perform one of the following to configure access to S3 buckets:
  - You have configured access to S3 buckets with a RAZ enabled environment.

    It is a best practice to enable RAZ to control access to your object store buckets. This allows you to use your CDP credentials to access S3 buckets, increases auditability, and makes object store data ingest workflows portable across cloud providers.

    1. Ensure that Fine-grained access control is enabled for your DataFlow environment.
    2. From the Ranger UI, navigate to the S3 repository.
    3. Create a policy to govern access to the S3 bucket and path used in your ingest workflow.

       Tip:

       The Path field must begin with a forward slash (/).
    4. Add the machine user that you have created for your ingest workflow to the policy you just created.

    For more information, see *Creating Ranger policy to use in RAZ-enabled AWS environment*.
  - You have configured access to S3 buckets using ID Broker mapping.

    If your environment is not RAZ-enabled, you can configure access to S3 buckets using ID Broker mapping.

    1. Access IDBroker mappings.
       a. To access IDBroker mappings in your environment, click Actions Manage Access .
       b. Choose the IDBroker Mappings tab where you can provide mappings for users or groups and click Edit.
    2. Add your CDP Workload User and the corresponding AWS role that provides write access to your folder in your S3 bucket to the Current Mappings section by clicking the blue + sign.

       Note: You can get the AWS IAM role ARN from the Roles Summary page in AWS and can copy it into the IDBroker role field. The selected AWS IAM role must have a trust policy allowing IDBroker to assume this role.
    3. Click Save and Sync.

List of required configuration parameters for the Kafka to S3 Avro ReadyFlow

When deploying the Kafka to S3 Avro ReadyFlow, you have to provide the following parameters. Use the information you collected in *Prerequisites*.

Table 1: Kafka to S3 Avro 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 and the object store. Ensure this user has the appropriate access rights in Ranger for the Kafka topic and ID Broker for object store access.
CDP Workload User Password	Specify the password of the CDP machine user or workload user you're using to authenticate against Kafka and the object store.
CDPEnvironment	DataFlow will use this parameter to auto-populate the Flow Deployment with Hadoop configuration files required to interact with S3. DataFlow automatically adds all required configuration files to interact with Data Lake services. Unnecessary files that are added won't impact the deployment process.
CSV Delimiter	If your source data is CSV, specify the delimiter here.

Parameter Name	Description
Data Input Format	Specify the format of your input data. Supported values are: • CSV • JSON • AVRO
Kafka Broker Endpoint	Specify the Kafka bootstrap servers string as a comma separated list.
Kafka Consumer Group ID	The name of the consumer group used for the the source topic you are consuming from.
Kafka Source Topic	Specify a topic name that you want to read from.
S3 Bucket	Specify the name of the S3 bucket you want to write to. The full path will be constructed from: s3a://#{S3 Bucket}/#{S3 Path}/${Kafka.topic}
S3 Bucket Region	Specify the AWS region in which your bucket was created. Supported values are: • us-gov-west-1 • us-gov-east-1 • us-east-1 • us-east-2 • us-west-1 • us-west-2 • eu-west-1 • eu-west-2 • eu-west-3 • eu-central-1 • eu-north-1 • eu-south-1 • ap-east-1 • ap-south-1 • ap-southeast-1 • ap-southeast-2 • ap-northeast-1 • ap-northeast-2 • ap-northeast-3 • sa-east-1 • cn-north-1 • cn-northwest-1 • ca-central-1 • me-south-1 • af-south-1 • us-iso-east-1 • us-isob-east-1 • us-iso-west-1
S3 Path	Specify the path within the bucket where you want to write to without any leading characters. The full path will be constructed from: s3a://#{S3 Bucket}/#{S3 Path}/${Kafka.topic}
Schema Name	Specify the schema name to be looked up in the 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.