Cloudera Data Visualization 7.2.9

Connecting to Data

Date published: 2020-10-30
Date modified: 2024-12-18

CLOUDERA

Legal Notice

Contents

Connecting to a data source in Cloudera Data Visualization

Cloudera Data Visualization allows you to create connections to many types of external data sources, enhancing your data analysis and visualization capabilities.

Cloudera Data Visualization currently supports the following connection types:

- Hive
- Impala
- MariaDB
- MySQL
- PostgreSQL
- Solr [Technical Preview]
- Spark SQL
- SQLite (not supported in Cloudera Data Warehouse)
- Snowflake [Technical Preview]

This range of supported connection types provides flexibility and versatility in connecting to various data sources for comprehensive data analysis.

When using Cloudera Data Visualization with Cloudera Data Warehouse, the data connection is automatically set up, but you can connect to various other data sources as well to suit your specific requirements.

In Cloudera AI, you can connect to an Impala or a Hive data warehouse, or tie in data from predictive Cloudera AI models.

Cloudera Data Visualization also includes a samples SQLite connection that contains example datasets. This connection is read-only by design, so sample datasets cannot be modified.

Related Information

Connections

Datasets

Data modeling

Creating a Cloudera AI data connection to Impala

Learn how to connect natively to data stored in Impala when using Cloudera Data Visualization in Cloudera AI.

About this task

Before you start using data modeling and visualization functions, you must connect to your data. The following steps show you how to create a new Cloudera AI data connection in Cloudera Data Visualization to an Impala data warehouse.

Note:

You must have the Manage data connections privilege or be an admin to be able to create and manage connections in Cloudera Data Visualization.

Setting user privileges requires administrator-level access. You can log in as an administrator using the default admin account with the following credentials:

- Username: vizapps_admin
- Password: vizapps_admin

When you create a connection, you automatically get privileges to create and manage the associated datasets. You can also build dashboards and visuals within these datasets.

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

Before you begin

If you are using a Cloudera Private Cloud Base cluster running Impala with Kerberos for authentication, make sure that Kerberos credentials are configured in Cloudera AI before creating a Cloudera AI data connection to the Impala data warehouse. This ensures seamless integration and authentication between Cloudera Data Visualization and the Impala cluster. If you add Kerberos credentials after launching the Cloudera Data Visualization app, you need to restart the app for the changes to take effect.

For more information on using Kerberos for authentication in Cloudera AI, see *Hadoop Authentication for AI Workspaces*.

Procedure

1. On the main navigation bar, click DATA.

The DATA interface opens, displaying the Datasets tab.

2. On the side menu bar, click NEW CONNECTION.

> **Note:** Only users with Manage data connections privilege or administrators can access the NEW CONNECTION button.

The Create New Data Connection modal window appears.

3. Choose Impala from the Connection type drop-down list and assign a name to your connection.

Create New Data Connection   ✕

Connection type	Impala ▾
Connection name	doc-test

Basic   Advanced   Parameters   Data

Hostname or IP address	Enter IP address of the server where your data resides
	(example: prod_db.yourcompany.com or 10.0.1.20)
Port #	443

Credentials

Username	
Password	············

TEST   CONNECT

In this example, the Impala connection is made through Knox. Knox always uses TLS encryption and port 443 is the default HTTPS port.

4. Enter the hostname or IP address of the running coordinator.

You can retrieve this information from the JDBC URL of the Impala DW.

5. Add 443 in the Port # field.

6. Enter your workload username and password as credentials.

7. Click the Advanced tab to configure additional details.

Create New Data Connection   ✕

| Connection type | Impala | ▾ |

| Connection name | doc-test |

Basic   **Advanced**   Parameters   Data

Connection mode   ○ Binary   ● HTTP

HTTP Path   [SQL path (default cliservice)]

Socket type   ○ Normal   ● SSL   ○ SSL with certificate

Authentication mode   ○ NoSasl   ● Plain   ○ LDAP   ○ Kerberos

Socket Timeout   [60]

Impersonation ⓘ   ☐ Enabled

Trusted ⓘ Impersonation   ☐ Enabled

Application Name   [viz]

[TEST]   [CONNECT]

a) For HTTP connection mode, locate the Impala Endpoint for the Data Hub.

Event History   Autoscale   Endpoints (3)   Tags (8)   Hardware   Network   Telemetry   Repository Details   Image Details   Recipes (0)
Database   Upgrade

Name	URL
CM-API	https://jingalls-test-dm-gateway.euph-aw.a465-9q4k.cloudera.site/jingalls-test-dm/cdp-proxy-api/cm-api/ ⧉
Impala	https://jingalls-test-dm-gateway.euph-aw.a465-9q4k.cloudera.site/jingalls-test-dm/cdp-proxy-api/impala/ ⧉
Impala	jdbc:impala://jingalls-test-dm-gateway.euph-aw.a465-9q4k.cloudera.site:443/;ssl=1;transportMode=http;httpPath=jingalls-test-dm/cdp-proxy-api/impala;AuthMech=3; ⧉

b) Copy and paste it into the HTTP Path field.

c) Set any additional details as required.

8. Check the Parameters and Data tabs for more configuration options.

Create New Data Connection ✕

Connection type	Impala ▾
Connection name	doc-test

Basic   Advanced   Parameters   **Data**

Concurrency ⓘ	55
Concurrency Per ⓘ **User**	5
Query Timeout ⓘ **(Minutes)**	22
Query Still Loading Warning (Seconds)	20
Row upload limit ⓘ	10000
Result Cache	☐ Enabled
Cache Retention Time (seconds)	86400

TEST   **CONNECT**

9. Once you finish configuring the settings, click TEST to check the connection.

10. Click CONNECT to establish the connection.

Results

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

Related Information

RBAC permissions

Setting role privileges

Promoting a user to administrator

Hadoop Authentication for ML Workspaces

Creating a Cloudera AI data connection to a Hive data warehouse

Learn how to connect natively to data stored in Hive when using Cloudera Data Visualization in Cloudera AI.

About this task

Before you start using data modeling and visualization functions, you must connect to your data. The following steps show you how to create a new Cloudera AI data connection in Cloudera Data Visualization to a Hive data warehouse.

Note:

You must have the Manage data connections privilege or be an admin to be able to create and manage connections in Cloudera Data Visualization.

Setting user privileges requires administrator-level access. You can log in as an administrator, using the default admin account with the following credentials:

- Username: vizapps_admin
- Password: vizapps_admin

When you create a connection, you automatically get privileges to create and manage the associated datasets. You can also build dashboards and visuals within these datasets.

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

Before you begin

If you are using a Cloudera Private Cloud Base cluster running Hive with Kerberos for authentication, make sure that Kerberos credentials are configured in Cloudera AI before creating a Cloudera AI data connection to the Hive data warehouse. This ensures seamless integration and authentication between Cloudera Data Visualization and the Hive cluster. If you add Kerberos credentials after launching the Cloudera Data Visualization app, you need to restart the app for the changes to take effect.
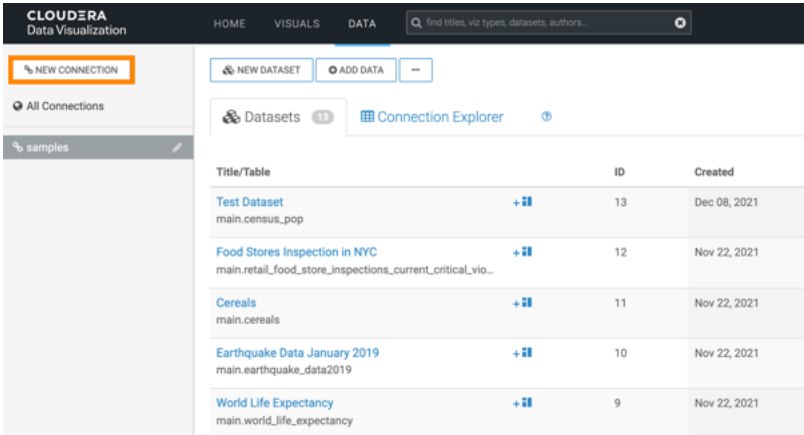
Procedure

1. On the main navigation bar, click DATA.

   The DATA opens, displaying the Datasets tab.
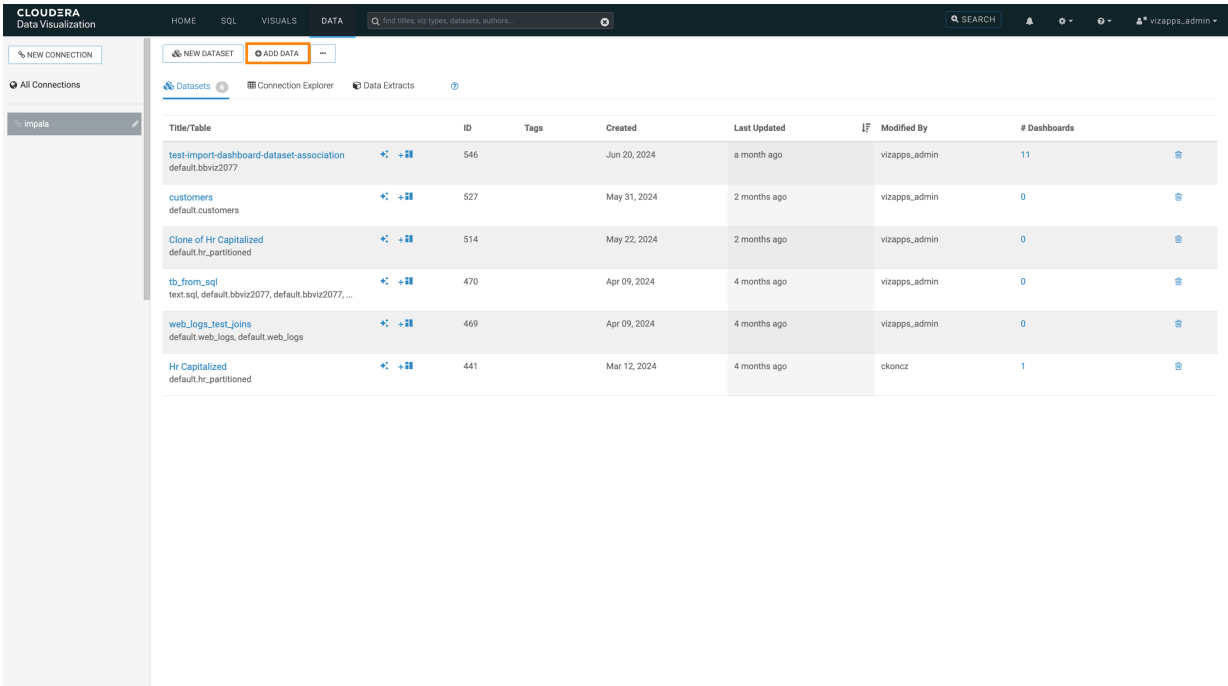
2. On the side menu bar, click NEW CONNECTION.

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

   The Create New Data Connection modal window appears.

3. Choose Hive from the Connection type drop-down list and assign a name to your connection.

Create New Data Connection   ✕

Connection type	Hive ▾
Connection name	doc-test

Basic   Advanced   Parameters   Data

Hostname or IP address   Enter IP address of the server where your data resides
(example: prod_db.yourcompany.com or 10.0.1.20)

Port #   443

Credentials

Username

Password   •••••••••••••

TEST   CONNECT

4. Enter the hostname or IP address of the running coordinator.

You can get the coordinator hostname from the JDBC URL of the Hive DW.

5. Use port 443.

6. Enter your workload username and password as credentials.

7. Click the Advanced tab to configure the additional details.

Create New Data Connection ✕

Connection type	Hive ▾
Connection name	doc-test

Basic   **Advanced**   Parameters   Data

Connection mode	○ Binary ◉ HTTP
HTTP Path	SQL path (default cliservice)
Access Token	Access token (optional)
Socket type	○ Normal ◉ SSL ○ SSL with certificate
Authentication mode	○ NoSasl ◉ Plain ○ LDAP ○ Kerberos
Socket Timeout	60
Impersonation ⓘ	☐ Enabled
Trusted ⓘ **Impersonation**	☐ Enabled
Application Name	viz

TEST   CONNECT

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

Create New Data Connection   ✕

Connection type	Hive ▾
Connection name	doc-test

Basic   Advanced   **Parameters**   Data

Parameter Name	Parameter Value	
hive.server2.async.exec.async.compile	false	🗑
Add new row		

TEST   CONNECT

9. Check the Data tab for more configuration options.

10. Once you finish configuring the settings, click TEST to test the connection.

11. Click CONNECT to establish the connection.

Results

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

Related Information

RBAC permissions

Setting role privileges

Promoting a user to administrator

Creating a Cloudera Data Warehouse data connection in Cloudera Data Visualization

Learn how to connect to data when using Cloudera Data Visualization in Cloudera Data Warehouse data service. You can connect Cloudera 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 Cloudera Data Warehouse URL has the same compute instance ID as your Virtual Warehouse.

> **Note:**
>
> To create and manage connections in Cloudera Data Visualization, you must have the Manage data connections privilege or hold administrative privileges. In Cloudera Data Warehouse, these are the members of the Admin Groups associated with the Cloudera Data Visualization instance.

When you create a connection, you automatically gain privileges to create and manage datasets associated with this connection, and to build dashboards and visuals within these datasets.

- For more information on the Manage data connections privilege, see *RBAC permissions*.
- For instructions on how to define privileges for a specific 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 Cloudera Data Visualization from the left navigation panel in Cloudera Data Warehouse.
2. On the main navigation bar, click DATA.

   The DATA interface appears, open on the Datasets tab.

3. In the side menu bar of DATA, click NEW CONNECTION.

> **Note:** The NEW CONNECTION button is only accessible to users assigned to roles with the 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, and provide a name for your connection.

Note: SQLite connection is not supported in Cloudera Data Warehouse, and the option is disabled in the connection list.

Create New Data Connection   ✕

Connection type	CDW Hive ▾
Connection name	doc-test

Basic   Advanced   Parameters   Data

Hostname or IP address — Enter IP address of the server where your data resides
(example: prod_db.yourcompany.com or 10.0.1.20)

Port # — 443

Credentials

Username

Password — ••••••••••••

TEST   CONNECT

5. Select a Cloudera Data Warehouse 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

6. Click the Advanced tab and configure the additional details.

> ⚠️ **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.

7. Click TEST.

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

8. 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 Cloudera Data Science Workbench data connection to a data warehouse

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

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 Cloudera Data Science Workbench data connection to a running Impala system.

> **Note:**
>
> To create and manage connections in Cloudera Data Visualization, you must have the Manage data connections privilege or hold administrative privileges.
>
> Setting user privileges requires administrator-level access. To log in as an administrator, you can use the default admin account with the following credentials:
>
> - Username: vizapps_admin
> - Password: vizapps_admin

When you create a connection, you automatically gain privileges to create and manage datasets associated with this connection, and to build dashboards and visuals within these datasets.

- For more information on the Manage data connections privilege, see *RBAC permissions*.
- For instructions on how to define privileges for a specific 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 the 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, and provide a name for the connection.

4. Enter the hostname or IP address of the running coordinator.

5. Under Port #, enter the port number.

6. Use your workload username and password as credentials.

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.

8. Click TEST.

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

9. 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

Managing data connections

You can change the properties of existing data connections and you can also delete connections that are not associated with any dataset.
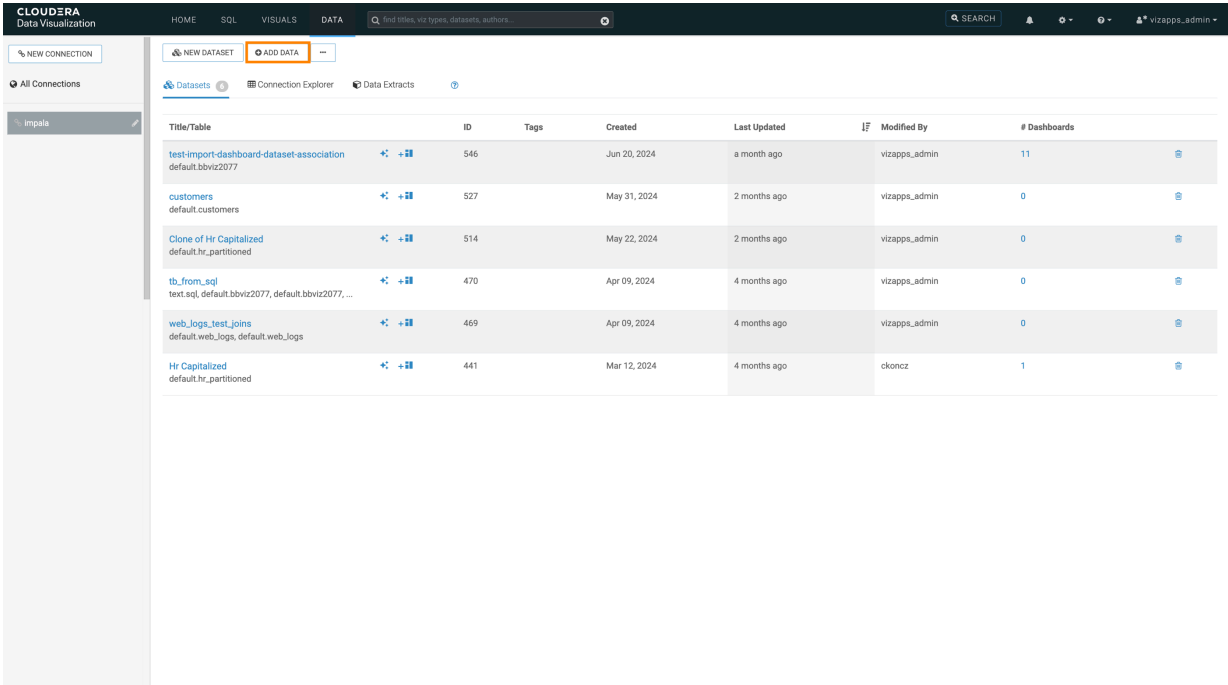
Editing a data connection

Learn how to edit a data connection in Cloudera 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.

Edit Data Connection

Connection type	Impala ▾
Connection name	ImpalaConnection

Basic	Advanced	Parameters	Cache

Hostname or IP address	10.02.40
	(example: prod_db.yourcompany.com or 10.0.1.20)
Port #	21050

Credentials

Username	admin
Password	••••••••••••••

TEST   CANCEL   DELETE CONNECTION   SAVE

Edit Data Connection

Connection type	Hive ▼
Connection name	HiveConnection

Basic   Advanced   Parameters   Cache

Hostname or IP address	20.01.20
	(example: prod_db.yourcompany.com or 10.0.1.20)
Port #	10000

Credentials

Username	admin
Password	•••••••••••••

TEST   CANCEL   DELETE CONNECTION   SAVE

4. At the bottom of the modal, click TEST.

5. If the connection is verified, click SAVE.

Connection Verified!   ✕

TEST   CANCEL   DELETE CONNECTION   SAVE

Results

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

Setting concurrency for a data connection

You can improve resource management in Cloudera Data Visualization by restricting the number of simultaneous connections on specific datasets.

About this task

Note: This setting is only applicable to Impala, Hive, and SQLite connections.

Procedure

1. Open the Edit Data Connection modal.

   For more information, see *Editing a data connection*.

2. Click the Data tab.

3. In the Concurrency field, provide a number to set the maximum limit for simultaneous requests.

   Note: The default maximum concurrency setting per connection is 100. You can change it to a different value in Site Settings Advanced Settings . If you change the default value, ensure that the new maximum concurrency aligns with the requirements of your data connection.

   ```
   DATACONNECTION_CONCURRENCY_DEFAULT = 100
   DATACONNECTION_CONCURRENCY_MAX = 100
   ```

4. In the Concurrency Per User field, provide a number to limit the maximum number of simultaneous requests a user can send.

   Note: You can use this setting to prevent a single user from blocking access for a connection. The default concurrency setting per user is 5, with a default maximum concurrency setting per user at 100. You can adjust these values in Site Settings Advanced Settings .

   ```
   DATACONNECTION_CONCURRENCY_USER_DEFAULT = 5
   DATACONNECTION_CONCURRENCY_USER_MAX = 100
   ```

5. Click SAVE to apply the changes.

Setting a warning if the query loads for too long

A data connection admin user can set the time limit for every connection before a warning is shown when data is loaded through a query.

About this task

For more information on creating or editing data connections, see *Connection to a data source* and *Editing a data connection*.

Procedure

1. Select the Data tab in the Data Connection modal.
2. Enter the number of seconds you want to pass before the warning appears in Query Still Loading Warning (Seconds).

It is also possible to have this message appear when viewing a dashboard, see *Dashboards*.

Related Information

Dashboards

Connecting to a data source in Cloudera Data Visualization

Editing a data connection

Deleting a data connection

Learn how you can remove an existing data connection in Cloudera Data Visualization.

About this task

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 the Edit Data Connection modal window, click DELETE CONNECTION.

Results

After the 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

Cloudera 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.
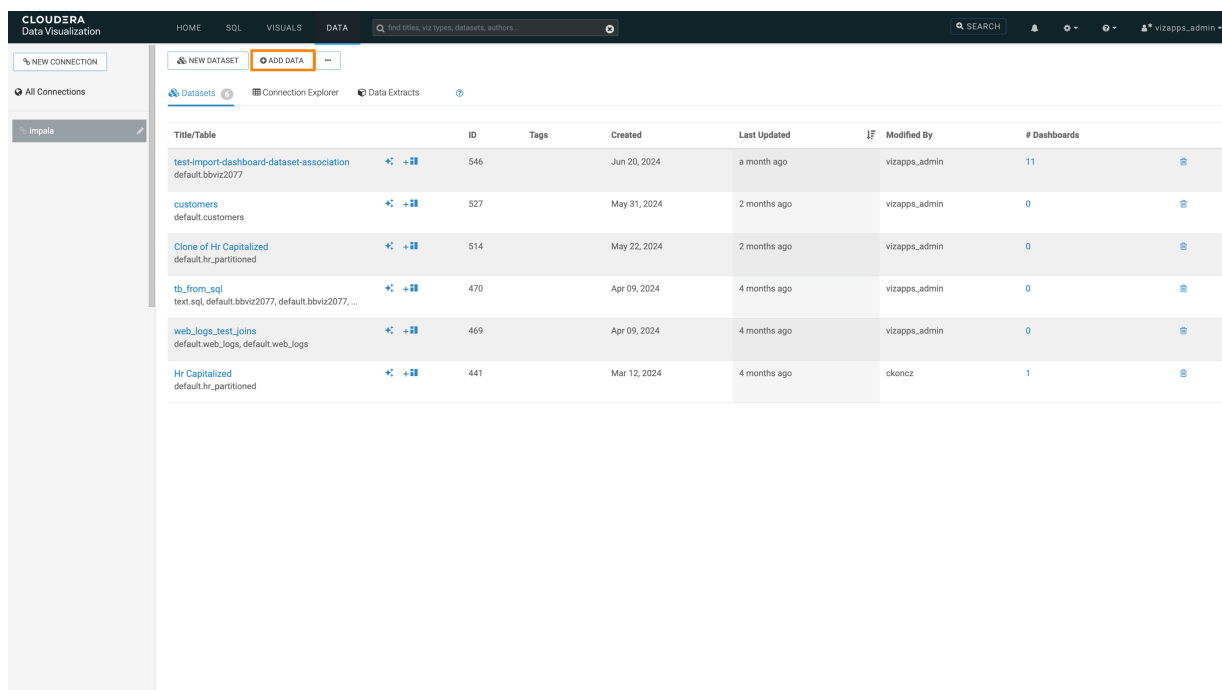
Discovering the Connection Explorer interface

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

Procedure

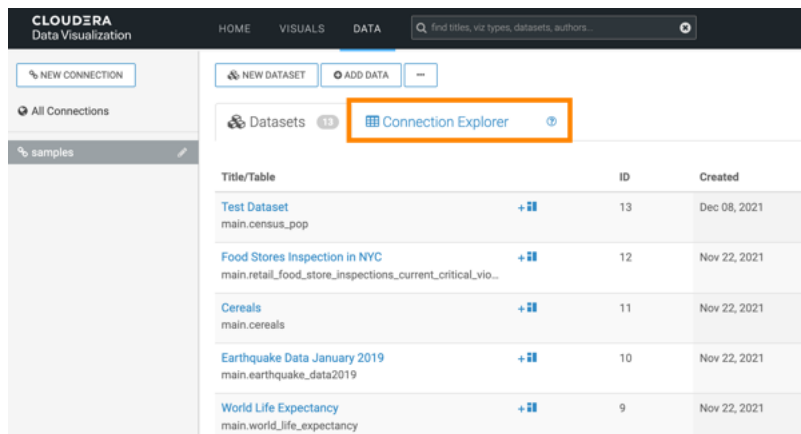**1.** On the main navigation bar, click DATA.

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, where you can explore the databases and tables available on the connection and you can also manage all functions related to the data on the connection. 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.
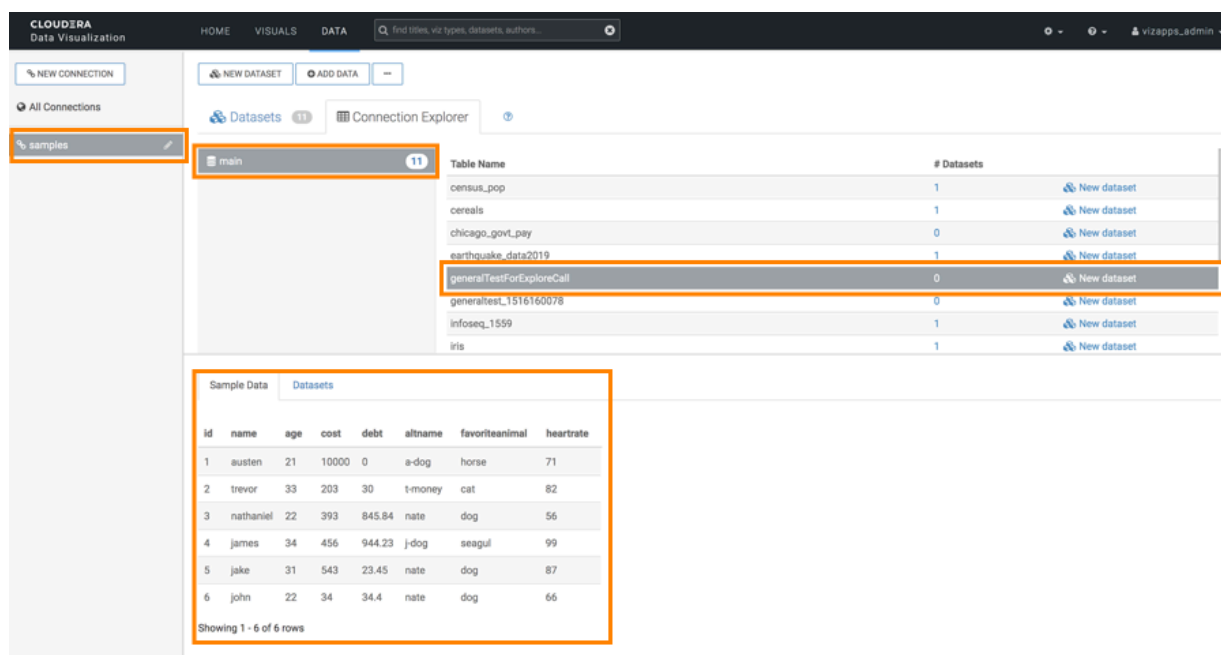
The Connection Explorer interface contains the following items:
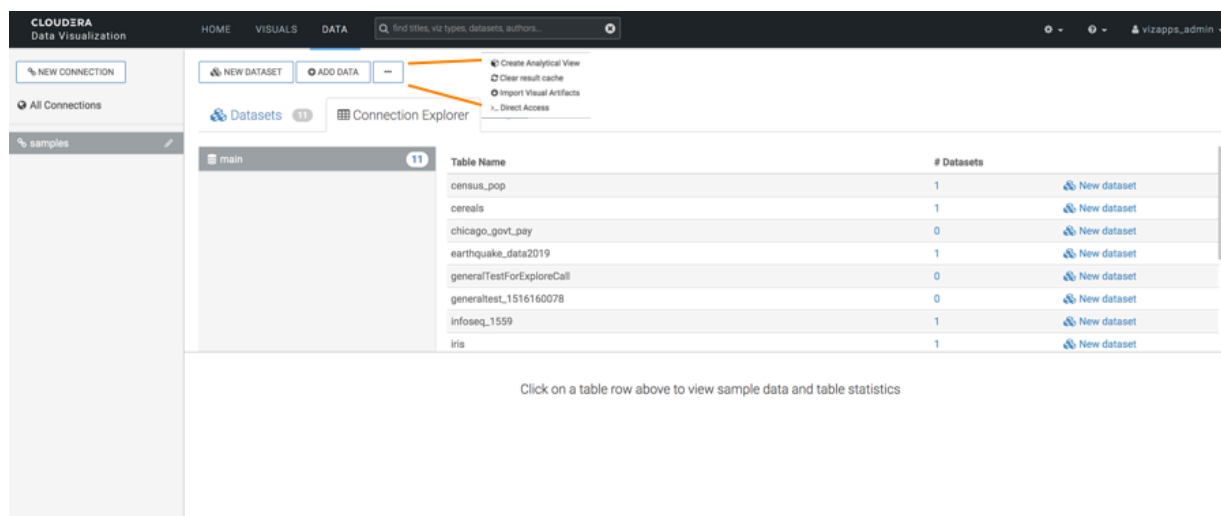
- New Connection is for connecting to any source of data.
- New Dataset is for creating datasets, which are necessary for developing dashboards and visuals. You can also start a new dataset from a specified table.
- For SQLite connections, the Add Data option enables you to introduce data that enriches your datasets from outside sources.
- The Supplemental menu, under the (ellipsis) icon, opens new options.

  a. For Impala connections, clicking the Clear result cache option under the supplemental menu reloads the full table definition.

  b. Import Visual Artifacts option under the supplemental menu enables you to restore or import visual artifacts from a backup JSON file.

  c. 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.

- 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.
- In the list of tables, the # Datasets column lists the number of datasets that use the particular table as their initial definition.
- 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.
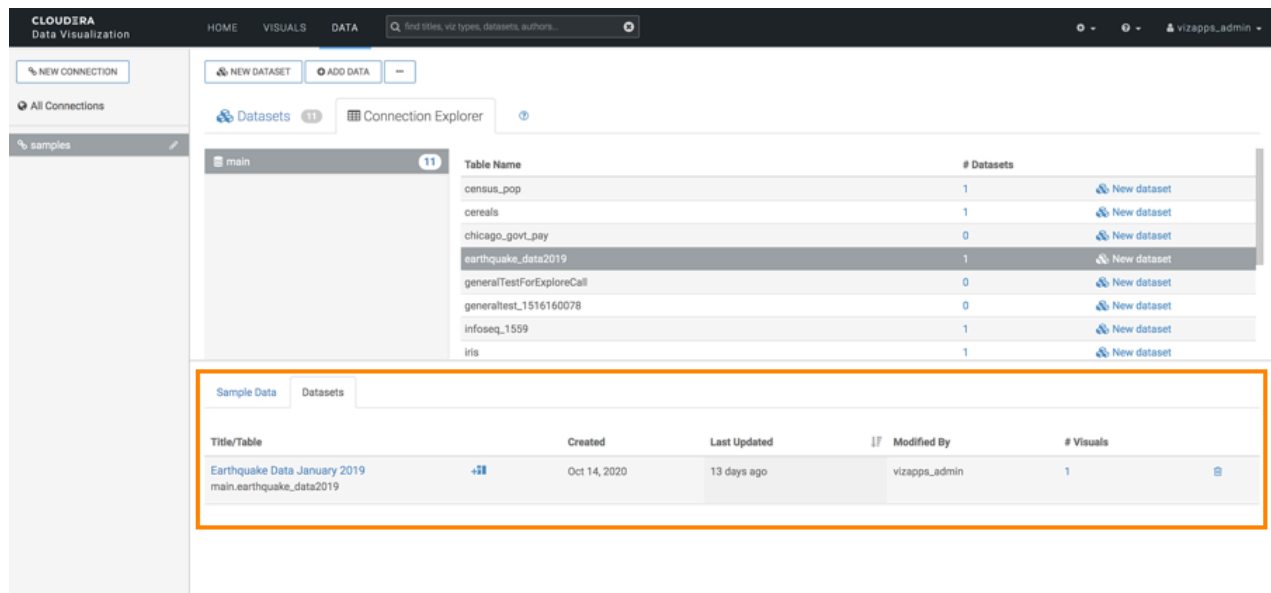
Using the Direct Access interface

The Direct Access interface of Cloudera 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 Cloudera Data Visualization.

Procedure

1. On the main navigation bar, click SQL.

   Alternatively, you can click DATA ••• >_ Direct Access .
2. Select the database in the left-side panel.

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 initiate the SQL query.

After the query runs, the results tabe shows the query results.

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

select * from main.us_counties limit 5

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

Showing 1 - 5 of 5 rows

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

For more information about the content available on the other tabs, see *Cloudera Data Visualization, SQL interface*.

Related Information

Cloudera Data Visualization, SQL interface

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.

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

Adding data

Experience the flexibility of importing and integrating data to enhance your data visualization projects. In Cloudera Data Visualization, you have the capability to add data from flat files. The imported data is processed as a new table and it is integrated into an existing data repository. Subsequently, you can use this data as a supplementary information source for your existing datasets by establishing new table joins. You can also create new datasets that reference the imported data as their primary tables.

Data import is available for the following connections:

- Hive
- Impala
- MariaDB
- MySQL
- PostgreSQL
- Spark SQL

   Note: Data import is not supported in Cloudera Data Warehouse as the connection is read-only.

- SQLite

  Note: This connection is not supported in Cloudera Data Warehouse.

- Snowflake [Technical Preview]

Cloudera Data Visualization supports two primary forms of data sources: CSV and URL.

You can use the Import Data functionality to enhance your existing datasets, incorporating this data by creating joins within the data model. The choice of creating a new dataset based on the imported data depends on your specific business requirements.

For importing data stored in Excel format, see *Preparing Excel files for data import*.

Importing data in CSV format

Learn how to add data to Cloudera Data Visualization from flat CSV files.

About this task

You can use comma-delimited files or files with other delimiting characters. If you want to import data stored in Excel format, see *Preparing Excel files for data import*.

This feature is available on the following connections:

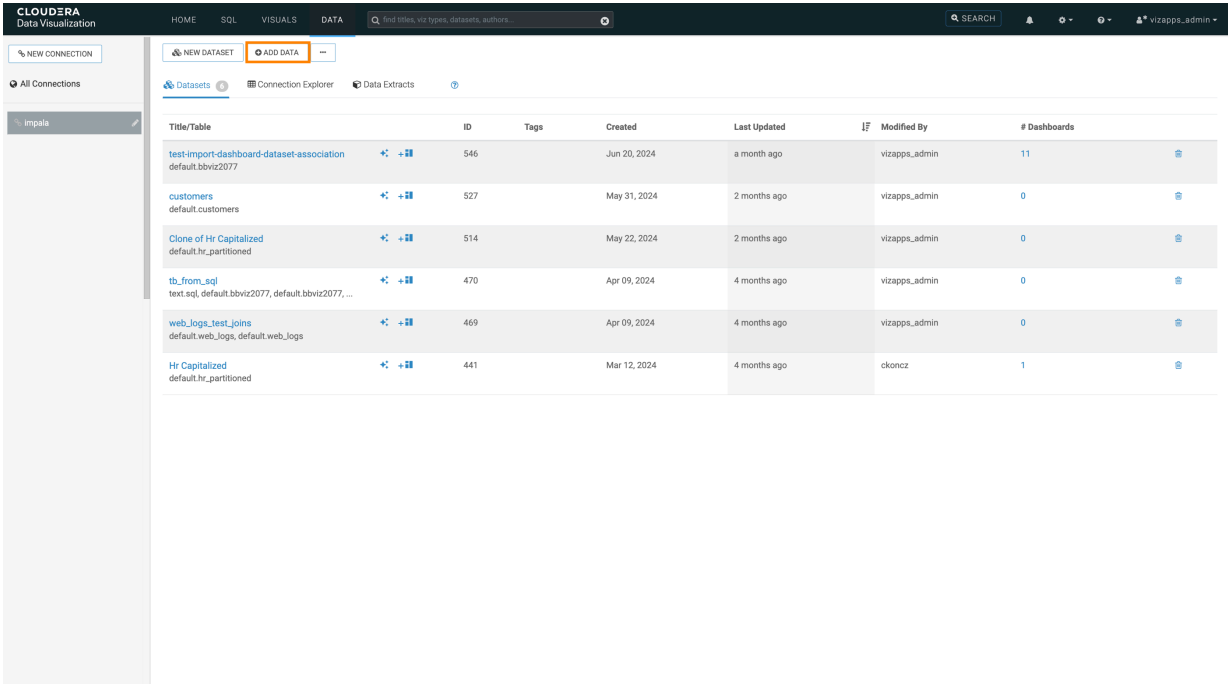| Enabled by default:	Disabled by default:
- SQLite - Hive - Impala - Snowflake [Technical Preview] - Solr [Technical Preview] - Spark SQL	- DuckDB - MS SQL - Teradata

Note: To import data into Cloudera Data Visualization, ensure that you use a data connection with write permissions.
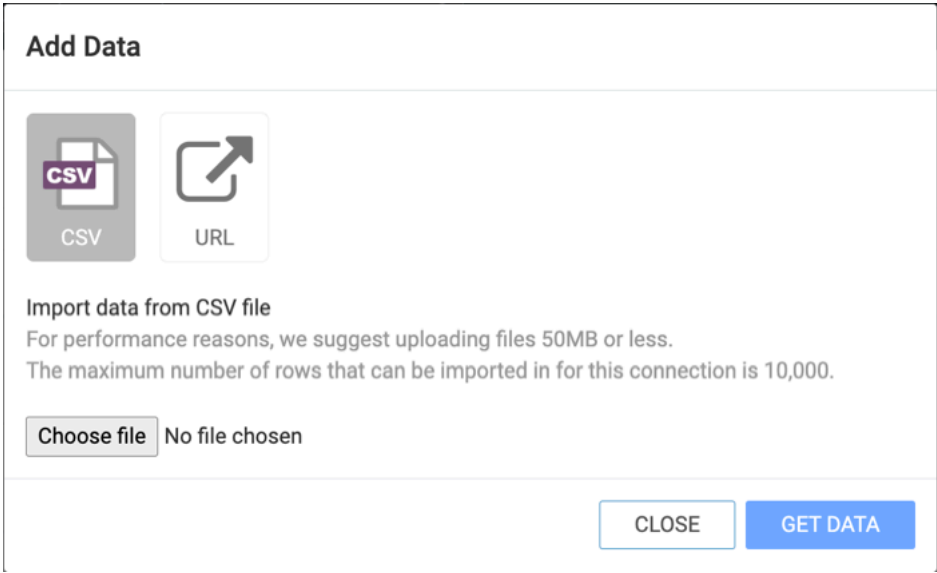
Procedure

1. On the main navigation bar, click DATA.

The Data view appears, showing the Datasets tab.

2. Click ADD DATA.

The Add Data modal window appears.

3. Add the CSV file that contains the data you want to import.

a) Select the CSV option and click Choose file.

b) Using your computer's file browser, locate and select the data file.

c) Click GET DATA.

4. In the Imported File modal window, ensure that the data you are adding is configured correctly.

   a) Verify the name of the file.

   b) Under Database, select which database you want to add data to.

   c) Under Table Name, specify a table. The system usually assigns a numerical suffix, which you can remove.

   d) Under Upon Import, select whether you only want to import data or create a dataset, or create a dataset and a dashboard based on the data.

   - If you select Create Dataset, you are directed to the Detail page of the dataset you have created after the import is completed.
   - If you select Create Dataset and Dashboard, you are directed to the newly created dashboard when the import is completed.

   e) For Column Delimiter, define the delimiter used in the source file. It can be Comma, Tab, Space, Semicolon, Colon, Pipe, Control A, or Other.

   f) Under Locale Setting, you can select one of the common locales: United States, United Kingdom, Sweden, Norway. Or you can select Other to add a separate field next to the locale setting, where you can specify the supported locale. For more information, see *Changing the locale setting of an imported file*.

   g) Under Options, select the data clean-up tasks that Cloudera Data Visualization provides prior to data import. These include Create "EXTERNAL" table, File contains headers, Fill missing columns, Skip malformed rows, and Use '\' as escape character.

   > **Note:** Cloudera Data Visualization can detect many items, for example the presence of header rows, automatically.

   h) In the Data Table preview, each column of the table appears with its data type and name.

   The data types available are Boolean, Integer, Real, String, and Timestamp.

   Cloudera Data Visualization detects the data type automatically, but you can modify them as needed (for example store numerical categories as strings).

   You can change the name of the table column before importing.

   i) Click APPLY CHANGES to save the adjustments you have made.

5. Click CONFIRM IMPORT.

The Confirm File Import modal window appears, where you have to click CONFIRM IMPORT again to finalize the import.

Confirm File Import   ✕

After confirming import, you can use this data in the following manner:

- As supplemental data to an existing dataset; in the Data interface, select the dataset and edit its Data Model to create new table joins.
- As the foundation of a new dataset; in the Data interface, click New Dataset.

This operation may take a long time. It cannot be canceled.

Are you sure you want to confirm the import?

                                                            CANCEL   CONFIRM IMPORT

6. To verify that the data has been imported correctly into your system, check the Connection Explorer interface.

Related Information

Preparing Excel files for data import

Changing the locale setting of an imported file

Adding data through URL

Learn how to how to add a data file using a URL in Cloudera Data Visualization.

About this task

Data files are typically in JSON format. The files may be comma-delimited or they use other delimiting characters.

Note: This feature is available on the following connections:
Enabled by default:

- SQLite
- Hive
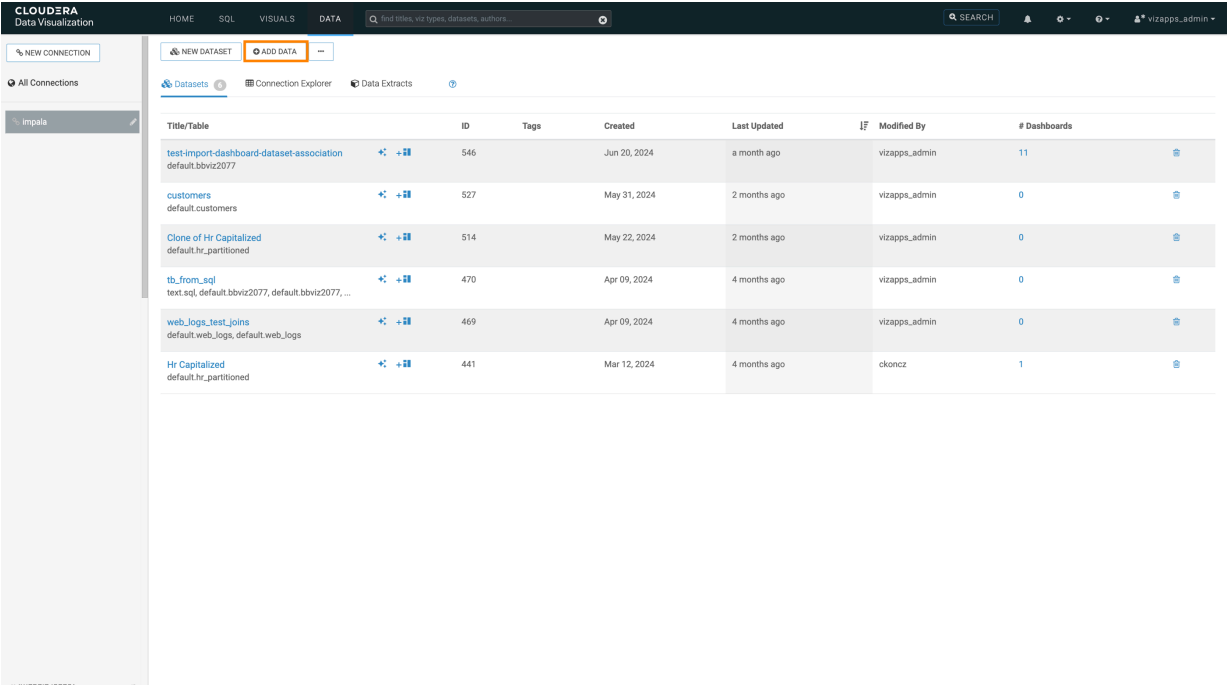- Impala
- Snowflake
- Spark SQL

Disabled by default:
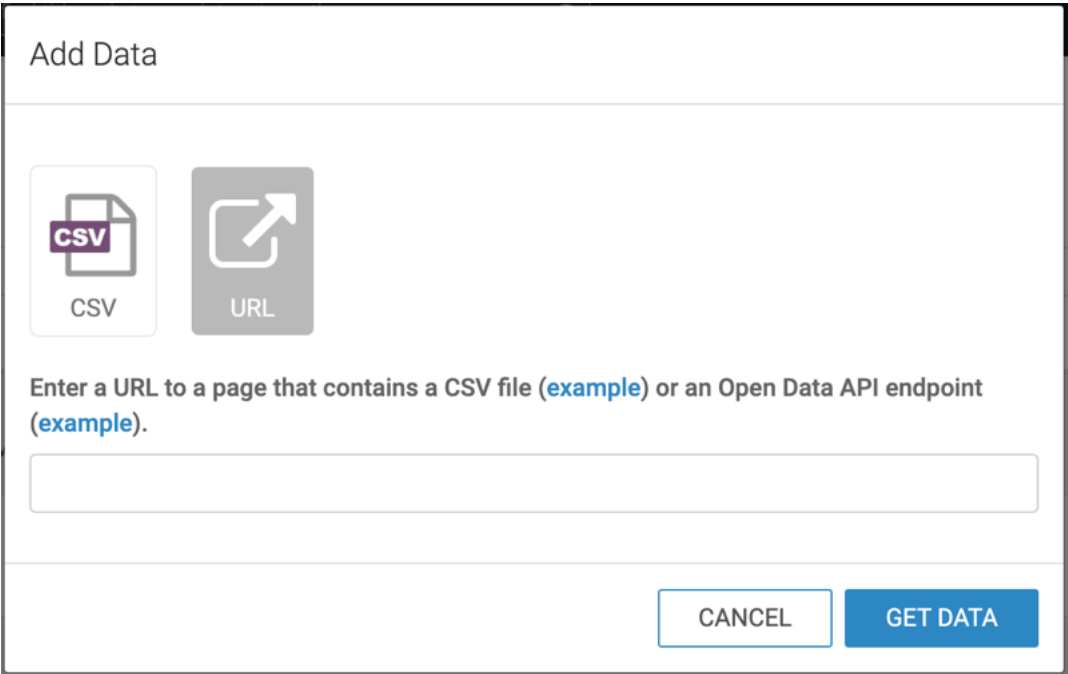
- DuckDB
- MS SQL
- Teradata

Procedure

1. On the main navigation bar, click DATA.

   The Data view appears, open on the Datasets tab.

2. Click ADD DATA.

   The Add Data modal window appears.

3. Select the URL option, and add the address of the file to the text window, and click GET DATA.

4. Ensure that your data is configured correctly before confirming the import:

   a. Under Database, specify documentation. Alternatively, select another database.

   b. Under Table Name, specify city_chicago_staff. The system typically assigns a numerical suffix that you can remove.

   c. Under Upon Import, select whether you only want to import the data or create a dataset and a dashboard based on the data.

   - If you select Create Dataset, you are taken to the Detail page of the dataset you have created when the import is completed.
   - If you select Create Dataset and Dashboard, you are taken to the newly created dashboard when the import is completed.

   d. For Column Delimiter, define the delimiter used in the source file. In this case, the column delimiter is Comma, but it can also be Tab, Space, Semicolon, Colon, Pipe, Control A, or Other. Make the necessary adjustments to this field.

   e. Under Locale Setting, you can select one of the common options: United States, United Kingdom, Sweden, Norway, or Other. Selecting Other adds a separate field next to the locale setting where you can specify the supported locale. For more information, see *Changing the locale setting of an imported file*.

   f. Under Options, select the data clean-up tasks that Cloudera Data Visualization provides prior to data import. These include File contains headers, Fill missing columns, Skip malformed rows, and Use '\' as escape character.

   > **Note:** Cloudera Data Visualization can detect many items automatically, such as the presence of header rows.

   g. In the Data Table preview, each column of the table appears with its data type and name. Cloudera Data Visualization detects the data type automatically, but you can change it in this interface prior to import. For example, you may wish to store numerical categories as strings. The data types are Boolean, Integer, Real, String, and Timestamp.

   h. You can change the name of the table column before importing.

5. Click APPLY CHANGES to ensure that all adjustments you made are saved.

6. Click CONFIRM IMPORT.

The Confirm File Import modal window appears, where you have to click CONFIRM IMPORT again.

Confirm File Import   ✕

After confirming import, you can use this data in the following manner:

- As supplemental data to an existing dataset; in the Data interface, select the dataset and edit its Data Model to create new table joins.
- As the foundation of a new dataset; in the Data interface, click New Dataset.

This operation may take a long time. It cannot be canceled.

Are you sure you want to confirm the import?

CANCEL   CONFIRM IMPORT
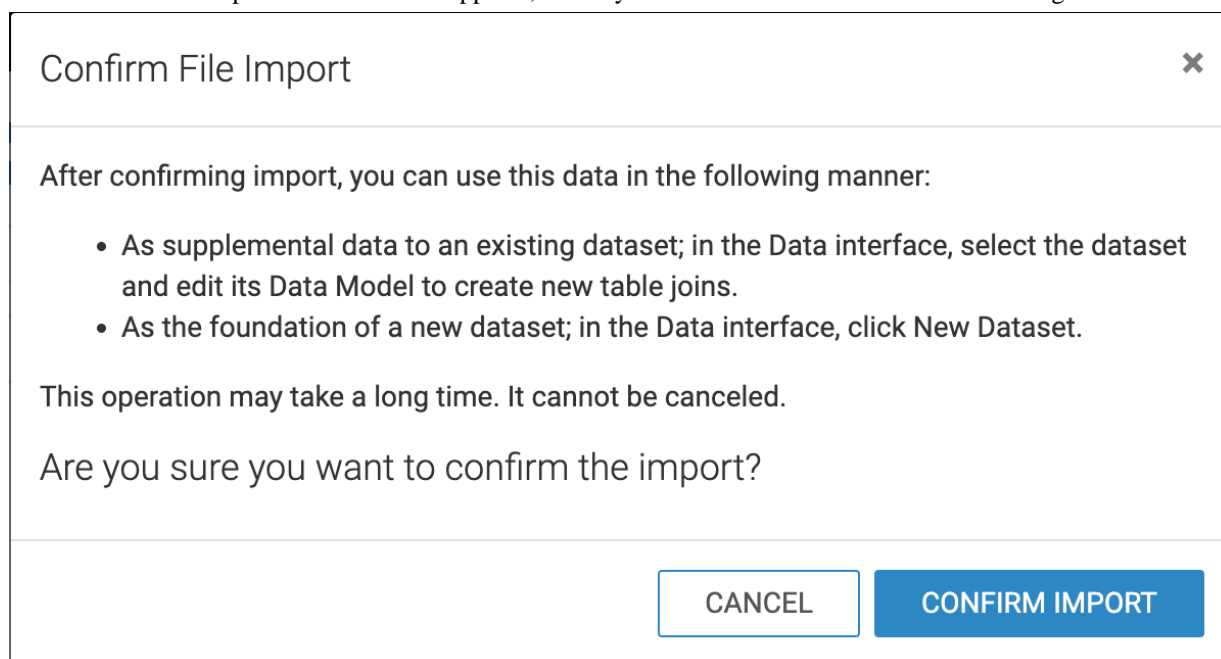
7. To verify that the data has been imported correctly into your system, check the Connection Explorer interface.

Related Information

Changing the locale setting of an imported file

Changing the locale setting of an imported file

When importing a data file, you can specify the localization for the data. This will specifically handle differences in specifying floating-point number.

Under Locale Setting, which is the default for your system, select one of the common options: United States, United Kingdom, Sweden, Norway, or Other.

Selecting Other brings adds a new text entry box. Clicking on the information icon opens a separate menu, where you can select another supported locale. In our case, we selected Russian, ru_RU.

Preparing Excel files for data import

Learn how you can import data from Microsoft Excel or other similar formats into Cloudera Data Visualization.

About this task

You must save the files in character-delimited format.

Procedure

1. In Microsoft Excel, open the file that contains the data you are importing.
2. Click Save as, then select Comma Separated Values (.csv).
3. Choose a file name, such as MyCSVdata, and click Save.