Cloudera DataFlow

S3 to CDW

Date published: 2021-04-06
Date modified: 2024-05-20

https://docs.cloudera.com/

Legal Notice

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: S3 to CDW..4

Prerequisites.. 4

List of required configuration parameters for the S3 to CDW ReadyFlow... 7

ReadyFlow: S3 to CDW

You can use the S3 to CDW Readyflow to consume CSV files from a source S3 location and write them as Parquet files to a destination S3 location and CDW Impala table.

This ReadyFlow consumes CSV files from a source S3 location, parses the schema by looking up the schema name in the CDP Schema Registry, converts the files into Parquet and writes them to a destination S3 location and CDW Impala table. You can specify the source S3 location, the target S3 location and the destination Impala table name. The ReadyFlow polls the source bucket for new files (it performs a listing periodically). Define KPIs on the failure_WriteS3 and failure_CreateCDWImpalaTable connections to monitor failed write operations.

Note: This ReadyFlow leverages CDP's centralized access control for cloud storage access. Make sure to either set up Ranger policies or an IDBroker mapping allowing your workload user access to the target S3 or ADLS location.

S3 to CDW ReadyFlow details	
Source	CDP Managed Amazon S3
Source Format	CSV
Destination	CDP Managed Amazon S3 and CDW Impala
Destination Format	Parquet

Prerequisites

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

For your S3 data ingest 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.
- 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.

For your Impala data ingest target

- You have activated your environment in Cloudera Data Warehouse. This automatically creates a default Database Catalog.
- You have created an Impala Virtual Warehouse referencing the default Database Catalog. Uncheck the Enable SSO setting.
- Select Copy the JDBC URL from the Virtual Warehouse UI. Use this connection string as the basis for the Impala JDBC URL parameter.
- For a RAZ enabled environment, you have assigned the CDP Workload User read and write access to the URL path where the Parquet files will be written via the Hadoop_SQL_URL policy.

Related Concepts

List of required configuration parameters for the S3 to CDW ReadyFlow

List of required configuration parameters for the S3 to CDW ReadyFlow

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

Table 1: S3 to CDW 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 are using to authenticate against the object stores (via IDBroker) and the schema registry.
Destination Database Driver	Upload the database driver jar file for your destination database.
Destination Database Driver Class Name	Specify the destination database driver class name, for example, com.cloudera.impala.jdbc.Driver for Impala databases.
Destination Database Table Name	Specify the destination database table name.
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}
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 B ucket}/#{Destination S3 Path}
Impala JDBC URL	Specify the Impala JDBC URL from the CDW Virtual Warehouse. Do not include the UID and PASSWORD connection parameters. The full JDBC URL will be constructed from: #{Impala JDBC URL};UID=#{CDP Workload User};PWD=#{CDP Workload User Password} where PWD value is provided by the dynamic property SENSITIVE.PWD in the Impala connection pool service configuration.
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.

Related Concepts

Prerequisites

Related Information

Deploying a ReadyFlow