

Getting Started with Public Cloud

Date published: 2024-01-01

Date modified: 2024-08-15

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

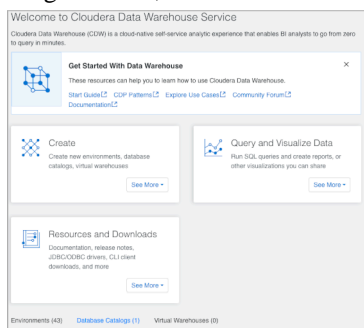
- Getting started in CDW.....4**
 - Prerequisites..... 5
 - Granting CDP users and groups access to CDW..... 5
 - Creating your first Virtual Warehouse.....6
 - Set up SQL AI Assistant.....6
 - Preparing the Microsoft Azure OpenAI service..... 7
 - Preparing the Amazon Bedrock Service..... 7
 - Preparing the OpenAI platform.....8
 - Prerequisites for configuring Hue SQL AI Assistant..... 8
 - Configure SQL AI Assistant in CDW..... 9
 - Service and model-related configurations for setting up the Hue SQL AI Assistant.....11

Getting started in CDW

In Cloudera Data Warehouse (CDW), you can deploy and administer cloud data warehousing. You can move workloads to the cloud securely and manage user access to warehouse data, or just parts of the data, through Apache Ranger fine-grained control. You learn how to get started in CDW.

As CDW administrator, you need privileges to access a CDP environment. As a CDW user, you need privileges to access warehouse data as described in the subtopics, *Prerequisites* and *Granting CDP users and groups access*.

To get started, click Data Warehouse on the CDP home page. The **Overview** page is displayed:



If you have permission to access an environment, you can create a Virtual Warehouse:

- A Hive Virtual Warehouse for Hive users
- An Impala Virtual Warehouse for Impala users
- Unified Analytics-enabled Virtual Warehouse.

Unified Analytics offers Hive/Impala-equivalent SQL syntax for many SQL queries, plus optimizations and enhancements.

You provide users with the URL of your Virtual Warehouse, and they can access the data you authorize using tools you download from **Resources and Downloads** and provide to them from **Resources and Downloads**, you can install or download the following software or drivers:

- DBT Hive and DBT Impala

Adapters for [using dbt](#) with Hive or Impala.

- CDP CLI

Install a tool to help you manage your Service Manager-managed cluster instances. The CLI can be used for automating cluster creation and termination.

- Hive JDBC Jar

Download the Hive JDBC driver for clients to connect to a Virtual Warehouse.

- Beeline CLI

Download the Beeline CLI tarball for clients to install and connect to a Virtual Warehouse.

- Workload Insights

Download this software for getting recommendations for creating materialized views.


- Impala JDBC/ODBC Driver

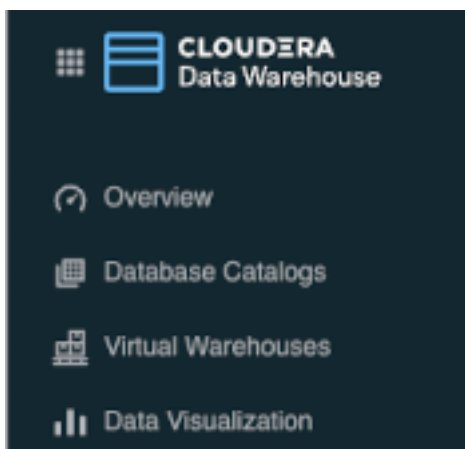
Download the Impala JDBC/ODBC driver for clients to connect to a Virtual Warehouse.

- Unified Analytics JDBC Driver

Download a JDBC driver for clients to connect to a Unified Analytics Virtual Warehouse.

Exiting CDW

In the left navigation, click  as shown in the following image:



Select Management Console to leave CDW to manage environments, users, and CDP services, or select a different menu item.

Prerequisites

To get started in Cloudera Data Warehouse (CDW), your data must conform to supported compression codecs, and you must obtain CDP resource roles to grant users access to an environment. Users can then get started on CDW tasks, such as activating the environment from CDW.

Unsupported compression

CDW does not support LZO compression due to licensing of the LZO library. You cannot query tables having LZO compression in Virtual Warehouses, which use CDW Impala or Hive LLAP engines.

CDP resource roles

You need to have the PowerUser role to change the following CDP resource roles that are associated with the CDW service.

- **DWAdmin:** This role enables users or groups to grant a CDP user or group the ability to activate, terminate, launch, stop, or update services in Virtual Warehouses.
- **DWUser:** This role enables users or groups to view and use CDW clusters (Virtual Warehouses) that are associated with specific environments.

A CDP PowerUser assigns these roles to users who require access to the Virtual Warehouses that are associated with specific environments. A default Database Catalog is associated with your environment. After granting these roles to users and groups, they then have Single Sign-On (SSO) access to the Virtual Warehouses.

Granting CDP users and groups access to CDW

You assign the DWAdmin and DWUser roles to users and groups who must use or manage Virtual Warehouses in Cloudera Data Warehouse (CDW) service on public cloud environments. These roles are not for accessing Kubernetes endpoints in the cluster.

Procedure

1. Log in to the CDP web interface.

2. In the left navigation menu, click Environments.
3. Search for the environment you want to grant access to, and then click the environment name.
4. In the environment's Clusters page, click the Actions drop-down menu on the right, and select Manage Access.
5. In the Access page, in the Select group or user text box, type the name of the user or group you want to add, and then click that user or group's name.
6. In Update Resource Roles for..., select the DWAdmin or DWUser or both roles, and then click Update Roles. You should receive a success message.
7. Click the environment name at the top of the page to navigate back to the Clusters page.
8. In the Clusters page, click Actions, and select Synchronize Users to FreeIPA.
Depending on how many users have access to the environment, this synchronization process can take a few seconds or a few minutes.

Results

The users and groups you have granted the DWAdmin or DWUser resource roles to can now use CDW service with the environment.

Creating your first Virtual Warehouse

After meeting prerequisites, you learn how to activate the environment from Cloudera Data Warehouse (CDW). You then create a Virtual Warehouse.

Before you begin

Before you can create a Virtual Warehouse in CDW service, you must have completed the following tasks:

- Register an environment with Management Console. This sets up your Data Lake. Depending on which cloud platform you are using, see [Register an AWS environment](#) or [Register an Azure environment](#).
- Grant the DWAdmin role to the user or group that needs to create a Virtual Warehouse. This user or group determines which environment and Data Lake your Virtual Warehouse uses. See [Granting CDP users access to Cloudera Data Warehouse service](#).

Procedure

1. Activate the environment that uses the Data Lake. See [Activating AWS environments](#) or [Activating Azure environments](#).
A default Database Catalog is created.
2. Add a new Virtual Warehouse. See [Creating a Virtual Warehouse](#).

Results

You can start executing workloads in the new Virtual Warehouse.

About setting up the Hue SQL AI Assistant

Administrators are required to set up and enable the SQL AI Assistant before Analysts can use it to generate, edit, optimize, and fix queries using natural language in Hue.

First, you must obtain clearance from your organization's infosec team to make sure it is safe to use the SQL AI Assistant because some of the table metadata and data, as mentioned in the previous section, is shared with the LLM.

Next, prepare one of the following AI platforms of your choice and then configure the SQL AI Assistant in Hue:

- Microsoft Azure OpenAI service,
- Amazon Bedrock service, or
- OpenAI platform

Preparing the Microsoft Azure OpenAI service

Microsoft Azure allows dedicated deployments of OpenAI GPT models. Using Azure's OpenAI service is much more secure than the publicly hosted OpenAI APIs because the data can be processed in your Virtual Private Cloud (VPC) network. Due to security considerations, Cloudera recommends that you use GPT models in Hue SQL AI Assistant with Azure's OpenAI service.

Procedure

1. Obtain a Microsoft Azure subscription by working with your organization's IT team.
Subscriptions vary based on your team and purpose.
2. Register to access the Azure OpenAI service.
Azure OpenAI requires registration and is currently only available to approved enterprise customers and partners. Customers who wish to use Azure OpenAI are required to submit a [registration form](#).
3. Create an Azure OpenAI resource in the Azure portal.
4. Go to Overview from the left navigation and obtain the resource URL and resource keys from the Develop tab under the resource details page.
You can use any one of the two available keys.
5. Go to Azure OpenAI Studio at <https://oai.azure.com/portal> and create your deployment under Management Deployments .
6. Select gpt-35-turbo-16k or higher.

What to do next

Enable the SQL AI Assistant in the Cloudera Data Warehouse service.

Related Information

[Configuring the SQL AI Assistant in Cloudera Data Warehouse](#)

Preparing the Amazon Bedrock Service

Amazon Bedrock is a fully managed service that makes foundation models from leading AI startups and Amazon available through an API.

Before you begin

You must have an AWS account with Bedrock access.

Procedure

1. Log in to the Amazon Bedrock service.
2. Obtain your access key and secret as follows:
 - a) Go to the IAM console: <https://console.aws.amazon.com/iam>.
 - b) Click on Users from the left menu and select the user you want to access.
 - c) Click on Security credentials.
 - d) Go to the Access keys section and note the access keys.
3. Establish Anthropic Claude access.
Claude from Anthropic is one of the best models available in Bedrock for SQL-related tasks. By default, Claude is not available on Bedrock. You need to place a special request for Claude.
After gaining access, you can try Claude in the text playground under the Amazon Bedrock service. If you are in the us-east-1 region, this must take you to <https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/text-playground>.

What to do next

Enable the SQL AI Assistant in the Cloudera Data Warehouse service.

Related Information

[Configuring the SQL AI Assistant in Cloudera Data Warehouse](#)

Preparing the OpenAI platform

Learn how to set up SQL AI Assistant on the OpenAI platform.

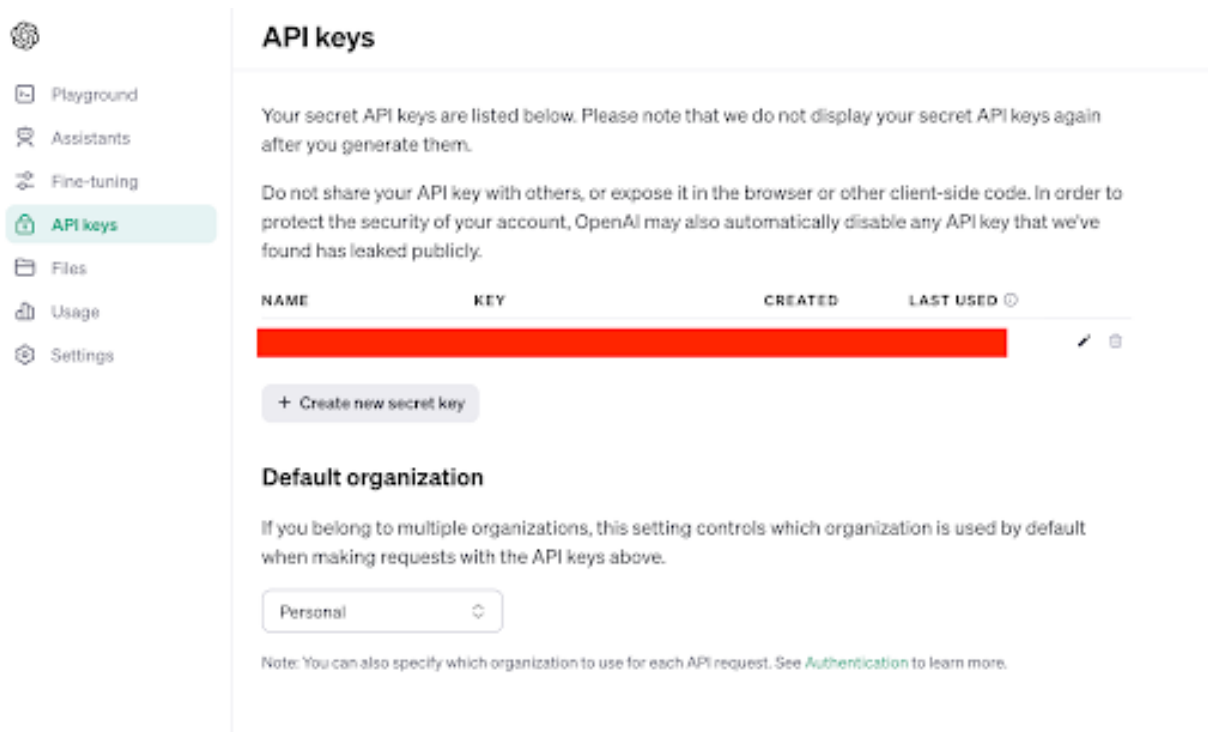
Before you begin

You must have created an account with the OpenAI platform.

Procedure

1. Log in to the OpenAI portal.
2. Obtain the API key by navigating to the API keys menu from the left navigation pane.

The key is required to integrate Hue SQL AI Assistant with the OpenAI service.



What to do next

Enable the SQL AI Assistant in the Cloudera Data Warehouse service.

Related Information

[Configuring the SQL AI Assistant in Cloudera Data Warehouse](#)

Prerequisites for configuring Hue SQL AI Assistant

To configure the SQL AI Assistant in Hue, you must pass the token required for connecting to the LLM service. Learn about the open and secure approaches to pass the tokens, and use the one that fits your organization policy.

(Recommended) Secure approach for passing a token to configure SQL AI Assistant in Cloudera Data Warehouse

In this approach, you use Kubernetes' method of distributing secrets. You first encode the credentials and then add the encoded bit as a data item in the HUE_AI_INTERFACE_TOKEN property. The token becomes available in the Hue pod as an environment variable.

About this task



Note: Secrets are lost when you rebuild the Virtual Warehouse. You need to redo this step to continue using encoded credentials.



Note: This method is supported only on the OpenAI platform and Microsoft Azure OpenAI service. It is unsupported on Amazon Bedrock.

Procedure

1. Use a base64 encoding tool to convert your token to a base-64 representation by running the following command:

```
echo -n ' [***MY-TOKEN***] ' | base64
```

Replace `[***MY-TOKEN***]` with the token value you want to encode.

2. Open a terminal session and run the following command to add the encoded secret:

```
kubectl edit secret hue-secret -n [***VIRTUAL-WAREHOUSE-NAMESPACE***]
```

Replace `[***VIRTUAL-WAREHOUSE-NAMESPACE***]` with the actual Virtual Warehouse ID (same as the namespace) in which you want to add the secret.

3. Add the encoded value returned for your token in the `HUE_AI_INTERFACE_TOKEN` property as follows:

```
...
apiVersion: v1
data:
  HADOOP_CREDSTORE_PASSWORD: [***ENCODED-HADOOP-CREDSTORE-PASSWORD***]
  HUE_AI_INTERFACE_TOKEN: [***ENCODED-TOKEN-VALUE***]
kind: Secret
```

Replace `[***ENCODED-TOKEN-VALUE***]` with the actual encoded value returned for your token.

Open approach for passing a token to configure Hue SQL AI Assistant

In this approach, you specify the token value in the `hue-safety-valve` field in Cloudera Data Warehouse. The credentials are saved in a configuration file in the plain text format.




Note:

- Cloudera recommends that you use the open approach to pass tokens in test deployments, for proof of concept use cases. Use the secure approach in production deployments.
- Amazon Bedrock does not support encoded tokens. Therefore, to use the Hue SQL AI Assistant on Amazon Bedrock, you must use the open approach to configure the SQL AI Assistant on Amazon Bedrock.

Configuring the SQL AI Assistant in Cloudera Data Warehouse

Learn how to configure the SQL AI Assistant in Cloudera Data Warehouse to use it in Hue.

Procedure

1. Log in to the Data Warehouse service as DWAdmin.
2. Go to the Virtual Warehouses tab, locate the Virtual Warehouse on which you want to enable this feature, and click  Edit .

- Go to Configurations Hue , select hue-safety-valve from the Configuration files drop-down menu, and add the following lines depending on the approach you are using for passing the token and your platform:

For Secure token

Azure

```
[desktop]
[[ai_interface]]
  service='azure'
  model_name='[***DEPLOYMENT-NAME***]'
  base_url="https://[***RESOURCE***].openai.azure.com/"
```

OpenAI

```
[desktop]
[[ai_interface]]
  service='openai'
```

For Open token

Azure

```
[desktop]
[[ai_interface]]
  service='azure'
  model_name='[***DEPLOYMENT-NAME***]'
  base_url="https://[***RESOURCE***].openai.azure.com/"
  token="[***RESOURCE-KEY***]"
```

AWS

```
[aws]
[[bedrock_account]]
  access_key_id='[***ACCESS-KEY***]'
  secret_access_key='[***SECRET-KEY***]'
  region='us-east-1'
[desktop]
[[ai_interface]]
  service='bedrock'
  model='claude'
```



Note: You can also use a newer model such as Claude 3 by specifying model='claude3'.

OpenAI

```
[desktop]
[[ai_interface]]
  service='openai'
  token='[***API-KEY***]'
```

- Click Apply Changes.

Results

You see ✨ Assistant on the Hue SQL editor.

Related Information

[Preparing the Microsoft Azure OpenAI service](#)

[Preparing the Amazon Bedrock Service](#)

[Preparing the OpenAI platform](#)

Service and model-related configurations for setting up the Hue SQL AI Assistant

Review the list of service, model, and semantic search-related configurations used for custom configuring the AI services and models you want to use with the SQL AI Assistant and how to specify them in the Hue Advanced Configuration Snippet in Cloudera Data Warehouse (CDW) web interface.

List of service and model-related configurations

You can configure the AI services and models you want to use by going to **CONFIGURATIONS** Hue hue-safety-valve and adding the following lines:

```
[desktop]
[[ai_interface]]
  [***CONFIG-KEY1***]= '[***VALUE***]'
  [***CONFIG-KEY2***]= '[***VALUE***]'
[[semantic_search]]
  [***CONFIG-KEY1***]= '[***VALUE***]'
  [***CONFIG-KEY2***]= '[***VALUE***]'
```

Specify the service and model-related configurations under the `[[ai_interface]]` section as listed in the following table:

AI interface config key	Description
service	API service to be used for AI tasks. AI is disabled when a service is not configured. For example, <code>gpt</code> and <code>ai_assistant</code> .
trusted_service	Indicates whether the LLM is trusted or not. Turn on to disable the warning. The default value is <code>False</code> .
model	The AI model you want to use for AI tasks. For example, <code>gpt</code> and <code>llama</code> .
model_name	The fully qualified name of the model to be used. For example, <code>gpt-3.5-turbo-16k</code> .
base_url	Service API base URL.
add_table_data	When enabled, sample rows from the table are added to the prompt. The default value is <code>True</code> .
table_data_cache_size	Size of the LRU cache used for storing table sample data.
auto_fetch_table_meta_limit	Number of tables to load from a database, initially.
token	Service API secret token.
token_script	Provides a secure way to get the service API secret token.

List of semantic search-related configurations

Specify the semantic search-related configurations used for RAG under the `[[semantic_search]]` section, as listed in the following table:

Semantic search config key	Description
relevancy	The technology you want to use for semantic search. Acceptable values are <code>vector_search</code> or <code>hybrid_search</code> .
embedding_model	The model you want to use for data-embedding. This must be compatible with SentenceTransformer.
cache_size	Size of the LRU cache used for storing embedding.