

Exploratory Data Science

Date published: 2022-06-10

Date modified: 2022-06-15

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

Exploratory Data Science.....	4
Data Discovery and Exploration Steps.....	6
Meeting the prerequisites.....	6
About sample data.....	7
Starting Exploratory Data Science and Visualization.....	8
Creating a Flight Details dashboard.....	9
Troubleshooting.....	11
Troubleshooting: 401 Unauthorized.....	11
Troubleshooting: 401 Unauthorized when accessing Hive.....	11
Troubleshooting: Existing connection name.....	12
Troubleshooting: Empty data page.....	12
Troubleshooting: Some connections not shown.....	12

Exploratory Data Science

The Exploratory Data Science pattern shows you how a data scientist or other data practitioners can kickstart a data science project. In this pattern, you discover the data available in your CDP environment, explore that data using SQL queries, and then set up some data visualizations to better understand the data — all of this from within the Cloudera Machine Learning (CML) interface.

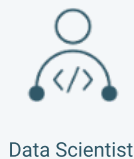
EXPLORATORY DATA SCIENCE

EXPLORE | QUERY | VISUALIZE



In the familiar CML interface, use data connections to access data sets stored in CDP Data Lakes. Explore the data with SQL queries, and then immediately create data visualizations.

Go from data to dashboard in 15 minutes



Data Scientist



Cloudera Machine Learning

Data
Connections



SQL



Visuals



We will show you how easy it is to find and explore data in CML.

Data Discovery and Exploration Steps

To follow this CDP pattern, you must have access to CDP Public Cloud and your IT team must have set up the Cloudera Data Warehouse and Cloudera Machine Learning services with other necessary prerequisites listed in the *Meeting the prerequisites* section. You will use sample data that is provided in the Virtual Warehouse.

Related Information

[Meeting the prerequisites](#)

Meeting the prerequisites

Before the Data Analysts can explore and query data, your central or departmental IT must have onboarded to CDP Public Cloud and must meet the requirements listed in this section.

To do in the CDP Management Console

1. Make sure that the Cloudera Data Lake is created and running. For more information, see [Creating an AWS environment with a medium duty data lake using the CLI](#).
2. Grant EnvironmentUser role to the Data Analyst user and synchronize the user to FreeIPA. For more information, see [Assigning account roles to users](#) and [Performing user sync](#).

To do in Cloudera Data Warehouse (CDW)

1. Make sure that the CDW environment is activated. For more information, see [Activating AWS environments](#).
2. Add an S3 bucket as an external bucket in the CDW environment with read-only access.



Important: You must add your own S3 bucket to the CDW environment.

- a. Go to CDW service Environments and click its edit icon.
 - b. On the Environment Details page, type the name of the AWS bucket you want to configure access to in the Add External S3 Bucket text box and select read-only access mode.
 - c. Click Add Bucket and then click Apply to update the CDW environment.
3. Create a non-default Database Catalog with the Load Demo Data option enabled.
 - a. Go to CDW service Database Catalogs and click Add New.
 - b. Specify a name, select an environment, select SDX from the Datalake drop-down list and enable the Load Demo Data option, and click CREATE.
 4. Create an Impala-based Virtual Warehouse with the Enable Data Visualization option and make sure it is in the running state.
 - a. Go to CDW service Virtual Warehouses and click Add New.
 - b. Select the type as IMPALA, select the Database Catalog associated with the Virtual Warehouse, select the size, click Enable Data Visualization, and then click CREATE.
 5. Enable the S3 File Browser for Hue.
 - a. Go to CDW service Virtual Warehouses, select your Virtual Warehouse and click edit.
 - b. On the **Virtual Warehouses** page, go to CONFIGURATIONS Hue and select hue-safety-valve from the drop-down list.
 - c. Add the following configuration for Hive or Impala Virtual Warehouse in the space provided and click APPLY:

```
[desktop]
# Remove the file browser from the blocked list of apps.
```

```
# Tweak the app_blacklist property to suit your app configuration.
app_blacklist=spark,zookeeper,hive,hbase,search,oozie,jobsub,pig,sqoop
,security

[aws]
[[aws_accounts]]
[[[default]]]
access_key_id=[ ***AWS-ACCESS-KEY*** ]
secret_access_key=[ ***SECRET-ACCESS-KEY*** ]
region=[ ***AWS-REGION*** ]
```

6. Enable a link to the Data VIZ application (Cloudera Data Visualization) from the Hue web interface or provide the Data VIZ application URL to the Data Analyst.
 - a. Go to CDW service Virtual Warehouses , select your Virtual Warehouse and click edit.
 - b. On the Virtual Warehouses details page, go to CONFIGURATIONS Hue and select hue-safety-valve from the drop-down list.
 - c. Add the following lines in the safety valve and click APPLY:

```
[desktop]
custom_dashboard_url=[ ***DATA-VIZ-URL*** ]
```

To do in Ranger

Grant the required DDL and DML Hadoop SQL policies to the Data Analyst user in Ranger.

1. Go to CDW service Database Catalogs , click the more option, and click Open Ranger.
2. On the **Ranger Service Manager** page, click Hadoop SQL.
3. Select the all - url policy.

The **Edit Policy** page is displayed.

4. Under the Add Conditions section, add the users under the Select User column and add permissions such as Create, Alter, Drop, Select, and so on from the Permissions column.
5. Scroll to the bottom of the page and click Save.

Other prerequisites

- Data in a file to be ingested is in a structured format, such as CSV, with headers.
- The data file is available to the Data Analyst on their computer or accessible on a shared drive with permissions already granted.
- Hue application URL is shared with the Data Analyst, if a link to the Data Visualization application is not enabled in Hue.

Related Information

[Data Discovery and Exploration Steps](#)

About sample data

For this pattern, we use a dataset about airlines and flights. This dataset is available in the Virtual Warehouse when you create a non-default Database Catalog with the Load Demo Data option enabled.

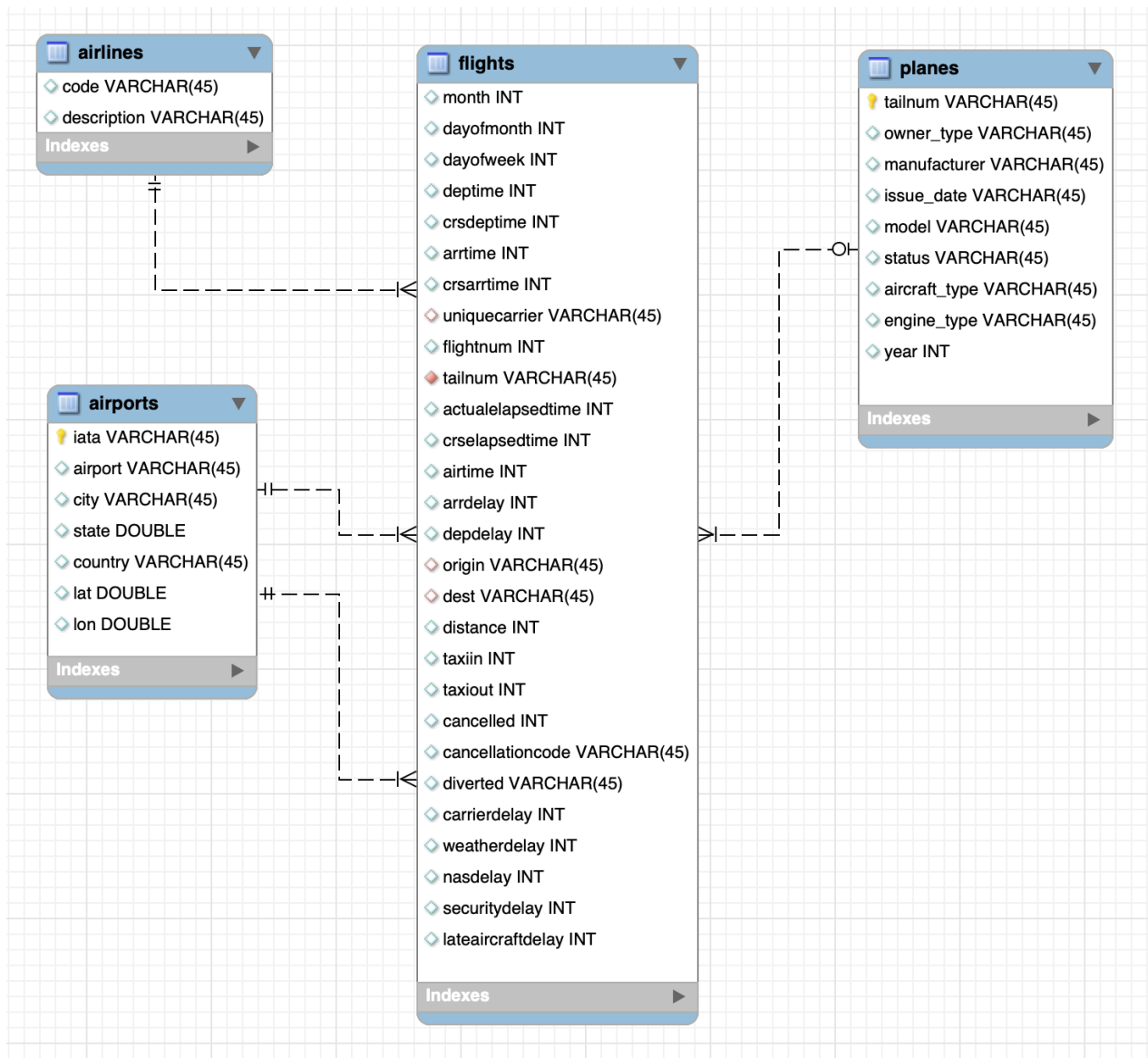
Airlines data (available in CDP)

The airline data consists of the following four tables:

- airlines
- airports
- flights

- planes

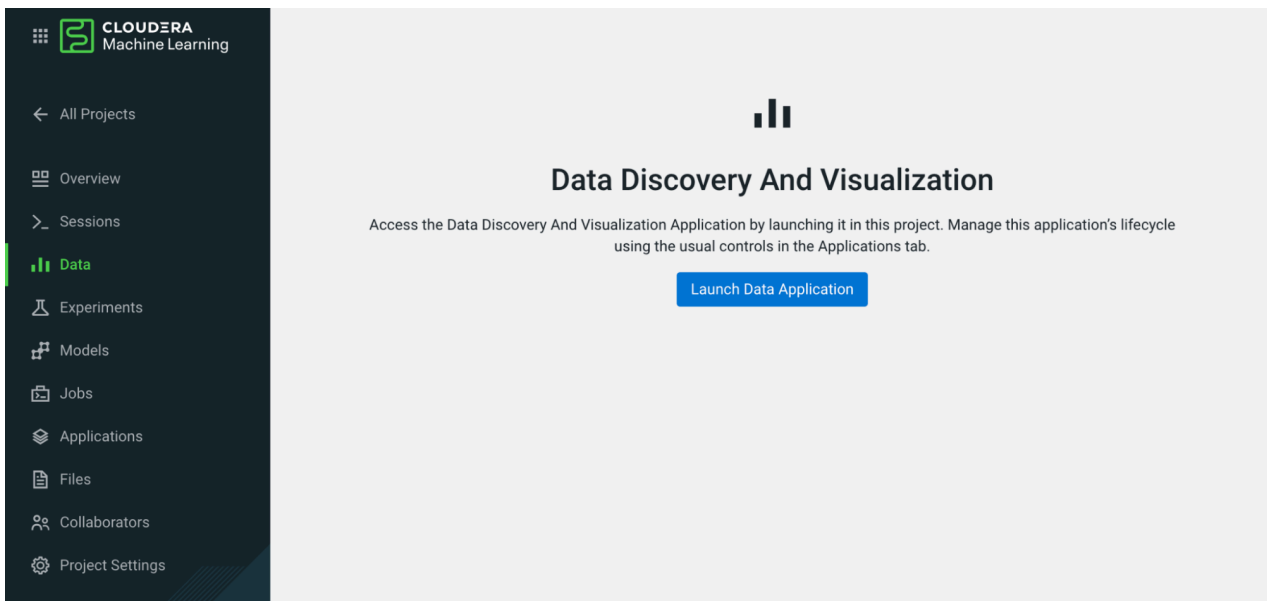
The airline data looks as follows:



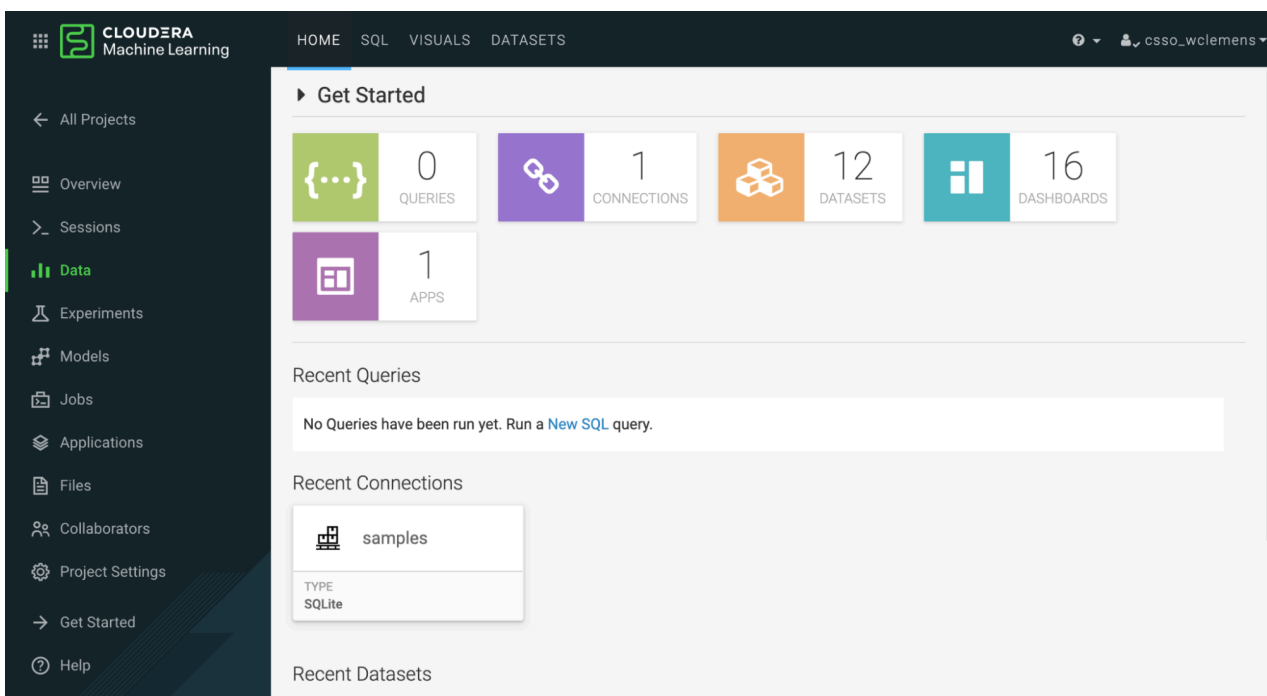
Starting Exploratory Data Science and Visualization

You can start the Exploratory Data Science and Visualization application within Cloudera Machine Learning (CML).

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 not, click Launch Data Application to start the application. It takes a few minutes to start the first time.



When the application starts, you can see four tabs at the top of the UI. From here, you can follow the steps in Get Started, or try some of the following tasks.



Creating a Flight Details dashboard

In this exercise, you will create a dashboard to show information about airlines and flight cancellations. You can create this dashboard using the Data Discovery and Visualization application integrated into Cloudera Machine Learning.

Procedure

1. In Data, select the SQL tab, then select the airline data connection.
2. In Database, select `airline_ontime_parquet`.

3. In Enter SQL below, enter this SQL statement:

```
select * from airline_ontime_parquet.flights
```

Enter the table name by clicking flights in the Tables list. The database name is automatically prepended in the SQL statement.

4. Click Run.

The SQL statement executes, and the results appear in the Results tab.

5. Click New Dashboard.

The Visuals tab opens.

6. In the Visuals menu, select the Table graphic type.

7. In the Dimensions menu, select uniquecarrier and drag it to the Dimensions shelf.

8. In the Measures menu, select Record Count and drag it to the Measures shelf.

9. In the Dimensions menu, select cancelled and drag it to the Measures shelf.

It appears as sum(cancelled).

10. On this sum(cancelled) item, in the Field Properties, click Alias, and rename the item to Cancel Count.

11. Drag the cancelled item again from the Dimensions menu to the Measures shelf.

12. For the second item, click the right arrow next to the label, and select Enter/Edit Expression.

The expression displays in an edit pane.

13. Change the expression to the following:

```
sum([cancelled])/count(1)*100 as "Cancel Percent"
```

14. Click Save.


Note that the name of the item is changed to "Cancel Percent".

15. Click on Record Count, then Alias, and change the name to Flight Count.


16. Click on uniquecarrier, then Alias, and change it to Carrier.

17. Click Refresh Visual.

The column names in the visual area are updated.

18. Click  and select Clone.

The table is copied to a new area in the visual designer.

19. In the new table, click Configure .

20. In the Visuals menu, select Bars to create a bar chart. The table transforms into a bar chart.

21. Move items around on the shelves so that they are arranged as shown:

- X Axis: Carrier
- Y Axis: Cancel Percent
- Tooltips: Cancel Count, Flight Count

22. Click Refresh Visual.

When you move the pointer onto a bar, tooltips showing data for that bar appear.

23. Click Cancel Percent, and in Field Properties, click Order and Top K.

24. Select Descending.

25. Click Refresh Visual.

The bars are reordered from largest to smallest.

26. In the Action menu on the graphic pane, select Edit Dataset SQL. The SQL editor page opens.

27. Click on the Dataset name field, and rename it to Flight Details. Click Apply, then click Apply again in the modal window that appears.

Results

This completes the Flight Details dashboard.

Troubleshooting

You may encounter the following issues.

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 4 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: 401 Unauthorized when accessing Hive

Problem: Session returns a 401: Unauthorized error when you are accessing a Hive database.

Check that the Hive data warehouse is working.

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

1. In the control plane, go to Data Warehouses.
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

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

The screenshot shows the 'Project Settings' interface with the 'Data Connections' tab selected. A red error banner at the top states: '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:'. Below this, a list of connections is shown. The first connection is named 'test' and is of type 'Hive Virtual Warehouse'. The 'Availability in Project' toggle is set to 'Available'. The 'Created At' timestamp is '08/30/2021 2:14 PM'. At the bottom right of the table, there are pagination controls showing '< 1 >'. Buttons for 'Sync with Workspace' and 'New Connection' are visible at the top right of the table area.

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	

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: Empty data page

You launch the application and see the spinner, but then the empty data page returns.

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.

Troubleshooting: Some connections not shown

The data application doesn't show all data connections.

See the error message to fix the workload password. To set the workload password, follow the procedure in *Setting the workload password*. You will need to restart the data application after setting the workload password.

Alternatively, you may need to click Sync under the Data Connections to get all of the data connections available at the project level.