Cloudera Data Flow

ADLS to Qdrant

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

CLOUDERA

Legal Notice

Contents

ReadyFlow overview: ADLS to Qdrant [Technical Preview]

You can use the ADLS to Qdrant [Technical Preview] ReadyFlow to consume PDF documents from ADLS, vectorize them using an OpenAI model and write the results to Qdrant.

This ReadyFlow consumes PDF documents from a source ADLS location, partitions the PDFs, chunks the data, vectorizes the data using an OpenAI embedding model, and stores the results in Qdrant vector DB. The default OpenAI model is 'text-embedding-ada-002'. An OpenAI API key and a Qdrant API key are required to run this flow. Define a KPI on the failure_WriteToQdrant connection to monitor failed write operations.

Note: This ReadyFlow is considered Technical Preview and is not designed for production use. The flow uses Python processors and must be run in NiFi 2.x.

To ensure correct functionality, make sure your Qdrant collection has been created with the same dimension as the 'text-embedding-ada-002' model (1536). The images within PDFs will be ignored.

Note:

This ReadyFlow leverages Cloudera on cloud'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 source ADLS location.

ADLS to Qdrant [Technical Preview] ReadyFlow details	
Source	Cloudera managed ADLS
Source Format	PDF
Destination	Qdrant
Destination Format	Vector DB

Prerequisites

Learn how to collect the information you need to deploy the ADLS to Qdrant [Technical Preview] ReadyFlow, and meet other prerequisites.

For your data ingest source

• You have your ADLS container, storage account, and path from which you want to ingest data.

- You have performed one of the following to configure access to the ADLS folders:

    - 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 on cloud credentials to access ADLS folders, 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 Data Flow environment.
        2. From the Ranger UI, navigate to the ADLS repository.
        3. 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 (/).

        4. 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.

        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 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.

        3. Click Save and Sync.

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.

For your data ingest target

- You have your OpenAI API key.
- You have your Qdrant instance and collection into which you want to ingest data.
- Your Qdrant collection was configured with the 'text-embedding-ada-002 model'.
- You have your Qdrant API key.

Related Concepts

Required parameters

Required parameters

When deploying the ADLS to Qdrant [Technical Preview] ReadyFlow, you have to provide the following parameters. Use the information you collected in *Prerequisites*.

Table 1: ADLS to Qdrant [Technical Preview] 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 (via IDBroker). Ensure this user has the appropriate access rights to the object store locations.

Parameter name	Description
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).
CDPEnvironment	The Cloudera Environment configuration resources.
OpenAI API Key	Specify the API key used to authenticate to OpenAI.
OpenAI Model Name	Specify the OpenAI model name to use for embedding the data. The default model is 'text-embedding-ada-002'.
Qdrant API Key	Specify the API key used to authenticate to Qdrant.
Qdrant Collection	Specify the name of the destination Qdrant collection.
Qdrant URL	Specify the URL of the destination Qdrant instance.
Source ADLS File System	Specify the name of the ADLS data container you want to read from. The full path will be constructed from: abfs://#{Source ADLS File System}@#{Source ADLS Storage Account}.dfs.core.windows.net/#{Source ADLS Path}
Source ADLS Path	Specify the path within the ADLS data container where you want to read from without any leading characters. The full path will be constructed from: abfs://#{Source ADLS File System}@#{Source ADLS Storage Account}.dfs.core.windows.net/#{Source ADLS Path}
Source ADLS Storage Account	Specify the source ADLS storage account name.

Related Concepts
Prerequisites
Related Information
Deploying a ReadyFlow