Cloudera DataFlow

S3 to S3 Avro with S3 Notifications

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

https://docs.cloudera.com/

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: S3 to S3 Avro with S3 Notifications

You can use the S3 to S3 Avro with S3 Notifications ReadyFlow to move your data between Amazon S3 buckets and to configure notifications.

This ReadyFlow consumes JSON, CSV, or Avro files from a source S3 location, converts the files into Avro and then writes them to a destination S3 location. You can specify the source format, the source and target location and the schema to use for reading the source data. The ReadyFlow is notified about the new files in the source bucket by configuring S3 notifications in AWS.

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

Moving data between S3 locations

You can use this ReadyFlow when you want to move data from a location in AWS S3 to another S3 location, and at the same time convert the data to Avro format. You need to specify the source format, the source and target location as well as the schema to use for handling the data. Your flow can consume JSON, CSV or Avro files from the source S3 location. It converts the files into Avro format and writes them to the destination S3 location. You define and store the data processing schema in the Schema Registry on a Streams Messaging Data Hub cluster. The data flow parses the schema by looking up the schema name in the Schema Registry.

Setting up S3 notifications

S3 buckets can be configured to generate notifications when new data arrives. You can configure this ReadyFlow to receive S3 notifications from an SQS queue so that new files are automatically moved from source to destination directory

Prerequisites

Learn how to collect the information you need to deploy the S3 to S3 Avro ReadyFlow, and meet other prerequisites.

For your data source and target

- You have the two S3 buckets and their paths as source and destination for the data movement.

- You have performed 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.
- Create an ID Broker Mapping for the CDP Workload User and the IAM role that you want to use in the SQS access policy.

  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 access to the SQS Queue.

     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.

- You have configured your Simple Queue Service URL for S3 bucket notifications.

  For detailed instructions, see Walkthrough: Configuring a bucket for notifications (SNS topic or SQS queue)

  Note:

  Update the access policy in *Step 1: Create an Amazon SQS Queue* with information about the ID Broker Mapping for the CDP Workload User and the IAM role that you want to use in the SQS access policy. The updated policy will look similar to:

```
{
      "Effect": "Allow",
      "Principal": {
        "AWS": "arn:aws:iam::[***bucket-owner-account-id***]:rol
e/[***IAM-ROLE***]"
      },
      "Action": [
        "sqs:ReceiveMessage",
        "sqs:DeleteMessage"
      ],
      "Resource": "arn:aws:sqs:[***region***]:[***queue-owner-
account-id***]:[***queue-name***]"
    }
```

- 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 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 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.

Related Concepts

List of required configuration parameters for the S3 to S3 Avro with S3 Notifications ReadyFlow

List of required configuration parameters for the S3 to S3 Avro with S3 Notifications ReadyFlow

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

Table 1: S3 to S3 Avro with S3 Notifications 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 the object stores (via IDBroker) and to the schema registry. Ensure this user has the appropriate access rights to the object store locations and to the schema registry.
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.

Parameter Name	Description
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.
Data Input Format	Specify the format of your input data. Possible values for this ReadyFlow are: • Avro • CSV • JSON
Destination S3 Bucket	Specify the name of the destination S3 bucket you want to write to. The full path will be constructed out of s3a://#{Destination S3 Bucket}/#{Destination S3 Path}/[subdirectory]/${filename} ([subdirectory] is the original directory of the file if the #{Source S3 Path} prefix is not specified)
Destination S3 Path	Specify the path within the destination bucket where you want to write to. The full path will be constructed out of s3a://#{Destination S3 Bucket}/#{Destination S3 Path}/[subdirectory]/${filename} ([subdirectory] is the original directory of the file if the #{Source S3 Path} prefix is not specified)
Notification SQS Queue URL	URL of the SQS Queue set up for delivering notifications about S3 object creation events. The SQS Queue must be in the same region as the source S3 Bucket and the bucket needs to be configured to send notifications to the queue.
Schema Name	Specify the schema name to be looked up in the Schema Registry used to parse the source files.
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.
Source S3 Bucket	Specify the name of the source S3 bucket you want to read from.
Source S3 Path	Specify the path within the source bucket where you want to read files from. If notifications are received with a source object key not matching the Source S3 Path (as a prefix), they will be filtered out and will not be loaded. The Source S3 Path will be stripped off from the source object key and will be replaced with the Destination S3 Path in the destination object key.
Source S3 Region	Specify the AWS region of your source S3 Bucket and S3 Event Notification SQS Queue. 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

Related Concepts

Prerequisites

Related Information

Deploying a ReadyFlow