Machine Learning

ML Discovery & Exploration

Date published: 2020-07-16
Date modified: 2022-04-11

https://docs.cloudera.com/

Legal Notice

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

Machine Learning Discovery and Exploration

ML Discovery and Exploration 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 CDP platform, or how to access them. The ML Discovery and Exploration feature automatically discovers the data sources available to you, from within the standard CML user interface.

With ML Discovery and Exploration, you can:

- Easily discover and connect to data sources (virtual warehouses or data lakes).
- Quickly set up an ML project to access a given data source.
- Start running SQL queries to explore the data.

Prerequisites

Machine Learning Discovery and Exploration has a few prerequisites.

- You must have Admin (write) permissions to set up or edit data connections. Only read permissions are needed to list data connections or view data connection details such as the library code.
- To manually set up data connections, you must have the JDBC URL (for Impala or Hive connections) or the Data Lake external directory URL (for Spark connections). You can obtain this from the option menu for the virtual warehouse to which you want to connect.
- One or more data warehouses must be already created in the environment. For more information on how to create a data warehouse, see *Adding a New Virtual Warehouse*.
- For Spark data connections, you must use ML Runtimes, and specifically the Spark Runtime Add-on must be available.
- Hive and Impala data connections also require ML Runtimes. Legacy engines may work, but are not supported.
- For Spark data connections, you must have permissions set correctly for external access to the S3 bucket.
- For both Hive and Impala connections, the HADOOP_USERNAME and WORKLOAD_PASSWORD must be set. The HADOOP_USERNAME is set automatically by the environment. To set the WORKLOAD_PASSWORD, see *Using data connections in a project*.
- For Hive data connections, make sure that SSO is disabled.

Related Information

Adding a New Virtual Warehouse

Using data connections in a project

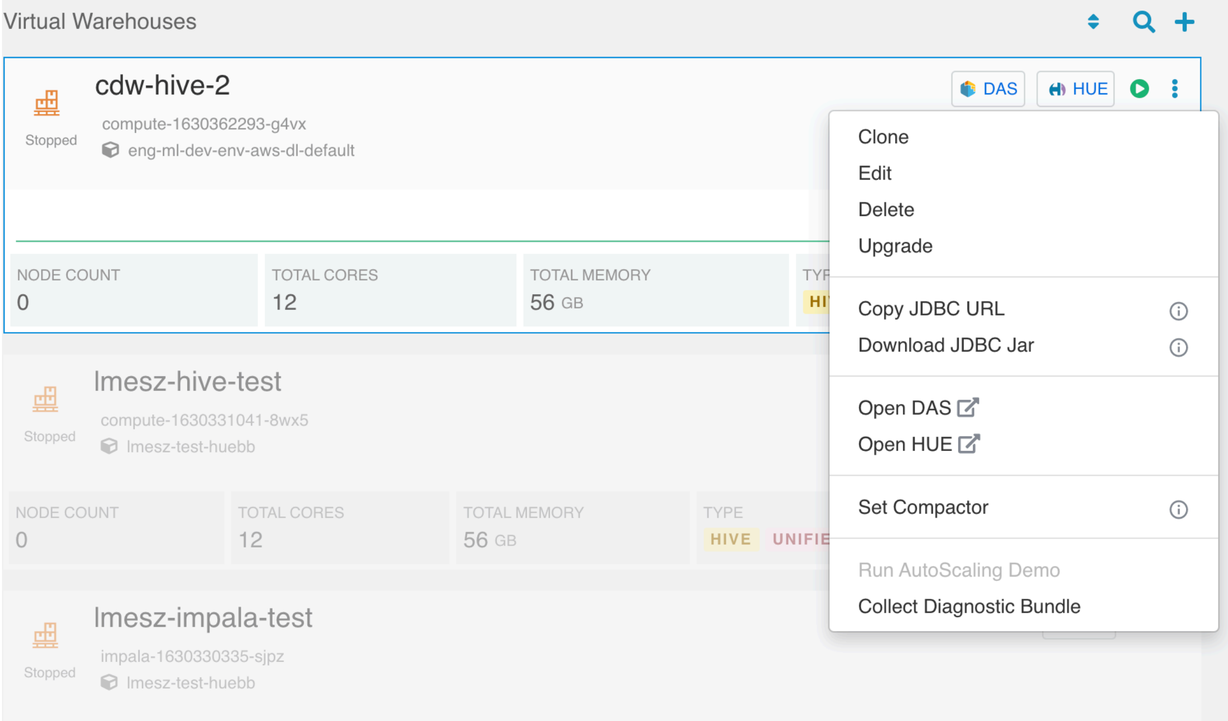
Set up a data connection manually

Data connections are automatically discovered. You can also set up a data connection manually, which works across CDP environments.

Procedure

1. Log into the CDP web interface and navigate to the Data Warehouse service.
2. In the Data Warehouse service, select Virtual Warehouses in the left navigation panel.

3. Select the options menu for the warehouse you want to access, and select Copy JDBC URL.



4. In Site Administration Data Connections , select New Connection.
5. Enter the connection name. You cannot have duplicate names for data connections within a workspace or within a given project.
6. Select the connection type:

   a. Hive Virtual Warehouse
   b. Impala Virtual Warehouse
7. Enter the JDBC URL for the data connection. You can obtain this from the virtual warehouse in Cloudera Data Warehouse.
8. (Optional) Enter the Virtual Warehouse Name. This is the name of the warehouse in Cloudera Data Warehouse.

Results

The data connection is available to users by default. To change availability, click the Available switch.

Data connection management

There are a few things to keep in mind about data connections.

- Manage data connections in a workspace

  At the workspace level, you can check the data connections that are available in a workspace. 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

  At the project level, you can see all data connections that were made available in the workspace when the project was created. The available connections can then further be marked as unavailable if so desired. You can update any changes to the connections that were made at the workspace level by clicking Sync with Workspace. Any changes made here only apply to your project.

- Data connection availability

  Keep in mind these two scenarios for setting data connection availability.

  1. If a workspace 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 Workspace button is clicked, the connection will appear in the project.
  2. If a workspace data connection is marked Available, and you then create a project, the connection shows up. If the workspace data connection is then toggled to Unavailable, and you click Sync with Workspace in the project, the data connection will remain available in the project.

Using data connections in a project

As a data scientist, you can manage the data connections within a project, and connect your code to the data.

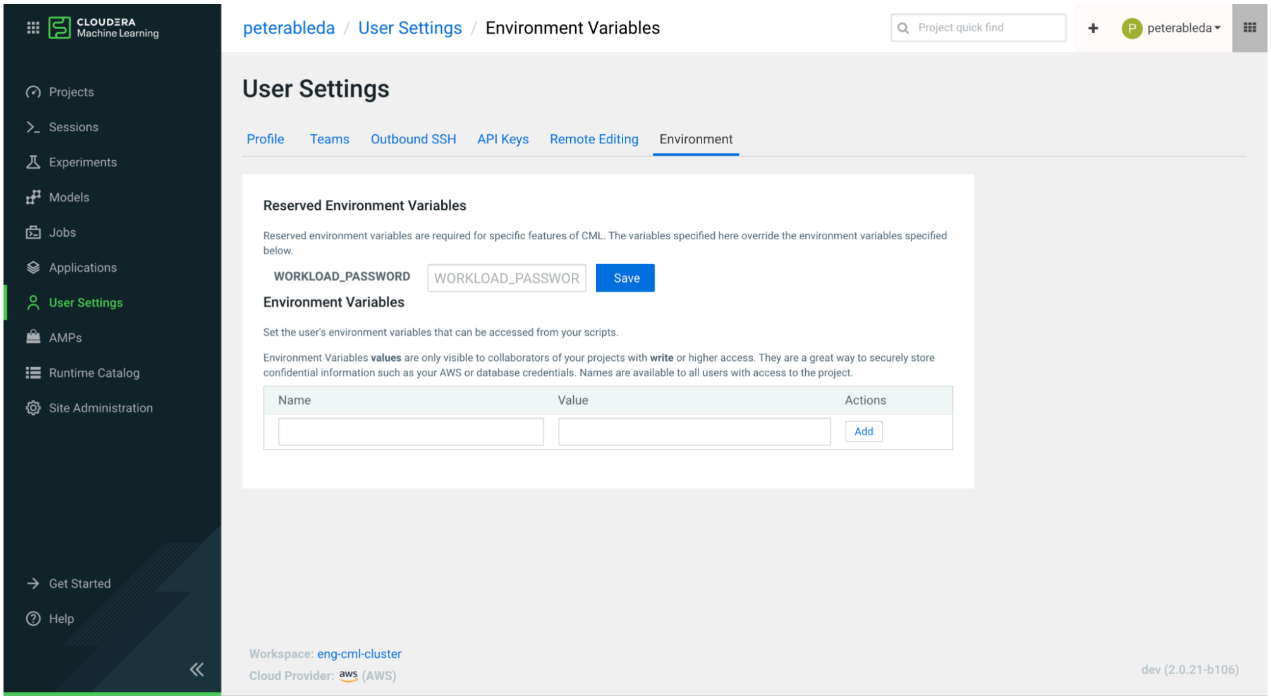
Ensure your environment is set up as follows:

- The project must be configured to use ML Runtimes.
- To use a Spark data connection, make sure you have the correct permissions set in the s3 bucket.
- Configure the WORKLOAD_PASSWORD environmental variable.

You need to set an environmental variable for the password (WORKLOAD_PASSWORD). This environmental variable is needed to access a Hive or Impala warehouse.

To set the password:

1. Go to User Settings Environment .
2. Under the Reserved Environment Variables set your WORKLOAD_PASSWORD.
3. Click Save.



Reserved Environment Variables

Reserved environment variables are required for specific features of CML. The variables specified here override the environment variables specified below.

WORKLOAD_PASSWORD: WORKLOAD_PASSWOR

Environment Variables

Set the user's environment variables that can be accessed from your scripts.

Environment Variables **values** are only visible to collaborators of your projects with **write** or higher access. They are a great way to securely store confidential information such as your AWS or database credentials. Names are available to all users with access to the project.

Name	Value	Actions
		Add

Workspace: eng-cml-cluster

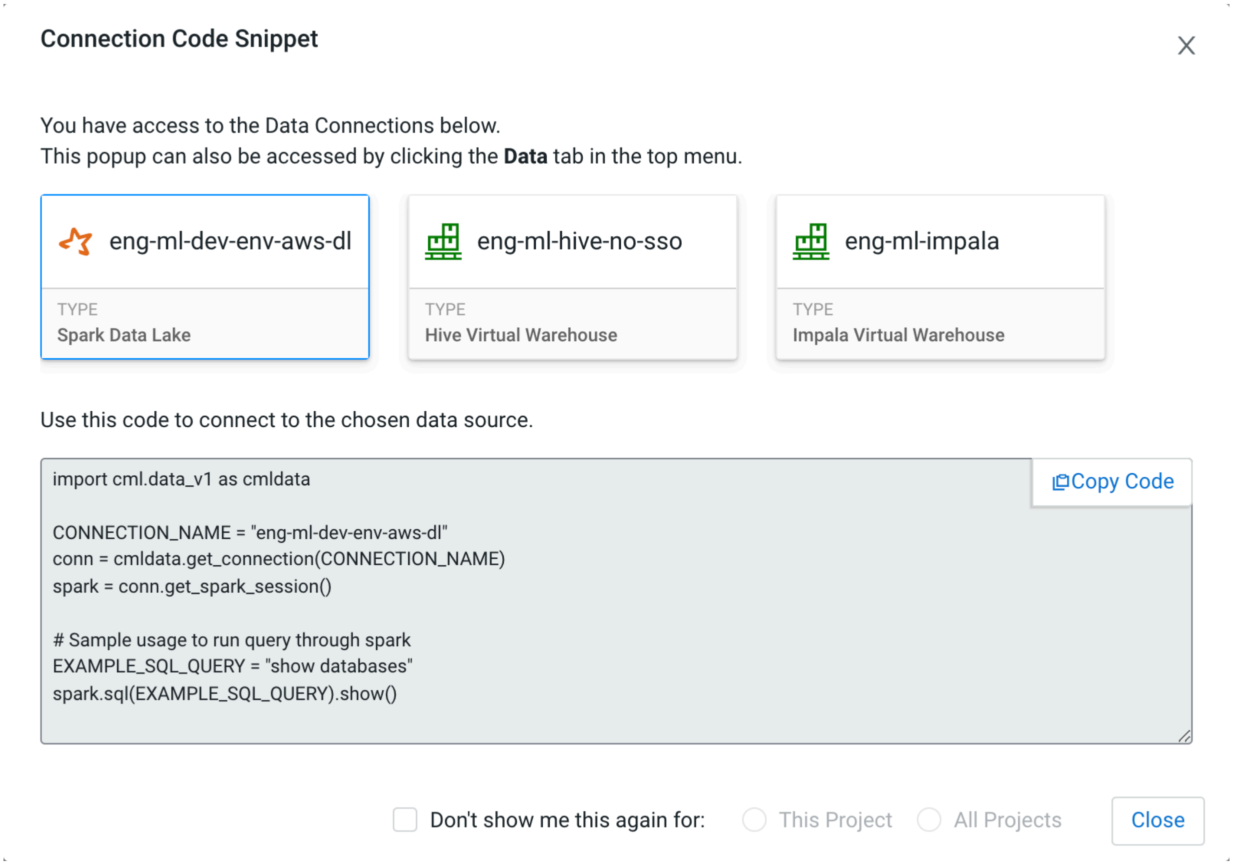
Cloud Provider: aws (AWS)

Using data connection snippets

As a data scientist, you can connect your project to data with a data connection snippet.

Procedure

1. Select New Project.
2. Enter the project name.
3. Start a new session. To use a Spark connection, make sure to select an ML Runtime engine.
4. In the Connection Code Snippet pane, select the data connection to use.



You have access to the Data Connections below.
This popup can also be accessed by clicking the **Data** tab in the top menu.

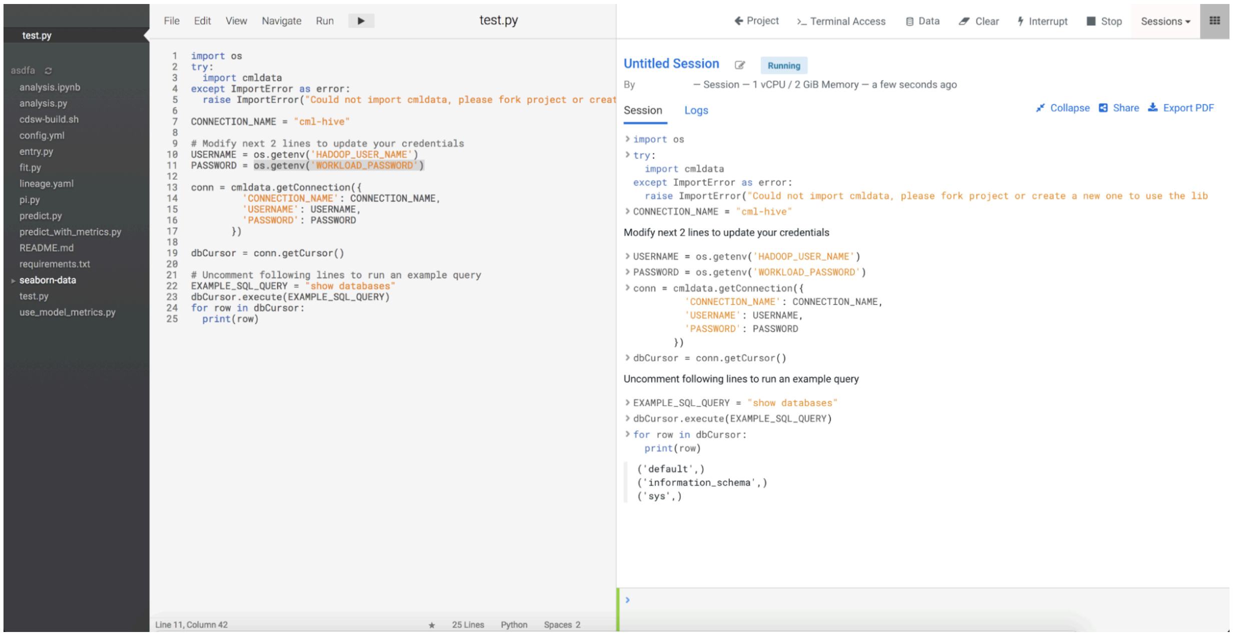
eng-ml-dev-env-aws-dl	eng-ml-hive-no-sso	eng-ml-impala
TYPE Spark Data Lake	TYPE Hive Virtual Warehouse	TYPE Impala Virtual Warehouse

Use this code to connect to the chosen data source.

```
import cml.data_v1 as cmldata

CONNECTION_NAME = "eng-ml-dev-env-aws-dl"
conn = cmldata.get_connection(CONNECTION_NAME)
spark = conn.get_spark_session()

# Sample usage to run query through spark
EXAMPLE_SQL_QUERY = "show databases"
spark.sql(EXAMPLE_SQL_QUERY).show()
```

□ Don't show me this again for: ○ This Project ○ All Projects

5. Select Copy Code.
6. Paste the snippet into your code.



```
import os
try:
  import cmldata
except ImportError as error:
  raise ImportError("Could not import cmldata, please fork project or creat

CONNECTION_NAME = "cml-hive"

# 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
       })

dbCursor = conn.getCursor()

# Uncomment following lines to run an example query
EXAMPLE_SQL_QUERY = "show databases"
dbCursor.execute(EXAMPLE_SQL_QUERY)
for row in dbCursor:
  print(row)
```

Line 11, Column 42   25 Lines   Python   Spaces 2

By — Session — 1 vCPU / 2 GiB Memory — a few seconds ago

```
> import os
> try:
    import cmldata
  except ImportError as error:
    raise ImportError("Could not import cmldata, please fork project or create a new one to use the lib
> CONNECTION_NAME = "cml-hive"
```

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
         })
> dbCursor = conn.getCursor()
```

Uncomment following lines to run an example query

```
> EXAMPLE_SQL_QUERY = "show databases"
> dbCursor.execute(EXAMPLE_SQL_QUERY)
> for row in dbCursor:
    print(row)

  ('default',)
  ('information_schema',)
  ('sys',)
```

7. Edit the username and password, if necessary. (This is not necessary for the Spark Data Lake connection, or if the WORKLOAD_PASSWORD environmental variable is set up for the user.)
8. Enter the data connection name.
9. Uncomment and edit the SQL query.
10. Select Run.

Results

The results of the SQL query display in the command window.

What to do next

When you have finished exploring the available data sources, you can select Don't show me this again for this project on the Connection Code Snippet pane, and it will no longer display when you start a session 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 Admin 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 Workspace 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 Workspace 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: you need to click Sync with Workspace at the project level to get the updated connections.

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, they can use curl to submit api requests to list the data connections.

A collaborator with write access can:

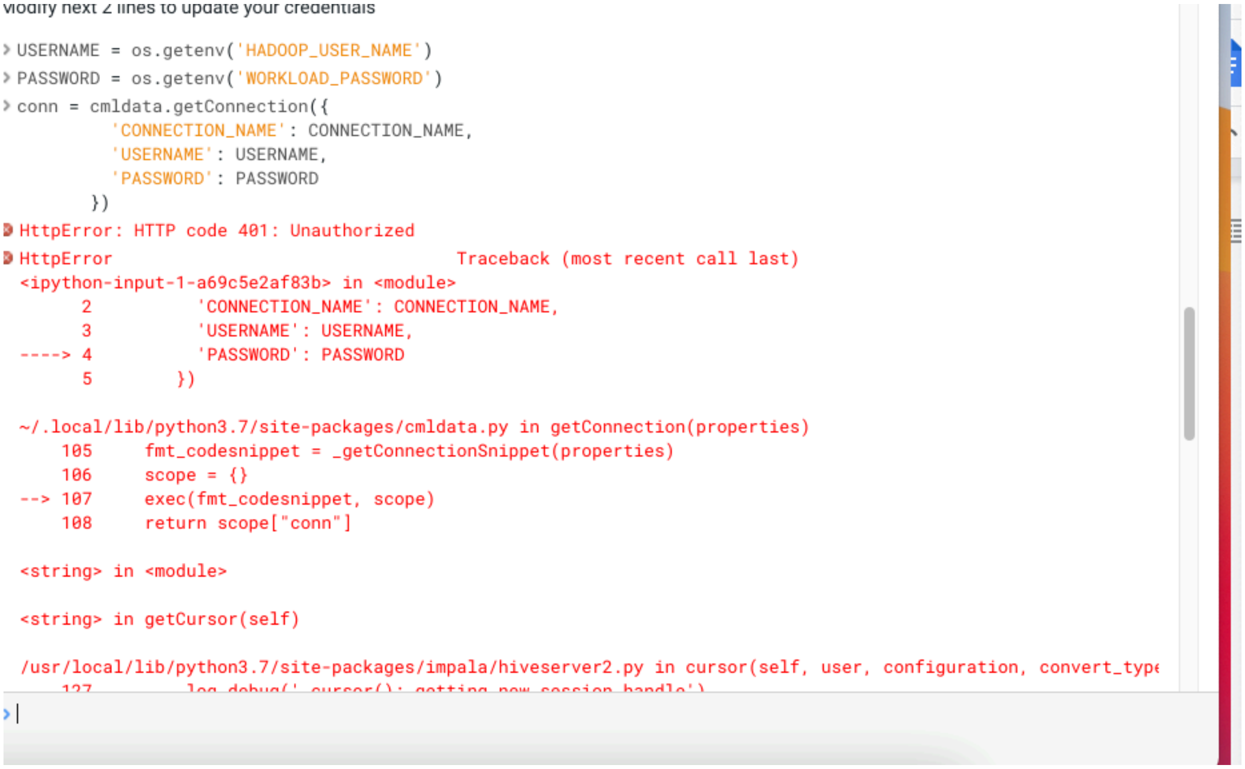
- Create data connections
- Edit data connections
- Delete data connections

- List data connections
- Get data connection details

Troubleshooting: 401 Unauthorized

Problem: Session returns a 401: Unauthorized error.

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.



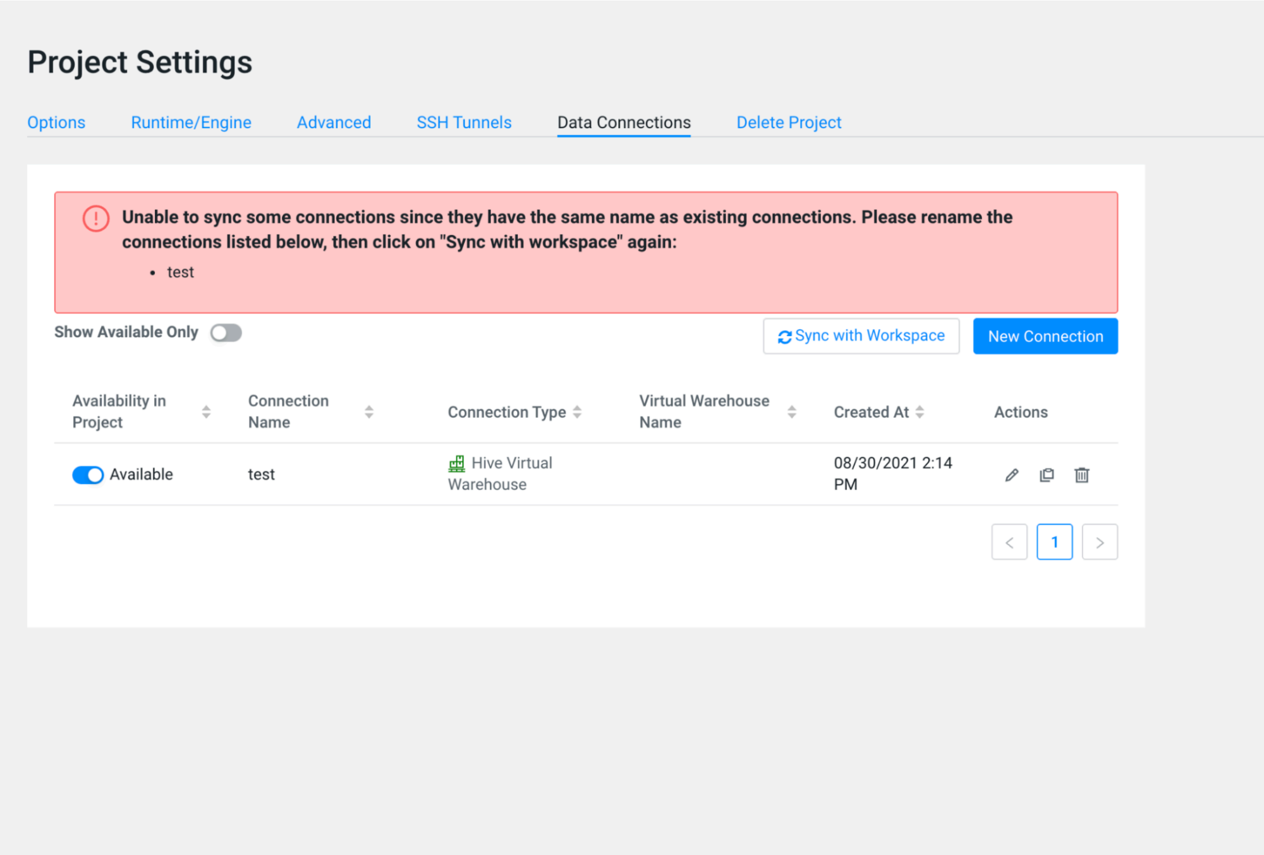
Modify next 2 lines to update your credentials

```
> 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: You need to set your workload password. Ensure you are using the correct credentials.

Troubleshooting: Existing connection name

Problem: When the user attempts to sync data connections, an error message displays, stating the crn or name is a duplicate.



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:

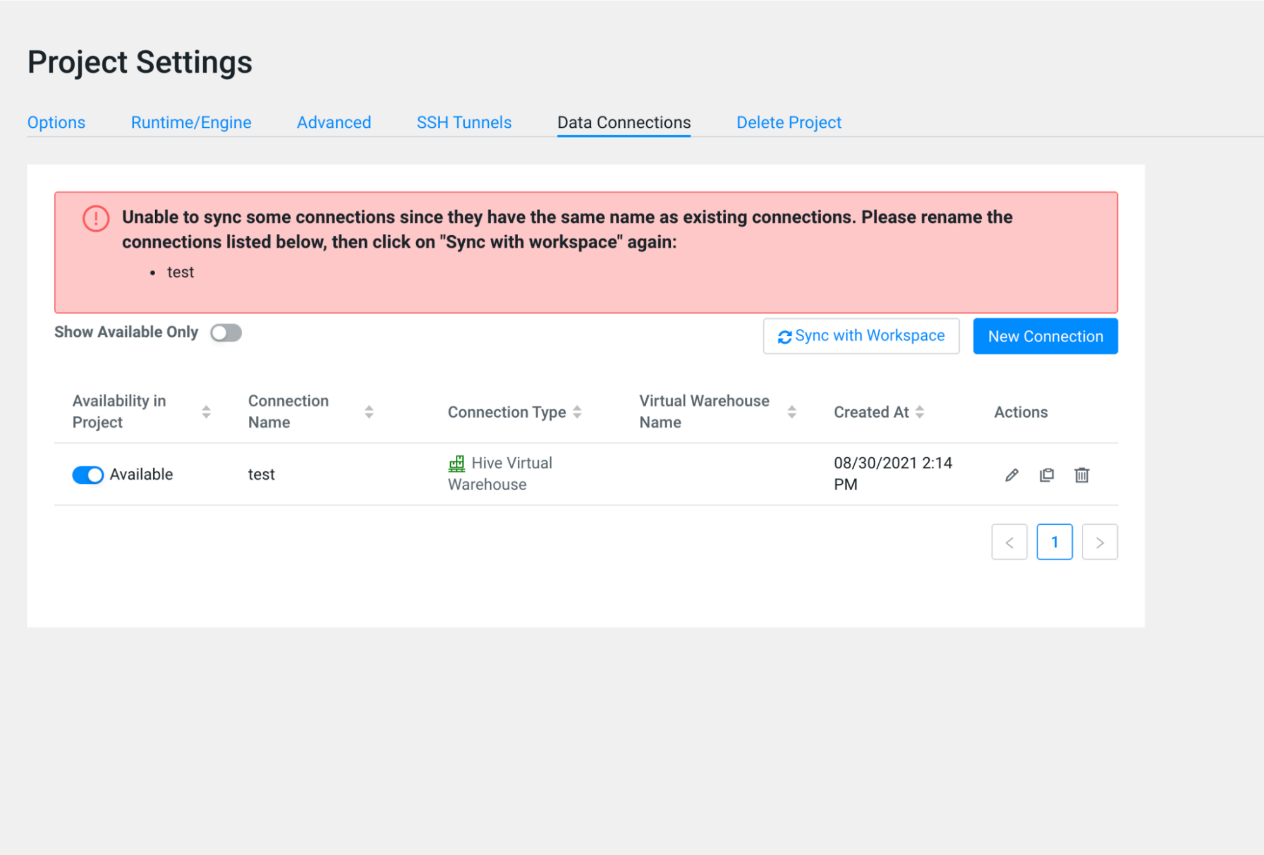
- test

Availability in Project	Connection Name	Connection Type	Virtual Warehouse Name	Created At	Actions
Available	test	Hive Virtual Warehouse		08/30/2021 2:14 PM	

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

Troubleshooting: Existing connection name

Problem: When the user attempts to sync data connections, an error message displays, stating the crn or name is a duplicate.



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

Availability in Project	Connection Name	Connection Type	Virtual Warehouse Name	Created At	Actions
Available	test	Hive Virtual Warehouse		08/30/2021 2:14 PM	

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