

Cloudera AI

Cloudera AI Discovery & Exploration

Date published: 2020-07-16

Date modified: 2025-10-31

CLOUDERA

<https://docs.cloudera.com/>

Legal Notice

© Cloudera Inc. 2025. 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

- Exploratory Data Science and Visualization..... 4**
- Prerequisites for Cloudera AI Discovery and Exploration..... 4**
- Starting Data Discovery and Visualization.....5**
 - Workarounds for Cloudera Data Visualization with Hive and Impala.....6
- Working with Data Discovery and Visualization.....9**
- Data connection management.....10**
- Managing default and backup data connections..... 11**
- API permissions for Projects.....11**
- Troubleshooting: 401 Unauthorized..... 12**
- Troubleshooting: 401 Unauthorized when accessing Hive or Impala virtual
warehouses..... 13**
- Troubleshooting: Existing connection name.....13**
- Troubleshooting: Empty data page..... 14**

Exploratory Data Science and Visualization

Exploratory Data Science and Visualization makes it simple for data scientists to get started on a data science project.

When you start a data science project, you are presented with a blank notebook, and no indication of what data sources are available on the Cloudera platform, or how to access them. The Exploratory Data Science and Visualization feature automatically discovers the data sources available to you, from within the standard Cloudera AI user interface.

The Exploration Data Science and Visualization experience enables all features for Exploratory Data Analysis. From the Data tab you can:

- Connect to data sources that are available in your project
- Explore the data with SQL to understand its basic shape and characteristics
- Create named datasets that you can reference later
- Quickly create visualizations of the data to understand its properties
- Create dashboards that you can share with your team

For more information on what you can do in the Data tab, see the documentation for *Cloudera Data Visualization*.

Related Information

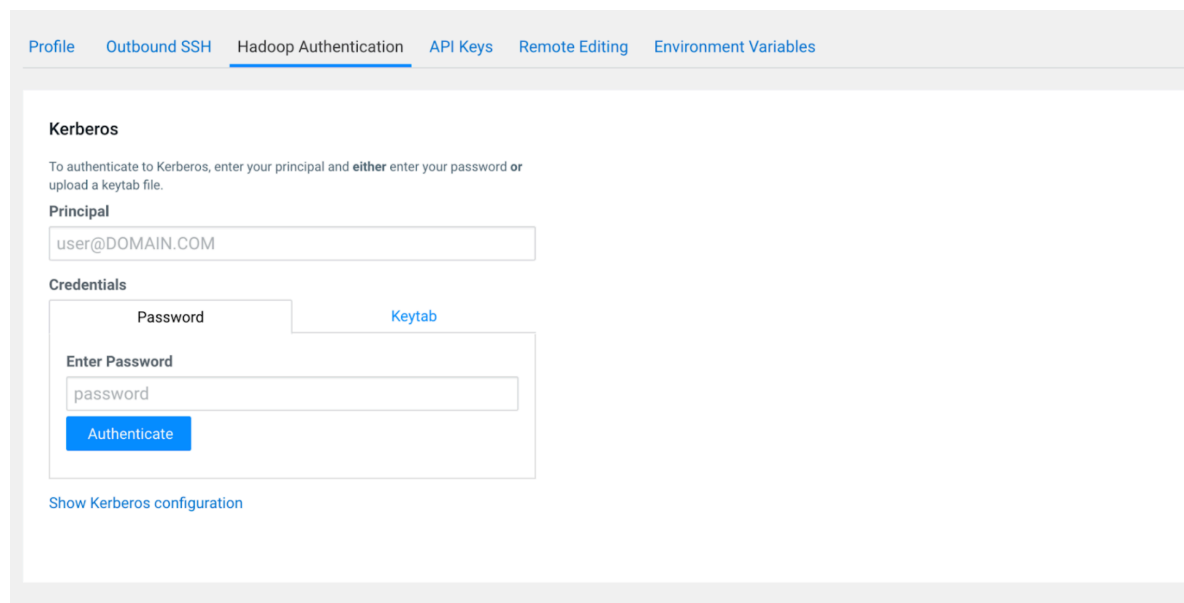
[Cloudera Data Visualization](#)

Prerequisites for Cloudera AI Discovery and Exploration

Cloudera AI Discovery and Exploration has a few prerequisites.

- For both Hive and Impala connections, the HADOOP_USERNAME must be set. The HADOOP_USERNAME is set automatically by the environment.
- Go to your **User Settings** page, select the **Hadoop Authentication** tab and log in.

Figure 1: Hadoop authentication



The screenshot shows the 'Hadoop Authentication' tab in the Cloudera AI user interface. The tab is part of a navigation bar that includes 'Profile', 'Outbound SSH', 'Hadoop Authentication', 'API Keys', 'Remote Editing', and 'Environment Variables'. The 'Hadoop Authentication' tab is active. Below the navigation bar, the 'Kerberos' section is displayed. It contains instructions: 'To authenticate to Kerberos, enter your principal and either enter your password or upload a keytab file.' There are two input fields: 'Principal' with the value 'user@DOMAIN.COM' and 'Credentials'. The 'Credentials' section has two tabs: 'Password' and 'Keytab'. The 'Password' tab is active, showing an 'Enter Password' field with the value 'password' and an 'Authenticate' button. A link 'Show Kerberos configuration' is at the bottom.

- To manually create Impala or Hive data connections, you must have the related JDBC URL. You can obtain this from the option menu for the virtual warehouse to which you want to connect. These warehouses must be already

created in the environment. For more information on how to create a data warehouse, see *Adding a New Virtual Warehouse*.

- To manually create a Spark connection, you need the Data Lake S3 bucket. This can be retrieved from the Data Lake page. For Spark data connections, you must use ML Runtimes, and specifically the Spark Runtime Add-on must be enabled before starting a workload (job or session).
- For Spark data connections, you must have permissions set correctly for external access to the S3 bucket.
- Hive and Impala data connections also require ML Runtimes. Legacy engines may work, but are not supported.
- There must be auto-discovered data connections, or your Administrator must create data connections to virtual warehouses or other data sources in your Cloudera environment.

Known Issue

- Only Hive virtual warehouses with SSO disabled are supported.

Related Information

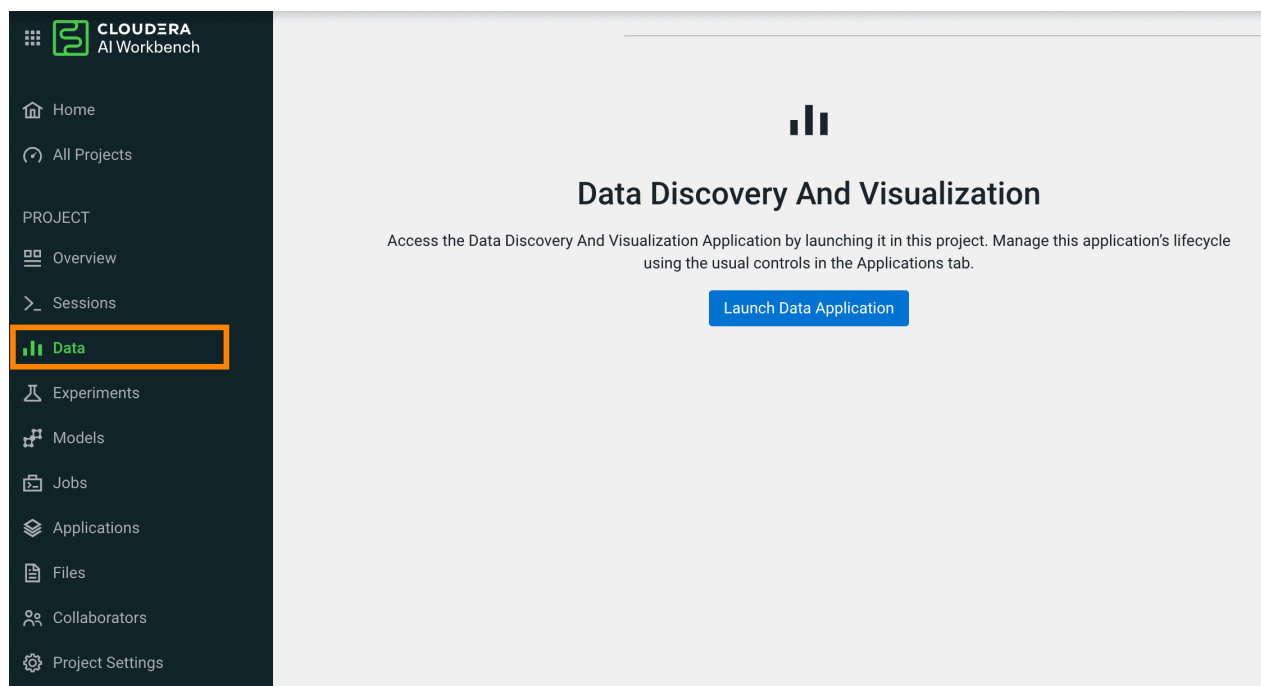
[Create a Dashboard](#)

[Adding a New Virtual Warehouse](#)

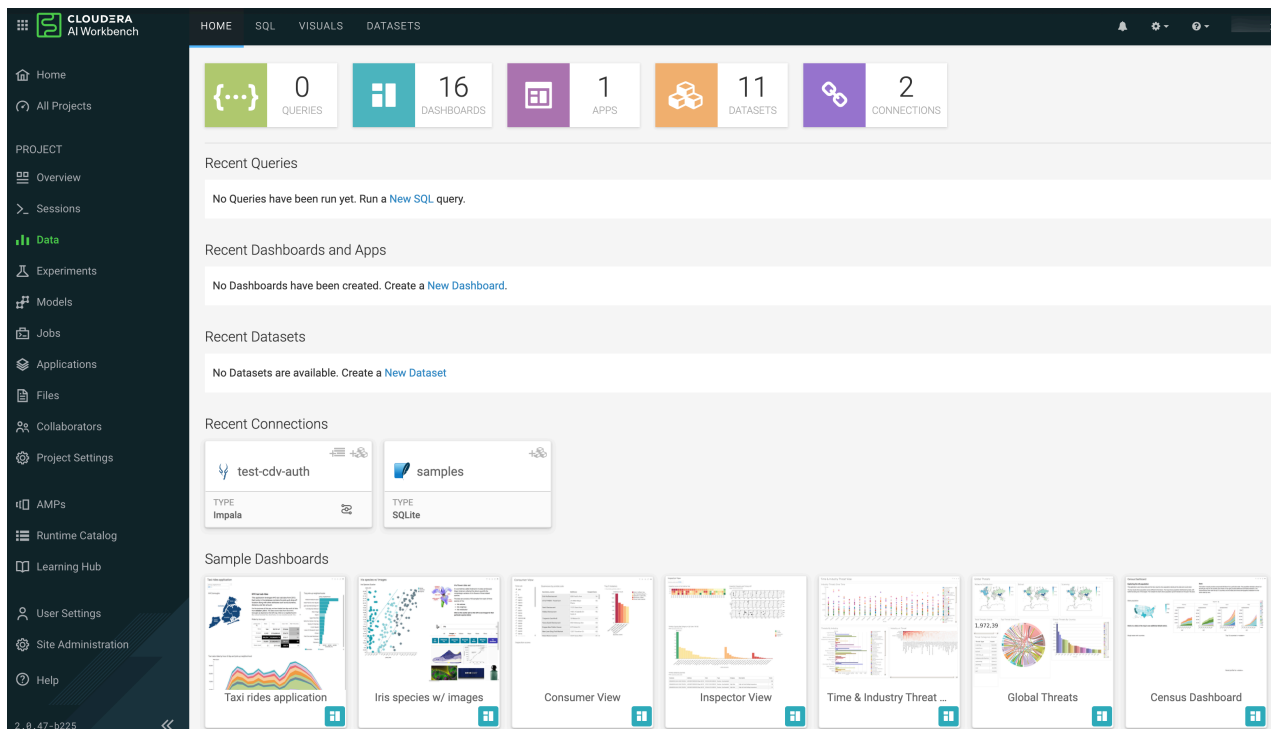
Starting Data Discovery and Visualization

You can start the Exploratory Data Science and Visualization application within Cloudera AI.

1. Create or open a project.
2. In the project, click Data in the left navigation pane.
3. If the application was previously started, it appears in the UI.
4. If it is not displayed, click Launch Data Application to start the application. It takes a few minutes to start for the first time.



When the application starts, you can see four tabs at the top of the UI. From here, you can create new dashboard or dataset or try some of the tasks described in the next topics.



Related Information

[Creating a dataset](#)

[Create a Dashboard](#)

[Troubleshooting: 401 Unauthorized](#)

Workarounds for Cloudera Data Visualization with Hive and Impala

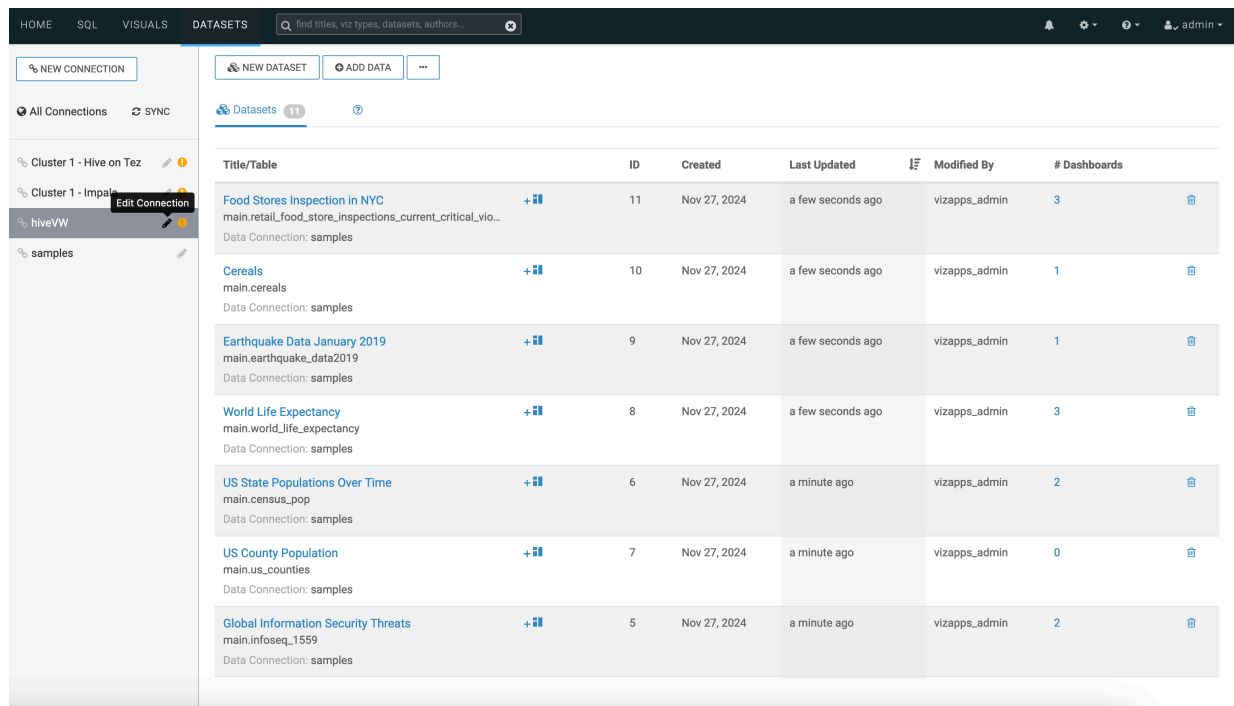
Follow the guidelines to start Cloudera Data Visualization on premises with Hive and Impala Virtual Warehouses and Hive and Impala Base.

Starting Cloudera Data Visualization on premises for Hive and Impala Virtual Warehouses

For starting Cloudera Data Visualization on premises for Hive and Impala Virtual Warehouses, switch to LDAP **Authentication mode** and add your username and password in the **Edit Data Connection** page:

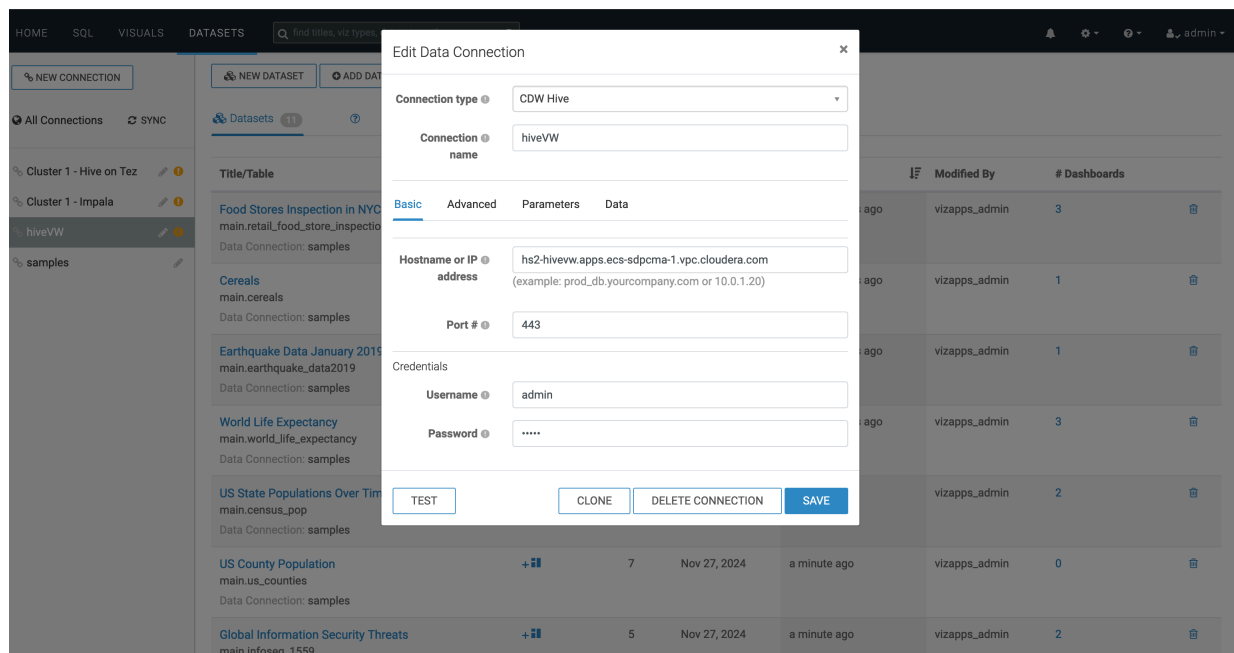
1. Select the required connection in the left navigation and click Edit Connection.

Figure 2: Selecting the required connection for editing



2. Provide your username and password.

Figure 3: Provide username and password

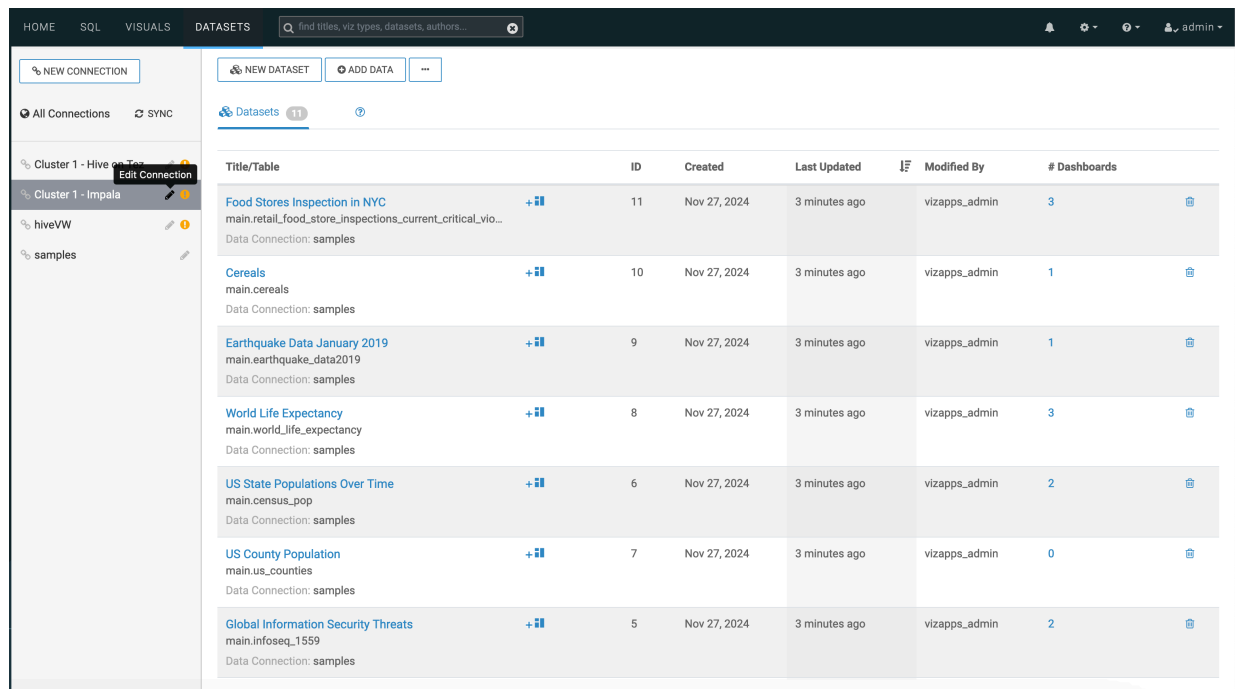


Starting Cloudera Data Visualization on premises for Hive and Impala

For starting Cloudera Data Visualization on premises for Hive and Impala switch to Kerberos **Authentication mode** in the **Edit Data Connection** page:

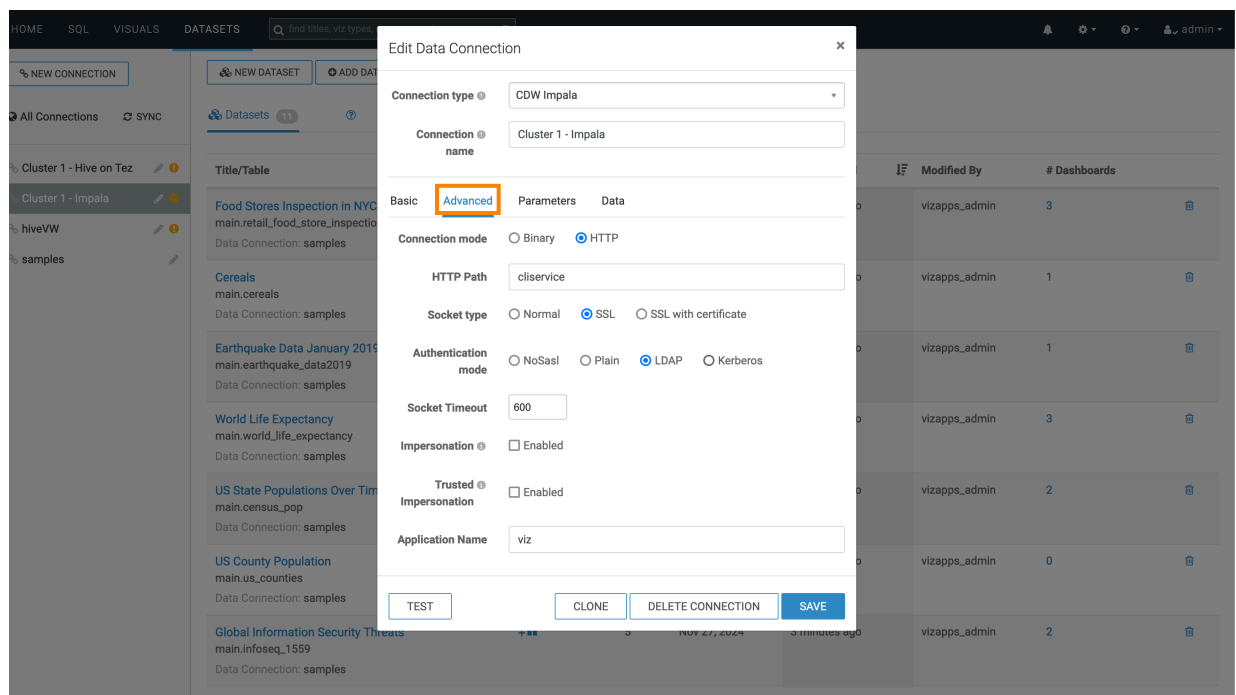
1. Select the required connection in the left navigation and click Edit Connection.

Figure 4: Selecting the required connection for editing



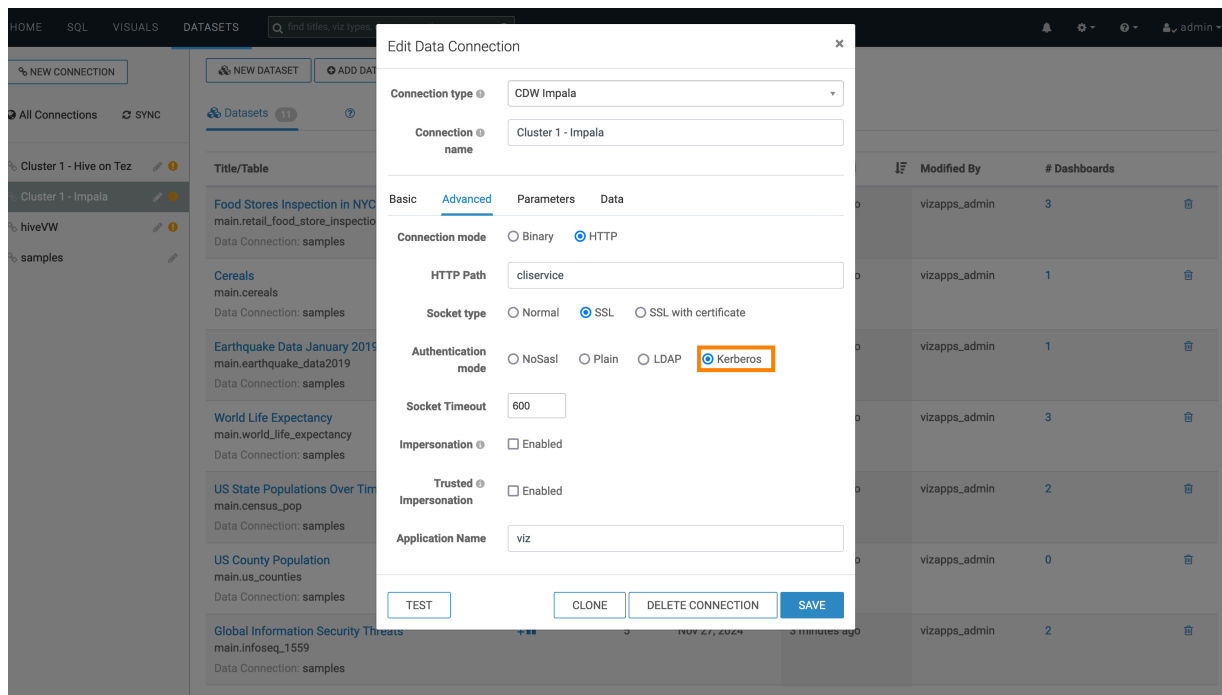
2. Select the **Advanced** tab in the **Edit Data Connection** window.

Figure 5: Selecting the Advanced tab



3. Select Kerberos as Authentication mode in the **Edit Data Connection** window.

Figure 6: Select Kerberos for authentication



Working with Data Discovery and Visualization

This section provides examples for tasks you can perform using Data Discovery and Visualization.

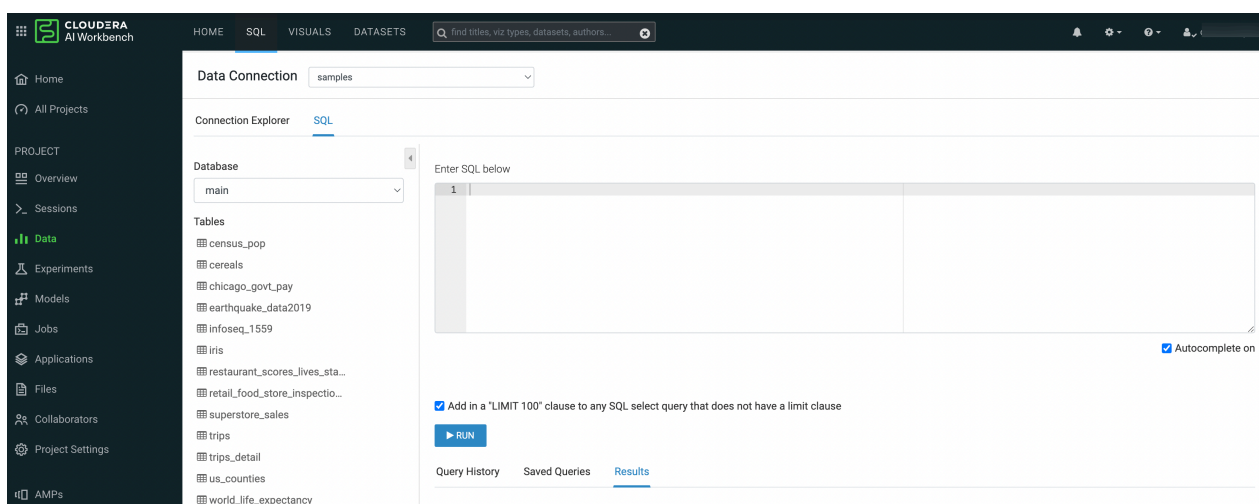
Run an SQL query

On the **SQL** tab, select a Data Connection from the dropdown box.



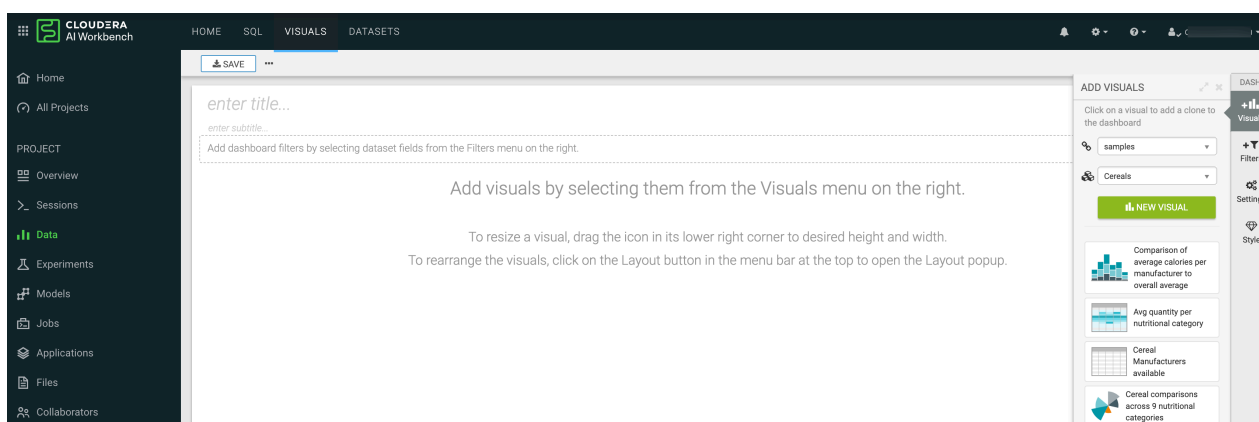
Note: If you do not see the data connections that you have configured in the Project, refer to the *Troubleshooting* details for more information.

Once you select a Data Connection, you can explore the available databases and tables. You can start writing an SQL query and click the table names to insert them into the query string. For more information, see *Creating a dataset*.



Build a dashboard

Click New Dashboard to access the Dashboard Designer. You can select from a wide variety of graphs, click to add them to your dashboard, and select the columns from your dataset to use in the visual. For more information, see *Create a Dashboard*.



Related Information

[Creating a dataset](#)

[Create a Dashboard](#)

[Troubleshooting: 401 Unauthorized](#)

Data connection management

Consider the following guidelines about data connection management.

- Manage data connections in a workbench

At the workbench level, you can check the data connections that are available in a workbench. In **Project Settings Data Connections**, check that your desired data source is present. You can also set the availability for any discovered connections, if necessary.

- Manage data connections in a project

All the data connections that were available in the workbench when the project was created will be automatically created in the project as well. The available connections can then further be marked as unavailable if desired so.

You can update any changes to the connections that were made at the workbench level by clicking Sync with Workbench. Any changes made here only apply to your project.

- Data connection availability

Consider these two scenarios for setting data connection availability.

1. If a workbench data connection is marked Unavailable, and you then create a project, the data connection will not appear in the project. If the connection is then changed to Available, and then the Sync with Workbench button is clicked, the connection will appear in the project.
2. If a workbench data connection is marked Available, and you then create a project, the connection shows up. If the workbench data connection is then toggled to Unavailable, and you click Sync with Workbench in the project, the data connection will remain available in the project.

Managing default and backup data connections

As a data scientist, you may want to set default and backup data connections (virtual warehouses). This makes it easy to manage the case where the default data source (virtual warehouse or data lake) becomes unavailable, for example.

Before you begin

You need Administrator privileges to perform this task. Here, the data connection names of **default** and **backup** are used as examples.

Procedure

1. In Site Administration Data connections, click Edit to change the name of the default data connection to default.
2. Change another connection to backup.
3. In Project Settings, click Sync with Workbench to update the names of the data connections. If two different data connections have the same name, an error occurs during synchronization.
4. If the default data connection becomes unavailable, the Workbench Administrator can go to the Data Connections tab, and rename the connections.

For example, after changing default to unavailable, change backup to default. Projects that use library code for their connection to the default data connection will continue to operate, because it is now using the new default connection.



Note: Select Sync with Workbench at the project level to make the updated connections available.

API permissions for Projects

The tasks available to collaborators depends on their level of access.

A collaborator with read-only access is able to do the following:

- List data connections
- Get data connection details



Note: The list of data connections is unavailable in the UI to viewer-only users because the Setting tab in a project is hidden for viewer collaborators. However, viewer-only users can use curl to submit API requests to list the data connections.

A collaborator with write access can:

- Create data connections
- Edit data connections

- Delete data connections
- List data connections
- Get data connection details

Troubleshooting: 401 Unauthorized

Troubleshoot the session returning a 401: Unauthorized error message.

When you are in a session and try to run the code for a Hive or Impala connection, the session returns a 401: Unauthorized HTTP error message.

Figure 7: 401: Unauthorized HTTP error message



```

modify next 2 lines to update your credentials

> USERNAME = os.getenv('HADOOP_USER_NAME')
> PASSWORD = os.getenv('WORKLOAD_PASSWORD')
> conn = cmldata.getConnection({
    'CONNECTION_NAME': CONNECTION_NAME,
    'USERNAME': USERNAME,
    'PASSWORD': PASSWORD
})

HttpError: HTTP code 401: Unauthorized
HttpError                                Traceback (most recent call last)
<ipython-input-1-a69c5e2af83b> in <module>
      2     'CONNECTION_NAME': CONNECTION_NAME,
      3     'USERNAME': USERNAME,
----> 4     'PASSWORD': PASSWORD
      5 })

~/local/lib/python3.7/site-packages/cmldata.py in getConnection(properties)
    105     fmt_codesnippet = _getConnectionSnippet(properties)
    106     scope = {}
--> 107     exec(fmt_codesnippet, scope)
    108     return scope["conn"]

<string> in <module>

<string> in getCursor(self)

/usr/local/lib/python3.7/site-packages/impala/hiveserver2.py in cursor(self, user, configuration, convert_type
    127         log.debug('cursor(): getting new session handle')
> |

```

Solution: Go to the Hadoop authentication page:

Figure 8: Access Hadoop authentication page

```

1 import cml.data.v1 as cmldata
2
3 CONNECTION_NAME = "hivel"
4 conn = cmldata.get_connection(CONNECTION_NAME)
5
6 ## Sample Usage to get pandas data frame
7 EXAMPLE_SQL_QUERY = "show databases"
8 dataframe = conn.get_pandas_dataframe(EXAMPLE_SQL_QUERY)
9 print(dataframe)
10 # Closing the connection
11 conn.close()
12
13 ## Other Usage Notes:
14
15 ## Alternate Sample Usage to provide different credentials as optional parameters
16 #conn = cmldata.get_connection(
17 #    CONNECTION_NAME, {"USERNAME": "someuser", "PASSWORD": "somepassword"}
18 #)
19
20 ## Alternate Sample Usage to get DB API Connection interface
21 #db_conn = conn.get_base_connection()
22
23 ## Alternate Sample Usage to get DB API Cursor interface
24 #db_cursor = conn.get_cursor()
25 #db_cursor.execute(EXAMPLE_SQL_QUERY)
26 #for row in db_cursor:
27 #    print(row)

```

```

> import cml.data.v1 as cmldata
> CONNECTION_NAME = "hivel"
> conn = cmldata.get_connection(CONNECTION_NAME)
KeyError: 'No password specified and JWT authentication is disabled. Please edit WORKLOAD_PASSWORD environment variable'
Traceback (most recent call last)
Cell In[1], line 1
----> 1 conn = cmldata.get_connection(CONNECTION_NAME)

File /opt/cmladdons/python/site-packages/cml/data_v1/data.py:136, in get_connection(dataconnection_name, param
134 if dataconn_props["type"] == "CUSTOM":
135     return _get_custom_connection(dataconnection_name, dataconn_props)
--> 136 return ConnectionClasses[dataconn_props["type"]](dataconn_props)

File /opt/cmladdons/python/site-packages/cml/data_v1/hiveconnection.py:17, in HiveConnection.__init__(self, pr
16 def __init__(self, properties):
--> 17     self.update_credentials(properties)
18     self.update_properties(properties)
19     self.hostname = properties.get("HOSTNAME")

File /opt/cmladdons/python/site-packages/cml/data_v1/hiveconnection.py:150, in HiveConnection.update_credentia
147     print("Invalid content in jwt file")
149 if properties["PASSWORD"] is None and isJwtEnabled is None:
--> 150     raise KeyError(
151         "No password specified and JWT authentication is disabled. Please edit WORKLOAD_PASSWORD enviri
152         + "variable under User Settings > Environment Variables"
153     )
154 return properties

KeyError: 'No password specified and JWT authentication is disabled. Please edit WORKLOAD_PASSWORD environment

```

Set your workload password. After setting the workload password, start a new session. If this error occurs in the Data tab, then restart the Data Discovery and Exploration application.

Troubleshooting: 401 Unauthorized when accessing Hive or Impala virtual warehouses

Troubleshoot the session returning a 401: Unauthorized error message when you are accessing a Hive database or Impala virtual warehouses.

Check that the Hive or Impala Virtual Warehouse is working.

Solution: Follow these steps to check that Hive or Impala data warehouse is running, and restart it if needed.

1. In the control plane, go to Cloudera Data Warehouse.
2. Click Virtual Warehouses, and select the data warehouse for the connection.
3. Check the status of that data warehouse, and make sure that it is running or in a good state.

Troubleshooting: Existing connection name

Troubleshoot receiving the error message that the CRN or the name is a duplicate, when attempting to synchronize data connections.

Project Settings

Options Runtime/Engine Advanced SSH Tunnels Data Connections Delete Project

⚠ Unable to sync some connections since they have the same name as existing connections. Please rename the connections listed below, then click on "Sync with workspace" again:

- test

Show Available Only ☐

[Sync with Workspace](#)
[New Connection](#)

Availability in Project	Connection Name	Connection Type	Virtual Warehouse Name	Created At	Actions
<input checked="" type="checkbox"/> Available	test	Hive Virtual Warehouse		08/30/2021 2:14 PM	Edit Copy Delete

[<](#)
[1](#)
[>](#)

Solution: This indicates a project connection (one that is not copied from the workbench) has the same name or CRN as a workbench connection. To resolve this, you need to change the name or the CRN of the data connection at the project level.

Troubleshooting: Empty data page

Troubleshoot receiving an empty data page with non-ending spinner after launching the application.

- On the applications tab, click **Data Discovery and Visualization application**, and check the logs. Check if another user stopped or restarted the application.
- On the applications tab, click on Settings Resource Profile , and check if the available resources are sufficient. You can increase the resources if necessary, then restart the application.