

Google Drive to S3/ADLS

Date published: 2021-04-06

Date modified: 2024-06-03

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: Google Drive to S3/ADLS.....	4
Prerequisites.....	4
List of required configuration parameters for the Google Drive to S3/ADLS ReadyFlow.....	7

ReadyFlow overview: Google Drive to S3/ADLS

You can use the Google Drive to S3/ADLS ReadyFlow to ingest data from a Google Drive location to a destination in Amazon S3 or Azure Data Lake Service (ADLS).

This ReadyFlow consumes files from Google Drive and Cloudera managed S3 or ADLS location. For the source, specify the Google service account key file in JSON and the Google Drive folder ID. You can choose whether to include objects in subfolders or to only include objects from the specified Google Drive folder. For the destination, specify the S3 or ADLS storage location and path.

The ReadyFlow polls the folder on Google Drive for new files (it performs a listing periodically). Failed S3 or ADLS write operations are retried automatically to handle transient issues. Define a KPI on the failure_WriteToS3/ADLS connection to monitor failed write operations.

Google Drive to S3/ADLS ReadyFlow details	
Source	Google Drive
Source Format	Any
Destination	Cloudera managed Amazon S3 or ADLS
Destination Format	Same as source

Prerequisites

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

For your data ingest source

- You have a Google Cloud account. If you do not have one, you can sign up for a free account [here](#).
- You have enabled Google Drive API in Google Cloud. For instructions, see [Enable Google Workspace APIs](#).
- You have granted access to the Google Drive folder.
 - In the Google Cloud Console, navigate to IAM & Admin Service Accounts .
 - Copy the email address of the service account you are going to use.
 - Navigate to the folder to be listed in Google Drive.
 - Right-click the folder and select Share.
 - Enter the service account email.
- You have your Google Drive folder ID.

Navigate to the folder to be listed in Google Drive and enter it. The URL in your browser includes the ID at the end of the URL. For example, if the URL is `https://drive.google.com/drive/folders/1trTraPVCnX5_TNwO8d9P_bz278xWOmGm`, the folder ID is `1trTraPVCnX5_TNwO8d9P_bz278xWOmGm`.

- You have data stored in the Google Drive folder that you want to move to an object store.

For Cloudera DataFlow

- You have enabled Cloudera DataFlow for an environment.

For information on how to enable Cloudera DataFlow for an environment, see [Enabling Cloudera DataFlow 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.
 - From the Management Console, go to the environment for which Cloudera DataFlow is enabled.
 - From the Actions drop down, click Manage Access.
 - 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.


- Give that user EnvironmentUser role.
- You have synchronized your user to the Cloudera Public Cloud environment that you enabled for Cloudera DataFlow.

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.
 - Give a user permission to add the ReadyFlow to the Catalog.
 - From the Management Console, click User Management.
 - Enter the name of the user or group you wish to authorize in the Search field.
 - Select the user or group from the list that displays.
 - Click Roles Update Roles .
 - 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.

- Give your user or group permission to deploy flow definitions.
 - From the Management Console, click Environments to display the Environment List page.
 - Select the environment to which you want your user or group to deploy flow definitions.
 - Click Actions Manage Access to display the Environment Access page.
 - Enter the name of your user or group you wish to authorize in the Search field.
 - Select your user or group and click Update Roles.
 - Select DFFlowAdmin from the list of roles.
 - Click Update Roles.
- Give your user or group access to the Project where the ReadyFlow will be deployed.
 - Go to DataFlow Projects .
 - Select the project where you want to manage access rights and click  More Manage Access .
- Start typing the name of the user or group you want to add and select them from the list.
- Select the Resource Roles you want to grant.
- Click Update Roles.
- Click Synchronize Users.

For your S3 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 Cloudera 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 Cloudera 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 Cloudera 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**.

For your ADLS data ingest target

- You have your ADLS container and path into which you want to ingest data.

- You have performed one of the following to configure access to your ADLS folder:
 - You have configured access to the ADLS folders with a RAZ enabled environment.

It is a best practice to enable RAZ to control access to your object store folders. This allows you to use your Cloudera Public Cloud credentials to access ADLS folders, increases auditability, and makes object store data ingest workflows portable across cloud providers.

- Ensure that Fine-grained access control is enabled for your Cloudera DataFlow environment.
- From the Ranger UI, navigate to the ADLS repository.
- Create a policy to govern access to the ADLS container and path used in your ingest workflow. For example: adls-to-adls-avro-ingest



Tip: The Path field must begin with a forward slash (/).

- Add the machine user that you have created for your ingest workflow to ingest the policy you just created.

For more information, see *Ranger policies for RAZ-enabled Azure environment*.

- You have configured access to ADLS folders using ID Broker mapping.

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

- Access IDBroker mappings.
 - To access IDBroker mappings in your environment, click **Actions Manage Access**.
 - Choose the IDBroker Mappings tab where you can provide mappings for users or groups and click **Edit**.
- Add your Cloudera Workload User and the corresponding Azure role that provides write access to your folder in ADLS to the Current Mappings section by clicking the blue + sign.



Note: You can get the Azure Managed Identity Resource ID from the Azure Portal by navigating to **Managed Identities Your Managed Identity Properties Resource ID**. The selected Azure MSI role must have a trust policy allowing IDBroker to assume this role.

- Click **Save and Sync**.

Related Concepts

[List of required configuration parameters for the Google Drive to S3/ADLS ReadyFlow](#)

List of required configuration parameters for the Google Drive to S3/ADLS ReadyFlow

When deploying the Google Drive to S3/ADLS ReadyFlow, you have to provide the following parameters. Use the information you collected in *Prerequisites*.

Table 1: Google Drive to S3/ADLS ReadyFlow configuration parameters

Parameter name	Description
CDP Workload User	Specify the Cloudera machine user or workload username that you want to use to authenticate to the object stores. Ensure this user has the appropriate access rights to the object store locations in Ranger or IDBroker.
CDP Workload User Password	Specify the password of the Cloudera machine user or workload user you are using to authenticate against the object stores (via IDBroker).
Destination S3 or ADLS Path	Specify the name of the destination S3 or ADLS path you want to write to. Make sure that the path starts with "/".

Parameter name	Description
Destination S3 or ADLS Storage Location	<p>Specify the name of the destination S3 bucket or ADLS container you want to write to.</p> <ul style="list-style-type: none">For S3, enter a value in the form: s3a://[***<i>Destination S3 Bucket</i>***]For ADLS, enter a value in the form: abfs://[***<i>Destination ADLS File System</i>***]@[***<i>Destination ADLS Storage Account</i>***].dfs.core.windows.net
Google Drive Folder ID	Specify the ID of the Google Drive folder you want to read from.
Google Service Account Key File	Upload the Google Service Account key file in JSON format.
Include Subfolder Files	<p>Specify whether to include a list of files from subfolders.</p> <p>Set to "true" to include files from subfolders</p> <p>Set to "false" to only list files from the specified Google Drive folder.</p> <p>The default value is <i>true</i>.</p>

Related Concepts

[Prerequisites](#)

[Related Information](#)

[Deploying a ReadyFlow](#)