

RAG Query Milvus

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: RAG Query Milvus [Technical Preview].....	4
Prerequisites.....	4
List of required configuration parameters for the RAG Query Milvus [Technical Preview] ReadyFlow.....	6

ReadyFlow overview: RAG Query Milvus [Technical Preview]

You can use the RAG Query Milvus [Technical Preview] ReadyFlow to query Milvus VectorDB with an embedded prompt.

This ReadyFlow generates and vectorizes a query from a user-specified question. It then performs a similarity search using one of five search metrics (L2, IP, COSINE, JACCARD, HAMMING) on the specified Milvus collection and returns the result with the shortest distance, extracting the specified fields. The flow then performs a lexical search based on the specified fields and returns all rows with the returned 'section', storing them as a text attribute. A prompt is created from the original user query and generated text document. The prompt is then passed to ChatGPT for an answer to the user query. A Milvus access token and an OpenAI API key are required to run this flow. Define a KPI on the failure_AskChatGPT connection to monitor failed responses.



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 the HuggingFace embedding model specified is the same model used to insert data in the queried Milvus collection and that the Milvus collection has been created with the expected schema.

RAG Query Milvus [Technical Preview] ReadyFlow details	
Source	Custom prompt-query
Source Format	Natural language
Destination	Milvus
Destination Format	Vector DB

Prerequisites

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

For your data ingest source

- You have a query in the form of a question that is relevant to the data in your Milvus collection.

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 data ingest target

- You have the OpenAI API key.
- Your Milvus version is 2.4.4.
- You have the Milvus access token.
- You have the name of the destination Milvus collection.
- You have the URI of the destination Milvus instance.

- Your Milvus collection schema has the following field names and field types:
 - id (*INT64*) You must enable Auto ID on this field.
 - text_embedding (*FLOAT_VECTOR*)
 - source (*VARCHAR*)
 - section (*VARCHAR*)
 - text (*VARCHAR*)
- The 'text_embedding' field in your destination Milvus collection is configured with the same dimensions as the default HuggingFace 'all-MiniLM-L12-v2' model (384).
- The source, section and text VARCHAR fields have been configured with Max Length values large enough to accommodate their respective values in your PDFs.
- Your Milvus collection has data loaded that is relevant to your prompt query. For example, PDFs related to the query have been loaded to Milvus using the “S3 to Milvus” or “ADLS to Milvus” ReadyFlows.



Note: The flow ends with a response from ChatGPT received by a placeholder processor (LogAttribute). Replace it with the processor(s) required to support your use case for the response data. This may be an integration with productivity applications like Slack or Discord, a web UI, or any other customer facing user interface.

-

Related Concepts

[List of required configuration parameters for the RAG Query Milvus \[Technical Preview\] ReadyFlow](#)

List of required configuration parameters for the RAG Query Milvus [Technical Preview] ReadyFlow

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

Table 1: RAG Query Milvus [Technical Preview] ReadyFlow configuration parameters

Parameter name	Description
HuggingFace Embedding Model	Specify the HuggingFace model name to use for embedding the data. The default model is 'all-MiniLM-L12-v2'.
Milvus Access Token	Specify the access token for the destination Milvus instance.
Milvus Collection Name	Specify the name of the destination Milvus collection.
Milvus Instance URI	Specify the URI of the destination Milvus instance.
OpenAI API Key	Specify the API key used to authenticate to OpenAI.
OpenAI Model Name	Specify the OpenAI model name. The default model is 'gpt-4o-mini'.
User Query	Specify the question to include in the ChatGPT prompt.

Related Concepts

[Prerequisites](#)

[Related Information](#)

[Deploying a ReadyFlow](#)