..

Connecting to Data

Date published: 2020-10-30
Date modified: 2022-09-21

https://docs.cloudera.com/

Legal Notice

Contents

Connecting to a data source in CDP Data Visualization

Data Visualization allows you to create connections to many types of external data sources.

When using Data Visualization in CDP Public Cloud with Cloudera Data Warehouse (CDW), the data connection is automatically set up, but you can connect to other data sources as well. In Cloudera Machine Learning (CML), you can connect to an Impala or a Hive data warehouse, or tie in data from predictive CML models.

The supported connection types are:

- Impala
- Hive
- Druid
- SQLite
- MySQL
- MariaDB
- PostgreSQL

Related Information

Connections

Datasets

Data modeling

Creating a CML data connection to an Impala data warehouse

Learn how to connect natively to data stored in Impala when using CDP Data Visualization in Cloudera Machine Learning (CML).

About this task

You must connect to your data prior to using the data modeling and visualization functions. The following steps show you how to create a new CML data connection to an Impala data warehouse.

Note:

Only users who have the Manage data connections privilege or administrators can create and manage connections in CDP Data Visualization.

You must be an administrator to be able to set privileges for a user. If you want to log in as an administrator, you can use the default admin account:

- username: vizapps_admin
- password: vizapps_admin

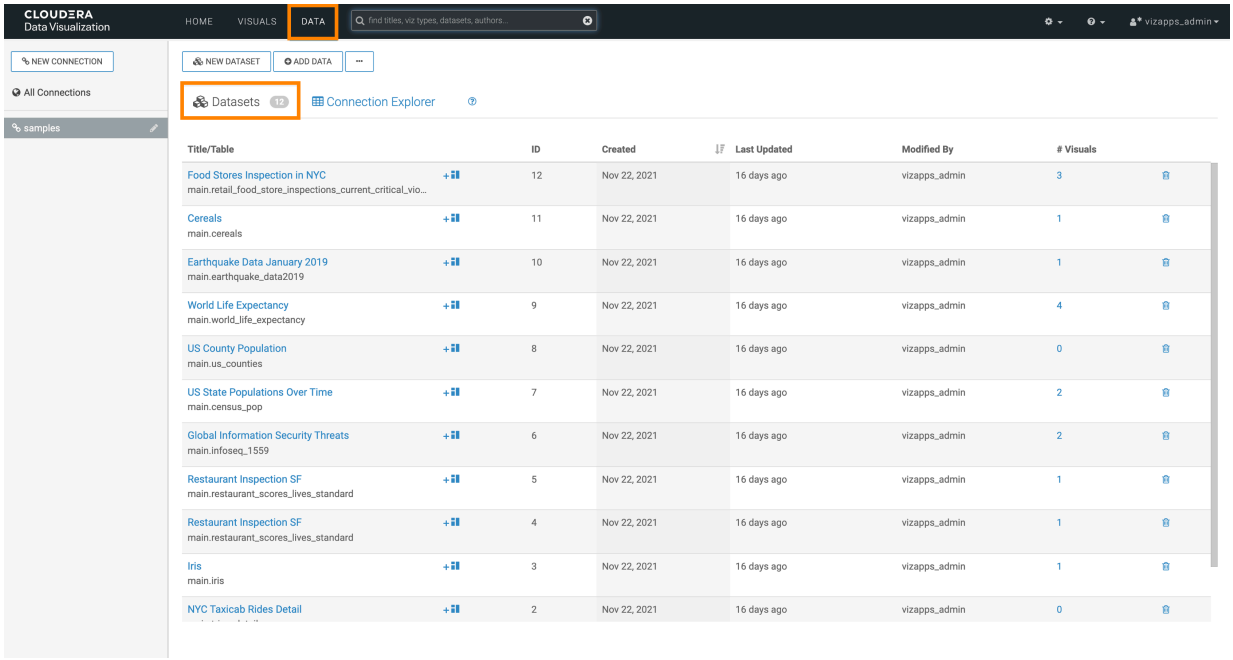
When you create a connection, you automatically have privileges to create and manage datasets on this connection, and to build dashboards and visuals in these datasets.

- For more information on the Manage data connections privilege, see *RBAC permissions*.
- For instructions on how to define privileges for a role, see *Setting role privileges*.
- For instructions on how to assign the administrator role to a user, see *Promoting a user to administrator*.

Procedure

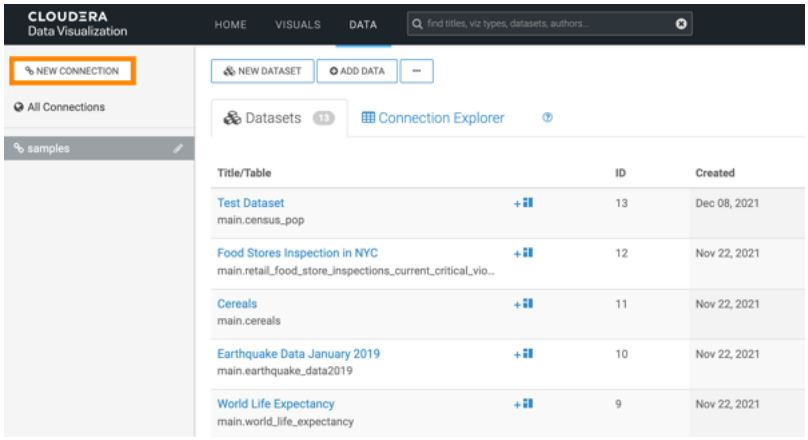
1. On the main navigation bar, click DATA.

   The DATA interface appears, open on the Datasets tab.

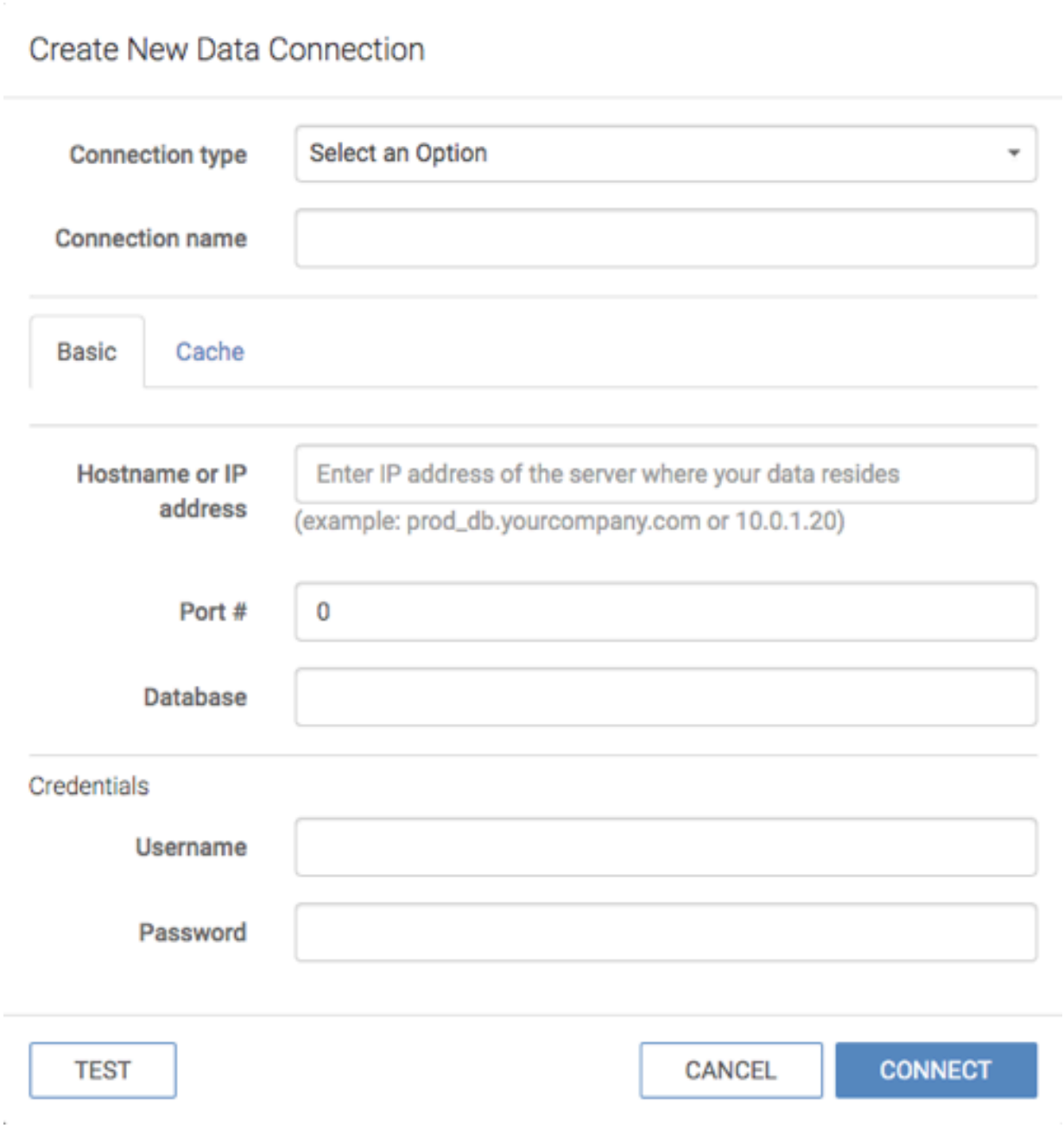


2. In the Data side menu bar, click NEW CONNECTION.

Note: The NEW CONNECTION button is only accessible to users assigned to roles with Manage data connections privilege and to administrators.



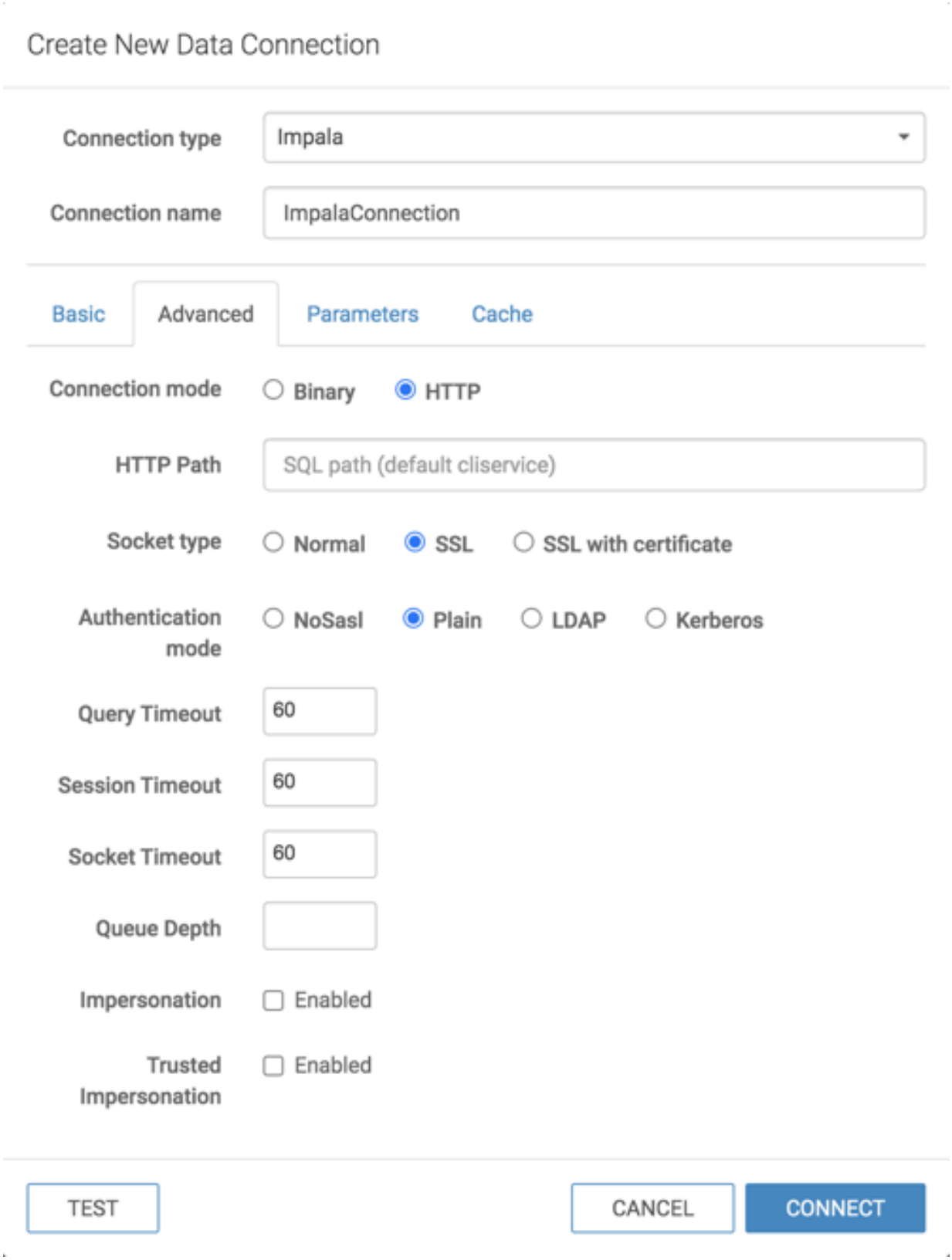
The Create New Data Connection modal window appears.



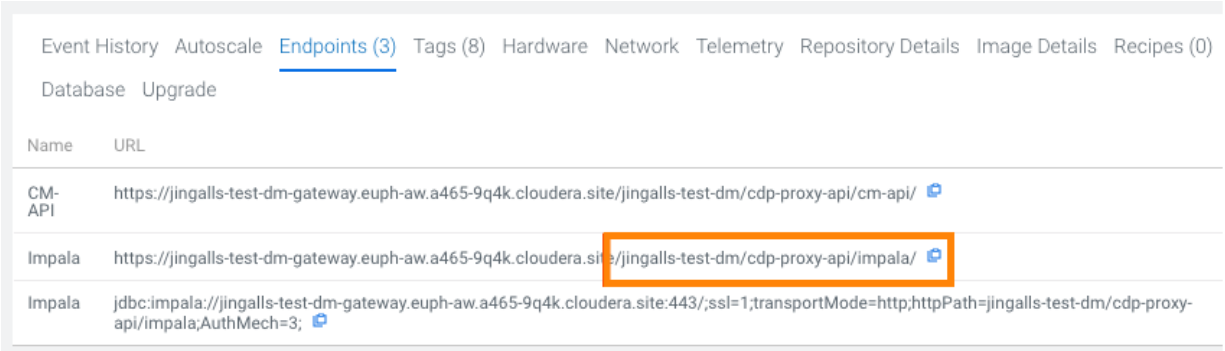
3. Select the Impala Connection type from the drop-down list and enter the hostname or IP address of the running coordinator.

   You can get the coordinator hostname from the JDBC URL of the Impala DW.
4. Enter 443 in the Port # field.
5. Enter your workload username and password as credentials.

6. Click the Advanced tab and make the selections below:



7. Locate the Impala Endpoint for the data hub.



8. Copy it and paste it into the HTTP Path field.
9. Click TEST to test the connection.
10. Click CONNECT to create the connection.

Results

You have set up a connection to a running Impala DW.

Related Information

RBAC permissions

Setting role privileges

Promoting a user to administrator

Creating a CML data connection to a Hive data warehouse

Learn how to connect natively to data stored in Hive when using CDP Data Visualization in Cloudera Machine Learning (CML).

About this task

You must connect to your data prior to using the data modeling and visualization functionalities. The following steps show you how to create a new CML data connection to a Hive data warehouse.

Note:

Only users who have the Manage data connections privilege or administrators can create and manage connections in CDP Data Visualization.

You must be an administrator to be able to set privileges for a user. If you want to log in as an administrator, you can use the default admin account:

- username: vizapps_admin
- password: vizapps_admin

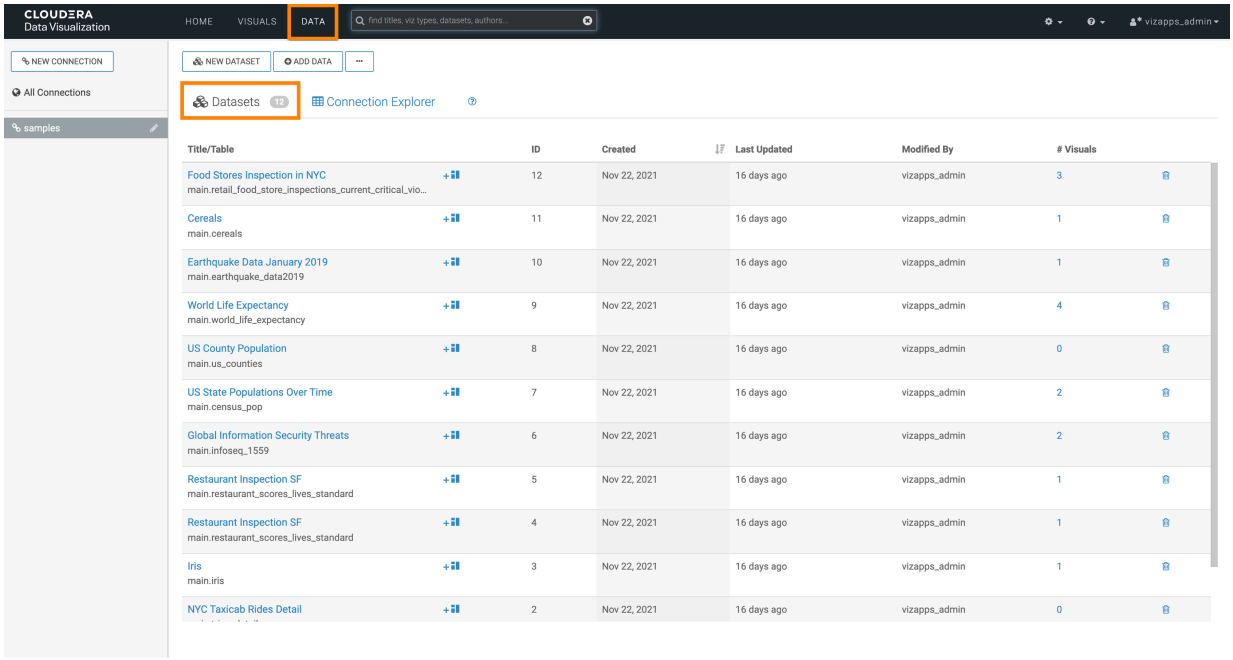
When you create a connection, you automatically have privileges to create and manage datasets on this connection, and to build dashboards and visuals in these datasets.

- For more information on the Manage data connections privilege, see *RBAC permissions*.
- For instructions on how to define privileges for a role, see *Setting role privileges*.
- For instructions on how to assign the administrator role to a user, see *Promoting a user to administrator*.

Procedure

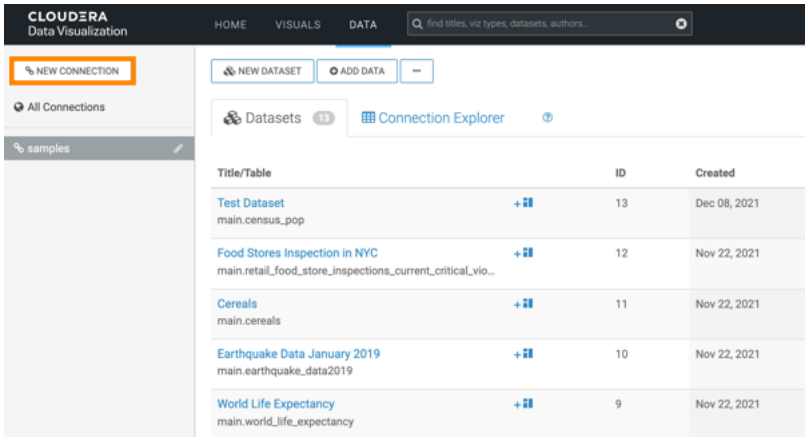
1. On the main navigation bar, click DATA.

   The DATA interface appears, open on the Datasets tab.

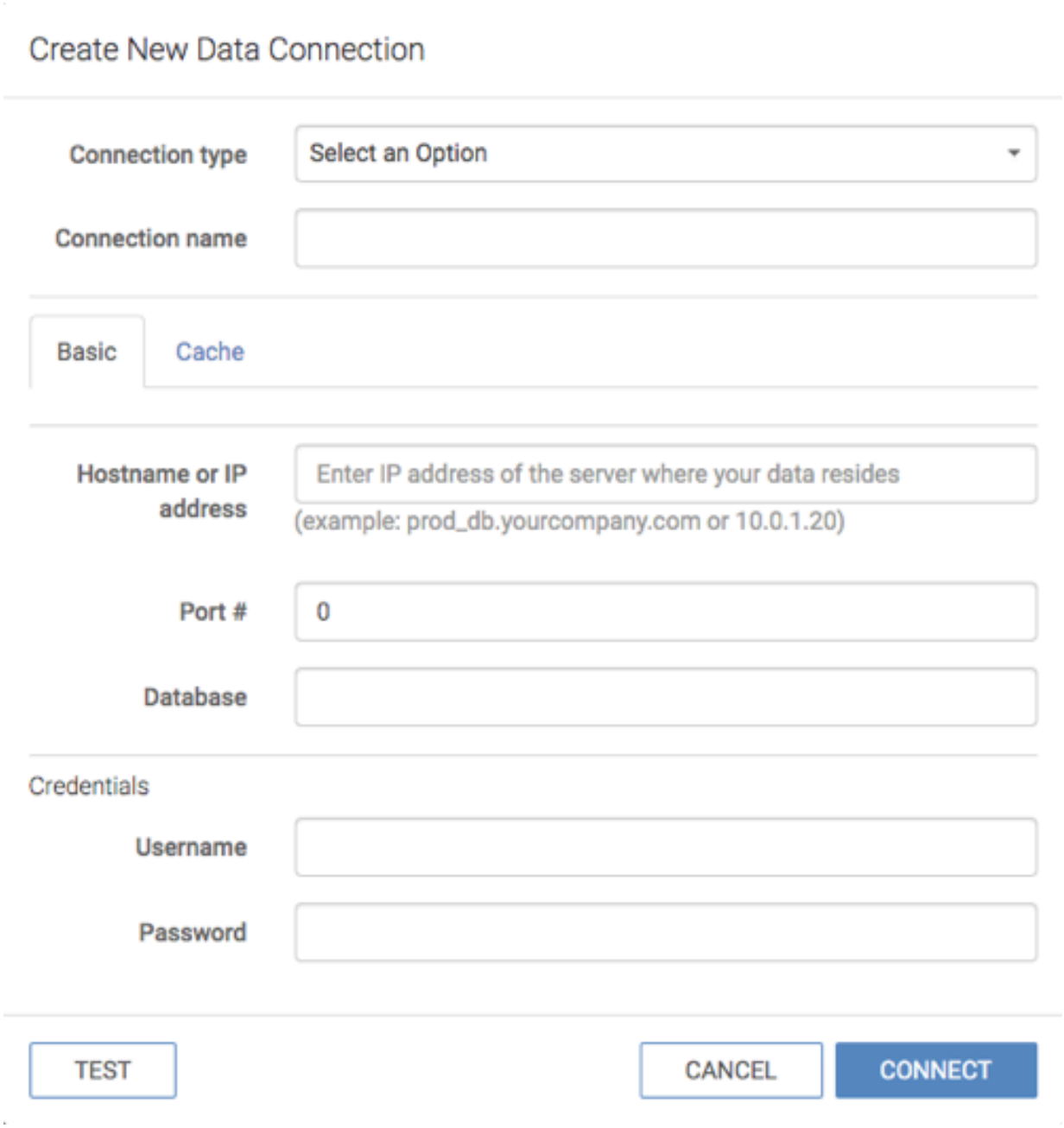


2. In the Data side menu bar, click NEW CONNECTION.

Note: The NEW CONNECTION button is only accessible to users assigned to roles with Manage data connections privilege and to administrators.



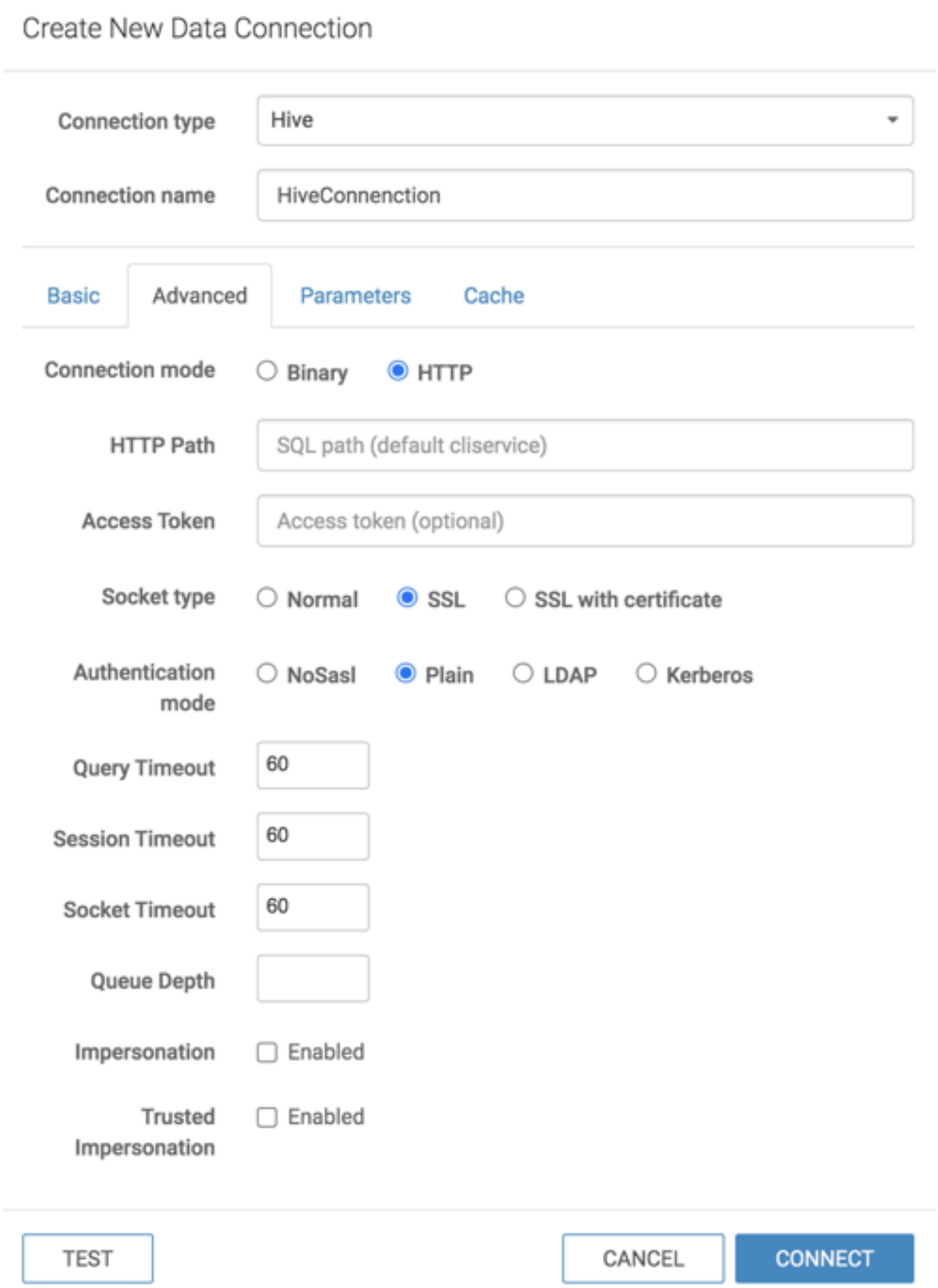
The Create New Data Connection modal window appears.



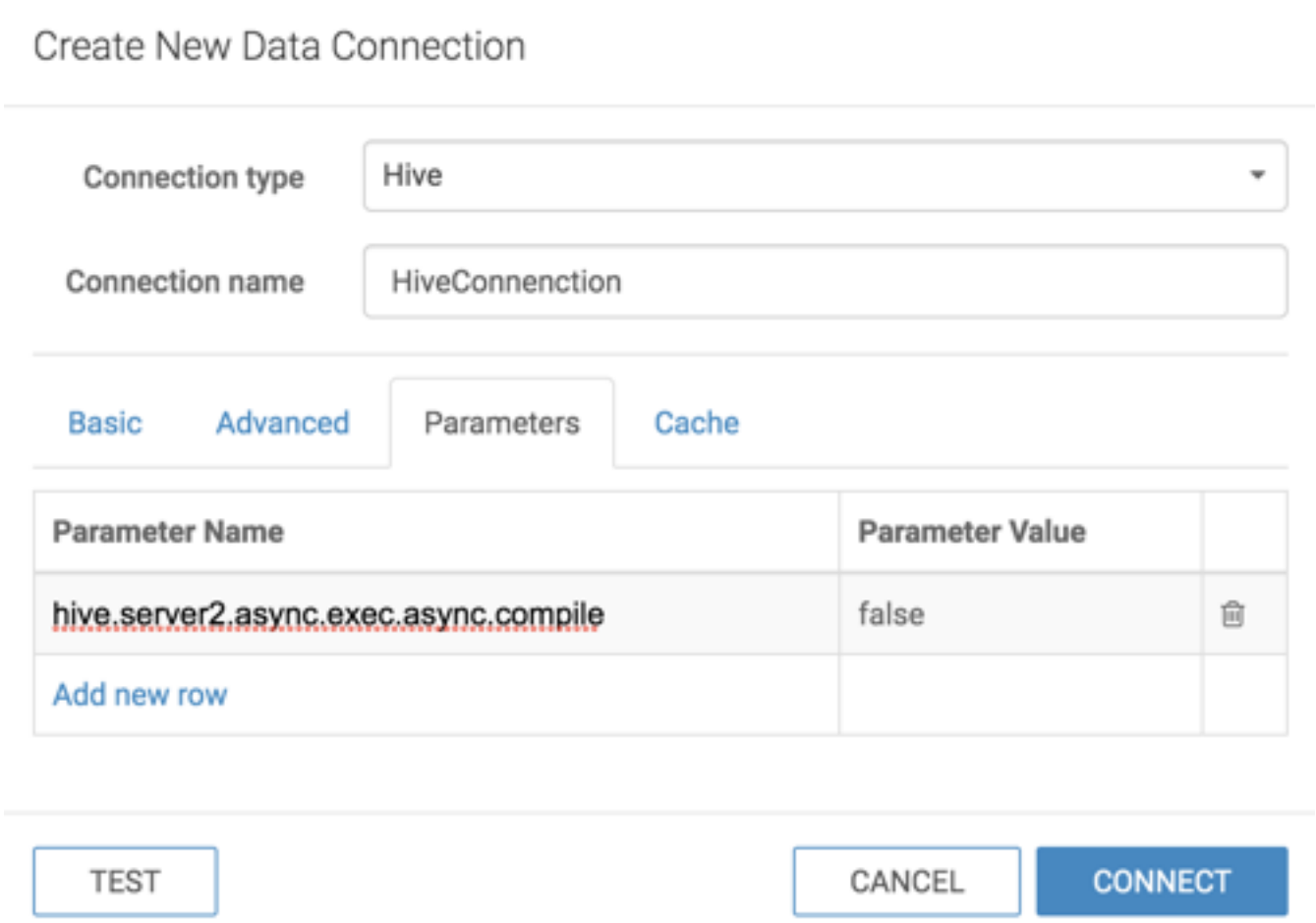
3. Select the Hive Connection type from the drop-down list and enter the hostname or IP address of the running coordinator.

   You can get the coordinator hostname from the JDBC URL of the Hive DW.
4. Use port 443.

5. Click the Advanced tab and make the selections below:



6. Click the Parameters tab and set the hive.server2.async.exec.async.compile parameter to false.



7. Use your workload username and password as credentials.
8. Click TEST and then CONNECT to create the connection.

Results

You have set up a connection to a running Hive DW.

Related Information

RBAC permissions

Setting role privileges

Promoting a user to administrator

Creating a CDW data connection in Data Visualization

Learn how to connect to data when using CDP Data Visualization in Cloudera Data Warehouse (CDW) data service. You can connect Data Visualization to a Virtual Warehouse to visualize your data. Similar to using a BI client, you can configure and connect to Virtual Warehouses from different clusters.

About this task

You must connect to your data prior to using the data modeling and visualization functions. You make the connection to the Virtual Warehouse when you select your warehouse in the steps below. The CDW Warehouse URL has the same compute instance ID as your Virtual Warehouse.

Note:

Only users who have the Manage data connections privilege or administrators can create and manage connections in CDP Data Visualization. In CDW, these are the members of the Admin Groups associated with the CDV instance.

When you create a connection, you automatically have privileges to create and manage datasets on this connection, and to build dashboards and visuals in these datasets.

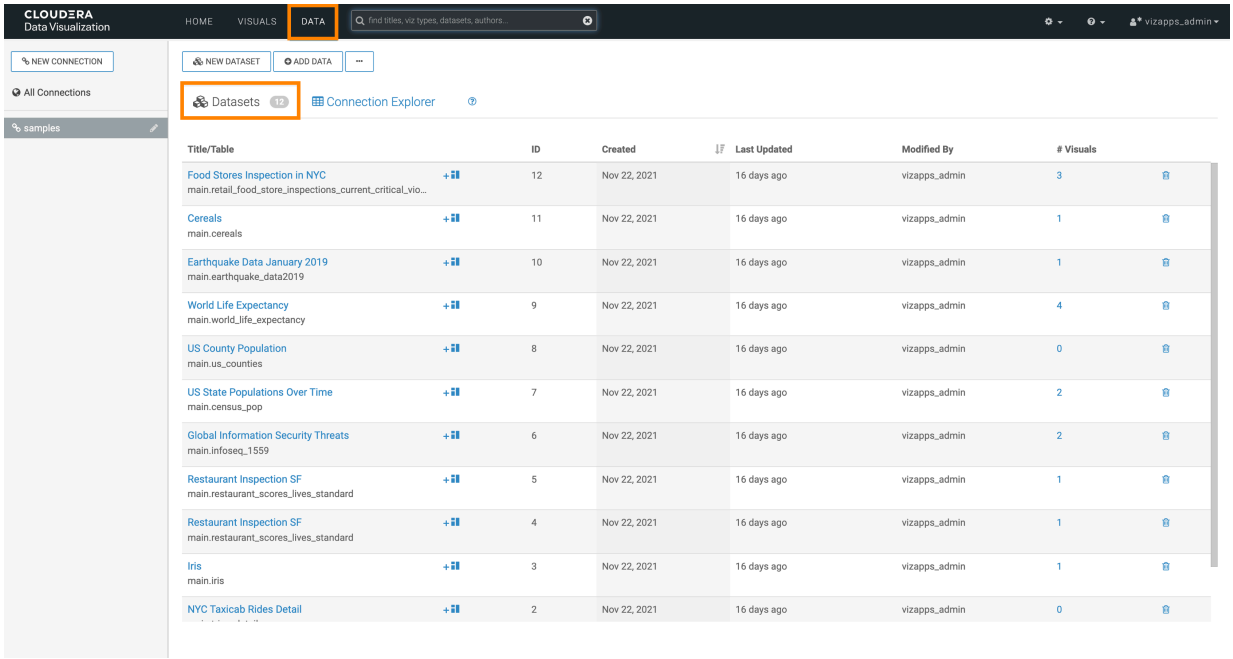
- For more information on the Manage data connections privilege, see *RBAC permissions*.
- For instructions on how to define privileges for a role, see *Setting role privileges*.
- For instructions on how to assign the administrator role to a user, see *Promoting a user to administrator*.

When you are creating a Hive or Impala data connection within the same cluster, the connection is considered secure and trusted, and the connection details can be auto populated with a default authentication user.

Procedure

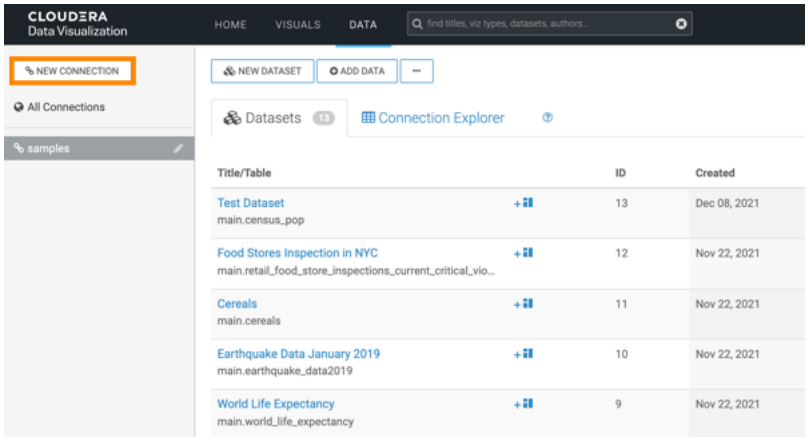
1. Start Data Visualization from the left navigation panel.
2. On the main navigation bar, click DATA.

   The DATA interface appears, open on the Datasets tab.

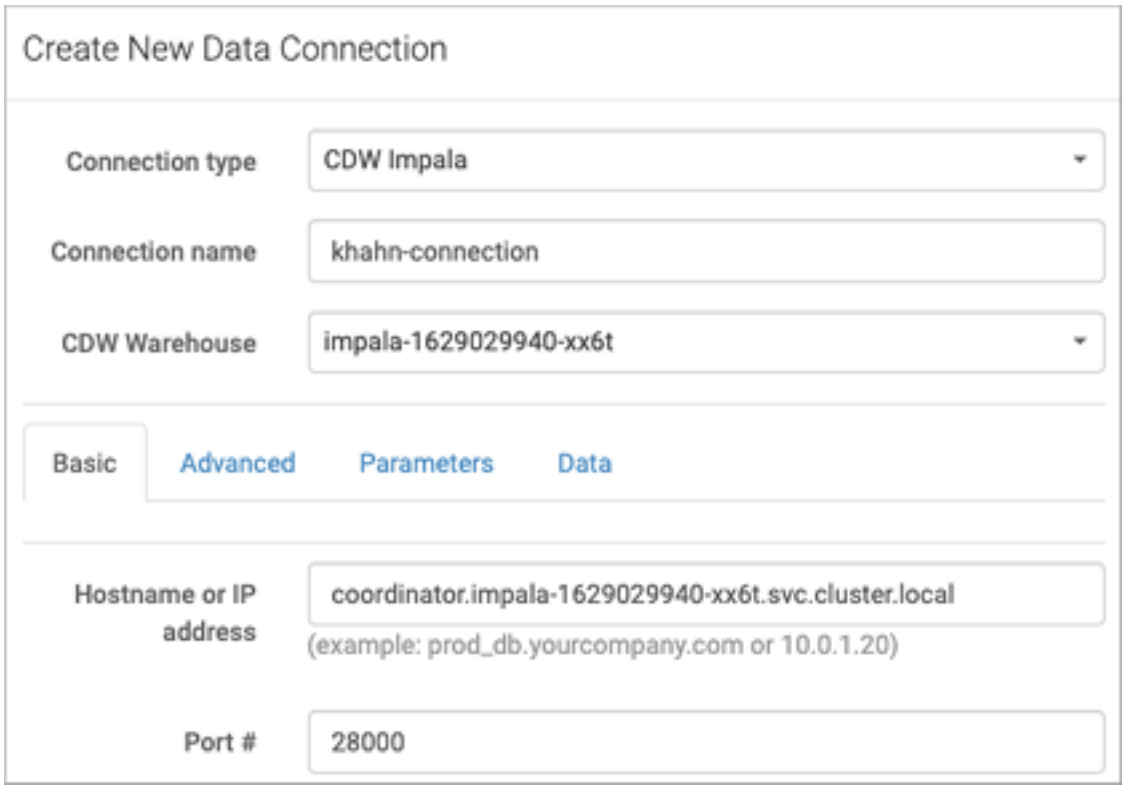


3. In the Data side menu bar, click NEW CONNECTION.

   Note: The NEW CONNECTION button is only accessible to users assigned to roles with Manage data connections privilege and to administrators.



   The **Create New Data Connection** modal window appears.



4. In Connection type, select CDW Hive or CDW Impala.
5. Provide a name for the connection.

6. Select a CDW Warehouse to connect to.

For Data connection within the same cluster

The following fields are auto populated:

- Hostname or IP address
- Port #
- Username

For Data connection outside the cluster

Enter the following information:

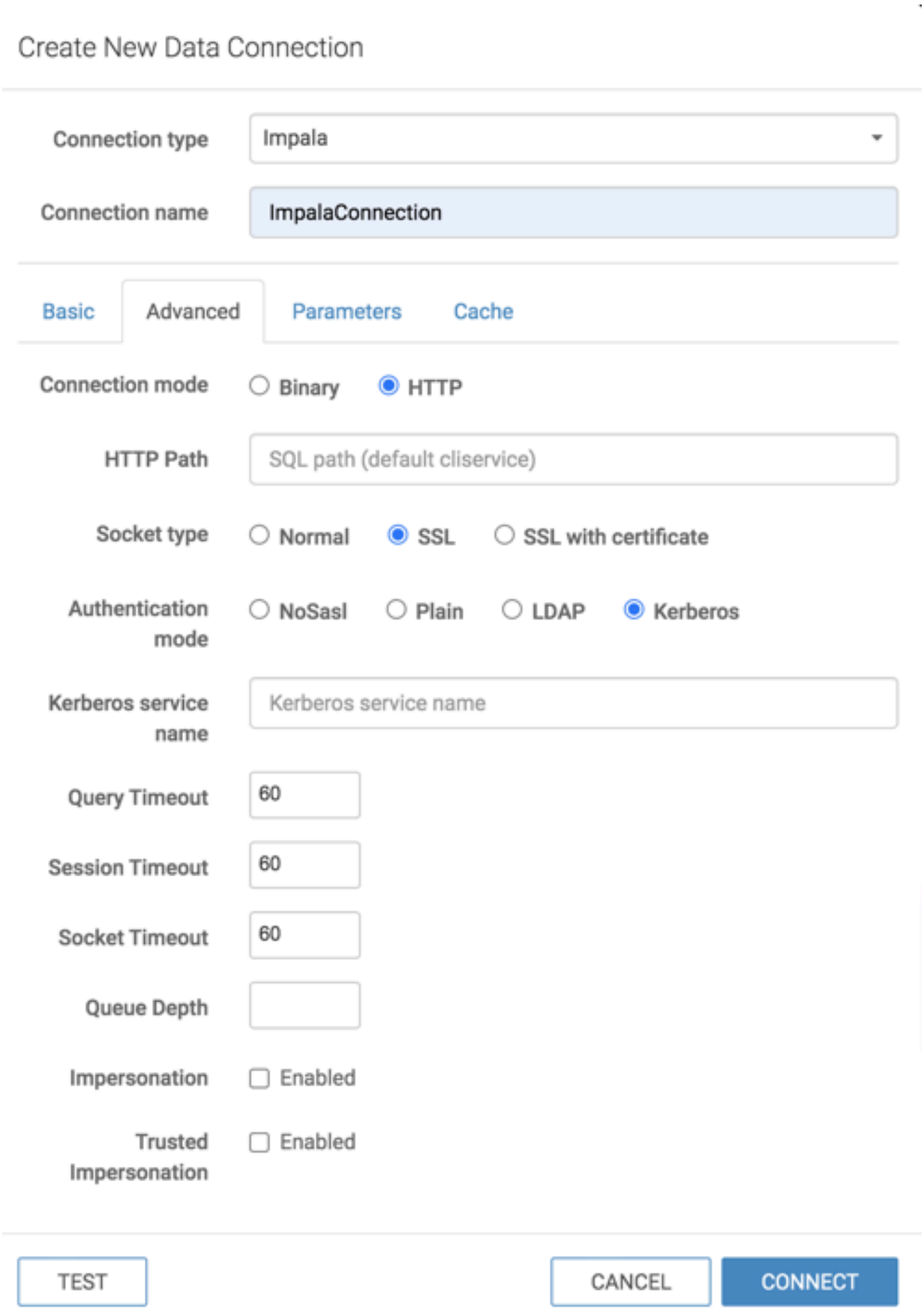
- Hostname or IP address
- Port #
- Username
- Password

7. Click the Advanced tab and make the appropriate selections.

   Important: Depending on the type of connection you are creating, there can be additional tabs in the Create New Data Connection modal window where you have to adjust further settings.

   The selections below are correct for the Cloudera iedh system.


Create New Data Connection
Connection type
Impala
Connection name
ImpalaConnection
Basic
Advanced
Parameters
Cache
Connection mode
Binary
HTTP
HTTP Path
SQL path (default cliservice)
Socket type
Normal
SSL
SSL with certificate
Authentication mode
NoSasl
Plain
LDAP
Kerberos
Kerberos service name
Kerberos service name
Query Timeout
60
Session Timeout
60
Socket Timeout
60
Queue Depth
Impersonation
Enabled
Trusted Impersonation
Enabled
TEST
CANCEL
CONNECT

8. Click TEST.

   If the connection is valid, the system returns a Connection Verified message.
9. Click CONNECT.

What to do next

You can create a data set, and then start creating visuals, dashboards, and applications. For more information, see *Creating datasets* and *Creating a visual*.

Related Information

RBAC permissions

Setting role privileges

Promoting a user to administrator

Creating datasets

Creating a visual

Creating a CDSW data connection to a data warehouse

Learn how to connect natively to data stored in a data warehouse when using CDP Data Visualization in Cloudera Data Science Workbench (CDSW).

About this task

You must connect to your data prior to using the data modeling and visualization functionalities. The following steps show you how to create a new CDSW data connection to a running Impala system.

Note:

Only users who have the Manage data connections privilege or administrators can create and manage connections in CDP Data Visualization. You must be an administrator to be able to set privileges for a user. If you want to log in as an administrator, you can use the default admin account:

- username: vizapps_admin
- password: vizapps_admin

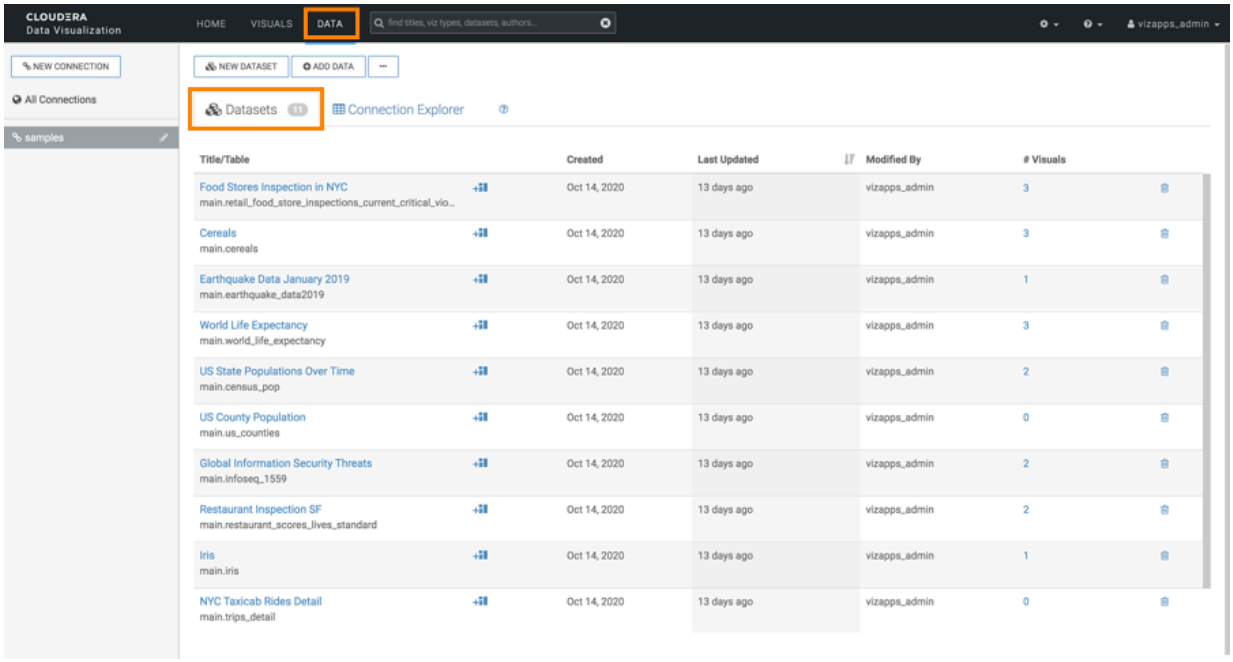
When you create a connection, you automatically have privileges to create and manage datasets on this connection, and to build dashboards and visuals in these datasets.

- For more information on the Manage data connections privilege, see *RBAC permissions*.
- For instructions on how to define privileges for a role, see *Setting role privileges*.
- For instructions on how to assign the administrator role to a user, see *Promoting a user to administrator*.

Procedure

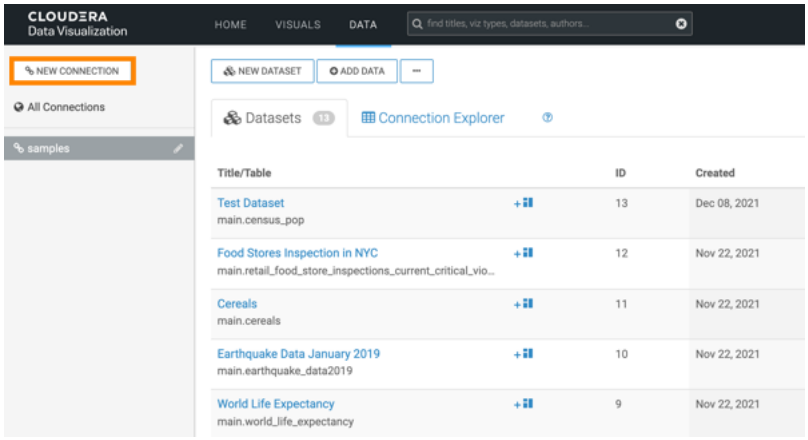
1. On the main navigation bar, click DATA.

   The DATA interface appears, open on the Datasets tab.

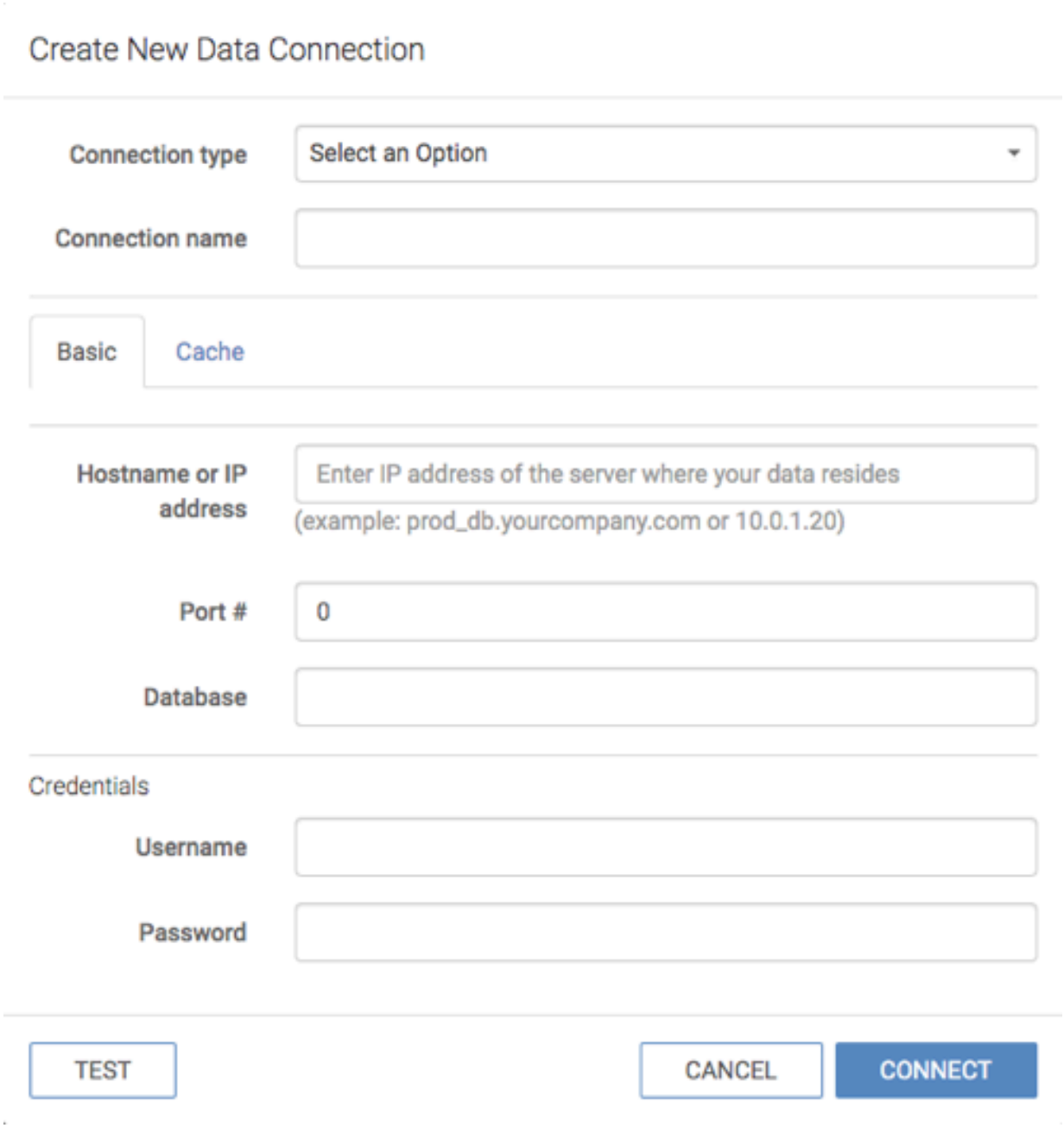


2. In the Data side menu bar, click NEW CONNECTION.

Note: The NEW CONNECTION button is only accessible to users assigned to roles with Manage data connections privilege and to administrators.



The Create New Data Connection modal window appears.

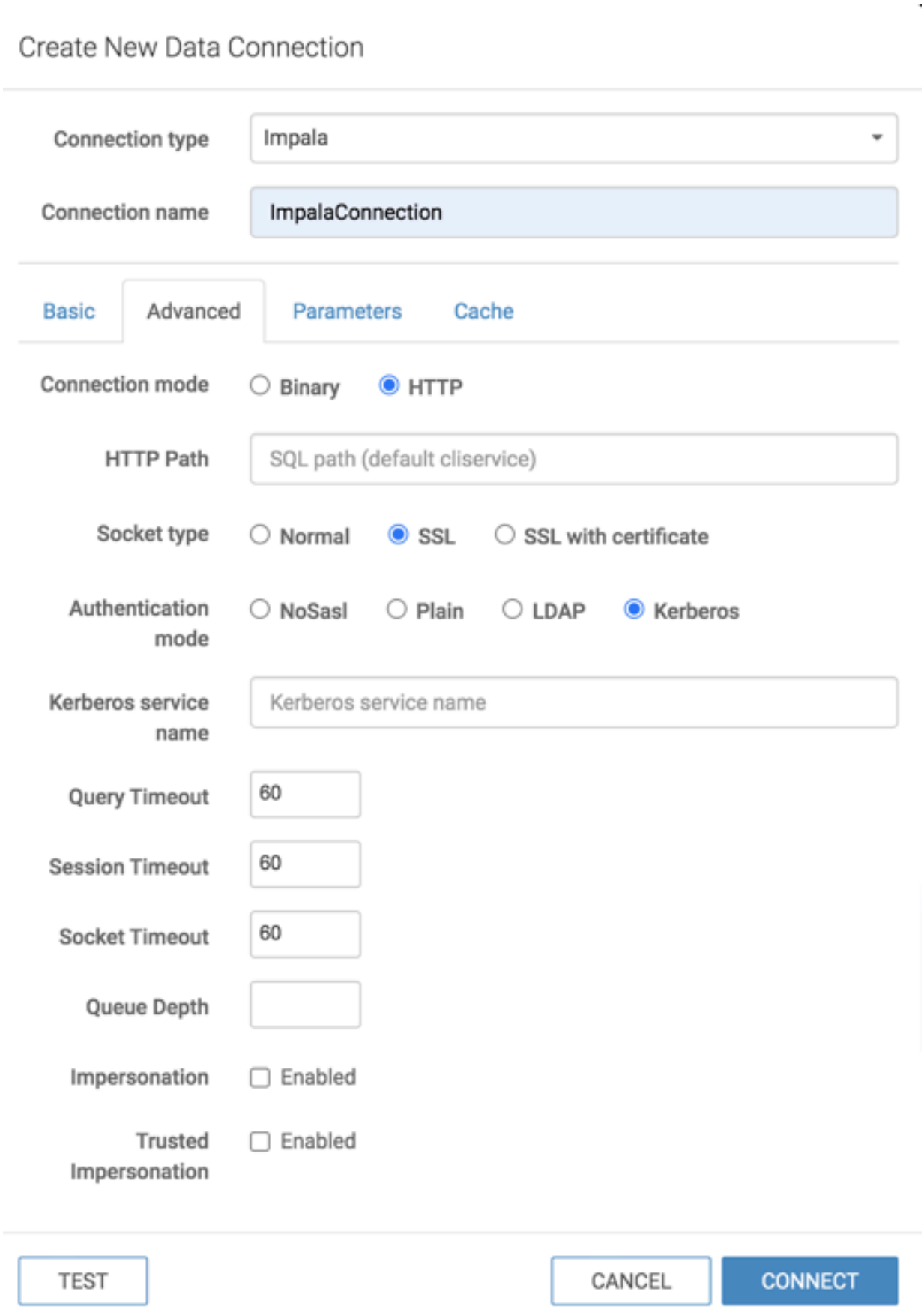


3. Select a Connection type from the drop-down list.
4. Provide a name for the connection.
5. Enter the hostname or IP address of the running coordinator.

   In this example, you can see to create an Impala connection. You can get the coordinator hostname from the JDBC URL of the Impala DW.
6. Under Port #, enter the port number.
7. Use your workload username and password as credentials.

8. Click the Advanced tab and make the appropriate selections.

Important: Depending on the type of connection you are creating, there can be additional tabs in the Create New Data Connection modal window where you have to adjust further settings.


Create New Data Connection
Connection type
Impala
Connection name
ImpalaConnection
Basic
Advanced
Parameters
Cache
Connection mode
Binary
HTTP
HTTP Path
SQL path (default cliservice)
Socket type
Normal
SSL
SSL with certificate
Authentication mode
NoSasl
Plain
LDAP
Kerberos
Kerberos service name
Kerberos service name
Query Timeout
60
Session Timeout
60
Socket Timeout
60
Queue Depth
Impersonation
Enabled
Trusted Impersonation
Enabled
TEST
CANCEL
CONNECT

9. Click TEST.

   If the connection is valid, the system returns a Connection Verified message.

10. Click CONNECT.

Results

You have set up a connection to a running data warehouse.

Related Information

RBAC permissions

Setting role privileges

Promoting a user to administrator

Editing a data connection

Learn how to edit a data connection in CDP Data Visualization.

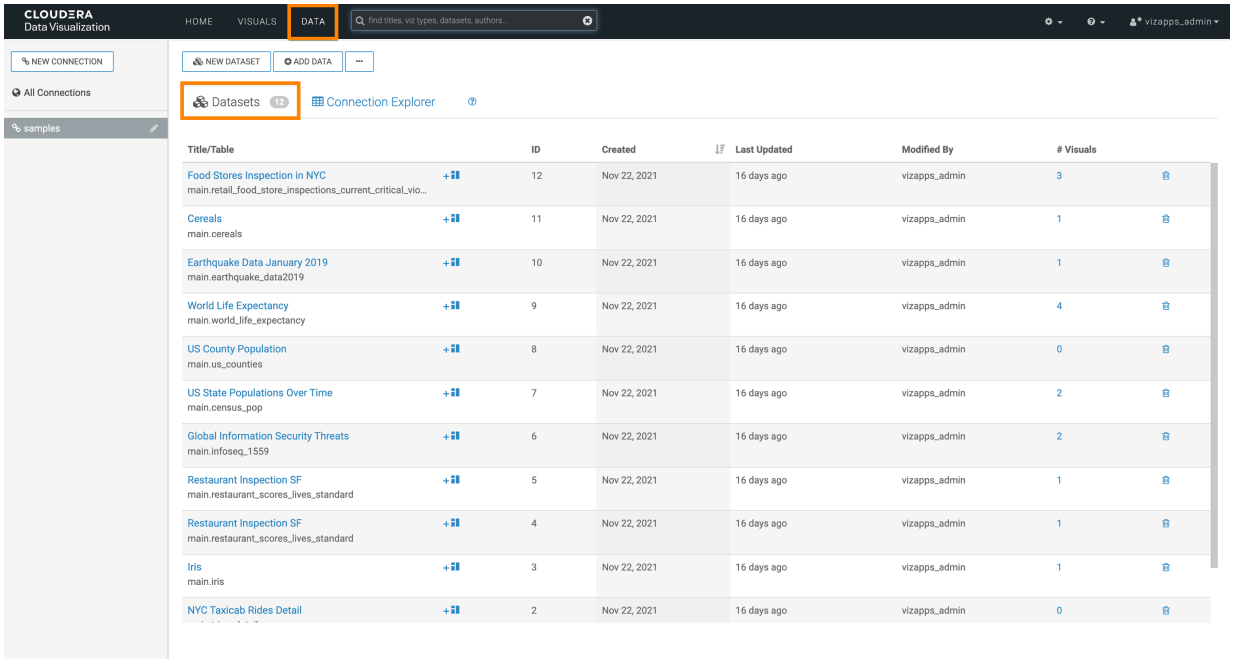
About this task

The following steps demonstrate how to edit existing data connections. The example shows changing an Impala connection to a Hive connection.

Procedure

1. On the main navigation bar, click DATA.

   The Data view appears, open on the Datasets tab.

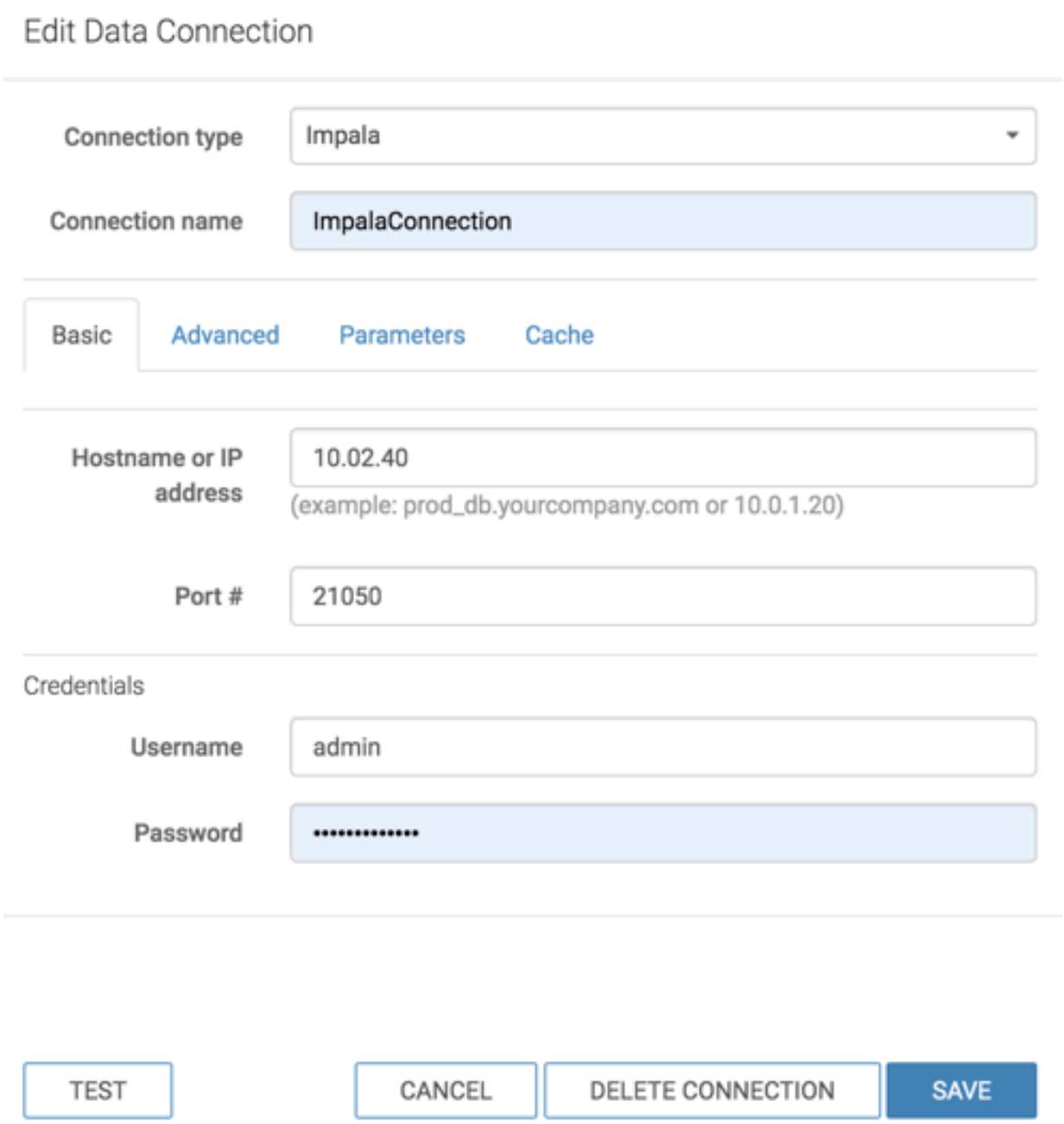


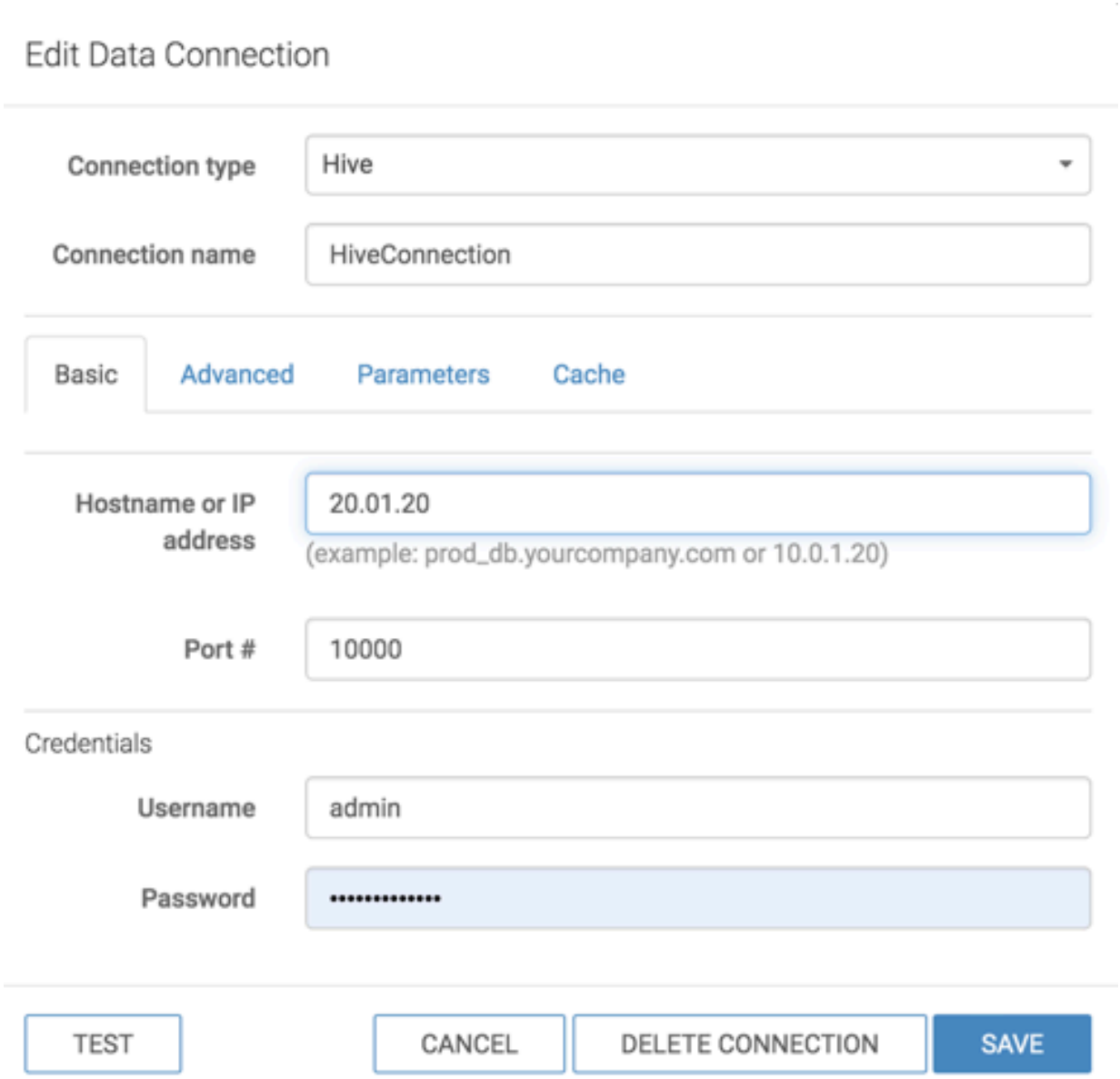
2. In the side bar, click the Edit button to the right of the connection you want to change.

   The Edit Data Connection modal window appears.

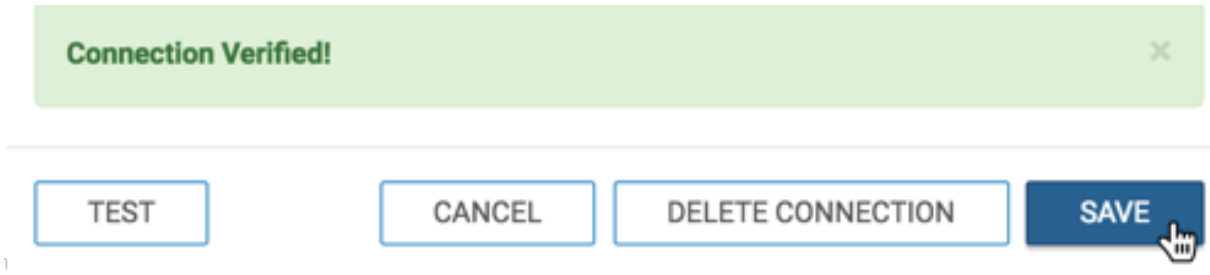
3. Edit the connnection details according to the connection type change you want to implement.

   In this example, an Impala connection is changed to a Hive connection.





4. At the bottom of the modal, click TEST.
5. If the connection is verified, click SAVE.



Results

After this operation succeeds, the name of the new type of connection appears on the side navigation bar.

Deleting a data connection

Learn how you can remove a data connection in CDP Data Visualization.

About this task

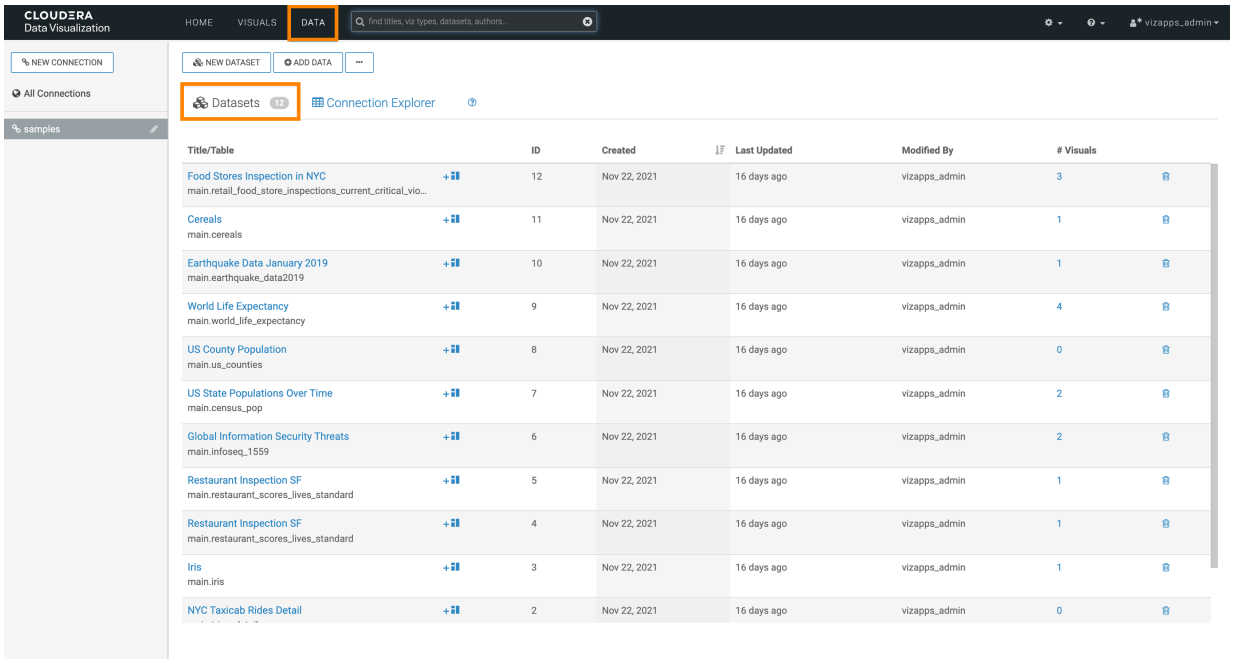
The following steps show you how to delete an existing data connection.

Tip: You can only delete connections that are not associated with any datasets. To learn how to delete datasets, see *Deleting datasets* .

Procedure

1. On the main navigation bar, click DATA.

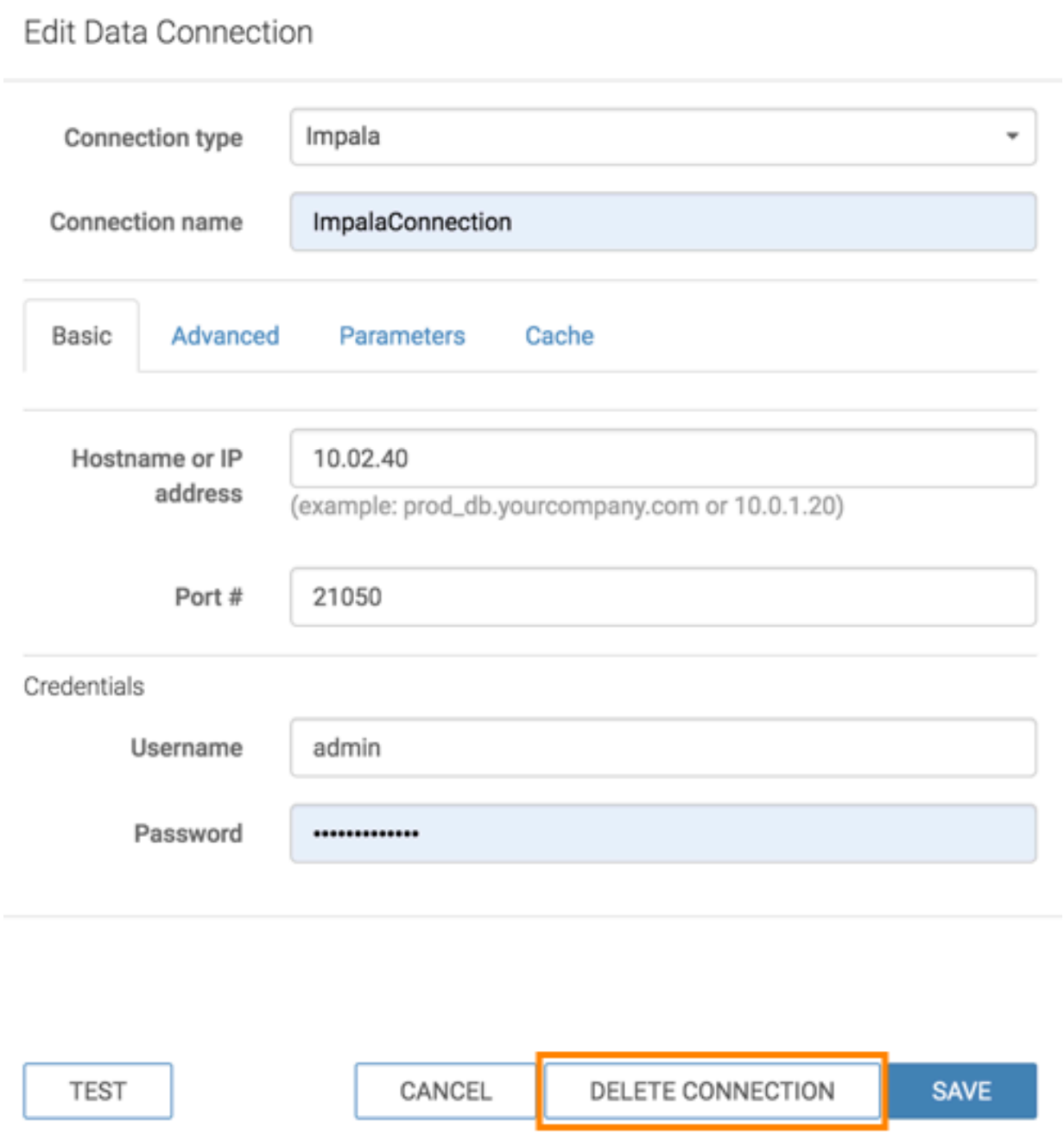
   The Data view appears, open on the Datasets tab.



2. In the side bar, click the Edit Connection (pencil) button to the right of the connection you want to delete.

   The Edit Data Connection modal window appears.

3. At the bottom of Edit Data Connection modal window, click DELETE CONNECTION.



Results

After this operation succeeds, the connection is deleted and its name no longer appears on the side navigation bar.

Related Information

Deleting datasets

Using the Connection Explorer

CDP Data Visualization enables you to view existing data connections and all data tables accessible through them. In the Connection Explorer interface, you can create new connections to data sources, preview that data, create new

datasets, navigate to these datasets, import supplemental data, and locate existing dashboards and visuals based on specific datasets.

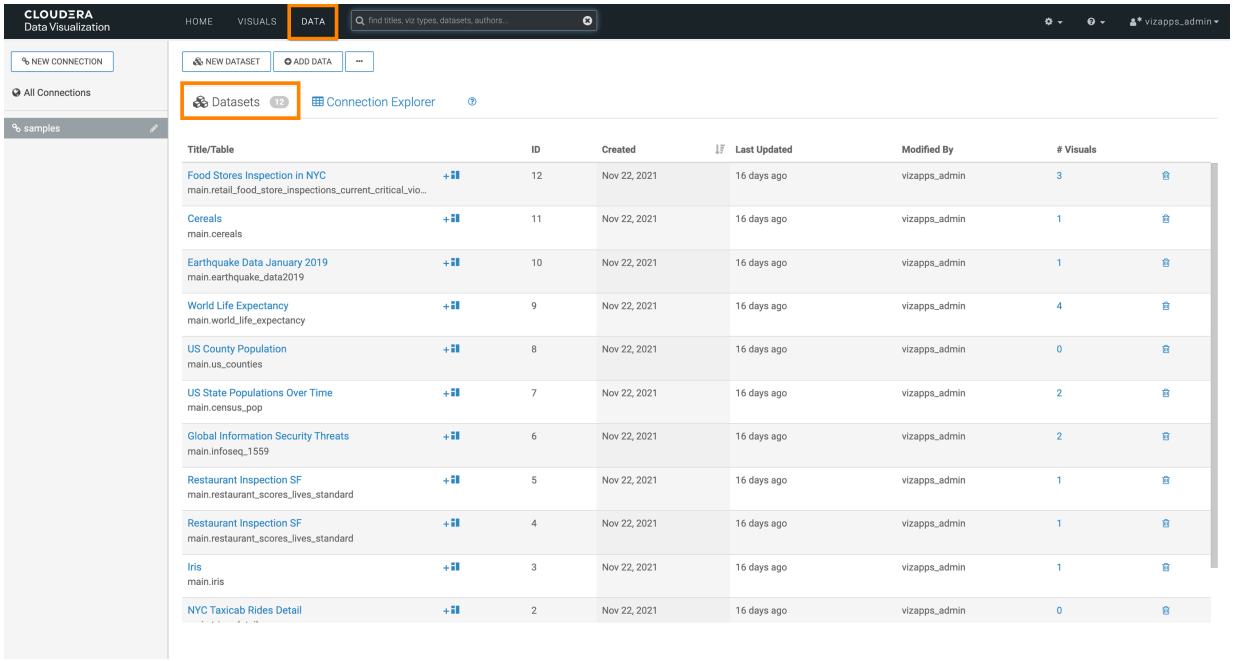
Discovering the Connection Explorer interface

Learn how you can navigate to the Connection Explorer interface and use it to connect to data in CDP Data Visualization.

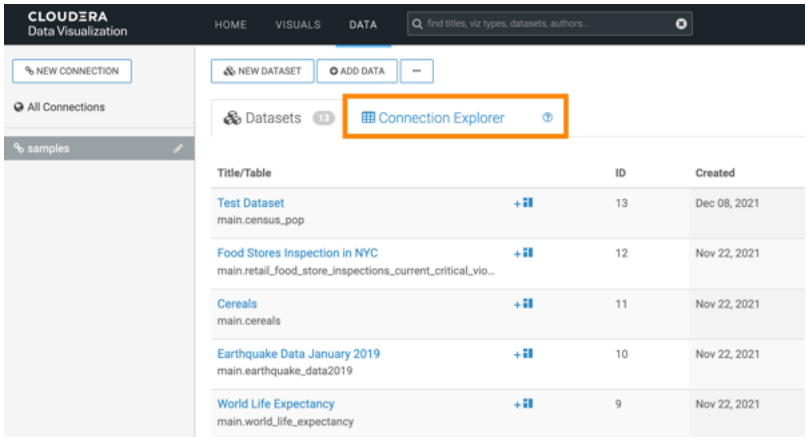
Procedure

1. On the main navigation bar, click DATA.

   The Data view appears, open on the Datasets tab. The Datasets tab lists all existing datasets on the connection.

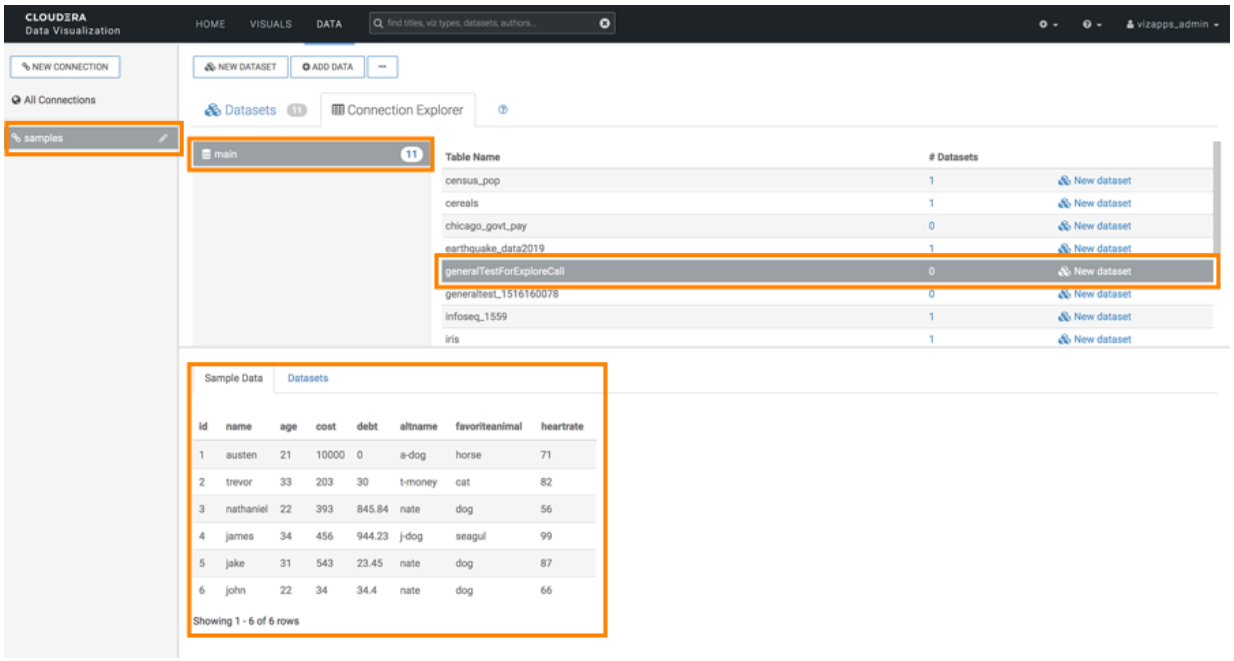


2. In the main area, click the Connection Explorer tab.

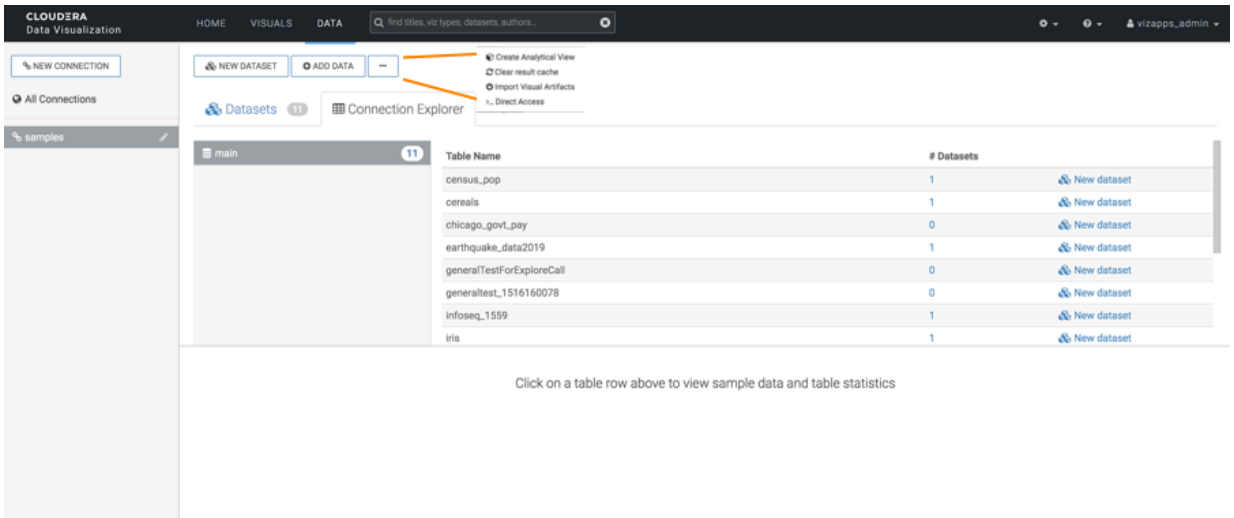


   The Connection Explorer interface appears. The Connection Explorer tab (current) enables you to explore the databases and tables available on the connection, and to manage all functions related to the data on the connection.

In this view, you can click a connection in left navigation that you want to explore and select a database. You can also select a specific table from that database, and explore its details.



The Connection Explorer interface contains the following items:



a. New Connection is for connecting to any source of data.

b. New Dataset is for creating datasets, which are necessary for developing dashboards and visuals. You can also use start a new dataset from a specified table.

c. For SQLite connections, the Add Data option enables you to introduce data that enriches your datasets from outside sources.

d. The Supplemental menu, under the (ellipsis) icon, opens new options.

1.
2. For Impala connections, clicking the Clear result cache option under the supplemental menu reloads the full table definition.
3. Import Visual Artifacts option under the supplemental menu enables you to restore or import visual artifacts from a backup *.json file.
4. Direct Access enables you to access data directly by running SQL queries. You can build datasets from specific SQL queries, as opposed to starting with an existing table.

e.

f.

g. The databases area of the screen shows all databases that you can access through the current connection. In our example, there is one called main (selected). Selecting a database shows its tables.

h. In the list of tables, the # Datasets column lists the number of datasets that use the particular table as their initial definition.

i. New Dataset is for creating a dataset on a specific table.

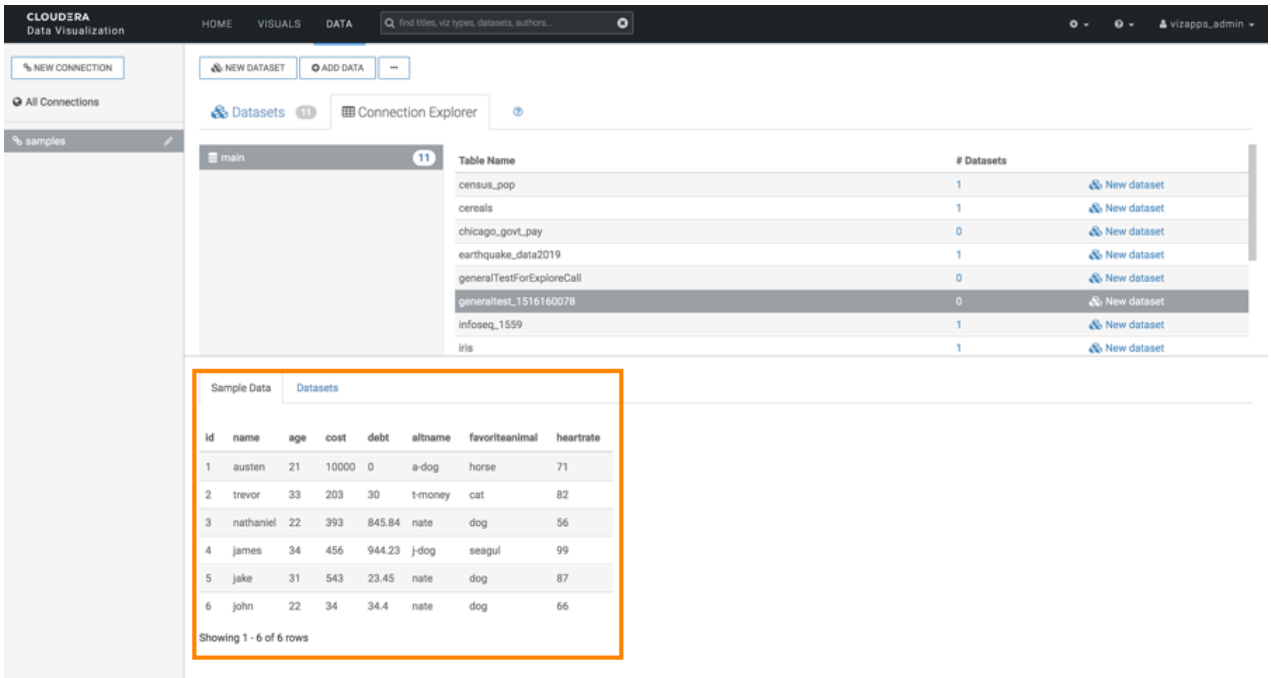
Previewing data table details

Learn how you can preview table details directly in the Connection Explorer interface.

To see more information about data tables in the Connection Explorer, click the row of a table. When you click a row, two tabs, Sample Data and Datasets appear below the list of tables.

Sample data

When you click a table, you can preview the table data in the Sample Data view.



Datasets

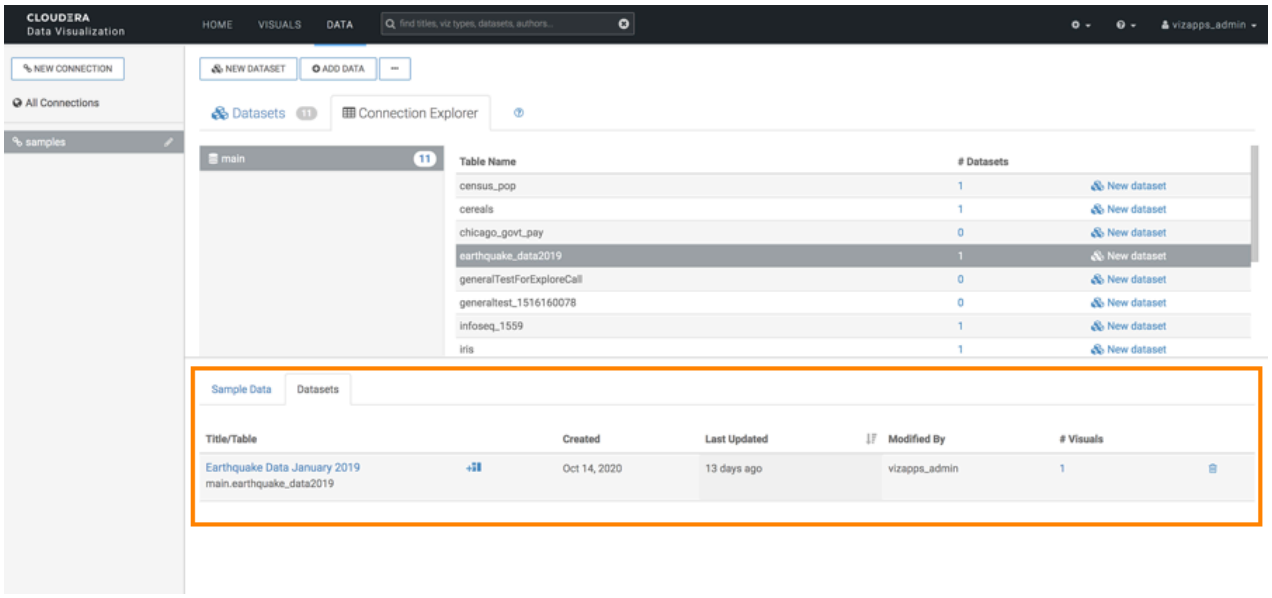
When you click a table, you can check the following data in the Datasets view:

- Title/Table
- Created date
- Last Updated date
- Modified by username
- # Visuals for a link to the dashboards and visuals based on the dataset.

You can also perform the following actions:

- Navigate directly to the dataset interface, where you can rename the dataset, modify fields and other parameters, create joins, and so on.
- Start a new dashboard based on the dataset.

- Order datasets based on any of the table columns.
- Delete datasets.



Using the Direct Access interface

The Direct Access interface of CDP Data Visualization enables you to run SQL queries on data connections directly on the DATA page. You can quickly examine the structure of tables that are available on the connection, build a query using standard SQL syntax, preview its results, and then create a dataset on the query. You can also download the data, if needed.

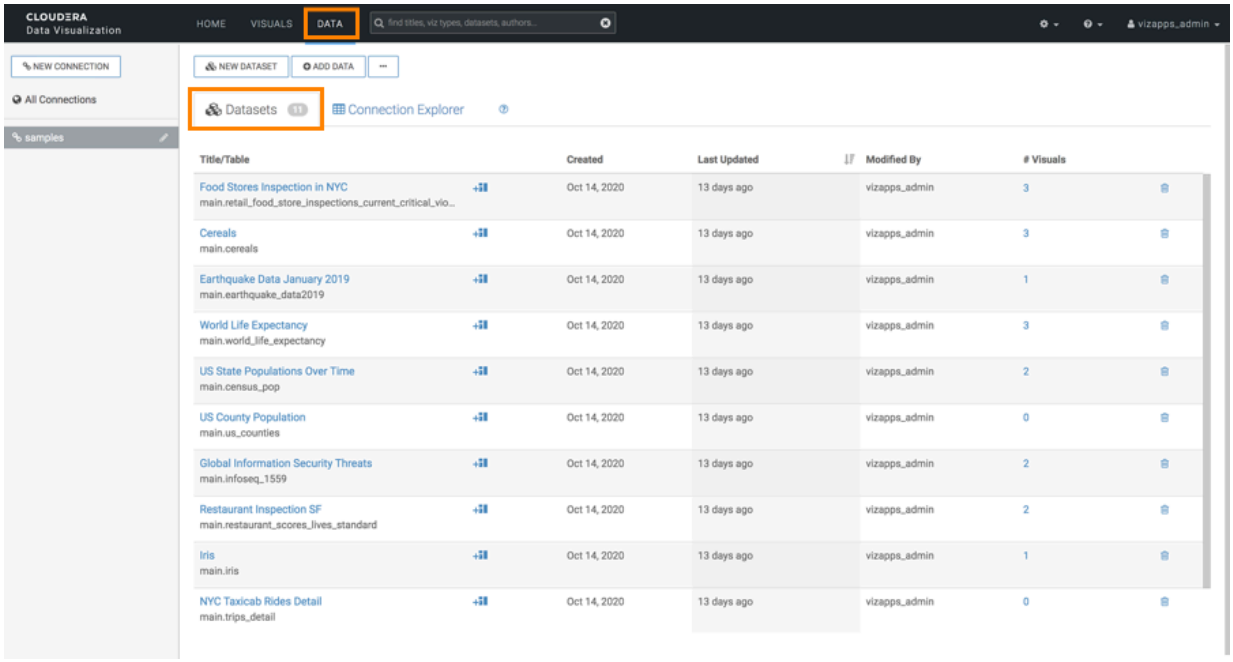
Running a SQL query in Direct Access

Learn how you can run a SQL query in the Direct Access interface of CDP Data Visualization.

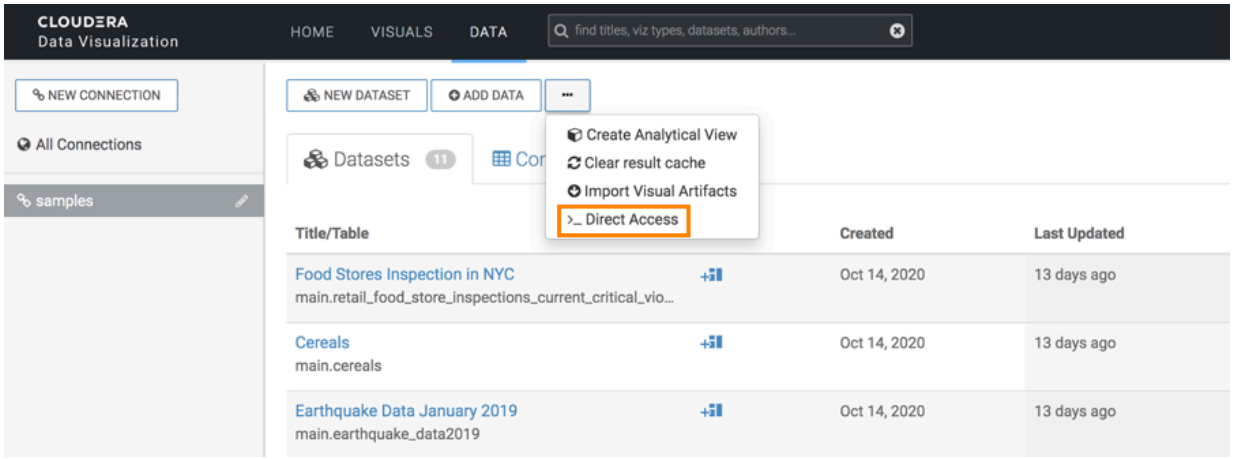
Procedure

1. On the main navigation bar, click DATA.

   The DATA view appears, open on the Datasets tab.



2. Open the Supplemental menu (ellipsis icon) and click >_ Direct Access.



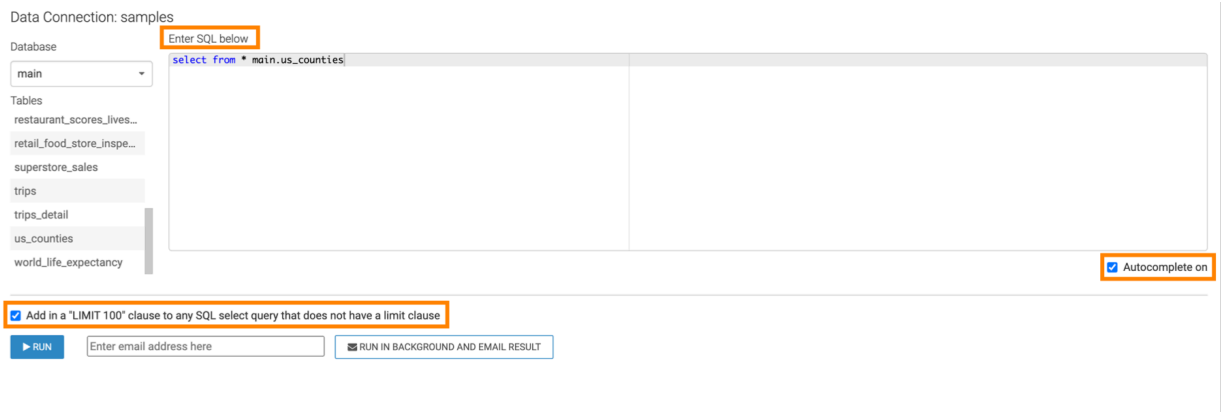
   The Direct Access interface appears, where you can select the database you want to access.

3. Add your SQL data query to the Enter SQL below code-enabled text box.

   The Autocomplete option is on by default. It validates the syntax of the SQL query you enter.

   If you want to limit the number of, you have two options:

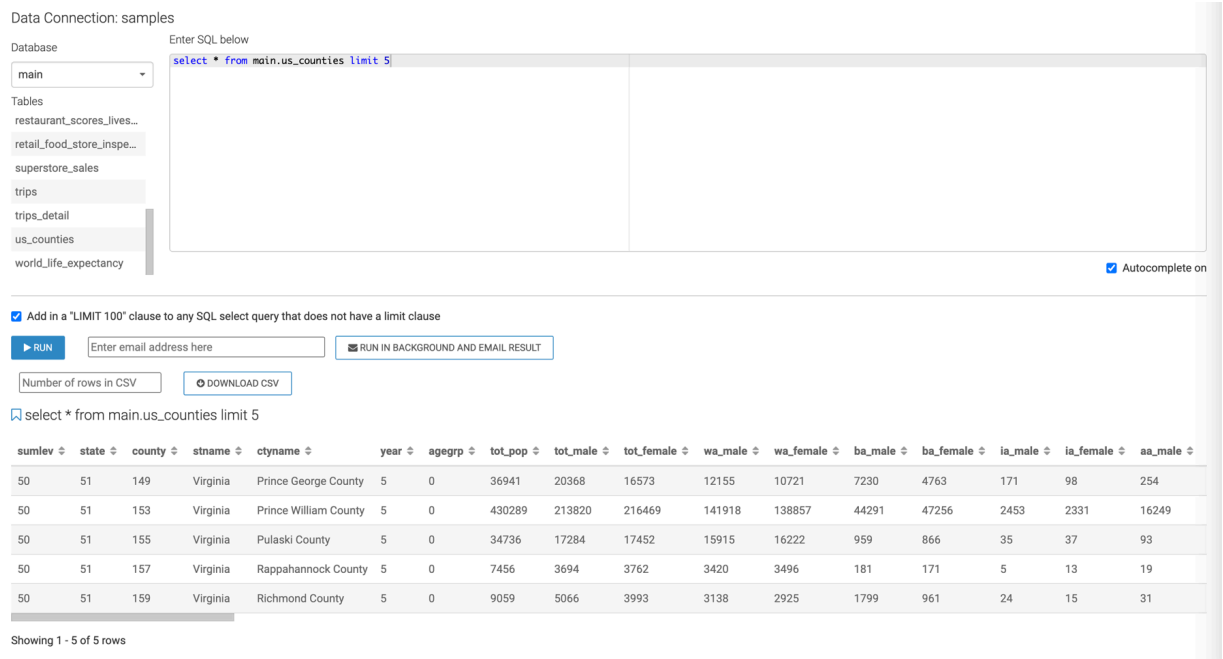
   a. You can add a limit clause in the SQL query syntax.
   b. You can mark the Add in a "LIMIT 100" clause to any SQL select query that does not have a limit clause option. This limitation is on by default. If you set a record limit in the SQL query, it will override this default option, even if it is checked.



4. Click RUN to execute the SQL query.

   After the query executes, the results area shows the query results.

   In this example, the following query has been run: select * from main.us_counties limit 5



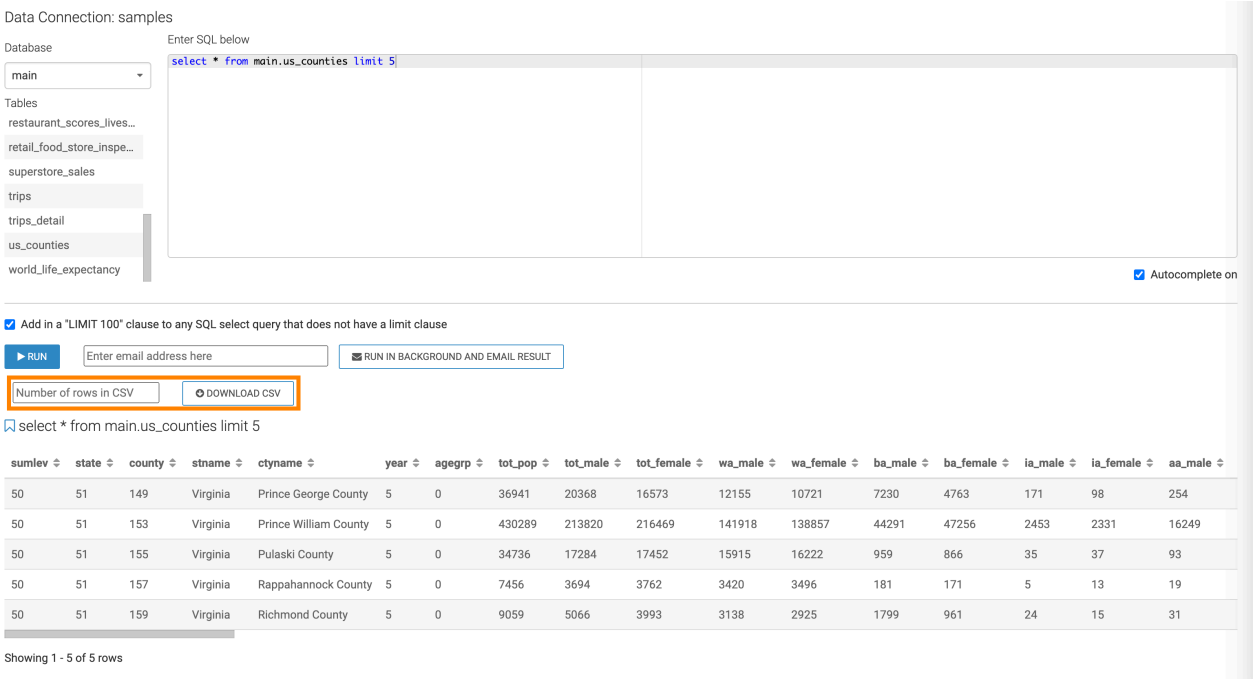
Note: If there is a error in the query the line number with the error will be displayed if the database returns it.

Downloading the results of a Direct Access query

After obtaining query results in the Direct Access interface, you can download the records in CSV format.

You can download the results in CSV format, by clicking DOWNLOAD CSV. The system saves the CSV file to your default download directory.

If you only want to download a subset of the query results, you can specify the Number of rows in CSV. This will limit the number of records in the CSV file.



sumlev	state	county	stname	ctyname	year	agegrp	tot_pop	tot_male	tot_female	wa_male	wa_female	ba_male	ba_female	ia_male	ia_female	aa_male
50	51	149	Virginia	Prince George County	5	0	36941	20368	16573	12155	10721	7230	4763	171	98	254
50	51	153	Virginia	Prince William County	5	0	430289	213820	216469	141918	138857	44291	47256	2453	2331	16249
50	51	155	Virginia	Pulaski County	5	0	34736	17284	17452	15915	16222	959	866	35	37	93
50	51	157	Virginia	Rappahannock County	5	0	7456	3694	3762	3420	3496	181	171	5	13	19
50	51	159	Virginia	Richmond County	5	0	9059	5066	3993	3138	2925	1799	961	24	15	31

The system saves the CSV file to your default download directory.