Collibra

Collibra Platform Self-Hosted

Data Catalog

Version 2023.11

Collibra

Collibra Platform Self-Hosted - Data Catalog

Release date: November 5, 2023

Revision date: February 21, 2024

You can find the most up-to-date technical documentation on our Documentation Center at

https://productresources.collibra.com/docs/collibra/latest/Content/Catalog/to_catalog.htm

Contents

Catalog submenu pages

The following table describes each of the submenu items of the Data Catalog application.

Page	Description
Data Catalog Home	The landing page when you click the Data Catalog tab. This page is designed to help you quickly and easily find Data Catalog-related assets.
Reports	All report assets.
Data Sets	All data sets shown as a set of tiles or as a table, with their name, description and, if there are any, connections to existing assets in Collibra.
Data Sources	Data sources that are used for data source registrations.
Data Dictionary	All data assets in Collibra.
Technology Assets	All technology assets in Collibra.
Metrics	Contains a variety of statistics related to how the assets of the Catalog are used.
Access Requests	The history of your access requests and their status.
Advanced Data Types	All advanced data types, which are used during a data source registration.

Data Catalog asset pages

The asset pages in Data Catalog provide information about assets. The information depends on the asset type and the asset type's assignment.

Catalog experience setting

Catalog experience is a setting that improves the user experience of the Data Catalog asset pages. The improvements include:

- Custom tabs that correspond to the page you are working on.
- A streamlined title bar showing general information.
- Quicker and easier navigation that requires less scrolling.

The Catalog experience setting is enabled by default. If required, you can disable it.

Page layout

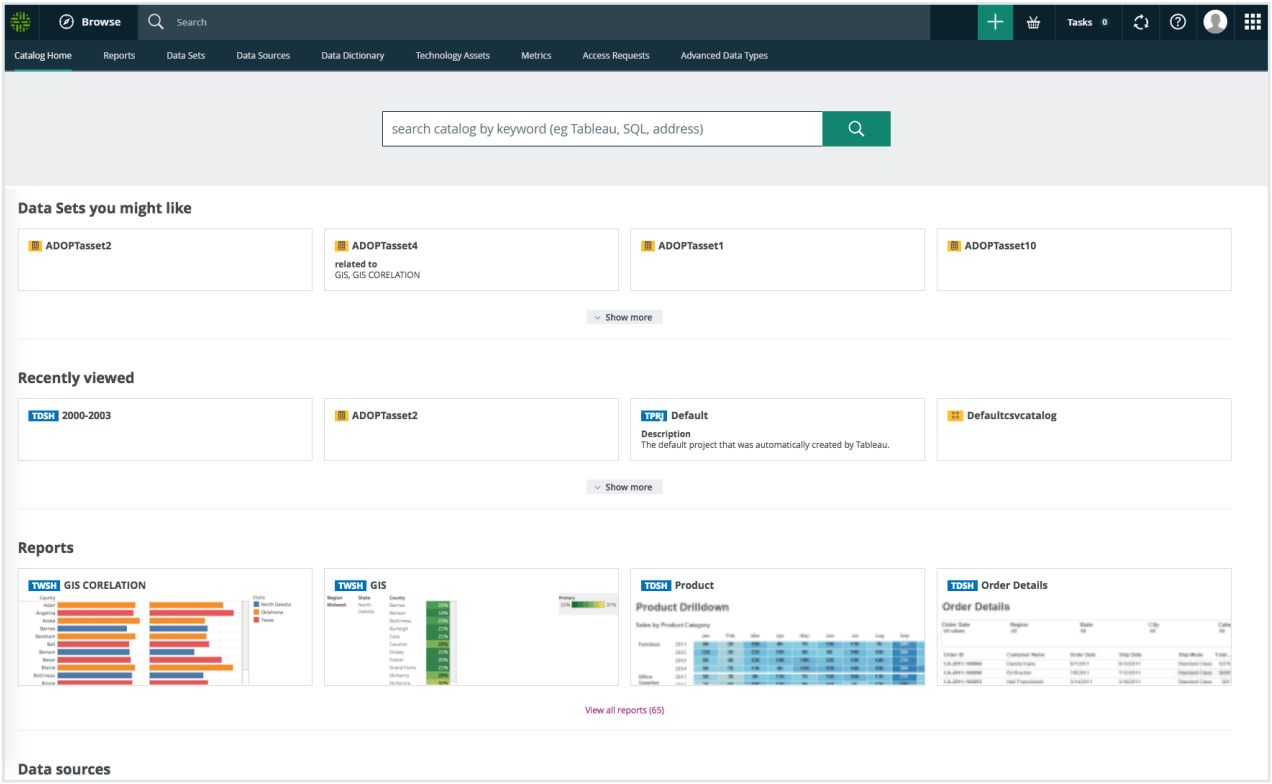
For more information on the Data Catalog asset pages, see the online version of this guide.

Data Catalog Home

The Collibra Data Catalog Home is the landing page when you click the Data Catalog tab. This page is designed to help you quickly and easily find Data Catalog-related assets.

> Note You need the Data Catalog global role or Data Catalog Author role to view Data Catalog Home.

The page is organized into five groupings, or sections, of assets and a Data Catalog-specific search field, as described in the following image and table.



Note The **Data sets you might like** section is enabled and disabled via Collibra Console. By default, it is enabled (shown) on the page. The other four sections are always shown and cannot be disabled. However, for any of the five sections, if there is no relevant data, nothing is shown on the page, including the section header.

Element name	Description
Search field	A Data Catalog-specific search field that you can use to find any asset in CollibraData Catalog, for example assets of asset types **Data Set**, **Schema**, **Table**, **Column**, **Tableau Workbook** and **Tableau View**. This search field works in the same manner as does the global search field, but it uses a default 'Data Catalog' filter.
Data Catalog Data Sets you might like	Shows up to four data sets you might be interested in, as determined by the recommender, which takes into account your data sets and the data sets of similar users. The **Show more** button enables you to view up to eight data sets on this page.

Element name	Description
Recently viewed	Shows the four most recently viewed Data Catalog-related assets. This section uses the Recent widget functionality. The **Show more** button enables you to view the eight most recently viewed assets.
Reports	Shows the four most recently created assets of asset type **Report** and its child asset types. Clicking the asset name takes you to the asset page. Clicking **View all reports** takes you to the Catalog reports page.
Data sources	Shows the four most recently created assets of asset type **Table**. Clicking the asset name takes you to the asset page. Clicking **View all data sources** takes you to the Data Sources page.
Data sets	Shows the four most recently created assets of asset type **Data Set**. Clicking the asset name takes you to the asset page. Clicking **View all data sets** takes you to the Data Sets overview page.

Recommenders in Data Catalog

The recommenders aim to suggest relevant business assets and data sets. Recommenders have to train regularly to update the recommendations. By default, this is done every night. Recommendations can be calculated on the basis of several algorithms. These algorithms also calculate an error margin for each recommendation, and eventually only the algorithm with the lowest error margin provides the recommendations.

You can edit the settings of the recommenders and matchers to optimize the recommendations.

> Note The recommender uses statistical information. Therefore, your recommendations will be empty or less useful if your company just started using Collibra Platform Self-Hosted.

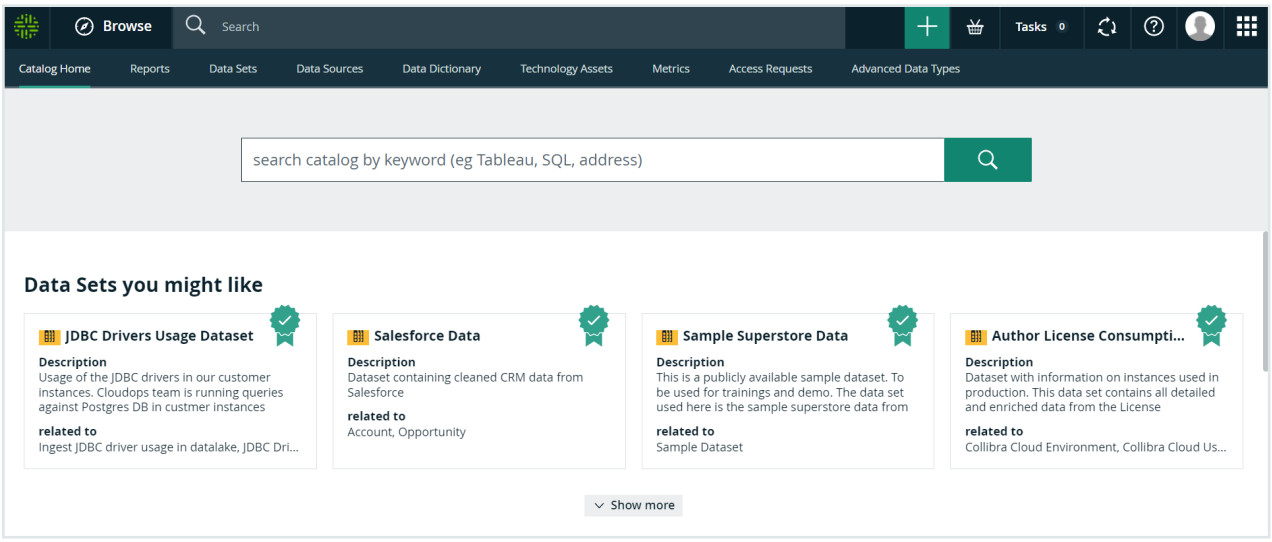
Recommendation of data sets to users

Description

The data set recommender recommends data sets to users, based on the data sets of similar users.

If you use some of the same data sets as some other users, you are probably also interested in data sets that they use but you don't. The recommendations are shown on Data Catalog Home.

Example



Strategy

The data set recommender compares the data sets used by the users to find relevant data sets. It roughly follows these steps:

1. See which data sets you are currently using.
2. Look for other users that also use your data sets.
3. See which data sets those users use, but you don't.
4. Recommend up to 9 of those data sets to you.

> Note
> If the recommender does not have enough data, for example if you just started using Collibra, it only considers 3 parameters:
>
> - Certified
> - Quality
> - Popularity (number of views of the data set asset page)

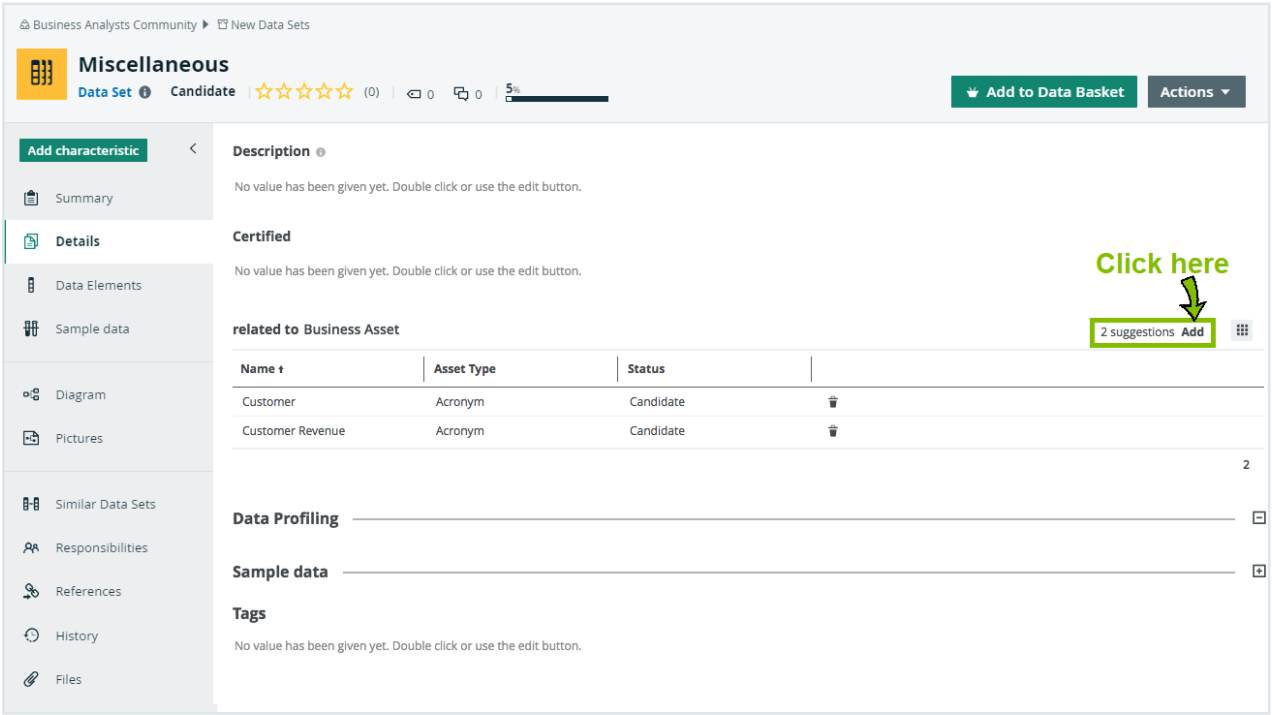
Recommendation of business assets to data sets

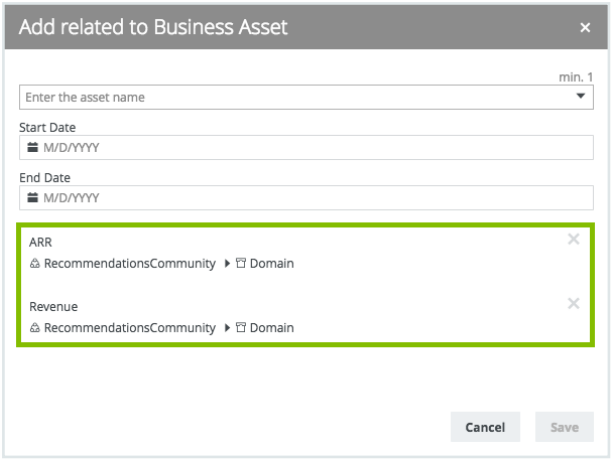
Description

The asset recommender recommends business assets to data sets, based on business assets it is related to.

If two data sets have relations to the same business assets, business assets related to only one of the two data sets may be relevant to the other data set as well.

Example





Strategy

The asset recommender uses the relation **data set related to business asset** set to find relevant assets. It roughly follows these steps:

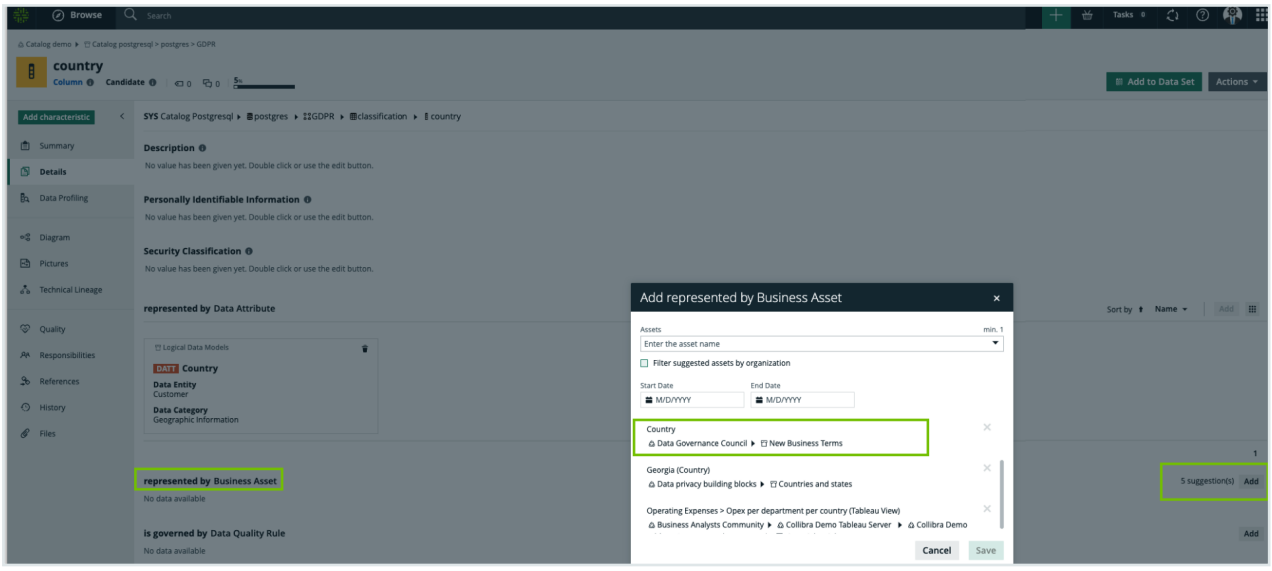
1. See which business assets are related to the current data set.
2. Look for other data sets related to those business assets.
3. See whether those data sets are also related to other business assets.
4. Recommend those business assets on the data set page and in the **Add related to** dialog box.

> Note If the recommender does not have enough data, for example if you just started using Collibra, it does not give you any recommendations.

Recommendation of business assets to column assets

Business assets are recommended to column assets based on the search engine in Collibra. The recommendations are shown in the section of **represented by business asset** relation.

Example



Recommendation of business assets to Tableau workbook assets and Tableau view assets

Business assets are recommended to Tableau workbook assets and Tableau view assets based on the search engine in Collibra. The recommendations are shown in the section of **report related to business asset** relation.

Matchers

The matchers aim to suggest assets and data sets that might be interesting for you.

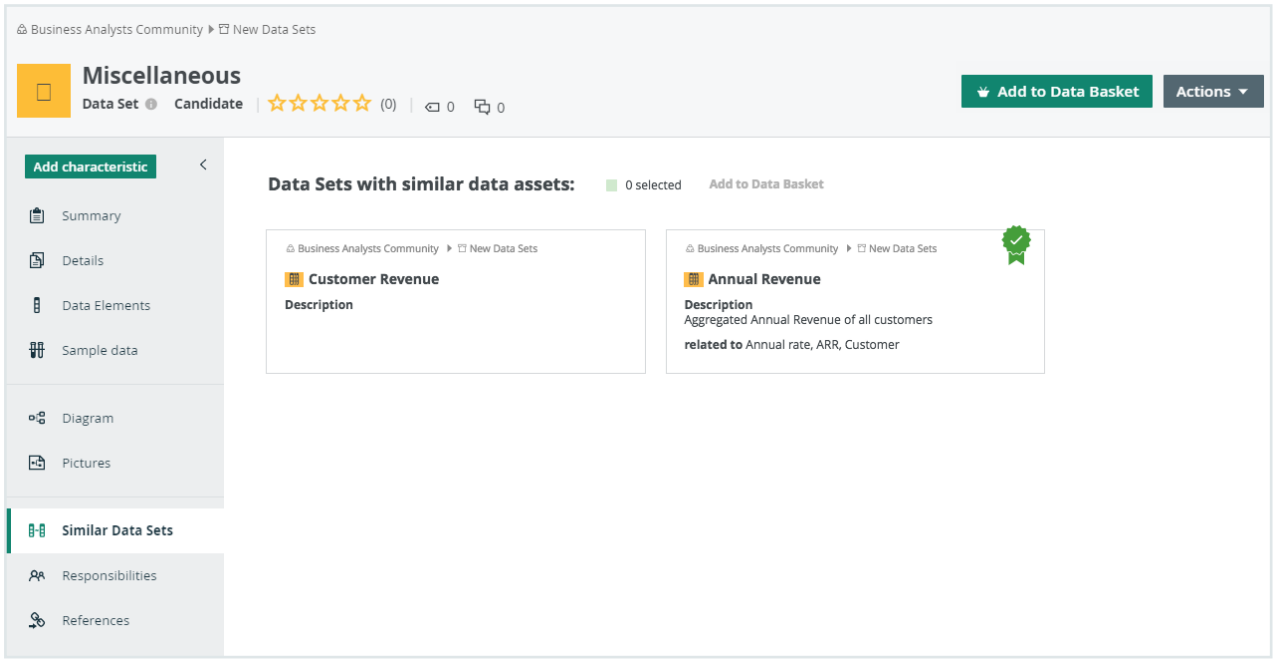
Matchers find similar data sets and schemas based on the name and the attributes.

You can edit the settings of the recommenders and matchers to optimize the recommendations.

> Note The matcher uses statistical information. Therefore, your recommendations will be empty or less useful if your company just started using Collibra Platform Self-Hosted.

Data set matcher

The data set matcher looks at the names and attributes of the column assets that a data set contains. It shows similar data sets on the data set asset page.



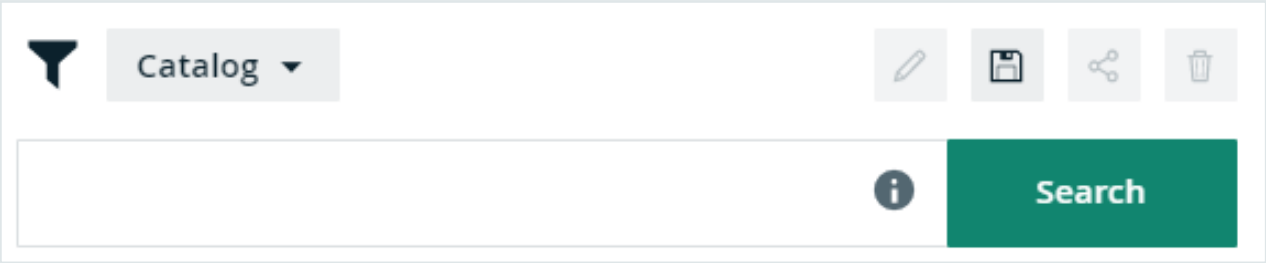
Schema matcher

The schema matcher is currently not used in Collibra.

Data Catalog Search

The Data Catalog Home page has a Data Catalog-specific search field that you can use to find assets in Data Catalog. When you launch a search from Data Catalog, the search page is the regular Collibra search page, but with the **Catalog** search filter applied.

> Note You need the Data Catalog global role or Data Catalog Author role to view the Data Catalog search page and use the Data Catalog Search.



In the search input field, you can type any text and press `Enter` or click **Search** to launch the search.

The search finds resources that contain a word that begins with your search text. For example, if you type *ca*, the search results could contain 'California' and 'Lewis Carroll', but not 'Meercat'.

You can also use wildcards and symbols to search, see Wildcards and symbols for searching.

Catalog reports

The **Reports** page is a view that shows:

- All **Report** assets.
- All packaged or manually created child asset types of **Report**, for example BI Report, Tableau View, and Looker Query.

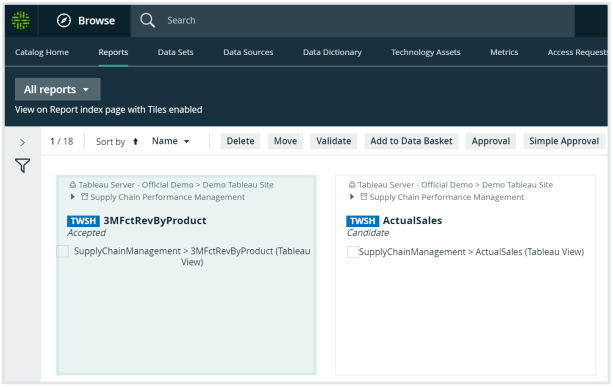
Report views

You can view the assets in table or tile display mode, and can perform all the same actions you can for any other table or set of tiles.

Reports in tile display mode

In tile display mode, you can do the following:

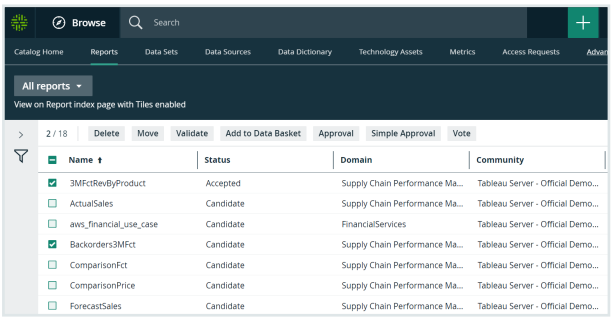
- Click an asset name to open the relevant asset page.
- Click anywhere else in the tile to select one or more assets. The list of available actions appears in the action toolbar.



Reports in table display mode

In table display mode, you can do the following:

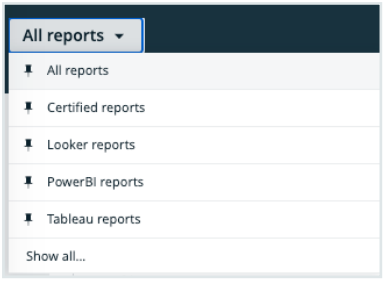
- Click an asset name to open the relevant asset page.
- Click anywhere else in the tile to select one or more assets. The list of available actions appears in the action toolbar.
- Edit cells in the table.



Filters

The default **All reports** view does not contain a filter, so it shows all Report assets. Some of the other packaged views do contain a filter. For example the **Certified reports** view only shows reports that are certified.

You can also create your own filter and, if necessary, save the filtered view as a new view. For example, you can create separate views for Report assets belonging to a specific source, for example Tableau, Looker or Power BI.



Data Catalog Data Sets

A data set is a logical, handpicked collection of data elements that can come from multiple data sources. For example, Customer Contact information. Data sets allow users to quickly know which data to use for a specific purpose and request access to it.

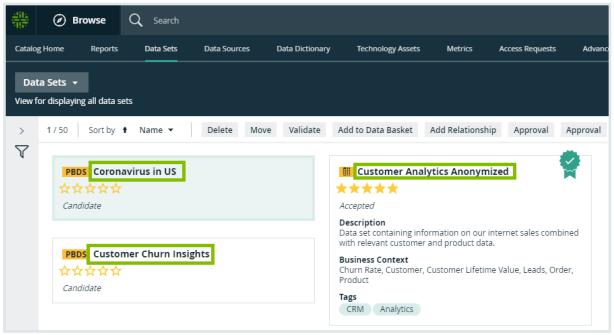
The Catalog **Data Sets** overview page displays existing data sets in a table or as tiles. The page displays the name of the data set, its description, its certification status, and, if there are any, connections to existing business assets in Collibra Platform Self-Hosted.

Data Sets overview page

The Data Sets overview page contains the data sets that are available in Collibra Platform Self-Hosted. You can view the data sets in table display mode or tile display mode.

Tile display mode

- Click a data set title to open its details.



- Click anywhere in the tile except for the title to select the data set. The list of actions that you can perform is displayed.

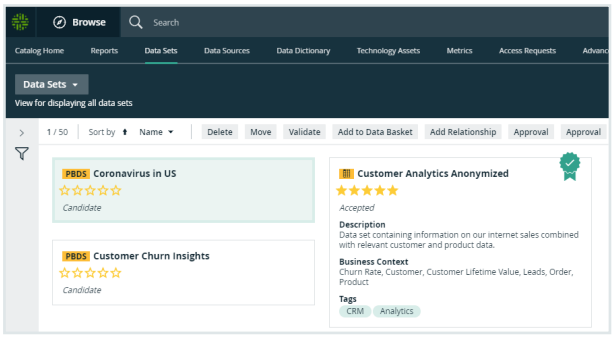
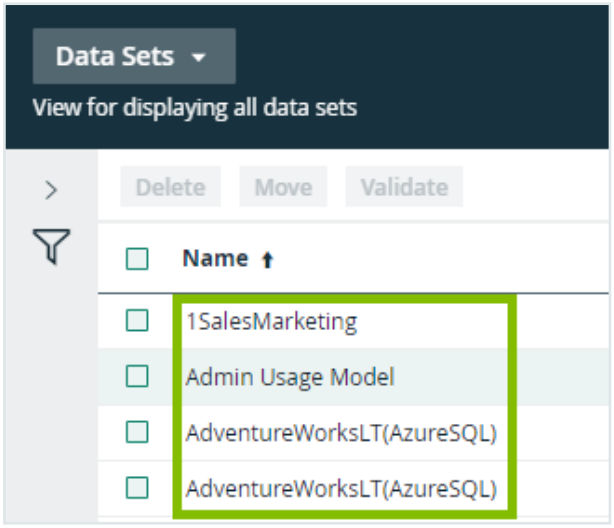
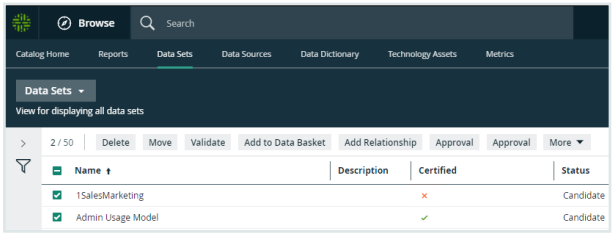


Table display mode

- Click a name of the data set to open its details.



- Select one or more data sets. The list of actions that you can perform is displayed.



> Note The Sample Data tab shows the first 100 columns of data. If you have more than 100 columns, they are not shown.

Data Set asset page

The **Data Set** asset page is basically the same as any asset page in Collibra Platform Self-Hosted with the following differences:

- The Data Set asset page has a special attribute, namely **Certified**. That attribute indicates whether a data set is certified or not. There are no restrictions for certifying a data set, except the ones your organization chooses. You decide when a data set can or has to be certified. For more information about how to do this, see Certify a data set.
- It contains suggestions for related Business Assets, based on the asset recommender.
- It contains a **Data Profiling** and **Sample data** section which contains respectively a data profile and sample data, if available.

You can perform the following actions on this page:

- Create a view
- Filter data
- Sort Catalog submenu pages
- Request access to data sets

  > Important
  > This workflow accepts by default only data sets that contain Column assets as data elements.

- Delete data sets

Creating data sets

In this section you can learn how to create a data set and how to add data to it.

Create a data set

You create data sets to add data to them.

Steps

1. On the main toolbar, click ＋.
   » The **Create** dialog box appears.
2. In the **Create** dialog box, click the **Asset** tab.
3. Click **Data Set**.
4. In the **Domain** field, select the domain to which you want to add one or more data sets.
5. In the **Name** field, type the name of the data set, press `Enter` to add other data set names.
6. Click **Create**.

Add data to a data set from an asset page

When you come across an asset that you want to add to a data set, you can add that asset from that asset page.

Prerequisites

- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a resource role with the Attribute > Add resource permission.

Steps

1. Navigate to an asset page of a schema, table or column asset.
2. In the upper-right corner, click **Add to Data Set**.
3. Enter the required information in the **Add data to data set** dialog box.
   - Existing data set:
     a. Select the data set.
     b. Click **Add to data set**.
   - New data set:
     a. Type a name in the **Data set name** field.
     b. Type a description in the **Data set description** field.
     c. Click **Create & Add data**.

Add data to a data set from the Data Sources or Data Dictionary page

You can add data to a data set from the Data Sources or Data Dictionary page.

Prerequisites

- For Data Dictionary: You have a global role with the Data Dictionary global permission, for example Data Dictionary.
- You have a resource role with the Attribute > Add resource permission.

Steps

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.
2. In the submenu, click **Data Sources** or **Data Dictionary**.
   If necessary, filter the list of data assets.
3. Select the check boxes of the data assets you want to add to a specific data set.

   Note
   - Some data assets are nested. If you select the top one, all its children are added as well.
   - Keep in mind that you can only add schemas, tables and columns.

4. Above the table, click **Add to Data Set**.
5. Enter the required information in the **Add data to data set** dialog box.
6. Click **Add to data set**.
   » A notification in the upper-right corner lets you know how many assets you have added to the data set.

Certify a data set

You can approve, endorse or guarantee the contents of a data set.

Steps

1. Navigate to the asset page of a data set that you want to certify.
2. Find the **Certified** characteristic and double-click the line of text below it.
3. Click in the field that is displayed.
4. Click **True**.
5. Click **Save**.

> Tip You can design a workflow to take care of the certification of a data set.

Delete data sets

If you no longer need a certain data set, you can delete it from the repository.

Steps

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.
2. In the submenu, click **Data Sets**
3. Search for the data sets that you want to delete.
   You can use the Filter pane or sort your data sets.
4. In table mode, select the check boxes of the data sets that you want to delete.
   In tile mode, hold the SHIFT key to select multiple data sets.
5. Click **Delete**.
6. Click **Yes** to confirm.

Data Sources page

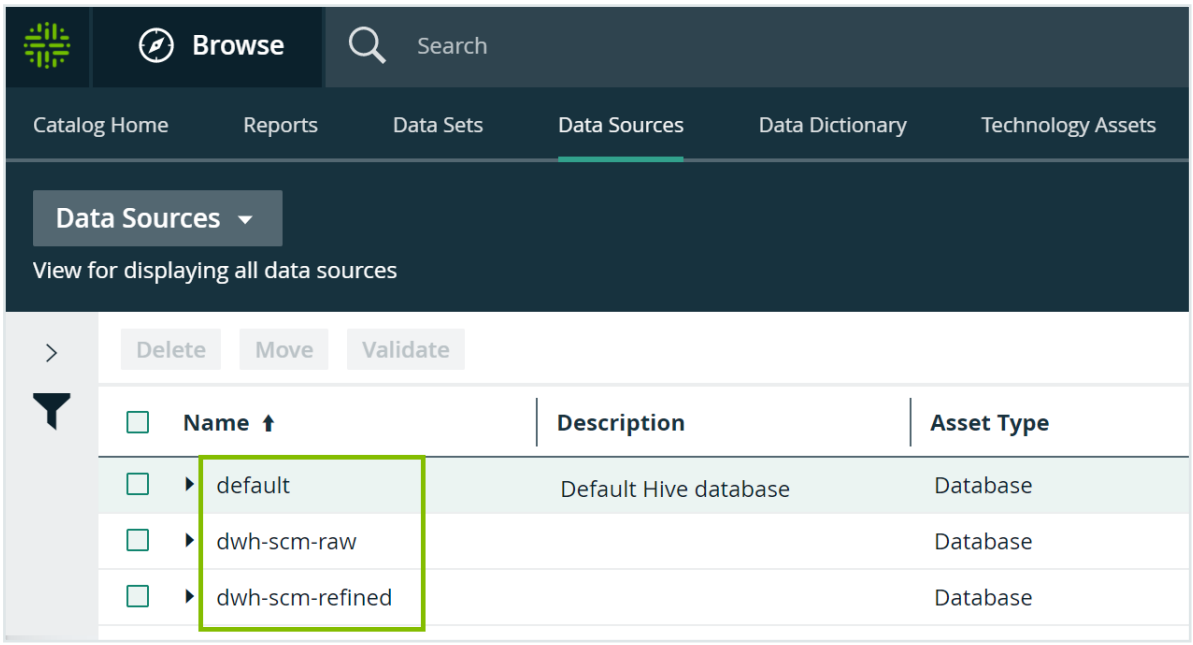
The **Data Sources** page is page that shows the asset types that are created by Database and S3 registrations. It's a combination of Data and Technology asset types.

You can view the assets in table display mode or tile display mode.

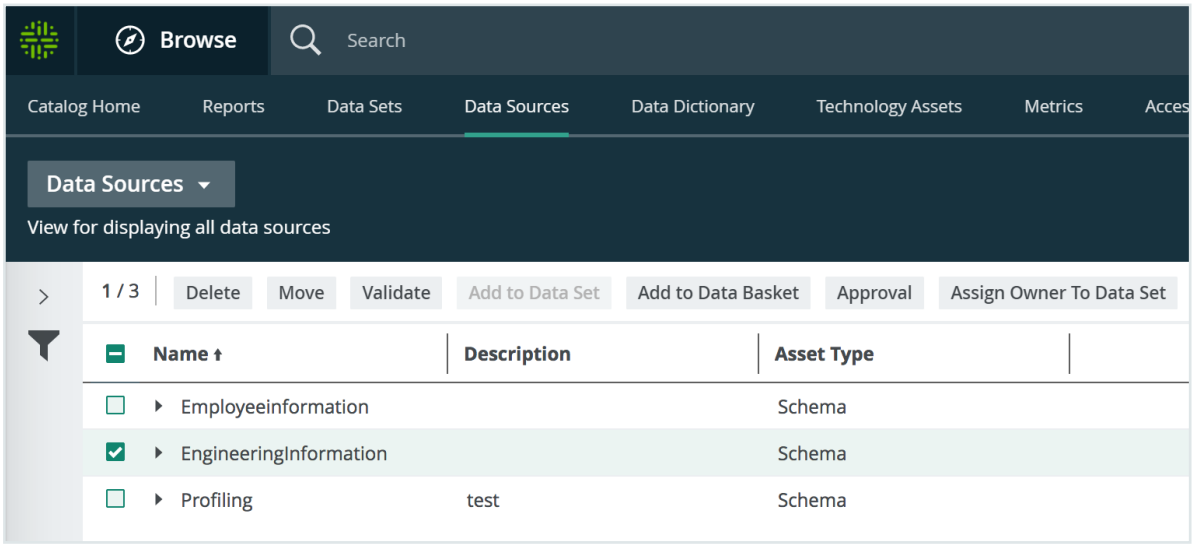
Data sources in table display mode

With hierarchies enabled, you can expand the assets to consult the structure of the data sources. If needed you can also show other asset types in the lower levels of the hierarchy.

- Click an asset name to open the relevant asset page.

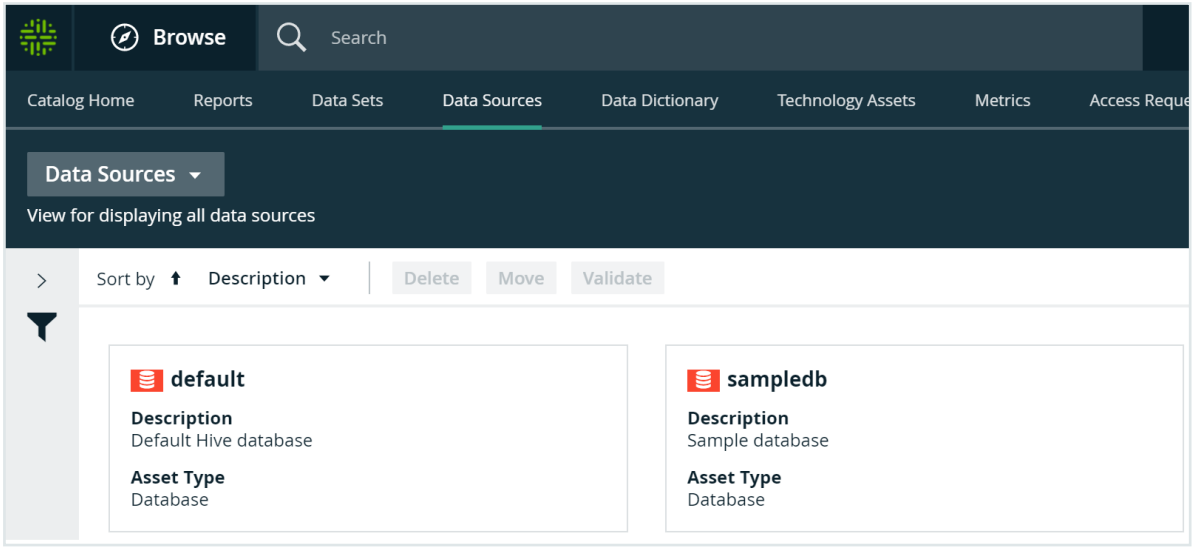


- Select one or more assets. The list of actions that you can perform is then displayed.

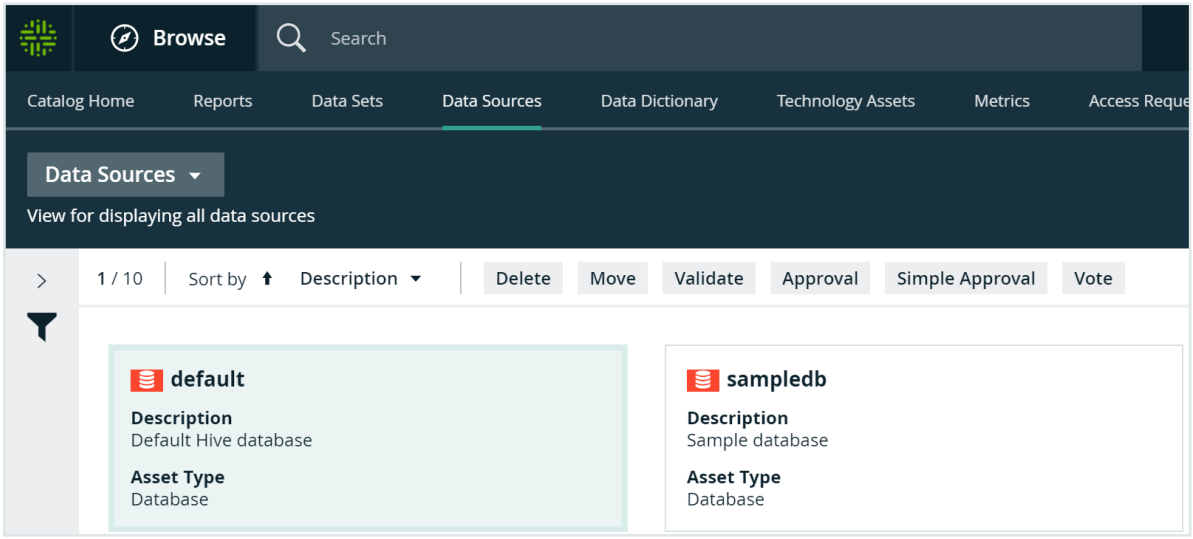


Data sources in tile display mode

- Click an asset name to open the relevant asset page.



- Click anywhere else in the tile to select the asset. The list of actions that you can perform is then displayed.



Data Dictionary page

The **Data Dictionary** page is a page that shows the assets of asset type Data Asset and its children asset types in Collibra Platform Self-Hosted.

You can view the assets in table display mode or tile display mode.

On this page, you can perform the following actions:

- Create a view
- Filter assets
- Sort assets by name, description and asset type
- Delete assets
- Move assets
- Add assets to a data set
- Start an asset workflow from an asset table, for assets

Technology Assets page

The **Technology Assets** page is a view that shows all assets of every technology asset type in Collibra Platform Self-Hosted.

You can view the assets in table display mode or tile display mode.

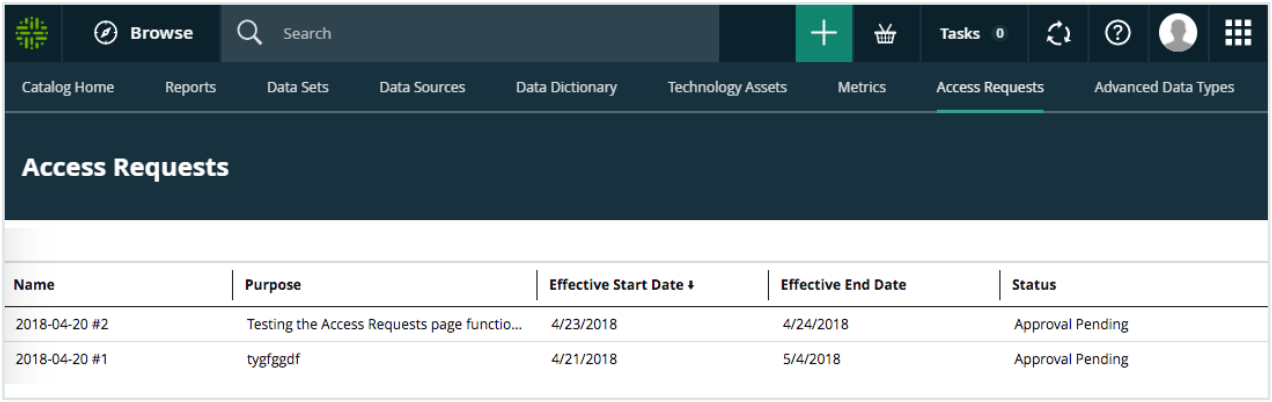
On this page, you can perform the following actions:

- Create views.
- Filter assets.
- Sort assets by name, description and asset type.
- Delete assets.
- Move assets to another domain.

Access Requests page

If you have requested access to one or more assets, the Access Requests page allows you to view the status of your requests. When you request access:

- The Request Assets Access workflow starts.
- A Data Usage asset is created in the Data Usages domain in your community.



The names of your requests are automatically generated with the date of your request.

Click the request name to open the asset page which shows all the information about your request.

If you have requested access to multiple assets, you can sort on any of the columns on the Access Requests page, to find a specific access request.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements.

Advanced data types

When you profile data when registering a data source, Collibra Platform Self-Hosted can detect some basic data types, such as numbers and text. Besides these basic data types, you can create your own advanced data types.

> Note Advanced data types are not taken into account when profiling via Edge. See About profiling and classification via Edge.

In this section, you learn how to work with advanced data types.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers

using GovCloud or on-prem environments. For more information, go to Announcements.

Data type detection

When you run a data profiling when registering a data source, Collibra Platform Self-Hosted tries to detect the data type of each column.

1. Collibra tries to match the fields of each column with every data type.
2. Collibra remembers the matches for each field, also if a field has multiple matches.
3. Collibra calculates the matching percentage of how many fields of the column match the same data type.
4. Collibra verifies the matching percentage against the data type detection threshold.

   Tip You can define the data type detection threshold in Collibra Console, see the Collibra Installation and Configuration Guide.

5. Collibra assigns the data type with the highest matching percentage to the source column, provided that the matching percentage exceeds the threshold.

Out of the box, there are several base data types such as integer, text and boolean. With each data profiling, these base data types are evaluated. If your data source contains special data types such as social security numbers or international bank account numbers, you can define them as advanced data types. In the data source registration wizard, you can then choose to also evaluate the data on these advanced data types.

Keep in mind that detecting advanced data types significantly increases the data profiling job execution time.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements.

Advanced data type management prerequisites

To manage advanced data types, you need the following prerequisites:

- Catalog role
- Advanced Data Type global permission

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements.

Create an advanced data type

If the basic data types, such as numbers and text, are not specific enough, you can create your own advanced data types.

Prerequisites

- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a global role with the Advanced Data Type > Add global permission.

Steps

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.
2. In the submenu, click **Advanced Data Types**.
3. Above the table, to the right, click **Add Advanced Data Type**.
4. In the **Add Advanced Data Type** dialog box, fill in the new data type properties.

Option	Description
Name	The name of the advanced data type. The name has to be unique, including the basic data types.

Option	Description
Description	The description of the advanced data type.

<table>
<tr><th>Option</th><th>Description</th></tr>
<tr><td>Base data type</td><td>The data type used as basis for the advanced data type:
Text
Geographical
True/False
Date
Time
Date and Time
Whole Number
Decimal Number
Array
N/A
Examples
<table>
<tr><th>Base data type</th><th>Field name</th><th>Patterns</th></tr>
<tr><td>Text</td><td>Email address</td><td>[a-z0-9]+[_a-z0-9\.-]*[a-z0-9]+@[a-z0-9-]+(\.[a-z0-9-]+)*(\.[a-z]{2,4})</td></tr>
<tr><td>Text</td><td>IP address</td><td>\b(?:(?:2(?:[0-4][0-9]|5[0-5])|[0-1]?[0-9]?[0-9])\.){3}(?:(?:2([0-4][0-9]|5[0-5])|[0-1]?[0-9]?[0-9]))\b</td></tr>
<tr><td>Date</td><td>Custom Date</td><td>yyyy-MM-dd</td></tr>
<tr><td>Time</td><td>Custom Time</td><td>HH mm</td></tr>
<tr><td>Date and Time</td><td>Custom Date and Time</td><td>MM/dd/yyyy HH:mm:ss</td></tr>
<tr><td>True/False</td><td>Boolean (French)</td><td>true: vrai, v
false: faux, f</td></tr>
</table></td></tr>
</table>

<table>
<tr><th>Option</th><th>Description</th></tr>
<tr><td>Advanced data type (variable field name)</td><td>The field name depends on the selected base data type.
<table>
<tr><th>Base data type</th><th>Field name</th><th>Description</th></tr>
<tr><td>Text</td><td>Regular expressions</td><td>List of regular expressions.
For more information about regular expressions, see regular-expressions.info.</td></tr>
<tr><td>Geographical</td><td>Regular expressions</td><td>List of regular expressions.
For more information about regular expressions, see regular-expressions.info.</td></tr>
<tr><td>Date</td><td>Date pattern</td><td>List of date patterns using the DateTimeFormatter format. See the official Java documentation.</td></tr>
<tr><td>Time</td><td>Time pattern</td><td>List of time patterns using the DateTimeFormatter format. See the official Java documentation.</td></tr>
<tr><td>Date and Time</td><td>Date and Time pattern</td><td>List of date and time patterns using the DateTimeFormatter format. See the official Java documentation.</td></tr>
<tr><td>Whole Number</td><td>Numeric format</td><td>Locale for the format of whole numbers.</td></tr>
<tr><td>Decimal Number</td><td>Numeric format</td><td>Locale for the format of decimal numbers.</td></tr>
<tr><td>True/False</td><td>◦ True values
◦ False values</td><td>◦ List of values that are accepted as True value.
◦ List of values that are accepted as False value.</td></tr>
<tr><td>Array or N/A</td><td></td><td>Not applicable for advanced data type detection.</td></tr>
</table></td></tr>
</table>

5. Click **Save**.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements.

Edit an advanced data type

If an existing advanced data type is incorrect, you can edit it.

Prerequisites

- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a global role with the Advanced Data Type > Update global permission.

Steps

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.
2. In the submenu, click **Advanced Data Types**.
3. In the row of the data type that you want to edit, click .
   The **Edit Advanced Data Type** dialog box appears.
4. Enter the required information.

Option	Description
Name	The name of the advanced data type. The name has to be unique, including the basic data types.
Description	The description of the advanced data type.

<table>
<tr><th>Option</th><th>Description</th></tr>
<tr><td>Base data type</td><td>The data type used as basis for the advanced data type:
Text
Geographical
True/False
Date
Time
Date and Time
Whole Number
Decimal Number
Array
N/A
Examples
<table>
<tr><th>Base data type</th><th>Field name</th><th>Patterns</th></tr>
<tr><td>Text</td><td>Email address</td><td>[a-z0-9]+[_a-z0-9\.-]*[a-z0-9]+@[a-z0-9-]+(\.[a-z0-9-]+)*(\.[a-z]{2,4})</td></tr>
<tr><td>Text</td><td>IP address</td><td>\b(?:(?:2(?:[0-4][0-9]|5[0-5])|[0-1]?[0-9]?[0-9])\.){3}(?:(?:2([0-4][0-9]|5[0-5])|[0-1]?[0-9]?[0-9]))\b</td></tr>
<tr><td>Date</td><td>Custom Date</td><td>yyyy-MM-dd</td></tr>
<tr><td>Time</td><td>Custom Time</td><td>HH mm</td></tr>
<tr><td>Date and Time</td><td>Custom Date and Time</td><td>MM/dd/yyyy HH:mm:ss</td></tr>
<tr><td>True/False</td><td>Boolean (French)</td><td>true: vrai, v
false: faux, f</td></tr>
</table></td></tr>
</table>

<table>
<tr><th>Option</th><th>Description</th></tr>
<tr><td>Advanced data type (variable field name)</td><td>The field name depends on the selected base data type.
<table>
<tr><th>Base data type</th><th>Field name</th><th>Description</th></tr>
<tr><td>Text</td><td>Regular expressions</td><td>List of regular expressions.
For more information about regular expressions, see regular-expressions.info.</td></tr>
<tr><td>Geographical</td><td>Regular expressions</td><td>List of regular expressions.
For more information about regular expressions, see regular-expressions.info.</td></tr>
<tr><td>Date</td><td>Date pattern</td><td>List of date patterns using the DateTimeFormatter format. See the official Java documentation.</td></tr>
<tr><td>Time</td><td>Time pattern</td><td>List of time patterns using the DateTimeFormatter format. See the official Java documentation.</td></tr>
<tr><td>Date and Time</td><td>Date and Time pattern</td><td>List of date and time patterns using the DateTimeFormatter format. See the official Java documentation.</td></tr>
<tr><td>Whole Number</td><td>Numeric format</td><td>Locale for the format of whole numbers.</td></tr>
<tr><td>Decimal Number</td><td>Numeric format</td><td>Locale for the format of decimal numbers.</td></tr>
<tr><td>True/False</td><td>◦ True values
◦ False values</td><td>◦ List of values that are accepted as True value.
◦ List of values that are accepted as False value.</td></tr>
<tr><td>Array or N/A</td><td></td><td>Not applicable for advanced data type detection.</td></tr>
</table>
</td></tr>
</table>

You cannot change the base data type.

5. Click **Save**.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements.

Delete one or more advanced data types

If you no longer use an advanced data type, you can delete it.

Prerequisites

- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a global role with the Advanced Data Type > Remove global permission.

Steps

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.
2. In the submenu, click **Advanced Data Types**.
3.

Single advanced data type	a. In the row of the data type that you want to delete, click . b. In the **Delete advanced data type** dialog box, click **Delete advanced data type**.
Multiple advanced data types	a. Select the check boxes in front of the advanced data types that you want to delete. b. On the action toolbar, click **Delete**. > Tip You can select all the visible assets at once by clicking the check box next to the **Name** column header. c. In the **Delete (x) advanced data type(s)** dialog box, click **Delete (x) advanced data type(s)**.

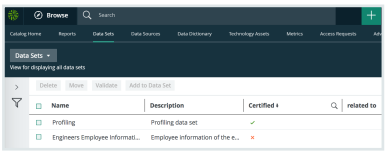
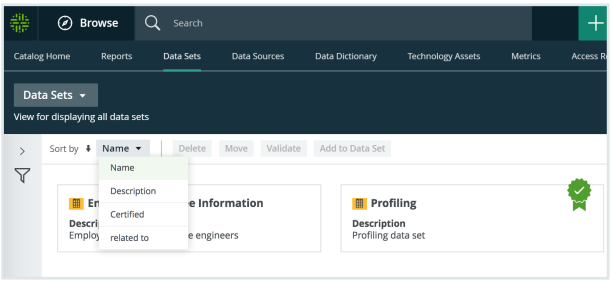
The data type attributes that contain the deleted advanced data type are reset to the base data type that was used for the advanced data type.

Sort Catalog submenu pages

You can reorder the data on Catalog pages, such as Reports, Data Sets, Data Sources and so on.

Steps

1. On the main menu, click ⋮⋮⋮, and then click **Catalog**.
   » The Catalog Home opens.
2. Click any of the items in the submenu, for example **Data Sets**.
3. Sort your data:

Table display mode	Tile display mode (if available)
Click any column header to sort the data based on that column. Click again to toggle between ascending and descending order. 	Click the **Sort by** arrow to sort ascending or descending, and click the drop-down list to select on which field you want to sort. 

Registering a data source

By registering a data source, you connect a data source to Collibra and make metadata of the data source available in Collibra. Based on that, you can create data sets that can then be used for reporting and analysis.

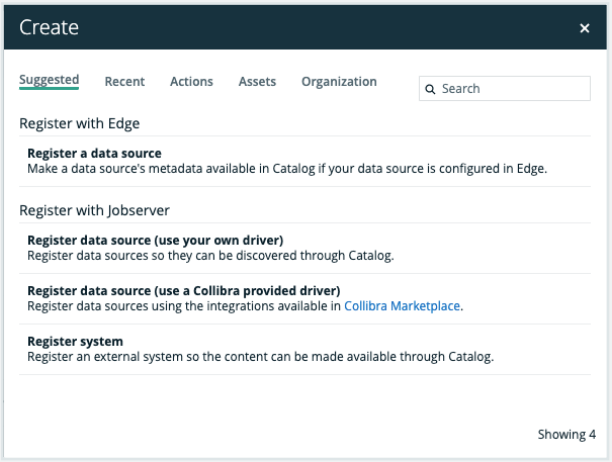
> Note If you are using a Collibra Data Intelligence Cloud environment with an on-premises Jobserver, the Jobserver version must be compatible with the cloud version. You can find the version of your Collibra Data Intelligence Cloud environment at the bottom of the sign-in window, for example 2023.11.0.

About registering a data source

By registering a data source, you connect a data source to Collibra and make metadata of the data source available in Collibra.

You can register a data source via Jobserver or via Edge.

Note
When you enable registering a data source via Edge, you can choose to register a data source using Edge or using Jobserver.



Differences between registering a data source via Jobserver or via Edge

The following table shows the differences between registering a data source via Jobserver or via Edge.

Part of process	Register a data source via Jobserver	Register a data source via Edge
Permissions	The required permissions to register a data source via Jobserver or via Edge are the same except for the following permission: You need a resource role with the following resource permissions on the **Schema** community: • Asset > add • Attribute > add • Domain > add • Attachment > add	The required permissions to register a data source via Edge or via Jobserver are the same except for the following permission: You need a global role with the View Edge connections and capabilities global permission.

Part of process	Register a data source via Jobserver	Register a data source via Edge
Registering a data source	When you register a data source via Jobserver, you have to enter all database connection properties in the **Register data source** dialog box.	Before you register a data source via Edge, you have to enable data source registration via Edge. You also need a JDBC connection to your data source and Edge capabilities with a JDBC Catalog JDBC ingestion capability template. When you register a data source in Data Catalog, you can then select which database you want to add to the JDBC connection.
Refreshing or synchronizing	After registering a data source, a Schema asset is created. On the Configuration tab page of the Schema asset page, you can refresh a data source.	After registering a data source, a Database asset is created. The Database asset has a relation of the type "Technology asset groups / is grouped by Technology asset" to the System asset that was selected when registering the data source. On the **Configuration** tab page of the Database asset page, you can synchronize one or many schemas.

Part of process	Register a data source via Jobserver	Register a data source via Edge
Profiling options	At the end of the registration process, you can select profiling options to create data profilingdata profiling and sample datasample data. The profiling data is automatically created after the refresh process. • Data profiling creates a summary of a data source that is registered with Data Catalog and determines the data type of columns in the data source. The summary mainly contains statistics and graphics to give the user an idea what the registered data is about. • Sample data is a set of randomly collected data from a data source. The purpose of showing sample data is to provide examples of the data so you know what to expect when you use the asset.	To be able to profile the data, you have to enable profiling and classification via Edge. After you have registered the data source, you can then select profiling options to create profiling data and data classes on a Database asset page. The metadata is profiled and classified automatically after synchronizing a schema or manually. Also to show sample data for a data source, extra setup is needed.

Difference between registering a data source and importing data

When you register a data source, Data Catalog reads and processes metadata of data sources that are not governed in Collibra Platform Self-Hosted. Collibra will create assets of the relevant types, such as Database, Table and Column.

> Example You register a data source that contains your financial data in a SAP HANA database. Afterwards, you can use the Collibra to manage the data, for example manage access control through data sets and use traceability to see your data lineage.

When you import data, you create or edit assets or complex relations, with their characteristics, from a view. Collibra will create assets of the type specified in the imported XLSX or CSV file.

> Example You import an XLSX file containing the most common business terms of your company. You can use Collibra to approve the terms and link them to more technical assets.

Naming convention

When you register a data source, Collibra follows a strict naming convention for the names of the new assets. Each asset has a display name and full name. You can freely edit the display name. However, you should never edit the full name, because Data Catalog may need it to refresh data sources. Editing the full name may cause unexpected results and break the synchronization process.

> Warning Editing the full name of the Database and Schema assets may lead to errors during the refresh process.

For information on Edge naming conventions, go to naming conventions.

Supported data sources for data source registration

Collibra Platform Self-Hosted supports several databases to register as a data source.

Configuration assets

When you register a database or system as a data source, you enter connection properties and other options. To store the configuration and connection properties, Data Catalog creates a special kind of asset, often called the configuration asset. Some of these assets show parts of the configuration on a dedicated Configuration tab page.

This list contains the most widely used configuration assets:

- Schema assets, if you register a data source using Jobserver
- Database assets, if you register a data source using Edge.
- S3 File System assets
- Tableau Server assets

Working with configuration assets

Even though you can import or export configuration assets with the import functionality or create them via the global create button, they would not contain any configuration. This means that, if you create a configuration asset in that way, you must also create the configuration and add it to the configuration asset. However, this is not possible for all configuration assets. For example, you cannot configure an S3 File System asset after creation. The only way to configure an S3 File System asset is by connecting to Amazon S3 and synchronizing its content. We highly recommend that you do not create configuration assets by importing them or via the global create button. Instead, use the appropriate procedure, such as registering a data source or registering a system.

> Warning If you delete a configuration asset, you also delete its configuration. Register your data source or system again to create a new configuration asset or contact support for more information.

Quartz Cron syntax

Cron is a software utility that specifies commands to run on a given schedule. This schedule is defined by a Cron pattern, which has a specific syntax that will be described in this section.

For example, you can refresh the schema of a data source or synchronize Tableau or Amazon S3 metadata outside office hours to reduce the impact of these actions on the performance of your environment.

> Note By default, you use Spring Cron expressions to schedule Collibra Console back-ups.

> Warning If you create an invalid Cron pattern, Collibra Platform Self-Hosted stops responding.

The Cron pattern consists of six or seven space-separated fields:

`<second> <minute> <hour> <day of the month> <month> <day of the week> <year>`

Position	Field	Mandatory	Allowed values	Allowed special characters	Examples
1	second	Yes	0-59	, - * /	• *10*: at the 10th second. • **/10*: every 10 seconds.
2	minute	Yes	0-59	, - * /	• *30*: at the 30th minute. • **/15*: every 15 minutes. • *5/10*: every 10 minutes starting at the 5th minute after the hour
3	hour	Yes	0-23	, - * /	• *10*: at 10 o'clock. • *8-10*: at 8,9 and 10 AM. • *6,18*: at 6 AM and at 6 PM.

Position	Field	Mandatory	Allowed values	Allowed special characters	Examples
4	day of the month	Yes	1-31	, - * ? / L W	• *3*: on the 3rd day of the month. • *1-4*: every first four days of the month. • *1,15*: the first day of the month and the 15th day of the month. • *L*: on the last day of the month. • *L-3*: on the third-to-last day of the month. • *15W*: on the nearest weekday to the 15th of the month. If the 15th is a Saturday, then the trigger will be on the 14th, if the 15th is a Sunday, then the trigger will be on the 16th. Note If the 1st day of the month is a Saturday, then *1W* corresponds to the 3rd day of the month, since the month is specified in the 5th value of the Cron expression. *LW*: on the last weekday of the month.
5	month	Yes	1-12 or JAN-DEC	, - * /	• *12*: in December. • *1-3*: every first three months of the year. • *JUL,AUG*: every July and August. Tip The names of the months are not case-sensitive.

Position	Field	Mandatory	Allowed values	Allowed special characters	Examples
6	day of the week	Yes	1-7 or SUN-SAT	, - * ? / L #	• *TUE*: every Tuesday. • *2-6*: every weekday, Monday to Friday. • *MON,WED,FRI*: every Monday, Wednesday and Friday. • *L*: on Saturday, the 7th day of the week. • *2L*: at the last Monday of the month. • *6#3*: on the 3rd Friday of the month. Tip The names of the days are not case-sensitive.
7	year	No	empty, 1970-2099	, - * /	• *<empty>*: if your schedule doesn't require a year, you can leave this value empty. • *2021*: in 2021. • *2021-2025*: in the years 2021, 2022, 2023, 2024 and 2025. • *2021,2022,2025*: in the years 2021, 2022 and 2025.

Special characters

Character	Description
*	Used to select all values within a field. Example * in the minute field corresponds with every minute.

Character	Description
?	Used to specify something in one of the two fields in which the character is allowed, but not the other, mainly used for days of the week. Example If you want your trigger to fire on a particular day of the month, for example the 10th, but don't care what day of the week that happens to be, you could put "10" in the day-of-month field, and "?" in the day of the week field.
-	Used to specify ranges. Example `10-12` in the hour field means “the hours 10, 11 and 12”.
,	Used to specify additional values. Example `MON,WED,FRI` in the day-of-week field means “the days Monday, Wednesday, and Friday”.
/	Used to specify increments. Example `0/15` in the seconds field means “the seconds 0, 15, 30, and 45”. And `5/15` in the seconds field means “the seconds 5, 20, 35, and 50". You can also leave out the number before /, which is equivalent to having 0 before / . `1/3` in the day-of-month field means “fire every 3 days starting on the first day of the month”.
L	Has different meaning in each of the two fields in which it is allowed. The value `L` in the **day-of-month field** means “the last day of the month” - day 31 for January, day 28 for February on non-leap years. You can also specify an offset from the last day of the month, such as “L-3” which would mean the third-to-last day of the calendar month. If you use `L` in the **day-of-week field** by itself, it means “7” or “SAT”. But if used in the day-of-week field after another value, it means “the last xxx day of the month” - for example “6L” means “the last Friday of the month”. When using the `L` option, it is important not to specify lists, or ranges of values, because you may get unexpected results.

Character	Description
W	Used to specify the weekday (Monday-Friday) nearest the given day. Example `15W` in the value for the day-of-month field, means the nearest weekday to the 15th of the month: • If the 15th is a Saturday, the trigger will fire on Friday the 14th. • If the 15th is a Sunday, the trigger will fire on Monday the 16th. • If the 15th is a Tuesday, then it will fire on Tuesday the 15th. However if you specify `1W` as the value for day-of-month, and the 1st is a Saturday, the trigger will fire on Monday the 3rd, as it will not 'jump' over the boundary of a month's days. The 'W' character can only be specified when the value in the day-of-month field specifies a single day, not a range or list of days. Tip The 'L' and 'W' characters can also be combined in the day-of-month field to yield 'LW', which translates to *"last weekday of the month"*.
#	Used to specify "the nth" XXX day of the month. Example `6#3` in the day-of-week field means "the third Friday of the month" (day 6 = Friday and "#3" = the 3rd one in the month). Other examples: `2#1` is the first Monday of the month and `4#5` is the fifth Wednesday of the month. Note that if you specify `#5` and there is not 5 of the given day-of-week in the month, then no firing will occur that month.

Example

- `0 0 * ? * * *` = the top of every hour of every day.
- `*/10 * * * * ?` = every ten seconds.
- `0 0 8-10 * * ? 2020` = 8, 9 and 10 o'clock of every day during the year 2020.
- `0 0 6,19 ? * *` = 6:00 AM and 7:00 PM every day.
- `0 0/30 8-10 ? * *` = 8:00, 8:30, 9:00, 9:30, 10:00 and 10:30 every day.
- `0 0 9-17 * * MON-FRI` = on the hour nine-to-five weekdays.
- `0 0 0 25 12 ?` = every Christmas Day at midnight, no matter what day of the week it is.
- `0 15 10 ? * 6L 2022-2025` = 10:15 AM on every Friday of every month during the years 2022, 2023, 2024 and 2025.
- `0 30 11 ? * 6#2` = 11:30 AM on the second Friday of every month.

Warning Quartz Cron only supports a value in either the 4th or the 6th position, but not in both. At the same time, both positions cannot be empty.

Foreign key ingestion

A foreign key, in relational databases, is a field in one table that refers to the primary key of another table. A primary key is a column or combination of columns, to uniquely identify table records.

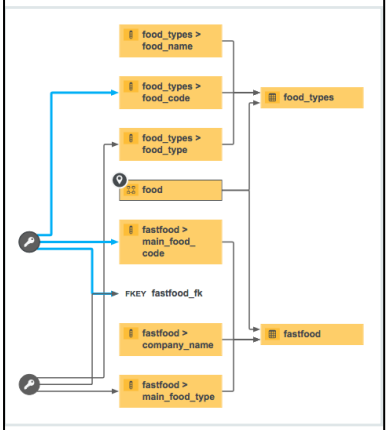
- The table with the primary key is referred to as the referenced table or parent table.
- The table with the foreign key is referred to as the child table.

Ingesting foreign keys

In Data Catalog, a foreign key is ingested as an asset of the Foreign key type. See Foreign Key asset page.

The Foreign key asset creates relations between columns of different tables. It consists of foreign key mappings between the parent and child table.

In the following example, you see an overview of the tables, columns and a foreign key:



- **food**: Schema that consists of two tables:
  - **food_types**: Table with the columns food_name, food_code and food_type.
  - **fastfood**: Table with the columns main_food_code, company_name and main_food_type.
- **fastfood_fk**: Foreign key asset consists of one or more foreign key mappings.
  - mapping 1 (marked with blue arrows):
    - Constrains the column **main_food_code** from the child table.
    - References the column **food_code** from the parent table.
  - mapping 2:
    - Constrains the column **main_food_type** from the child table.
    - References the column **food_type** from the parent table.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Registering a data source via Jobserver

By registering a data source via Jobserver, you connect a data source to Collibra. With this, you can make metadata of the data source available in Collibra.

During the data source registration process, you create a Schema asset. Via this asset, you can refresh the metadata of the data source.

Tip You can also register a data source via Edge.

Data source ingestion steps

The following table shows the steps required for data source ingestion.

Step	What?	Description
1	Register a data source	Registering a data source creates a connection between your data source and Collibra. It makes metadata of the data source available in Collibra. Note You can register a data source using a Collibra-provided driver or your own driver.
2	Ingestion	After registering a data source, Collibra creates a **Physical Data Dictionary** domain and new assets of the type Schema, Table and Column, corresponding to the data in your data source. Note Once you used a connection to successfully register a data source via Jobserver, you cannot change the connection properties. See Error when managing connection properties of a driver for Jobserver.
3	Refresh a data source	Refreshing the schema of a registered data source updates the metadata of the data source in Collibra. You typically do this when the data in a registered data source has been updated. Tip You can do this manually or automatically at fixed intervals.

Profiling data options

When you register your data source, you can choose profiling options for the registered data.

Option	Description
Store Data Profile	Option to perform data profiling on the registered data.
Detect advanced data types	Option to detect advanced data types in the data source.
Store Sample Data	Option to extract sample data from the registered data.

Option	Description
Tables excluded from registration	Database tables that will not be ingested. Note • If required, you can exclude multiple tables. To do this, press *Enter* after typing a value and then type the next. • You can use an asterisk (*) as wildcard to select multiple tables. For example, if you want to exclude the tables that all start with act_, you can enter *act_**. • The table names are case sensitive. • You can add or remove tables from this list by refreshing the schema. • The Table assets that are created after ingestion have an attribute type called Table Type that defines the type of table that is declared in the data source. For example, TABLE, VIEW,...

After registering a data source

When the registration is complete:

- A message at the top right tells you that data source registration is complete. A domain and Schema asset are immediately created and an ingestion job is started.
- You can immediately add the registered data source to a data set by clicking the corresponding link in the confirmation message.
- The ingestion job creates assets that represent the metadata of the data source.

  > Note Table assets that are created after ingestion have an attribute type called Table Type that defines the type of table that is declared in the data source. For example, TABLE, VIEW,...

- A workflow to assign a technical steward to the new domain is started. This is a simple packaged workflow that you can edit to fit your organization's needs. When you have assigned a technical steward, that technical steward has to set the security classification and indicate whether the data elements contain personally identifiable information (PII).

Register a data source using a Collibra-provided driver

You can register a database as a data source using one of the JDBC drivers provided by Collibra Marketplace.

Tip You can also do this with your own JDBC driver.

Warning
- This operation should only be executed by your database administrator.
- The Collibra-provided drivers have been tested with Collibra Data Governance Center version 5.7.5. In older versions, you might encounter unexpected behavior.

Steps

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.
2. On the main toolbar, click +.
   » The **Create** dialog box appears.
3. In the **Create** dialog box, click **Register data source (use a Collibra provided driver)**.
4. If there is no JDBC driver available, add and configure the driver of your preference.
5. In the **Register data source** dialog box, enter the required information.

Field	Description
Process on	The jobserver used for ingesting.
Schema name	This name is used in Collibra as schema asset and must therefore be unique.
Schema description	The description of the schema. This is used as description of the schema asset.
Data owner	The owner of the registered data in Collibra.

6. Click **Next**.
7. Enter the database connection properties.

Option	Description
JDBC driver version	The JDBC driver to connect to your database. Note By default, you see the name of the driver that was used last.
Connect via	The jobserver used for ingesting.
<Configuration properties>	The connection properties as defined in your JDBC driver. Note For more information on the connection details of supported data sources, see JDBC connection details.
Store credentials	Select this option to store the credentials to access the database. With a schema refresh, you can clear this option again.
Username	Username to access the database. Note This field is ignored if your data source uses Cyberark, Kerberos or NTLM.
Password	Corresponding password to access the database. Note This field is ignored if your data source uses Cyberark, Kerberos or NTLM.
Schedule data refresh	Enable or disable a schedule to automatically refresh the data registration.
Cron pattern	Schedule of the data refresh as a Cron pattern. If you create an invalid Cron pattern, Collibra Platform Self-Hosted stops responding.
Time zone	The time zone of the database.

> Note If Collibra DGC cannot connect to the database, you cannot continue the data source registration wizard.

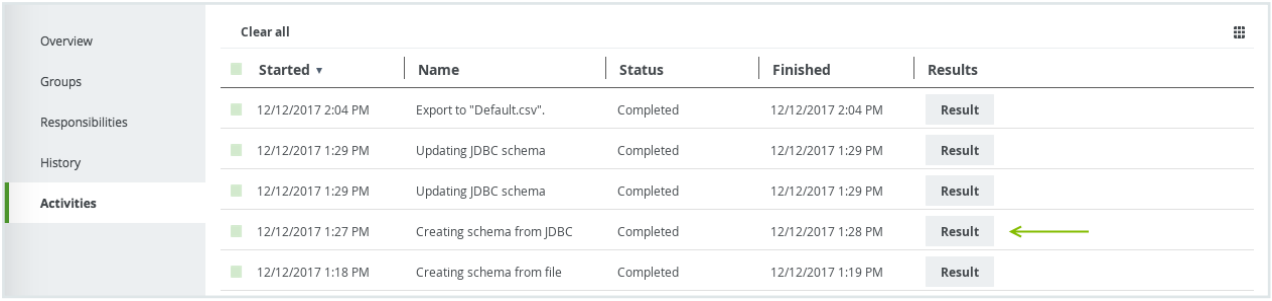
8. Click **Next**.
9. Select the data profiling options.

Option	Description
Store Data Profile	Option to perform data profiling on the registered data.
Detect advanced data types	Option to detect advanced data types in the data source.
Store Sample Data	Option to extract sample data from the registered data.
Tables excluded from registration	Database tables that will not be ingested. Note ◦ If required, you can exclude multiple tables. To do this, press *Enter* after typing a value and then type the next. ◦ You can use an asterisk (*) as wildcard to select multiple tables. For example, if you want to exclude the tables that all start with act_, you can enter *act_**. ◦ The table names are case sensitive. ◦ You can add or remove tables from this list by refreshing the schema. ◦ The Table assets that are created after ingestion have an attribute type called Table Type that defines the type of table that is declared in the data source. For example, TABLE, VIEW,...

10. Click **Create**.

What's next?

The data source is registered and the data is automatically ingested. The ingestion of data is executed in a job. You can see this job in the list of activities.



Click the **Result** button to open the data profiling results.

Tip

- If the database contains foreign keys, they will be registered as new assets of the Foreign Key asset type. Assets of this type contain the complex relation, which is the link between all column assets that are part of the foreign key definition.
  However, the complex relation is not created if a column is part of a table that is added to the list of **Tables excluded from registration**.
- If you exclude a table during the schema refresh, the corresponding table, column assets and foreign key mapping will be deleted.

Manage Collibra-provided JDBC drivers

To register a database as a data source you need a JDBC driver. You can use one of the JDBC drivers provided by Collibra Marketplace.

This allows you to do the following:

- Edit an existing JDBC driver.
- Install a new JDBC driver for a data source type that has an existing JDBC driver, for example Oracle12c.
- Install a new JDBC driver for a data source type that doesn't have a JDBC driver yet, for example Amazon EMR.

Tip You can also do this with your own JDBC drivers.

Warning
- This operation should only be executed by your database administrator.
- The Collibra-provided drivers have been tested with Collibra Data Governance Center version 5.7.5. In older versions, you might encounter unexpected behavior.

Steps

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.
2. On the main toolbar, click +.
   » The **Create** dialog box appears.
3. In the **Create** dialog box, click **Register data source (use a Collibra provided driver)**.
4. If a JDBC driver is already installed for your data source, do the following:
   a. Enter the schema properties.

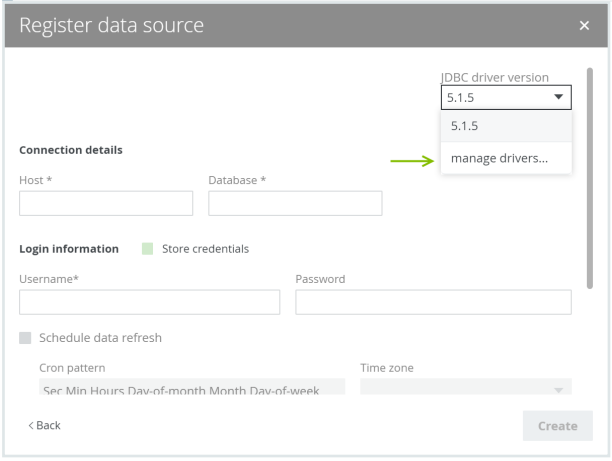
Field	Description
Schema name	This name is used in Collibra as schema asset and must therefore be unique.
Schema description	The description of the schema. This is used as description of the schema asset.
Data owner	The owner of the registered data in Collibra.

   b. Click **Next**.

In the **JDBC driver version** field, click **manage drivers...**.

> Note By default, you see the name of the driver that was used last.

c.



5. Perform one of the following steps:
   - Click **Add JDBC Driver** if you want to create a new JDBC driver.
   - Click ✎ if you want to edit an existing JDBC driver.
6. Enter the required information.

Field	Description
JDBC Driver Version Name	The name of the JDBC driver. Tip As a best practice, we recommend you use a strict naming convention which includes the data source and a version number. For example: Google BigQuery 1.5 or MySQL 5.9.
⬆ Upload	Button to upload the relevant files for the data source. Note If you downloaded the JDBC driver from Collibra Marketplace, make sure to unzip the downloaded ZIP file before uploading it to Collibra Data Governance Center. Note The JDBC driver has to be in JAR format.

Field	Description
Driver files	This table contains a list of uploaded files. You can remove a driver file by clicking .

7. Click **Next**.
8. Configure the JDBC connection.

> Note For more information on the connection details of supported data sources, see JDBC connection details.

9. Click **Create**.

What's next?

You can now complete the data source registration wizard for Collibra-provided JDBC drivers.

Register a data source using your own driver

You can register a database as a data source using one of your own drivers.

> Tip You can also do this with a Collibra-provided JDBC driver.

This operation should only be executed by your database administrator.

Prerequisites

- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have set up the JDBC driver of your source data, for example MySQL.
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- If you are using a Collibra Data Intelligence Cloud environment with an on-premises Jobserver, both must have the same installer version. You can find the installer

version of your Collibra Data Intelligence Cloud environment at the bottom of the sign-in window of its Collibra Console, for example 5.9.1-0

- You have a resource role with the following resource permissions on the **Schema** community:
  - Asset > add
  - Attribute > add
  - Domain > add
  - Attachment > add
- You have the permissions to retrieve the metadata of the following database components through the JDBC Driver Database Metadata methods:
  - Schemas
  - Tables
  - Columns
  - Primary keys
  - Foreign keys

> Note
> - For the list of supported databases and versions, go to Databases supported versions.
> - For the JDBC connection details of the various databases, go to JDBC connection details.

Steps

1. On the main menu, click ⠿, and then click **Catalog**.
   » The Catalog Home opens.
2. On the main toolbar, click +.
   » The **Create** dialog box appears.
3. In the **Register data source** dialog box, click the type of your data source.
4. If there is no JDBC driver available, add and configure the driver of your preference.
5. In the **Register data source** dialog box, enter the required information.

Field	Description
Process on	The jobserver used for ingesting.
Schema name	This name is used in Collibra as schema asset and must therefore be unique.

Field	Description
Schema description	The description of the schema. This is used as description of the schema asset.
Data owner	The owner of the registered data in Collibra.

6. Click **Next**.
7. Enter the database connection properties.

Option	Description
JDBC driver version	The JDBC driver to connect to your database.
Connect via	The jobserver used for ingesting.
Database	Name of the database. This field is not available for all data sources.
Host	Hostname to access the database.
Port	Port to access the database.

<table>
<tr><th>Option</th><th>Description</th></tr>
<tr><td><Configuration properties></td><td>
The connection properties as defined in your JDBC driver.

Note For more information on the connection details of supported data sources, see JDBC connection details.

If you want to use Kerberos authentication, you also need the following connection properties.
<table>
<tr><th>Label</th><th>Description</th></tr>
<tr><td>Principal</td><td>The Kerberos principal identity.</td></tr>
<tr><td>Kerberos realm</td><td>The Kerberos realm name.</td></tr>
<tr><td>Login context name</td><td>The login context name that is used as the index to the configuration.</td></tr>
<tr><td>Jaas file name</td><td>The name of the Jaas file.</td></tr>
<tr><td>Kerberos configuration file</td><td>The configuration file containing specific properties for Kerberos authentication.</td></tr>
</table>
If you want to use NTLM authentication, you also need the following connection properties.
<table>
<tr><th>Label</th><th>Description</th></tr>
<tr><td>Security</td><td>The security that enables the authentication</td></tr>
<tr><td>Authentication scheme</td><td>The used authentication scheme, which is NTLM.</td></tr>
</table>
If you want to use CyberArk authentication, you need the following connection properties.
</td></tr>
</table>

<table>
<tr><th>Option</th><th>Description</th></tr>
<tr><td></td><td>
<table>
<tr><th>Label</th><th>Description</th></tr>
<tr><td>Keystore file</td><td>The name of the keystore file. The keystore must contain the client key and client certificate or certificate chain.

If <code>defaultTruststore</code> is set to <code>false</code>, the keystore has to contain the trusted CA certificate needed to validate the server certificate offered by CyberArk.

The value must have the following format: <code>file://<keystore-file name.jks></code>.

Example
<code>file://cyberark-keystore.jks</code></td></tr>
<tr><td>Keystore password</td><td>The password required to open the keystore.</td></tr>
<tr><td>Default truststore</td><td>The indication of the default truststore. The default value is set to <code>False</code>.

◦ <code>False</code>: The certificate is validated through the keystoreFile property.
◦ <code>True</code>: The certificate is validated through the default truststore from the Java JRE. This is recommended when CyberArk is set up to offer a server certificate that can be validated by a public CA (certification authority).</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><th>Option</th><th>Description</th></tr>
<tr><td></td><td>
<table>
<tr><th>Label</th><th>Description</th></tr>
<tr><td>CyberArk address</td><td>The host and port number through which the CyberArk server is accessible. The format of the address is <code>hostname:port</code>.

Example
<code>my.cyberark.com:5502</code></td></tr>
<tr><td>CyberArk application ID</td><td>The application ID as defined in CyberArk.

This ID should be provided by your network or system administrator.</td></tr>
<tr><td>CyberArk query</td><td>The CyberArk query.

This query should be provided by your network or system administrator.</td></tr>
</table>
</td></tr>
<tr><td>Store credentials</td><td>Select this option to store the credentials to access the database. With a schema refresh, you can clear this option again.</td></tr>
<tr><td>Username</td><td>Username to access the database.

Note This field is ignored if your data source uses any authentication method other than credentials.</td></tr>
<tr><td>Password</td><td>Corresponding password to access the database.

Note This field is ignored if your data source uses any authentication method other than credentials.</td></tr>
<tr><td>Schedule data refresh</td><td>Enable or disable a schedule to automatically refresh the data registration.</td></tr>
</table>

Option	Description
Cron pattern	Schedule of the data refresh as a Quartz Cron pattern. Warning If you create an invalid Cron pattern, Collibra Platform Self-Hosted stops responding.
Time zone	The time zone of the database.

Note If Collibra cannot connect to the database, you cannot continue the data source registration wizard.

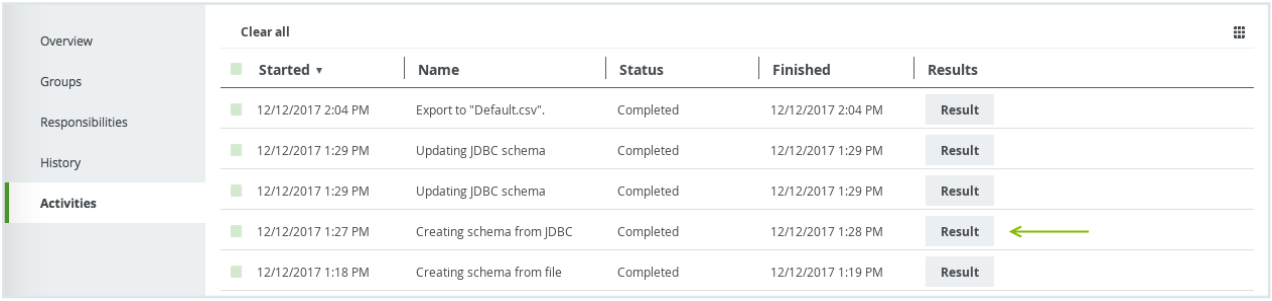
8. Click **Next**.
9. Select the data profiling options.

Option	Description
Store Data Profile	Option to perform data profiling on the registered data.
Detect advanced data types	Option to detect advanced data types in the data source.
Store Sample Data	Option to extract sample data from the registered data.
Tables excluded from registration	Database tables that will not be ingested. Note ◦ If required, you can exclude multiple tables. To do this, press *Enter* after typing a value and then type the next. ◦ You can use an asterisk (*) as wildcard to select multiple tables. For example, if you want to exclude the tables that all start with act_, you can enter *act_**. ◦ The table names are case sensitive. ◦ You can add or remove tables from this list by refreshing the schema. ◦ The Table assets that are created after ingestion have an attribute type called Table Type that defines the type of table that is declared in the data source. For example, TABLE, VIEW,...

10. Click **Create**.

What's next?

The data source is registered and the data is automatically ingested. The ingestion of data is executed in a job. Go to the list of activities to follow up on the progress,



Click the **Result** button to open the data profiling results.

Tip

- If the database contains foreign keys, they will be registered as new assets of the **Foreign Key** asset type. Assets of this type contain the complex relation, which is the link between all column assets that are part of the foreign key definition.
  However, the complex relation is not created if a column is part of a table that is added to the list of **Tables excluded from registration**.
- If you exclude a table during the schema refresh, the corresponding table, column assets and foreign key mapping will be deleted.

Register an Excel file as data source

Note If you are using a Collibra Data Intelligence Cloud environment with an on-premises Jobserver, the Jobserver version must be compatible with the cloud version. You can find the version of your Collibra Data Intelligence Cloud environment at the bottom of the sign-in window, for example 2023.11.0.

Prerequisites

- You have downloaded an Excel file.
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- You have a resource role with the following resource permissions:
  - Asset > add
  - Attribute > add
  - Domain > add
  - Attachment > add

Steps

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.
   Or open any asset of the type Schema, Data Set, Table, Column or Tableau Server.
2. On the main toolbar, click +.
   » The **Create** dialog box appears.
3. In the **Create** dialog box, click **Register data source (use your own driver)**.
   » The **Register data source (use your own driver)** dialog box appears.
4. In the **Register data source** dialog box, click **Excel**.
5. Enter the data source configuration.

Field	Description
Process on	The jobserver used for ingesting.
Schema name	This name is used in Collibra as schema asset and must therefore be unique.
Schema description	The description of the schema. This is used as description of the schema asset.
Data owner	The owner of the registered data in Collibra.

6. Click **Next**.

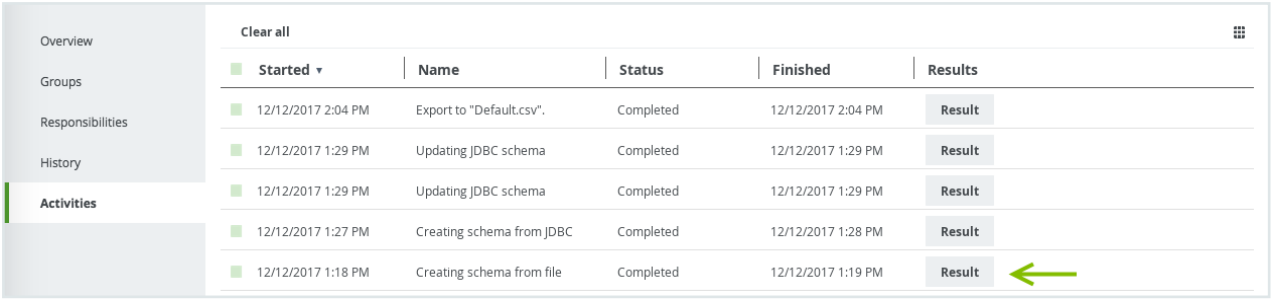
7. Select the data profiling options.

Option	Description
Store Data Profile	Option to perform data profiling on the registered data.
Detect advanced data types	Option to detect advanced data types in the data source.
Store Sample Data	Option to extract sample data from the registered data.
Tables excluded from registration	Database tables that will not be ingested. Note ◦ If required, you can exclude multiple tables. To do this, press *Enter* after typing a value and then type the next. ◦ You can use an asterisk (*) as wildcard to select multiple tables. For example, if you want to exclude the tables that all start with act_, you can enter *act_**. ◦ The table names are case sensitive. ◦ You can add or remove tables from this list by refreshing the schema. ◦ The Table assets that are created after ingestion have an attribute type called Table Type that defines the type of table that is declared in the data source. For example, TABLE, VIEW,...

8. Click **Create**.

What's next?

The data source is registered and the data is automatically ingested. The ingestion of data is executed in a job. You can see this job in the list of activities.



Click the **Result** button to open the data profiling results.

If you have selected the option to perform data profiling and/or extract sample data, you can go to the schema page to verify if this process has completed in the **Synchronization Status** field. Refresh the schema page until the **Synchronization Status** field has disappeared.

Note that there Collibra may have resolved some small issues:

Use case	Behavior
Missing column name	If the file is missing a column name, a default name will be given, _c + index. The index is the column position in the file starting with 0. For example, **_c4** corresponds with the fifth column in the file.
Duplicate column name	If the file has duplicate column names, the column names will be appended with an index. The index is the column position in the file, starting with 0. For example, **mycol1** and **mycol3** are columns 2 and 4 in the file, each with the column name **mycol**.
Empty sheet	If the Excel file has empty sheets, they are not registered.

Register a CSV file as data source

> Note If you are using a Collibra Data Intelligence Cloud environment with an on-premises Jobserver, the Jobserver version must be compatible with the cloud version. You can find the version of your Collibra Data Intelligence Cloud environment at the bottom of the sign-in window, for example 2023.11.0.

Prerequisites

- You have downloaded a CSV file.
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- You have a resource role with the following resource permissions:
  - Asset > add
  - Attribute > add
  - Domain > add
  - Attachment > add

Steps

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.
   Or open any asset of the type Schema, Data Set, Table, Column or Tableau Server.
2. On the main toolbar, click +.
   » The **Create** dialog box appears.
3. In the **Create** dialog box, click **Register data source (use your own driver)**.
   » The **Register data source (use your own driver)** dialog box appears.
4. In the **Register data source** dialog box, click **Csv**.
5. Enter the data source configuration.

Field	Description
Process on	The jobserver used for ingesting.
Schema name	This name is used in Collibra as schema asset and must therefore be unique.
Schema description	The description of the schema. This is used as description of the schema asset.
Data owner	The owner of the registered data in Collibra.

6. Click **Next**.

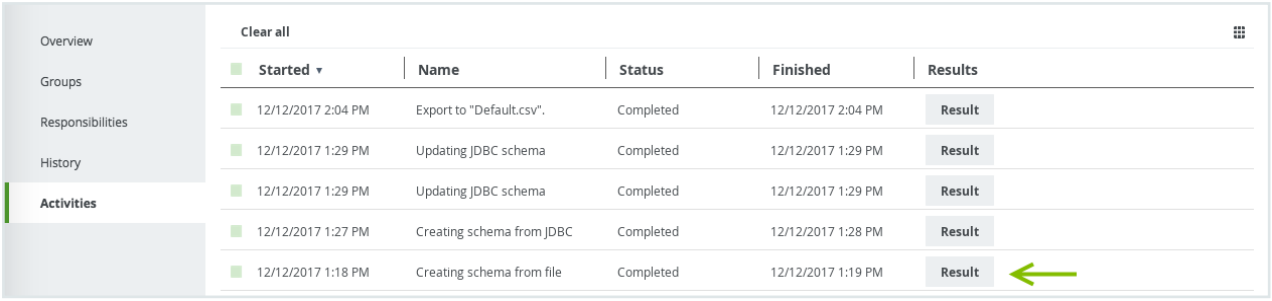
7. Select the data profiling options.

Option	Description
Store Data Profile	Option to perform data profiling on the registered data.
Detect advanced data types	Option to detect advanced data types in the data source.
Store Sample Data	Option to extract sample data from the registered data.
Tables excluded from registration	Database tables that will not be ingested. Note ◦ If required, you can exclude multiple tables. To do this, press *Enter* after typing a value and then type the next. ◦ You can use an asterisk (*) as wildcard to select multiple tables. For example, if you want to exclude the tables that all start with act_, you can enter *act_**. ◦ The table names are case sensitive. ◦ You can add or remove tables from this list by refreshing the schema. ◦ The Table assets that are created after ingestion have an attribute type called Table Type that defines the type of table that is declared in the data source. For example, TABLE, VIEW,...

8. Click **Create**.

What's next?

The data source is registered and the data is automatically ingested. The ingestion of data is executed in a job. You can see this job in the list of activities.



Click the **Result** button to open the data profiling results.

> Note
> - Empty rows in the CSV file are ignored. As a consequence, they do not count towards the row count or missing value count.
> - You can define the format of empty values by configuring the data profiling behavior. However, if a field is empty in the CSV file, it will be considered empty even if it does not match the format defined in the configuration.

If you selected the option to perform data profiling and/or extract sample data, you can verify that the process was completed in the Synchronization Status field on the schema asset page. Refresh the schema page until the **Synchronization Status** field disappears.

Note that there Collibra may have resolved some small issues:

Use case	Behavior
Missing column name	If the file is missing a column name, a default name will be given, _c + index. The index is the column position in the file starting with 0. For example, **_c4** corresponds with the fifth column in the file.
Duplicate column name	If the file has duplicate column names, the column names will be appended with an index. The index is the column position in the file, starting with 0. For example, **mycol1** and **mycol3** are columns 2 and 4 in the file, each with the column name **mycol**.
Empty sheet	If the Excel file has empty sheets, they are not registered.

Manage your own JDBC drivers

To register a database as a data source you need a JDBC driver. You can use one of your own JDBC drivers.

For more information, see Supported data sources for data source registration.

This allows you to do the following:

- Edit an existing JDBC driver.
- Install a new JDBC driver for a data source type that has an existing JDBC driver, for example Oracle12c.
- Install a new JDBC driver for a data source type that doesn't have a JDBC driver yet, for example Amazon EMR.

> Tip You can also do this with a Collibra-provided JDBC driver that you download from Collibra Marketplace.

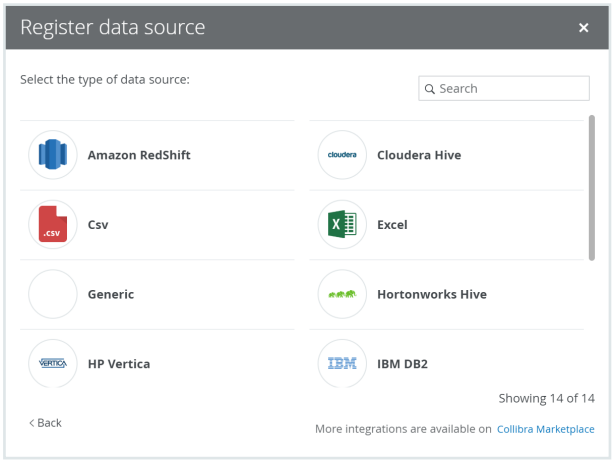
This operation should only be executed by your database administrator.

Prerequisites

- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have downloaded the JDBC driver of your choice as an archive file (for example, ZIP or JAR).
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- You have a resource role with the following resource permissions on the **Schema** community:
  - Asset > add
  - Attribute > add
  - Domain > add
  - Attachment > add

Steps

1. On the main menu, click ⋮⋮⋮, and then click **Catalog**.
   » The Catalog Home opens.
2. On the main toolbar, click +.
   » The **Create** dialog box appears.
3. In the **Create** dialog box, click **Register data source (use your own driver)**.
4. In the **Register data source** dialog box, click the type of your data source.

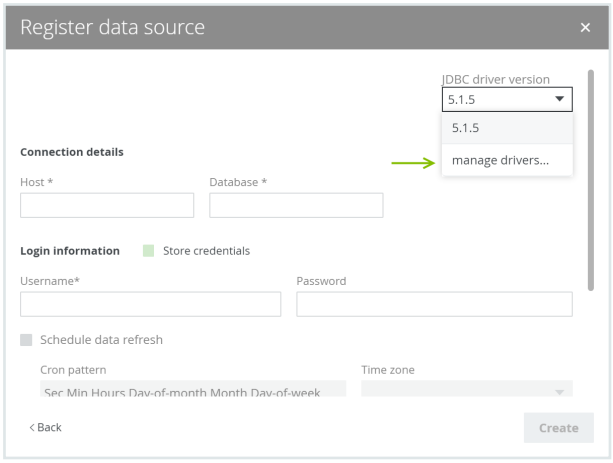


5. If a JDBC driver is already installed for your data source:
   a. Enter the schema properties.

Field	Description
Schema name	This name is used in Collibra as schema asset and must therefore be unique.
Schema description	The description of the schema. This is used as description of the schema asset.
Data owner	The owner of the registered data in Collibra.

   b. Click **Next**.

c. In the **JDBC driver version** field, click **manage drivers...**.



6. Perform one of the following steps:
   - Click **Add JDBC Driver** if you want to create a new JDBC driver.
   - Click ✎ if you want to edit an existing JDBC driver.
7. Enter the required information.

Field	Description
JDBC Driver Version Name	The name of the JDBC driver. Tip As a best practice, we recommend you use a strict naming convention which includes the data source and a version number. For example: Google BigQuery 1.5 or MySQL 5.9.
⬆ Upload	Button to upload the relevant files for the data source. The JDBC driver should be in JAR or ZIP format with a valid Java archive structure. For authentication with CyberArk, you also need to upload a keystore file in JKS format. Note When you click the button, an **Open** dialog box appears. By default, the dialog box filters on JAR, ZIP and CONF files. However, you can change the filter to show all files. For Hortonworks Hive with Kerberos authentication, you need two files: **jaas.conf** and **krb5.conf**.

Field	Description
Driver files	This table contains a list of uploaded files. You can remove a driver file by clicking .

8. Click **Next**.
9. Configure the JDBC connection.

> Note For more information on the connection details of supported data sources, see JDBC connection details of your own drivers.

10. Click **Create**.

What's next?

You can now complete the data source registration wizard.

JDBC connection details of your own drivers

In this section, you will see the connection details needed to register a data source or manage your own JDBC driver.

> Note About the **Connection properties** table:
>
> - The **Label** column is the value that will appear in the connection details dialog box of the **Data Source Registration** wizard.
> - The **Property** column contains the parameters in which the user input will be saved.

Amazon Redshift

Label	Property	Mandatory
Hostname	host	Yes

Label	Property	Mandatory
Port	port	Yes
Database	database	Yes
Schema	schema	Yes

Cloudera Hive

Label	Property	Mandatory
URL (hostname:port)	host	Yes
Principal	principal	Yes
Schema	schema	Yes

Hortonworks Hive

Label	Property	Mandatory
URL (hostname:port)	host	Yes
Schema	schema	Yes

HP Vertica

Label	Property	Mandatory
Hostname	host	Yes
Port	port	Yes
Database	database	Yes
Schema	schema	Yes

IBM DB2

Label	Property	Mandatory
Hostname	host	Yes
Port	port	Yes
Database	database	Yes
Schema	schema	Yes

MapR Hive

Label	Property	Mandatory
URL (hostname:port)	host	Yes
Schema	schema	Yes

Microsoft SQL Server

Label	Property	Mandatory
Hostname	host	Yes
Port	port	Yes
Database	databaseName	Yes
Schema	schema	Yes

MySQL

Label	Property	Mandatory
Hostname	host	Yes
Port	port	Yes
Database	database	Yes

Oracle DB

Label	Property	Mandatory
Hostname	host	Yes
Port	port	Yes
SID	sid	Yes
Schema	schema	Yes

PostgreSQL

Label	Property	Mandatory
Hostname	host	Yes
Port	port	Yes
Database	database	Yes
Schema	schema	Yes

Teradata

Label	Property	Mandatory
Hostname	host	Yes
Port	port	Yes
Database	database	Yes
Schema	schema	Yes

Authentication methods

Certain authentication methods require additional connection properties.

NTLM

If you want to use NTLM authentication, you also need the following connection properties.

Label	Property	Mandatory
Security	*integratedSecurity* must be value `True`.	Yes
Authentication scheme	*authenticationScheme* must be value *NTLM*.	Yes

Kerberos

If you want to use Kerberos authentication, you also need the following connection properties.

Label	Property	Mandatory
Principal	principal	Yes
Kerberos realm	realm	Yes

Label	Property	Mandatory
Login context name	loginContextName You can find the value for this property in the jaas.conf file.	Yes
Jaas file name	com.collibra.jobserver.dto.catalog.JdbcConnection.jaasConfig	Yes
Kerberos configuration file	com.collibra.jobserver.dto.catalog.JdbcConnection.krbConfig	Yes

Cyberark

If you want to use CyberArk authentication, you need the following connection properties. If you use one of the CyberArk connection properties, Data Catalog automatically uses CyberArk authentication.

Label	Property	Mandatory
Keystore file	keystoreFile	Yes
Keystore password	keystorePass	Yes
Default truststore	defaultTruststore	No
CyberArk address	cyberarkAddress	Yes
CyberArk application ID	cyberarkAppId	Yes
CyberArk query	cyberarkQuery	Yes

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Authentication

If you register a database as data source or manage a JDBC driver, you can use various authentication methods to access your data source.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

CyberArk authentication

CyberArk is middleware to manage authentication and is used to provide access to various data sources. You can use CyberArk to let Data Catalog access and ingest data sources with username and password authentication.

> Note You can only authenticate to data sources using username and password authentication.

Setting up CyberArk authentication

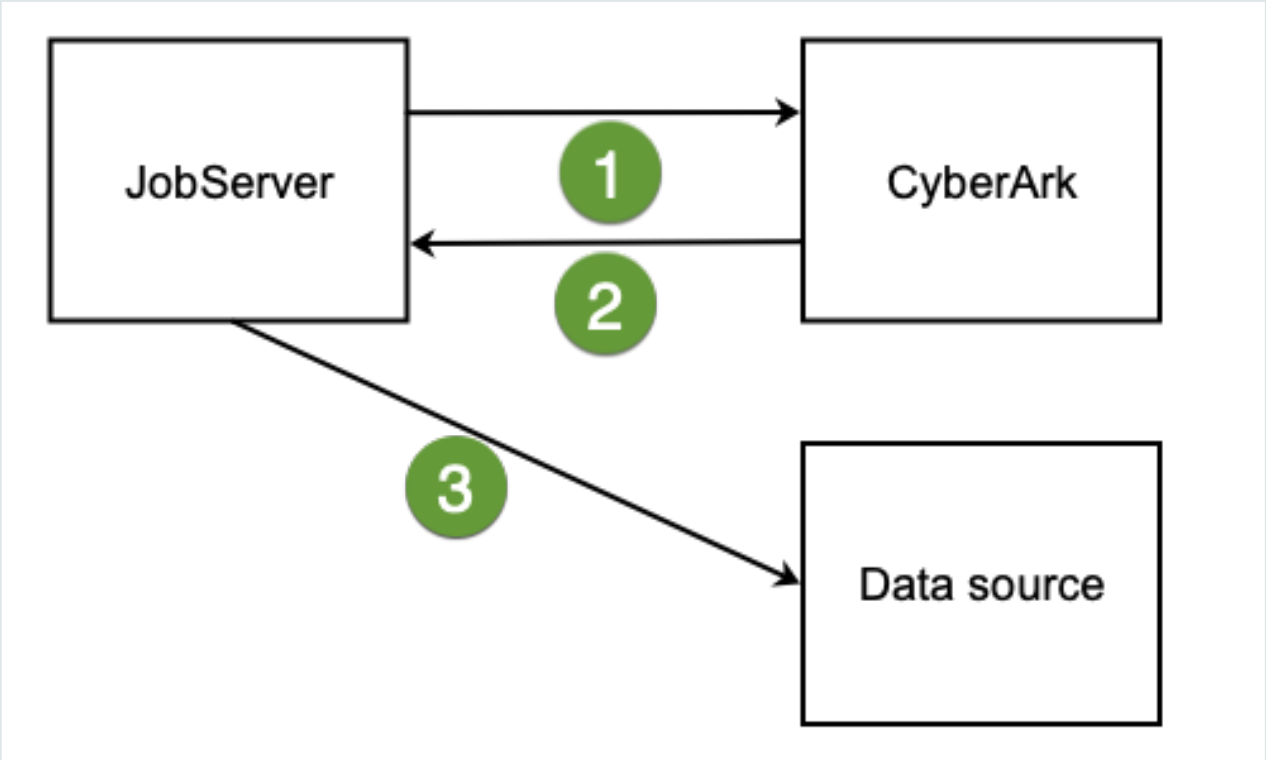
You set up CyberArk authentication when you register your data source or manage your JDBC driver. When you register your data source or manage your JDBC driver, you only provide the username, the password you need to authenticate to the data source is stored in CyberArk and is retrieved by the Jobserver. When you ingest a data source using CyberArk authentication, the Jobserver uses certificate-based mutual authentication to authenticate to CyberArk.

> Note The connection to CyberArk is only supported over HTTPS.

To authenticate via CyberArk, you have to enable CCP WebService in CyberArk and keep the default name AIMWebService unchanged. You also have to provide your own CyberArk certificates via a JKS keystore that you upload to Collibra when you register your

data source or manage your JDBC driver. The JKS keystore contains the CyberArk client certificates, the private key and, if required, a server certificate.

Authentication workflow



Step	Action
1	The Jobserver requests credentials from CyberArk through a certificate-based mutual authentication.
2	CyberArk provides the Jobserver with a username and password.
3	The Jobserver uses these credentials to authenticate to a data source.

Configuration

If you want to use CyberArk authentication, you need the following connection properties. If you use one of the CyberArk connection properties, Data Catalog automatically uses CyberArk authentication.

Label	Property	Description	Mandatory
Keystore file	keystoreFile	The name of the keystore file. The keystore must contain the client key and client certificate or certificate chain. If `defaultTruststore` is set to `false`, the keystore has to contain the trusted CA certificate needed to validate the server certificate offered by CyberArk. The value must have the following format: `file://<keystore-file name.jks>`. Example `file://cyberark-keystore.jks`	Yes
Keystore password	keystorePass	The password required to open the keystore.	Yes
Default truststore	defaultTruststore	The indication of the default truststore. The default value is set to `False`. • `False`: The certificate is validated through the keystoreFile property. • `True`: The certificate is validated through the default truststore from the Java JRE. This is recommended when CyberArk is set up to offer a server certificate that can be validated by a public CA (certification authority).	No
CyberArk address	cyberarkAddress	The host and port number through which the CyberArk server is accessible. The format of the address is `hostname:port`. Example `my.cyberark.com:5502`	Yes

Label	Property	Description	Mandatory
CyberArk application ID	cyberarkAppId	The application ID as defined in CyberArk. This ID should be provided by your network or system administrator.	Yes
CyberArk query	cyberarkQuery	The CyberArk query. This query should be provided by your network or system administrator.	Yes

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

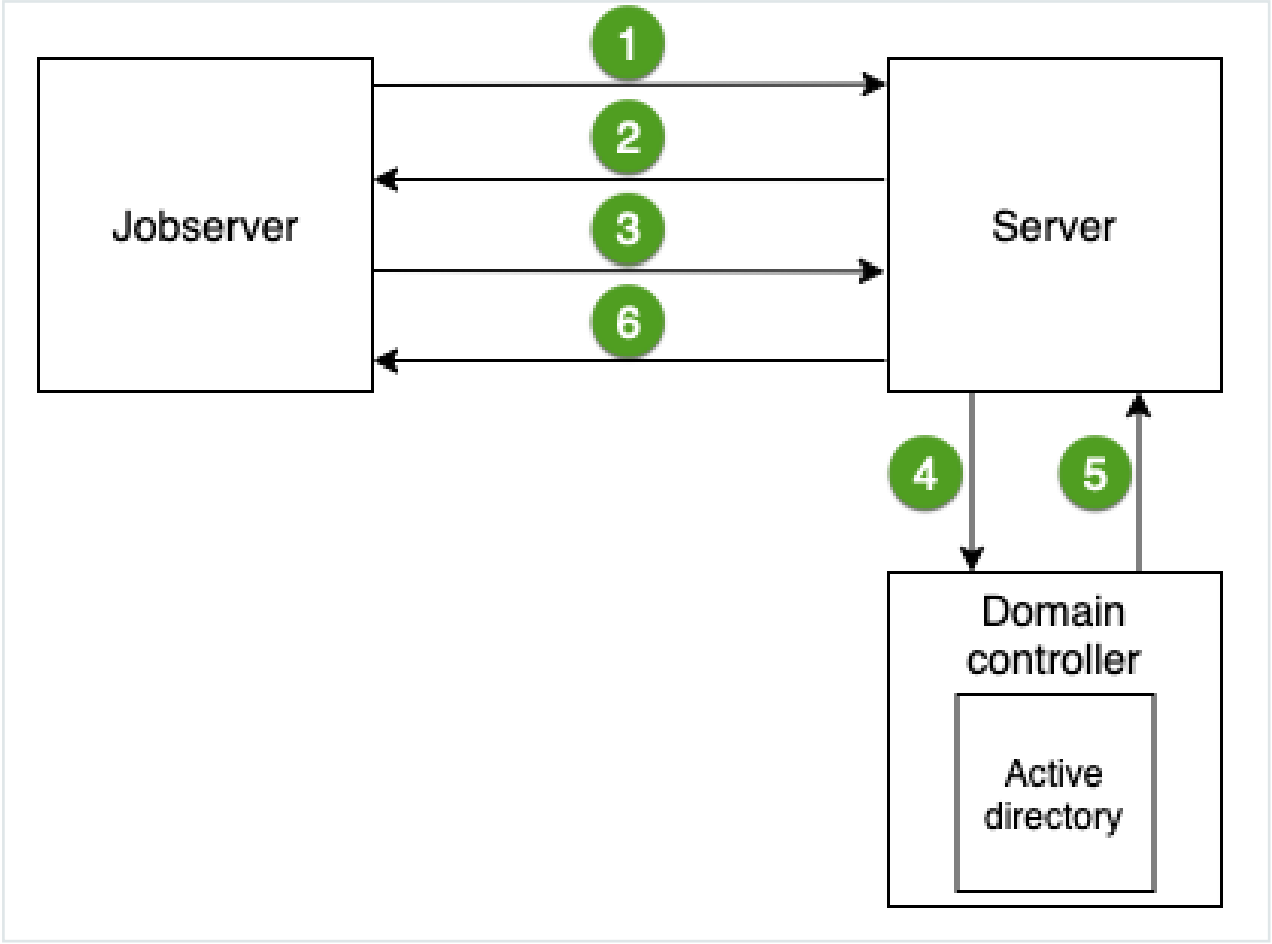
NTLM authentication

NTLM is an authentication protocol used on networks that include systems running the Windows operating system and on stand-alone systems. It uses a challenge-response authentication to connect to the Microsoft SQL Server data source. For more information, see the Microsoft NTLM user guide.

If you have a Microsoft SQL Server data source that uses NTLM authentication, you have to set up specific connection properties when you register the data source or manage the JDBC driver.

Authentication workflow

When you ingest a Microsoft SQL Server data source using NTLM authentication, the Jobserver connects to the server to request access. The server then sends a challenge for the Jobserver to encrypt and send back. The domain controller validates that response and gives the Jobserver access to the data source.



Step	Action
1	The Jobserver requests access to the Microsoft SQL Server data source.
2	The server sends a challenge message to the Jobserver to identify the Jobserver.
3	The Jobserver sends a response back to the server.
4	The server sends the challenge and response message to the domain controller.
5	The active directory on the domain controller validates the challenge and response message and sends the result to the server.
6	The server gives the Jobserver permission to access the data source.

Configuration

If you want to use NTLM authentication, you also need the following connection properties.

Label	Property	Description	Mandatory
Security	*integratedSecurity* must be value `True`.	The security that enables the authentication	Yes
Authentication scheme	*authenticationScheme* must be value *NTLM*.	The used authentication scheme, which is NTLM.	Yes

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Kerberos authentication

You can use Kerberos authentication for registering a Hive data source, for example Cloudera Hive, Hortonworks Hive or MapR Hive.

Authentication type

We only support Kerberos username and password authentication, not keytab. Ensure that you configure this in the **jaas.conf** file by setting the **useKeyTab** option to *false*.

In the following jaas.conf example, **Client** is the value of the **loginContextName** field when you configure the Kerberos connection configuration.

Example

```
Client {
com.sun.security.auth.module.Krb5LoginModule required
useKeyTab=false
useTicketCache=true;
};
```

If there are multiple entries in this configuration file, ask the database administrator or network administrator which one to use. For more information about the Jaas login configuration file, see the Java documentation.

Example krb5.conf

The following is an example configuration file of Kerberos.

```
[libdefaults]
  renew_lifetime = 7d
  forwardable = true
  default_realm = MY.REALM
  ticket_lifetime = 24h
  dns_lookup_realm = false
  dns_lookup_kdc = false
  default_ccache_name = /tmp/krb5cc_%{uid}

[logging]
  default = FILE:/var/log/krb5kdc.log
  admin_server = FILE:/var/log/kadmind.log
  kdc = FILE:/var/log/krb5kdc.log

[realms]
  MY.REALM = {
    kdc = <kdc.my.realm>
    admin_server = <kadmin.my.realm>
  }
```

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.

For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Enable debug for Kerberos authentication issues

If an error occurs during the Kerberos authentication, you can enable debugging to track the root cause of the error.

To enable debugging for the Kerberos authentication:

1. On the server that hosts the Jobserver service, open the file **context_jvm.conf** in **<drive>/collibra/spark-jobserver/conf** for editing.
2. Is the following parameter present in the file: `-Dsun.security.krb5.debug`
   - Yes: Set its value to *true*.
   - No: Add the following line to the file: `-Dsun.security.krb5.debug=true`
3. Save and close the file.
4. Restart the Jobserver service.

The default log file in which to look for Kerberos authentication issues is **<drive>/collibra_data/logs/context_<context-name>/spark-job-server.log**.
In general, you list the **context_<context-name>** directories and pick the most recent one.

Tip After resolving the authentication issues, set the parameter to *false*.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Cancel a data ingestion job

If you are the one that started the data ingestion job, you can cancel it while the data ingestion job is still running.

Prerequisites

- You have registered a data source.
- You have started the ingestion job.

Steps

1. On the main menu, click , then **Show more**.
   » Your profile page opens on the **Activities** tab page.
2. Click next to the ingestion job to cancel it.

> Note When the job is finished, the icon changes into a icon. You can't cancel the ingestion job anymore.

» The data ingestion job is canceled.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

About refreshing a schema

Refreshing a schema is the process of updating the metadata of a registered data source in Collibra Platform Self-Hosted.

You can refresh a schema manually or automatically at fixed intervals. This is particularly useful if the content of the data source changes regularly.

In this section, you can find the relevant actions to successfully refresh a schema.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Refresh the schema of a registered data source

You can refresh a schema of registered data to update the data and the profiling. It can also be useful to do this to change data types to force the profiling to use the correct type.

> Tip You can also refresh the schema automatically via a schedule.

Prerequisites

- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have set up the JDBC driver of your source data, for example MySQL.
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- If you are using a Collibra Data Intelligence Cloud environment with an on-premises Jobserver, both must have the same installer version. You can find the installer version of your Collibra Data Intelligence Cloud environment at the bottom of the sign-in window of its Collibra Console, for example 5.9.1-0
- You have a resource role with the following resource permissions on the **Schema** community:
  - Asset > add
  - Attribute > add
  - Domain > add
  - Attachment > add
- You have the permissions to retrieve the metadata of the following database components through the JDBC Driver Database Metadata methods:
  - Schemas
  - Tables
  - Columns
  - Primary keys
  - Foreign keys

> Note
> - For the list of supported databases and versions, see Databases supported versions.
> - For the JDBC connection details of the various databases, see JDBC connection details.

Steps

1. Open the Schema asset.
   a. On the main menu, click , and then click **Catalog**.
      » The Catalog Home opens.
   b. In the submenu, click **Data Dictionary** and select the **All Schemas** view.
   c. Click the schema that you want to refresh.

   > Tip You can also use the Collibra Platform Self-Hosted search function to look up your schema.

2. In the view bar, to the right, click **Actions** → **Refresh**.
   » The **Refresh Schema** dialog box appears.

   > Tip If Catalog experience is disabled, the **More** menu is shown instead of **Actions**.

3. Enter the required information.
   This dialog box varies with the data source:
   - Relational database

     > Note
     > - If you exclude a table during the schema refresh, you will delete the corresponding table, column assets and the foreign key mapping (complex relation).
     > - If you clear the **Store credentials** option, the credentials are no longer stored.

   - CSV file
   - Excel file

   This step may take some time.

4. Click **Save & Refresh**.
   » The refresh of the schema starts, you can follow the refresh job in the list of activities.

What's next?

- The representation of the schema is updated: Data Catalog creates, edits and deletes assets as needed.
  - This can lead to refresh conflicts. See Resolve schema refresh conflicts via Jobserver.
  - If you had deleted assets manually, Data Catalog usually doesn't create them again if you refresh the schema. However, if the assets are required to represent the schema structure, Data Catalog can create them again.

    > Example
    > You ingested a schema that contains a table and three columns. In Data Catalog, this is represented by a Schema asset, a Table asset and three Column assets.
    >
    > Additionally, the following relations are created between the relevant assets:
    >
    > - Schema contains/is part of Table
    > - Table contains/is part of Column
    >
    > In the actual data source, the columns are physically inside the table. However, in Data Catalog, they are separate assets linked by relations. As a consequence, you can delete the Table asset without deleting the Column assets. If you did that, Data Catalog creates the Table asset again if you refresh the schema, because the Table asset is needed for the relations to the Column assets.

- If the data source has new values and you selected the checkboxes to store sample data and data profile information, new sample data is generated and all profiling information is updated.
  If you did not select the **Store Sample Data** checkbox, any previously gathered sample data is removed. If you did not select the **Store Data Profile** checkbox, any previously gathered data profiling information is removed.

- Data types or categorical attributes that you changed manually are not updated when you refresh the schema.

  > Note If you change the data type back to the original value assigned by the profiler, Data Catalog can update it if you refresh the schema.

- If you use this schema of the data source for Tableau stitching, you have to restitch after each schema refresh to make sure that all relations are up to date.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Schedule a schema refresh

You can refresh a schema manually, but you can also create a schedule to refresh a schema on a regular basis.

You can only create a refresh schedule for schemas of databases that are registered as a data source, not from CSV or Excel files.

> Tip You can schedule the refresh during the data source registration process or afterwards via the Schema asset.

> Note
> - To enable a scheduled schema refresh, you have to save the credentials in the configuration of a data source registration.
> - The refresh schedule uses Quartz Cron expressions.
> - If you use the schema for Tableau stitching, you have to restitch after each schema refresh to make sure that all relations are up-to-date.

Prerequisites

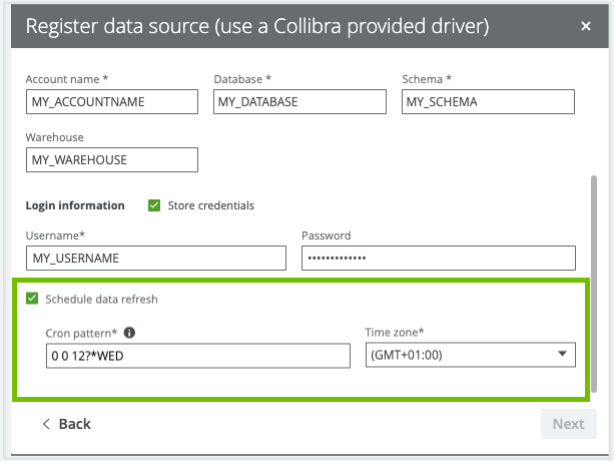
- You have registered a data source.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a role with the following resource permissions on the **Schema** community:
  - Asset: add
  - Attribute: add
  - Domain: add
  - Attachment: add

> Note These permissions are always necessary when registering a data source.

Schedule the refresh during the data source registration process

You can create the refresh schedule when you register a data source.

> Example
> When you register a Snowflake data source in Collibra Platform Self-Hosted, you can create a refresh schedule by selecting **Schedule data refresh**. You can then enter the CRON pattern *0 0 12?*WED* to refresh every Wednesday at 12:00:00 PM.



Schedule the refresh via the Schema asset

You can create the refresh schedule when you refresh the schema of a registered data source via the Schema asset.

1. Open the Schema asset.
   a. On the main menu, click ⠿, and then click Catalog.
      » The Catalog Home opens.
   b. In the submenu, click **Data Dictionary** and select the **All Schemas** view.
   c. Click the schema that you want to refresh.

   > Tip You can also use the Collibra Platform Self-Hosted search function to look up your schema.

2. In the view bar, to the right, click **Actions** → **Refresh**.
   » The **Refresh Schema** dialog box appears.

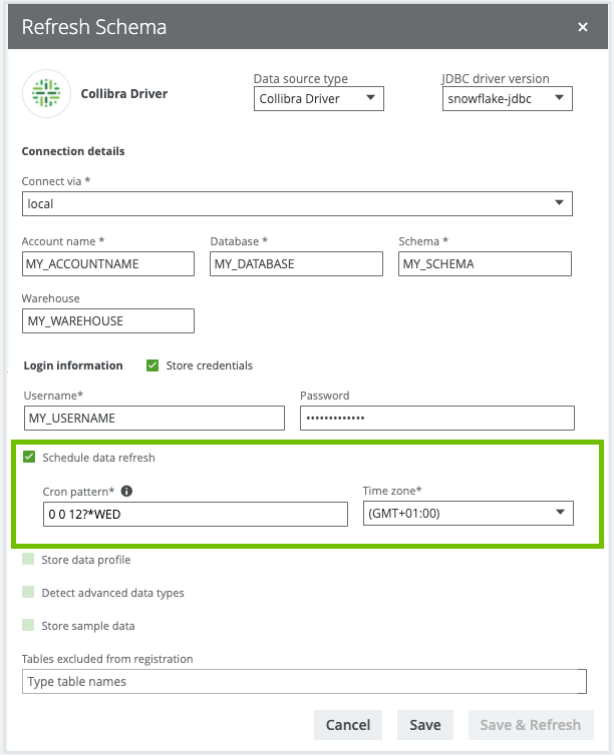
   > Tip If Catalog experience is disabled, the **More** menu is shown instead of **Actions**.

3. In the **Login information** section, check **Store credentials** and enter the username and password you use to access your data source.
   » Your credentials are used to automatically connect to your data source and refresh the metadata in Collibra Platform Self-Hosted.
4. Select **Schedule data refresh**.
5. Enter the required information.

Option	Description
Cron pattern	Schedule of the data refresh as a Quartz Cron pattern. Warning If you create an invalid Cron pattern, Collibra Platform Self-Hosted stops responding.
Time zone	The time zone of the database.

6. Click **Save**.

Example
When you refresh a schema of a registered data source, you can create a refresh schedule by selecting **Schedule data refresh**. You can then enter the CRON pattern *0 0 12?*WED* to refresh every Wednesday at 12:00:00 PM.



Remove a synchronized schema

You can remove a synchronized schema and the associated synchronized assets. To do so, you need to delete the Schema, Table and Column assets of the synchronized schema.

Required permissions

- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a global role with the View Edge connections and capabilities global permission, for example, Edge integration engineer.
- You have a resource role with the **Configure external system** resource permission, for example, Owner.

Steps

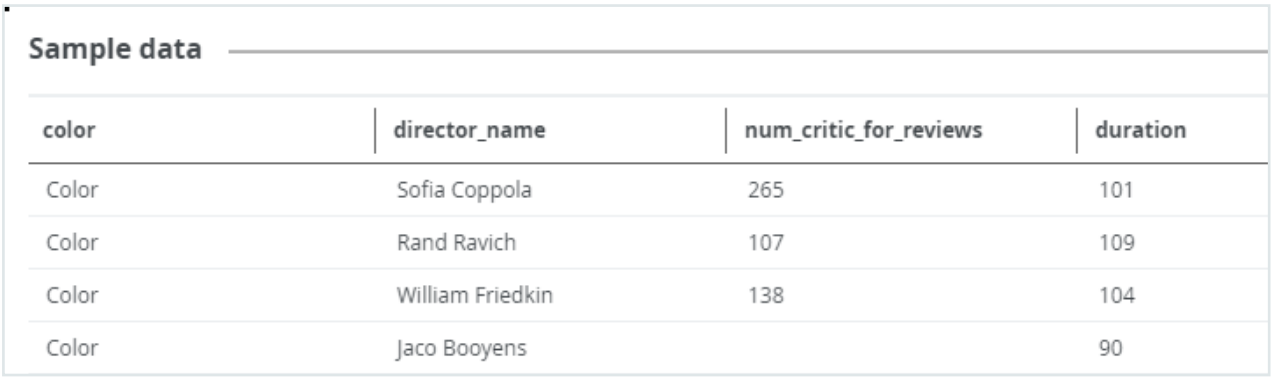
1. Look up the target domain of the synchronized schema:
   a. Open a Database asset page.
   b. In the tab pane, click **Configuration**.
   c. Select the synchronized schema and check the Target Domain.
2. Navigate to the target domain.
   A domain can contain content from more than one schema.
   a. If the domain contains only assets that can be removed, delete the domain.
   b. If the domain contains assets from other schemas as well, select the check-boxes of the relevant assets one by one and delete the assets.

   > Tip To know which assets below to a specific schema in a domain, you can create a filter, for example by using the asset name: "Full Name <starts with> {edgeConnectionName}>{databaseName}> {schemaName}".

3. Go back to the Database asset and click **Configuration**.
4. Click the **Refresh List** icon.
   » Once the refresh is completed, the schema will no longer have the check symbol ( ) and will show the status "Not Synchronized".

Sample data

Sample data is a set of randomly collected data from a data source. Sample data can be displayed for Table, Column, or Data Set assets. The purpose of showing sample data is to provide examples of the data so you know what to expect when you use the asset.



Sample data

color	director_name	num_critic_for_reviews	duration
Color	Sofia Coppola	265	101
Color	Rand Ravich	107	109
Color	William Friedkin	138	104
Color	Jaco Booyens		90

You can only view sample data for an asset:

- If the sample data feature is active.
- If you have the required permissions.
- If the asset is a Table, Column, or Data Set asset.

  > Note Currently, you can only request sample data via Edge for Table and Column assets.

- If sample data is available for the asset.

Sample data is available in:

Asset type	If Catalog experience is active, you can see the sample data in:	If Catalog experience is not active, you can see the sample data in:
Table	**Summary** tab pane **Sample data** tab pane	**Details** tab pane **Sample data** tab pane
Column	**Summary** tab pane **Data profiling** tab pane	**Details** tab pane **Sample data** tab pane
Data Set	**Summary** tab pane **Sample data** tab pane	**Details** tab pane **Sample data** tab pane

> Tip
> In Table and Data Set assets, you only see sample data for columns for which you have the required permission. If you do not have access, you see the text <sensitive> in the column instead of sample data.

The way Collibra handles sample data depends on how the assets are added in Collibra and how the sample data is collected:

	Assets are created by registering a data source via Edge.	**Assets are created by registering a data source via Jobserver.**	**Assets are manually added or imported.**
Sample data for an asset is uploaded via the Catalog REST API - Profiling.	The sample data is stored in the Collibra cloud repository. The sample data is displayed to all users with the required permissions.	The sample data is stored in the Collibra cloud repository. This sample data is also used for data classification via the Data Classification Platform. The sample data is displayed to all users with the required permissions.	The sample data is stored in the Collibra cloud repository. The sample data is displayed to all users with the required permissions.
Sample data is collected and stored when the data source is registered via Jobserver. See Configure the use of sample data via Jobserver.	Not applicable.	The sample data is stored in the Collibra cloud repository. This sample data is also used for data classification via the Data Classification Platform. The sample data is displayed to all users with the required permissions.	Not applicable.

	Assets are created by registering a data source via Edge.	Assets are created by registering a data source via Jobserver.	Assets are manually added or imported.
Sample data can be manually requested for an asset that is registered via the Edge register data source process.	The randomly collected sample data is cached on the Edge site for 24-48 hours. No sample data is stored in the Collibra cloud repository. The sample data is only displayed to users with the required permissions and if the sample data has been requested.	Not applicable.	Not applicable.

	Assets are created by registering a data source via Edge.	Assets are created by registering a data source via Jobserver.	Assets are manually added or imported.
	Note • Currently, you can only request sample data via Edge for Table and Column assets. • We randomly collect rows from the data source. The data of the randomly collected rows, however, is not switched around, we display all data for each randomly collected row.		

For details on the process, go to Understanding the process to display sample data.
For details on the sample data limitations and guidelines, go to Limitations and guidelines.

Required permissions to view sample data

To view sample data for an asset, you need:

- View permission on the asset.
  View permission is required to access the asset in general.
- Resource permission: **Asset > Data > View Samples**.
  View Samples permission is needed to see the sample data.

> Tip
> In Table and Data Set assets, you only see sample data for columns for which you have the required permission. If you do not have access, you see the text <sensitive> in the column instead of sample data.

Configuring the use of sample data

Sample data limitations and guidelines

- Sample data via Edge may require additional Edge site memory, CPU and disk space.
- Currently, you can only request sample data via Edge for Table and Column assets.
- For performance reasons, the number of samples to display must be less than 1,000.
  This limit is configurable in the **Maximum number of samples** setting, in the Data Profiling section. The default value is 100. The maximum value is 1,000.
  Go to Configure the use of sample data via Edge or Configure the use of sample data via Jobserver.
- For performance reasons, avoid sampling tables with more than 1,500 columns.
  This limit is not configurable at the moment.
- The sampling feature always uses push-down sampling if push-down sampling is available for the data source. Push-down sampling increases the sample data extraction speed.
  We advise to only allow sampling on data sources that support push-down sampling. To know if your data source allows for push-down sampling (called partial scan in Edge), go to Data sources supported by Edge or Overview of Collibra-provided JDBC drivers (Jobserver).

> Note If you try sampling on a data source that does not allow push-down sampling, the sample data extraction time is proportional to the database table size. The bigger the table, the longer it will take to retrieve the samples.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on sampling via Edge, go to Configure the use of sample data via Edge.

Configure the use of sample data via Jobserver

You must configure your Collibra environment if you want to display sample data for data sources registered via Jobserver.

	Configuration step	More details
1	Ensure the users have the required permissions.	Required permissions to view sample data

	Configuration step	More details
2	In the Service Configuration settings, • Set the **Data Profiling** setting **Maximum number of samples** to a value higher than 0. • Define the maximum number of characters that you want to collect per sample in **Maximum value length**.	Show how ## Prerequisites • You have the **ADMIN** or **SUPER** role in Collibra Console. • You have the **SUPER** role in Collibra Console. • You have the **ADMIN** or **SUPER** role in Collibra Console. ## Steps 1. Open the DGC service settings for editing: a. Open Collibra Console. » Collibra Console opens with the **Infrastructure** page. b. In the tab pane, expand an environment to show its services. c. In the tab pane, click the Data Governance Center service of that environment. d. Click **Configuration**. e. Click **Edit configuration**. 2. Open the DGC service settings for editing: a. Open Collibra Console. » Collibra Console opens with the **Infrastructure** page. b. In the tab pane, expand an environment to show its services. c. In the tab pane, click the Data Governance Center service of that environment. d. Click **Configuration**. e. Click **Edit configuration**.

	Configuration step	More details
		3. Go to the **Data profiling** section. 4. Make sure the setting **Maximum number of samples** is higher than 0. The default value is 100. The maximum value is 1,000. For more information, go to DGC service configuration: options. 5. In **Maximum value length**, define the maximum number of characters that you want to collect per sample. We don't recommend increasing this number as it may affect the stability of the system. 6. Click **Save all**.
3	When you register or refresh the data source via Jobserver, select the option **Store Sample Data**.	Register a data source via Jobserver

For detailed information on the sample data process, go to Understanding the process to display sample data.

Add the Catalog JDBC Sampling capability

After you have configured the settings for sample data, and you have created a JDBC Edge connection for your data source, you need to add the Catalog JDBC Sampling capability to the connection.

- The **Catalog JDBC Sampling** capability consists of two possible operations:
  - Extracting the sample data, which collects the data from a data source and caches the data on the Edge site.
  - Reading the sample data, which reads sample data from the Edge cache and returns it as a result of an API call or displays it in an asset page.
- Once the capability is selected, define the JDBC connection to which the capability applies.

Before you begin

- You have created and installed an Edge site.
- You have created a JDBC connection for your data source.

Required permissions

- You have a global role that has the **System administration** global permission.
- You have a global role that has the **Manage connections and capabilities** global permission, for example, Edge integration engineer.
- You have a global role that has the **Register profiling information** global permission.

Steps

1. Open an Edge site.
   a. On the main menu, click , and then click **Settings**.
      » The Collibra settings page opens.
   b. In the tab pane, click **Edge**.
      » The **Sites** tab opens and shows a table with an overview of the Edge sites.
   c. In the table, click the name of the Edge site whose status is **Healthy**.
      » The Edge site page opens.
2. In the **Capabilities** section, click **Add capability**.
   » The **Add capability** page is shown.
3. Enter the required information.

Field	Description	Required
Capability	This section contains general information about the capability.	
Name	The name of the Edge capability.	✓ Yes
Description	The description of the Edge capability.	✗ No

Field	Description	Required
Capability template	The capability template. The value that you select in this field determines which sections appear on the page. Select the following Edge capability: `Catalog JDBC Sampling`	✓ Yes
Connection	This section contains information to connect to the data source.	
JDBC connection	The connection to the data source.	✓ Yes
General	This section contains general information about logging.	
Debug	An option to automatically send Edge infrastructure log files to Collibra Platform Self-Hosted. By default, this option is set to *false*. > Note We highly recommend to only send Edge infrastructure log files to Collibra Platform Self-Hosted when you have issues with Edge. If you set it to *true*, it will automatically revert to *false* after 24h.	✗ No
Log level	An option to determine the verbosity level of Catalog connector log files. By default, this option is set to *No logging*.	✗ No

4. Click **Create**.
   » The capability is added to the Edge site.
   » The fields become read-only.

What's next?

With the correct settings and permissions in place, users can start requesting sample data for the data source.

Delete sample data

The way to remove sample data for a data source depends on how the sample data is made available.

- For Jobserver, perform one of the following:
  - Refresh the related schema and don't select the **Store Sample Data** checkbox.
    As a result, any previously gathered sample data is removed from the Collibra cloud repository.
  - Call the Catalog profiling REST API with an empty array for the samples parameter.
    As a result, any previously gathered sample data is removed from the Collibra cloud repository.
- For Catalog profiling REST API, call the Catalog profiling REST API with an empty array for the samples parameter.
  As a result, any previously gathered sample data is removed from the Collibra cloud repository.
- For Edge, you cannot delete sample data.
  Sample data for data sources registered via Edge is not stored in the Collibra cloud repository, it is cached on the Edge site. Every day, Edge deletes all sample data that is older than 24 hours from its cache.

  > Note If a data source was previously connected to Jobserver or if sample data was pushed using the Catalog profiling REST API, and the data source is now an Edge data source, sample data may still be stored in the Collibra cloud repository for this data source. If you want to remove this sample data, call the Catalog profiling REST API with an empty array for the samples parameter.

Example of API code that deletes sample data from the Collibra cloud repository

In the example code:

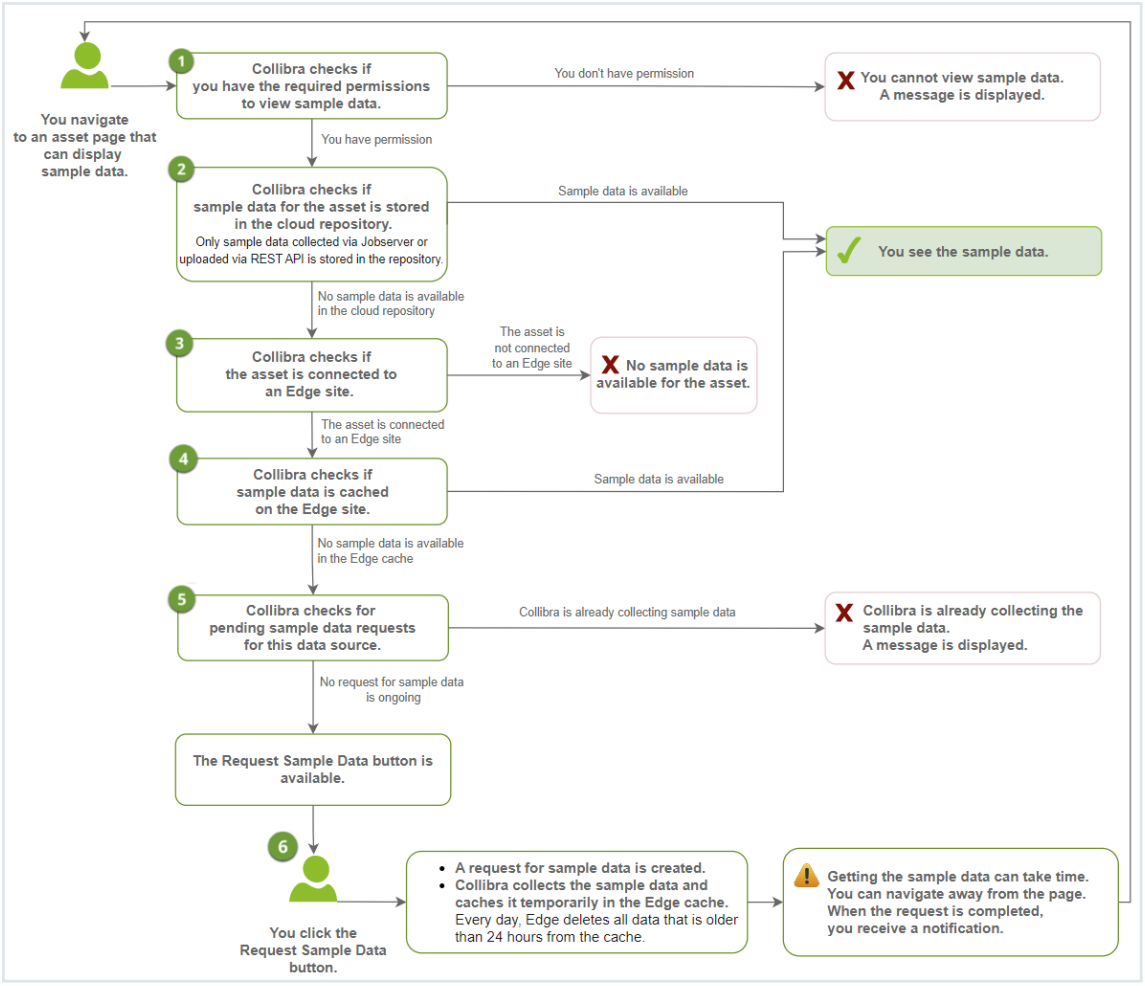
- replace <your_environment> by the name of your environment.
- replace the assetIdentifier section by any combination that uniquely identifies the asset for which you want to delete the sample data.

Example

```
PATCH https://<your_
environment>.collibra.com/rest/catalog/1.0/profiling/columns
{
   "columnProfiles": [
      {
         "assetIdentifier": {
         "assetName": "Catalog postgresql>catalog_
         postg>GDPR>Consumers>Process_id(column)",
         "communityName": "Catalog demo",
         "domainName": "Catalog postgresql > catalog_postg >
         GDPR"
         },
         "samples": {
         "samples": null
         }
      }
   ]
}
```

Understanding the process to display sample data

If you open a Column, Table or Data Set asset page, Collibra performs a series of checks to determine if sample data is displayed.



> **Note**
> Currently, you can request sample data via Edge only for Table and Column assets.

Check or Action	Description	Positive outcome	Negative outcome
1	Collibra checks if you have the required permissions to view sample data.	You have the required permissions: • The process continues with the next check.	You don't have the required permissions: • You cannot see the sample data and a message appears on the page. • The process stops.

Check or Action	Description	Positive outcome	Negative outcome
2	Collibra checks if sample data is stored in the Collibra cloud repository. Tip This is only possible if: • Sample data was extracted during the registration of the data source via Jobserver. • Sample data was uploaded by using the Catalog REST API - Profiling.	Sample data is available in the Collibra cloud repository: • The sample data is visible in the page. • The process stops.	No sample data is available in the Collibra cloud repository: • The process continues with the next check.
3	Collibra checks if the asset is connected to an Edge site. An asset is connected to an Edge site when the asset has been registered via the Edge Catalog data source registration process. Only adding the Catalog JDBC Sampling capability to your Edge site is not enough. The asset is connected to the Edge site via its related Database asset.	The asset is connected to an Edge site: • The process continues with the next check.	No sample data is available for the asset: • You cannot see the sample data. • The process stops.

Check or Action	Description	Positive outcome	Negative outcome
4	Collibra checks if sample data is available in the cache of Edge. This is possible if sample data has been requested before and the cache has not been cleared in the meantime. Once a day, Edge deletes all data that is older than 24 hours from its cache.	Sample data is available in the cache: • The sample data is visible in the page. Important It can take some time for the sample data to be displayed. Don't navigate away from the page while the process is ongoing. • The process stops.	No sample data is available in the cache: • The process continues with the next check.
5	Collibra checks if a sample data request is pending for the data source.	A sample data request is pending for the data source: • You need to wait until the sample data has been collected and cached in the Edge site. • The process stops.	No sample data request is pending for the data source: • The button **Request Sample Data** appears on the page and in the **Action** drop-down list. • The process stops until you click the button.

Check or Action	Description	Positive outcome	Negative outcome
6	You click the **Request Sample Data** button.	• A 'Request sample data' job is launched and added to the Activities list and job queue. • When Edge is ready for the job, the job starts. • Sample data is randomly collected and temporarily made available in the cache of the Edge site. We randomly collect rows from the data source. The data of the randomly collected rows, however, is not switched around, we display all data for each randomly collected row. If you request sample data for a column, sample data is collected and cached for the entire table. Every day, Edge deletes all data that is older than 24 hours from its cache. This means the sample data remains available between 24 and 48 hours. • Because it can take time for the job to start and complete, you can navigate away from the page while the job is in progress. • Once the job is completed, you receive a notification. • The process stops. Note Columns mapped to following java.sql.Types are excluded from the sampling queries: ARRAY, BINARY, BLOB, CLOB, DATALINK, DISTINCT, JAVA_OBJECT, LONGVARBINARY, NCLOB, NULL, OTHER, REF, REF_CURSOR, ROWID, SQLXML, STRUCT, VARBINARY.	

Troubleshooting sample data

Tip Make sure your Edge site meets the Sample data requirements. For information, go to Edge hardware requirements to show sample data.

Message: To ensure data security, sample data is currently not visible

Issue:
When you open sample data for an asset, no data is displayed and you see the following message in the page: `To ensure data security, sample data is currently not visible.`

Possible reasons:

- You don't have the required permissions to view sample data.
- The **Catalog JDBC Sampling** capability has not been defined for the data source Edge connection.

Solution:

- Request the required permissions.
- If the sampling capability is missing, add the **Catalog JDBC Sampling** capability for the data source.

Message codes

Code	Description	Possible reasons	Solution
200	This code indicates the sample data processes ran correctly. Also when no sample data is available in the data source, this code is provided.		

Code	Description	Possible reasons	Solution
400	This message appears if: • Something is wrong with the provided asset ID or • The sampling capability is not installed on the Edge site. The error message will specify the problem.	• The asset exists but the asset is not a column or table • The table has no columns. • Something is wrong in the relationship of the column, table or database, like a column asset that was not ingested but manually created and no relationship has been defined. • The **Catalog JDBC Sampling** capability has not been defined for the data source Edge connection.	• If it concerns a wrong asset, provide a valid column or table asset id. • If the sampling capability is missing, add the **Catalog JDBC Sampling** capability for the data source.
401	This message appears if you are not authenticated to use the sampling API.	The authentication failed.	Provide valid credentials.
403	This message appears if you lack permission to any of the columns within the requested asset.	You do not have the required permissions. Both View permission and View Samples permission are needed to see sample data for an asset.	Verify the user has the required permissions.
404	This message appears if the asset cannot be found.	The asset does not exist.	Provide an existing column or table asset id.

Code	Description	Possible reasons	Solution
503	This message appears if the Edge service gets a timeout or fails.	The Edge service is not available.	Verify that the Edge site is still online and healthy. If not, check the Edge logs to get a better understanding of the issue. If the problem persists, contact Collibra Support for assistance.

Error message: Generic API exception PayloadTooLarge

Issue:
When you open sample data for a Table asset, you receive the following message:`com.collibra.edge.management.exceptions.PayloadTooLarge: The payload size should be below 102400 bytes.`

Reason: The table contains too many columns.

Solution: You can open an individual Column asset to request its sample data.

Error message: There is no matching sampling capability found

Issue:
You receive the following error message:
`There is no matching sampling capability found for connection [connection_id].`

Reason:
This message appears when you open a Column or Table asset page for a data source that has been registered via Edge but for which the Edge site doesn't have an associated Edge capability for sampling.

Solution:
To solve the issue, add the **Catalog JDBC Sampling** capability for the data source. The message provides the id of the Edge connection linked to the data source.

No sample data is displayed

There are many conditions that can result in no sample data being displayed. Before reporting an issue, check the following:

Reason	Description	Solution
The setting **Maximum number of samples** is set to 0.	The sampling feature is disabled and no samples are displayed.	Set the **Data Profiling** setting **Maximum number of samples** to a value higher than 0. For details, Configuring the use of sample data.
The sampling capability is missing for your Edge data source.	Samples can only be extracted if the sampling capability is set for the data source on the corresponding Edge site.	Add the **Catalog JDBC Sampling** capability for the data source.
The asset for which you want to collect sample data has no data.	There is no data to show for the asset.	
No sample data is stored in the Collibra cloud repository. (not applicable for data sources registered via Edge)	• For Jobserver data sources, sample data is only available in the Collibra cloud repository if the **Store Sample Data** option was selected during the registration of the data source. • For assets created without Jobserver or Edge registration, sample data is only available if they were uploaded to the Collibra cloud repository via the Catalog Profiling REST API.	Configuring the use of sample data

You do not see the Request Sample Data button

Issue:
You want to see sample data for an asset but you cannot request it. The **Request Sample Data** button is not available.

Reason:
The possible reasons are:

- You don't have the required permissions.
- This asset was not created via the Edge Catalog data source registration process.
- There is no sample data available in the data source. In that case, you see an empty table.

I cannot request sample data for a data set via Edge

Currently, you can only request sample data via Edge for Table and Column assets.

You always get old sample data for a data source registered via Edge

Issue:
You always see old sample data for a data source registered via Edge.

Reason:
Sample data stored in the Collibra cloud repository takes precedence over sample data extraction by Edge. Sample data can be available for an Edge data source in the Collibra cloud repository if this data source was previously connected to Jobserver or if sample data was pushed using the Catalog profiling REST API for the data source. For more information on the process, go to Understanding the process to display sample data.

Solution:
If you want to remove samples from the Collibra cloud repository, go to Delete sample data.

Collecting the sample data is very slow via Edge

- It can take some time to read and display the sample data available in the Edge cache.
- The sample data extraction time via Edge is influenced by multiple factors. For example: table size, number of columns in a table, number of samples to collect, maximum length of samples, and push-down sampling mechanism available for the data source. For more details, go to Sample data limitations and guidelines.
- Maybe a lot of parallel sample data requests are ongoing. This happens when a lot of users want to see sample data at the same time.

  > Tip If you experience issues in this situation, you can decrease the number of Edge data sources for which the sampling capability is enabled.

You want to retrieve sample data log files

For data sources registered via Edge, Edge logs are generated when sample data is extracted from the data source and cached on the Edge site. The logs start with this text: "Writing cache samples with the key...".
Looking at the Edge logs within a 2-day period should give information on the sampling activity.

> Example
> ```
> Writing cache samples with the key
> 'catalog.sample.6385e23cb1ae443a7786c555108d8bb028d23dee39e76ce
> 3169eaa9cdacb1ed3'
> "Cache write sample for table 'Snowflake>SNOWFLAKE_SAMPLE_
> DATA>TPCDS_SF100TCL>CALL_CENTER'
> ```

Quality extraction

The quality extraction functionality allows you to ingest Collibra Data Quality & Observability user-defined rules, metrics, and dimensions into Collibra Data Catalog for registered data sources by using the DQ Connector Edge capability.

> Warning You can use this functionality only if you have the same data sources registered in Collibra Collibra Data Quality & Observability and you have Edge enabled in your Collibra Platform Self-Hosted.

About DQ Connector

The native DQ Connector brings Collibra Data Quality & Observability into your Collibra Platform Self-Hosted. The DQ Connector is an Edge capability template that helps you integrate your Collibra Data Quality & Observability user-defined rules, metrics, and dimensions into Collibra Data Catalog.

> Note To extract data quality statistics from CollibraCollibra Data Quality & Observability both Data Catalog and Collibra Data Quality & Observability must ingest the same data source.

DQ Connector requirements

- Collibra Platform Self-Hosted 2021.07 or newer.
  - Edge
- Collibra Collibra Data Quality & Observability 2.15 or newer.
  - Existent data quality statistics for the selected data source.

DQ Connector configuration

1. Connect to a Collibra Data Quality & Observability source:
    a. Create a Collibra Data Quality & Observability Edge site.
    b. Connect to your Collibra Data Quality & Observability source.
    c. Add ingestion capabilities to your Collibra Data Quality & Observability connection.
    d. Configure destinations for Collibra Data Quality & Observability assets.
    e. Add Collibra Data Quality & Observability characteristics to assets.
    f. Add a DQ Connector capability.
2. Register Collibra Data Quality & Observability Edge connections in Data Catalog:
    a. Create a Data Catalog System Asset.
    b. Register the Collibra Data Quality & Observability data source in Data Catalog.

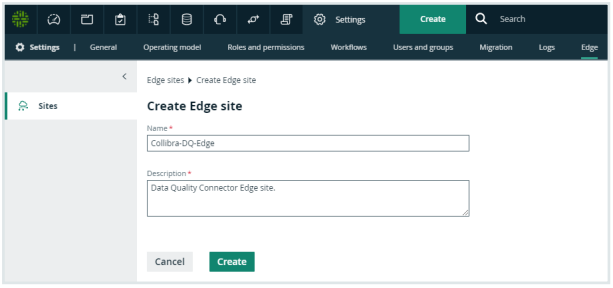
Connect to a Collibra Data Quality & Observability source

Because the DQ Connector is an Edge capability, you must be able to ingest data via Edge. For information about enabling and configuring Edge, see the Edge Configuration guide.

Create a Collibra Data Quality & Observability Edge site

Create an Edge site with the following properties:

Field	Description
Name	The name of the Edge site, for example **Collibra-DQ-Edge**. Do not use spaces or special characters. This field is mandatory and the name must be globally unique.
Description	The description of the Edge site. We recommend to put at least basic location information of the Edge site. This field is mandatory.



Install the Collibra Data Quality & Observability Edge site

Follow the instructions for your environment to Install an Edge site.

> Note This process automatically creates an Edge user, which you use later in the setup process.

Connect to your Collibra Data Quality & Observability source

Create a connection for each CollibraCollibra Data Quality & Observability data source you want to synchronize.

Section	Property	Value
Connection settings	Name	The same name as the Collibra Collibra Data Quality & Observability connection name.
	Description	The description of the JDBC connection. This field is also visible when you register content.
	Connection provider	The connection provider, which determines the available connection parameters. Same as Collibra Collibra Data Quality & Observability.

Section	Property	Value
Connection parameters Example for **Username / Password JDBC drive**	Username	The same username as the Collibra Collibra Data Quality & Observability connection username.
	Password	The same password as the Collibra Collibra Data Quality & Observability connection password.
	Driver class name	The same driver name as the Collibra Collibra Data Quality & Observability connection driver name.
	Driver Jar	The same driver JAR file as from Collibra Collibra Data Quality & Observability.
	Connection string	The same URL as the Collibra Collibra Data Quality & Observability connection URL.

Add ingestion capabilities to your Collibra Data Quality & Observability connection

You must add a Catalog JDBC ingestion Edge capability template for each connection you have created to extract and process data for your data source.

Field	Description	Required
Capability	This section contains general information about the capability.	
Name	The name of the Edge capability.	✓ Yes
Description	The description of the Edge capability.	✕ No

Field	Description	Required
Capability template	The capability template. The value that you select in this field determines which sections appear on the page. Select the following Edge capability: `Catalog JDBC ingestion`	✓ Yes
Connection	This section contains information to connect to the data source.	
JDBC connection	The connection to the data source.	✓ Yes
JDBC data source type (Deprecated)	Deprecated field. The field was used to indicate the type of the data source. You no longer need to change this field. The required value is automatically identified. Note The automatically identified value is not shown in this page.	✓ Yes
Supports schemas	A text field where you have to enter *True* to enable database registration of data sources that have no schema. If the data source has schemas, you can ignore this field. Tip If the data source does not have a schema, Data Catalog creates a Schema asset with the same name as the full name of the database.	✗ No

Field	Description	Required
Others	This section can contain additional capability properties. Warning Adding additional properties can have a significant impact on your Edge site. Only add or update them together with Collibra Support.	✕ No

Field	Description	Required

Field	Description	Required
General	This section contains general information about logging.	
Debug	An option to automatically send Edge infrastructure log files to Collibra Platform Self-Hosted. By default, this option is set to *false*. Note We highly recommend to only send Edge infrastructure log files to Collibra Platform Self-Hosted when you have issues with Edge. If you set it to *true*, it will automatically revert to *false* after 24h.	✕ No
Log level	An option to determine the verbosity level of Catalog connector log files. By default, this option is set to *No logging*.	✕ No

Configure destinations for Collibra Data Quality & Observability assets

Collibra Data Quality & Observability rules, metrics and dimensions require their own domains in Data Catalog. If you don't have existing domains for data quality or wish to use new ones for the quality extraction purpose, create a domain for each type of data quality asset:

- Rules: **Rulebook Domain**
- Metrics: **Business Asset Domain**
- Dimensions: **Business Asset Domain**

Assign permissions for Collibra Data Quality & Observability domains

Edge must have the correct resource permissions to manage assets inside the dedicated Collibra Data Quality & Observability domains. For each dedicated domain, assign the **Technical Steward** role to the Edge user.

> Note The Edge user is automatically created when you install the Edge site.

Add Collibra Data Quality & Observability characteristics to assets

To show Collibra Data Quality & Observability statistics for your data source, assign the following characteristic types to the **Table** and **Column** asset types:

Asset type	Characteristic type
Table	governed by Governance Asset
Column	is governed by Data Quality Rule

Add a DQ Connector capability

The DQ Connector facilitates the communication with Collibra Collibra Data Quality & Observability. Add a DQ Connector capability to your Collibra Data Quality & Observability Edge site:

Field	Description	Required
Capability	This section contains general information about the capability.	
Name	The name of the Edge capability.	✓ Yes
Description	The description of the Edge capability.	✗ No
Capability template	The capability template. The value that you select in this field determines which sections appear on the page. Select the following capability template to ingest Collibra Data Quality & Observability user-defined rules, metrics, and dimensions into Collibra Data Catalog: `DQ Connector`	✓ Yes
DQ	This section contains information about the Collibra Data Quality & Observability connection.	
Base URL	Your Collibra Data Quality & Observability URL	✓ Yes

Field	Description	Required
Username	The Collibra Data Quality & Observability username for this connection.	✓ Yes
Password	The Collibra Data Quality & Observability password for this connection.	✓ Yes
Encryption options	Select the type of encryption to use. Default: *To be encrypted by Edge management server*.	
Issuer of the JWT	If you have selected *Encrypted with public key*, enter your JWT issuer.	✗ No
Collibra metadata model	This section contains information about where to ingest Collibra Data Quality & Observability assets.	
DQ Rules domain id	The UUID of the **Rulebook Domain** for the ingested Collibra Data Quality & Observability rules.	✓ Yes
DQ Metrics domain id	The UUID of the **Rulebook Domain** for the ingested Collibra Data Quality & Observability metrics.	✓ Yes
DQ Dimensions domain id	The UUID of the **Governance Asset Domain** for the ingested Collibra Data Quality & Observability dimensions.	✓ Yes
Default DQ Dimension name	The default **Data Quality Dimension**, for example *Accuracy*, *Completeness*, *Consistency* and so on. Default: *Completeness*.	✓ Yes
DQ Metric classified by DQ Dimension relation type id	The UUID of the **Data Quality Metric classified by / classifies Data Quality Dimension** relation. If left unspecified, this relation will not be added.	✗ No
Assets are imported in batches of this size	The batch size of the ingestion. Default: *5000*.	✓ Yes

Next steps

- Register a Collibra Data Quality & Observability source in Data Catalog.

Register a Collibra Data Quality & Observability source in Data Catalog

To make the Collibra Data Quality & Observability metadata available in Collibra Data Catalog, you must register the data source for each Collibra Collibra Data Quality & Observability data source you want to synchronize.

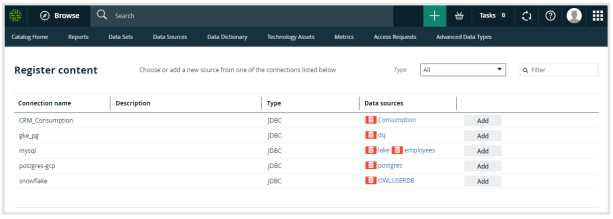
Create a Data Catalog System Asset

As a prerequisite to registering a data source in Data Catalog, you must create a System asset for each connected data source with the following properties:

Field	Value
Type	**System**
Domain	The domain to which the new assets will belong. You can only create a asset type in any domain of a domain type that is assigned to a selected asset type.
Name	The same name as the CollibraCollibra Data Quality & Observability connection name.

Register the Collibra Data Quality & Observability data source in Data Catalog

Register each Collibra Data Quality & Observability source in Data Catalog.



Next steps

- Extract Collibra Data Quality & Observability metadata.

Extract Data Quality metadata

After you completed the DQ Connector configuration, you can start ingesting Collibra Data Quality & Observability metadata.

Prerequisites

- You have configured the metadata synchronization properties for the data source.

Steps

1. Open a Database asset page.
2. In the tab pane, click **Configuration**.
3. In the **Quality extraction** section, do one of the following:
    - To select schemas for data quality synchronization:
        i. Click **Edit**.
            » The **Data quality** column becomes editable.
        ii. Select whether to synchronize the available schemas.
        iii. Click **Save**.
    - To synchronize the selected schemas:
        i. Select the schema name to see its configuration.
        ii. Click **Synchronize**.
            » The synchronization job is started for the selected schemas.

Chapter 4

Data profiling

About data profiling

Data profiling creates a summary of a data source that is registered with Data Catalog and determines the data type of columns in the data source. The summary mainly contains statistics and graphics to give the user an idea what the registered data is about.

You can create profiling results by:

- Registering a data source via Jobserver or via Edge, and choosing to profile the data.
- Importing profiling results via the Catalog API.

You can find the profiling results in Table and Column asset pages.

Profiling process

You can profile data via Edge or via Jobserver. The following table shows the differences.

Part of process	Profiling via Edge	Profiling via Jobserver
Data size	There is no data size limit. The Edge site calculates the profiling statistics while reading the data.	There is a limit on the size of the data that is used to calculate the profiling statistics. By default, this is 10 GB.
Connectivity	Collibra connects to an Edge site. The Edge site is installed in the customer's environment, close to the data source. The Edge site communicates to Collibra Platform Self-Hosted and other 3rd party systems using an HTTPS connection.	Jobserver requires an HTTP proxy to support reverse connectivity.
Register a data source	You can profile the data only after you registered a data source and synchronized one or more schemas via Edge. You can start the profiling process via the **Configuration** tab page on the Database asset page.	When registering a data source via Jobserver, options are available to profile the data and create sample data.

Part of process	Profiling via Edge	Profiling via Jobserver
Anonymizing data	Profiling happens on the Edge site. The profiling results are automatically anonymized for columns with the Text or Geo data type before they are sent to Data Catalog. It is not possible to disable the anonymization of these data types. An administrator can also decide to anonymize the profiling results for all columns.	Settings are available to enable the anonymization of the profiling results.
Classification	The classification process starts at the same time as the profiling of the data.	The classification process does not start together with the profiling.
Deleting data profiling results	Once data profiling results are available, you can only delete them by deleting the assets.	To delete data profiling results for a schema, refresh the schema without storing the data profile. Go to We have announced the end of life of Jobserver and all related Jobserver integrations for September 30, 2024, with the exception of Public Sector customers using GovCloud or on-prem environments. For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge..

Only using part of the data to create profiling results

Push down sampling (Jobserver) or **partial scan (Random Rows)** (Edge) means that the task of creating a set of data to profile is delegated to the data source itself and allows to only use part of the data to create profiling results.

- The data source randomly selects data to profile and transfers it to the Jobserver or the Edge site in one fetching process.
  If the Jobserver cache storage is reached, the fetching process can be stopped. Because the data source already created the data randomly, the omitted data can be ignored without lowering the representativeness of the data.
- Push down sampling or partial scan can be done using dynamic SQL query, if the data source supports it. For an overview, see Overview of Collibra-provided JDBC drivers.

Push down sampling or partial scan drastically increases the performance of collecting data to profile.

Push down sampling is not used by default on Jobserver. To use push down sampling, do the following:

Step	When	Description
1	Manage the driver	Add the **pushDownSampling** connection property.

Step	When	Description
2	Register your data source	Follow the usual steps to register a data source, but include the following options: 1. Enter a value for the **pushDownSampling** connection property. Note ◦ The value must be between *100* and *1 000 000*. Your data source creates the set of data to profile from that amount of rows. ◦ If the size of the amount of rows exceeds the limit of the cache storage (Collibra recommends 10 to 20 GB), the amount of rows is reduced. ◦ If you typed a value that is bigger than the amount of rows in the data source, the entire data source is used to create the profiling results. 2. Select **Store Data Profile** and, optionally, **Store Sample Data** to profile via Jobserver.

Random Rows is an option when you profile and classify a data source that allows partial scan. For details about the options, go to Profiling and classification options.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on profiling via Edge, go to Profiling via Edge.

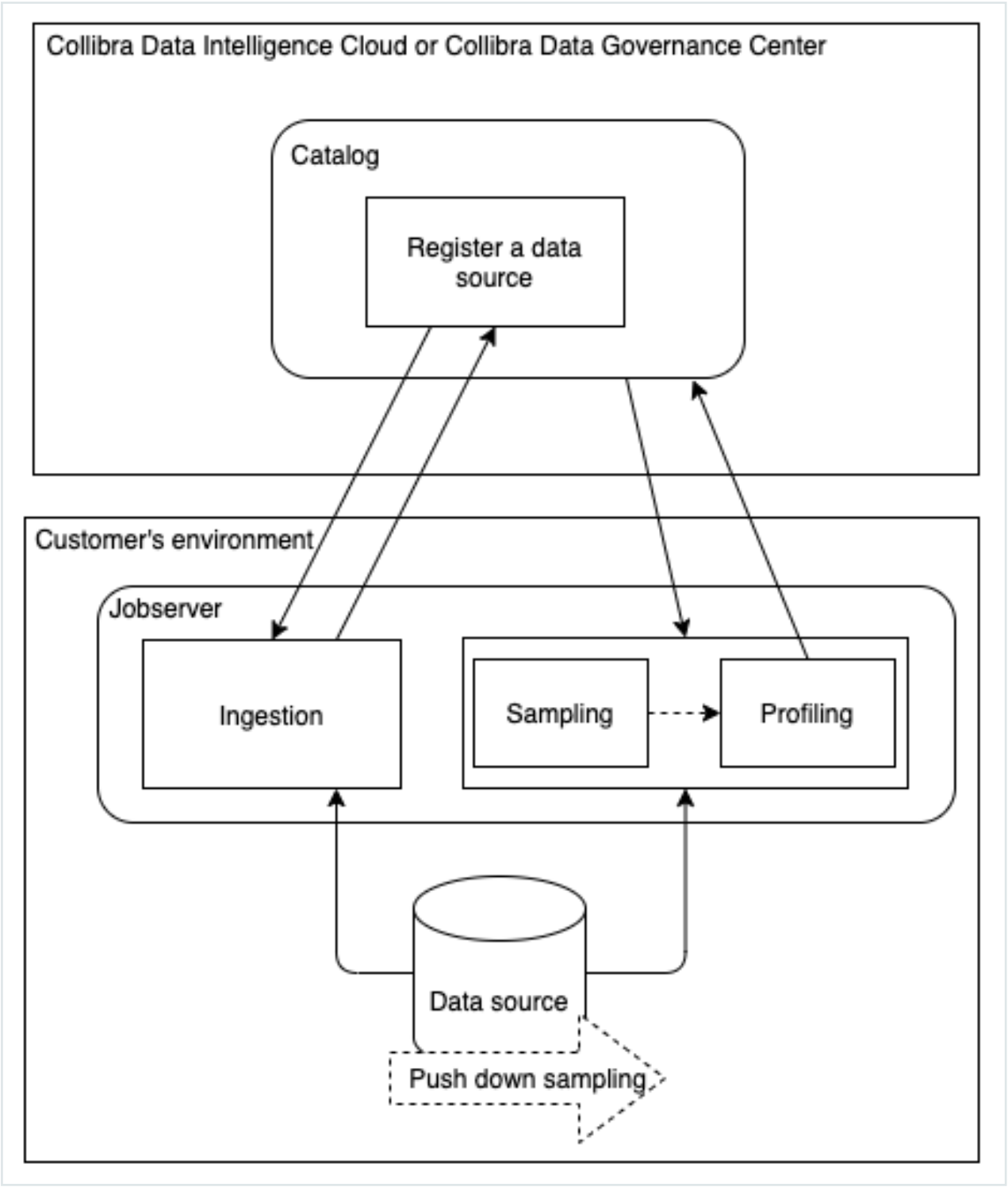
Profiling via Jobserver

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on profiling via Edge, go to Profiling via Edge.

About profiling via Jobserver

Profiling process via Jobserver

When you register a data source via Jobserver, Data Catalog triggers the ingestion process.
By default, the complete data set is transferred to Jobserver. Then Jobserver creates a representative subset of the data to profile, based on your data source. Jobserver then profiles that data and sends the profiling results to Data Catalog. You can enable the **Anonymize data** option to hash or remove profiling results that can be considered sensitive.


Collibra Data Intelligence Cloud or Collibra Data Governance Center
Catalog
Register a data source
Customer's environment
Jobserver
Ingestion
Sampling
Profiling
Data source
Push down sampling

Data used to create profiling results via Jobserver

To create the profiling results, Data Catalog uses a representative set of the data from the data source.

> Note This data is not the same as the sample data that can be available for an asset.

If you register a data source via Jobserver, the data that will be used by data profiling is created when you register the data source.

- If you use Jobserver without push down sampling:
  First, the complete data set is transferred to Jobserver. Then Jobserver creates the set of the data to be profiled. This is sometimes called sampling.
  The size is determined by the **Table profiling data size** setting in Collibra Console or the Services Configuration section of the Collibra settings. By default, the size is 10 GB.
- If you use Jobserver with push down sampling (also called partial scan):
  The data source itself creates the set of data to profile and sends it to Jobserver. The data source creates the set of data from randomly selected rows. If the Jobserver cache storage is reached, the process stops. Because the data source already created the set of data randomly, the omitted data can be ignored without lowering the representativeness of the sample.

  > Warning Push down sampling can be done using dynamic SQL query, if the data source supports it. To verify if your data source allows push down sampling, see Collibra-provided JDBC drivers.

  > Tip Push down sampling drastically increases the performance of collecting data to profile.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on profiling via Edge, go to Profiling via Edge.

Anonymization of data via Jobserver

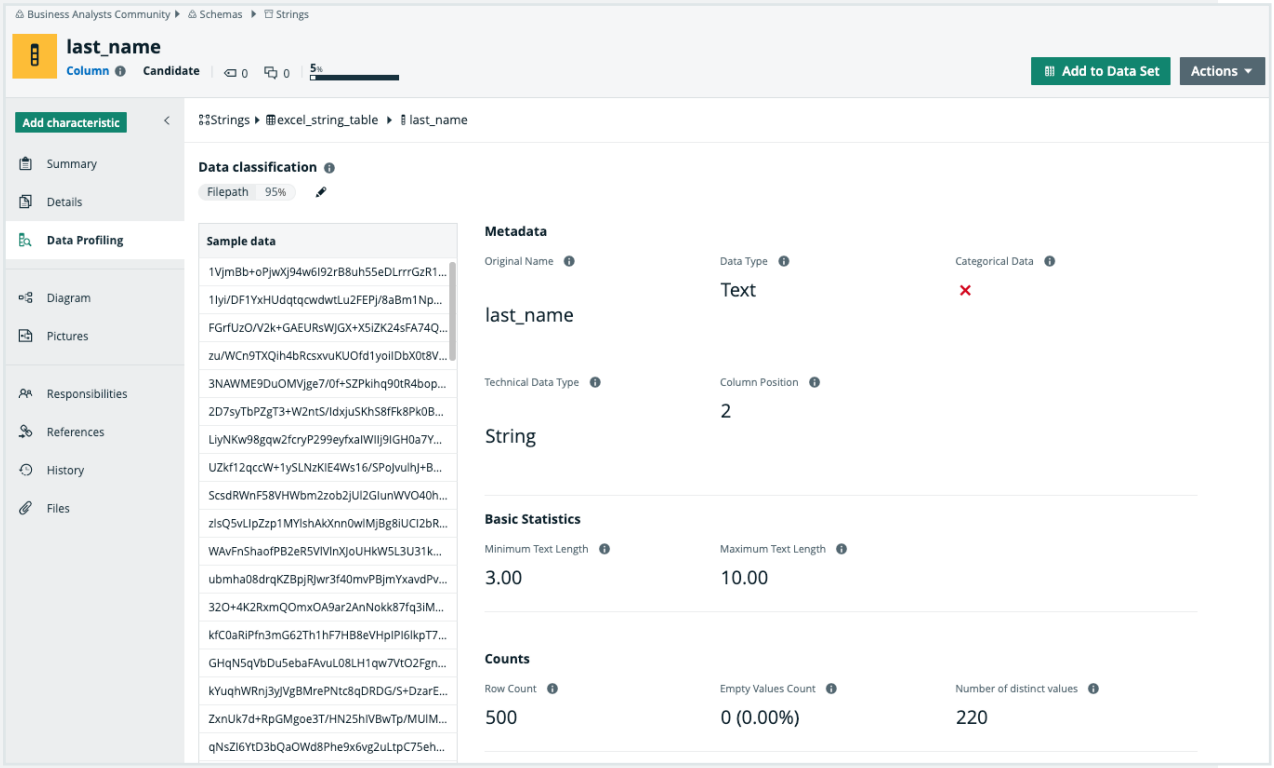
> Note Jobserver does not automatically anonymize sample data and profiling results.

To ensure that your sensitive data is not stored in the cloud, you must enable the Anonymize data option in Collibra Console. This option is by default disabled.
With this option enabled:

- Collibra anonymizes the content of columns with data of the type Text and Geo immediately at the end of the profiling process. As a result, data samples and the values that are shown in the data distribution charts are replaced by a random hash value for columns that contain these data types. Attributes that could contain sensitive data, like attributes of the type Mode or Percentiles, are no longer calculated for columns with data type Text or Geo.
- Identical values in a column get the same hash value so that you can still recognize the values as identical.

> Note
> Collibra detects the data type of a column during profiling and only anonymizes the data if the data type attribute is Text or Geo. However, if Collibra detects a data type that does not correctly correspond with the actual data type, some data may not have been anonymized or may have been wrongfully anonymized. To solve this, you can manually modify the column's data type and profile again.

Example You have enabled the Anonymize data option in Collibra Console for Jobserver and have profiled a column that has data type Text. If you go to the **Summary** or **Data Profiling** tab, all textual and geographical data has been removed or replaced by hashed values:



Warning Currently, if you enable the data anonymization process you can no longer use automatic data classification via the Data Classification platform. However, you can still classify and anonymize profiling results if you use Edge.

Tip If you profile and classify via Edge, the profiling results for columns with data type Text or Geo are automatically anonymized. You can anonymize all columns by enabling the Anonymize Edge profiling results for all data types feature.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Data Profiling Result dialog box with Jobserver

When you registered a data source via Jobserver, and you click the **Result** button of a data source registration activity, the **Data Profiling Results** dialog box opens.
A data source registration activity can be:

- Creating schema from JDBC
- Creating schema from file
- Updating JDBC schema
- Updating Excel schema
- Updating CSV schema

The **Data Profiling Results** dialog box contains the following information:

Item	Description
Schema	Name of the schema as added to Collibra Platform Self-Hosted.
Status	Status of the data source registration job.
Start time	Date and time when the data source registration job has started.
End time	Date and time when the data source registration job has completed.
Duration	Elapsed time of the data source registration job.
Ingestion Details	Summary of the job, including error messages and the list of tables and columns that have been ingested.
Profiling Details	The number of tables that have been correctly profiled.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on profiling via Edge, go to Profiling via Edge.

Modify the column data type of registered data

When Collibra Platform Self-Hosted creates a data profile of registered data, it detects the data type of each column. It's possible that Collibra detects a data type that does not correctly correspond with the actual data type, for example the Text data type is detected for a column, but the actual data in the column are dates.

For more information about the data type detection, see We have announced the end of life of Jobserver and all related Jobserver integrations for September 30, 2024, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements..

You can update the data type of each column to ensure that the data is properly managed in Collibra.

> Note If you use the Jobserver to register a data source and you have enabled the Anonymize data option in Collibra Console, Collibra detects the data type of a column during profiling and only anonymizes the data if the data type attribute is Text or Geo. Other data types are not anonymized. If you use Edge to register a data source, these columns are automatically anonymized.

Prerequisites

- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a resource role with the Attribute > Update resource permission.

Steps

There are two ways to modify a column's data type:

- In the data sources table.
- On the Column's asset page.

In the data sources table

1. On the main menu, click ⠿, and then click Catalog.
   » The Catalog Home opens.
2. In the submenu, click **Data Sources**.
3. Add the **Data Type** column to the table.
4. Expand the schema and table to see the columns.
5. Double-click in the **Data Type** column and choose the correct data type.
6. Click ✔ to apply the change.

On the Column asset page

1. Look up the column via the **Search** function.

   > Tip If you don't know the exact name of the column name, you can find it via **Data Catalog** → **Data Dictionary** and select the **All Schemas** view. Then click the schema that contains the column and click the column whose data type you want to update.

2. In the tab pane, click **Data Profiling**.
3. In the **Metadata** section, double-click the value of the **Data Type** parameter.
4. Select the correct type from the list.
5. Click **Save**.

When you refresh the schema, this change is not overridden.

Enable profiling for Edge

To enable Edge profiling and classification of synchronized metadata in Data Catalog, you need to run a command and enable multiple settings.

Before you begin

- You have enabled Database registration via Edge.
- You have run the command to enable classification on your Edge site.

Required permissions

- You have the **ADMIN** or **SUPER** role in Collibra Console.
- You have the **SUPER** role in Collibra Console.
- You have the **ADMIN** or **SUPER** role in Collibra Console.

Steps

1. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.
   c. In the tab pane, click the Data Governance Center service of that environment.
   d. Click **Configuration**.
   e. Click **Edit configuration**.
2. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.
   c. In the tab pane, click the Data Governance Center service of that environment.
   d. Click **Configuration**.
   e. Click **Edit configuration**.
3. In the **Data profiling** section, enter the required information:

Setting	Description
Database profiling via Edge	An option to enable profiling and classifying of synchronized metadata via Edge instead of Jobserver. ◦ ✔ True: Profiling and classification via Edge. ◦ ✖ False: Profile via Jobserver and classify via the Data Classification Platform. Note You can enable Database profiling via Edge only if you also enabled Database registration via Edge.

Setting	Description
Maximum duration of a profiling Edge job	The maximum time duration, in minutes, that a profiling Edge job can run before Data Profiling stops the job. The default value is 20,160 minutes, 2 days. You can increase this limit to a maximum of 4 days.
Parallel schema profiling via Edge	The maximum number of schemas that Edge can profile at the same time. By default, the value of this setting is 4. This means Edge processes four profiling jobs at a time. This can have a huge positive impact on the performance of the profiling activity. You can increase this number to a maximum of 16. Note ◦ If you increase this number to more than four jobs, make sure that your Edge site resources are aligned with the extra requests it will receive. ◦ If you decrease this number and the running number of jobs exceeds the limit, no job will be canceled. Instead, there won't be any room to schedule a new job until at least one running job is completed. Example The parallel schema profiling via Edge setting is set to 4. ◦ For 1 database that contains 3 schemas, we will process all 3 schemas at the same time. ◦ For 2 databases that contain 4 schemas in total, we will process all 4 schemas at the same time. ◦ For 1 database that contains 8 schemas, we will start with 4 schemas and then proceed to the next ones as soon as a job is completed.
Anonymize profiling data for all data types in Edge	Enable this option to anonymize all Edge profiling results stored in Collibra. ◦ ✓ True: Profiling results via Edge are anonymized for all columns. ◦ ✕ False (default): Profiling results via Edge are anonymized only for columns with the Text or Geo data type.

> Note
> - You don't need to enable the**Anonymize data (Jobserver)** setting because this setting is not relevant for Edge.
> - You don't need to enable the **Enable Data Classification** setting in the **Data Classification configuration** section. This setting relates only to the Data Classification Platform.
> If this setting is set to `true`, the **Classify** button is available on Column and Table asset pages. This button allows you to classify data via the Data Classification Platform. However, when using profiling and classification via Edge, you don't need the Data Classification Platform.

4. Click **Save all**.

What's next?

Continue the configuration for profiling and classification.

Data profiling information

If you create a data profile of registered data, profiling results are generated in the table and column assets.

- If you use Jobserver to register the data source, data profiling information depends on the profile options that you selected when you registered the data source.
- If you use Edge to register the data source, most information is only available after you specifically profiled the data. For an overview of the data that becomes available after the registration of a data source via Edge, see Data source registration information.

Column attribute	Profiling option (Jobserver)	Statistics	Description	Retrieved from JDBC property
Column Name	No option selected	N/A	The column name of the registered table.	COLUMN_NAME

Column attribute	Profiling option (Jobserver)	Statistics	Description	Retrieved from JDBC property
Data Type	Store Data Profile If you want to have Advanced Data Type detected, select Detect advanced data types	N/A	The data type of the column. This type is detected by the profiling process. This can differ from the **Technical Data Type** value. For example, if a database has a column with text as data type, and the column contains only integer values, the profiling process will set the *Whole Number* data type instead of text. If you enable the Anonymize data option in Collibra Console, Collibra anonymizes data in Column assets that have data type Text and Geo. If the profiling process has detected a wrong data type, you can update it afterwards. By default, Collibra anonymizes profiling results in Column assets that have data with the Text or Geo data type. However, is possible to anonymize the profiling results for all columns. For more information, go to Anonymization via Edge.	

Column attribute	Profiling option (Jobserver)	Statistics	Description	Retrieved from JDBC property
Description from Source	No option selected	N/A	The description of the column in the data source.	REMARKS
Row Count	Store Data Profile	Exact	The number of rows in the data source.	
Empty Values Count	Store Data Profile	Exact	The number of rows that are empty.	
Number of distinct values	Store Data Profile	Exact or approximate depending on column cardinality	The number of unique values in the column.	

Column attribute	Profiling option (Jobserver)	Statistics	Description	Retrieved from JDBC property
Chart	Store Data Profile	Depending on chart type	This column displays whether a chart was generated ( ) for the column or not (no icon available). If you hover over the icon, a preview of the chart appears. If you hover over a data point in the preview, extra data appears for the data point. The chart type varies per data type. Following charts available: • Frequency chart • Histogram that shows distribution • Probability distribution curve Note Charts are not available for the following data types: • Data type = Text and Categorical Data = false • Data type = Array • Data type = N/A	

Column attribute	Profiling option (Jobserver)	Statistics	Description	Retrieved from JDBC property
Frequency	Store Data Profile	Exact or approximate depending on column cardinality	A bar chart showing frequency data.	
Distribution - Histogram	Store Data Profile	Approximate	A histogram showing the representation of the distribution of numerical data.	
Distribution - Probability distribution curve	Store Data Profile	Approximate	A curve showing the representation of the probability distribution of numerical data.	
Technical Data Type	No option selected	N/A	Data type of the column as defined in the source. This value can differ from the **Data Type** value.	TYPE_NAME
Descriptive statistics (decile, percentile, quartiles)	Store Data Profile	Approximate	The value of the calculated statistic of the registered data.	
Categorical Data	Store Data Profile	Exact or approximate depending on column cardinality	Indication whether the data in the column is categorical or not. For example, if 100 000 rows are registered and there are only five distinct values, then the data is considered to be categorical.	

Column attribute	Profiling option (Jobserver)	Statistics	Description	Retrieved from JDBC property
Categories	Store Data Profile	Exact or approximate depending on column cardinality	List of detected categories. This column has only values if the data is categorical.	
Char octet Length	No option selected	N/A	Maximum number of bytes in a character type's column.	CHAR_OCTET_LENGTH
Column Position	No option selected	N/A	The index of the column in the source table.	ORDINAL_POSITION
Is Auto Incremented	No option selected	N/A	Indication whether the data in the column is auto-incremented or not.	IS_AUTOINCREMENT
Is Generated	No option selected	N/A	Indication whether the data in the column is generated or not.	IS_GENERATEDCOLUMN
Is Nullable	No option selected	N/A	Indication whether the column can store NULL values or not.	IS_NULLABLE
Is Primary Key	No option selected	N/A	Indication whether the column is a primary key or not.	True if the primary keys resultSet contains the COLUMN_NAME
Maximum Text Length	Store Data Profile	Exact	The length of the longest text value in the column, including white spaces.	

Column attribute	Profiling option (Jobserver)	Statistics	Description	Retrieved from JDBC property
Maximum Value	Store Data Profile	Exact	The maximum value in the column.	
Mean	Store Data Profile	Exact	The mean of all the values in the column, excluding empty rows.	
Median	Store Data Profile	Exact	The median value of the column.	
Minimum Text Length	Store Data Profile	Exact	The length of the shortest text value in the column.	
Minimum Value	Store Data Profile	Exact	The minimum value in the column.	
Mode	Store Data Profile	Exact or approximate depending on column cardinality	The value with the highest frequency for categorical data.	
Number Of Fractional Digits	No option selected	N/A	The number of fractional digits.	DECIMAL_DIGITS
Primary Key Name	No option selected	N/A	The name of the primary key composed by the column.	PK_NAME
Size	No option selected	N/A	The size of the column in the table.	COLUMN_SIZE
Standard Deviation	Store Data Profile	Exact	The statistical standard deviation of numeric values.	

Column attribute	Profiling option (Jobserver)	Statistics	Description	Retrieved from JDBC property
Variance	Store Data Profile	Exact	The statistical variance of numeric values.	
Sample	Store Sample Data	N/A	A random sample of the data set that represents the entire data set. Note In Edge, viewing sample data is not linked to the profiling feature. See sample data.	

Table attribute	Profiling option (Jobserver)	Statistics	Description	From JDBC property
Table Name	No option selected	N/A	The table name in the data source.	TABLE_NAME
Table Type	No option selected	N/A	The table type in the data source, such as TABLE or VIEW.	TABLE_TYPE
Description from Source	No option selected	N/A	The description of the table in the data source.	REMARKS

Data profiling of a table

The location of the data profiling information of a Table asset depends on the Catalog experience setting.

- If the setting is enabled, the information is displayed in **Columns** tab page.
- If Catalog experience is not enabled, the **Data Profiling** tab page displays the information.

The following profiling information is available by default:

- Name
- Data Type
- Row Count
- Empty Values Count
- Number of distinct values
- Chart

For information about these columns and columns that you can add, go to Data profiling information.

You can customize the table by clicking the Display options icon (⠿).
For example, to add more columns, click ⠿ → ✎ **Fields** and then click **Select fields**.

Data profiling of a column

In the **Data Profiling** tab of a Column asset, you can see the details of the column.

The details are grouped in some fixed sections:

Section	Content
Metadata	Contains the metadata of the column, such as data type, column name and so on.
Counts	Contains basic content information, such as number of rows and number of distinct values.
Basic Statistics	Contains the basic statistics of the data, such as minimum and maximum value.

Depending on the column's data type, you can find extra sections:

Section	Content
Quantiles	Contains the descriptive statistics of the data. This section is only available if the data type is numerical.
Categorical Data	Contains the values of the different categories. If there are too many values, only the first 50 and last 50 values are displayed.
Chart	Displays the statistics in a graphical way. The chart type varies per data type: • bar chart: textual, boolean, and numerical data that is considered categorical (Categorical Data = true). • data distribution chart: numerical and date and time data.

Note
If you use Jobserver, you can anonymize columns with data type Text or Geo by enabling the Anonymize data feature in Collibra Console.
If you use Edge, the profiling results for columns with Text or Geo data type are automatically anonymized. You can anonymize all columns by enabling the **Anonymize Edge profiling results for all data types** feature.

Data profiling charts

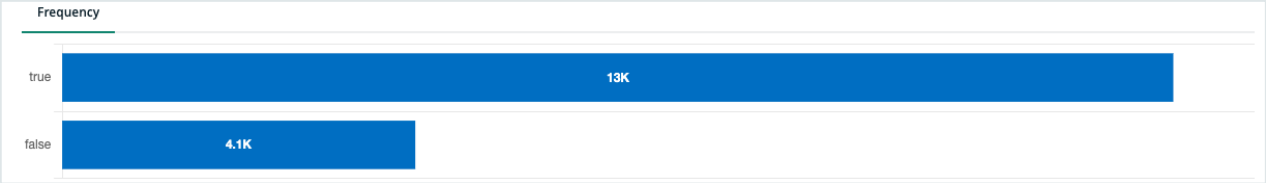
The data profiling process provides a view on the registered data by means of bar charts and distribution charts.

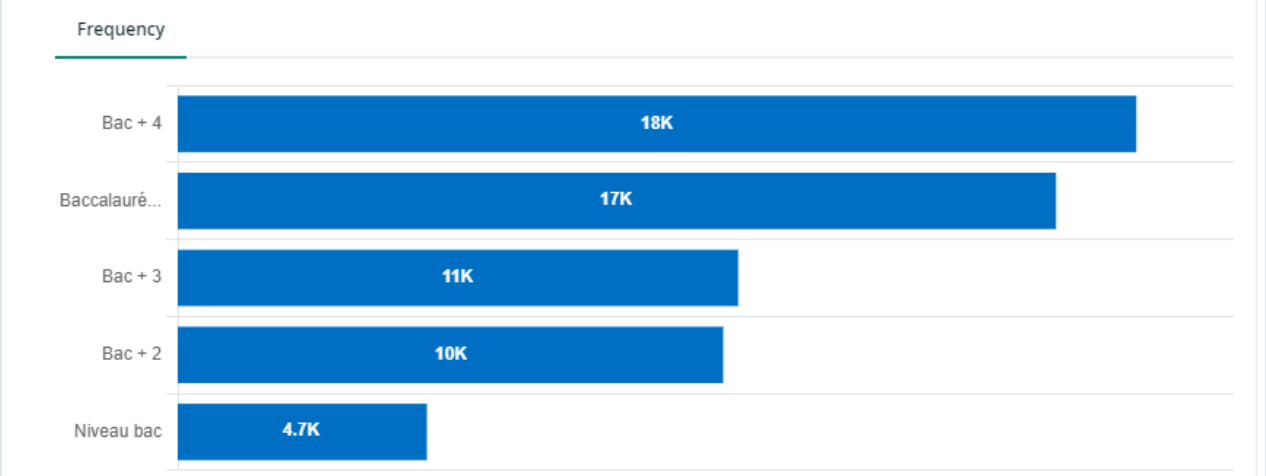
Tip
- In some charts, you can zoom in by selecting the area of your preference. Click the **Reset zoom** button to return to the original chart view.
- The charts use the abbreviations K for thousand and M for million.

Bar chart

A bar chart, or frequency chart, displays the most and least frequent values of a column along with their number of occurrences.

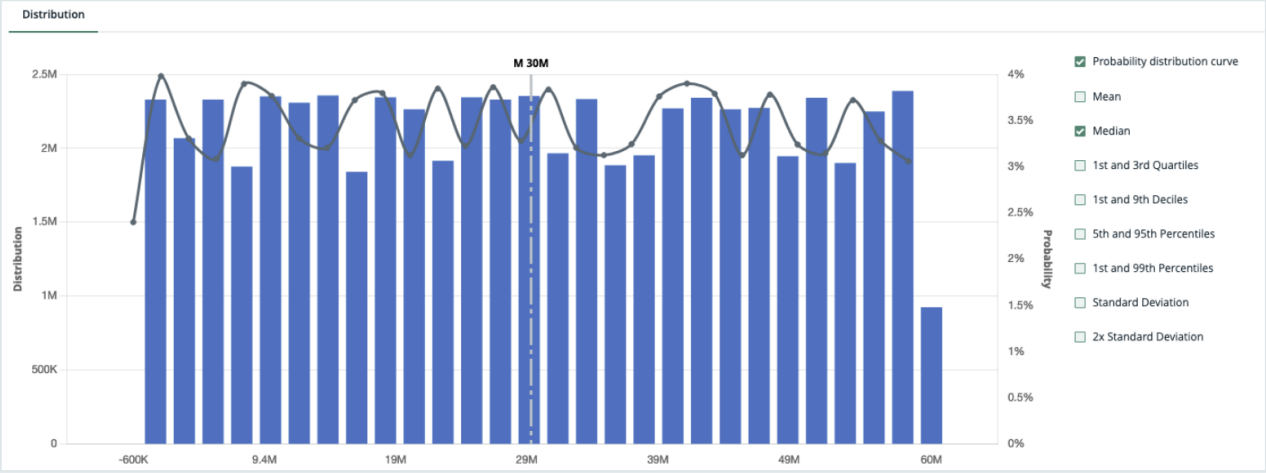
This chart is available if the data type is boolean, numerical or text, and is considered categorical (Categorical Data = true).

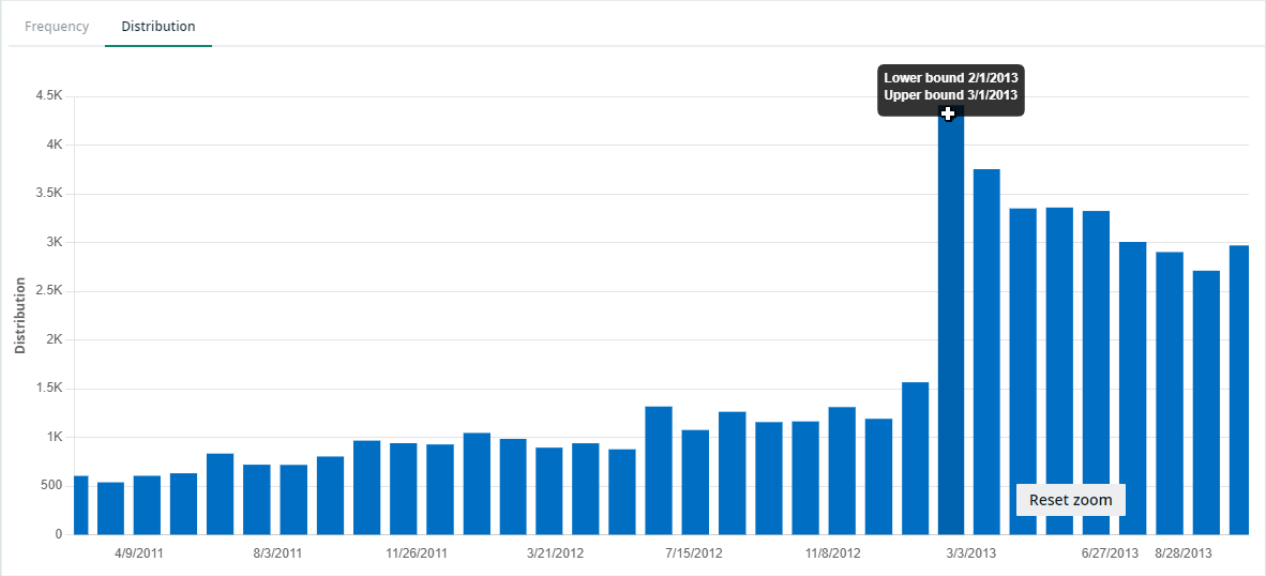




Data distribution chart

The data distribution chart, or histogram, displays how data is distributed.
This chart is available if the data type is numerical or a date.
In this chart, you can receive extra information such as the mean, standard deviation and so on, by selecting the option at the right of the chart.





For information on other profiling results, go to Data profiling information.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements.

Automatic Data Classification via the Cloud Data Classification Platform

When you register a data source, you can store a data profile and sample data. This is required if you want to classify columns in the data set. The Cloud Data Classification Platform predicts the data classes of selected columns and sends them back to Collibra Platform Self-Hosted, where you confirm or reject the suggested data classes. The Cloud Data Classification Platform uses your feedback to retrain the platform and improve future data classifications.

> Warning If you want to use the Cloud Data Classification Platform, request it via your Collibra contact or create a support ticket. See also We have announced the end of life of Jobserver and all related Jobserver integrations for September 30,

2024, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements..

Limitations

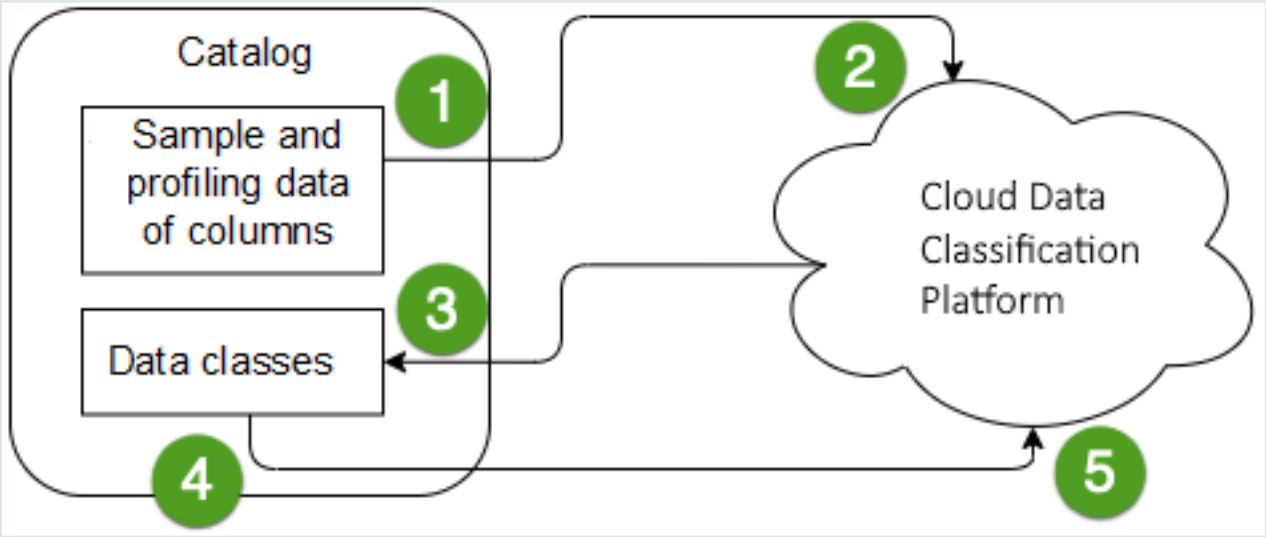
- Automatic data classification via the Cloud Data Classification Platform is a cloud service. Only if your on-premises environment can reach the cloud service, you can use it.
- Out-of-the-box, automatic data classification can predict several data classes. However, you can also create user-defined data classes to increase its prediction quality.
- The only supported language for data classes is English.
- The Cloud Data Classification Platform needs sample data and profiling data to be able to predict the data classes.

  > Note You can create sample data and profiling data by registering a data source and choosing to create sample data and profiling data or by importing the data via the Catalog API.

- The Cloud Data Classification Platform only works for columns of data sources that are registered in Data Catalog with sample data and profiling data.

Automatic data classification flow via the Cloud Data Classification Platform

In the following schema, you can see the different steps of an automatic data classification flow via the Cloud Data Classification Platform.



Step	Description
1	You select the columns that you want to classify and send their sample and profiling data to the Cloud Data Classification Platform. See We have announced the end of life of Jobserver and all related Jobserver integrations for September 30, 2024, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements.
2	The Cloud Data Classification Platform predicts the data classes of the columns.
3	The Cloud Data Classification Platform sends the data classes to Collibra.
4	You provide feedback by accepting or rejecting the predicted data class of each column or by adding your own new classes. The Cloud Data Classification Platform can predict multiple data classes for one column. If the prediction is accurate, you can accept multiple data classes for one column.
5	Your data class selections are sent to the Cloud Data Classification Platform . The Cloud Data Classification Platform stores your selections, along with the associated sample data, to retrain the classification model and improve future classification predictions.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements.

Cloud Data Classification Platform setup

The Cloud Data Classification Platform requires specific setup.

Before you begin

- Data Catalog experience is enabled in the DGC service configuration.
  - » This will give you access to the improved Schema asset page.
- You are using profiling data within Data Catalog.

Steps

Request the use of the Cloud Data Classification Platform via your Collibra contact or create a support ticket.

Note
- Be aware that after you accept the predicted data classes, all sample data and profiling data is stored on the Cloud Data Classification Platform.
- We recommend to use a Cloud Data Classification Platform running in the same region as your Collibra environment.

What's next?

Enable the Cloud Data Classification Platform.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements.

Classify columns

By classifying columns, Collibra's Automatic Data Classification predicts their data structures, after which, you can accept or reject the prediction.

> Note This information is specific to the Cloud Data Classification Platform. For information on classifying via Edge, see About profiling and classification via Edge.

You can classify columns via the:

- Database asset page
- Schema asset page
- Table asset page

> Tip You can also use the physical data connector to manually select a data class for individual columns.

Prerequisites

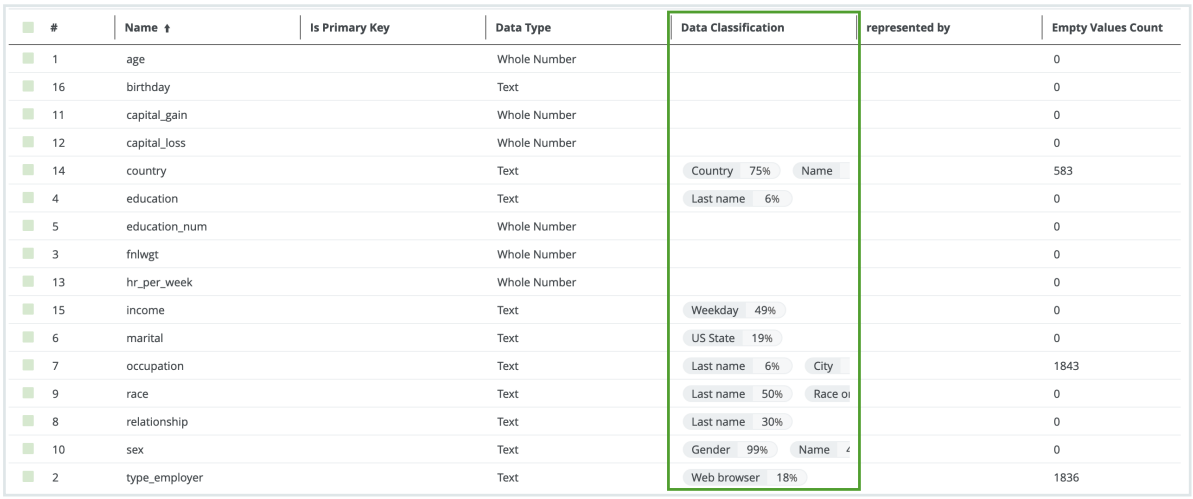
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have created a support ticket via Zendesk to access to the Automatic Data Classification platform.
- You have configured the Cloud Data Classification Platform.
- You have the correct permissions to classify tables and columns.
- You have registered a data source, including these options:
  - Store Data Profile
  - Store Sample Data
- Data Catalog experience is enabled in the DGC service configuration.
  » This will give you access to the improved Schema asset page.

Via the Database asset page

1. Open the Database asset that contains the tables and columns in the schema you want to classify.

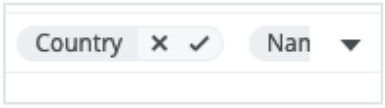
a. On the main menu, click ⠿, and then click Catalog.
   » The Catalog Home opens.
b. In the subpages, click **Technology Assets**.
c. Filter on the Database asset type.

2. Open the relevant database, and then click **Actions** → **Classify**.
   » You can follow the status of the classification in Activities.
3. Open the database asset with the classified columns.
4. Add the Data Classification column to the table.
   » In the **Data Classification** column, you find the suggested data classes.



#	Name	Is Primary Key	Data Type	Data Classification	represented by	Empty Values Count
1	age		Whole Number			0
16	birthday		Text			0
11	capital_gain		Whole Number			0
12	capital_loss		Whole Number			0
14	country		Text	Country 75% Name		583
4	education		Text	Last name 6%		0
5	education_num		Whole Number			0
3	fnlwgt		Whole Number			0
13	hr_per_week		Whole Number			0
15	income		Text	Weekday 49%		0
6	marital		Text	US State 19%		0
7	occupation		Text	Last name 6% City		1843
9	race		Text	Last name 50% Race o		0
8	relationship		Text	Last name 30%		0
10	sex		Text	Gender 99% Name		0
2	type_employer		Text	Web browser 18%		1836

5. Hover over the classification percentages and accept (✓) or reject (✕) the suggested data class.



- Accepting the classification leaves the classification in the list.
- Rejecting the classification removes the result from the data classification list.

Via the Schema asset page

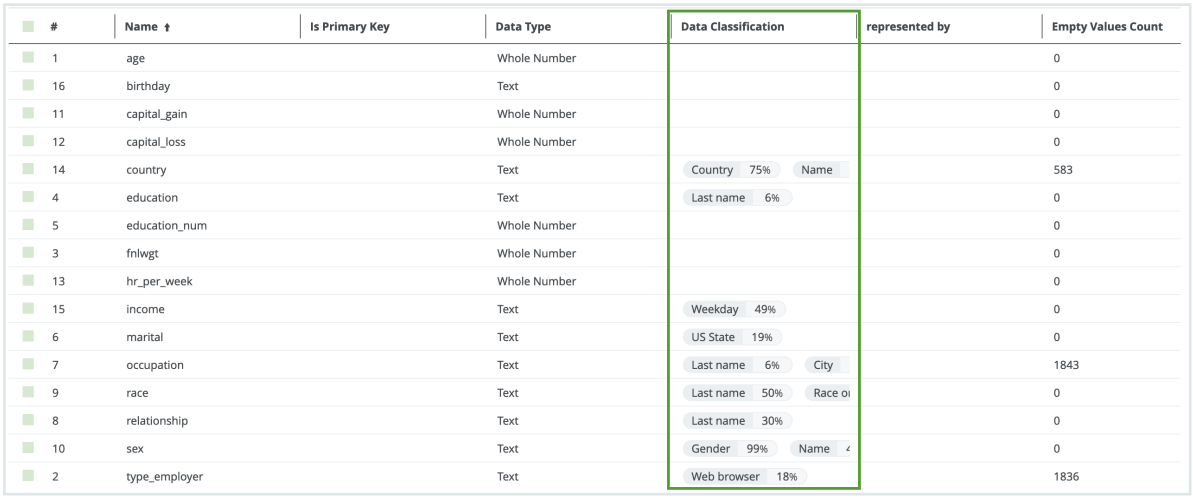
1. Open the Schema asset that contains the tables and columns that you want to classify.
   a. On the main menu, click ⠿, and then click Catalog.
      » The Catalog Home opens.
   b. In the subpages, click **Data Sources**.
   c. Click the relevant schema.

2. Click the Tables tab.
3. Select one or more tables from the schema.
4. To classify all columns in the table, click **Actions** → **Classify**.

   > Tip To classify one or more specific columns, select the columns, then click **Actions** → **Classify**.

   » You can follow the status of the classification job in Activities.
5. Open the Table asset with the classified columns.
6. Add the Data Classification column to the table.

   » In the **Data Classification** column, you find the suggested data classes.



#	Name	Is Primary Key	Data Type	Data Classification	represented by	Empty Values Count
1	age		Whole Number			0
16	birthday		Text			0
11	capital_gain		Whole Number			0
12	capital_loss		Whole Number			0
14	country		Text	Country 75% Name		583
4	education		Text	Last name 6%		0
5	education_num		Whole Number			0
3	fnlwgt		Whole Number			0
13	hr_per_week		Whole Number			0
15	income		Text	Weekday 49%		0
6	marital		Text	US State 19%		0
7	occupation		Text	Last name 6% City		1843
9	race		Text	Last name 50% Race o		0
8	relationship		Text	Last name 30%		0
10	sex		Text	Gender 99% Name		0
2	type_employer		Text	Web browser 18%		1836

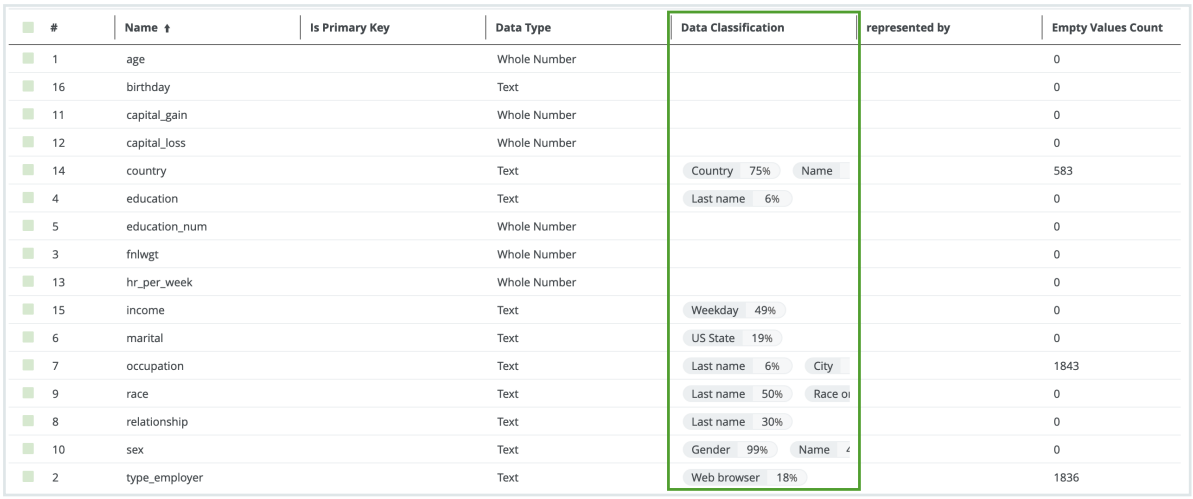
7. Hover over the classification percentages and accept (✔) or reject (✖) the suggested data class.

Via the Table asset page

1. Open a Table asset that has columns you want to classify.
2. On the Table asset page, do one of the following:
   a. To classify all columns in the table, click **Actions** → **Classify** in the upper right corner.
   b. To classify specific columns in the table, select the columns and click **Actions** → **Classify** in the upper right corner.

      » You can follow the status of the classification job in Activities.

3. Open the relevant table, and then add the Data Classification column to the table.
   » In the **Data Classification** column, you find the suggested data classes.



#	Name	Is Primary Key	Data Type	Data Classification	represented by	Empty Values Count
1	age		Whole Number			0
16	birthday		Text			0
11	capital_gain		Whole Number			0
12	capital_loss		Whole Number			0
14	country		Text	Country 75% Name		583
4	education		Text	Last name 6%		0
5	education_num		Whole Number			0
3	fnlwgt		Whole Number			0
13	hr_per_week		Whole Number			0
15	income		Text	Weekday 49%		0
6	marital		Text	US State 19%		0
7	occupation		Text	Last name 6% City		1843
9	race		Text	Last name 50% Race o		0
8	relationship		Text	Last name 30%		0
10	sex		Text	Gender 99% Name		0
2	type_employer		Text	Web browser 18%		1836

4. Hover over the classification percentages and accept (✓) or reject (✕) the suggested data class.

Unified Data Classification method (Beta)

The Unified Data Classification method is a data classification method on Edge based on data classes that you can configure and modify based on your own needs.

About the Unified Data Classification method

> Important Unified Data Classification is in beta testing. Only activate this feature in your Test environments. Don't enable it in Production environments yet because it's not fully ready.

Why do we need a new data classification method?

- Organizations want to create custom data classes that can be used and detected by the automatic data classification process on Edge.

- Running the data classification together with profiling, like the Edge data classification method, isn't aligned with an organization's needs. The data class of a specific column hardly ever changes, whereas the profiling statistics do. In that sense, classification should not be run as often as profiling.
- The Cloud Data Classification Platform can no longer remain available due to issues with error-control in machine learning. With machine learning, it is hard to understand why a column is classified in a specific way and to solve issues for incorrect classifications.

What is the Unified Data Classification method?

The Unified Data Classification method:

- Works via Edge and requires specific setup.
  Because the data doesn't leave your organization's network, the automatic data classification process is more secure. The samples used during the automatic data classification process are stored between 24 and 48 hours in the Edge Site cache. They are not transferred to Collibra.
- Saves time during the profiling activity.
  The classification no longer starts with the profiling activity. You can start a separate classification process for a specific asset with a dedicated **Classify** button.
- Relies on classification rules specified for each data class.
  This means that the classification engine no longer relies on machine learning, which will make issues and changes more transparent. This also provides more flexibility and allows for customizations.
- Delivers optional out-of-the-box data classes.
  This means you decide which out-of-the-box data classes you want to use. It also allows you to adjust the provided data classes to your own needs, like changing the name or changing the classification rules.
- Works via a new REST API.
  With the new REST API, you can manage data classes and start the classification.
- Will replace the current data classification via Edge and data classification via the Cloud Data Classification Platform over time.

Important
- The automatic data classification process is not available in on-premises environments. You can create data classes and manually classify your data.
- Data classes and classifications created via the Unified Data Classification method are separated from the old data classes and classifications. The old data classes and classifications are no longer visible if you enable Unified Data Classification.The opposite is also true.
- When Unified Data Classification becomes generally available, a classification migration process will be put in place. This process will transfer all old data classes and data classifications to the new Unified Data Classification method. Old transferred data classes will need to be updated to include classification rules to work with the new automatic data classification method.

Tip You can follow a training and watch videos via Collibra University.

Beta limitations

In this phase, you cannot:

- Merge data classes.
- Search on data classes and classifications.
- Migrate existing data classes and existing data classifications to the new data classification method.

The global permissions "Classification / Data Classes / Classify" and "Classification > Data Classes > Read" are not enforced yet, meaning you can assign them but they are not yet taken into account. For more information, go to Enable Unified Data Classification.

Enable the Unified Data Classification method

Important Unified Data Classification is in beta testing. Only activate this feature in your Test environments. Don't enable it in Production environments yet because it's not fully ready.

Before you can start configuring and using the Unified Data Classification method, you have to enable it in your environment.

Before you start

- You have created and installed an Edge site.
- You have created a JDBC connection for your data source.
- You have registered your data source.

Required permissions

- You have a global role that has the **System administration** global permission.
- You have a global role that has the **Manage connections and capabilities** global permission, for example, Edge integration engineer.

Steps

1. Enable the **Unified Classification enabled** setting.
   This setting makes the Unified Data Classification method, available for use.

   > Important
   > If you enable this setting, all existing data classes and classifications become unavailable. If you deactivate the setting again, you revert back to your previous Data Classification setup and the previously defined data classes and classifications are available again.

   Show how

   ### Required permissions

   - You have the **ADMIN** or **SUPER** role in Collibra Console.
   - You have the **SUPER** role in Collibra Console.
   - You have the **ADMIN** or **SUPER** role in Collibra Console.

   Steps

a. Open the DGC service settings for editing:
    i. Open Collibra Console.
        » Collibra Console opens with the **Infrastructure** page.
    ii. In the tab pane, expand an environment to show its services.
    iii. In the tab pane, click the Data Governance Center service of that environment.
    iv. Click **Configuration**.
    v. Click **Edit configuration**.
b. Open the DGC service settings for editing:
    i. Open Collibra Console.
        » Collibra Console opens with the **Infrastructure** page.
    ii. In the tab pane, expand an environment to show its services.
    iii. In the tab pane, click the Data Governance Center service of that environment.
    iv. Click **Configuration**.
    v. Click **Edit configuration**.

c. In the **Beta features** section, enter the required information:

Setting	Description
Unified Classification enabled	Enables the new Unified Data Classification method on Edge. ■ ✔ True: The environment uses the new classification method, Unified Data Classification. This has an impact on the available data classes, the required capabilities, and the way you classify data. Note All existing data classes and classifications become unavailable. Tip A migration process will be available in a future release. This process will transfer all old data classes and classifications to the Unified Data Classification method. Old transferred data classes will need to be updated to include classification rules to work with the new automatic data classification method. ■ ✕ False: (default): The beta feature is not enabled. If you deactivate the setting, you revert back to your previous data classification setup and the previously defined data classes and classifications become available again.

d. Click **Save all**.

2. For each data source that you want to classify, add the **Catalog Data Classification** capability to the Edge connection.

   Show how

   a. Open an Edge site.
      i. On the main menu, click ⠿, and then click ⚙ **Settings**.
         » The Collibra settings page opens.
      ii. In the tab pane, click **Edge**.
         » The **Sites** tab opens and shows a table with an overview of the Edge sites.
      iii. In the table, click the name of the Edge site whose status is **Healthy**.
         » The Edge site page opens.
   b. In the **Capabilities** section, click **Add capability**.
      » The **Add capability** page is shown.

c. Enter the required information.

Field	Description	Required
Capability	This section contains general information about the capability.	
Name	The name of the Edge capability.	✓ Yes
Description	The description of the Edge capability.	✗ No
Capability template	The capability template. The value that you select in this field determines which sections appear on the page. Select the following Edge capability: `Catalog Data Classification`	✓ Yes
Connection	This section contains information to connect to the data source.	
JDBC connection	The connection to the data source.	✓ Yes
General	This section contains general information about logging.	
Debug	An option to automatically send Edge infrastructure log files to Collibra Platform Self-Hosted. By default, this option is set to *false*. > Note We highly recommend to only send Edge infrastructure log files to Collibra Platform Self-Hosted when you have issues with Edge. If you set it to *true*, it will automatically revert to *false* after 24h.	✗ No
Log level	An option to determine the verbosity level of Catalog connector log files. By default, this option is set to *No logging*.	✗ No

d. Click **Create**.
   » The capability is added to the Edge site.
   » The fields become read-only.

3. Give the Edge Site user the following global permissions:

    - Classification > Data Classes > List Values.
    - View Permissions > View All.
4. Give your data stewards the global permissions they need. The available permissions are:
    - Classification > Data Classes > Add
    - Classification > Data Classes > Update
    - Classification > Data Classes > Remove
    - Classification > Data Classes > Read.
      The read global permission is not enforced yet, meaning you can assign it but it's not yet taken into account.
    - Classification > Data Classes > Classify.
      The Classification / Data Classes / Classify global permission is not enforced yet, meaning you can assign it but it's not yet taken into account.
5. To view the **Classify** button in an asset page, you need the Catalog global permission and the following resource permissions:
    - Column asset (Asset > Attribute):
        - Add
        - Remove
        - Update
    - Table asset (Asset > Attribute):
        - Add
        - Remove
        - Update

What's next?

Users with the correct permissions can now start configuring the data classes and using the Unified Data Classification method.

Configuring data classes in the Unified Data Classification method

> Important Unified Data Classification is in beta testing. Only activate this feature in your Test environments. Don't enable it in Production environments yet because it's not fully ready.

You can create data classes via the asset pages where you can update the classification or via the **Data Classification** page in the Stewardship application.
To update and delete data classes, especially the data classification rules, you must use the **Data Classification** page in the Stewardship application.

> Note Currently, you cannot merge data classes.

About data classes in the Unified Data Classification method

Data classes are the different groups you want to use to classify your data, for example, email, phone number, and web browser. You can create, update, and remove data classes. You can also import out-of-the-box data classes and update them.
Currently, you cannot merge data classes using the Unified Data Classification method.

A data class in the Unified Data Classification method (Beta) consists of the following elements:

Data class element	Description
Name	The name of the data class.

Data class element	Description
Enabled	Switch to indicate whether this data class needs to be taken into account during the classification process. If a data class is not enabled, the automatic data classification doesn't consider this data class and the data class is also unavailable when you manually classify a column. However, if a column is already classified with that data class, the classification is still valid. This option can be useful if the data class is not ready for use or if it is in testing phase.
Description	The description of the data class.
Details	
Minimum confidence threshold	The confidence percentage that must be reached for the data class to be considered as a classification result. The confidence percentage is the percentage of values in the column that match the classification rule, for example, the regular expression. Enter a value between 0 and 100. The default value is 0. Example If you add value 80 in this field, the data class is returned by the automatic classification process only if the confidence percentage reaches 80 percent or higher. Tip Confidence scores of 0 are never taken into account.

Data class element	Description

Include empty values	Indicates if you want to include empty values in the confidence percentage calculation. The possible values are:
	• ✔ True True: If the value is set to true, empty values are taken into account by the classification process when calculating the confidence percentage of a matching data class.
	• ✘ False False (Default): If the value is set to false, only the non-empty values are taken into account by the classification process when calculating the confidence percentage of a matching data class.
	This option can be used to receive an accurate confidence score for all data in a column.
	Example
	You have a column Z with 40 empty values and 60 phone numbers. You have a data class A with a regular expression to detect US phone numbers.
	• If you set this option to False and you classify column Z, data class A could be suggested with a confidence percentage of 100.
	• If you set this option to True and you classify column Z , data class A could be suggested but with a confidence percentage of only 60.
	Important Some regular expressions are constructed to allow a match with empty values. This means that, through the regular expression, empty values can be matched to the data class, which affects the confidence score.
	Example:
	This expression won't match empty values with the email data class:
	`^([a-zA-Z0-9._%\-]+@[a-zA-Z0-9.\-]+\.[a-zA-Z]{2,6})$`
	This expression will match empty values with the email data class:
	`^([a-zA-Z0-9._%\-]+@[a-zA-Z0-9.\-]+\.[a-zA-Z]{2,6})*$`

Examples	Some examples of values that match the classification rule for the data class.

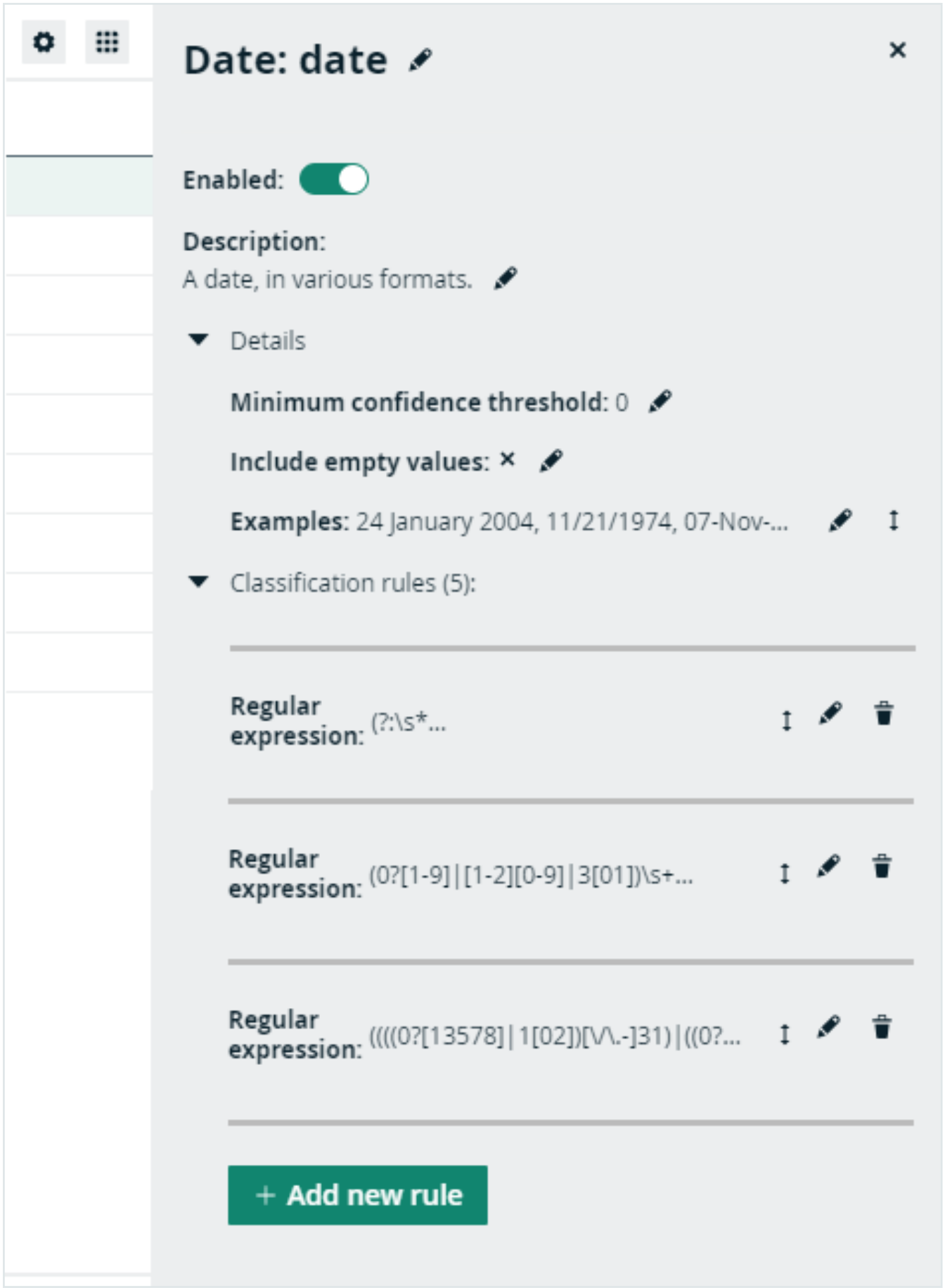
Data class element	Description
Classification rules	A data classification rule is used by the data classification process to calculate the confidence score, which is a percentage that indicates the likelihood that the data class fits the data in an asset. A data class can contain multiple data classification rules. Each rule is verified against the data, and the data class is assigned as soon as one of the rules applies. > Example You have defined the email data class with a regular expression. However, the values “unknown”, “invalid”, and “missing” are also acceptable email values in your data source. You can add a list of values as a second rule on the email data class. For the full example, go to Example: Configuring a data class with two classification rules. If you are using Collibra on-premises and Unified Data Classification, don't add any classification rules because automatic classification is not available.
Type	The possible values are: Regular expression or List of values. Depending on your selection other fields appear.

Data class element	Description
Regular expression	This field appears if you select a classification rule of the type Regular expression. A regular expression, also referred to as regex or regexp, is a sequence of characters that specifies a match pattern in text. Multiple regular expression grammar variants exist. We use the Java variant. Example A regular expression for an email address can be `^[a-zA-Z0-9._%\-]+@[a-zA-Z0-9.\-]+\.[a-zA-Z]{2,6}$` Tip • Multiple websites provide guidelines and examples of regular expressions, for example, Regexlib and RegexBuddy, or even ChatGPT. • You can also test your regular expression on various websites, for example, Regex101 (Select the **Java 8** option in the **Flavor** panel). The referenced websites serve only as examples. The use of ChatGPT or other generative AI products and services is at your own risk. Collibra is not responsible for the privacy, confidentiality, or protection of the data you submit to such products or services, and has no liability for such use. Important The required format of the regular expression is different between the UI and the API. In the API backslashes must be added twice (escaped). In the UI, this is not needed. For example: In the UI, use `^\+?\d{1,3}?\d{1,4}$`, and in the API, use `^\\+?\d{1,3}?\\d{1,4}$`.

Data class element	Description

Values	This field appears if you select a classification rule of the type List of values.
	
	Add the values that define a specific data class.
	
	Example
	A data class for T-shirt sizes based on a list of values could be:
	
	`S`
	
	`M`
	
	`L`
	
	`small`
	
	`medium`
	
	`large`
	
	Important
	• The number of values in a list is limited to 1,000. Later, you will be able to add larger lists by uploading a file. This is not yet available in this phase.
	• Add only one value per line.
	• The maximum number of characters in a single list value is 10,000.
	• Don't add any leading or trailing blank characters in a value.
	• The values are not case-sensitive, the value "small" in the list will also be a match with the values "Small" and "SMALL".

Description	A description of the classification rule.


Date: date
Enabled:
Description:
A date, in various formats.
Details
Minimum confidence threshold: 0
Include empty values: ×
Examples: 24 January 2004, 11/21/1974, 07-Nov-...
Classification rules (5):
Regular expression: (?:\s*...
Regular expression: (0?[1-9]|[1-2][0-9]|3[01])\s+...
Regular expression: ((((0?[13578]|1[02])[\/\.-]31)|((0?...
+ Add new rule

> Tip
> The maximum number of rules in a data class is 25.

About out-of-the-box data classes

Out-of-the-box data classes are created by Collibra. You can decide if and which out-of-the-box data classes you want to use. This allows you to have only the data classes you are interested in and to reduce the risk of similar, overlapping data classes.

> Example In the out-of-the-box , we have the data classes: Credit card and Credit card: Visa. Both are overlapping because a Visa credit card is also a credit card. You can decide which data class you want to use depending on the granularity you need.

Once an out-of-the-box data class has been imported, it's considered as a regular data class, which means you can edit the data class and change its classification rules.

If you import out-of-the-box data classes again, we'll detect whether you have data classes with the same name. We'll inform you about this and indicate whether their rules are different. The following statuses are available:

Status	Description
New	This data class is not yet available in your environment. You can import this data class without any risks that you erase existing data.
Exists (no changes)	A data class with the same name is already available in your environment and the definition of the different classification rules are the same. This data class will not be imported, even if you select it for import. Data classes with this status are, by default, deselected for import.
Exists (changed)	A data class with the same name but with different classification rules is already available in your environment. You can import this data class, but the current classification rules will be replaced by those in the out-of-the-box data class. Also the classification rule description will be updated. Data classes with this status are, by default, deselected for import.

> Important
> When we compare data classes to check if they were changed, we compare only the classification rules.
>
> - Global properties, such as data class description, confidence score threshold, and examples are not taken into account.
>   If you import the out-of-the-box data class, these properties are not updated.
> - Classification rule descriptions are not taken into account.
>   However, if you import the out-of-the-box data class, the classification rules, including the classification rule descriptions, are updated.

Create a data class

> Important Unified Data Classification is in beta testing. Only activate this feature in your Test environments. Don't enable it in Production environments yet because it's not fully ready.

You can create data classes via the asset pages where you can update the classification or via the **Data Classification** page in the Stewardship application.
To update and delete data classes, especially the data classification rules, you must use the **Data Classification** page in the Stewardship application.

If you are using Collibra on-premises and Unified Data Classification, don't add any classification rules because automatic classification is not available.

Before you begin

You have enabled the Unified Data Classification method.

Required permissions

You have a global role that has the **Data Classes > Add** global permission.
You have a global role that has the **Data Classes > Update** global permission.
You have a global role that has the **Data Classes > Remove** global permission.

Steps

Watch a video

1. On the main menu, click , and then click **Stewardship**.
2. Click the **Data Classification** tab.
3. If the data class doesn't exist yet:
    a. Click **Add**.
    b. Type the name of the data class and press *Enter*.
    c. Click **Create**.
4. Select the data class that you want to configure.

5. The data class parameters appear in a pane on the right-hand side.
6. Optionally, add a description by clicking the **Edit** icon next to the **Description** field.
7. Open the **Details** section.
8. Complete the fields as required.

Data class element	Description
Minimum confidence threshold	The confidence percentage that must be reached for the data class to be considered as a classification result. The confidence percentage is the percentage of values in the column that match the classification rule, for example, the regular expression. Enter a value between 0 and 100. The default value is 0. Example If you add value 80 in this field, the data class is returned by the automatic classification process only if the confidence percentage reaches 80 percent or higher. Tip Confidence scores of 0 are never taken into account.

Data class element	Description

Include empty values	Indicates if you want to include empty values in the confidence percentage calculation.
	The possible values are:
	◦ ✔ True True: If the value is set to true, empty values are taken into account by the classification process when calculating the confidence percentage of a matching data class.
	◦ ✖ False False (Default): If the value is set to false, only the non-empty values are taken into account by the classification process when calculating the confidence percentage of a matching data class.
	This option can be used to receive an accurate confidence score for all data in a column.
	
	Example
	You have a column Z with 40 empty values and 60 phone numbers. You have a data class A with a regular expression to detect US phone numbers.
	◦ If you set this option to False and you classify column Z, data class A could be suggested with a confidence percentage of 100.
	◦ If you set this option to True and you classify column Z , data class A could be suggested but with a confidence percentage of only 60.
	
	Important Some regular expressions are constructed to allow a match with empty values. This means that, through the regular expression, empty values can be matched to the data class, which affects the confidence score.
	Example:
	This expression won't match empty values with the email data class:
	`^([a-zA-Z0-9._%\-]+@[a-zA-Z0-9.\-]+\.[a-zA-Z]{2,6})$`
	This expression will match empty values with the email data class:
	`^([a-zA-Z0-9._%\-]+@[a-zA-Z0-9.\-]+\.[a-zA-Z]{2,6})*$`

Examples	Some examples of values that match the classification rule for the data class.

9. To change a value, click the **Edit** icon .
   To save the value, click the Save icon.
10. Open the **Classification rules** section.
11. Click **Add new rule**.
    A data class without a classification rule can be used only for manual classification. You need to add at least one classification rule to allow data classification based on the data class.
12. From the **Type** list, select the type of classification rule that you want to add to the data class.
    Depending on your selection, extra fields appear.
13. Complete the fields as required.

Data class element	Description
Regular expression	This field appears if you select a classification rule of the type Regular expression. A regular expression, also referred to as regex or regexp, is a sequence of characters that specifies a match pattern in text. Multiple regular expression grammar variants exist. We use the Java variant. Example A regular expression for an email address can be `^[a-zA-Z0-9._%\-]+@[a-zA-Z0-9.\-]+\.[a-zA-Z]{2,6}$` Tip ◦ Multiple websites provide guidelines and examples of regular expressions, for example, Regexlib and RegexBuddy, or even ChatGPT. ◦ You can also test your regular expression on various websites, for example, Regex101 (Select the **Java 8** option in the **Flavor** panel). The referenced websites serve only as examples. The use of ChatGPT or other generative AI products and services is at your own risk. Collibra is not responsible for the privacy, confidentiality, or protection of the data you submit to such products or services, and has no liability for such use. Important The required format of the regular expression is different between the UI and the API. In the API backslashes must be added twice (escaped). In the UI, this is not needed. For example: In the UI, use `^\+?\d{1,3}?\d{1,4}$`, and in the API, use `^\\+?\d{1,3}?\\d{1,4}$`.

Data class element	Description

Values	This field appears if you select a classification rule of the type List of values.
	Add the values that define a specific data class.
	Example
	A data class for T-shirt sizes based on a list of values could be:
	`S`
	`M`
	`L`
	`small`
	`medium`
	`large`
	Important
	◦ The number of values in a list is limited to 1,000. Later, you will be able to add larger lists by uploading a file. This is not yet available in this phase.
	◦ Add only one value per line.
	◦ The maximum number of characters in a single list value is 10,000.
	◦ Don't add any leading or trailing blank characters in a value.
	◦ The values are not case-sensitive, the value "small" in the list will also be a match with the values "Small" and "SMALL".

Description	A description of the classification rule.

14. Click **Save**.
    The classification rule for the data class is configured.
    A new section appears. If you expand the section, the details are shown.
15. If needed click **Add new rule** to add another classification rule to the data class.
    The maximum number of rules in a data class is 25.

What's next?

Import out-of-the-box data classes

Go to some examples

Import out-of-the-box data classes

> Important Unified Data Classification is in beta testing. Only activate this feature in your Test environments. Don't enable it in Production environments yet because it's not fully ready.

The Unified Data Classification method provides a set of out-of-the-box data classes. If you want to use an out-of-the-box data class, you can import it and then update it as desired.

Before you begin

You have enabled the Unified Data Classification method.

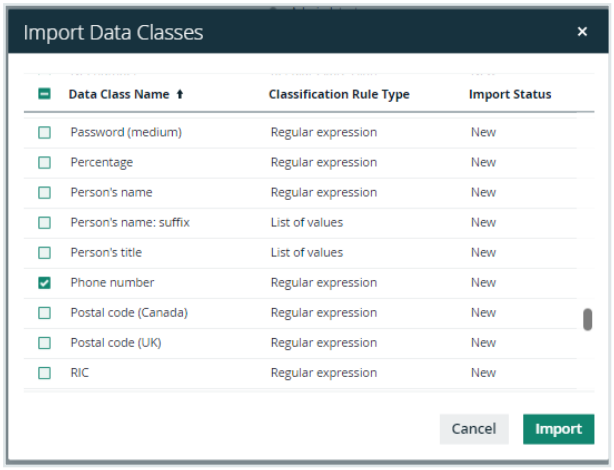
Required permissions

You have a global role that has the **Data Classes > Add** global permission.
You have a global role that has the **Data Classes > Update** global permission.
You have a global role that has the **Data Classes > Remove** global permission.

Steps

1. On the main menu, click ⠿, and then click **Stewardship**.
2. Click the **Data Classification** tab.
3. Click **Import**.
   » The **Import Data Classes** dialog box opens listing all the available out-of-the-box data classes and their status. For information on the possible statuses, go to About out-of-the-box data classes.

4. Select the data classes that you want to import.



5. If you have selected existing data classes that are reported as changed, confirm the import action.
6. Click **Import**.
   The selected data classes are imported and become available. Any selected, existing data classes are updated with the out-of-the-box classification rules. For more details, go to About out-of-the-box data classes. You can update the imported data class completely based on your needs.

> Note Existing classifications are not affected by this process.

What's next?

Go to some examples

Classify assets via the Unified Data Classification method

> Important Unified Data Classification is in beta testing. Only activate this feature in your Test environments. Don't enable it in Production environments yet because it's not fully ready.

Manually classify assets

To add a data class to a Column asset, navigate to the Column asset and select the data class for the column.
For an example, go to Example: Manual classification.

The data classes in the drop-down list are the ones defined after you have activated the Unified Data Classification method. For information on configuring new data classes, go to Configuring data classes.

Automatically classify assets

When you start the automatic data classification process, the process verifies the data in the column or columns against the data classes, and makes classification suggestions with a confidence score. This score is an estimation based on data samples that the data classification process collects. A deviation from the exact score is possible.

If the Unified Data Classification setting has been enabled, the automatic classification process uses the newly defined data classes. For information on configuring new data classes, go to Configuring data classes.

> Important
> - The automatic data classification process needs at least six values that can be checked, to classify a column.
>   Example: For data class A, you define a regular expression and indicate you don't want to consider empty values.
>   If you then classify a column with a lot of null values and five non-null values, the column won't get classified, even if the non-null values match data class A.
> - The automatic data classification process will extract a maximum of 1,000 values from the data source. The samples are stored between 24 and 48 hours in the Edge Site cache. They are not transferred to the Collibra. If the Edge Site cache already contains at least 100 samples for this data source, the automatic data classification process will use those.

To start this classification process for one column:

1. Navigate to the related Column asset.
2. Click the **Data Profiling** tab page.

3. Click the **Classify** button.
   » The data classification process starts.
   » If a data class matches the data in the column, a classification suggestion will be assigned to the Column asset with a confidence percentage.

To start the classification process for one or more columns from a Table, Schema, or Database asset:

1. Navigate to the Table, Schema, or Database asset.
2. Select **Actions** → **Classify**.
   » The data classification process starts.
   » If a data class matches a column in the Table asset, a data classification suggestion will be assigned to the Column asset with a confidence percentage.

What's next?

Go to some examples

Accepting or rejecting automatic data classification suggestions

When Unified Data Classification predicts data classes for a column, the suggestions are visible in the **Data Classification** column in the Table and Column asset pages.

#	Name	Is Primary Key	Data Type	Data Classification	represented by	Empty Values Count
1	age		Whole Number			0
16	birthday		Text			0
11	capital_gain		Whole Number			0
12	capital_loss		Whole Number			0
14	country		Text	Country 75% Name		583
4	education		Text	Last name 6%		0
5	education_num		Whole Number			0
3	fnlwgt		Whole Number			0
13	hr_per_week		Whole Number			0
15	income		Text	Weekday 49%		0
6	marital		Text	US State 19%		0
7	occupation		Text	Last name 6% City		1843
9	race		Text	Last name 50% Race o		0
8	relationship		Text	Last name 30%		0
10	sex		Text	Gender 99% Name		0
2	type_employer		Text	Web browser 18%		1836

- If no data classes are suggested for a column, the automatic data classification process could not predict the data class.
- Sometimes multiple data classes can be suggested.
- The percentage next to the data class indicates the confidence level of the suggestion.
  If automatic data classification acceptance and rejection is active, data classification suggestions with a confidence level within the defined thresholds will be accepted or rejected automatically.

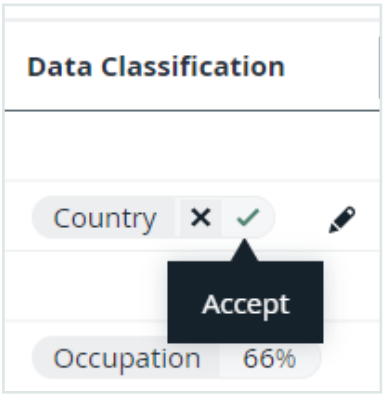
You can accept or reject the data classification suggestions, or add a new data class.

Accept or reject a data classification suggestion

- If you accept the suggested data class, the data class is added to the column.
- If you reject the suggested data class, the data class is removed from the column.

> Important Once you rejected the data class, it won't be suggested again by the automatic data classification method for the column.

To manually accept or reject a data class, hover over the data class and click the appropriate icon.



If automatic data classification acceptance and rejection is active, data classification suggestions with a confidence level within the defined thresholds will be accepted or rejected automatically.

Unified Data Classification examples

> Important Unified Data Classification is in beta testing. Only activate this feature in your Test environments. Don't enable it in Production environments yet because it's not fully ready.

To show you the possibilities in the new data classification method, let's go through some examples.

Scenario

We have registered a data source. By identifying the data class of the data, we get an indication of the type of data, which makes the creation of relations between the physical and logical layer much easier.

1. We have configured the environment for the new classification method, Unified Data Classification.
2. We have registered and synchronized a data source with multiple tables via Edge.

We can now start using the data classification method.

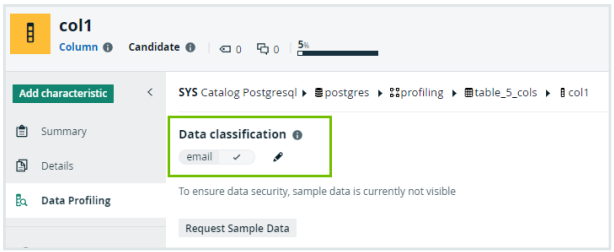
Example: Manual classification

You know that a specific column contains email addresses. You want to manually add the data class, email, to this column.

Steps

1. Navigate to the Column asset.
2. Open the **Data Profiling** tab page.
3. In **Data Classification**, click the **Edit** icon .
   » A drop-down list with all possible data classes appears.
4. Enter the name of the data class you want to add, for example, *email*.
5. If the data class is available from the list, select the option.
   If the data class is not available in the list, press *Enter*, to create the data class.

6. Click **Save**.
   » The Column asset has been manually classified as email.



   » If the data class email didn't exist yet, it is now available for other Column assets.

What's Next?

You can now configure the Email data class so it can be used by the automatic data classification method, Unified Data Classification.

Example: Configuring a data class based on a regular expression and starting the automatic classification for a column

You want to configure the Email data class you manually created and assigned, so this data class can be assigned automatically by the Unified Data Classification method.

Before you begin

Make sure you know which regular expression you want to use for the data class. For more information and references to useful resources, go to Add a data class.

Steps

1. Configure the email data class.
   a. On the main menu, click , and then click **Stewardship**.
   b. Click the **Data Classification** tab.
   c. Select the email data class row.
      » The data class parameters appear in a pane on the right-hand side.
   d. Open the **Details** section.

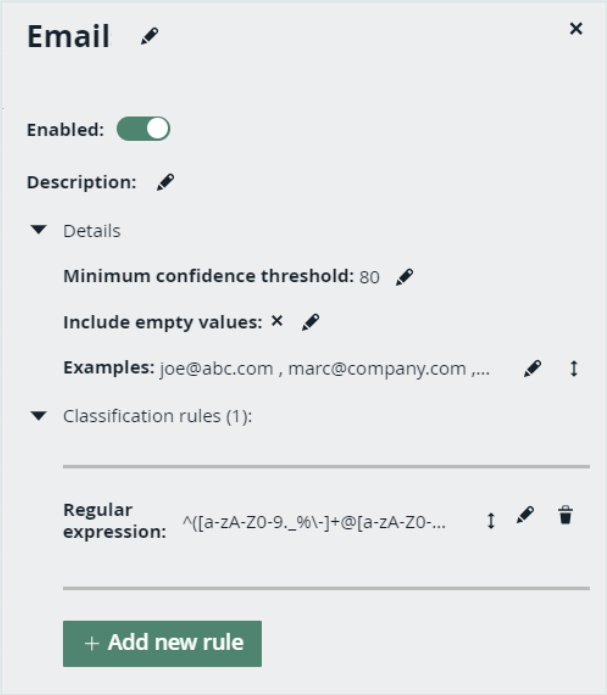
e. Complete the fields as required.
For information on the fields, go to Add a data class.

Data class parameter	Description
Minimum confidence threshold	We set this value to 80.
Allow empty values	We leave this field as the default value (False).
Examples	We add the following examples: `joe@abc.com , marc@company.com , jsmith@b.cc`

f. Open the **Classification rules** section.
g. Click **Add new rule**.
h. In **Type**, select the **Regular expression** option.
» Extra fields appear.
i. Complete the fields as required.
For information on the fields, go to Add a data class.

Data class parameter	Description
Regular expression	We add: `^([a-zA-Z0-9._%\-]+@[a-zA-Z0-9.\-]+\.[a-zA-Z]{2,6})$`
Description	We leave this field empty.

j. Click **Save**.
» The classification rule for the Email data class is configured.
» The **Rule details** section appears. If you expand the section, you see the details.



2. Start the classification.
    a. Navigate to a Column asset with email data.
    b. Click the **Data Profiling** tab page.
    c. Click **Classify**.
        » The data classification process starts. For more information, go to Automatically classify assets.
        » The Email data classification suggestion will be assigned to the Column asset with a confidence percentage. For more information, go to accepting and rejecting data classification suggestions.

What's Next?

You can now configure additional data classes to be used in the automatic classification for a column, table or schema.

Example: Configuring a data class based on a regular expression, importing a data class, and starting the automatic classification for a table

You want to add two new data classes in the Unified Data Classification method:

- Add an extra data class, Date in dd/mm/yyyy format
- Import the out-of-the-box data class, Phone number.

Before you begin

Make sure you know which regular expressions you want to use for the data classes. For more information and references to useful resources, go to Add a data class.

Steps

1. Create and configure the Date data class.
    a. On the main menu, click , and then click **Stewardship**.
    b. Click the **Data Classification** tab.
    c. Add the data class.
        i. Click **Add**.
        ii. Add the Name of the data class. In our case, Date.
        iii. Press *Enter* to add the data class.
        iv. Click **Create**.
            » The data class is created and is available in the list.
    d. Define the data class parameters.
        i. In the **Data Classification** tab, select the row of the new data class.
            » The data class parameters appear in a pane on the right-hand side.
        ii. Optionally, add a description by clicking the Edit icon next to the **Description** field.
        iii. Open the **Details** section.
        iv. Complete the fields as required.
        For information on the fields, go to Configuring data classes.

Data class parameter	Description
Minimum confidence threshold	We set this value to 80.

Data class parameter	Description
Include empty values	We leave this field as the default value (False).
Examples	For Date, we add the following examples: `23/11/2026, 09/02/2023`

v. Open the **Classification rules** section.
vi. Click **Add new rule**.
vii. In the **Type** list, select **Regular expression**.
» Extra fields appear.
viii. Complete the fields as required.
For information on the fields, go to Configuring data classes.

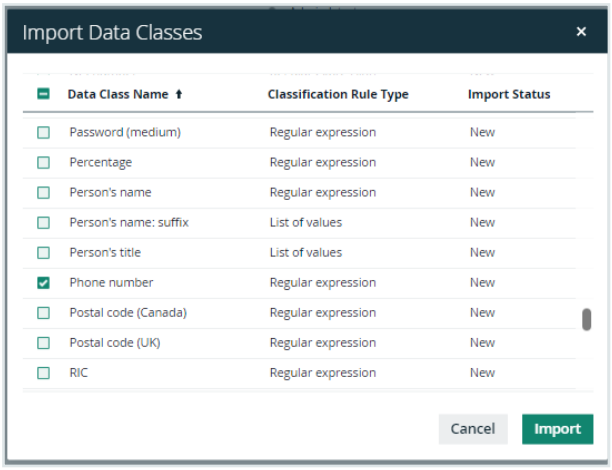
Data class parameter	Description
Regular expression	For Date, we add the following expression: `(0[1-9]|[12][0-9]|3[01])\/(0[1-9]|1[1,2])\/(19|20)\d{2}`
Description	We leave this field empty.

ix. Click **Save**.
» The classification rule for the data class is configured.
» If you expand the **Classification rules** section, you see the details.

2. Import the Phone number data class.

a. On the main menu, click , and then click **Stewardship**.
b. Click the **Data Classification** tab.
c. Click **Import**.
» A dialog box opens, listing all the out-of-the-box data classes and their status. For information on the possible statuses, go to About out-of-the-box data classes.

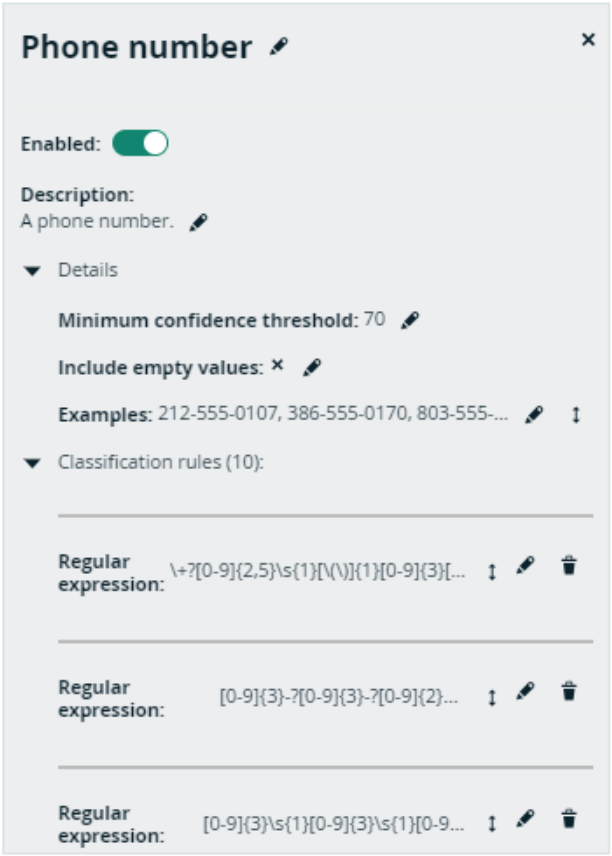
d. Clear all data classes, except **Phone number**.



e. Click **Import**.

» The data class is added.

» If you click the data class, you see the details.



3. Start the automatic classification.

a. Navigate to a Table asset.
b. Select **Actions** → **Classify**.
   » The data classification process starts. For more information, go to Automatically classify assets
   » If a data class matches a column in the Table asset, a data classification suggestion will be assigned to the Column asset with a confidence percentage. For more information, go to accepting and rejecting data classification suggestions.

What's Next?

You can now configure an additional data class that is based on a list of values instead of a regular expression.

Example: Configuring a data class based on a list of values and starting the automatic classification for a table

You want to create an extra data class for T-shirt sizes in the Unified Data Classification method. Once that is done, you want to start the classification process for a full table.

Before you begin

Make sure you know which values you used in the organization to refer to T-shirt sizes. In this case, we consider: XS, M, L, XL, XXL, XXXL, Extra small, Small, Medium, Large, X-Large, XX-Large, XXX-Large, 2XL, 3XL.
For more information, go to Add a data class.

Steps

1. Create and configure data class T-shirt size.
   a. On the main menu, click ⠿, and then click **Stewardship**.
   b. Click the **Data Classification** tab.

c. Add the data class.
    i. Click **Add**.
    ii. Add the Name of the data class. In our case, T-shirt size.
    iii. Press *Enter* to add the data class.
    iv. Click **Create**.
        » The data class has been created and is available in the list.
d. Define the data class parameters.
    i. In the **Data Classification** tab, select the row of the new data class.
        » The data class parameters appear in a pane on the right-hand side.
    ii. Optionally, add a description by clicking the Edit icon next to the **Description** field.
    iii. Open the **Details** section.
    iv. Complete the fields as required.
    For information on the fields, go to Configuring data classes.

Data class parameter	Description
Minimum confidence threshold	We set this value to 80.
Include empty values	We leave this field as the default value (False).
Examples	`Small, L`

    v. Open the **Classification rules** section.
    vi. Click **Add new rule**.
    vii. In the **Type** list, select **List of values**.
        » Extra fields appear.
    viii. Complete the fields as required.
    For information on the fields, go to Configuring data classes.

Data class parameter	Description

Values	We add the following list. Each value must start on a new line.
	`XS`
	`S`
	`M`
	`L`
	`XL`
	`XXL`
	`XXXL`
	`extra small`
	`small`
	`medium`
	`large`
	`X-large`
	`XX-large`
	`XXX-large`
	`2XL`
	`3XL`

Description	We leave this field empty.

ix. Click **Save**.
» The classification rule for the data class is configured.
» If you expand the **Classification rules** section, you see the details.

2. Start the automatic classification.
a. Navigate to a Table asset.
b. Select **Actions** → **Classify**.
» The data classification process starts. For more information, go to

Automatically classify assets

» If a data class matches a column in the Table asset, a data classification suggestion will be assigned to the Column asset with a confidence percentage. For more information, go to accepting and rejecting data classification suggestions.

> Important The values are not case-sensitive, the value “small” in the list will also be a match with the values “Small” and “SMALL”.

> Example A column contains the values `petite`, `s`, `L`, `xl`, `XL`, `unknown`, `unknown`, and `no size`. After the automatic data classification, the column will be classified as a T-shirt size with a confidence score of 50% because half of the values in the column are part of the list of values. Note that the character case didn’t affect the result.

What's Next?

You can also add an extra classification rule to an existing data class.

Example: Configuring a data class with two classification rules

You want to update your existing data class, email, in the Unified Data Classification method. This data class is now based on a regular expression, but you also want to add a list of supported values.

Before you begin

Make sure you know which regular expressions you want to use for the data classes. For more information and references to useful resources, go to Add a data class.

Steps

1. On the main menu, click ⠿, and then click **Stewardship**.
2. Click the **Data Classification** tab.

3. Select the email data class row.
   » The data class parameters appear in a pane on the right-hand side.
4. Open the **Details** section.
5. Click **Add new rule**.
6. In the **Type** list, select **List of values**.
   » Extra fields appear.
7. Complete the fields as required.
   For information on the fields, go to Configuring data classes.

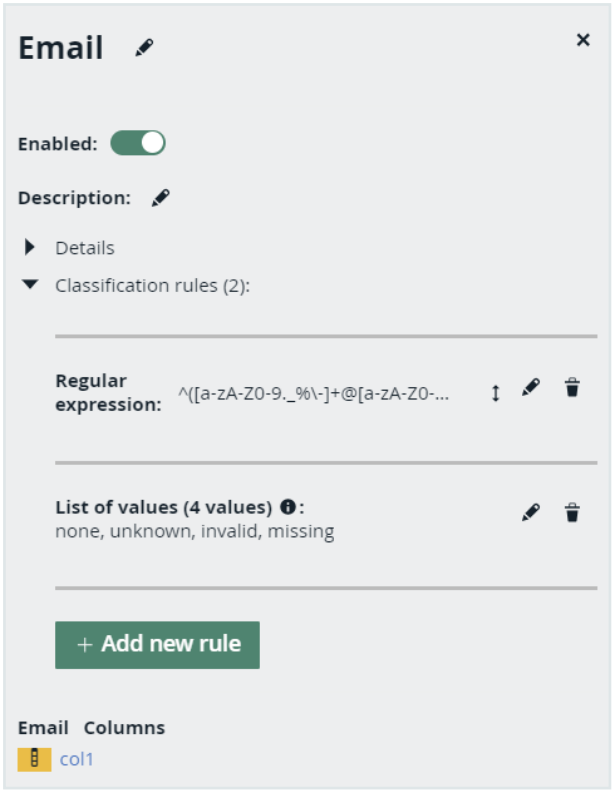
Data class parameter	Description

Values	We add the following list. Each value must start on a new line.
	`none`
	`unknown`
	`invalid`
	`missing`

Description	We leave this field empty.

8. Click **Save**.
   » The extra classification rule for the Email data class is configured.

» If you expand the **Classification rules** section, you see the details.



What's Next?

For other examples, go to Examples.

Troubleshooting the Unified Data Classification method

Check out the Collibra Support portal for Unified Data Classification troubleshooting information.

Data Classification dashboard

The Data Classification dashboard shows all of the data classes available in your environment.

You can use the Data Classification dashboard to:

- See information about data classes.
- Add, merge, and delete data classes.
- Link data classes to data concepts and data attributes.

About the Data Classification dashboard

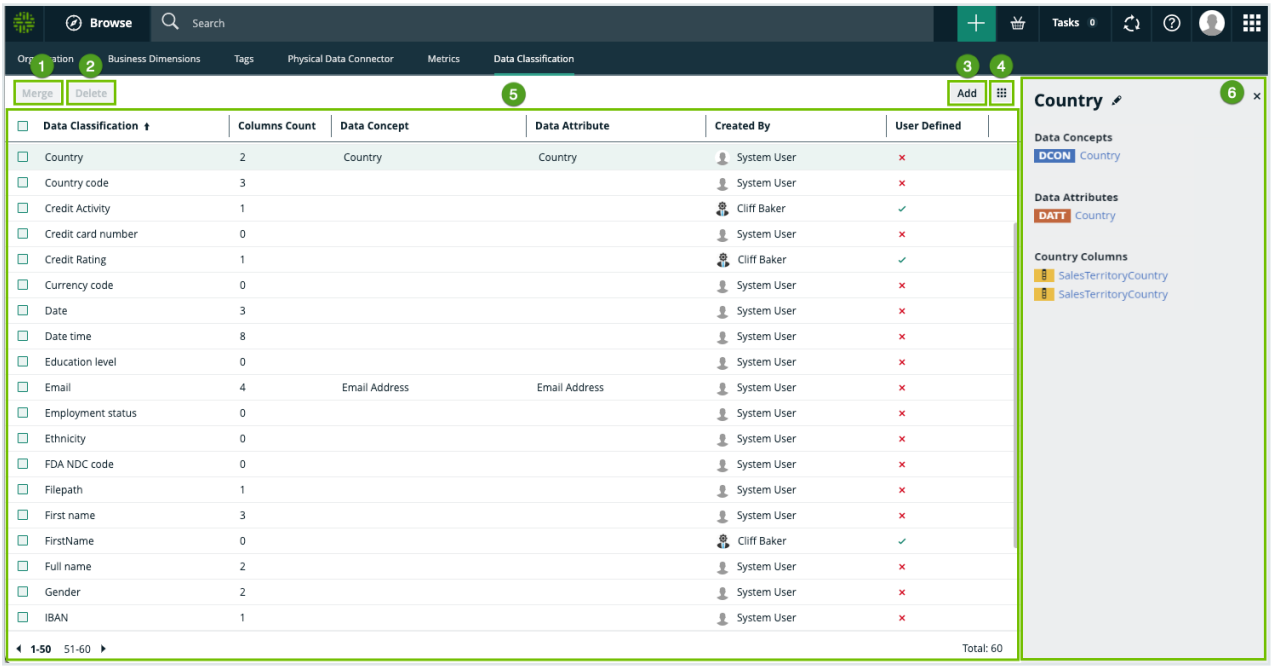
The Data Classification Dashboard shows the list of data classes in your Collibra environment and gives you more control and visibility on them.

When you make changes via the Data Classification dashboard, feedback is automatically sent to the Cloud Data Classification Platform. The feedback is not used when you use data classification via Edge.

To open the dashboard, go to the Stewardship application and click **Data Classification**.

Tip

- If you use the Cloud Data Classification Platform, the packaged data classes do not appear unless automatic data classification has been enabled and configured, and the synchronization process, to make the packaged data classes available in Collibra, has run once. The synchronization process runs once a day.
- If you use data classification via Edge, the packaged data classes do not appear unless automatic data classification has been enabled and configured, and you started the profiling and classification process for a data source.



No.	Name	Description
1	Merge button	A button to merge multiple data classes.
2	Delete button	A button to delete one or more data classes.
3	Add button	A button to manually add a new data class.
4	Table menu ( )	The table menu contains buttons to manage the columns shown.

No.	Name	Description
5	Table with packaged and manually created data classes	A table that shows all the data classes that exist in your environment. You can also view details about each data class.
	Data Classification	The data class name. You can manually add, merge, edit or remove the data classes
	Column Count	The number of columns classified as the associated data class.
	Data Concept	The name of the associated Data Concept assets. It connects the data class to your business asset model.
	Data Attribute	The name of the associated Data Attribute assets. It connects the data class to your logical data model.
	Created By	The name of the user who created the class. If the data class is a packaged data class, the user is the *System User*.
	Created On	The date the data class was created.
	Last Modified By	The name of the user who made the last change.
	Last Modified On	The date the data class was last changed.
	User Defined	Indicates if the data class was automatically or manually created.
6	Side pane	A side pane that provides you with extra details about the selected data class.

View data class information via the Data Classification dashboard

You can view data class information on the Data Classification dashboard.

Prerequisites

- You have configured Automatic Data Classification via the Cloud Data Classification Platform or via Edge.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

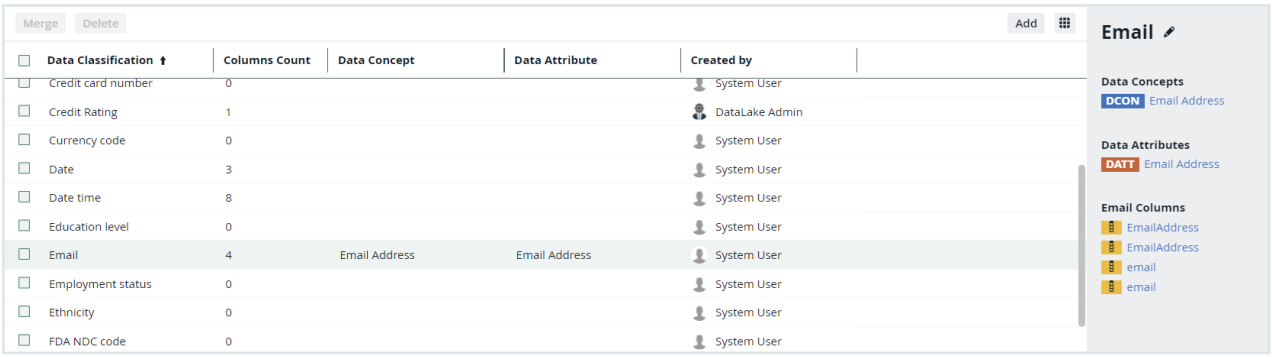
Tip
- If you use the Cloud Data Classification Platform, the packaged data classes do not appear unless automatic data classification has been enabled and configured, and the synchronization process, to make the packaged data classes available in Collibra, has run once. The synchronization process runs once a day.
- If you use data classification via Edge, the packaged data classes do not appear unless automatic data classification has been enabled and configured, and you started the profiling and classification process for a data source.

Steps

1. On the main menu, click , then Stewardship.
2. In the submenu, click **Data Classification**.
3. Click on a row.
   » The data class information appears in the side pane.

The Data Class side pane

The Data Class side pane provides extra data class information. When you click the row of a data class in the Data Classification Dashboard, the data class information appears in the side pane.



In the side pane, you find the following information:

Attribute	Description
Data class name	The name of the selected data class. You can edit the name by clicking .
Data Concepts	The list of data concepts that are associated with the data class. This section is only shown if there are associated data concepts.
Data Attributes	The list of data attributes that are associated with the data class. This section is only shown if there are associated data attributes.
<Data class> Columns	The list of columns that are classified with the selected data class. When there are too many columns to show, you can follow a **See all** link. This opens a search results page with all corresponding columns. This section is only shown if there are columns with the selected data class.

Add data classes

Collibra contains a large number of packaged data classes. If a certain data class is not available, you can add it. Data classes that are defined manually are user-defined data classes.

> Tip You can also create new data classes from a Table or Column asset.

Prerequisites

- You have configured Automatic Data Classification via the Cloud Data Classification Platform or via Edge.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

Steps

1. On the main menu, click ⠿, then Stewardship.
2. In the submenu, click **Data Classification**.
   » The table with all data classes is shown.
3. Above the table to the right, click **Add**.
4. Enter the name of a data class and press `Enter`.
   The name of the data class is case-sensitive and it can contain spaces.
   You can enter multiple data classes.
5. Click **Create**.
   » The classes are added.
   If you are using Jobserver, the classes are automatically sent to the Cloud Data Classification Platform.
   If you are using Edge, the classes are not used to retrain the classification process.
6. Optionally you can link the new classes to a Data Concept or Data Attribute asset.
   a. In the **Data Concept** column, click ✎.
   b. Click in the **Select** field.
      » The list with existing Data Concept assets appears.
   c. Select one or more Data Concept assets from the drop-down list and click ✓.
   d. Do the same in the **Data Attribute** column.

Merge data classes

You can merge two or more data classes via the Data Classification Dashboard. For example, if you have the data classes Email, E-mail and email address, then it is recommended to merge them into the packaged data class Email.

Not only will it keep your data classes list clean, but it will give better results when Collibra performs data classification on ingested data.

> Note You cannot merge two or more packaged data classes, but you can merge user-defined data classes in a packaged data class. Packaged data classes appear in the **Created By** column as *System User*.

Prerequisites

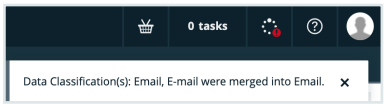
- You have configured Automatic Data Classification via the Cloud Data Classification Platform or via Edge.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

Steps

1. On the main menu, click , then Stewardship.
2. In the submenu, click **Data Classification**.
3. Select the checkboxes next to the data classes you want to merge.
4. Above the table, click **Merge**.
5. Select the data class you want to merge the selected data classes into.

   > Note
   > - You cannot merge packaged data classes and you can also not merge a packaged data class into a user-defined data class.
   > - The data class attributes Columns Count, Data Concept and Data Attributes are also merged. You can update the list of Data Concepts and Data Attributes after the merge.

6. Click **Merge**.



Edit data classes

You can edit the name of a data class via the Data Classification Dashboard side pane.

Prerequisites

- You have configured Automatic Data Classification via the Cloud Data Classification Platform or via Edge.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

Steps

1. On the main menu, click , then Stewardship.
2. In the submenu, click **Data Classification**.
   » The table with all data classes is shown.
3. Click in the row of the data class that you want to edit.
   » The data class information appears in the side pane.
4. In the side pane, click next to the data class name.
5. Enter a new name.
6. Click **Save**.
   » The name of the data class is updated.

Delete a data class

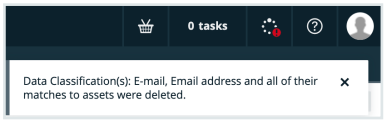
You can delete a data class from the Data Classification dashboard if it has become obsolete. Note that this is an irreversible action.

Prerequisites

- You have configured Automatic Data Classification via the Cloud Data Classification Platform or via Edge.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

Steps

1. On the main menu, click ⠿, then Stewardship.
2. In the submenu, click **Data Classification**.
3. Select the checkboxes next to the data classes you want to delete.
   You cannot delete packaged data classes. These data classes appear in the **Created By** column as *System User* or in the **User Defined** column with ✕.
4. Above the table, click **Delete**.
5. Click **Delete Data Classification**.



Connect data classes to data layers

You can use the Classification Dashboard to connect data classes to the logical and conceptual data layers.

Prerequisites

- You have configured Automatic Data Classification via the Cloud Data Classification Platform or via Edge.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

Steps

1. On the main menu, click ⠿, then Stewardship.
2. In the submenu, click **Data Classification**.
3. In the **Data Concept** or **Data Attribute** column, click ✎.
4. Click in the **Select** field.
   » The list with existing Data Concept or Data Attribute assets is shown.

5. Click ✔.

   » The Classification Dashboard creates a relationship between the data class and the logical and conceptual data layers. Column assets that have this data class will be connected to these data layers via their mutual relationship to the data class. Direct relationships between physical and logical information can then be created via Collibra workflows or other methods.

Guided Stewardship

Guided Stewardship is a set of features designed to help Data Stewards simplify the process of creating connections between physical data assets and their associated logical and conceptual assets. By establishing reliable and fully-connected data structures within your Collibra environment, you can trace relationships across all layers of representation and understand your data in a more complete way.

Guided Data Stewardship operating model

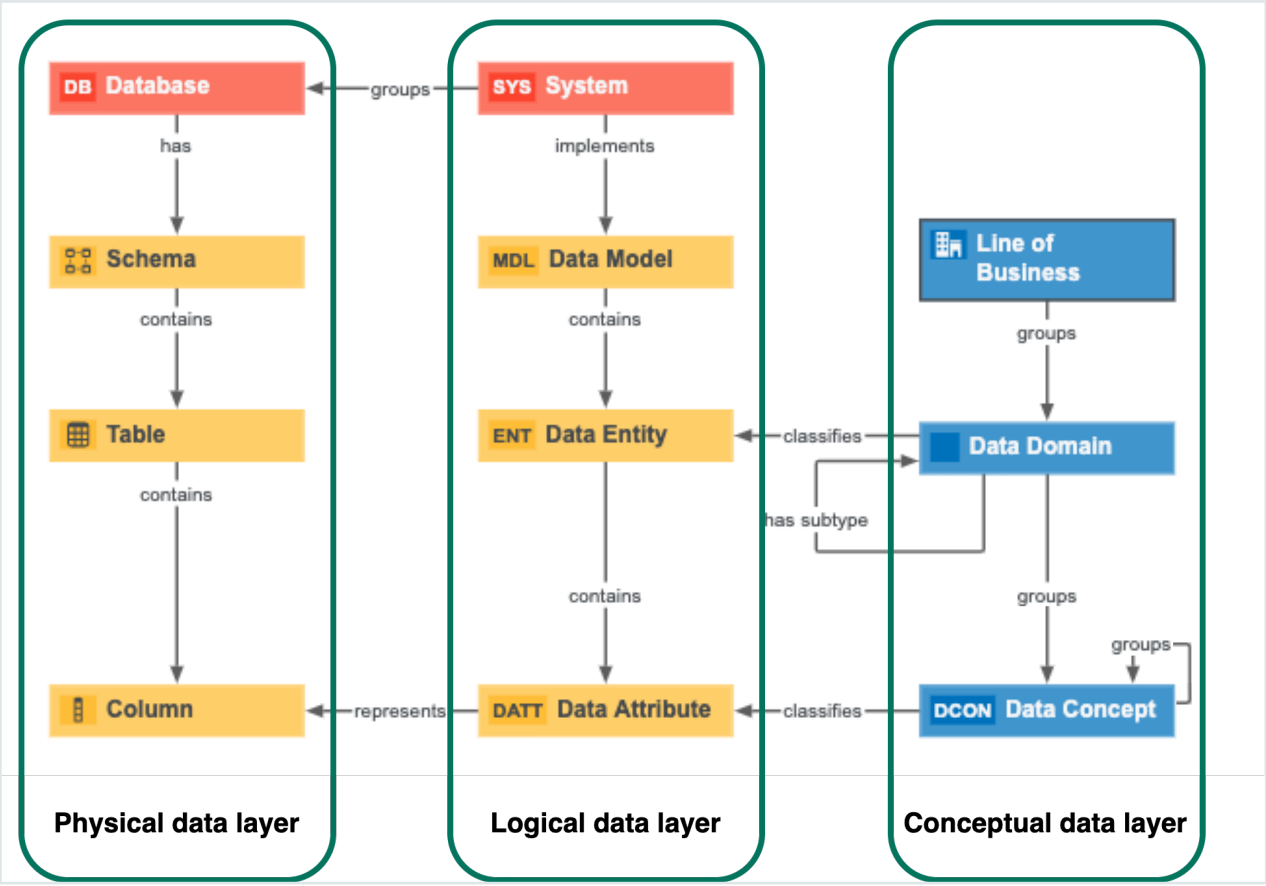
The Guided Data Stewardship operating model defines the structure of the information in Catalog. For this reason, the Guided Data Stewardship operating model is sometimes also referred to as the Data Catalog operating model.

Three data layers

The operating model consists of three data layers, representing the three different structural data layers that exist in typical organizations:

- The conceptual data layer represents the overarching structure of objects and elements in your data landscape.
- The logical data layer represents the context-dependent data structures in your organization.
- The physical data layer represents the actual data in your data environment.

The following image shows a complete view of the Data Catalog operating model. It identifies all of the relevant asset types, per data layer, and the relationships that bind them together in the Collibra Data Governance Center.



> Note Database and System assets are Technology assets that represent the highest level over physical data and logical data organization.

Conceptual data layer

The conceptual data layer is the highest level of organization in the Data Catalog operating model. It represents the overarching structure of objects and elements within an organization's data landscape. It is where you define concepts, such as Customer and Product and their component fields, without direct reference to system-specific implementations.

Organization of the conceptual data layer is based on many-to-many relationships, which makes the conceptual data layer more concise and flexible than tree-like arrangements that rely strictly on one-to-one and one-to-many relationships.

The conceptual data layer consists of the following asset types:

- Line of Business
- Data Domain
- Data Concept

Line of Business asset type

The Line of Business asset type is the highest level of abstraction in the conceptual data layer. Also known as business unit or business area, it represents a specific area of business in an organization.

> Example Finance, Sales, Retail, Investment Management

Key relation type

Line of Business assets are:

Related to...	Via the relation type...	Description
Data Domain assets	Line of Business groups / is grouped by Data Domain	Many-to-many relation, whereby: • A Line of Business asset can group many Data Domain assets. • A Data Domain asset can be grouped by many Line of Business assets.

Data Domain asset type

Data domains, also known as data categories or subject areas, are high-level, theoretical representations of your data. They represent the structure of concepts in data environments and contain all the different nuances of corresponding business terms.

> Example Customer, Employee, User, Order, Product

Key relation types

Data Domain assets are:

Related to...	Via the relation type...	Description
Line of Business assets	Business Asset groups / is grouped by Business Asset	Many-to-many relation, whereby: • A Line of Business asset can group many Data Domain assets. • A Data Domain asset can be grouped by many Line of Business assets.
Data Concept assets	Business Asset groups / is grouped by Business Asset	Many-to-many relation, whereby: • A Data Domain asset can group many Data Concept assets. • A Data Concept asset can be grouped by many Data Domain assets.
Other Data Domain assets	Data Domain has subtype / is subtype of Data Domain	One-to-many relation, whereby: • A Data Domain asset can have many subtype Data Domain assets. • A Data Domain asset can be the subtype of only one Data Domain asset.

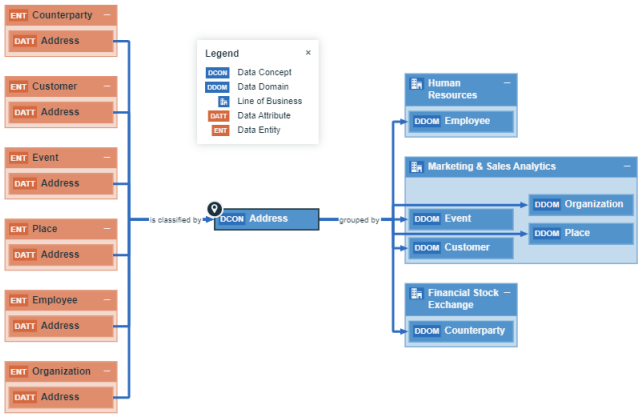
Data Concept asset type

A Data Concept asset is a high-level theoretical representation of your data and describes one aspect of one or more data domains. These assets represent the most common concepts that are used to organize database content. They allow users to define a context-independent representation of the structure of an organization's data.

They are the most granular level of context-independent structure users can establish within the conceptual data layer, and are comparable to columns in the physical data layer.

> Example Address, Name, ID number, Phone number, Price, Year

For example, if you have a Data Concept asset for Address then this might correlate to a Data Entity asset for Customer Address, Supplier Address and Employee Address.



Key relation types

Data Concept assets are:

Related to...	Via the relation type...	Description
Data Domain assets	Business Asset groups / grouped by Business Asset	Many-to-many relation, whereby: • A Data Concept asset can be grouped by many Data Domain assets. • A Data Domain asset can group many Data Concept assets.
Other Data Concept assets	Business Asset groups / grouped by Business Asset	Many-to-many relation, whereby: • A Data Concept asset can group, and be grouped by, many Data Concept assets.
Data Attribute assets	Business Dimension classifies / is classified by Asset	Many-to-one relation, whereby: • A Data Concept asset can classify many Data Attribute assets. • A Data Attribute asset can be classified by only one Data Concept asset.

Organization based on many-to-many relations

The conceptual data layer is organized such that the relationships between Lines of Business and Data Domain assets, and between Data Domain and Data Concept assets,

are many-to-many relationships.

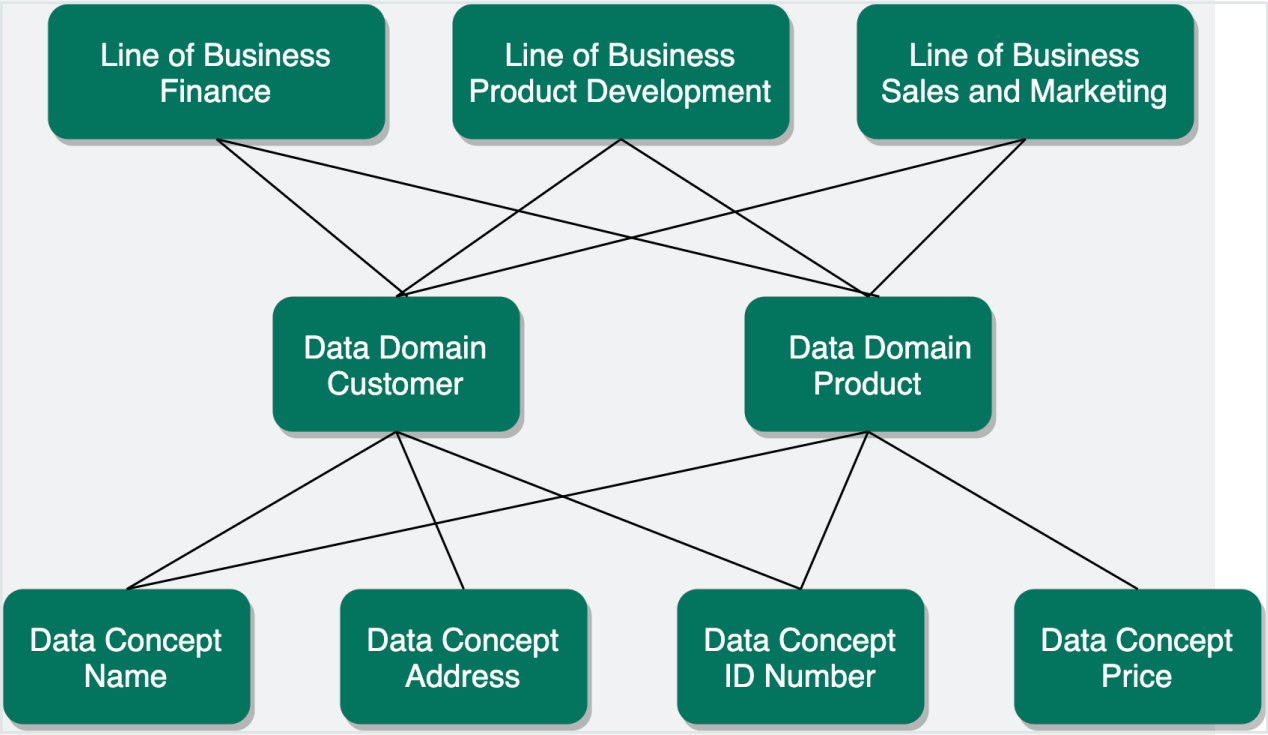
This graph-based approach, based on many-to-many relationships, makes the conceptual data layer more concise and flexible.

Example

In this example, we've identified three lines of business, each of which groups both the Customer data domain and Product data domain. In turn, each data domain groups several data concepts, some of which are grouped by both data domains.

Both data domains group the Name and ID Number data concepts. This is conceivable because Name and ID Number, as Data Concept assets, are abstract representations of these two concepts, rather than specific implementations of them, which are described in the logical data layer and implemented by System assets.

In this way, information stored in the conceptual data layer is kept to a minimum and the Data Domain and Data Concept assets are referred to as often as necessary.



In summary, Line of Business, Data Concept and Data Domain assets are independent assets that do not, by nature, encapsulate or organize the structure of other assets. The Name and ID Number Data Concept assets exist independently of the Data Domain assets that group them. A Customer can have a Name and a Product can have a Name, but you need only one Data Concept asset to encapsulate the idea of “name”.

Conceptual data layer versus the Business Glossary

This section examines the differences and relation between the conceptual data layer and the Collibra Business Glossary.

Business terms: context-dependent representations of business concepts

In short, the Business Glossary is a system that helps organizations govern their business terms.

> Example Let's consider the business term Customer, within a multinational consumer goods organization that deals with different consumer groups in different cultural contexts. This organization uses business terms to create a shared understanding of Customer, across different geographical regions. Its offices around the world create their own business terms to encapsulate the specific cultural complexity of a customer, in their own way. Its various business units also have their own definitions, to address different operational, legal and compliance demands.

Business terms are a flexible tool that account for complex business and organizational structures. Anything can be represented by a business term, including the nuanced representations specific to different languages, cultures and branches of business.

Data, on the other hand, can be more explicitly defined and grouped. While there may be several ways to describe Customer, based on cultural and geographic nuance, when we consider data, a customer can be uniquely identified, defined and grouped. This is where the conceptual data layer comes in.

The conceptual data layer: context-independent representation of the structure of data

A data domain is a container for other data domains and data concepts that encompass associated terminology and definitions that an organization intends to govern.

> Example Customer Master Data, Product Master Data, Reference Data

While business terms represent Customer in the context of a specific language, culture or branch of business, a customer data domain represents the structure of Customer in a data environment, and encapsulates all of the different nuances of the business term. By abstracting the idea of Customer in a data domain, one can start to consider how customers can be represented by physical data.

The same applies to data concepts, such as Year, Date, Address, and Name. While there may be many business terms that represent Year, across different teams and geographies, the data concept encapsulates all of them and creates a layer of abstraction that allows you to define high-level data structures.

Logical data layer

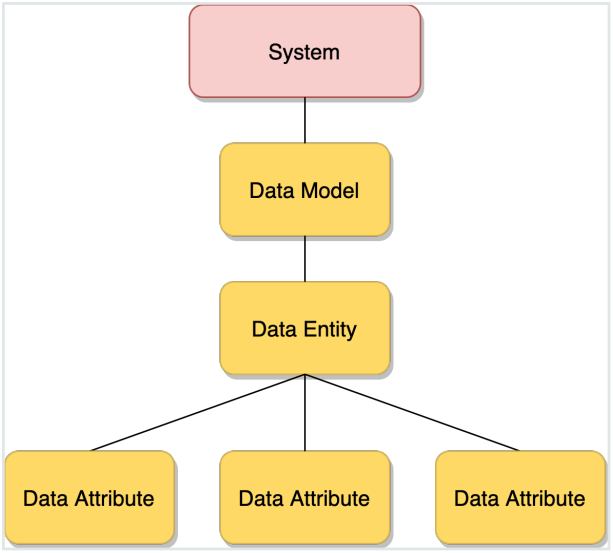
The logical data layer defines data structures within an organization's systems, whereas the conceptual data layer represents context-independent data structures within an organization.

The Data Entity-Data Attribute structure is closely related to the Data Domain-Data Concept structure of the conceptual data layer. The main difference between the two is that the conceptual data layer is context-independent, whereas the logical data layer describes the structure in an individual System.

The logical data layer consists of the following asset types:

- Data Model
- Data Entity
- Data Attribute

The logical data layer can be visualized as a tree-like structure, starting with a high-level System and Data Model assets, and branching out with implementation-specific Data Entity and Data Attribute assets.



> Note Although the System asset type is a Technology Asset, it adds higher-level structure to the logical data layer and is considered part of the logical data layer.

Data Model asset type

The Data Model asset is the highest level of organizational structure in the logical data layer, and defines the specific structure of data in a System.

Key relation types

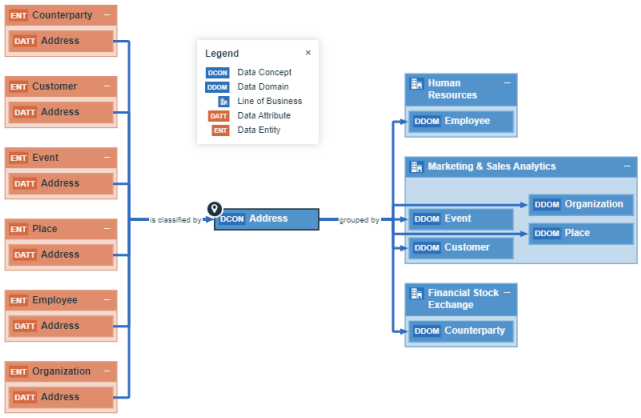
Data Model assets are:

Related to...	Via the relation type...	Description
System assets	System implements / is implemented in Data Model	One-to-one relation, whereby: • A System asset can implement only one Data Model asset. • A Data Model asset can be implemented in only one System asset. Note The one-to-one nature of this relationship is what makes Data Models – and, therefore, the entire logical data layer – context-dependent, as opposed to the context-independent conceptual data layer.
Data Entity assets	Data Model contains / is contained in Data Entity	One-to-many relation, whereby: • A Data Model asset can contain many Data Entity assets. • A Data Entity asset can be contained in only one Data Model asset.

Data Entity asset type

Data Entity assets are the logical data layer and correlate to Data Domain assets of the conceptual data layer. Data Entity assets can be thought of as system-specific implementations of Data Domain assets.

For example, if you have a Data Concept asset for Address then this might correlate to a Data Entity asset for Customer Address, Supplier Address and Employee Address.



Key relation types

Data Entity assets are:

Related to...	Via the relation type...	Description
Data Model assets	Data Entity is part of / contains Data Model	One-to-many relation, whereby: • A Data Entity asset can be part of or contained in only one Data Model asset. • A Data Model asset can contain multiple Data Entity assets.
Data Domain assets	Data Domain (Business Dimension) classifies / is classified by Data Entity (Asset)	Many-to-many relation, whereby: • A Data Domain asset can classify many Data Entity assets. • A Data Entity asset can be classified by many Data Domain assets.
Data Attribute assets	Data Entity contains / is part of Data Attribute	One-to-many relation, whereby: • A Data Entity asset can contain many Data Attribute assets. • A Data Attribute asset can be part of or contained in only one Data Entity asset.

Data Attribute asset type

Data Attribute assets are the logical data layer and correlate to Data Concept assets of the conceptual data layer. They can be thought of as system-specific implementations of Data Concept assets.

Key relation types

Data Attribute assets are:

Related to...	Via the relation type...	Description
Data Entity assets	Data Entity contains / is part of Data Attribute	One-to-many relation, whereby: • A Data Entity asset can contain many Data Attribute assets. • A Data Attribute asset can be contained by only one Data Entity asset.
Data Concept assets	Data Concept classifies / is classified by Data Attribute	One-to-many relation, whereby: • A Data Concept asset can classify many Data Attribute assets. • A Data Attribute asset can be classified by only one Data Concept asset.

Physical data layer

The physical data layer represents the actual data – the schemas, tables and columns – in an organization's systems.

The physical data layer consists of the following asset types:

- Schema
- Table
- Column

Note

- Although the Database asset type is a Technology Asset, it is considered part of the physical data layer.
- The Schema, Table and Column assets in a Collibra Data Intelligence Cloud environment are almost never created manually; rather, they are automatically created via the Data Catalog ingestion process, when registering a data source.

Schema asset type

A Schema is the highest level of physical structure in a Database. It defines, in a formal language, the structure of the tables and columns in the database.

Key relation types

Schema assets are:

Related to...	Via the relation type...	Description
Database assets	Database has / belongs to Schema	One-to-many relation, whereby: • A Database asset can have many Schema assets. • A Schema asset can belong to only one Database asset.
Table assets	Schema contains / is part of Table	One-to-many relation, whereby: • A Schema assert can contain many Table assets. • A Table asset can be part of only one Schema asset.

Table asset type

Table assets represent the physical tables in a data environment.

Key relation types

Tables assets are:

Related to...	Via the relation type...	Description
Schema assets	Table is part of / contains Schema	One-to-many relation, whereby: • A Table asset can be a part of only one Schema asset. • A Schema asset can contain many Table assets.
Column assets	Table contains / is part of Column	One-to-many relation, whereby: • A Table asset can contain many Column assets. • A Column asset can be a part of only one Table asset.

Column asset type

Column assets represent the physical columns in a data environment. It is the lowest level of definition in the physical data layer.

Key relation types

Column assets are:

Related to...	Via the relation type...	Description
Table assets	Column is part of / contains Table	One-to-many relation, whereby: • A Column asset can be a part of only one Table asset. • A Table asset can contain many Column assets.
Data Attribute assets	Data Attribute represents / represented by Column	One-to-many relation, whereby: • A Data Attribute asset can represent many Column assets. • A Column asset can be represented by only one Data Attribute asset.

Technology Assets

Two Technology Assets are included in the Data Catalog operating system:

- System, which is part of the logical data layer.
- Database, which is part of the physical data layer.

Database asset type

Database assets represent the physical databases in your data environment. They are the highest level of physical data organization in a data environment. Database assets should have specific names, and implement specific technologies, such as PostgreSQL.

Key relation types

Database assets are:

Related to...	Via the relation type...	Description
System assets	System groups / is grouped by Database	One-to-many relation, whereby: • A System asset can group many Database assets. • A Database asset can be grouped by only one System asset.
Schema assets	Database has / belongs to Schema	One-to-many relation, whereby: • A Database asset can have many Schema assets. • A Schema asset can belong to only one Database asset.

System asset type

System assets represent executable software that an organization uses to automate business functions that help run the business smoothly and efficiently. Systems can be any commercially available or privately developed software that is running in your environment.

> Example CRM, ERP and EDW software

Key relation types

System assets are:

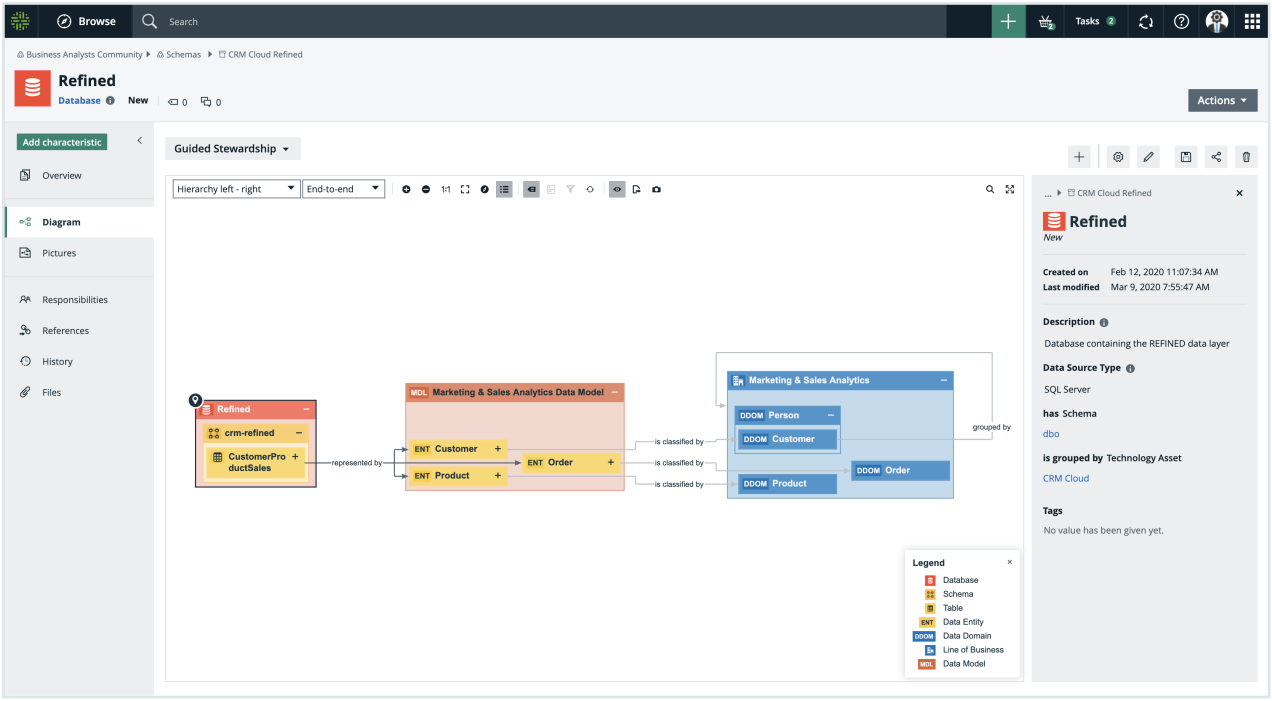
Related to...	Via the relation type...	Description
Data Model assets	System implements / is implemented in Data Model	One-to-one relation, whereby: • A System asset can implement only one Data Model asset. • A Data Model asset can be implemented by only one System asset.
Database assets	System groups / is grouped by Database	One-to-many relation, whereby: • A System asset can group many Database assets. • A Database asset can be grouped by only one System asset.

Guided Data Stewardship diagram views

For assets in the Guided Data Stewardship operating model, there are two packaged diagram views: Guided Data Stewardship and Guided Data Stewardship - Data Concept. These diagram views show the relation types that bind assets, as established through the Physical Data Connector.

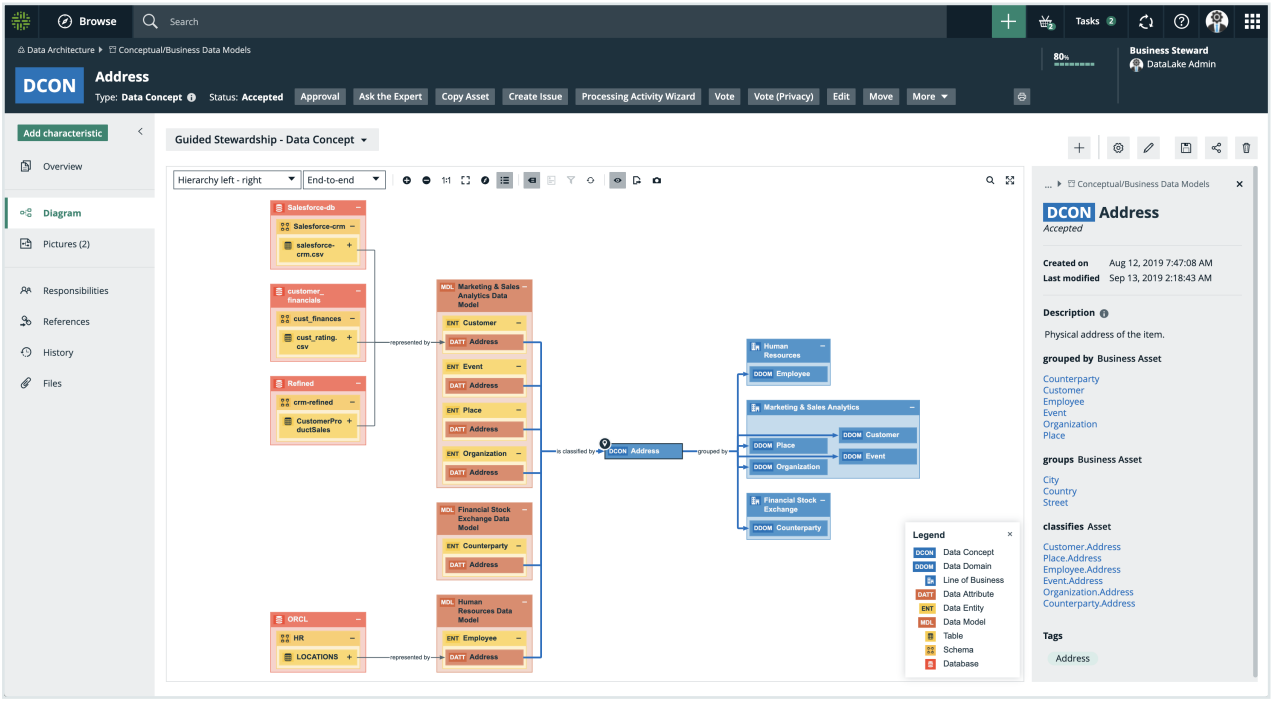
Guided Data Stewardship view

The Guided Data Stewardship view is the default diagram view designed to help you visualize direct and indirect relations across the entire data environment. For the logical data layer, this view shows the relation types that bind the Data Model, Data Entity, and Data Attribute assets. For the conceptual data layer, it shows the Line of Business and Data Domain assets.



Guided Data Stewardship- Data Concept view

The Guided Data Stewardship - Data Concept view is the default diagram view for Data Concept assets only. This diagram view shows the logical and physical data associated with a Data Concept.



For more information, see Diagram views.

Physical Data Connector

The Physical Data Connector shows a high-level overview of database information on which you can filter.

You can use the Physical Data Connector to:

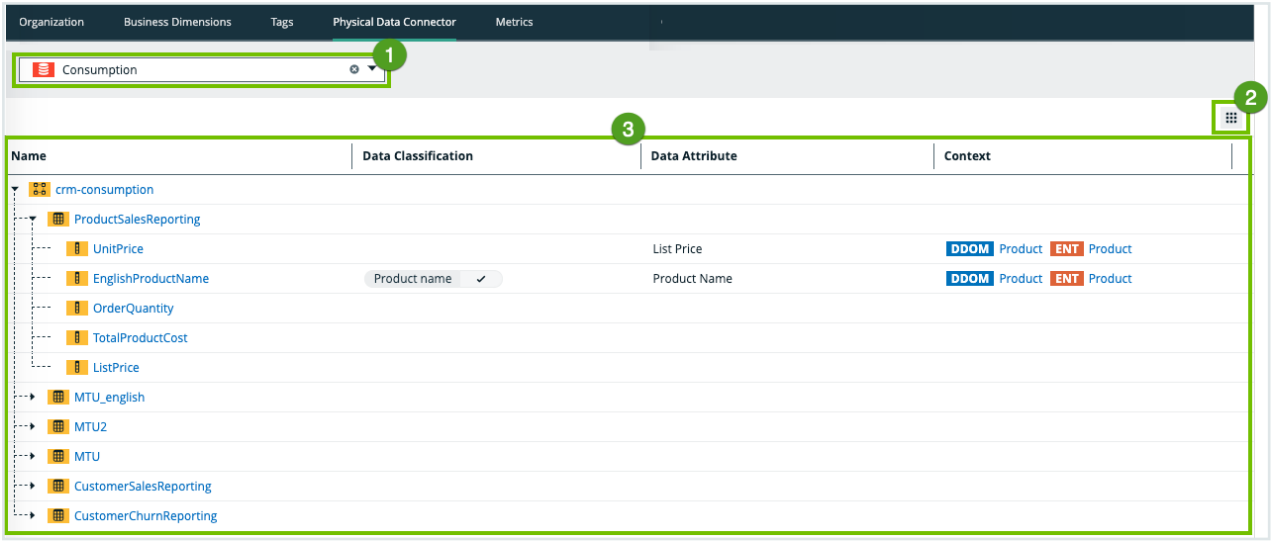
- Connect the Data Catalog physical data layer to the logical data layer.
- Manually classify columns.

About the Physical Data Connector

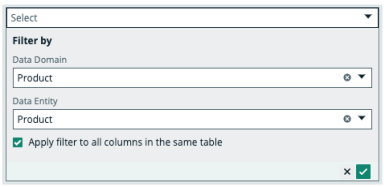
The Physical Data Connector shows a table with a high-level overview of database information. The table has a tree-like structure that enables you to drill down to the column level of a database. It shows the connection between the physical data layer and the logical data layer and enables you to find Data Attribute assets that relate to individual Column assets.

> Note In the physical data connector, a Column asset is only visible if it has a parent Table asset and a parent Schema asset. For more details, go to Guided Data Stewardship operating model. A parent Database asset is not required.

You access the Physical Data Connector via the Physical Data Connector subpage on the Stewardship tab.



No.	Name	Description
1	Drop-down	A drop-down list to filter on a specific database.
2	Table menu	The table menu contains buttons for actions you can perform on the table.
	⠿	A button to manage the columns shown.

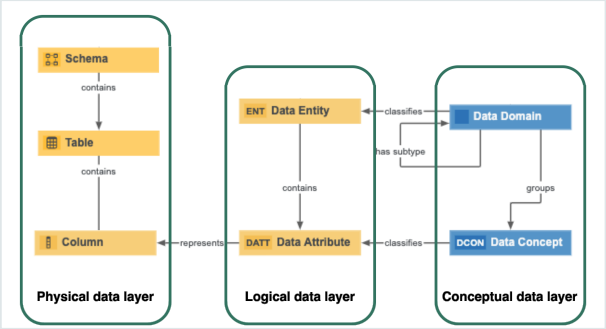
No.	Name	Description
3	Table with database information	A table that shows the content of the registered database and the connections between the physical data layer and logical data layer.
	Name	The name of the asset and the icon of the asset type. If you click on the asset, the asset page opens. To sort assets alphabetically, click on the column header.
	Data Classification	The data class of an asset. You can manually add, edit or remove the data class of a Column asset. You can also approve or reject suggested classes
	Data Attribute	The Data Attribute asset linked to the Column asset via relation type "Data Attribute represents / represented by Column". When you filter on a Data Domain or Data Entity, the other drop-down lists dynamically update to only show content that relates to your filter. You can select the **Apply filter to all columns in the same table** checkbox to use the same filters to link a Data Attribute to other Column assets in the same table.  Tip The physical data connector enables you to quickly connect Data Attribute assets to Column assets. However, you can also connect the physical data layer to the logical data layer via Data Catalog's asset pages by adding a relation of the type Data Attribute represents / represented by Column.
	Context	The context of the data. This field is read-only and is filled with the Data Domain asset and Data Entity asset related to the Data Attribute asset, if a relation exists.

Physical Data Connector relation types

The Physical Data Connector enables you to easily connect the physical data layer to the logical data layer by filtering on the conceptual data layer.

The Physical Data Connector uses the following relation types to connect assets from the different data layers:

- Business Dimension (Data Domain) classifies / is classified by Asset (Data Entity)
- Business Asset (Data Domain) groups / grouped by Business Asset (Data Concept)
- Data Domain has subtype / is subtype of Data Domain
- Business Dimension (Data Concept) classifies Asset (Data Attribute)
- Data Entity contains Data Attribute
- Data Attribute represents Column
- Schema contains Table
- Table contains Column



Manually classify columns

The Physical Data Connector enables you to manually add, edit or remove a data class of a Column asset. This is useful, for example, if Automatic Data Classification missed some data classes.

> Tip You can also automatically classify all columns in a table using Automatic Data Classification.

Prerequisites

- You have configured Automatic Data Classification for the DGC service.
- You have the correct permissions to classify tables and columns.
- You have registered a data source.
- Data Catalog experience is enabled in the DGC service configuration.

Steps

1. On the main menu, click ⠿, then Stewardship.
2. In the submenu, click **Physical Data Connector**.
3. In the drop-down list, filter on a database.
   » The table shows all ingested schemas in the database. You can use the asset tree to drill down to the column level of the database.
4. In the asset tree, find the Column asset that you want to classify.
5. In the Data Classification column, click ✎.
6. Click in the **Select** field.
   » The list with existing data classes appears.
7. In the **Select** field, use the drop-down list to find a data class or enter a new data class name and press `Enter`.

   > Note
   > - Data classes are case-sensitive.
   > - You can add more data classes if applicable, but avoid it as much as possible.
   > - If you created a new data class, it is automatically sent to the Data Classification Platform.
   > - We recommend that you only add one data class to a column.

8. Click ✔.
   » The data class is automatically accepted (✔).

Connect physical data to logical data

You can use the Physical Data Connector to easily connect a Column asset to a Data Attribute asset via the relation type Data Attribute represents / represented by Column.

A Column asset represents the lowest level of the physical data layer, while a Data Attribute asset represents the lowest level of the logical data layer.

> Tip You can also add a relation of the type Data Attribute represents / represented by Column via a Data Attribute's or Column's asset page.

Prerequisites

- You have registered a data source.

Steps

1. On the main menu, click , then Stewardship.
2. In the submenu, click **Physical Data Connector**.
3. In the drop-down list, filter on a database.
   » The table shows all ingested schemas in the database. You can use the asset tree to drill down to the column level of the database.
4. In the asset tree, find the Column asset that you want to link to a Data Attribute asset.
5. In the **Data Attribute** column, click .
   » A Data Attribute drop-down list with two filters appears.
6. Link a Data Attribute asset to the Column asset based on the Data Domain and Data Entity filter.
   a. Optionally, select a Data Domain asset and Data Entity asset that are related to the Data Attribute.
      » When you filter on a Data Domain asset or Data Entity asset, the other drop-down lists are dynamically updated to only show content related to your filter.
   b. If you want to use the same filters to find Data Attribute assets for other Column assets in the same table, select the **Apply filter to all columns in the same table** checkbox.
   c. Select the correct Data Attribute asset in the drop-down list.

      > Note You can only select one Data Attribute asset. The Data Attribute asset must exist in your Collibra environment.

   d. Click ✓ to accept the Data Attribute asset.

» The Data Attribute asset is now linked to the Column asset via the relation type "Data Attribute represents / represented by Column". This relation is also shown on the asset pages of the Column and Data Attribute assets.
» If there is a Data Domain asset and Data Entity asset that is related to the Data Attribute asset, they are shown in the Context column. If you used the filters in the Data Attribute column, the same assets as your filters are shown in the Context column.

Warning If you click 🗑 to delete a Data Attribute asset in the physical data connector overview, you also delete the relation between the Column asset and the Data Attribute asset from the respective asset pages.

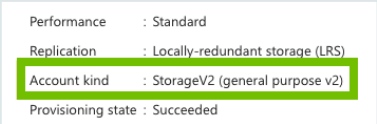
Working with Azure Data Lake Storage

About the Azure Data Lake Storage file system integration

The Azure Data Lake Storage file system integration allows for the registration of Azure Data Lake Storage (ADLS) as a data source in Collibra and the synchronization of the metadata. ADLS is a service provided on Microsoft Azure Blob Storage. After the integration, the files and directories of the ADLS file system are represented in Collibra by specific asset types, retaining the original names.

Important

- The ADLS integration supports Azure Data Lake Storage Gen2.
  Azure Data Lake Storage Gen1 is not supported. To verify which Azure version you are using, check the **Account Kind** in the Overview section in your Azure storage account details. StorageV2 indicates you are using Gen2.



- You can integrate an Azure Data Lake Storage file system only via Edge.

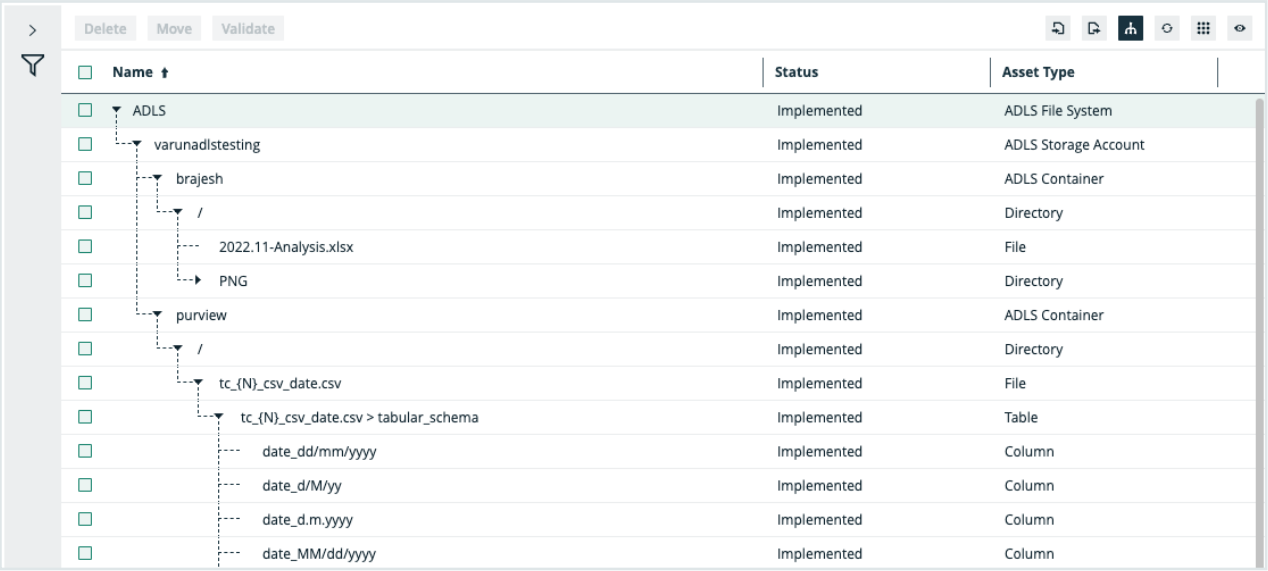
For detailed information on Microsoft Azure Data Lake Storage Gen2, go to the Azure documentation.

About Microsoft Purview

The ADLS integration supports Microsoft Purview, a service used for schema discovery. This allows you to integrate the schemas, tables and columns from the files into one single File asset in Collibra rather than multiple File assets. For more details, go to the ADLS operating model.

Important

- Even if you use Microsoft Purview to integrate schemas and tables, we don't currently support profiling and sampling.
- Currently, the ADLS integration can ingest up to 100,000 assets from Purview.



For detailed information on Microsoft Purview, go to the Purview documentation.

What's Next?

Steps overview: Integrate an Azure Data Lake Storage file system

To learn about the ADLS integration and watch videos, follow our University course.

Azure Data Lake Storage asset types and operating model

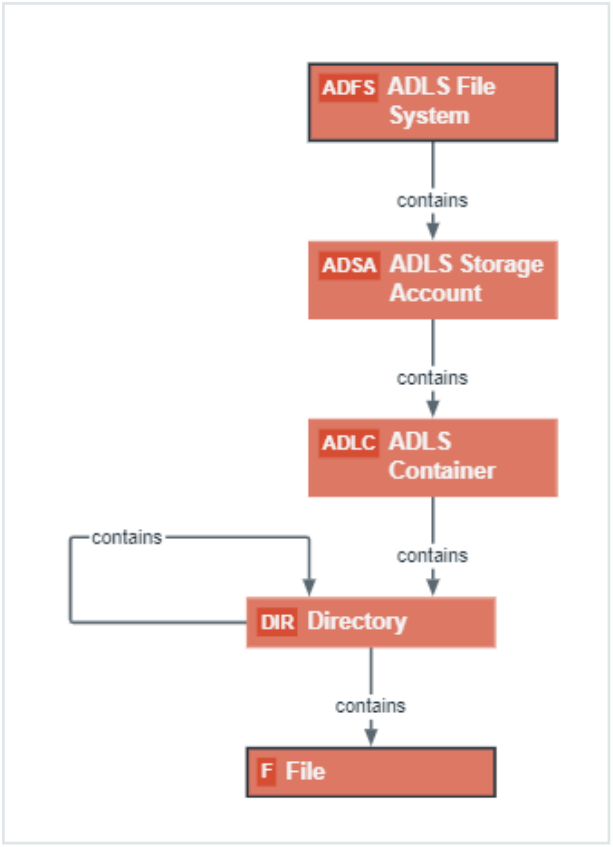
The Azure Data Lake Storage file system integration of Collibra Platform Self-Hosted uses a specific subset of asset types. All of these come out-of -the-box with your software.

Asset type	Description
Technology Asset ▸ System ▸ File Storage ▸ ADLS File System	An asset type that represents Azure Data Lake Storage file system.

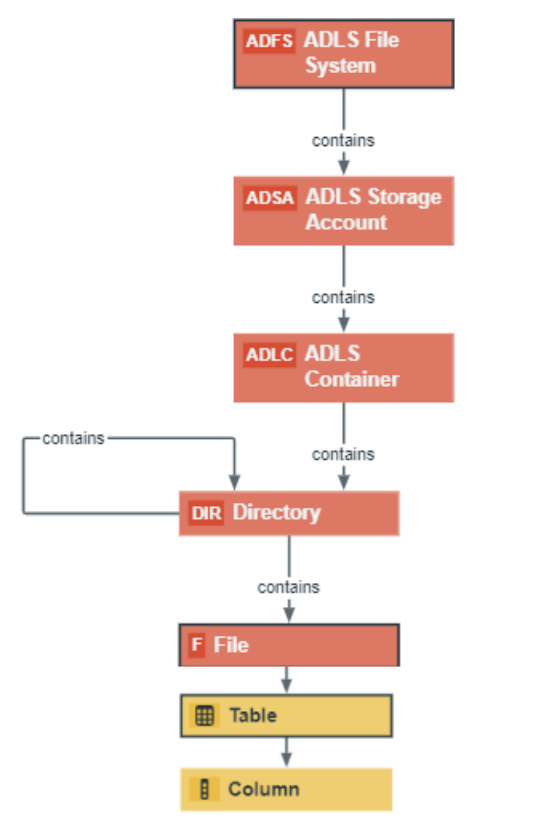
Asset type	Description
Technology Asset ▸ File Container ▸ ADLS Storage Account	An asset type that represents an Azure Data Lake Storage Account, which is a logical unit of storage containing Azure Data Lake Container objects.
Technology Asset ▸ File	A piece of information technology (hardware, software, database, software platform) that helps an organization to run a business application.
Technology Asset ▸ File Container ▸ ADLS Container	An asset type that represents an Azure Data Lake Container, which is a logical unit of storage containing Azure Data Lake Storage objects.
Technology Asset ▸ File Container ▸ Directory	A collection of data that is treated by a computer as a unit, for the purposes of input and output.
Data Asset ▸ Data Structure ▸ Table	An implementation of data entities in columns and rows, in a given database system. It is the basic structure of a relational database. Examples: Account_tbl, CUST_ADDR
Data Asset ▸ Data Element ▸ Column	An atomic unit of data that can be stored in a database table. Examples: FST_NM, EMPID

ADLS operating model

The following image shows the relations between ADLS asset types and the cardinality of the relation types in the assets' assignment.


ADFS ADLS File System
contains
ADSA ADLS Storage Account
contains
ADLC ADLS Container
contains
contains
DIR Directory
contains
F File

Note
If you use Microsoft Purview, a File asset can contain Table and Column assets.



For information on the data that is integrated, go to Integrated Azure Data Lake Storage data.

Steps overview: Integrate an Azure Data Lake Storage file system

Before you begin

If you want to use Microsoft Purview to ingest schemas and table, you need to register and scan your ADLS folders in Microsoft Purview. For more information, go to the Microsoft Purview documentation.
For an overview of the supported file types for scanning, go to the Microsoft Purview documentation.

Steps

#	Step	Description
1	Enable the ADLS synchronization via Edge and give the Edge site user the required permissions.	Allows the integration of an ADLS file system via Edge.
2	Create an ADLS connection to Edge.	Creates a connection to Azure in your Edge site.
3	Add the ADLS synchronization capability to Edge.	Adds the ADLS synchronization capability to the Azure Edge connection. The capability allows to retrieve data from the ADLS file system.
4	Register the ADLS file system.	Creates the initial structure of a Storage Catalog domain and ADLS File System asset in the selected parent community.
5	Connect the ADLS file system asset to the Edge ADLS capability.	Links the registered ADLS file system to the Edge capability.
6	Create crawlers.	Creates crawlers that define the folders you want to synchronize.
7	Synchronize the ADLS file system.	You can manually synchronize the ADLS file system or you can add a synchronization schedule to automatically synchronize it. As a result, the ADLS metadata is integrated.

To learn about the ADLS integration and watch videos, follow our University course.

Enable the Azure Data Lake Storage file system integration via Edge

You can enable the integration of an Azure Data Lake Storage (ADLS) file system via Edge.

Prerequisites

- You have the **ADMIN** or **SUPER** role in Collibra Console.
- You have the **SUPER** role in Collibra Console.
- You have the **ADMIN** or **SUPER** role in Collibra Console.

Steps

1. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.
   c. In the tab pane, click the Data Governance Center service of that environment.
   d. Click **Configuration**.
   e. Click **Edit configuration**.
2. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.
   c. In the tab pane, click the Data Governance Center service of that environment.
   d. Click **Configuration**.
   e. Click **Edit configuration**.
3. In the **Register data source** section, enter the required information:

Setting	Description
ADLS synchronization via Edge	An option to enable the integration of an Azure Data Lake Storage (ADLS) file system via Edge. ◦ ✓ True: You can integrate an Azure Data Lake Storage (ADLS) file system via Edge. ◦ ✕ False: You cannot integrate an Azure Data Lake Storage (ADLS) file system via Edge. Set this option to True.

4. Click **Save all**.

What's next?

You can now create a connection to Azure in your Edge site.

Add the ADLS synchronization capability

After you have created a connection to the Azure Data Lake Storage (ADLS) file system in your Edge site, you have to add the ADLS synchronization capability to the connection.

Before you begin

- You have created and installed an Edge site.
- You have created a connection to ADLS in your Edge site.

Required permissions

You have a global role that has the **Manage connections and capabilities** global permission, for example, Edge integration engineer.

Steps

1. Open an Edge site.
   a. On the main menu, click , and then click **Settings**.
      » The Collibra settings page opens.
   b. In the tab pane, click **Edge**.
      » The **Sites** tab opens and shows a table with an overview of the Edge sites.
   c. In the table, click the name of the Edge site whose status is **Healthy**.
      » The Edge site page opens.
2. In the **Capabilities** section, click **Add capability**.
   » The **Add capability** page is shown.
3. Enter the required information.

Field	Description	Required
Capability	This section contains general information about the capability.	

Field	Description	Required
Name	The name of the Edge capability.	✓ Yes
Description	The description of the Edge capability.	✗ No
Capability template	The capability template. The value that you select in this field determines which sections appear on the page. Select the following Edge capability: `ADLS synchronization`	✓ Yes
ADLS service account	This section contains the information on how to connect to Azure Data Lake Storage.	
Azure Connection	The ADLS connection to be used.	✓ Yes
Microsoft Purview Account Name	The name of your Microsoft Purview account. If you enter a Purview account name, the integration uses Microsoft Purview for the integration.	✗ No
Save Input Metadata	If you select this option the metadata extracted from the data source will be saved in a file that can be used for troubleshooting. Select this option only on request of Collibra Support.	✗ No
Max Schema Level	For columns that have a structured technical data type, Array or Struct, you can register the structure of the data. This is supported for AVRO, CSV, JSON, ORC, PARQUET, PSV, SSV, TSV, TXT, and XML. In this field, enter the maximum level of the structure you want to see. For example, 3. > Note If you include a high number of levels, this can have an impact on the integration performance.	✗ No

Field	Description	Required
Advanced Configuration	This section contains configuration options that can help when investigating issues with the capability. Important Only complete the fields **Logging configuration**, **Memory (MiB)**, and **JVM arguments** on request of or together with Collibra Support.	✕ No

4. Click **Create**.
   » The capability is added to the Edge site.
   » The fields become read-only.

What's next?

You can now register the ADLS file system.

Registering and synchronizing Azure Data Lake Storage

Register an Azure Data Lake Storage file system

You can register an Azure Data Lake Storage (ADLS) file system in Data Catalog.

Before you begin

You have added an Edge capability with the ADLS capability template.

Required permissions

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.

Steps

1. On the main menu, click ⠿, and then click **Catalog**.
   » The Catalog Home opens.
2. On the main toolbar, click +.
   » The **Create** dialog box appears.
3. In the **Create** dialog box, click **Register a data source**.
   » The **Register content** page opens.
4. In the **Connection name** column, search for the connection name you want to use to integrate the ADLS file system.
5. Click **Add** for the connection you want to use.
   » The **Register Azure Data Lake Storage file system** page opens.
6. Enter the required information.

Field	Description
Community	The parent community in which the initial ADLS structure must be created.
File system name	The name for ADLS file system asset.
Description	The description to provide extra information about the file system. This is used as the **Description** attribute of the ADLS File System asset.
Owner	The owner name of the data in the created community.

7. Click **Register**.
   » An ADLS File System asset is created.
   » A Storage Catalog domain is created with the same name as the ADLS File System asset.

What's next?

You can now connect the ADLS File System asset to the ADLS Edge connection.

Create a crawler for Azure Data Lake Storage

By creating a crawler for Azure Data Lake Storage (ADLS), you can specify which directories you want to synchronize.

Before you begin

- You have registered an ADLS file system.
- You have connected the ADLS File System asset to the ADLS Edge capability.

Required permissions

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.

Steps

1. Open the ADLS File System asset.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, click **Create crawler**.
   » The **Create crawler** dialog appears.
4. Enter the required information.

Field	Description
Domain	The domain in which the assets of the ADLS file system are to be created.
Name	The name you want to give to the crawler in Collibra.

Field	Description
Include path	The case-sensitive path to a directory of a directory in ADLS. All objects and subdirectories of this path are taken into account during the synchronization. Use the following structure to refer to the path: `https://<storage account name>.blob.core.windows.net/<container name>/<blob name>`. Note The include path is case-sensitive. Example `https://myaccount.blob.core.windows.net/mycontainer/myblob` `https://myaccount.blob.core.windows.net/$root/myblob` refers to the root container. For information on working with root containers, go to the ADLS documentation.

Field	Description
Exclude patterns	A case-sensitive pattern that represents the objects that are included via the **Include path**, but that you want to exclude from the synchronization. When you define a pattern, you can use the following rules: ◦ `*` matches zero or more characters. ◦ `**` matches zero or more directories in a path. ◦ `?` matches one character. Note ◦ The exclude patterns are case-sensitive. ◦ The Exclude patterns apply only to files, not folders. Example ◦ `comm/*.jsp` matches all .jsp files in the comm path. ◦ `comm/t?st.jsp` matches comm/test.jsp but also comm/tast.jsp or comm/txst.jsp. ◦ `commm/**/test.jsp` matches all test.jsp files in the comm path. ◦ `org/framework/**/*.jsp` matches all .jsp files in the org/framework path. ◦ `org/**/servlet/test.jsp` matches org/framework/servlet/test.jsp but also org/framework/testing/servlet/test.jsp and org/servlet/test.jsp.
Add pattern	Button to add additional exclude patterns.
Add path	Button to add an additional Include path.

5. Click **Create**.

What's next?

You can now synchronize ADLS file system manually or define a synchronization schedule.

About synchronizing Azure Data Lake Storage

Synchronizing Azure Data Lake Storage file system (ADLS) is the process of ingesting metadata from a selected ADLS repository and making the data available in Collibra Platform Self-Hosted.

When you synchronize ADLS, the content of your repository is analyzed and represented in Collibra by means of assets and their characteristics. Collibra also takes the defined crawlers into account.

You can synchronize manually, or you can automate it by adding a synchronization schedule.

You can only synchronize one ADLS File System at a time.

- If a synchronization job is in progress and a second one is triggered, manually or automatically, it will be queued.
- If a synchronization job is still running and a new synchronization of the same ADLS File System is triggered (manually or automatically), the running synchronization will continue and the new synchronization request is ignored.

Synchronize Azure Data Lake Storage manually

You can manually start a synchronization job of an ADLS File System asset. This can be useful if you want to test your crawlers, or if you want to synchronize immediately. You can also add a synchronization schedule to synchronize automatically.

Before you begin

- You have registered an ADLS file system.
- You have connected the ADLS File System asset to the ADLS Edge capability.
- If needed, you have defined crawlers.

Required permissions

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.

Steps

1. Open the ADLS File System asset.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, click **Synchronize now**.
   » A notification indicates synchronization has started.

   » The synchronization job appears in the **Activities** list as a bulk synchronization.

   When the synchronization finishes, the resulting assets, including their attributes and relations, are created, edited or deleted in the selected domain(s) and in the Data Sources page of Data Catalog.

> Note In case of a partial synchronization caused by a temporary communication issue, the status of the assets that cannot be synchronized is set to **Missing from source**. Their previous status is restored, if they are found in the source system during the next fully successful synchronization.

What's next?

You can view a summary of the results from the Activities list.
You can view the assets in their domain.

Add an Azure Data Lake Storage synchronization schedule

To keep the content of Collibra Platform Self-Hosted synchronized with your ADLS file system, you can synchronize manually or create a schedule to automatically do this with a fixed interval.

> Note You can only create one synchronization schedule.

Before you begin

- You have registered an ADLS file system.
- You have connected the ADLS File System asset to the ADLS Edge capability.
- If needed, you have defined crawlers.

Required permissions

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.

Steps

1. Open the ADLS File System asset.
2. In the tab pane, click **Configuration**.
3. In the **Synchronization schedule** section, click **Add Schedule**.

4. Enter the required information.

Field	Description
Repeat	The interval when you want to synchronize automatically. The possible values are: **Daily**, **Weekly**, **Monthly**, and **Cron expression**.
Cron	The Quartz Cron expression that determines when the synchronization takes place. This field is only visible if you select `Cron expression` in the **Repeat** field.
Every	The day on which you want to synchronize, for example, Sunday. This field is only visible if you select `Weekly` in the **Repeat** field.
Every first	The day of the month on which you want to synchronize, for example, Tuesday. This field is only visible if you select `Monthly` in the **Repeat** field.
At	The time at which you want to synchronize automatically, for example, 14:00. ◦ You can only schedule on the hour. For example, you can add a synchronization schedule at 8:00, but not at 8:45. If you try to add it at 8:45, we will default it to 8:00. Use a cron expression if you don't want to schedule on the hour. ◦ This field is only visible if you select `Daily`, `Weekly`, or `Monthly` in the **Repeat** field.
Time zone	The time zone for the schedule.

5. Click **Save**.

What's next?

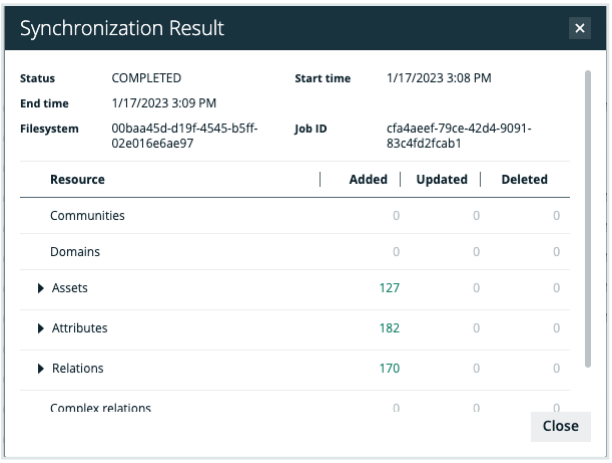
At every defined time, the synchronization starts.
You can view the assets in their domain.

View the summary of an Azure Data Lake Storage synchronization

After you synchronized an ADLS file system, you can view the summary of the results. This shows the impact of the synchronization on the assets in Collibra Platform Self-Hosted.

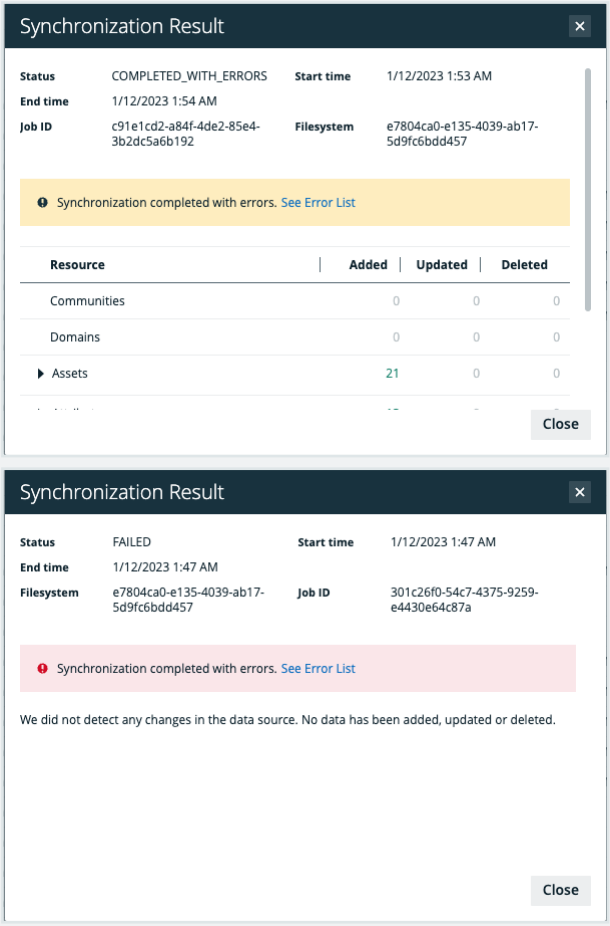
Steps

1. Open the Activities list.
2. In the row containing the ADLS synchronization job, click **Result**.
   » The **Synchronization Results** dialog box appears.



Note

- If anything changed, the information about the total number of resources that were added, modified or removed as a result of the synchronization are displayed.
- In case of errors, you can receive additional information about the error.



Tip The **Job ID** is useful when troubleshooting your synchronization process with Collibra Support.

For information on the result, go to Integrated Azure Data Lake Storage data.

Integrated Azure Data Lake Storage data

After you have synchronized the data, the integration of the Azure Data Lake Storage file system is completed, and the resulting assets are available in the domain that was

specified in the crawler.

> Warning Do not move the assets to another domain. Doing so may lead to errors during future synchronizations.

> Tip ADLS synchronization relies on UUIDs.

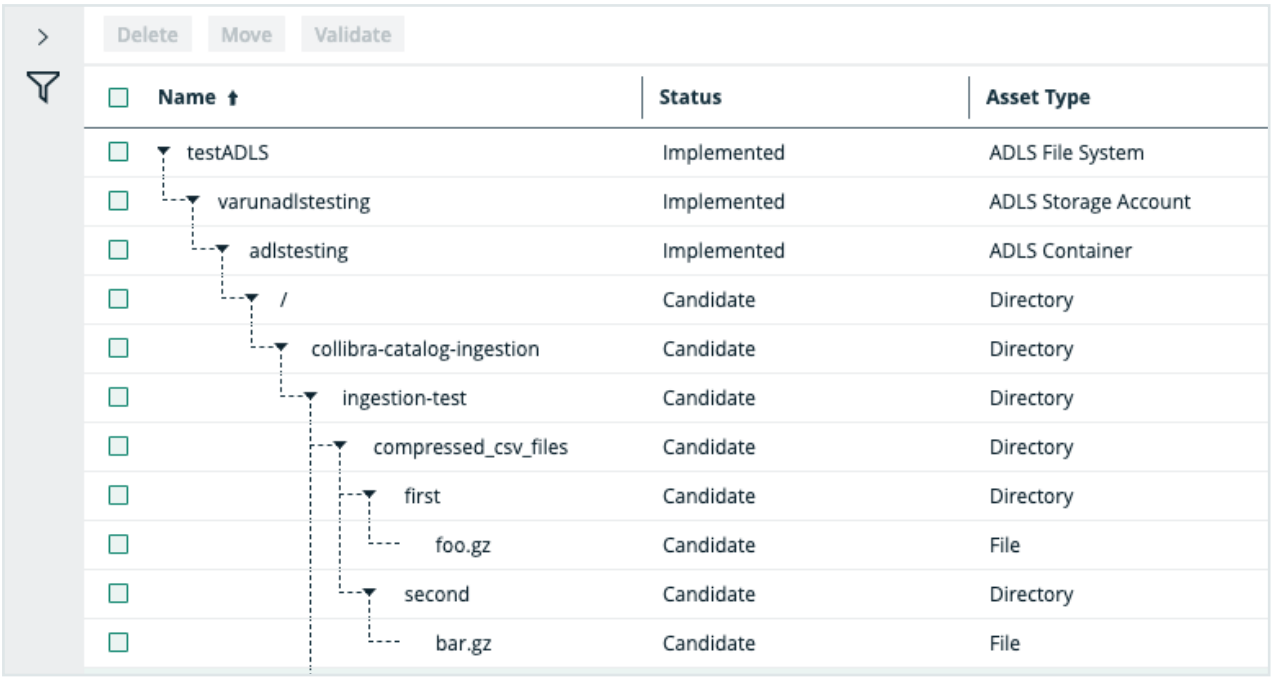
> Note In case of a partial synchronization caused by a temporary communication issue, the status of the assets that cannot be synchronized is set to **Missing from source**. Their previous status is restored, if they are found in the source system during the next fully successful synchronization.

By default, the assets are shown in a plain list, but you can enable a multi-path hierarchy to show it in a tree structure. The resulting assets depend on whether you use Microsoft Purview.

Synchronization results without Microsoft Purview

For the best result, use the following relations when you define a multi-path hierarchy:

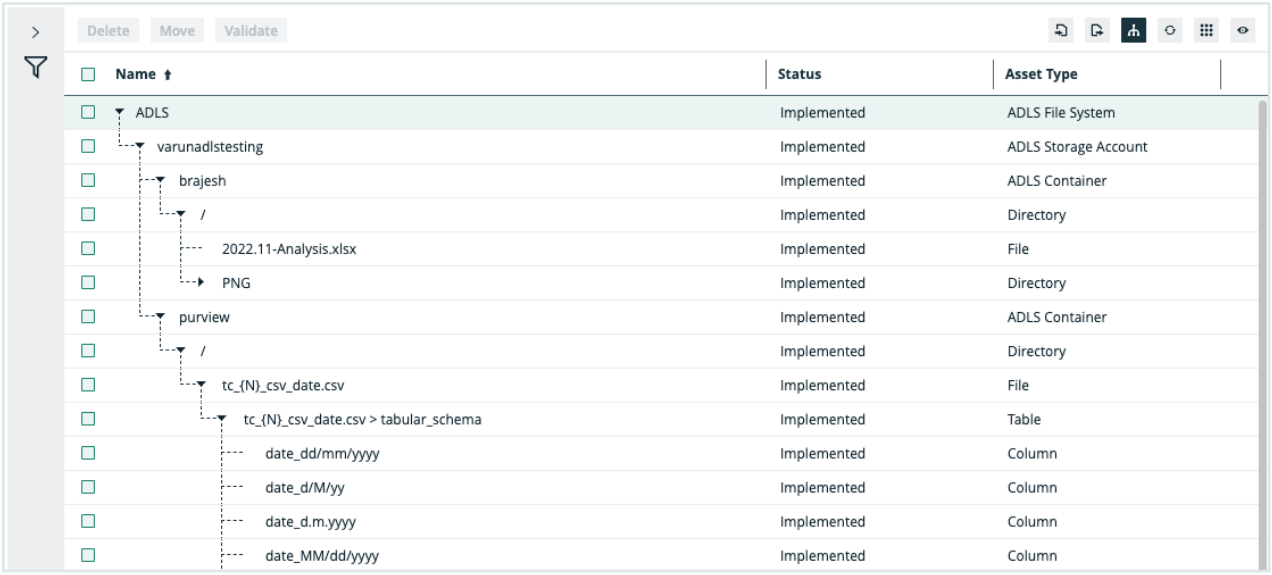
- File Storage contains File Container
- File Container contains File Container
- Directory contains Directory
- File Container contains File



Synchronization results with Microsoft Purview

For the best result, use the following relations when you define a multi-path hierarchy:

- File Storage contains File Container
- File Container contains File Container
- Directory contains Directory
- File Container contains File
- File contains Table
- Table contains Column



Synchronized metadata per asset type

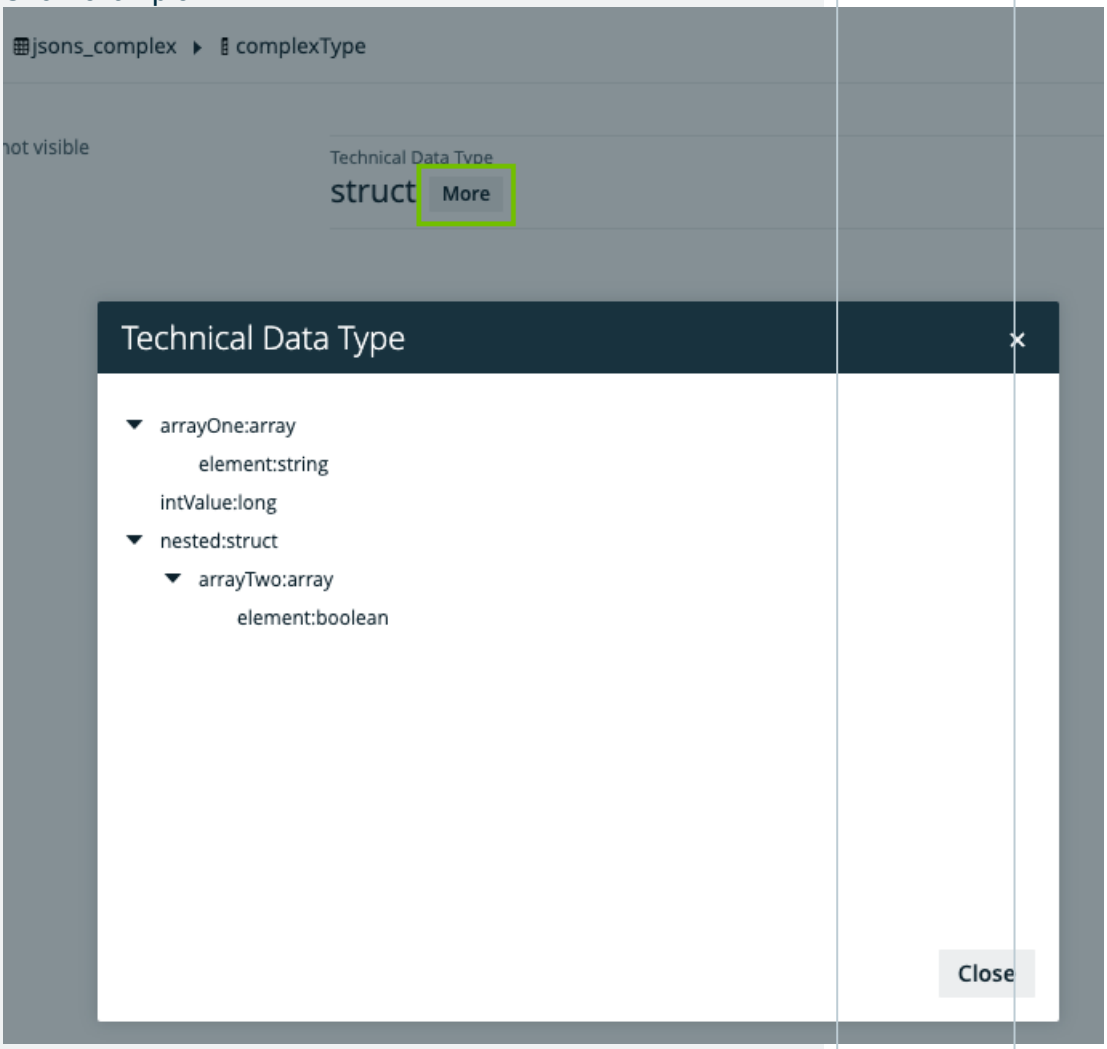
This table shows the metadata for each ADLS asset type.

Asset type	Synchronized metadata	Resource ID
ADLS Storage Account	File Storage contains / is part of File Container	00000000-0000-0000-0001-002600000-000
ADLS Container	Location	00000000-0000-0000-0000-000000000-203
	File Container contains / is part of File Container	00000000-0000-0000-0001-002600000-001

Asset type	Synchronized metadata	Resource ID
Directory	File Container contains / is part of File Container	00000000-0000-0000-0001-002600000-001
	Directory contains / is part of directory	00000000-0000-0000-0001-002600000-003
File	File Type	00000000-0000-0000-0001-002500000-012
	Size	00000000-0000-0000-0001-000500000-009
	File Container contains / contained in	00000000-0000-0000-0000-000000007-060

Asset type	Synchronized metadata	Resource ID
Table	Description	00000000-0000-0000-0000-000000003-114
	File contains / is part of Table	00000000-0000-0000-0001-002600000-002

Asset type	Synchronized metadata	Resource ID
Column	Description	00000000-0000-0000-0000-000000003-114
	Column Position	00000000-0000-0000-0001-000500000-020

Asset type	Synchronized metadata	Resource ID
	Technical Data Type **Tip** For columns that have a structured technical data type, Array or Struct, you can click the **More** button in the Column asset to see the structure of the data in a dialog box. This is supported for AVRO, CSV, JSON, ORC, PARQUET, PSV, SSV, TSV, TXT, and XML. In the capability settings, you define the maximum level you want to see in the structure. Show example 	00000000-0000-0000-0000-000000000219

Asset type	Synchronized metadata	Resource ID
	Column is part of / contains Table	00000000-0000-0000-0000-000000007-042

Troubleshooting Azure Data Lake Storage integration

Where do I find the Edge Site Id and Job Id?

If you report an error with Azure Data Lake Storage (ADLS) integration, the Customer Support team can ask you for the Edge Site Id and Job Id. The team needs this information to access details about the error.

To retrieve the Job Id, see View the summary of an Azure Data Lake Storage synchronization.

To retrieve the Site Id:

1. Go to Settings.
2. In the Edge section, click Sites.
3. Click the name of the Edge site.
4. The Edge site Id is available in the ID field.

You receive the error when synchronizing ADLS: You are not allowed to perform this action.

Issue: You receive the following error: `Error while processing crawler catalogingestion: Import job failed with message. You are not`

`allowed to perform this action..`

Reason: The ADLS synchronization capability does not store any user credentials. It calls the Import API using the Edge site user credentials. By default, the Edge site user cannot add any new assets in Collibra.

Solution: Make sure to give following permission to the Edge site user: 'Manage all resources'. To do so, go to **Settings** → **Global Permissions** and select the **Resources** → **Manage all resources** permission for the Edge site role.

You receive an error when synchronizing ADLS indicating there is an issue with the Service Principal permissions

Issue: You receive an error when synchronizing ADLS indicating there is an issue with the Service Principal permissions.

Possible reason: Something is wrong in the configuration of the Service Principal account or the definition of the ADLS Include path.

Troubleshooting:

1. Install the Azure Command Line Interface (CLI) using the Azure documentation.
2. Open Azure CLI and login as the service principal.
   Use the following format: `az login --service-principal -u <app-id> -p <password-or-cert> --tenant <tenant>`
   For more information, go to the Azure documentation.
   Output example:

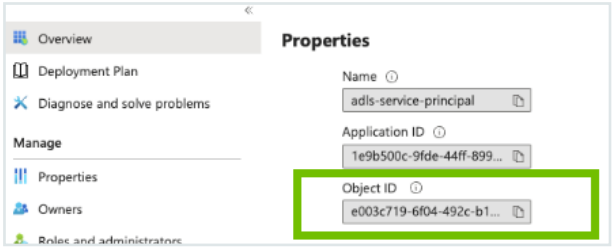
```
[
  {
    {"cloudName": "AzureCloud",
    "homeTenantId": "********************************",
    "id": "********************************",
    "isDefault": true,
    "managedByTenants": [
      {
      "tenantId": "********************************"
```

```
        }
      ]
      "name": "power-bi_dev_tv_testing",
      "state": "Enabled",
      "tenantId": "3******************************",
      "user": {
        "name": "19592009-c7ac-4573-8ac5-701d4529ef23",
        "type": "servicePrincipal"
      }
    }
  ]
```

3. Ask to list the assigned roles.
   Use the following code: `az role assignment list --all`
   Output: You receive a list of permissions and roles. We expect that you see the roleDefinitionNames "Reader" and "Storage Blob Data Reader".
   For more information, go to the Azure documentation.
4. Ask to list the containers
   Use the following format: `az storage container list --auth-mode login --account-name <account name> | grep name`
   Output: You receive a list of containers. We expect that you see the container in which your data is located, and which you reference in the Include Path. https://<storage account name>.blob.core.windows.net/<container name>/<blob name>.
   For more information, go to the Azure documentation.
5. Check if the directory/blob you reference in the Include Path exists.
   Use following format: `az storage fs directory exists -n <directory> -f <file system> --account-name <account name> --auth-mode login`
   Output: We expect that you get True as output.
   For more information, go to the Azure documentation.
6. To ingest purview collections into Collibra, the service principal needs read permissions on the Purview collection. You can give the add “Collection admins“ permission using CLI.
   a. Login in to CLI.
      Use the following code: `az login`
   b. Install the Azure Purview CLI module.
      Use the following code: `az extension add --name purview`
   c. Run Purview
      Use the following format: `az purview account add-root-collection-`

```
admin --name <account name> --object-id <Service principal
Object Id> --resource-group <SampleResourceGroup> >
```

You can find the Service principal Object Id in the IAM account overview.



7. If your ADLS storage is private, make sure that the **Allow Azure services on the trusted services list to access this storage account** checkbox in the **Networking → Firewalls and virtual networks** is selected.
   Your ADLS storage is private if you selected the **Public networking access** option **Enabled from selected virtual networks and IP adresses**.

You receive the error when synchronizing ADLS: Crawler path does not exist

Issue: During the synchronization, you receive the following error: `Crawler path does not exist on ADLS ....`

Possible reason: The property **Hierarchical namespace** is probably not enabled in your ADLS account.

Solution: In Microsoft Azure, open your storage account, and in **Overview → Properties**, enable the Blob service property **Hierarchical namespace**.

You receive an error: Offset should not be greater than 100000

Issue: During the synchronization, you receive the following error: `Illegal argument: Offset should not be greater than 100000..`

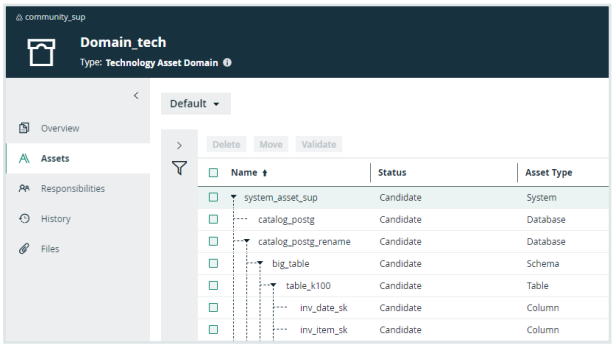
Reason: Currently, the ADLS integration can ingest only up to 100,000 assets from Purview.

Working with Databricks

In Collibra Platform Self-Hosted, you can work with Databricks in multiple ways. You can register individual Databricks databases via the Databricks JDBC driver, and you can integrate all metadata of the databases from Databricks Unity Catalog.

Ways to work with Databricks

In Collibra Platform Self-Hosted, you can work with Databricks in two ways.
You can register individual Databricks databases via the Databricks JDBC driver, and you can integrate all metadata of the databases from Databricks Unity Catalog.
It is important to understand the difference between these ways of working because the result in Collibra is different.

Possible way to work with Databricks	Result in Collibra
Integrating metadata from Databricks Unity Catalog	If you integrate Databricks Unity Catalog, you integrate the metadata of all databases in the Databricks Unity Catalog metastore into Collibra Platform Self-Hosted. The resulting assets represent the Databricks databases, schemas, tables and columns. Note • Because we only integrate the metadata, you cannot get sample data for the columns and tables, nor profile and classify them. If you want to do that, you need to register the Databricks database via the Databricks JDBC driver. For information, go to combining the integration and the JDBC driver. • The Databricks Unity Catalog integration supports the integration of following tables: EXTERNAL, MANAGED, STREAMING_TABLE, and VIEW tables. Important You can integrate Databricks Unity Catalog only via Edge. You cannot integrate Databricks Unity Catalog via Jobserver. Example 

Possible way to work with Databricks	Result in Collibra
Registering a Databricks data source via the Databricks JDBC connector	If you register a specific Databricks data source via the Databricks JDBC connector, the resulting assets represent the columns and the tables in the Databricks database. You can retrieve sample data, and can profile and classify the data.

Combining the two ways of working with Databricks

The two possibilities don't cancel each other out. You can use both ways to show the information you want in Collibra Platform Self-Hosted. You can use the integration of Databricks Unity Catalog to quickly get an overview of all your Databricks databases in Collibra Platform Self-Hosted. Once you have a better view on the important databases, you can register them individually via the JDBC driver.

Combining the two ways of working with Databricks
1. You first register a Databricks data source via the Databricks JDBC connector. 2. If you then integrate Databricks Unity Catalog, the integration: ◦ Skips the assets that have been registered via JDBC. ◦ Adds the new information from Databricks Unity Catalog.

Combining the two ways of working with Databricks

1. You first integrate Databricks Unity Catalog.
2. If you then register a Databricks data source via the JDBC connector, the registration adds all data source assets.
   This results in duplicate assets.
3. If you then integrate Databricks Unity Catalog again, we advise to exclude the databases that you registered via JDBC. You can do this by using the **Filters and Domain Mapping** property in the Databricks Unity Catalog capability. The integration:
   - Adds the new information from Databricks Unity Catalog.
   - Skips the assets that have been registered via JDBC.
   - If any assets were removed or excluded from the integration, they are marked as **Missing from source**. You can manually remove them.

Example
Your Databricks Unity Catalog consists of the three databases: A, B, C.

1. You integrate the metadata from Data Unity Catalog.
   This results in the Database assets: A, B, C.
2. You want to register the metadata of database C to access the profiling and classification results.
   The JDBC registration results in a Database asset **C'**, with the same metadata as C.
3. You integrate the metadata again (because there have been updates).
   - If you don't exclude database C, all databases will be updated, except for **C'**.
   - If you exclude database C, database C will receive the "Missing from source" status, and you can manually remove these assets.

From that moment, you should:

- For A and B, use the Databricks Unity Catalog integration with exclude rules for C, to update the metadata.
- For **C'**, use the synchronization via JDBC to update the metadata.

Important Use the same System asset for the integration and the registration. Otherwise, assets will be duplicated.

For more information about Databricks, go to the Databricks documentation.

Integrating Databricks Unity Catalog

Databricks Unity Catalog is a technical catalog on Databricks side that provides schema information for all the Databricks databases that are available in the connected Databricks instances.

About the Databricks Unity Catalog integration via Edge

Databricks Unity Catalog is a technical catalog on Databricks side that provides schema information for all the Databricks databases that are available in the connected Databricks instances.

If you integrate Databricks Unity Catalog, you integrate the metadata of all databases in the Databricks Unity Catalog metastore into Collibra Platform Self-Hosted. The resulting assets represent the Databricks databases, schemas, tables and columns.

Note
- Because we only integrate the metadata, you cannot get sample data for the columns and tables, nor profile and classify them. If you want to do that, you need to register the Databricks database via the Databricks JDBC driver. For information, go to combining the integration and the JDBC driver.
- The Databricks Unity Catalog integration supports the integration of following tables: EXTERNAL, MANAGED, STREAMING_TABLE, and VIEW tables.

Tip
You can follow a course on this Databricks Unity Catalog integration via Collibra University.

Important
- You can integrate Databricks Unity Catalog only via Edge. You cannot integrate Databricks Unity Catalog via Jobserver.

- You cannot retrieve sample data or profile and classify the data for the Tables and Column assets created via the Databricks Unity Catalog integration. If you want to do that, you need to register the Databricks database via the Databricks JDBC driver.

Why use the Databricks Unity Catalog integration?

The Databricks Unity Catalog integration allows to get all the metadata from Databricks Unity Catalog into Collibra in one action, which means you quickly get an overview of all your Databricks databases in Collibra Platform Self-Hosted. You can also register Databricks databases into Collibra Platform Self-Hosted via the Databricks JDBC connection. However, you must register each database individually.

Important
If you used the JDBC Databricks driver to register a specific Databricks database before, the related Database assets are not integrated again when you run the Databricks Unity Catalog integration.

For more details on the different ways of working with Databricks in Collibra and how to combine the integration and individual registration of databases, go to Ways to work with Databricks.

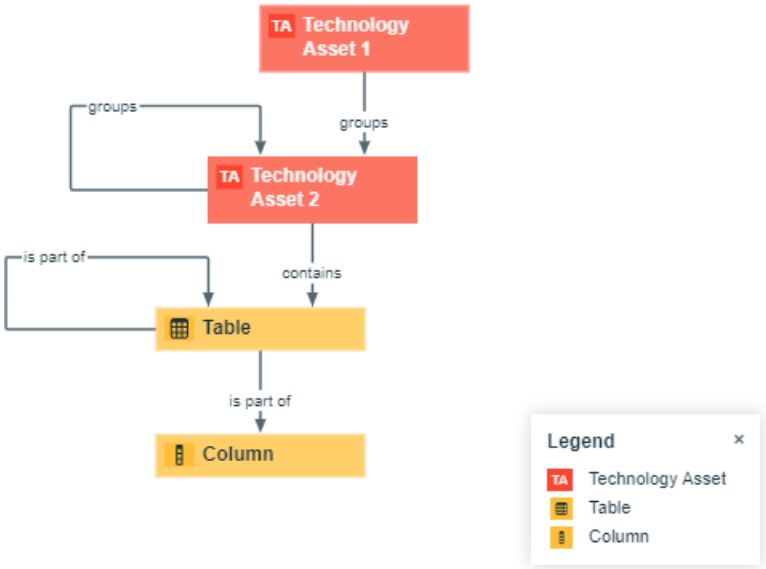
For more information about Databricks Unity Catalog, go to the Databricks Unity Catalog documentation.

Databricks Unity Catalog assets types and operating model

The Databricks Unity Catalog integration of Collibra Platform Self-Hosted uses a specific subset of asset types. All of these come out-of -the-box.

Asset type	Description
Technology Asset ▸ System	Executable software that you can buy commercially off the shelf (COTS), or build internally, to automate one or more business functions that help run a business smoothly and efficiently. Examples: CRM, ERP, EDW
Technology Asset ▸ Database	A collection of data that is systematically organized or structured, to make it easy to create, update, and query the information. Examples: Ora_DGC_V45, SalesDB2020
Data Asset ▸ Data Structure ▸ Schema	An asset that contains the location of specific data. It provides all the details that are required for setting up a connection to a database or server.
Data Asset ▸ Data Structure ▸ Table	An implementation of data entities in columns and rows, in a given database system. It is the basic structure of a relational database. Examples: Account_tbl, CUST_ADDR
Data Asset ▸ Data Element ▸ Column	An atomic unit of data that can be stored in a database table. Examples: FST_NM, EMPID

Databricks Unity Catalog operating model



Steps overview: Integrate Databricks Unity Catalog via Edge

#	Step	Description
1	Enable the Databricks file system registration and synchronization via Edge.	Allows the integration of Databricks via Edge.
2	Give the Edge Site user the required permissions.	Ensures the Edge Site user can integrate the metadata.
3	Create a Databricks connection to your Edge site.	Creates a connection to Databricks in an Edge site.
4	Add the Databricks Unity Catalog capability to the Edge site.	Adds the Databricks Unity Catalog capability to the Edge connection. The capability allows you to retrieve data from Databricks Unity Catalog.

#	Step	Description
5	Synchronize Databricks Unity Catalog.	You can manually synchronize Databricks Unity Catalog or add a synchronization schedule. Once the synchronization is completed, the metadata is integrated.

Enable the Databricks integration via Edge

You can enable the registration and synchronization of Databricks Unity Catalog via Edge.

Requirements and permissions

- You have the **ADMIN** or **SUPER** role in Collibra Console.
- You have the **SUPER** role in Collibra Console.
- You have the **ADMIN** or **SUPER** role in Collibra Console.

Steps

1. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.
   c. In the tab pane, click the Data Governance Center service of that environment.
   d. Click **Configuration**.
   e. Click **Edit configuration**.
2. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.
   c. In the tab pane, click the Data Governance Center service of that environment.
   d. Click **Configuration**.
   e. Click **Edit configuration**.

3. In the **Register data source** section, activate this setting:

Setting	Description
Databricks Unity Catalog synchronization via Edge	An option to enable the integration of Databricks Unity Catalog via Edge. ◦ ✔ True: You can register and synchronize Databricks Unity Catalog via Edge. ◦ ✘ False: You cannot integrate Databricks Unity Catalog. Set this option to True.

4. Click **Save all**.

What's next?

Give the Edge Site user the required permissions and create a Databricks connection to your Edge site.

Add the Databricks Unity Catalog capability

After you have created a connection to Databricks in your Edge site, you have to add the Databricks Unity Catalog capability to the connection.

Before you start

- You have created and installed an Edge site.
- The integration of Databricks via Edge has been enabled.
- You have created a connection to Databricks in your Edge site.

Required permissions

- You have a global role that has the **Manage connections and capabilities** global permission, for example, Edge integration engineer.

Steps

1. Open an Edge site.
   a. On the main menu, click ⠿, and then click ⚙ **Settings**.
      » The Collibra settings page opens.
   b. In the tab pane, click **Edge**.
      » The **Sites** tab opens and shows a table with an overview of the Edge sites.
   c. In the table, click the name of the Edge site whose status is **Healthy**.
      » The Edge site page opens.
2. In the **Capabilities** section, click **Add capability**.
   » The **Add capability** page appears.
3. Enter the required information.

Field	Description	Required
Capability	This section contains general information about the capability.	
Name	The name of the Edge capability.	✓ Yes
Description	The description of the Edge capability.	✗ No
Capability template	The capability template. The value that you select in this field determines which sections appear on the page. Select the following Edge capability: `Databricks Unity Catalog synchronization`	✓ Yes
Databricks Connection		
Databricks Connection	The Databricks connection to be used.	✓ Yes
Configuration	This section contains information on how to connect to Databricks Unity Catalog.	
Save input metadata	If you select this option the metadata extracted from the data source will be saved in a file that can be used for troubleshooting. Select this option only on request of Collibra Support.	✗ No

Field	Description	Required
Exclude Schemas (will be removed soon)	Comma-separated list of the schemas that you don't want to integrate. Note The listed schemas will be excluded for all databases. We recommend using this field to list the schemas that are automatically generated in a database, such as information_schema and default, and which you don't want to integrate.	✕ No

Field	Description	Required
Filters and Domain Mapping (Beta)	Text in JSON format to include or exclude databases and schemas, and to configure domain mappings. This feature is intended for non-production use at this moment, as it is in Beta. ◦ The text must be in JSON format and can contain an include and an exclude block. You can use any JSON validator to verify the format. Collibra is not responsible for the privacy, confidentiality, or protection of the data you submit to such JSON validators, and has no liability for such use. ◦ In the include block, you can specify the domain in which specific catalogs or schemas must be ingested. The format is: `"Catalog/Database > schema ": "domain ID"`. For example, `"HR > address-schema": "30000000-0000-0000-0000-000000000000"`. ◦ In the exclude block, you can specify the catalogs or schemas that you don't want to ingest. For example, `"* > test"`. ◦ The exclude block has priority over the include block. ◦ If the include block is not present, we ingest all assets into the same domain as the System asset. ◦ If there is no explicit domain mapping for a schema, we use the domain specified for the database. ◦ You can use the keyword `default` as a domain ID. In that case, the catalog or schema will be ingested in the same domain as the System asset. ◦ A match with a database has priority over a match with a schema. ◦ The integration fails before the synchronization starts, if one or more domain IDs specified in the include block don't exist. ◦ The integration fails before the synchronization starts if a domain ID is left empty in the include block. ◦ You can use the ? and * wildcards in the catalog and schema names. If a catalog or schema matches multiple lines, the most detailed match is taken into account.	✕ No

Field	Description	Required
	Example `{ "include": { "HR": "20000000-0000-0000-0000-000000000000", "HR > address-schema": "30000000-0000-0000-0000-000000000000", "Orders > fk*": "40000000-0000-0000-0000-000000000000", "Orders > *": "50000000-0000-0000-0000-000000000000", "* > profiling": "60000000-0000-0000-0000-000000000000", "sales": "default" }, "exclude": ["testDB", " * > information_schema"] }` In this example: ◦ Assets from the "HR" database will be ingested into the domain with ID "20000000-0000-0000-0000-000000000000". However, all assets from the "HR > address-schema" schema will be ingested into the domain with id "30000000-0000-0000-0000-000000000000". ◦ All assets from the "Orders" database with schemas starting with fk (fk*) will be ingested into the domain with ID "40000000-0000-0000-0000-000000000000", and all other assets from the "Orders" database will be ingested into the domain with ID "50000000-0000-0000-0000-000000000000". ◦ All assets from the "sales" database will be ingested in the same domain as the System asset. ◦ Assets from the "profiling" schema will be ingested into the domain with ID "60000000-0000-0000-0000-000000000000". However, the "profiling" schema in the database "Orders" will be ingested in the domain with ID "50000000-0000-0000-0000-000000000000" because a database match has priority over a schema match. ◦ All assets from the "testDB" database will be excluded. ◦ All assets from the "information_schema" schema in all databases will be excluded.	

Field	Description	Required
Extensible Properties Mapping (Beta)	Via the **Extensible Properties Mapping** field, Databricks Unity Catalog allows you to add additional properties to Catalog, Schema, and Table objects. Important ◦ This feature is intended for non-production use at this moment, as it is in Beta. ◦ If you use this feature, make sure to set up all required characteristic assignments for the asset types. Two possible JSON formats are available. ◦ Version 0.1: This version allows you to ingest custom properties only. You can ingest the values from the Properties field from Catalog, Schema, and Table objects into specific attributes in Collibra assets. You do this by adding the mapping between the Properties fields for the objects in Databricks Unity Catalog and the Collibra attribute IDs to ingest the data in, using a JSON string. ▪ The text must be in JSON format and can contain a Catalogs, Schemas, and Tables block. The Catalogs block refers to Database assets, the Schemas block to Schema assets, and the Tables block to Table assets. ▪ In each block, you specify the property name and the attribute ID to which you want to map the value in the property. The format is: `"[property name]": "[attribute resource ID]"`. For example, `"Description from source system": "00000000-0000-0000-0001-000500000074"`.	✕ No

Field	Description	Required
	Example `{ "catalogs": { "color": "00000000-0000-0000-0000-000000001234", "Description from source system": "00000000-0000-0000-0001-000500000074" }, "schemas": { "File Location": "00000000-0000-0000-0001-000500000004" }, "tables": { "delta.lastCommitTimestamp": "00000000-0000-0000-0000-000000003114" } }` In this example: ■ In the Database assets that we create, we'll add the Color value in attribute 00000000-0000-0000-0000-000000001234, and the Description from Source value in attribute 00000000-0000-0000-0001-000500000074. ■ In the Schema assets that we create, we'll add the File Location value in attribute 00000000-0000-0000-0001-000500000004. ■ In the Table assets that we create, we'll add the delta.lastCommitTimestamp value in attribute 00000000-0000-0000-0000-000000003114. ◦ Version 0.2: This version allows you to ingest both default system properties and custom properties. You can ingest the most values from the Details page from Catalog, Schema, and Table objects into specific attributes in Collibra assets. You do this by adding the mapping between the fields for the objects in Databricks Unity Catalog and the Collibra attribute IDs to ingest the data in, using a JSON string.	

Field	Description	Required
	■ The text must be in JSON format. ■ A Version block referencing 0.2 must be added. ■ A Catalogs, Schemas, and Tables block can be added. The Catalogs block refers to Database assets, the Schemas block to Schema assets, and the Tables block to Table assets. ■ Inside a Catalogs, Schemas, or Tables block, you can add a systemAttributes and a customParameters block. systemAttributes refers to the default system properties. customParameters refers to the custom properties. ■ In each block, you specify the property name and the attribute ID to which you want to map the value in the property. The format is: `"[property name]": "[attribute resource ID]"`. For example, `"Description from source system": "00000000-0000-0000-0001-000500000074"`. Following system properties are supported: ■ Catalogs: "browse_only", "catalog_type" , "connection_name" , "created_at" , "created_by" , "isolation_mode" , "metastore_id" , "provider_name" , "provisioning_info" , "securable_kind" , "securable_type" , "share_name" , "storage_location" , "storage_root" , "updated_at" , and "updated_by". ■ Schemas: "catalog_type", "created_at" , "created_by" , "metastore_id" , "storage_location" , "storage_root" , "updated_at" , and "updated_by". ■ Tables: "access_point", "created_at" , "created_by" , "data_access_configuration_id" , "data_source_format" , "deleted_at" , "metastore_id" , "sql_path" , "storage_credential_name" , "storage_location" , "updated_at" , "updated_by" , and "view_definition".	

Field	Description	Required
	Example `{ "version": 0.2, "catalogs": { "systemAttributes": { "metastore_id": "00000000-0000-0000-0000-000000004224" }, "customParameters": { "color": "00000000-0000-0000-0000-000000001234", "Description from source system": "00000000-0000-0000-0001-000500000074" } }, "schemas": { "customParameters": { "File Location": "00000000-0000-0000-0001-000500000004" } }, "tables": { "systemAttributes": { "metastore_id": "00000000-0000-0000-0000-000000004224" }, "customParameters": { "delta.lastCommitTimestamp": "00000000-0000-0000-0000-000000003114" } } }` In this example: ■ In the Database assets that we create, we'll add the metastore_id value in attribute "00000000-0000-0000-0000-000000004224", the Color value in attribute 00000000-0000-0000-0000-000000001234, and the Description from Source value in attribute 00000000-0000-0000-0001-000500000074. ■ In the Schema assets that we create, we'll add the File Location value in attribute 00000000-0000-0000-0001-000500000004. ■ In the Table assets that we create, we'll add the metastore_id value in attribute "00000000-0000-0000-0000-000000004224" and the delta.lastCommitTimestamp value in attribute	

Field	Description	Required
Advanced Configuration	This section contains configuration options that can help when investigating issues with the capability. Important Only complete the fields **Logging configuration**, **Memory (MiB)**, and **JVM arguments** on request of or together with Collibra Support.	✕ No

4. Click **Create**.
   » The capability is added to the Edge site.
   » The fields become read-only.

What's next?

You can now synchronize Databricks Unity Catalog.

Synchronizing Databricks Unity Catalog

After Edge is ready to integrate Databricks Unity Catalog, you can start the synchronization.

Synchronize Databricks Unity Catalog

Synchronizing Databricks Unity Catalog is the process of integrating metadata from the databases connected to Databricks Unity Catalog and making the data available in Collibra Platform Self-Hosted.

You can synchronize manually, or you can automate it by adding a synchronization schedule.

Before you begin

- You have created a connection to Databricks in your Edge site.
- You have added the Databricks Unity Catalog capability for the connection.

- You know in which System asset you want to add the Databricks Unity Catalog assets.
  - If you have registered Databricks databases via the JDBC driver before, use the same System asset.
  - If you never registered Databricks databases before, create a new System asset manually and use that one.

Required permissions

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a global role with the View Edge connections and capabilities global permission, for example, Edge integration engineer.

Manually synchronize Databricks Unity Catalog

1. On the main menu, click ⠿, and then click **Catalog**.
   » The Catalog Home opens.
2. On the main toolbar, click +.
   » The **Create** dialog box appears.
3. In the **Register with Edge** section of the **Create** dialog box, click **Register a data source**.
   » The **Register content** page opens.
4. In the **Connection name** column, locate the Databricks connection that you used when you added the Databricks Unity Catalog capability and click the link in the **Data sources/Capabilities** column.
   » The Databricks Unity Catalog capability configuration page opens.
5. In the **Configuration Section** section, click **Add Configuration**.
6. Select the System asset in which you want to add the Databricks assets.
7. Click **Save Configuration**.
8. Click **Synchronize**.
   » A notification indicates the synchronization has started.

Add a synchronization schedule

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.
2. On the main toolbar, click +.
   » The **Create** dialog box appears.
3. In the **Register with Edge** section of the **Create** dialog box, click **Register a data source**.
   » The **Register content** page opens.
4. In the **Connection name** column, locate the Databricks connection that you used when you added the Databricks Unity Catalog capability and click the link in the **Data sources/Capabilities** column.
   » The Databricks Unity Catalog capability configuration page opens.
5. In the **Configuration Section** section, click **Add Configuration**.
6. Select the System asset in which you want to add the Databricks assets.
7. Click **Save Configuration**.
8. In the **Synchronization schedule** section, click **Add schedule**.
9. Enter the required information and click **Save**:

Field	Description
Repeat	The interval when you want to synchronize automatically. The possible values are: **Daily**, **Weekly**, **Monthly**, and **Cron expression**.
Cron	The Quartz Cron expression that determines when the synchronization takes place. This field is only visible if you select `Cron expression` in the **Repeat** field.
Every	The day on which you want to synchronize, for example, Sunday. This field is only visible if you select `Weekly` in the **Repeat** field.
Every first	The day of the month on which you want to synchronize, for example, Tuesday. This field is only visible if you select `Monthly` in the **Repeat** field.

Field	Description
At	The time at which you want to synchronize automatically, for example, 14:00. ◦ You can only schedule on the hour. For example, you can add a synchronization schedule at 8:00, but not at 8:45. If you try to add it at 8:45, we will default it to 8:00. Use a cron expression if you don't want to schedule on the hour. ◦ This field is only visible if you select `Daily`, `Weekly`, or `Monthly` in the **Repeat** field.
Time zone	The time zone for the schedule.

What's next?

The synchronization job integrates the metadata of all databases, schemas, tables and columns.
After the synchronization:

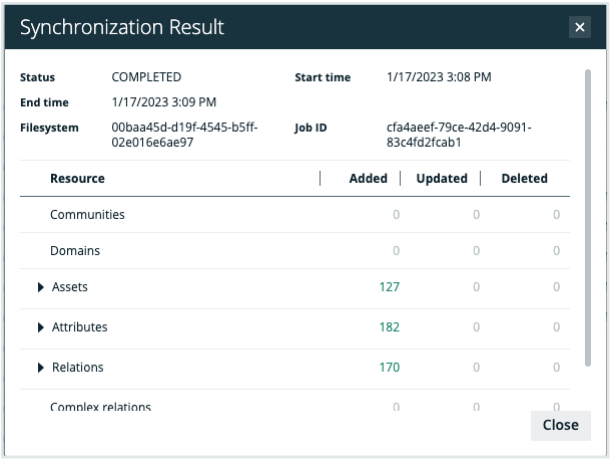
- You can view a summary of the results from the Activities list.
- The resulting assets get a relation to the System asset that you selected.
  For information on the integrated data, go to Integrated Databricks data.

View the summary of a Databricks Unity Catalog synchronization

After you synchronized Databricks Unity Catalog, you can view the summary of the results. This shows the impact of the synchronization on the assets in Collibra Platform Self-Hosted

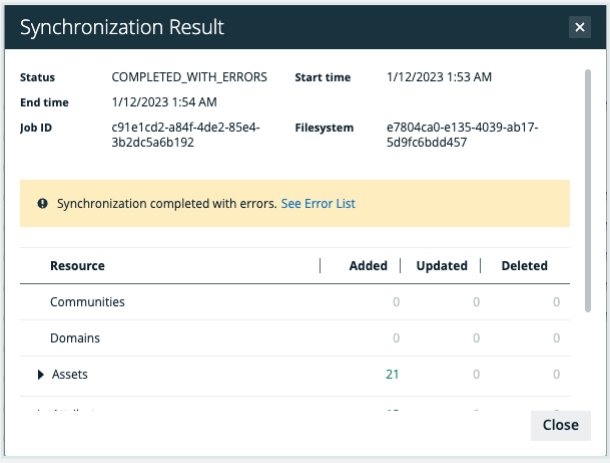
Steps

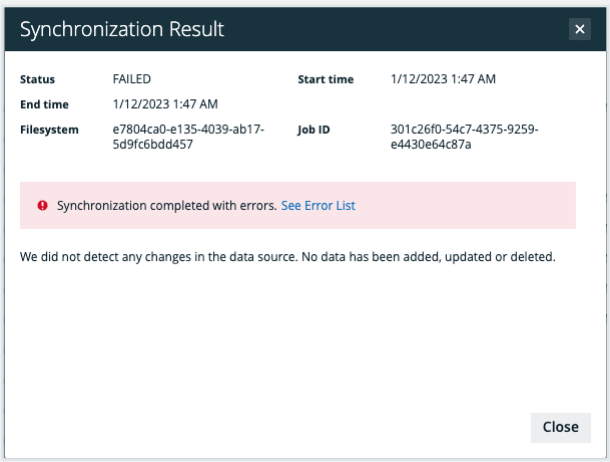
1. Open the Activities list.
2. In the row containing the Databricks Unity Catalog synchronization job, click **Result**.
   The **Synchronization results** dialog box appears.



Note

- If anything changed, the information about the total number of resources that were added, modified, or removed as a result of the synchronization are displayed.
- In case of errors, you can receive additional information about the error.





Tip The **Job ID** is useful when troubleshooting your synchronization process with Collibra Support.

For information on the resulting assets, see Integrated Databricks data.

Integrated Databricks Unity Catalog data

After the synchronization, the resulting Database, Schema, Table, and Column assets are available in the domain where the provided System asset is located.

Important

- If you move the resulting Table and Column assets to another domain and you run the integration again, the Table and Column assets will be moved back to their initial domain. However, if you move the resulting Database or Schema asset to another domain, the Database asset will remain in the new domain. To move all resulting assets to another location permanently, select another System asset in the current synchronization configuration or create a new capability with a synchronization configuration that integrates the data in the new location.

  Example
  You created System asset A in Domain A and synchronized Databricks.
  As a result, Table A and Column A have been added to Domain A.
  Then, you manually moved Table A and column A to Domain B.
  When you synchronize Databricks again, Table A and Column A will move back to Domain A.

- The Databricks Unity Catalog integration uses different naming conventions compared to the Edge JDBC naming conventions. The applied naming conventions are:

Asset type	Naming convention	Example
Database	domainName>catalogName	ay-tech-domain-4>oleg-test
Schema	databaseFullName>schemaName	ay-tech-domain-4>oleg-test>demo

Asset type	Naming convention	Example
Table	schemaFullName>tableName	ay-tech-domain-4>oleg-test>demo>dinner
Column	tableFullName>columnName (column)	ay-tech-domain-4>oleg-test>demo>dinner>recipe (column)

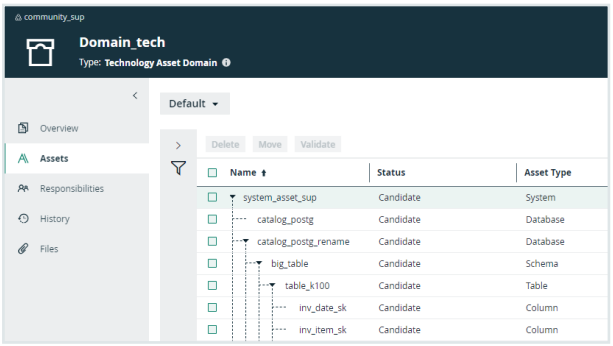
Tip Databricks synchronization relies on UUIDs.

Note In case of a partial synchronization caused by a temporary communication issue, the status of the assets that cannot be synchronized is set to **Missing from source**. Their previous status is restored, if they are found in the source system during the next fully successful synchronization.

By default, the assets are shown in a plain list, but you can enable a multi-path hierarchy to show it in a tree structure. For the best result, use the following relations in the multi-path hierarchy:

1. Technology Asset groups Technology Asset
2. Database contains Table
3. Technology Asset has Schema
4. Schema contains Table
5. Table contains Column

The following image shows the resulting hierarchical table.



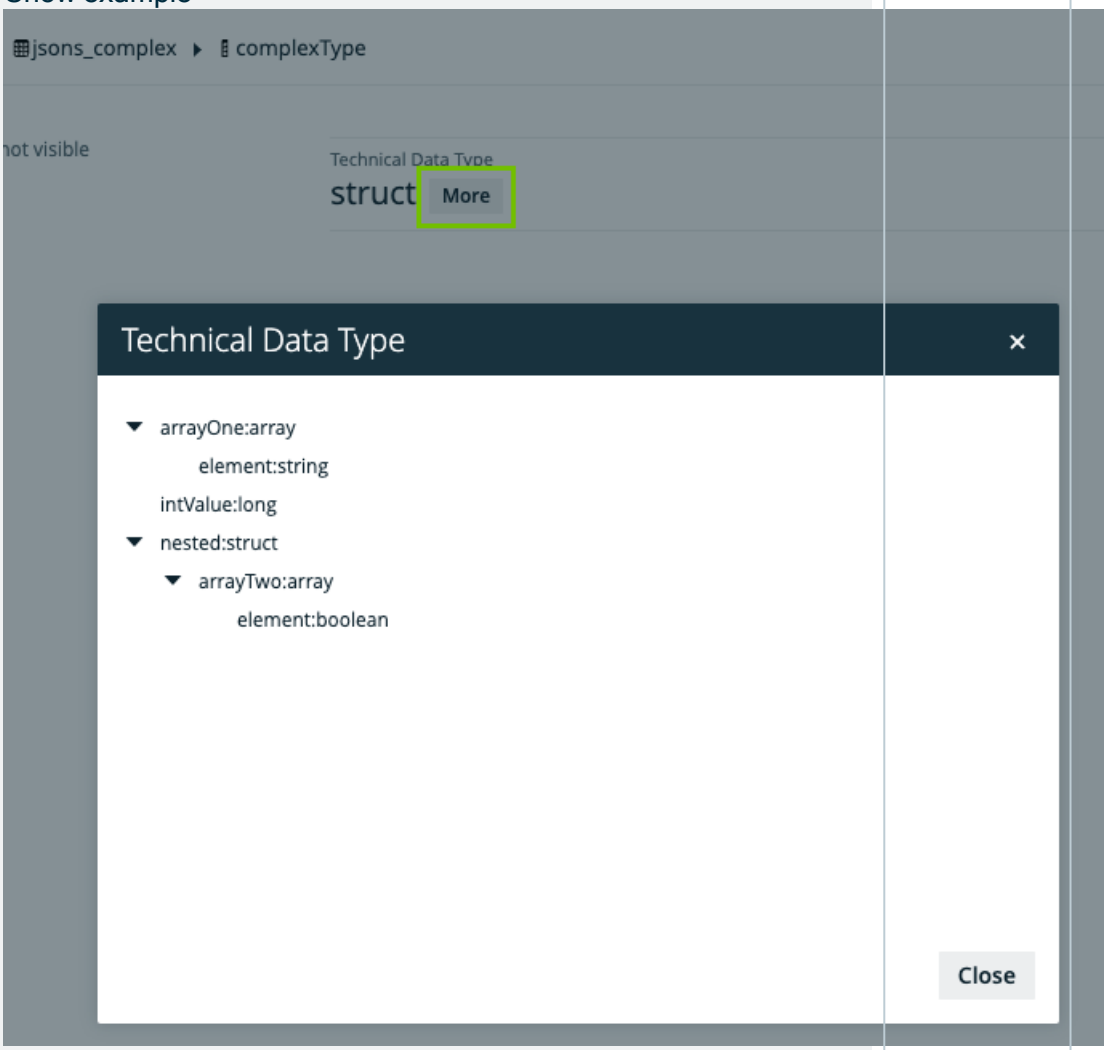
Synchronized metadata per asset type

This table shows the metadata for each Databricks asset type.

Asset type	Synchronized metadata	Resource ID
Data-base	Description from source system	00000000-0000-0000-0001-000500000-074
	Any extensible properties defined via the capability.	
	Technology Asset groups / is grouped by Technology Asset	00000000-0000-0000-0000-000000007-054
Schema	Description from source system	00000000-0000-0000-0001-000500000-074
	Any extensible properties defined via the capability.	
	Technology Asset has / belongs to Schema	00000000-0000-0000-0000-000000007-024

Asset type	Synchronized metadata	Resource ID
Table	Description from source system	00000000-0000-0000-0001-000500000074
	Any extensible properties defined via the capability.	
	Schema contains / is part of Table	00000000-0000-0000-0000-000000007043

Asset type	Synchronized metadata	Resource ID
Column	Description from source system	00000000-0000-0000-0001-000500000-074
	Column Position	00000000-0000-0000-0001-000500000-020
	Is Nullable	00000000-0000-0000-0001-000500000-011
	Is Primary Key	00000000-0000-0000-0001-000500000-015
	Primary Key Name (if the column is the primary key)	00000000-0000-0000-0001-000500000-016
	Original Name	00000000-0000-0000-0001-000500000-032

Asset type	Synchronized metadata	Resource ID
	Technical Data Type **Tip** For columns that have a structured technical data type, Array or Struct, you can click the **More** button in the Column asset to see the structure of the data in a dialog box. Show example  	00000000-0000-0000-0000-000000000219

Asset type	Synchronized metadata	Resource ID
Foreign Key	Column is part of / contains Table	00000000-0000-0000-0000-000000007-042
	Foreign Key Mapping (if the column is part of a foreign key)	00000000-0000-0000-0000-000000007-504
	Tip The full name of the Foreign Key asset has the following pattern : table_full_name > foreign_key_name (foreign_key)	
	Foreign Key Mapping	00000000-0000-0000-0000-000000007-504

Troubleshooting Databricks Unity Catalog integration

Where do I find the Edge Site ID and Job ID?

If you report an error with the Databricks integration, the Customer Support team can ask you for the Edge Site ID and Job ID. The team needs this information to access details about the error.

To retrieve the Job ID, go to View the summary of a Databricks Unity Catalog synchronization.

To retrieve the Site ID:

1. Go to **Settings**.
2. In the **Edge** section, click **Sites**.
3. Click the name of the Edge site.
4. The Edge site ID is available in the **ID** field.

You receive an error when synchronizing Databricks

Issue: You receive the following error: `Error while processing crawler catalog ingestion: Import job failed with message. You are not allowed to perform this action..`

- Reason: The synchronization capability does not store any user credentials. It calls the Import API using the Edge site user credentials. By default, the Edge site user cannot add any new assets in Collibra.
- Solution: Give extra permissions to the Edge Site user. To do so, go to **Settings** → **Global Permissions** and select the **Resources** → **Manage all resources** permission for the Edge site role.

Issue: You receive the following error: `A mapping for the external system id '{your-system-asset-id}' and resource '{resource-id}' already exists.`

- For the reason and solution, go to this knowledge article.

Registering a Databricks file system via the Databricks JDBC connector and Edge

If you register a specific Databricks data source via the Databricks JDBC connector, the resulting assets represent the columns and the tables in the Databricks database.
You can retrieve sample data, and can profile and classify the data.

Before you begin

- You have enabled the following settings:
  - Database registration via Edge to allow registering a data source via Edge.
  - Database profiling via Edge to allow profiling and classification via Edge.
- You have created and installed an Edge site.

Steps

	Step	What?	Description	Results
Preparation	1	Add a Databricks JDBC connection to your Edge site	Adds a Databricks JDBC connection to your Edge site.	
	2	Add the following capabilities: • Catalog JDBC ingestion • JDBC Profiling • If you also want to collect sample data, Catalog JDBC Sampling.	Adds the required capabilities to the Databricks connection	
Setup	3	Register the data source	Registering a data source creates the structure for the metadata in Collibra.	• A **Physical Data Dictionary** domain containing a Database asset is created. • A list of available schemas is created on the **Configuration** tab page of the Database asset.
	4	Configure the synchronization of your data source	Making a selection of schemas and tables that you want to ingest.	The information on the **Configuration** tab page of the Database asset is filled in.

	Step	What?	Description	Results
Registration	5	• Synchronize one or more schemas manually • Add a synchronization schedule to synchronize automatically	Synchronizing the schema of a registered data source to make the metadata available in Collibra.	Schema, Table, Column, and Foreign Keys assets are created in the specified domain, and registration data becomes available.
	6	If needed, profile and classify the synchronized data.	Data profiling creates a summary of a data source that is registered with Data Catalog and determines the data type of columns in the data source. The summary mainly contains statistics and graphics to give the user an idea what the registered data is about. Classification analyzes and predicts the content of registered data sources based on a subset of the data itself, helping you to easily gain insights on what kinds of data you have and where it resides.	The Table and Column assets contain profiling information and the Columns are classified.

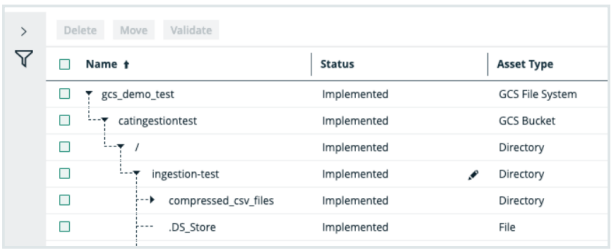
For general information on working with Databricks, go to Ways to work with Databricks.

Working with Google Cloud Storage

The Google Cloud Storage file system integration allows for the registration of Google Cloud Storage (GCS) as a data source in Collibra and the synchronization of the metadata. GCS is a service provided in the Google Cloud Platform (GCP).

About the Google Cloud Storage file system integration via Edge

The Google Cloud Storage file system integration allows for the registration of Google Cloud Storage (GCS) as a data source in Collibra and the synchronization of the metadata. GCS is a service provided in the Google Cloud Platform (GCP).
After synchronization, the files and directories of the GCS file system are represented in Collibra by specific asset types, retaining the original names.



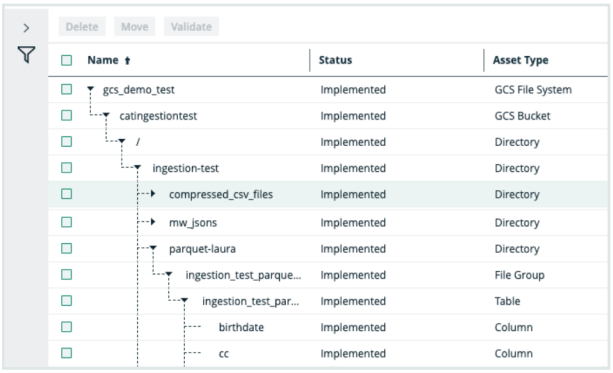
Important
- You cannot profile and classify the integrated columns and tables.
- You can only integrate a Google Cloud Storage file system via Edge, not via Jobserver.

For more information about these Google products, go to the Google Cloud Storage documentation and Google Dataplex documentation.

About Google Dataplex

The GCS integration supports Google Dataplex, a service used for schema discovery. This allows you to integrate the schemas, tables and columns from the files and create a File Group asset in Collibra rather than multiple File assets.

Important
- The Dataplex zone in which the GCS buckets are registered must be in the same project as the GCP service account.
- For integrations of Dataplex with multi-region or dual-region GCS buckets, we query all Dataplex lakes and zones that are located in the regions of the buckets and in which a Dataplex service is available. The composition of multi-regions and dual-regions, as well as the availability of a Dataplex service are hard-coded. If new regions are added or if a Dataplex service is made available in new regions, Dataplex information from these regions will not be registered until a new version of the GCS integration feature is released.



For information on how to add a GCS asset to a Dataplex Zone that can then be discovered by our GCS integration, go to the Google Dataplex documentation. For information on the supported data types, go to the data types Google documentation.

> Note When you add a bucket to Dataplex and Dataplex identifies schemas (tables and columns) for files in the bucket, these tables and columns are also added automatically to BigQuery by Dataplex.

Google Cloud Storage assets, domain types and operating model

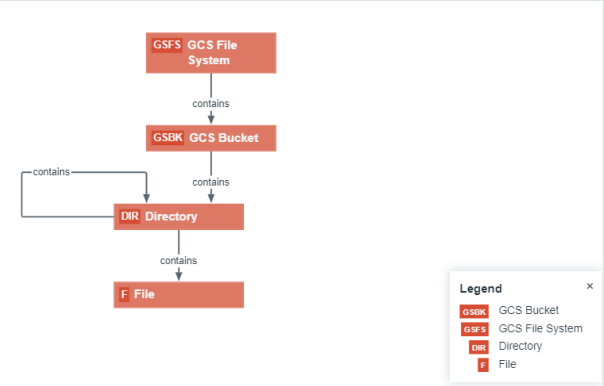
The Google Cloud Storage file system integration of Collibra Platform Self-Hosted uses a specific subset of asset types. All of these come out-of -the-box with your software.

> Note The File Group asset will only be available if you use Google Dataplex.

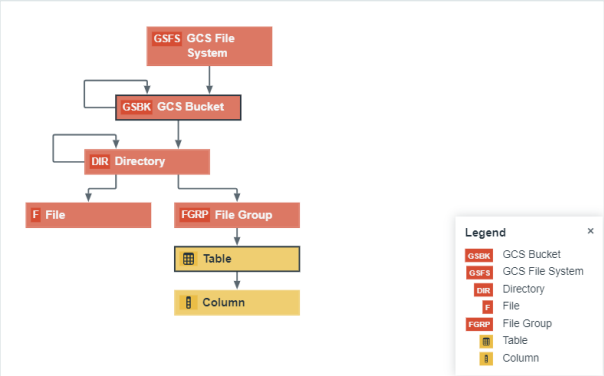
Asset type	Description	Domain type
Technology Asset ▸ File Container	An asset type that represents a Cloud File Container.	• Storage Catalog • Technology Asset Domain
Technology Asset ▸ File Container ▸ Directory	A collection of data that is treated by a computer as a unit, for the purposes of input and output.	• Storage Catalog • Technology Asset Domain
Technology Asset ▸ File Container ▸ GCS Bucket	An asset type that represents an Google Cloud Storage bucket which is a logical unit of storage containing Google Cloud Storage objects.	Storage Catalog
Technology Asset ▸ File Group	A collection of physical files which together represent a single logical file.	Storage Catalog

Asset type	Description	Domain type
Technology Asset ▸ System ▸ File Storage ▸ GCS File System	An asset type that represents a Google Cloud Storage file system.	Storage Catalog

GCS operating model



Note
If you use Google Dataplex, a Directory can contain a File Group asset instead of File assets.



Steps overview: Integrate a Google Cloud Storage file system via Edge

You can configure Collibra to register and synchronize a Google Cloud Storage (GCS) file system via Edge.

> Tip If you are using schemas with table files that you want to integrate as File Group assets with tables and columns instead of File assets, you can use Google Dataplex. The Dataplex zone in which the GCS buckets are registered must be in the same project as the GCP service account. For information on how to add a GCS asset to a Dataplex Zone that can then be discovered by the our GCS integration, go to the Google Dataplex documentation.

#	Step	Description
1	Enable the Google Cloud Storage file system registration and synchronization via Edgeand give the Edge Site user the required permissions.	Define that you want to integrate GCS via Edge. Note If you have defined an outbound (forward) proxy on your Edge site, the integration will take that configuration into account when connecting to the data source. The following proxies are supported for GCS: • Path through (No authentication) • Path through (Basic authentication) • MITM (No authentication) • MITM (Basic authentication) • No proxy for noProxy hosts defined by Edge
2	Create a GCP connection to your Edge site.	Create a connection to the Google Cloud Platform (GCP) in an Edge site.
3	Add a GCS synchronization capability to your Edge site.	Add the GCS synchronization capability to the GCP Edge connection. The capability allows to retrieve data from the GCS file system.
4	Register a GCS file system.	Create the initial structure of a Storage Catalog domain and GCS File System asset in the selected parent community.

#	Step	Description
5	Connect the GCS file system asset to the Edge capability.	Link the registered GCS file system to the Edge capability.
6	Create crawlers.	Create crawlers to define the folders that you want to synchronize.
7	Synchronize GCS.	You can manually synchronize GCS or you can add a synchronization schedule to automatically synchronize it.

Preparing Edge for Google Cloud Storage

Before you can register and synchronize Google Cloud Storage via Edge, you need to prepare your Edge site.

> Note
> If you have defined an outbound (forward) proxy on your Edge site, the integration will take that configuration into account when connecting to the data source. The following proxies are supported for GCS:
>
> - Path through (No authentication)
> - Path through (Basic authentication)
> - MITM (No authentication)
> - MITM (Basic authentication)
> - No proxy for noProxy hosts defined by Edge

Enable the Google Cloud Storage file system integration via Edge

You can enable the registration and synchronization of a Google Cloud Storage (GCS) file system via Edge.

Prerequisites

- You have the **ADMIN** or **SUPER** role in Collibra Console.
- You have the **SUPER** role in Collibra Console.
- You have the **ADMIN** or **SUPER** role in Collibra Console.

Steps

1. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.
   c. In the tab pane, click the Data Governance Center service of that environment.
   d. Click **Configuration**.
   e. Click **Edit configuration**.
2. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.
   c. In the tab pane, click the Data Governance Center service of that environment.
   d. Click **Configuration**.
   e. Click **Edit configuration**.
3. In the **Register data source** section, enter the required information:

Setting	Description
Google Cloud Storage synchronization via Edge	An option to enable Google Cloud Storage file system registration and synchronization via Edge. ◦ ✓ True: You can register and synchronize a Google Cloud Storage file system via Edge. ◦ ✕ False: You can't register a Google Cloud Storage file system via Edge.

4. Click **Save all**.

What's Next?

Give the Edge Site user the required permissions.

Required Edge User permissions

Make sure the Edge Site global role has the Manage all resources global permission.
The GCS synchronization capability does not store any user credentials. It calls the Import API using the Edge site user credentials. By default, the Edge Site user cannot add any new assets in Collibra.

What's next?

You can create the Edge connection for GCS

Create a Google Cloud Platform connection to an Edge site

After you created and installed an Edge site, you can create a connection to the Google Cloud Platform (GCP).

Before you begin

- You have created and installed an Edge site.
- You have enabled the GCS integration via Edge.

Required permissions

- You have a global role that has the **Manage connections and capabilities** global permission, for example, Edge integration engineer.
- You need a Google Cloud Platform Service Account that can read the Google Cloud Storage (GCS) file system that you want to integrate. This means the Service Account must have the permissions to list buckets (`storage.buckets.list`) and objects in a bucket (`storage.objects.list`). For information on GCP, go to the Google documentation.
- If you use Dataplex, the Service Account must be able to detect file schemas in GCS resources from Dataplex. This means the Service Account must have the following permissions `dataplex.*.get` and `dataplex.*.list`, for example, via the

Dataplex Viewer role. For information on GCP service account, go to the Google documentation. For information on Dataplex roles, go to the Google documentation.

Steps

1. Open an Edge site.
   a. On the main menu, click ⠿, and then click ⚙ **Settings**.
      » The Collibra settings page opens.
   b. In the tab pane, click **Edge**.
      » The **Sites** tab opens and shows a table with an overview of the Edge sites.
   c. In the table, click the name of the Edge site whose status is **Healthy**.
      » The Edge site page opens.
2. In the **Connections** section, click **Create connection**.
   » The **Create connection** page appears.
3. Enter the required information.

Field	Description	Required
Connection settings	This section contains the general settings of your connection.	
Name	The name of the Edge connection for Google Cloud Platform.	✓ Yes
Description	The description of the connection.	✕ No
Connection provider	The connection provider, which determines the available connection parameters. Select the **GCP connection** to connect to Google Cloud Platform.	✓ Yes
Connection parameters	This section contains the settings to connect to your data source.	

Field	Description	Required

GCP Service Account Credentials JSON	The account to connect to the GCP.
	Add the full content of the service account key JSON file.
	
	Example
	`{`
	`"type": "service_account",`
	`"project_id": "PROJECT_ID",`
	`"private_key_id": "KEY_ID",`
	`"private_key": "-----BEGIN PRIVATE KEY-----\nPRIVATE_KEY\n-----END PRIVATE KEY-----\n",`
	`"client_email": "SERVICE_ACCOUNT_EMAIL",`
	`"client_id": "CLIENT_ID",`
	`"auth_uri": "https://accounts.google.com/o/oauth2/auth",`
	`"token_uri": "https://accounts.google.com/o/oauth2/token",`
	`"auth_provider_x509_cert_url": "https://www.googleapis.com/oauth2/v1/certs",`
	`"client_x509_cert_url": "https://www.googleapis.com/robot/v1/metadata/x509/SERVICE_ACCOUNT_EMAIL"}`
	
	Ensure the service account has the required permissions.
	For more information about service account keys, go to the Google documentation.
		✓ Yes

Encryption options	Select the type of encryption used to store the Secret Access Key. Default: *To be encrypted by Edge management server*.	✓ Yes
Additional parameters	Your connection to GCP does not require any additional parameters. Delete the existing blank property.	✗ No

4. Click **Create**.
   » The connection is added to the Edge site.

What's next?

You can now add the GCS synchronization capability to an Edge site.

Add the GCS synchronization capability

After you have created a connection to the Google Cloud Platform (GCP) in your Edge site, you have to add the GCS synchronization capability to the connection.

Before you start

- You have created and installed an Edge site.
- You have enabled the GCS integration via Edge.
- You have created a connection to the Google Cloud Platform (GCP) in your Edge site.

Required permissions

- You have a global role that has the **Manage connections and capabilities** global permission, for example, Edge integration engineer.

Steps

1. Open an Edge site.
   a. On the main menu, click , and then click **Settings**.
      » The Collibra settings page opens.
   b. In the tab pane, click **Edge**.
      » The **Sites** tab opens and shows a table with an overview of the Edge sites.
   c. In the table, click the name of the Edge site whose status is **Healthy**.
      » The Edge site page opens.
2. In the **Capabilities** section, click **Add capability**.

» The **Add capability** page appears.

3. Enter the required information.

Field	Description	Required
Capability	This section contains general information about the capability.	
Name	The name of the Edge capability.	✓ Yes
Description	The description of the Edge capability.	✗ No
Capability template	The capability template. The value that you select in this field determines which sections appear on the page. Select the following Edge capability: `GCS synchronization`	✓ Yes
GCP service account	This section contains information on how to connect to Google Cloud Storage.	
GCP Connection	The GCP connection to be used.	✓ Yes
Configuration	This section contains information on the configuration of the crawlers.	
Maximum number of files per crawler	The maximum number of files that can be registered per crawler. The default value is 1,000.	✓ Yes
Save input metadata	Select the checkbox if you want to save the input metadata extracted from the data source in ZIP files. The files can be useful for troubleshooting. Select this option only on request of Collibra Support. The Collibra Support team can provide the location of the saved ZIP files after the synchronization. This checkbox is not selected by default.	✗ No

Field	Description	Required
Integrate Schemas from Dataplex	Select the checkbox if you want to integrate the schemas from Dataplex based on the crawler path that will be specified in the GCS integration configuration. If the checkbox is not selected, no Dataplex data will be ingested. This checkbox is selected by default.	✕ No
Project IDs	Add a comma-separated list of the Project IDs where Dataplex is enabled. The capability will search in these projects for schemas based on the crawler path that will be specified in the GCS integration configuration. If the Project IDs field is empty, the integration will search in the project included in the provided GCP Service Account Credentials JSON.	✕ No
Advanced Configuration	This section contains configuration options that can help when investigating issues with the capability. Important Only complete the fields **Logging configuration**, **Memory (MiB)**, and **JVM arguments** on request of or together with Collibra Support.	✕ No

4. Click **Create**.
   » The capability is added to the Edge site.
   » The fields become read-only.

What's next?

You can now register a GCS file system.

Registering and synchronizing Google Cloud Storage

After Edge is ready to integrate the Google Cloud Storage file system, you can start the registration and synchronization.

Register a Google Cloud Storage file system

You can register a Google Cloud Storage (GCS) file system in Data Catalog.

Required permissions

- The Edge site role has the Manage all resources global permission.
  The GCS synchronization capability does not store any user credentials. It calls the Import API using the Edge site user credentials. By default, the Edge site user cannot add any new assets in Collibra. Therefore, you need to give this role this permission.
- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.

Steps

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.
2. On the main toolbar, click +.
   » The **Create** dialog box appears.
3. In the **Create** dialog box, click **Register a data source**.
   » The **Register content** page opens.
4. In the Connection name column, search for the connection name you want to use to integrate the GCS file system.
5. Click **Add** for the connection you want to use.
   » The **Register Google Cloud Storage file system** page opens.
6. Enter the required information.

Field	Description
Community	The parent community in which the initial GCS structure must be created.
File system name	The name for GCS file system asset.

Field	Description
Description	The description to provide extra information about the file system. This is used as the **Description** attribute of the GCS File System asset.
Owner	The owner name of the data in the created community.

7. Click **Register**.
   » A GCS File System asset is created.
   » A Storage Catalog domain is created with the same name as the GCS File System asset.

What's next?

You can now connect the GCS File System asset to the GCS Edge connection.

Connect a GCS File System asset to the GCS synchronization capability

To retrieve data from Google Cloud Storage (GCS), you have to create a connection between the GCS File System asset and the Edge capability you want to use.
You always have to do that after registering a new GCS File System.
You can also edit the connection settings, for example, if you want to use another capability than the one you originally used.

Before you begin

- You have added the Edge GCS synchronization capability.
- You have registered a GCS file system.

Required permissions

- The Edge site role has the Manage all resources global permission.
  The GCS synchronization capability does not store any user credentials. It calls the Import API using the Edge site user credentials. By default, the Edge site user cannot add any new assets in Collibra. Therefore, you need to give this role this

permission.
- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a global role with the View Edge connections and capabilities global permission, for example, Edge integration engineer.

Steps

1. Open the GCS File System asset.
2. In the tab pane, click **Configuration**.
3. In the **Connection details** section, click **Edit connection details**.
4. Select the Edge capability you want to link this asset to.
   In the list, both the Edge site and the capability name are displayed. For example: edge-qa > gcs1-synch.
5. Click **Save**.

What's next?

You can now create crawlers.

Create a crawler for Google Cloud Storage

By creating a crawler for Google Cloud storage (GCS), you can specify which folders you want to synchronize.

Before you begin

- You have registered a GCS file system.
- You have connected the GCS File System asset to the GCS Edge capability.

Prerequisites

- You have a global role with the Catalog global permission, for example, Catalog Author.

- You have a resource role with the **Configure external system** resource permission, for example, Owner.

Steps

1. Open the GCS File System asset.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, click **Create crawler**.
   » The **Create crawler** dialog appears.

4. Enter the required information.

Field	Description
Domain	The domain in which the assets of the GCS file system are to be created.
Name	The name you want to give to the crawler in Collibra.
Include path	The case-sensitive path to a directory of a bucket in GCS. All objects and subdirectories of this path are taken into account during the synchronization. Use the following structure to refer to the path: gs://{bucketname}/{path(optional)} Example In GCS, one of the buckets is called "marketing" with directory "mkt". ◦ To include the whole bucket, the path must be: gs://marketing ◦ To only include the "mkt" directory of that bucket, the path must be: gs://marketing/mkt/
Exclude patterns	A pattern that represents the objects that are included via the **Include path**, but that you want to exclude from the synchronization. When you define a pattern, you can use the following rules: ◦ `*` matches zero or more characters. ◦ `**` matches zero or more directories in a path. ◦ `?` matches one character. Example ◦ `comm/*.jsp` matches all .jsp files in the comm directory. ◦ `comm/t?st.jsp` matches comm/test.jsp but also comm/tast.jsp or comm/txst.jsp. ◦ `commm/**/test.jsp` matches all test.jsp files in the comm path. ◦ `org/framework/**/*.jsp` matches all .jsp files in the org/framework path. ◦ `org/**/servlet/test.jsp` matches org/framework/servlet/test.jsp but also org/framework/testing/servlet/test.jsp and org/servlet/test.jsp.
Add pattern	A button to add additional exclude patterns.

5. Click **Create**.

Example on how the Include path and the Exclude patterns work together

In bucket1 of the GCS system, the following files exist:

myfolder/departments/finance.json
myfolder/departments/market-us.json
myfolder/departments/market-emea.json
myfolder/departments/market-ap.txt
myfolder/employees/hr.json
myfolder/employees/john.csv
myfolder/employees/jane.csv
myfolder/employees/juan.txt
myfolder/report.xlsx
rubbish.txt

Below, you find the results for several Include path and Exclude patterns combinations:

Include path	Exclude pattern	What does it mean?	Result
gs://bucket1/	<none>	All files in gs://bucket1 are taken into account.	myfolder-/de-partments/finance.json myfolder-/departments/market-us.json myfolder-/departments/market-emea.json myfolder-/departments/market-ap.txt myfolder-/employees/hr.json myfolder-/employees/john.csv myfolder-/employees/jane.csv myfolder-/employees/juan.txt myfolder/report.xlsx rubbish.txt

Include path	Exclude pattern	What does it mean?	Result
gs://bucket1	<none>	All files in gs://bucket1 are taken into account.	myfolder-/de-partments/finance.json myfolder-/departments/market-us.json myfolder-/departments/market-emea.json myfolder-/departments/market-ap.txt myfolder-/employees/hr.json myfolder-/employees/john.csv myfolder-/employees/jane.csv myfolder-/employees/juan.txt myfolder/report.xlsx rubbish.txt
bucket1	<none>	None of the files are taken into account because the Include path is not correct.	<none>

Include path	Exclude pattern	What does it mean?	Result
gs://bucket1/	*.txt **.json	All files in gs://bucket1/ are taken into account, except: • the TXT files in the main folder gs://bucket1 • the JSON files the main folder gs://bucket1	myfolder/departments/finance.json myfolder/departments/market-us.json myfolder/departments/market-emea.json myfolder/departments/market-ap.txt myfolder/employees/hr.json myfolder/employees/john.csv myfolder/employees/jane.csv myfolder/employees/juan.txt myfolder/report.xlsx
gs://bucket1/	**/*.txt myfolder-/employees/*.json	All files in gs://bucket1/ are taken into account, except: • theTXT files in all sub-folders of gs://bucket1 • the JSON files in sub-folder gs://buck-et1/my-folder/employees/	myfolder/departments/finance.json myfolder/departments/market-us.json myfolder/departments/market-emea.json myfolder/employees/john.csv myfolder/employees/jane.csv myfolder/report.xlsx

Include path	Exclude pattern	What does it mean?	Result
gs://bucket1/	myfolder/**/*txt	All files in gs://bucket1/ are taken into account, except the TXT files in all subfolders of gs://bucket1/myfolder/.	myfolder/departments/finance.json myfolder/departments/market-us.json myfolder/departments/market-emea.json myfolder/employees/hr.json myfolder/employees/john.csv myfolder/employees/jane.csv myfolder/report.xlsx rubbish.txt
gs://bucket1/myfolder	employees/* myfolder-/departments/*	All files in gs://bucket1/myfolder/ are taken into account except: • all files in all sub-folders of gs://bucket1/my-folder/employees • all files in gs://bucket1/my-folder/myfolder/departments/	myfolder/departments/finance.json myfolder/departments/market-us.json myfolder/departments/market-emea.json myfolder/departments/market-ap.txt myfolder/report.xlsx
gs://bucket1/my-folder/departments	*json	All files in gs://bucket1/myfolder/departments are taken into account except all JSON files in this folder.	myfolder/departments/market-ap.txt

Include path	Exclude pattern	What does it mean?	Result
gs://bucket1/	**/j???.*	All files in gs://bucket1/ are taken into account, except the files starting with j followed by three characters, from all sub-folders in bucket1.	myfolder/departments/finance.json myfolder/departments/market-us.json myfolder/departments/market-emea.json myfolder/departments/market-ap.txt myfolder/employees/hr.json myfolder/report.xlsx rubbish.txt
gs://bucket1	myfolder/**	All files in gs://bucket1/ are taken into account expect for the files in myfolder/	rubbish.txt

What's next?

You can now synchronize GCS manually or define a synchronization schedule.

About synchronizing Google Cloud Storage

Synchronizing Google Cloud Storage(GCS) is the process of ingesting metadata from a selected GCS repository and making the data available in Collibra Platform Self-Hosted.

When you synchronize GCS, the content of your repository is analyzed and represented in Collibra by means of assets and their characteristics. Collibra also takes the defined crawlers into account.

You can synchronize manually, or you can automate it by adding a synchronization schedule. You can only synchronize one GCS File System at a time.

- If a synchronization job is in progress and a second one is triggered, manually or automatically, the second job is queued.

- If a synchronization job is still running and a new synchronization of the same GCS File System is triggered (manually or automatically), the running synchronization continues and the new synchronization request is ignored.

After the synchronization, the resulting assets are in the domain that was specified in the crawler. For information on the integrated data, go to Integrated Google Cloud Storage data.

Synchronize Google Cloud Storage manually

You can manually start a synchronization job of a GCS File System asset. This can be useful if you want to test your crawlers, or if you want to synchronize immediately. You can also add a synchronization schedule to synchronize automatically.

Before you begin

- You have registered a GCS file system.
- You have connected the GCS File System asset to the GCS Edge capability.
- If needed, you have defined crawlers.

Required permissions

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.

Steps

1. Open the GCS File System asset.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, click **Synchronize now**.
   » A notification indicates synchronization has started.

   » The synchronization job appears in the **Activities** list as a bulk synchronization.

When the synchronization finishes, the resulting assets, including their attributes and relations, are created, edited or deleted in the selected domain(s) and in the Data Sources page of Data Catalog.
If one of the directories in GCS doesn't have a name, we will create a unique name for the asset in Collibra.

> Note In case of a partial synchronization caused by a temporary communication issue, the status of the assets that cannot be synchronized is set to **Missing from source**. Their previous status is restored, if they are found in the source system during the next fully successful synchronization.

What's next?

You can view a summary of the results from the Activities list.
You can view the assets in their domain.

Add a Google Cloud Storage synchronization schedule

To keep the content of Collibra Platform Self-Hosted synchronized with your Google Cloud Storage (GCS) file system, you can synchronize manually or create a schedule to automatically do this with a fixed interval.

> Note You can only create one synchronization schedule.

Before you begin

- You have registered a GCS file system.
- You have connected the GCS File System asset to the GCS Edge capability.
- If needed, you have defined crawlers.

Required permissions

- You have a resource role with the **Configure external system** resource permission, for example, Owner.

- You have a global role with the Catalog global permission, for example, Catalog Author.

Steps

1. Open the GCS File System asset.
2. In the tab pane, click **Configuration**.
3. In the **Synchronization schedule** section, click **Add Schedule**.
4. Enter the required information.

Field	Description
Repeat	The interval when you want to synchronize automatically. The possible values are: **Daily**, **Weekly**, **Monthly**, and **Cron expression**.
Cron	The Quartz Cron expression that determines when the synchronization takes place. This field is only visible if you select `Cron expression` in the **Repeat** field.
Every	The day on which you want to synchronize, for example, Sunday. This field is only visible if you select `Weekly` in the **Repeat** field.
Every first	The day of the month on which you want to synchronize, for example, Tuesday. This field is only visible if you select `Monthly` in the **Repeat** field.
At	The time at which you want to synchronize automatically, for example, 14:00. ◦ You can only schedule on the hour. For example, you can add a synchronization schedule at 8:00, but not at 8:45. If you try to add it at 8:45, we will default it to 8:00. Use a cron expression if you don't want to schedule on the hour. ◦ This field is only visible if you select `Daily`, `Weekly`, or `Monthly` in the **Repeat** field.
Time zone	The time zone for the schedule.

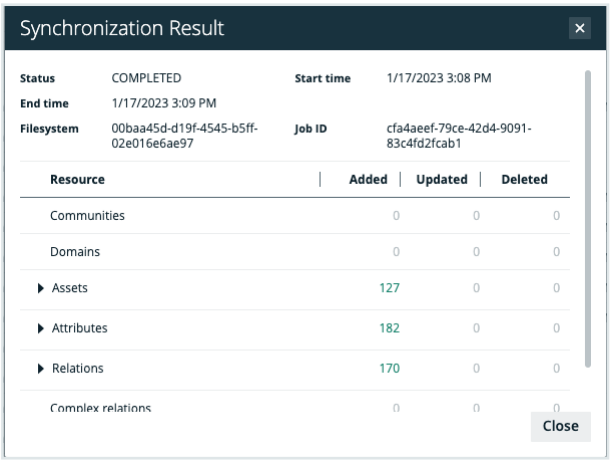
5. Click **Save**.

View the summary of a Google Cloud Storage synchronization

After you synchronized Google Cloud Storage (GCS), you can view the summary of the results. This shows the impact of the synchronization on the assets in Collibra Platform Self-Hosted

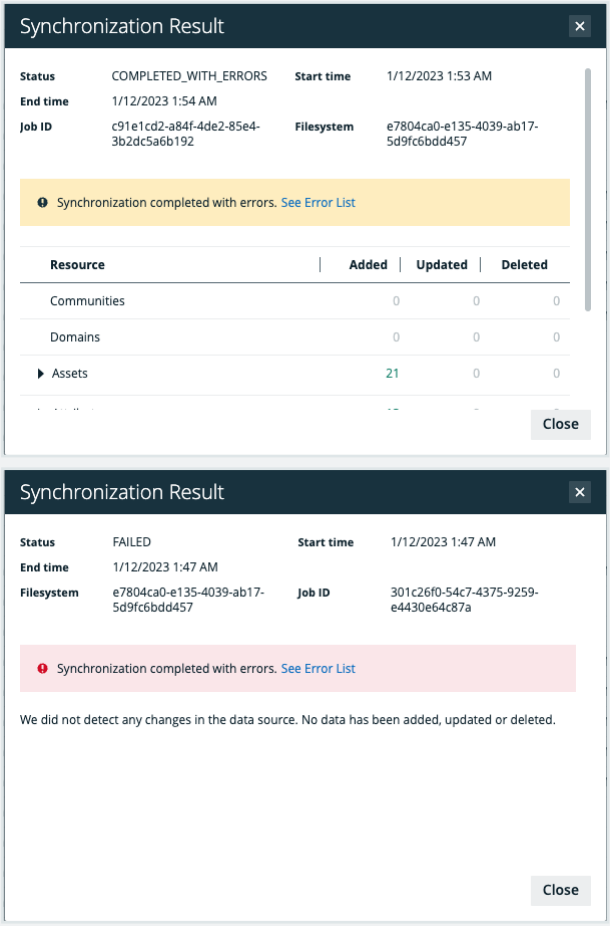
Steps

1. Open the Activities list.
2. In the row containing the GCS synchronization job, click **Result**.
   » The **GCS synchronization results** dialog box appears.



Note

- If anything changed, the information about the total number of resources that were added, modified or removed as a result of the synchronization are displayed.
- In case of errors, you can receive additional information about the error.



Tip The **Job ID** is useful when troubleshooting your synchronization process with Collibra Support.

For information on the resulting assets, see Integrated Google Cloud Storage data.

Integrated Google Cloud Storage data

After the synchronization, the resulting assets are in the domain that was specified in the crawler.

> Warning Do not move the assets to another domain. Doing so may lead to errors during future synchronizations.

> Tip GCS synchronization relies on UUIDs.

> Note In case of a partial synchronization caused by a temporary communication issue, the status of the assets that cannot be synchronized is set to **Missing from source**. Their previous status is restored, if they are found in the source system during the next fully successful synchronization.

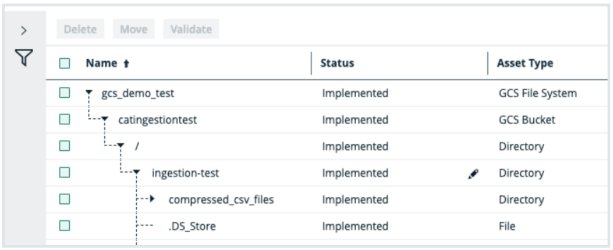
By default, the assets are shown in a plain list, but you can enable a multi-path hierarchy to show it in a tree structure. The resulting assets depend on whether you use Google Dataplex.

Synchronization results without Google Dataplex

For the best result, we recommend that you use the following relations:

1. File Storage contains File Container
2. File Container contains File Container
3. File Container contains File
4. Directory contains Directory

The following images shows the resulting hierarchical table.

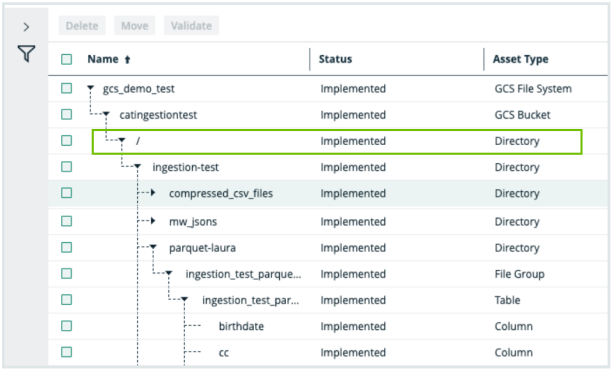


Synchronization results with Google Dataplex

For the best result, we recommend that you use the following relations:

1. File Storage contains File Container
2. File Container contains File Container
3. File Container contains File
4. Directory contains Directory
5. Directory contains File Group
6. File Group contains Table
7. Table contains Column

The following images shows the resulting hierarchical table.



The synchronization creates a Directory asset named /. This is needed to ensure the asset can contain the File Group assets.

Synchronized metadata per asset type

This table shows the metadata for each GCS asset type.

Asset type	Synchronized metadata	Resource ID
GCS Bucket	File Storage contains/ is part of File Container	00000000-0000-0000-0001-002600000000
	Location	00000000-0000-0000-0000-000000000203

Asset type	Synchronized metadata	Resource ID
Directory	URL	00000000-0000-0000-0000-000000000258
	File Container contains/ is part of File Container	00000000-0000-0000-0001-002600000001
	Directory contains/ is part of Directory	00000000-0000-0000-0001-002600000003
File Group	File Type	00000000-0000-0000-0001-002500000012
	Directory contains/ is part of File Group	00000000-0000-0000-0001-002600000004
File	URL	00000000-0000-0000-0000-000000000258
	File Container contains/ contained in File	00000000-0000-0000-0000-000000007060
Table	Description	00000000-0000-0000-0000-000000003114
	File Group contains/ is part of Table	00000000-0000-0000-0001-002600000005

Asset type	Synchronized metadata	Resource ID
Column	Description	00000000-0000-0000-0000-000000003114
	Technical Data Type	00000000-0000-0000-0000-000000000219
	Column Position	00000000-0000-0000-0001-000500000020
	Is Nullable	00000000-0000-0000-0001-000500000011
	Column is part of/ contains Table	00000000-0000-0000-0000-000000007042

Troubleshooting Google Cloud Storage integration

Where do I find the Edge Site ID and Job ID?

If you report an error with Google Cloud Storage (GCS) integration, the Customer Support team can ask you for the Edge Site Id and Job ID. The team needs this information to access details about the error.

To retrieve the Job ID, see View the summary of a Google Cloud Storage synchronization.

To retrieve the Site ID:

1. Go to Settings.
2. In the Edge section, click **Sites**.
3. Click the name of the Edge site.
4. The Edge site ID is available in the ID field.

You receive an error when synchronizing GCS

Issue: You receive the following error: `Error while processing crawler catalogingestion: Import job failed with message. You are not allowed to perform this action..`

Reason: The GCS synchronization capability does not store any user credentials. It calls the Import API using the Edge site user credentials. By default, the Edge site user cannot add any new assets in Collibra.

Solution: Give extra permissions to the Edge site user. To do so, go to **Settings** → **Global Permissions** and select the **Resources** → **Manage all resources** permission for the Edge site role.

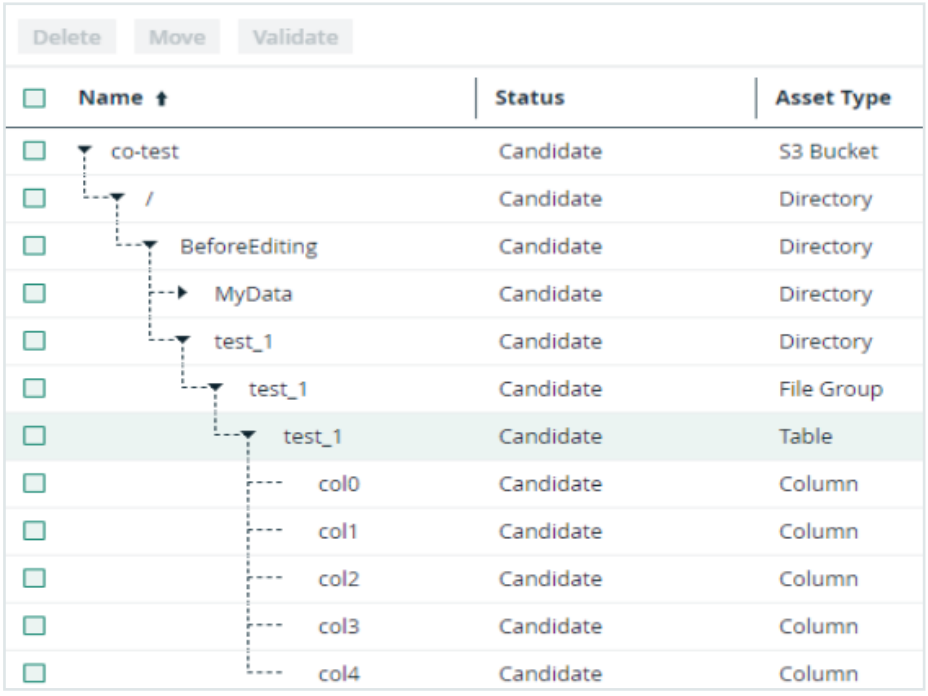
Chapter 8

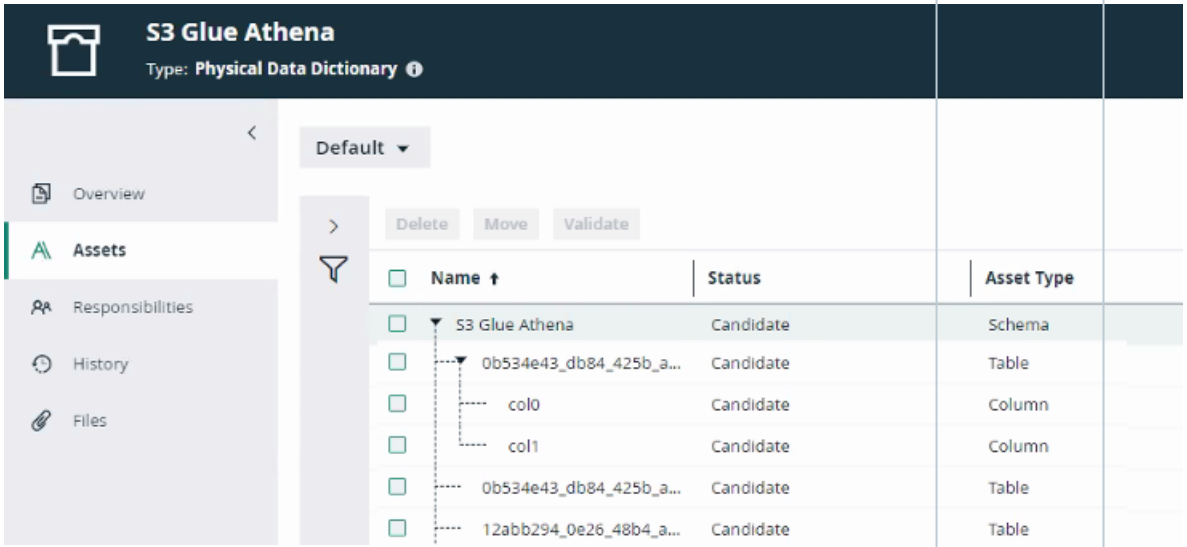
Working with Amazon S3

Amazon S3 or Amazon Simple Storage Service is an online object storage service hosted by Amazon. For more information, visit the Amazon S3 documentation.

Two ways to work with Amazon S3

Amazon S3 or Amazon Simple Storage Service is an online object storage service hosted by Amazon. For more information, visit the Amazon S3 documentation.
In Collibra Platform Self-Hosted, you can either integrate or register an Amazon S3 file system. It's important to understand the difference between integrating and registering because the result in Collibra is different.

Possible way to work with Amazon S3	Result in Collibra	Steps
Integrating an Amazon S3 file system	If you integrate an Amazon S3 file system, the resulting assets represent the Amazon S3 folder structure by means of Amazon S3 Bucket, Directory, File, Table and Column assets. Note that you can't profile and classify the columns and tables. Example 	• Via Edge • via Job-server

Possible way to work with Amazon S3	Result in Collibra	Steps
Registering an Amazon S3 file system via the AWS Glue JDBC connector	If you register an Amazon S3 file system via the AWS (Amazon Web Services) Glue JDBC connector, the resulting assets represent the columns and the tables in Amazon S3 without the folder context. You can profile and classify the data, but the folder structure of your Amazon S3 environment isn't represented in Data Catalog.The AWS Glue JDBC connector leverages the Athena JDBC driver. Example 	• via Edge • via Jobserver

Integrating an Amazon S3 file system

If you integrate an Amazon S3 file system, the resulting assets represent the Amazon S3 folder structure by means of Amazon S3 Bucket, Directory, File, Table and Column assets. Note that you can’t profile and classify the columns and tables.

About integrating an Amazon S3 file system

The Amazon S3 file system integration allows for the registration of Amazon S3 as a data source in Collibra and the synchronization of metadata in Amazon S3. After the synchronization, the files and directories of Amazon S3 are represented in Collibra by specific asset types, retaining the original names. However, not all file types are fully supported.

> Note
> - You can restrict the AWS regions to which Data Catalog is allowed to connect. This step is recommended for efficient synchronization.
> - If you integrate an Amazon S3 file system, you can't profile or classify data. If you want to be able to profile and classify the data, go to Two ways to work with Amazon S3.

You can integrate Amazon S3 file systems via Edge or via Jobserver.

Amazon S3 asset and domain types

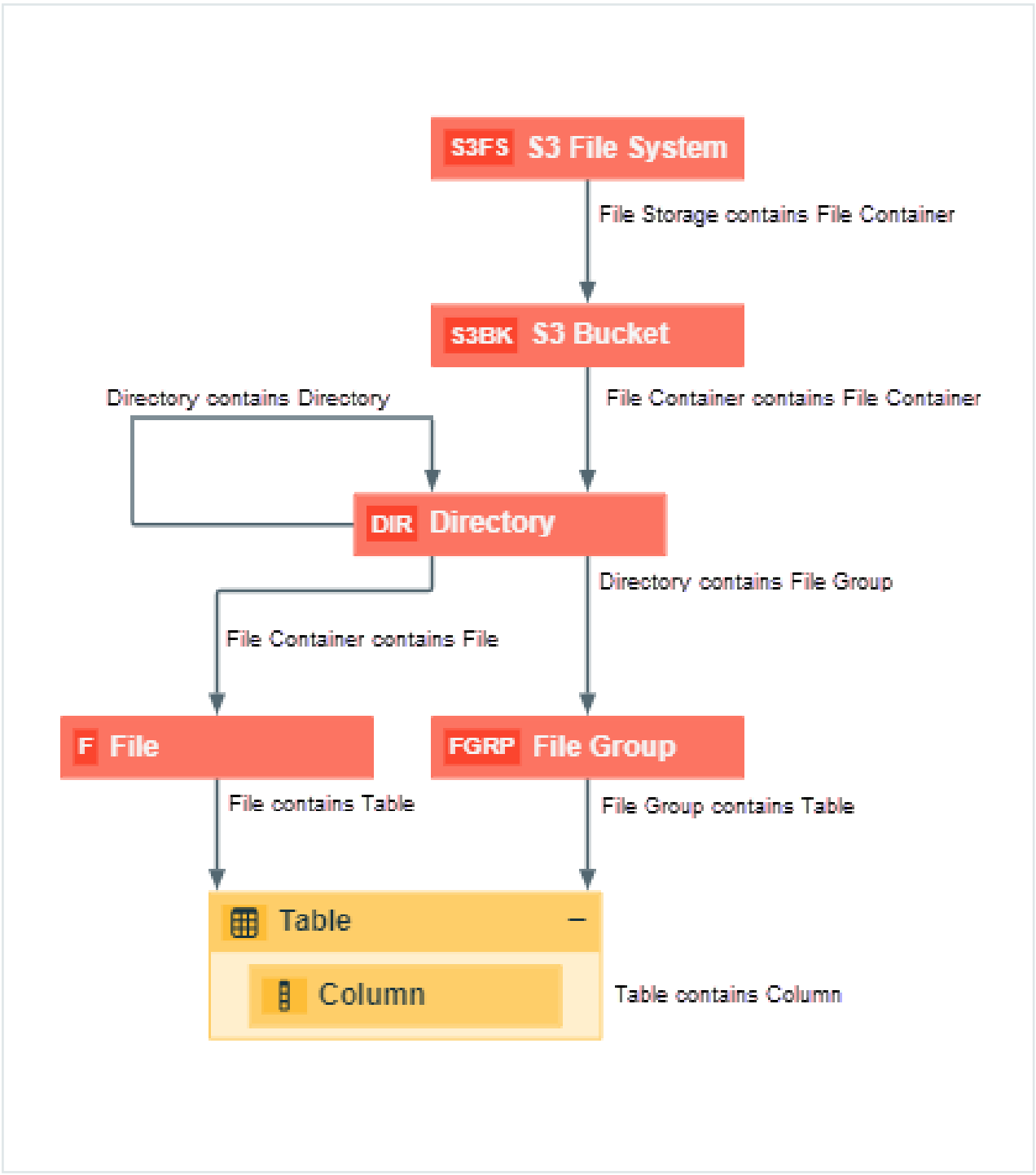
The Amazon S3 file system integration of Collibra Platform Self-Hosted uses a specific subset of asset types. All of these come out of the box with your software.

Asset type	Description	Domain type
Data Asset ▸ Data Element ▸ Column	An atomic unit of data that can be stored in a database table. Examples: FST_NM, EMPID	• Physical Data Dictionary • Storage Catalog
Data Asset ▸ Data Structure ▸ Table	An implementation of data entities in columns and rows, in a given database system. It is the basic structure of a relational database. Examples: Account_tbl, CUST_ADDR	• Physical Data Dictionary • Storage Catalog

Asset type	Description	Domain type
Data Asset ▸ Data Structure ▸ Table ▸ Database View	A Database View is a virtual table based on the result-set of an SQL statement.	• Physical Data Dictionary • Storage Catalog
Technology Asset ▸ File Container	An asset type that represents a Cloud File Container.	• Storage Catalog • Technology Asset Domain
Technology Asset ▸ File Container ▸ Directory	A collection of data that is treated by a computer as a unit, for the purposes of input and output.	• Storage Catalog • Technology Asset Domain
Technology Asset ▸ File Container ▸ S3 Bucket	An asset type that represents an Amazon S3 Bucket, which is a logical unit of storage containing Amazon S3 Objects.	Storage Catalog
Technology Asset ▸ File Group	A collection of physical files which together represent a single logical file.	Storage Catalog
Technology Asset ▸ System ▸ File Storage	An asset type that represents a Cloud File Storage bucket.	Storage Catalog
Technology Asset ▸ System ▸ File Storage ▸ S3 File System	Amazon S3 (Simple Storage Service) file system abstraction.	Storage Catalog

Amazon S3 operating model

The following image shows the relations between S3 asset types and the cardinality of the relation types in the assets' assignment.



Amazon S3 supported file types

Amazon S3 can contain a wide range of objects in different file types. However, not all file types are fully supported due to limitations of AWS Glue.

The following list shows the file types that are supported by Collibra Platform Self-Hosted. Note that other file types may work properly as well. For an exhaustive list of supported file types, see the AWS Glue documentation.

- AVRO
- ORC
- PARQUET
- JSON
- BSON
- XML
- ION
- COMBINED_APPACHE
- APACHE
- LINUX_KERNEL
- RUBY_LOGGER
- SQUID
- REDISMONLOG
- REDISLOG
- CSV
- ZIP
- TAR
- RAR
- GZ
- JAR

Integrating an Amazon S3 file system via Edge

Restrict AWS regions

You can restrict the AWS regions to which Collibra Data Catalog can connect.

> Note When there is no restriction, the S3 integration will make requests to all possible AWS regions, which could result in long synchronization times.

Prerequisites

- You have the **ADMIN** or **SUPER** role in Collibra Console.
- You have the **SUPER** role in Collibra Console.
- You have the **ADMIN** or **SUPER** role in Collibra Console.

Steps

1. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.
   c. In the tab pane, click the Data Governance Center service of that environment.
   d. Click **Configuration**.
   e. Click **Edit configuration**.
2. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.
   c. In the tab pane, click the Data Governance Center service of that environment.
   d. Click **Configuration**.
   e. Click **Edit configuration**.

3. In the **Register data source** section, enter the required information:
4.

Setting	Description
AWS regions restriction	A list of AWS regions Data Catalog is allowed to connect to. For example, *eu-west-3* and *us-east-2*. For a list of all AWS locations, see the AWS documentation. ◦ If you want to allow Collibra to make a connection to any AWS region, leave the field empty. ◦ If you remove a region from this list and the region was previously used for an S3 integration, you may want to delete the Glue database from the previously used region manually. By default, Collibra does not remove it. The Glue database has the following naming convention: collibra_catalog_<Asset Id>_<Domain Id> For example: collibra_catalog_d3174a88-5ffe-4d50-8fbe-7bf0832ec3af_5d198ce9-4e56-4d0e-a885-58204da50741 ◦ When using Edge, a warning is added to the logs if an invalid region is detected in the restricted regions list.

5. Click **Save all**.

Enable the Amazon S3 file system integration via Edge

You can enable the integration of an Amazon S3 file system via Edge.

Prerequisites

- You have the **ADMIN** or **SUPER** role in Collibra Console.
- You have the **SUPER** role in Collibra Console.
- You have the **ADMIN** or **SUPER** role in Collibra Console.

Steps

1. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.

c. In the tab pane, click the Data Governance Center service of that environment.
d. Click **Configuration**.
e. Click **Edit configuration**.

2. Open the DGC service settings for editing:
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, expand an environment to show its services.
   c. In the tab pane, click the Data Governance Center service of that environment.
   d. Click **Configuration**.
   e. Click **Edit configuration**.
3. In the **Register data source** section, enter the required information:

Setting	Description
Amazon S3 synchronization via Edge	An option to enable Amazon S3 file system registration and synchronization via Edge. ◦ ✔ True: You can register and synchronize an Amazon S3 file system via Edge. ◦ ✖ False: You can only register an Amazon S3 file system via Jobserver. Note Enabling the registration of an Amazon S3 file system via Edge does not prevent you from registering an Amazon S3 file system via Jobserver.

4. Click **Save all**.

Managing crawlers

A crawler is an automated script that ingests data from Amazon S3 to Data Catalog.

You can create, edit and delete crawlers in Collibra Platform Self-Hosted. When you synchronize Amazon S3, the crawlers are created in AWS Glue and executed. Each crawler crawls a location in Amazon S3 based on its include path. The results are stored in one AWS Glue database per domain assigned to one or more crawlers. Those databases are ingested in Data Catalog in the form of assets, attributes and relations. The databases are stored in AWS Glue until the next synchronization. At that moment, they are deleted

and recreated. The crawlers in AWS Glue are deleted immediately after the synchronization is finished.

> Important
> - If you completed the **Glue database configuration** parameter in the capability, you need to create a dummy crawler. A dummy crawler is a crawler with an invalid include path, such as s3://dummy. This crawler won't be taken into account when you run the synchronization.
> In a future release, we'll remove the need for a dummy crawler.
> - By default, AWS Glue allows up to 25 crawlers per account. For more information, see the AWS Glue documentation. This has consequences for Collibra:
>   - If you created crawlers in AWS Glue directly, Collibra can create less crawlers for synchronization.
>   - Because Collibra creates the crawlers in AWS Glue during synchronization, you should avoid having 25 or more crawlers in one S3 File System asset.
>   - You can synchronize several S3 File System assets simultaneously, but if the total number of crawlers exceeds the maximum amount in AWS Glue, synchronization will fail. Since Collibra deletes the crawlers from AWS Glue after synchronization, it is safer to synchronize each S3 File System asset at a unique time.
> - Crawlers in AWS Glue can crawl multiple buckets, but in Collibra, each crawler can only crawl a single bucket.

Create a crawler

You can create a crawler for an S3 File System asset in Data Catalog.

Prerequisites

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, click **Create crawler**.
   » The **Create crawler** dialog box appears.

4. Enter the required information.

Field	Description
Domain	The domain in which the assets of the S3 file system are created. More information about linking domains to crawlers: ◦ A specific Storage Catalog domain is created automatically when the S3 File System asset is created. That domain is selected by default. However, you can manually create a new Storage Catalog domain and select it. ◦ If multiple crawlers point to the same domain, then all assets are created in the same domain. ◦ If multiple crawlers point to different domains, then all assets are created in their respective domains. ◦ If multiple crawlers from the same S3 File System asset overlap and point to different domains, then overlapping assets are created in each domain. ◦ If multiple crawlers from the same S3 File System asset overlap and point to the same domain, then overlapping assets are created once in that domain. ◦ If crawlers from multiple S3 File System assets overlap and point to different domains, then overlapping assets are created in each domain. ◦ If crawlers from multiple S3 File System assets overlap and point to the same domain, then overlapping assets are created once in the domain and the S3 Bucket asset has a relation to both S3 File System assets.

Field	Description
Name	The name of the crawler in Collibra. More information about crawler names: ◦ You cannot use the same name for two crawlers in the same S3 File System asset. ◦ The name of the corresponding crawler in AWS Glue will contain this name. Its name will follow the following convention: `collibra_catalog_<s3fs asset id>_<name_of_the_crawler_in_Collibra>`. ◦ The crawler name must be compliant with the AWS Glue limitations: ▪ It has to match the single-line string pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\t]*`. ▪ The length should be between 1 and 255 bytes long, including the fixed prefix that Collibra adds. That means that you can use roughly 65 characters, depending on the characters that were used. Warning This restriction is imposed by Amazon S3, which allows up to 255 bytes, including the prefix added by Collibra. If you enter too many characters and exceed the byte limit, synchronization fails.
Include path	The case-sensitive path to a directory of a bucket in Amazon S3. All objects and subdirectories of this path are crawled. For more information and examples, see the AWS Glue documentation.
Exclude patterns	Glob pattern that represents the objects that are in the include path, but that you want to exclude. For more information and examples, see the AWS Glue documentation.
Add pattern	Button to add additional exclude patterns.

5. Click **Create**.

What's next?

You can now synchronize Amazon S3 manually or define a synchronization schedule.

Edit a crawler

You can edit a crawler of an S3 File System asset in Data Catalog. For example, you can do this if you want to edit the exclude pattern.

Prerequisites

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, in the row of the crawler that you want to edit, click .

» The **Edit crawler** window appears.

4. Enter the required information.

Field	Description
Domain	The domain in which the assets of the S3 file system are created. More information about linking domains to crawlers: ◦ A specific Storage Catalog domain is created automatically when the S3 File System asset is created. That domain is selected by default. However, you can manually create a new Storage Catalog domain and select it. ◦ If multiple crawlers point to the same domain, then all assets are created in the same domain. ◦ If multiple crawlers point to different domains, then all assets are created in their respective domains. ◦ If multiple crawlers from the same S3 File System asset overlap and point to different domains, then overlapping assets are created in each domain. ◦ If multiple crawlers from the same S3 File System asset overlap and point to the same domain, then overlapping assets are created once in that domain. ◦ If crawlers from multiple S3 File System assets overlap and point to different domains, then overlapping assets are created in each domain. ◦ If crawlers from multiple S3 File System assets overlap and point to the same domain, then overlapping assets are created once in the domain and the S3 Bucket asset has a relation to both S3 File System assets.

Field	Description
Name	The name of the crawler in Collibra. More information about crawler names: ◦ You cannot use the same name for two crawlers in the same S3 File System asset. ◦ The name of the corresponding crawler in AWS Glue will contain this name. Its name will follow the following convention: `collibra_catalog_<s3fs asset id>_<name_of_the_crawler_in_Collibra>`. ◦ The crawler name must be compliant with the AWS Glue limitations: ▪ It has to match the single-line string pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\t]*`. ▪ The length should be between 1 and 255 bytes long, including the fixed prefix that Collibra adds. That means that you can use roughly 65 characters, depending on the characters that were used. Warning This restriction is imposed by Amazon S3, which allows up to 255 bytes, including the prefix added by Collibra. If you enter too many characters and exceed the byte limit, synchronization fails.
Include path	The case-sensitive path to a directory of a bucket in Amazon S3. All objects and subdirectories of this path are crawled. For more information and examples, see the AWS Glue documentation.
Exclude patterns	Glob pattern that represents the objects that are in the include path, but that you want to exclude. For more information and examples, see the AWS Glue documentation.
Add pattern	Button to add additional exclude patterns.

5. Click **Save**.

Delete a crawler

You can delete a crawler from an S3 File System asset.

> Note If you delete an S3 File System asset that contains one or more crawlers, the crawlers are also deleted.

Prerequisites

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, in the row of the crawler that you want to delete, click .
   » The **Delete Crawler** confirmation message appears.
4. Click **Delete crawler**.

Synchronizing Amazon S3

When you synchronize Amazon S3, the content of your Amazon S3 repository is analyzed and represented by means of assets and their characteristics.
You can synchronize manually, or you can automate it by adding a synchronization schedule by means of a cron expression.

- You can only synchronize one S3 File System at a time. If a synchronization job is in progress and a second one is triggered, manually or automatically, it will be queued.
- If a synchronization job is still running and a new synchronization of the same S3 File System is triggered (manually or automatically), the running synchronization will continue and the new synchronization request is ignored.

Technically, the synchronization happens in several steps:

- If you **did not** completed the **Glue database configuration** parameter in the capability:

a. Collibra creates crawlers in AWS Glue, based on the crawlers defined in Collibra.
b. If AWS Glue contains databases with metadata from a previous synchronization, the databases are deleted.
c. Each AWS Glue crawler crawls a location in Amazon S3 based on its include path. For each domain assigned to one or more crawlers, AWS Glue creates a database with the crawling results.
d. Collibra ingests those databases and creates assets, attributes and relations as required to match the metadata.
   The resulting assets are in the domain that was specified in the crawler.
e. The AWS Glue crawlers are deleted.

- If you **did** completed the **Glue database configuration** parameter in the capability:

  Collibra ingests those databases defined in **Glue database configuration** parameter and creates assets, attributes and relations as required to match the metadata.
  The resulting assets are added to the domain specified in the parameter.

  The glue database is never deleted, even if the **Delete Glue database left after previous synchronization of the file system** parameter is selected in the capability.

> Warning Do not move the assets to another domain. Doing so may lead to errors during future synchronizations. This is a known limitation.

Naming convention

Synchronizing Amazon S3 relies on a naming convention to match assets during the synchronization process. We highly recommend that you not change the S3 File System asset's full name.

> Warning Editing full name of the S3 File System assets may lead to errors during the synchronization process.

Synchronize Amazon S3 manually

You can manually start a synchronization job of an S3 File System asset. This can be useful if you want to test your crawlers, or if you want to synchronize immediately.

> Tip You can also add a synchronization schedule to synchronize automatically.

Prerequisites

- You have registered an Amazon S3 file system.
- You have a programmatic AWS user and IAM role with the required permissions.
- You have created one or more crawlers.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a resource role with the Configure external system resource permission on the community or domain that contains the S3 File System, for example Owner.
- You have a role with the following resource permissions on the S3 community you created when you registered an Amazon S3 file system:
  - Asset: add
  - Attribute: add
  - Domain: add
  - Attachment: add

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, click **Synchronize now**.
   » A notification indicates synchronization has started.

   » The synchronization job appears in the **Activities** list as a bulk synchronization.

   » The **Synchronization schedule** section displays the time of the last synchronization.

» Once the synchronization is completed, you can view a summary of the results from the **Activities** list and you can view the assets in their domain. For more information, go to Integrated Amazon S3 data.

> Note In case of a partial synchronization caused by a temporary communication issue, the status of the assets that cannot be synchronized is set to **Missing from source**. During the next fully successful synchronization, the assets are removed or their previous status is restored, depending on their actual status in the source system.

Add an S3 synchronization schedule

To keep the content of Collibra Platform Self-Hosted synchronized with your Amazon S3 File System, you can synchronize manually or create a schedule to automatically do this with a fixed interval.

> Note You can only create one synchronization schedule.

Prerequisites

- You have a resource role with the Configure external system resource permission on the community or domain that contains the S3 File System, for example Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.
- You have a programmatic AWS user and IAM role with the required permissions.
- You have created one or more crawlers.
- You have a role with the following resource permissions on the S3 community you created when you registered an Amazon S3 file system:
  - Asset: add
  - Attribute: add
  - Domain: add
  - Attachment: add

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Synchronization schedule** section, click **Add Schedule**.
4. Enter the required information.

Field	Description
Repeat	The interval when you want to synchronize automatically. The possible values are: **Daily**, **Weekly**, **Monthly**, and **Cron expression**.
Cron	The Quartz Cron expression that determines when the synchronization takes place. This field is only visible if you select `Cron expression` in the **Repeat** field.
Every	The day on which you want to synchronize, for example, Sunday. This field is only visible if you select `Weekly` in the **Repeat** field.
Every first	The day of the month on which you want to synchronize, for example, Tuesday. This field is only visible if you select `Monthly` in the **Repeat** field.
At	The time at which you want to synchronize automatically, for example, 14:00. ◦ You can only schedule on the hour. For example, you can add a synchronization schedule at 8:00, but not at 8:45. If you try to add it at 8:45, we will default it to 8:00. Use a cron expression if you don't want to schedule on the hour. ◦ This field is only visible if you select `Daily`, `Weekly`, or `Monthly` in the **Repeat** field.
Time zone	The time zone for the schedule.

5. Click **Save**.

Edit an S3 synchronization schedule

You can edit the synchronization schedule of an Amazon S3 File System asset. For example, you can do this if you think the synchronization job runs too often or not often enough.

Prerequisites

- You have a resource role with the Configure external system resource permission on the community or domain that contains the S3 File System, for example Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.
- You have a programmatic AWS user and IAM role with the required permissions.
- You have created one or more crawlers.
- You have a role with the following resource permissions on the S3 community you created when you registered an Amazon S3 file system:
  - Asset: add
  - Attribute: add
  - Domain: add
  - Attachment: add

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Synchronization schedule** section, click **Edit Schedule**.

4. Enter the required information.

Field	Description
Repeat	The interval when you want to synchronize automatically. The possible values are: **Daily**, **Weekly**, **Monthly**, and **Cron expression**.
Cron	The Quartz Cron expression that determines when the synchronization takes place. This field is only visible if you select `Cron expression` in the **Repeat** field.
Every	The day on which you want to synchronize, for example, Sunday. This field is only visible if you select `Weekly` in the **Repeat** field.
Every first	The day of the month on which you want to synchronize, for example, Tuesday. This field is only visible if you select `Monthly` in the **Repeat** field.
At	The time at which you want to synchronize automatically, for example, 14:00. ◦ You can only schedule on the hour. For example, you can add a synchronization schedule at 8:00, but not at 8:45. If you try to add it at 8:45, we will default it to 8:00. Use a cron expression if you don't want to schedule on the hour. ◦ This field is only visible if you select `Daily`, `Weekly`, or `Monthly` in the **Repeat** field.
Time zone	The time zone for the schedule.

5. Click **Save**.

Remove an S3 synchronization schedule

You can remove a synchronization schedule from an Amazon S3 File System asset to stop automatically synchronizing Amazon S3.

Prerequisites

- You have a resource role with the Configure external system resource permission on the community or domain that contains the S3 File System, for example Owner.

- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.
- You have a programmatic AWS user and IAM role with the required permissions.
- You have created one or more crawlers.
- You have a role with the following resource permissions on the S3 community you created when you registered an Amazon S3 file system:
  - Asset: add
  - Attribute: add
  - Domain: add
  - Attachment: add

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Synchronization schedule** section, click **Remove Schedule**.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

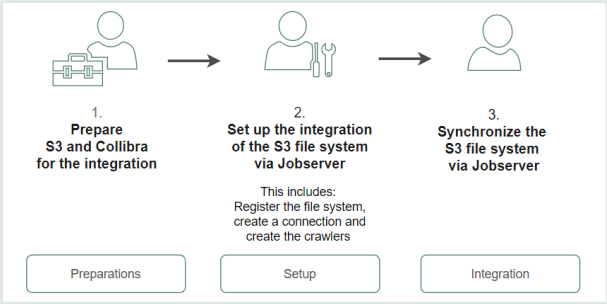
Integrating an Amazon S3 file system via Jobserver

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Integrate an Amazon S3 file system via Jobserver

The Amazon S3 file system integration allows for the registration of an Amazon S3 file system as a data source and synchronization of Amazon S3 metadata in Collibra, representing the full Amazon S3 file structure in Collibra.

Follow the steps below to integrate an Amazon S3 file system via Jobserver.



> Tip You can also follow a training and watch videos via Collibra University.

	Step	What?	Description	Results
Preparations	1	Prepare the Amazon S3 file system for integration via Jobserver	Prepares the S3 file system for integration in Data Catalog.	You have access keys that you can use during the integration.
	2	Restrict AWS regions	Makes sure the regions to collect data from are known.	Collibra knows which regions to look at.

	Step	What?	Description	Results
Setup	3	Register the Amazon S3 file system as a data source	Creates an initial structure for the integration.	A Storage Catalog domain and S3 File System asset become available in the selected parent community.
	4	Connect to Amazon S3	Sets up the connection to Amazon S3.	The connection is available.
	5	Create crawlers	Creates crawlers to find and ingest the data of Amazon S3.	The crawlers to collect metadata from Amazon S3 are available.
Integration	6	Synchronize Amazon S3	Runs the crawlers to ingest the metadata of Amazon S3.	The metadata of Amazon S3 is available in Collibra. By default, the assets are shown in a plain list, but you can create a hierarchy to show it in a tree structure.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Preparing S3

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Prepare S3 file system for Jobserver

Before you integrate an S3 file system via Jobserver, you need to prepare S3 for the integration. You need to:

- Create a custom policy and a programmatic user
  As a result, you will receive access keys that you have to use during the integration or registration.
- Create an Identity and Access Management role. This is the role that will be used by the crawlers.

Password encryption

Collibra's integration of Amazon S3 does not use a separate encryption services, but reuses the Collibra DGC core service encryption method. This method uses the AES/CBC/PKCS5Padding transformation to encrypt your passwords when you connect to Amazon S3.

Required Amazon Web Services (AWS)

Collibra relies on **AWS Glue** and **AWS Identity and Access Management** to ingest and synchronize data.

AWS Glue

AWS Glue is an Amazon cloud service to perform extract-transform-load (ETL) processes on data, stored in data sources such as Amazon S3.
AWS Glue has the following components:

- Glue crawlers:
  Glue crawlers analyze and describe a wide range of data sources such as Amazon S3 or MySQL. However, Data Catalog only uses them for the Amazon S3 file system integration.
- Glue database:
  Glue crawlers store their results in a database in the form of tables and columns. Both the tables and columns in the Glue database contain metadata that describes

the content of Amazon S3. Data Catalog reads those databases for data ingestion. The name of the created Glue database is *collibra_catalog_<S3 File System-ID>_<Domain-ID>*.

- ETL processes:
  The ETL processes can extract data from a data source, process that data, for example, categorize and clean it and produce output. This component is currently not used by Data Catalog.

Though you need an AWS account, you do not have to work in AWS Glue directly because Collibra does everything for you. For more information about AWS Glue, see the AWS Glue documentation.

> Note Collibra only uses AWS Glue to ingest data from Amazon S3. All other features, such as crawling other data sources or ETL processes are not integrated.

AWS Identity and Access Management

Collibra uses the AWS Identity and Access Management (IAM) service to manage access to Amazon S3 and AWS Glue. Similar to AWS Glue, you need an AWS account to use the IAM service, but after setting up the required users and roles, you do not have to work directly with IAM. For more information about IAM, see the IAM documentation.

You need two things in IAM:

- An AWS programmatic user to access Amazon S3 and AWS Glue.
- An IAM role for the crawlers.

Programmatic user

Collibra needs programmatic access to Amazon S3 and AWS Glue by means of a user. The following policies and permissions are required:

- AWSGlueServiceRole (AWS managed policy)
  If you don't want to use this out-of-the-box AWS managed policy, you will need to work with AWS support to define a more restrictive policy.
- pass_role (inline policy)
  You can use the following JSON content:

```
{
    "Version": "2012-10-17",
    "Statement":
    [
        {
            "Sid": "VisualEditor0",
            "Effect": "Allow",
            "Action": "iam:PassRole",
            "Resource": "*"
        }
    ]
}
```

For more information about creating a user with programmatic access, see the IAM documentation.

IAM role

AWS Glue Crawlers need an IAM role to allow the crawlers to execute an operation on your behalf. The "pass_role" permission policy of the programmatic user is used to assign this role to the crawler.

You need at least the following parameters:

- Trusted entities: glue.amazonaws.com
- Policies:
  - AmazonS3ReadOnlyAccess (AWS managed policy, required when you need to access a private S3 bucket.)
  - AWSGlueServiceRole (AWS managed policy)

> Note You can provide more restrictive permissions to the IAM role, if dictated by your security requirements. Your AWS subject matter expert can create the appropriate permission set using the steps in the IAM documentation. We recommend that you test a crawler with an IAM role that has these permissions in the AWS console, to ensure that it is successful before you use the IAM role in Collibra.

You can also use the IAM role for role-based access control, to authenticate to Amazon AWS without manually entering a user ID and secret access key.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Configure role-based Amazon S3 access control for Jobserver

When you register an Amazon S3 file system, you can authenticate to Amazon S3 based on an IAM role. As a result, you can connect to Amazon S3 without an access key ID and secret access key.

Prerequisites

- You have access to the AWS IAM console.
- You have access to the Amazon EC2 console.
- You have an Amazon EC2 instance.

Steps

1. In AWS Identity and Access Management, do the following:
   a. Create a new IAM role or select an existing IAM role.
   b. Attach the following policies to the IAM role:
      - AWSGlueServiceRole (AWS managed policy)
      - pass_role (inline policy)
        You can use the following JSON content:

```
{
 "Version": "2012-10-17",
 "Statement":
 [
  {
   "Sid": "VisualEditor0",
   "Effect": "Allow",
   "Action": "iam:PassRole",
   "Resource": "*"
  }
```

```
  ]
}
```

2. In the Amazon EC2 console, attach the IAM role to the Amazon EC2 instance.
3. Install the Jobserver service on the Amazon EC2 instance node.
    - Linux
    - Windows

More information

If the credentials in the Amazon EC2 instance can't be used to authenticate, you can create a credentials file and save it in the **user_home/.aws/** folder. The credentials file should look like this:

```
[default]
aws_access_key_id = <access key ID>
aws_secret_access_key = <secret access key>
```

For more information, see the AWS developer guide.

> Warning Do not use a credentials file unless absolutely necessary.

What's next?

You can now connect to Amazon S3 via the jobserver service on the Amazon EC2 instance node.

Restrict AWS regions

You can restrict the AWS regions to which Collibra Data Catalog can connect.

> Note When there is no restriction, the S3 integration will make requests to all possible AWS regions, which could result in long synchronization times.

Prerequisites

- You have the **ADMIN** or **SUPER** role in Collibra Console.
- You have the **SUPER** role in Collibra Console.
- You have the **ADMIN** or **SUPER** role in Collibra Console.

Steps

1. Open the DGC service settings for editing:
    a. Open Collibra Console.
        » Collibra Console opens with the **Infrastructure** page.
    b. In the tab pane, expand an environment to show its services.
    c. In the tab pane, click the Data Governance Center service of that environment.
    d. Click **Configuration**.
    e. Click **Edit configuration**.
2. Open the DGC service settings for editing:
    a. Open Collibra Console.
        » Collibra Console opens with the **Infrastructure** page.
    b. In the tab pane, expand an environment to show its services.
    c. In the tab pane, click the Data Governance Center service of that environment.
    d. Click **Configuration**.
    e. Click **Edit configuration**.

3. In the **Register data source** section, enter the required information:
4.

Setting	Description
AWS regions restriction	A list of AWS regions Data Catalog is allowed to connect to. For example, *eu-west-3* and *us-east-2*. For a list of all AWS locations, see the AWS documentation. ◦ If you want to allow Collibra to make a connection to any AWS region, leave the field empty. ◦ If you remove a region from this list and the region was previously used for an S3 integration, you may want to delete the Glue database from the previously used region manually. By default, Collibra does not remove it. The Glue database has the following naming convention: collibra_catalog_<Asset Id>_<Domain Id> For example: collibra_catalog_d3174a88-5ffe-4d50-8fbe-7bf0832ec3af_5d198ce9-4e56-4d0e-a885-58204da50741 ◦ When using Edge, a warning is added to the logs if an invalid region is detected in the restricted regions list.

5. Click **Save all**.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Register an Amazon S3 file system

To integrate an Amazon S3 file system, you register the Amazon S3 file system in Data Catalog to create a S3 File System asset .
The newly created S3 file system asset does not automatically connect to Amazon S3. You create a connection manually in the S3 File System asset.

Prerequisites

- You have a resource role with the **Configure external system** resource permission, for example, Owner.

- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a role with the following resource permissions on the S3 community you create when you registered an Amazon S3 file system:
  - Asset: add
  - Attribute: add
  - Domain: add
  - Attachment: add

Steps

1. On the main menu, click ⠿, and then click **Catalog**.
   » The Catalog Home opens.
2. On the main toolbar, click +.
   » The **Create** dialog box appears.
3. In the **Create** dialog box, click **Register system**.
   » The **Register system** page appears.
4. In the **Register system** page, click **Amazon S3**.
   » The **Register Amazon S3 file system** dialog box appears.
5. Enter the required information.

Field	Description
Community	The parent community in which the initial Amazon S3 structure will be created.
File system name	The name for the S3 file system asset.
Description	The description to provide extra information about the file system. This is used as the Description attribute of the S3 File System asset.
Owner	The owner name of the data in the created community.

6. Click **Register**.
   » An S3 File System asset is created.
   » A Storage Catalog domain is created with the same name as the S3 File System asset.
   » The configuration page of the S3 File System asset is automatically opened.

What's next?

You can now connect to Amazon S3.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Connect a file system asset to Amazon S3 via Jobserver

To retrieve data from Amazon S3, you have to connect via an S3 File System asset. You always have to do that after registering a new Amazon S3 File System. You can also edit the settings afterwards, for example, if you want to use another Jobserver than the one you originally selected.

Prerequisites

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- You have a programmatic AWS user and IAM role with the required permissions.

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Connection details** section, click **Edit connection details**.

4. Enter the required information.

Field	Description
Connect via	The Jobserver used for synchronizing.
Access key ID	The access key ID of the programmatic AWS user.
Secret access key	The secret access key of the programmatic AWS user.
IAM role	The IAM role to be assigned to the crawlers.

5. Click **Save**.

What's next?

You can now create crawlers.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Managing crawlers

A crawler is an automated script that ingests data from Amazon S3 to Data Catalog.

You can create, edit and delete crawlers in Collibra Platform Self-Hosted. When you synchronize Amazon S3, the crawlers are created in AWS Glue and executed. Each crawler crawls a location in Amazon S3 based on its include path. You can make an S3 bucket accessible for crawlers from the same or other AWS accounts than the account in which the S3 bucket is located. The results are stored in one AWS Glue database per domain assigned to one or more crawlers. Those databases are ingested in Data Catalog in the form of assets, attributes and relations. The databases are stored in AWS Glue until the next synchronization. At that moment, they are deleted and re-created. The crawlers in AWS Glue are deleted immediately after as the synchronization is finished.

Important

- If you completed the **Glue database configuration** parameter in the capability, you need to create a dummy crawler. A dummy crawler is a crawler with an invalid include path, such as s3://dummy. This crawler won't be taken into account when you run the synchronization.
  In a future release, we'll remove the need for a dummy crawler.
- By default, AWS Glue allows up to 25 crawlers per account. For more information, see the AWS Glue documentation. This has consequences for Collibra:
  - If you created crawlers in AWS Glue directly, Collibra can create less crawlers for synchronization.
  - Because Collibra creates the crawlers in AWS Glue during synchronization, you should avoid having 25 or more crawlers in one S3 File System asset.
  - You can synchronize several S3 File System assets simultaneously, but if the total number of crawlers exceeds the maximum amount in AWS Glue, synchronization will fail. Since Collibra deletes the crawlers from AWS Glue after synchronization, it is safer to synchronize each S3 File System asset at a unique time.
- Crawlers in AWS Glue can crawl multiple buckets, but in Collibra, each crawler can only crawl a single bucket.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Create a crawler

You can create a crawler for an S3 File System asset in Data Catalog.

Prerequisites

For Jobserver

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- You have connected an S3 File System asset to Amazon S3.

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, click **Create crawler**.

» The **Create crawler** dialog box appears.

4. Enter the required information.

Field	Description
Domain	The domain in which the assets of the S3 file system are created. More information about linking domains to crawlers: ◦ A specific Storage Catalog domain is created automatically when the S3 File System asset is created. That domain is selected by default. However, you can manually create a new Storage Catalog domain and select it. ◦ If multiple crawlers point to the same domain, then all assets are created in the same domain. ◦ If multiple crawlers point to different domains, then all assets are created in their respective domains. ◦ If multiple crawlers from the same S3 File System asset overlap and point to different domains, then overlapping assets are created in each domain. ◦ If multiple crawlers from the same S3 File System asset overlap and point to the same domain, then overlapping assets are created once in that domain. ◦ If crawlers from multiple S3 File System assets overlap and point to different domains, then overlapping assets are created in each domain. ◦ If crawlers from multiple S3 File System assets overlap and point to the same domain, then overlapping assets are created once in the domain and the S3 Bucket asset has a relation to both S3 File System assets.

Field	Description
Name	The name of the crawler in Collibra. More information about crawler names: ◦ You cannot use the same name for two crawlers in the same S3 File System asset. ◦ The name of the corresponding crawler in AWS Glue will contain this name. Its name will follow the following convention: `collibra_catalog_<s3fs asset id>_<name_of_the_crawler_in_Collibra>`. ◦ The crawler name must be compliant with the AWS Glue limitations: ▪ It has to match the single-line string pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\t]*`. ▪ The length should be between 1 and 255 bytes long, including the fixed prefix that Collibra adds. That means that you can use roughly 65 characters, depending on the characters that were used. Warning This restriction is imposed by Amazon S3, which allows up to 255 bytes, including the prefix added by Collibra. If you enter too many characters and exceed the byte limit, synchronization fails.
Include path	The case-sensitive path to a directory of a bucket in Amazon S3. All objects and subdirectories of this path are crawled. For more information and examples, see the AWS Glue documentation.
Exclude patterns	Glob pattern that represents the objects that are in the include path, but that you want to exclude. For more information and examples, see the AWS Glue documentation.
Add pattern	Button to add additional exclude patterns.

5. Click **Create**.

What's next?

You can now synchronize Amazon S3 manually or define a synchronization schedule.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Edit a crawler

You can edit a crawler of an S3 File System asset in Data Catalog. For example, you can do this if you want to edit the exclude pattern.

Prerequisites

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- You have connected an S3 File System asset to Amazon S3.

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, in the row of the crawler that you want to edit, click .

» The **Edit crawler** window appears.

4. Enter the required information.

Field	Description
Domain	The domain in which the assets of the S3 file system are created. More information about linking domains to crawlers: ◦ A specific Storage Catalog domain is created automatically when the S3 File System asset is created. That domain is selected by default. However, you can manually create a new Storage Catalog domain and select it. ◦ If multiple crawlers point to the same domain, then all assets are created in the same domain. ◦ If multiple crawlers point to different domains, then all assets are created in their respective domains. ◦ If multiple crawlers from the same S3 File System asset overlap and point to different domains, then overlapping assets are created in each domain. ◦ If multiple crawlers from the same S3 File System asset overlap and point to the same domain, then overlapping assets are created once in that domain. ◦ If crawlers from multiple S3 File System assets overlap and point to different domains, then overlapping assets are created in each domain. ◦ If crawlers from multiple S3 File System assets overlap and point to the same domain, then overlapping assets are created once in the domain and the S3 Bucket asset has a relation to both S3 File System assets.

Field	Description
Name	The name of the crawler in Collibra. More information about crawler names: ◦ You cannot use the same name for two crawlers in the same S3 File System asset. ◦ The name of the corresponding crawler in AWS Glue will contain this name. Its name will follow the following convention: `collibra_catalog_<s3fs asset id>_<name_of_the_crawler_in_Collibra>`. ◦ The crawler name must be compliant with the AWS Glue limitations: ▪ It has to match the single-line string pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\t]*`. ▪ The length should be between 1 and 255 bytes long, including the fixed prefix that Collibra adds. That means that you can use roughly 65 characters, depending on the characters that were used. Warning This restriction is imposed by Amazon S3, which allows up to 255 bytes, including the prefix added by Collibra. If you enter too many characters and exceed the byte limit, synchronization fails.
Include path	The case-sensitive path to a directory of a bucket in Amazon S3. All objects and subdirectories of this path are crawled. For more information and examples, see the AWS Glue documentation.
Exclude patterns	Glob pattern that represents the objects that are in the include path, but that you want to exclude. For more information and examples, see the AWS Glue documentation.
Add pattern	Button to add additional exclude patterns.

5. Click **Save**.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.

For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Delete a crawler

You can delete a crawler from an S3 File System asset.

> Note If you delete an S3 File System asset that contains one or more crawlers, the crawlers are also deleted.

Prerequisites

- You have a resource role with the **Configure external system** resource permission, for example, Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- You have connected an S3 File System asset to Amazon S3.

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, in the row of the crawler that you want to delete, click .
   » The **Delete Crawler** confirmation message appears.
4. Click **Delete crawler**.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.

For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

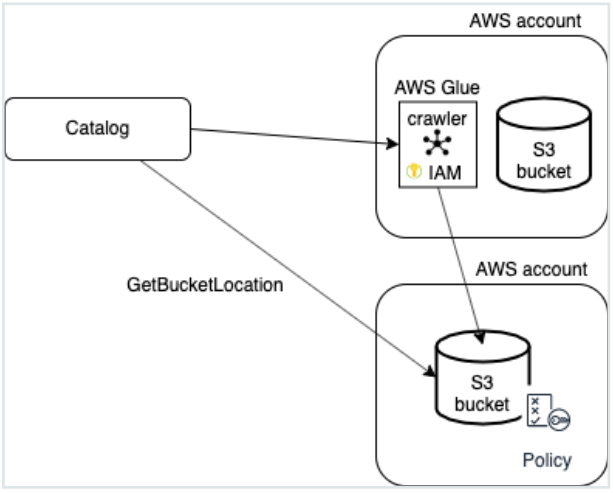
Cross-account crawling

If you use Jobserver, you can make an S3 bucket accessible for crawlers from other AWS accounts than the account in which the S3 bucket is located. To access the external S3 bucket, the programmatic user and the IAM crawling role must be defined in the AWS main account.

Policy

A policy must be attached to the external S3 bucket to allow:

- the AWS Glue crawler to access and perform S3 actions on an external S3 bucket from another AWS account.
- Data Catalogto execute the S3 GetBucketLocation API on an external S3 bucket via the programmatic user.



You can use the following JSON content:

```
{
  "Version": "2012-10-17",
  "Statement": [
```

```
    {
        "Sid": "collibra-jobserver-access",
        "Effect": "Allow",
        "Principal": {
            "AWS": "arn:aws:iam::<enter_id>:role/collibra-job-
server-s3-role"
        },
        "Action": "s3:*",
        "Resource": [
            "arn:aws:s3:::crawler-name",
            "arn:aws:s3:::crawler-name/*"
        ]
    },
    {
        "Sid": "collibra-jobserver-access",
        "Effect": "Allow",
        "Principal": {
            "AWS": "arn:aws:iam::<enter_id>:user/collibra-job-
server"
        },
        "Action": "s3:getBucketLocation",
        "Resource": [
            "arn:aws:s3:::*"
        ]
    }
  ]
}
```

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Synchronizing Amazon S3

When you synchronize Amazon S3, the content of your Amazon S3 repository is analyzed and represented by means of assets and their characteristics.
You can synchronize manually, or you can automate it by adding a synchronization schedule by means of a cron expression.

- You can only synchronize one S3 File System at a time. If a synchronization job is in progress and a second one is triggered, manually or automatically, it will be queued.

- If a synchronization job is still running and a new synchronization of the same S3 File System is triggered (manually or automatically), the running synchronization will continue and the new synchronization request is ignored.

Technically, the synchronization happens in several steps:

1. Collibra creates crawlers in AWS Glue, based on the crawlers defined in Collibra.
2. If AWS Glue contains databases with metadata from a previous synchronization, the databases are deleted.
3. Each AWS Glue crawler crawls a location in Amazon S3 based on its include path. For each domain assigned to one or more crawlers, AWS Glue creates a database with the crawling results.
4. Collibra ingests those databases and creates assets, attributes and relations as required to match the metadata.
   The resulting assets are in the domain that was specified in the crawler.

   > Warning Do not move the assets to another domain. Doing so may lead to errors during future synchronizations. This is a known limitation.

5. The AWS Glue crawlers are deleted.

Naming convention

Synchronizing Amazon S3 relies on a naming convention to match assets during the synchronization process. We highly recommend that you not change the S3 File System asset's full name.

> Warning Editing full name of the S3 File System assets may lead to errors during the synchronization process.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Synchronize Amazon S3 manually

You can manually start a synchronization job of an S3 File System asset. This can be useful if you want to test your crawlers, or if you want to synchronize immediately.

> Tip You can also add a synchronization schedule to synchronize automatically.

Prerequisites

- You have registered an Amazon S3 file system.
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- You have a programmatic AWS user and IAM role with the required permissions.
- You have connected an S3 File System asset to Amazon S3.
- You have created one or more crawlers.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a resource role with the Configure external system resource permission on the community or domain that contains the S3 File System, for example Owner.
- You have a role with the following resource permissions on the S3 community you created when you registered an Amazon S3 file system:
  - Asset: add
  - Attribute: add
  - Domain: add
  - Attachment: add

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Crawlers** section, click **Synchronize now**.
   » A notification indicates synchronization has started.

   » The synchronization job appears in the **Activities** list as a bulk synchronization.

» The **Synchronization schedule** section displays the time of the last synchronization.

» Once the synchronization is completed, you can view a summary of the results from the **Activities** list and you can view the assets in their domain. For more information, go to Integrated Amazon S3 data.

Note In case of a partial synchronization caused by a temporary communication issue, the status of the assets that cannot be synchronized is set to **Missing from source**. During the next fully successful synchronization, the assets are removed or their previous status is restored, depending on their actual status in the source system.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Add an S3 synchronization schedule

To keep the content of Collibra Platform Self-Hosted synchronized with your Amazon S3 File System, you can synchronize manually or create a schedule to automatically do this with a fixed interval.

Note You can only create one synchronization schedule.

Prerequisites

- You have a resource role with the Configure external system resource permission on the community or domain that contains the S3 File System, for example Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.
- You have a programmatic AWS user and IAM role with the required permissions.

- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- You have connected an S3 File System asset to Amazon S3.
- You have created one or more crawlers.
- You have a role with the following resource permissions on the S3 community you created when you registered an Amazon S3 file system:
  - Asset: add
  - Attribute: add
  - Domain: add
  - Attachment: add

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Synchronization schedule** section, click **Add Schedule**.

4. Enter the required information.

Field	Description
Repeat	The interval when you want to synchronize automatically. The possible values are: **Daily**, **Weekly**, **Monthly**, and **Cron expression**.
Cron	The Quartz Cron expression that determines when the synchronization takes place. This field is only visible if you select `Cron expression` in the **Repeat** field.
Every	The day on which you want to synchronize, for example, Sunday. This field is only visible if you select `Weekly` in the **Repeat** field.
Every first	The day of the month on which you want to synchronize, for example, Tuesday. This field is only visible if you select `Monthly` in the **Repeat** field.
At	The time at which you want to synchronize automatically, for example, 14:00. ◦ You can only schedule on the hour. For example, you can add a synchronization schedule at 8:00, but not at 8:45. If you try to add it at 8:45, we will default it to 8:00. Use a cron expression if you don't want to schedule on the hour. ◦ This field is only visible if you select `Daily`, `Weekly`, or `Monthly` in the **Repeat** field.
Time zone	The time zone for the schedule.

5. Click **Save**.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Edit an S3 synchronization schedule

You can edit the synchronization schedule of an Amazon S3 File System asset. For example, you can do this if you think the synchronization job runs too often or not often

enough.

Prerequisites

- You have a resource role with the Configure external system resource permission on the community or domain that contains the S3 File System, for example Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.
- You have a programmatic AWS user and IAM role with the required permissions.
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- You have connected an S3 File System asset to Amazon S3.
- You have created one or more crawlers.
- You have a role with the following resource permissions on the S3 community you created when you registered an Amazon S3 file system:
  - Asset: add
  - Attribute: add
  - Domain: add
  - Attachment: add

Steps

1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Synchronization schedule** section, click **Edit Schedule**.

4. Enter the required information.

Field	Description
Repeat	The interval when you want to synchronize automatically. The possible values are: **Daily**, **Weekly**, **Monthly**, and **Cron expression**.
Cron	The Quartz Cron expression that determines when the synchronization takes place. This field is only visible if you select `Cron expression` in the **Repeat** field.
Every	The day on which you want to synchronize, for example, Sunday. This field is only visible if you select `Weekly` in the **Repeat** field.
Every first	The day of the month on which you want to synchronize, for example, Tuesday. This field is only visible if you select `Monthly` in the **Repeat** field.
At	The time at which you want to synchronize automatically, for example, 14:00. ◦ You can only schedule on the hour. For example, you can add a synchronization schedule at 8:00, but not at 8:45. If you try to add it at 8:45, we will default it to 8:00. Use a cron expression if you don't want to schedule on the hour. ◦ This field is only visible if you select `Daily`, `Weekly`, or `Monthly` in the **Repeat** field.
Time zone	The time zone for the schedule.

5. Click **Save**.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on the integration of S3 via Edge, go to Integrating an Amazon S3 file system via Edge.

Remove an S3 synchronization schedule

You can remove a synchronization schedule from an Amazon S3 File System asset to stop automatically synchronizing Amazon S3.

Prerequisites

- You have a resource role with the Configure external system resource permission on the community or domain that contains the S3 File System, for example Owner.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have registered an Amazon S3 file system.
- You have a programmatic AWS user and IAM role with the required permissions.
- You have configured one or more Jobservers in Collibra Console. If there is no available Jobserver, the **Register data source** actions will be grayed out in the global create menu of Collibra Platform Self-Hosted.
- You have connected an S3 File System asset to Amazon S3.
- You have created one or more crawlers.
- You have a role with the following resource permissions on the S3 community you created when you registered an Amazon S3 file system:
  - Asset: add
  - Attribute: add
  - Domain: add
  - Attachment: add

Steps

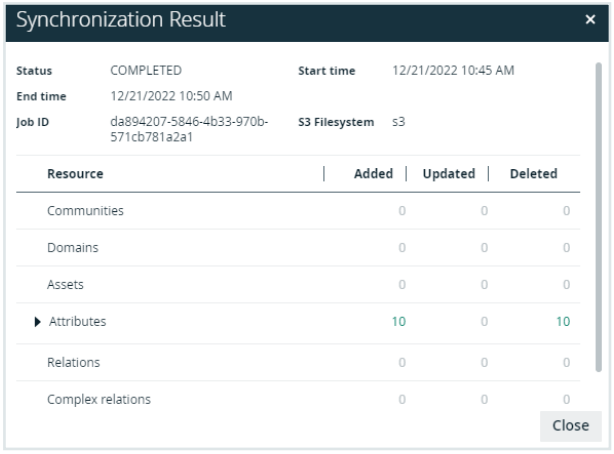
1. Open an S3 File System asset page.
2. In the tab pane, click **Configuration**.
3. In the **Synchronization schedule** section, click **Remove Schedule**.

View the summary of an Amazon S3 synchronization

After you synchronized Amazon S3, you can view the summary of the results. This shows the impact of the synchronization on the assets in Collibra Platform Self-Hosted

Steps

1. Open the Activities list.
2. In the row containing the S3 synchronization job, click **Result**.
   » The **Synchronization Result** dialog box appears.



Note
- The dialog box contains information about the total number of resources that were added, modified or removed as a result of the synchronization.
- In case of an error, the dialog box contains additional information about the error.

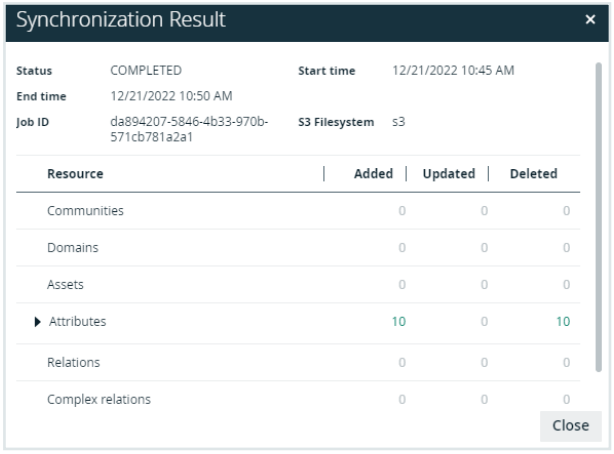
Tip The **Job ID** is useful when troubleshooting your synchronization process with Collibra Support.

View the summary of an Amazon S3 synchronization

After you synchronized Amazon S3, you can view the summary of the results. This shows the impact of the synchronization on the assets in Collibra Platform Self-Hosted

Steps

1. Open the Activities list.
2. In the row containing the S3 synchronization job, click **Result**.
   » The **S3 synchronization results** dialog box appears.



Note
- The **Ingestion Details** section contains information about the total number of resources that were added, modified or removed as a result of the synchronization.
- In case of an error, the **Ingestion Details** section contains additional information about the error.

Tip The **Job ID** is useful when troubleshooting your synchronization process with Collibra Support.

Integrated Amazon S3 data

After you have synchronized the data, the integration of the Amazon S3 file system is completed.

Synchronization results

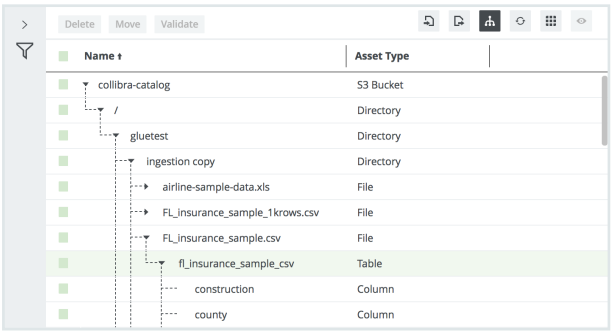
After synchronization, the resulting assets are in the domain that was specified in the crawler.

Warning Do not move the assets to another domain. Doing so may lead to errors during future synchronizations. This is a known limitation.

By default, the assets are shown in a plain list, but you can enable a multi-path hierarchy to show it in a tree structure. For the best result, we recommend that you use the following relations:

1. File Container contains File Container
2. Directory contains Directory
3. File container contains File
4. Directory contains File Group
5. File contains Table
6. File Group contains Table
7. Table contains Column

The following images shows the resulting hierarchical table.



> Note In case of a partial synchronization caused by a temporary communication issue, the status of the assets that cannot be synchronized is set to **Missing from source**. During the next fully successful synchronization, the assets are removed or their previous status is restored, depending on their actual status in the source system.

Synchronized metadata per asset type

This table shows the metadata for each Amazon S3 asset type.

Asset type	Synchronized metadata	Resource ID
S3 Bucket	URL	00000000-0000-0000-0000-000000000258
	Location	00000000-0000-0000-0000-000000000203
	File Storage contains/ is part of File Container	00000000-0000-0000-0001-002600000000
Directory	URL	00000000-0000-0000-0000-000000000258
	File Container contains/ is part of File Container	00000000-0000-0000-0001-002600000001
	Directory contains/ is part of Directory	00000000-0000-0000-0001-002600000003
File Group	URL	00000000-0000-0000-0000-000000000258
	File Type	00000000-0000-0000-0001-002500000012
	Document Size	00000000-0000-0000-0000-000000000259
	Number of Files	00000000-0000-0000-0001-002500000001
	Directory contains/ is part of File Group	00000000-0000-0000-0001-002600000004

Asset type	Synchronized metadata	Resource ID
File	URL	00000000-0000-0000-0000-000000000258
	File Type	00000000-0000-0000-0001-002500000012
	Document Size	00000000-0000-0000-0000-000000000259
	File Container contains/ contained in File	00000000-0000-0000-0000-000000007060
Table	Glue database name	00000000-0000-0000-0001-000500000066
	Glue table name	00000000-0000-0000-0001-000500000067
	File contains/ is part of Table	00000000-0000-0000-0001-002600000002
	File Group contains/ is part of Table	00000000-0000-0000-0001-002600000005
Column	Technical Data Type	00000000-0000-0000-0000-000000000219
	Column Position	00000000-0000-0000-0001-000500000020
	Column is part of/ contains Table	00000000-0000-0000-0000-000000007042

Delete an S3 File System asset

You can delete an integrated S3 File System asset from Collibra Platform Self-Hosted.

> Note
> - The crawlers of the S3 File System asset are deleted.
> - The assets that were created by the synchronization are not deleted.

Prerequisites

- You have registered an Amazon S3 file system.
- You have a global role with the Catalog global permission, for example, Catalog Author.
- You have a resource role with the Asset > Remove resource permission.

Steps

1. Open an S3 File System asset page.
2. On the view toolbar, click **Actions** → **Delete**.
   » The **Delete Confirmation** dialog box appears.

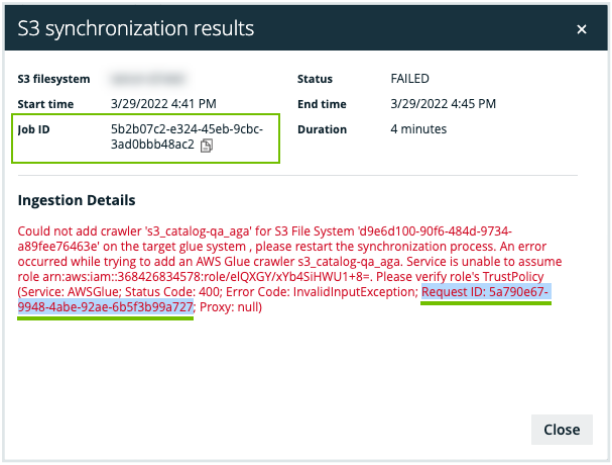
   > Tip If Catalog experience is disabled, the **More** menu is shown instead of **Actions**.

3. Click **Delete S3 File System**.

Troubleshooting: S3 file system integration

Where do I find the **Job ID** and **Request ID** for AWS troubleshooting?

The S3 synchronization results dialog box includes the Job ID. When an S3 synchronization fails, the results includes a detailed error message with the **Request ID**.



> Tip Share the **Request ID** with AWS support to understand why the specific request is failing in AWS. This is typically useful to troubleshoot IAM permission issues in your AWS environment.

Message **Could not add/change/delete crawler '<crawler name>' for S3 File System '<asset name>'.**

You can find more information about the actual problem in the Jobserver logs. The problem is usually described in the AWS SDK error message.

Cause	Description	Solution
Incorrect or too limited IAM permissions for the programmatic user defined in the connection details.	While connecting, the verification process only checks that the user can log in, but it doesn't verify permissions. Any further operation may therefore fail if the IAM permissions are wrong or too limited. This also applies to the AWS regions. Collibra checks the credentials in the default region, based on the region AWS SDK. Because the IAM service is global, that is sufficient in most cases. However, it is possible to put constraints on specific regions, including the AWS SDK default region.	Edit the IAM permissions or connect to Amazon S3 with another IAM user or role.

Cause	Description	Solution
Maximum number of crawlers in AWS Glue reached.	When you synchronize Amazon S3, Collibra creates crawlers in AWS Glue and executes them. After synchronization, they are deleted. By default, each AWS Glue account can only store 25 crawlers. This number can be reached easily, especially if the customer uses AWS Glue apart from Collibra.	• Delete one or more crawlers. • Create an advanced crawler by tweaking the include path and the exclude patterns. • Create additional S3 File System assets and divide the required crawlers between the assets. Then synchronize them at different times. • Synchronize different S3 File Systems at different times. • Ask Amazon support to increase that number. For more information, see the AWS Glue documentation.
Bucket does not exist	Typo in a bucket name - bucket doesn't exist.	Edit the crawler's include path to correct the bucket name.
No permission to access the bucket in Amazon S3.	This includes buckets that exist but belong to different accounts.	Request permission or delete the relevant crawler.
Unsupported AWS region.	S3 ingestion in Collibra Data Catalog relies on AWS Glue to analyze S3 buckets. However, AWS Glue is currently not supported in all AWS regions, which may lead to failing crawling creation. The log will display an UnknownHostException.	This is a built-in limitation of AWS Glue. For the list of supported regions for AWS Glue, see the AWS documentation.

Cause	Description	Solution
Incorrect AWS region.	AWS regions can be restricted so that S3 ingestion and synchronization in Collibra Data Catalog can only be performed in the regions your AWS account has access to. Example You will get an error message when: • A user with a European account tries to perform S3 ingestion in AWS region Canada. • A user with a European account tries to synchronize S3 buckets for AWS regions Europe and Canada. • A user with a Chinese and Canadian account tries to synchronize buckets for AWS regions Ireland and Canada.	This is a security measure. The AWS regions to which Collibra Data Catalog is allowed to connect can be restricted via Collibra Console.

Example `[2018-08-03 13:50:38,347] INFO .agent.SprayRoutesProvider [] [] - output: (500 Internal Server Error,{"messageCode":"s3_bucketDoesntExist","messageArguments":["qsdgqsbqfdscs"]})`

Message **Value not allowed. The connection details of the S3 File System are incorrect.**

Cause	Description	Solution
The credentials for the AWS user are incorrect.	This message appears when the credentials for the AWS user are incorrect. The access key ID and/or secret access key are wrong.	Pay attention that they do not contain trailing spaces.

Cause	Description	Solution
Your AWS account doesn't have access to an AWS region where the S3 bucket is located.	This message appears when you add an AWS region in Collibra Console to which your AWS account doesn't have access and then try to ingest an S3 file system.	Make sure that you have access to the AWS region where the S3 bucket is located.

Glue Crawler fails with an **Internal Service Exception** error message

This is an AWS Glue crawler error. For possible steps to resolve the issue, see the AWS documentation.

Glue Crawler failed and AWS logs show an **Internal server error** message

When checking the logs in Jobserver you may notice that one or more crawlers failed in AWS Glue. In that case, you need to open the AWS console and check the crawlers list in AWS Glue. Because crawlers are deleted from AWS Glue after ingestion, you will have to manually re-create the crawlers and run them again before proceeding. The failing crawler has a red exclamation mark and the Failed status. You can check the logs for more information.

Sometimes, the logged message just shows an "Internal server error". The only way to get more information is to contact the Amazon helpdesk. However, we noticed such errors often happen in the following situations

- The number of files to crawl is very large (> 100k)
- There is a series of very small files to crawl (>100).

In both cases, the problem is caused by AWS Glue. All Amazon services are protected against DDoS attacks and they throw throttling exceptions when too many operations are done in a specific time frame. Unfortunately this limit also applies between Amazon services. In this specific case, the AWS Glue database service is denying requests from the AWS Glue crawler service, which causes the crawling process to abort. Because this

is an inherent Amazon limitation, Collibra cannot fix this problem. A possible work-around is to use more S3 File System assets with more restricted include paths.

Error message **The AWS Access Key Id you provided does not exist in our records** though credentials are accepted

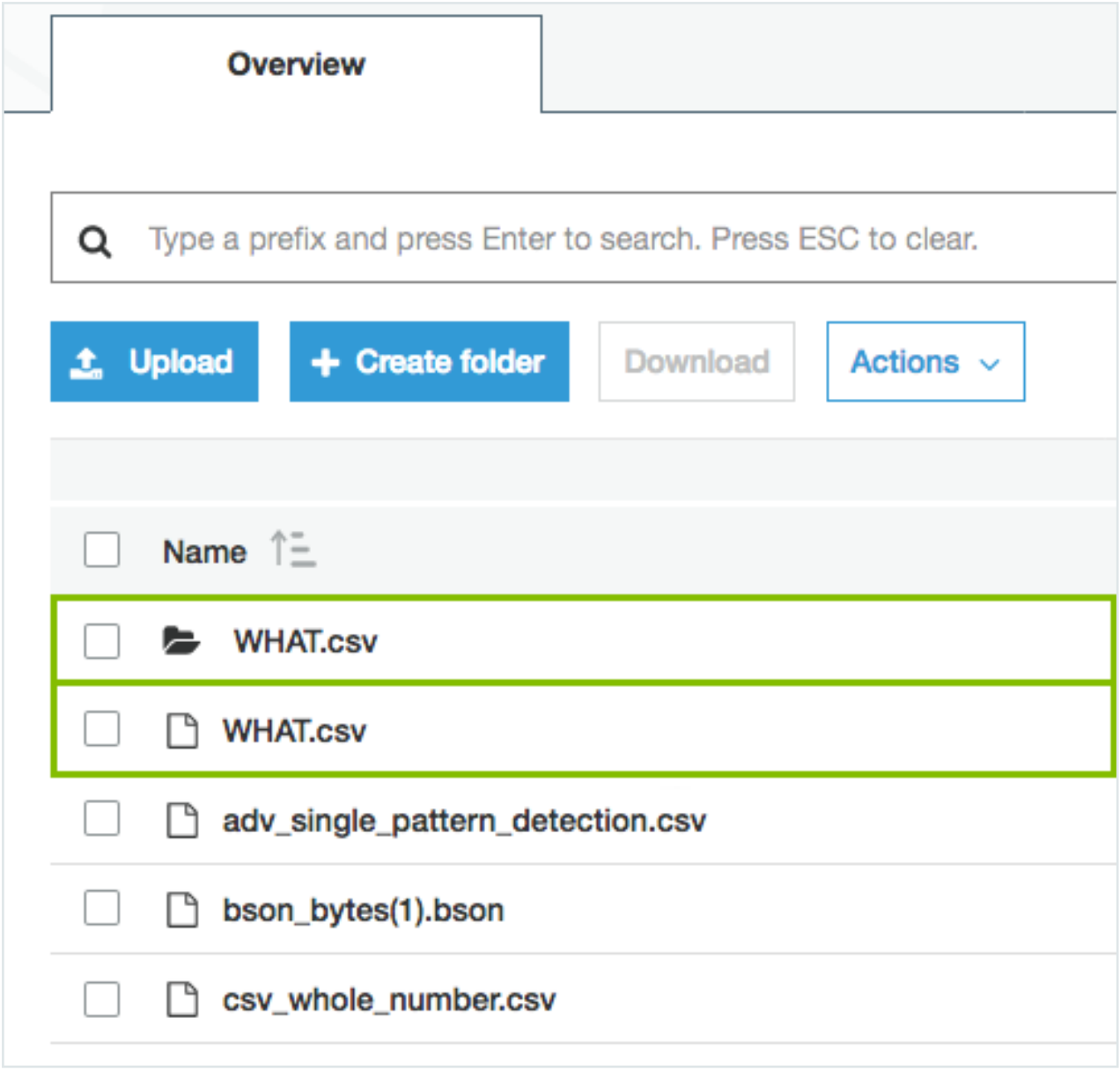
A user may be able to store S3 credentials in the S3 File System asset, though he cannot synchronize Amazon S3, create, edit or delete crawlers. The following message appears:

```
The AWS Access Key Id you provided does not exist in our
records.
(Service: Amazon S3; Status Code: 403; Error Code: Inval-
idAccessKeyId; ...
```

This may be caused by insufficient permissions on AWS Glue services. For more information, see About the Amazon S3 file system integration.

Synchronization fails when a directory contains a file and a directory with the same name (known issue)

In Amazon S3, you can use periods (.) in the name of a directory. As a consequence, you can give the directory a name that is identical to a file name, for example, Collibra.txt. However, if this happens, ingestion fails. This is a known issue.



Synchronizing an S3 File System fails with a **relationMaxLimitReachedTarget** message in the logs

This error comes from a broken relation in the assets tree. An asset created by S3 ingestion gets more than one parent asset. For example, a File asset has more than one parent directory or a Directory asset has more than one parent directory.

This typically happens when a user moves S3 assets to a different domain and then starts a synchronization. In that case, the ingestion jobs try to recreate the missing assets in the

original domain while old relations are still present. This can lead to an inconsistency in the relation tree.

We strongly recommend that you never move assets created by S3 ingestion to another domain.

> Example
> You work in domain called Amazon, which contains a Directory asset called Main.
> The Main Directory asset has a child asset of the File type, called Names.
> You move the Main Directory asset to another domain called Local.
> When you synchronize again, Data Catalog first recreates the Main Directory asset in the Amazon domain and then it updates the Names File asset.
> As a consequence, the Names File has 2 parent directories, which is a relation cardinality error.

Synchronizing Amazon S3 fails because you don't have the necessary permissions

In Collibra Data Intelligence Cloud 2020.11 and newer and Collibra Data Governance Center 5.7.7 and newer, Collibra checks the permissions of the AWS user when you synchronize Amazon S3. Synchronizing Amazon S3 fails if the AWS user does not have the necessary permissions.

A dialog box shows the following message:

```
Could not get/delete Glue database for S3 File System <name-of-
Amazon-S3-file-system>, please make sure you have all the neces-
sary permissions.
```

Solution: Give the AWS programmatic user the permission policy **AWSGlueServiceRole**. This is an AWS managed policy. If you don't want to use this out-of-the-box AWS managed policy, you will need to work with AWS support to define a more restrictive policy. An example of such policy is:

Example

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "VisualEditor0",
            "Effect": "Allow",
            "Action": [
                "glue:GetCrawler",
                "glue:GetCrawlers",
                "glue:DeleteDatabase",
                "glue:GetTables",
                "glue:DeleteCrawler",
                "glue:StopCrawler",
                "s3:ListBucket",
                "glue:GetDatabases",
                "glue:CreateCrawler",
                "glue:GetDatabase",
                "iam:PassRole",
                "glue:StartCrawler",
                "glue:BatchDeleteTable",
                "s3:GetBucketLocation"
            ],
            "Resource": "*"
        }
    ]
}
```

No assets are created after the synchronization job is completed

This is usually because AWS Glue didn't find any suitable files to process. A typical problem is a typo in the include path or exclude patterns. AWS Glue does not fail when an include path points to a directory that doesn't exist. Also, always verify there are no leading or trailing spaces in those fields.

Only part of the expected files or file groups are integrated

Jobs in Collibra can only succeed or fail. It's possible that some of the crawlers are correctly defined while others contain errors, such as a typo in an include path or an unsupported AWS region. In that case, the activity is marked as successful, though part of

it didn't succeed. Currently, the only way to confirm this is to read the log files of Collibra and Jobserver or Edge.

> Note When you start synchronization, the crawlers are created in AWS Glue. Once the crawlers are created, they are executed. If Collibra cannot create one or more crawlers, synchronization fails immediately. If the crawlers are created successfully, but fail later, synchronization only fails if all crawlers fail.

Partial ingestion or update of assets

It is possible to store a very large number of files in S3 buckets, hence leading to a large number of assets, attributes and relations to ingest into Data Catalog. To optimize memory and speed, the ingestion process is not transactional as a whole. It works with small transactional batches. If ingestion fails and aborts after some batches are already executed, it is possible that the ingested data is incomplete (if it is the first synchronization) or only partly updated (if it is not the first synchronization). In this case, it's advised to fix the problem and resynchronize as soon as possible.

Some of the folders and files in Amazon S3 are not visible in Collibra

You may notice that the content of your Amazon S3 does not always match the content in Collibra. Some folders from Amazon S3 may not appear in Collibra and some files are merged or split into different assets. This is not a bug in Collibra. When you synchronize Amazon S3, you create and execute crawlers in AWS Glue. Those crawlers create a table with metadata. That table is ingested in Collibra and is the basis for the relevant assets.

However, the crawlers in AWS Glue have some specific behavior to deal with partitioned tables. When the majority of schemas at a folder level are similar, the AWS Glue crawler creates partitions of a table instead of separate tables. Based on that information, the assets in Collibra are created.

See the AWS Glue documentation for more information about folders and tables in Amazon S3 and what happens when a crawler runs.

JSON ingestion shows partial value in technical data type attributes (known issue)

For security reasons, all values that contain information between < and > characters are automatically trimmed by Collibra. However, if JSON is ingested by AWS Glue, the technical data type attribute contains those characters to represent the JSON structure. As a consequence, the value is trimmed and thus invalid. In future releases of Collibra, several attribute types will be changed to the plain text kind to avoid this issue.

The file size or other property is not filled in for file xxx.yyy

AWS Glue only provides the file size for known file types, called "classifiers" in the AWS Glue terminology. Files that are classified as Unknown are registered but won't have any property associated. For the list of built-in classifiers, see the AWS Glue documentation.

The table name has a strange hash-code at the end

AWS Glue appends a hash code to differentiate two different files of the same name but different directories, for example, csv_boolean_csv_ fe8de80c6f9a2b31463801aa2778a427. This name, including the hash code, is actually transferred to Data Catalog.

A file is wrongly considered a File Group

AWS Glue preferably considers a directory as a data set when possible. This leads to a File Group being created in Data Catalog. There are multiple cases where it considers (possibly wrongly) one or more files as a File Group. Unfortunately, those rules are not clearly defined in AWS Glue documentation. Collibra noticed that AWS Glue considers a directory as a data set in the following cases:

- A directory only contains one file that belongs to a known classifier (file type).
- All files contained in a directory (including sub-directories) expose a similar schema (for example, all CSV files with columns of text type).

If you use Jobserver, experiment with include paths and exclude patterns of the crawlers. For example, if a crawler wrongly takes a directory with subdirectories as a single File Group, the official work-around is to add crawlers with the subdirectories as include paths.

Unfortunately, this work-around requires a lot of manual work and is limited by the maximum number of crawlers in AWS Glue (25 by default, but this can be expanded on request).

If you are using Edge, check to solution in the troubleshooting item: File Groups get the status Missing from Source

File Groups get the status **Missing from Source** after the S3 synchronization via Edge

File Group assets can receive the status "missing from source" if the behavior of the AWS crawler is not consistent, meaning AWS classifies files as File Group one day and classifies them as File on another day.
If this happens, File Group assets are created during the first synchronization but no longer exist after the second synchronization, resulting in the status "Missing from source".

Solution:
If you are using Edge, you can add custom parameter **file-group-as-file** to your S3 Edge capability. By adding the custom parameter, the S3 synchronization will always ingests File groups as File assets. The custom parameter is:

- Name: file-group-as-file
- Value: true

Problem setting up Lake Formation and S3 synchronization via Edge

You aren't able to see all S3 buckets when choosing a storage location for which to review, grant or revoke user permissions for Lake Formation. For more details, go to Prepare S3 file system for Edge.

Solution:
Provide extra cross region sharing on AWS side. For information, go to the Lake Formation documentation.

Message **Resource does not exist or requester is not authorized to access requested permissions** when setting up Lake Formation via Edge

When you're adding access permissions for specific storage locations, you receive the following message “Resource does not exist or requester is not authorized to access requested permissions“.

Solution:
Go to **AWS Lake Formation** → **Administration** → **Administrative roles and tasks** and add the IAM user as Data Lake administrator. For more details, go to Prepare S3 file system for Edge.

Registering an Amazon S3 file system via the AWS Glue JDBC connector

If you register an Amazon S3 file system via the AWS (Amazon Web Services) Glue JDBC connector, the resulting assets represent the columns and the tables in Amazon S3 without the folder context.
You can profile and classify the data, but the folder structure of your Amazon S3 environment isn't represented in Data Catalog.

> Note The AWS Glue JDBC connector leverages the Athena JDBC driver.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on the registration of S3 via Edge, go to Register an Amazon S3 file system via the AWS Glue JDBC connector and Edge.

Register an Amazon S3 file system via the AWS Glue JDBC connector and Jobserver

The Amazon S3 file system registration via the AWS Glue JDBC connector allows for the registration of an Amazon S3 file system as a data source and the synchronization of Amazon S3 metadata in Collibra, representing the S3 tables and columns in Collibra.

Follow the steps below to register an Amazon S3 file system via Jobserver.

Step	What?	Description	Results
1	Create an AWS Glue connector	Creates a AWS (Amazon Web Services) Glue JDBC connection needed to register the Amazon S3 file system as a data source. Note The AWS Glue JDBC connector leverages the Athena JDBC driver.	The AWS Glue JDBC connection is available.
2	Register the S3 file system as a data source via Jobserver	Makes metadata of the data source available.	A Physical Data Dictionary domain and new assets of the type Schema, Table and Column, corresponding to the data in your data source are available.
3	Refresh the registered data source	Updates the metadata and profiling of a registered data source.	

Catalog workflows

To keep the information flows in Data Catalog configurable, workflows are used. You can configure the out-of-the-box workflows. However, note that these workflows are designed to work together. If you change one workflow, ensure the other Data Catalog workflows are still functioning because they may depend on one another.

Name	Description
Assign Owner To Data Set	This process automates adding owners to data sets. This workflow is automatically triggered when a new Data Set asset is created.
Cancel Process	This process notifies the concerned users of a workflow cancellation.
Escalation Process	This process is the default mechanism for the escalation of user tasks in workflows.
Post Data Ingestion Workflow	This process facilitates assigning the Owner and Technical Steward for newly ingested Schema assets. This workflow is automatically triggered when a new Schema asset is created and after a data source is registered.
Propose New Business Asset	This process facilitates the creation of new Business Assets in the **Data Governance Council** community.
Propose New Data Asset	This process facilitates the creation of new Data Assets in the **Data Governance Council** community.
Propose New Technology Asset	The Propose New Technology Asset workflow allows you to create a new Technology asset in Collibra Platform Self-Hosted. By default, the asset is added to the **Data Governance Council** community, in the **New Applications** domain.

Name	Description
Request Assets Access	The Request Assets Access workflow allows you to request access to assets that are referenced in your Data Basket. All data owners have to approve the request before you can access the assets. Important • This workflow accepts by default only data sets that contain Column assets as data elements. • This workflow replaces the Request Data Sets Access workflow. However, if you restore a 5.4.x backup or older, the old Requests Data Sets Access workflow will overwrite the Request Assets Access workflow. In that case, you have to deploy the Requests Assets Access workflow again and apply all possible customizations. The workflow calculates the name of the asset by combining the creation date with a sequential number for that day, for example 2019-09-30 #1 and sets the asset characteristics according to the data submitted through the start form. The user who started the workflow receives the Requester role. The user with an Owner role approves the request for each data set and the Owner or Technical Steward provides access to the data set elements. For information on when this workflow is used, go to request access to data sets and reports. For more information on the workflow itself, go to the Collibra Developer Portal.
Simple Approval	The Simple Approval workflow is a single-step process that allows you to approve an asset in Collibra Platform Self-Hosted.
Voting Sub-Process	The Voting Sub-Process is a workflow that can be called by other workflows when users need to vote. It is used within other packaged workflows such as the Approval Process, the Simple Approval or the Issue Management workflow. You can use this sub-process in new custom workflows. The result is a true or false boolean that is provided to the parent workflow.

Tip For more information about these workflows or workflows in general, go to the Collibra Developer Portal.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Catalog Troubleshooting

If you are experiencing general issues with Data Catalog, consult the articles in this section.
For specific issues with one of the Data Catalog features, go to the dedicated section. For example:Collibra Data Lineage troubleshooting, Sample data, Google Cloud Storage, S3, and so on.

How to enable logging for data ingestion

If you want to troubleshoot issues with data ingestion, you have to enable logging for data ingestion.

By default, logging for data ingestion is disabled because your data can be exposed. For more information, go to Environment log settings for Repository services.

> Warning If you have investigated the data ingestion issue, don't forget to revert all the changes from this section.

Steps

1. Open the Data Governance Center logging settings.
   a. Open Collibra Console.
      » Collibra Console opens with the **Infrastructure** page.
   b. In the tab pane, click the **Data Governance Center** service of the environment whose log settings you need.
   c. Click **Logs**.
   d. Above the table, to the right, click **Settings**.
2. Click **Add logger**.
   » The **Add logger** dialog box appears.

3. Enter the required information.

Field	Description
Logger name	The name of the logger. Enter one of the following: ◦ `com.collibra.jobserver.client` ◦ `com.collibra.catalog.core.datausage.impl` ◦ `com.collibra.catalog.core.schema.impl` ◦ `com.collibra.catalog.core.schema.impl.ingestion` ◦ `com.collibra.catalog.core.schema.impl.profiling` ◦ `com.collibra.catalog.core.schema.impl.report`
Logger level	The amount of log entries you want in the logs. Select **DEBUG**.

4. Click **Add logger**.
5. Repeat this until you have added all the loggers.
   » If you create a diagnostics file for DGC files, the logs for the added packages will be included. No passwords nor user names are available in full in these logs.

What's next?

You can create and download a diagnostics file.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

The Jobserver logs are out of memory

When the Jobserver log files are out of memory, the logs that are created during ingestion or profiling are deleted immediately after they are created.

Solution

1. Stop the environment for which you want to update the memory settings.
2. Open a terminal session on the server that hosts the jobserver.
3. Open the **/opt/collibra/spark-jobserver/conf/logback.xml** file and do the following.
   a. In the logger section, update the properties to match the below information:

```
<logger name="com.collibra.jdbc" level="DEBUG"/>
<logger name="org.apache.spark" level="WARN"/>
<logger name="akka" level="WARN"/>
<logger name="com.collibra.catalog.profilers" level-
l="DEBUG"/>
<logger name="spark.jobserver.context" level="WARN"/>
<logger name="com.collibra.catalog.ingestion" level-
l="DEBUG"/>
<logger name="com.collibra.catalog.profiling.anonymization"
level="DEBUG"/>
<logger name="com.collibra.jobserver.job" level="DEBUG"/>
<logger name-
e="com.collibra.awsconnector.services.s3glue.impl" level-
l="DEBUG"/>
```

   b. Remove any other loggers.
   c. Save and close the file.
4. Start the environment again.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Ingestion out-of-memory error in Jobserver

When you upload a JDBC driver larger than 50 MB or when you have uploaded multiple JDBC drivers, you may encounter an out-of-memory error. Due to this problem, the jobserver does not release the memory needed to store the driver in memory.

Resolution

To solve this problem, you have to increase the memory of the Jobserver application, for example, increase it to 3 GB.

1. Stop the environment for which you want to update the memory settings.
2. Open a terminal session on the server that hosts the jobserver.
3. Open the file **<drive>/collibra/spark-jobserver/conf/jobserver.conf** for editing.
4. Look up the parameter **driver-memory**.
5. Edit the parameter value, for example, *3G*, corresponding with 3 GB.
   The default value is 2G.
6. Save and close the file.
7. Open the file **<drive>/collibra_data/spark-jobserver/config/server.json** for editing.
8. Look up the parameter **jobserverMemory**.
9. Edit the parameter value, for example, *2048M*, corresponding with 2 GB.
   The default value is 1024M.
10. Save and close the file.
11. Start the environment again.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Error when managing connection properties of a driver for Jobserver

Issue

When you want to change the properties of a connection used to register data sources via Jobserver, you receive the following error message:

```
CollibraIllegalStateException: jdbcDriverCannotBeUpdatedWhenLinked
when trying to delete or edit a JDBC driver.
```

or

```
You cannot update the driver because it is linked to a Schema Asset
```

Reason

Once you have successfully used a connection to register a data source via Jobserver, you cannot update the connection properties anymore.

Solution

If you want to change the properties for a driver, you need to create a new driver:

1. Open a schema registered via the driver you want to update.
2. Go to **Actions** → **Refresh**.
3. In **JDBC driver version**, select **Manage drivers...**.
4. Create a new driver for the data source.
   As a best practice for the name of the drivers, use a naming convention which includes the data source and the JDBC driver version number. For example: Google BigQuery 8257 or MySQL 8257. If you want to use the same driver version with other properties, add an extra number. For example: Google BigQuery 8257 v2.
   For details on the properties, see Manage Collibra-provided JDBC drivers.
5. Save the new driver.
   » The new driver is automatically applied to the schema.
6. For each schema that uses the old driver, go to **Actions** → **Refresh**, and select the new driver.

Missing schema name during data ingestion

If you ingest a data source with a new JDBC driver, you can receive an error "No schema has been specified".

> Note In the stacktrace you can see a "`CollibraIllegalArgumentException`" message.

Solution

Make sure that you defined a schema property for the new JDBC driver.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Different versions for Collibra and Jobserver

You can install the services of a Collibra Platform Self-Hosted environment on multiple nodes. If you do so, make sure that you use the same installer on all the nodes. This also applies to upgrading an environment.

If your environment has different versions for the Data Governance Center and Jobserver services, the following errors will occur when you run an ingestion.

- Spark Context's logs
  ```
  [2017-11-07 07:27:15,608] WARN nalRequestDataDeserializer []
  [akka://JobServer/user/jobManager-c7-8eec-de0c02029808] - Pack-
  age com.collibra.jobserver.dto.catalog.ingestion, different ver-
  sion detected: client uses version 1.2.4-SNAPSHOT, server uses
  version 1.2.2-SNAPSHOT
  ```
- Collibra logs
  ```
  20:21:43.407 [Procedure Manager] WARN
  ```

```
c.c.j.c.i.s.StateDeserializer - Package com.-
collibra.jobserver.dto.catalog.profiling, different version
detected: client uses version 1.1.10, server uses version 1.1.8
```

Solution

Install all the Collibra services with the same installer.

Error when refreshing a Schema registered via Jobserver

Issue

If you manually or automatically refresh a schema registered via Jobserver, you receive the following error:
`Server connection failed - java.lang.ClassNotFoundException:...`

Reason

This can happen for PostgreSQL, Oracle, and SQL Server data sources that are registered via Jobserver and **Register data source [use your own driver]**. The message means that the JAR file used for the connection is not available. The file was probably removed during the upgrade to 2022.11. For information on the reason, go to Removing outdated drivers during upgrade to 2022.11.

Solution

Go to Update a driver after the 2022.11 upgrade to solve the issue.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.

For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Resolve schema refresh conflicts via Jobserver

> Note This information only applies to Jobserver. For information on how Edge handles differences between the original schema and the updated schema, see About synchronizing schemas.

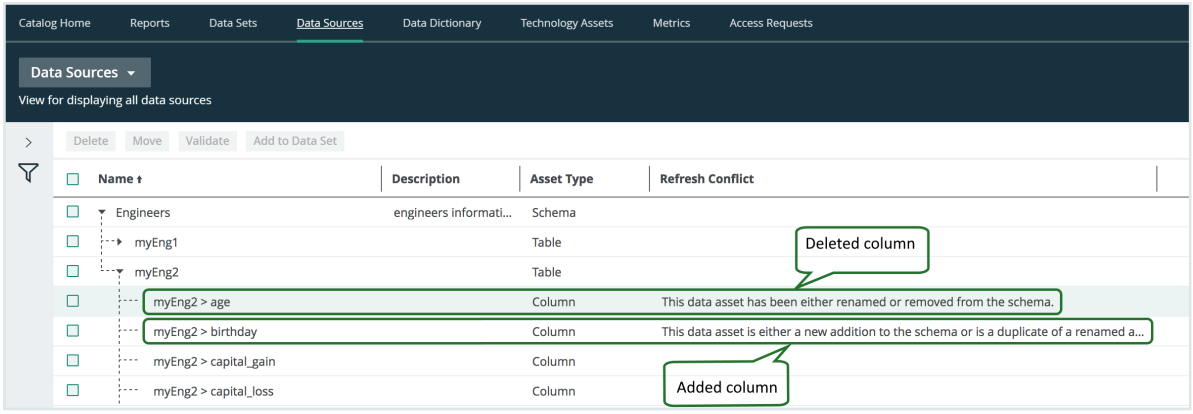
If you refresh a schema via Jobserver, the ingestion process detects differences between the original schema, already in Collibra Platform Self-Hosted, and the updated schema.

If columns or tables have been added to or removed from the schema, the process will create or delete the corresponding Column and Table assets in Collibra. However, the ingestion process results in a refresh conflict if one or more columns or tables were added and others were removed. If that happens, it adds a Refresh conflict attribute to all added and removed columns or tables. You have to resolve these conflicts before you can refresh the schema again. If you do not resolve the refresh conflicts, any future attempts to refresh the data source will fail.

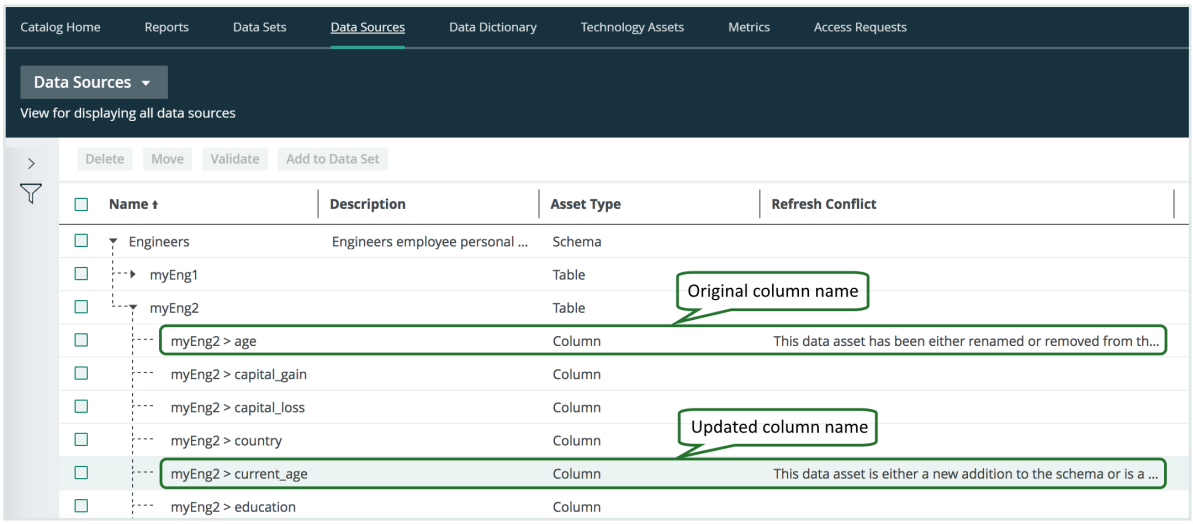
To see if there are any conflicts after a refresh, you have to add the **Refresh Conflict** field to the **Data Sources** view of the schemas.

You may come across the following scenarios:

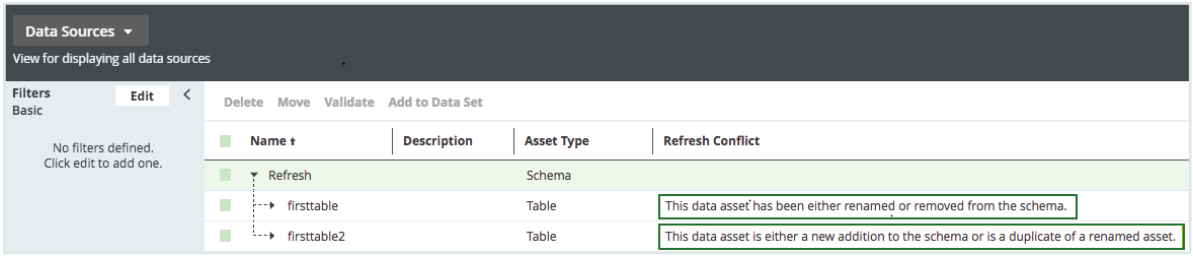
- A column is deleted from the schema and another one is added to the schema:
  a. You have to manually delete the column asset.
  b. You have to remove the **Refresh conflict** attribute from the added column asset.



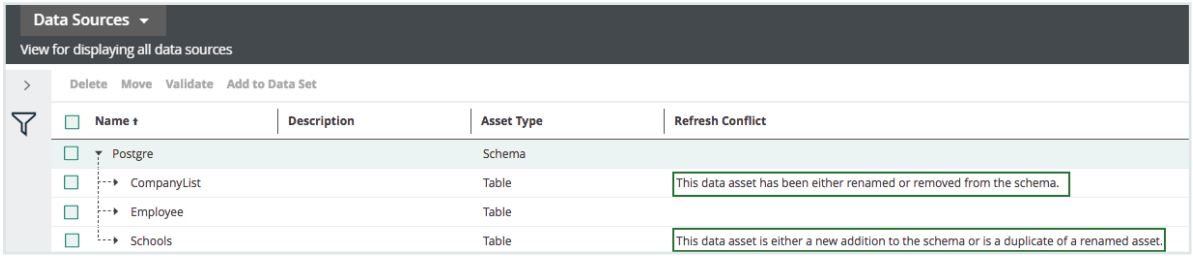
- A column is renamed in the schema:
  a. You have to remove the column asset with the updated column name.
  b. You have to rename the original column name to the newly ingested column name and delete the **Refresh Conflict** attribute.



- A column is deleted from the schema: this is automatically detected by the refresh operation. No further action is required of you.
- A column is added to the schema: this is automatically detected by the refresh operation. No further action is required of you.
- A table is renamed in the schema:
  a. You have to manually delete the renamed new table and all the columns contained in the table.
  b. You have to manually rename the existing old table and all the columns contained in the table.



- A table is deleted from the schema and another table is added to the schema:
  a. You have to manually delete the deleted table and all the columns in the table.
  b. You have to manually delete the Refresh Conflict attribute for the added table.



> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Resolve a schema refresh conflict when columns are added and deleted at the same time

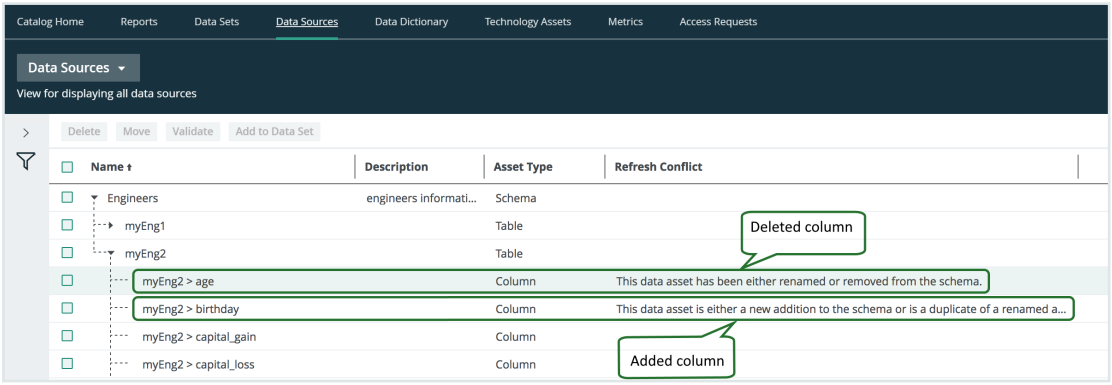
If you refresh a schema, the ingestion process will detect conflicts if the data source has the following changes:

- A column has been removed.
- A column has been added.

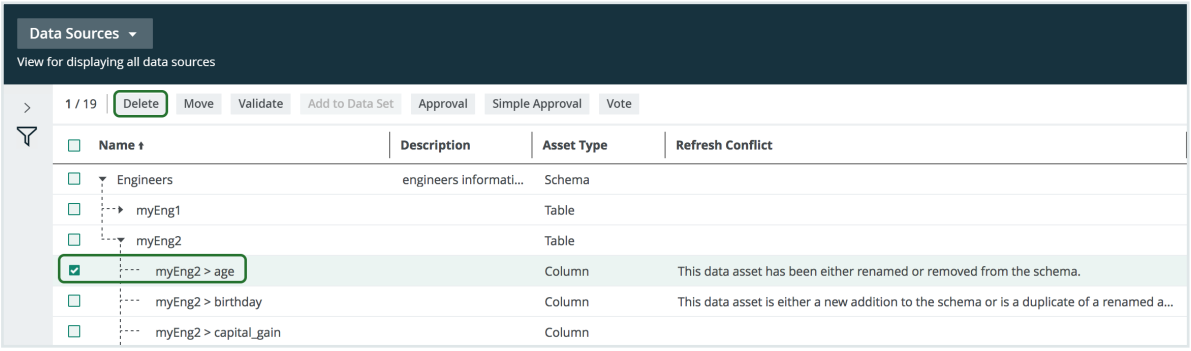
In the following example, the ingested schema has a column **age** and in the updated schema, the column **age** is removed and a column **birthday** is added.

To resolve such a refresh conflict, follow these steps:

1. Look up the data source with the search function or as follows:
   a. On the main menu, click ⠿, and then click Catalog.
      » The Catalog Home opens.
   b. In the submenu, click **Data Sources**.
   c. Optionally, add the **Refresh Conflict** column to the table.
   d. In the table, expand the relevant schema and table to find the columns with refresh conflicts.

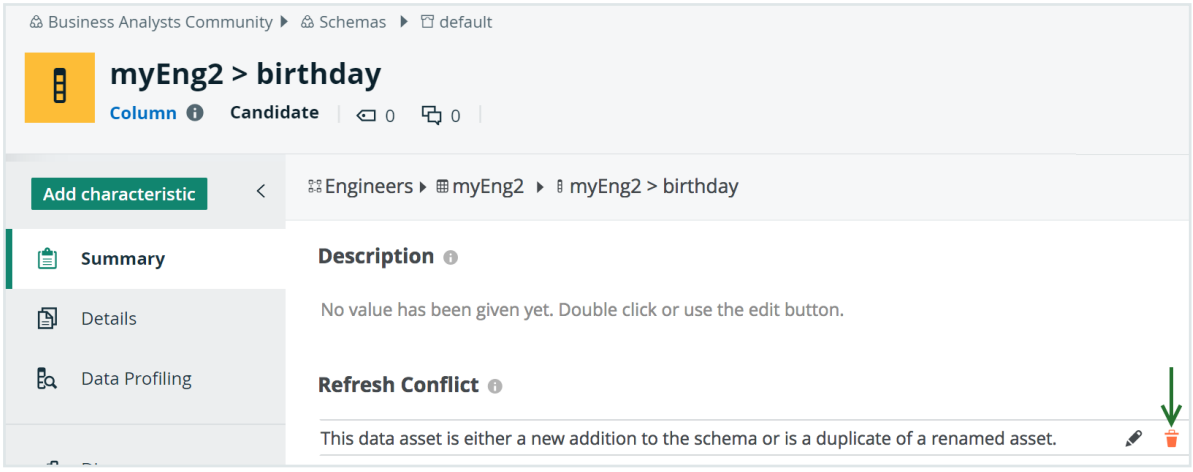


2. Select the column that is removed from the data source. In this example it is the **age** column.
   If necessary, select all column assets that are removed from the data source.
3. Above the table click **Delete**.

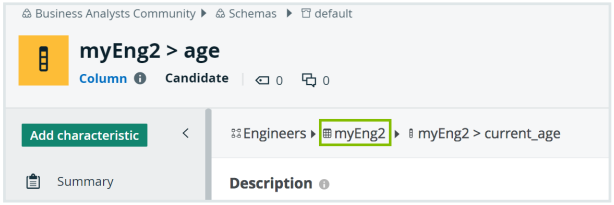


4. Click **Yes** to confirm the deletion of the column.
5. Click the name of the added column name.
   » The column asset page appear.

6. In the **Refresh Conflict** section of the column asset page, hover over the message and click 🗑 on the right-hand side.



7. Click **Yes** to confirm the deletion of the attribute.
8. Click the browser's **Back** button to return to the **Data Sources** view of the table.
   You can also click on the breadcrumb, as shown in the following image, to open the table asset page of the ingested schema.'



9. Repeat steps 5 to 8 for all other added columns.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

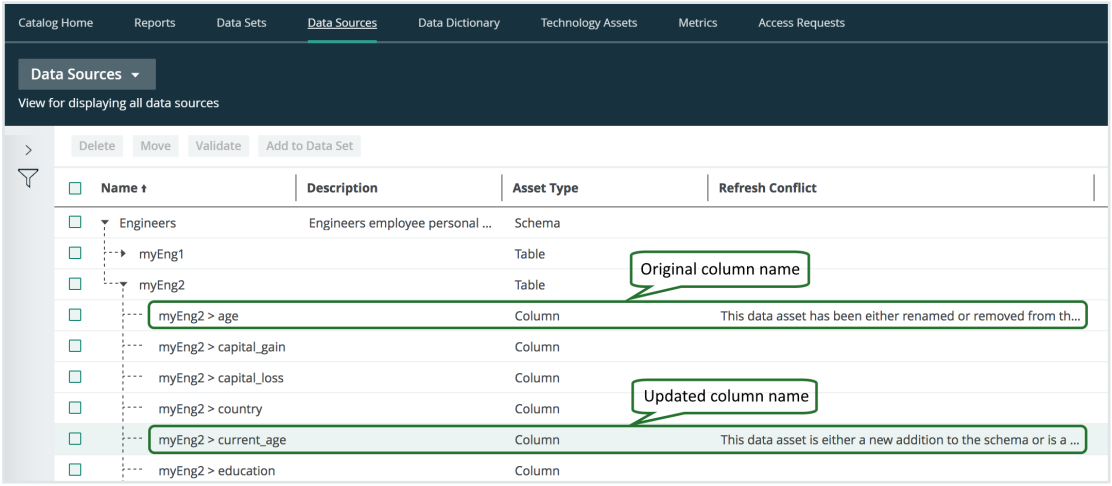
Resolve a schema refresh conflict for a renamed column

If you refresh a schema where the data source contains a column that has been renamed, the ingestion process will detect a conflict. In the following example, the ingested schema

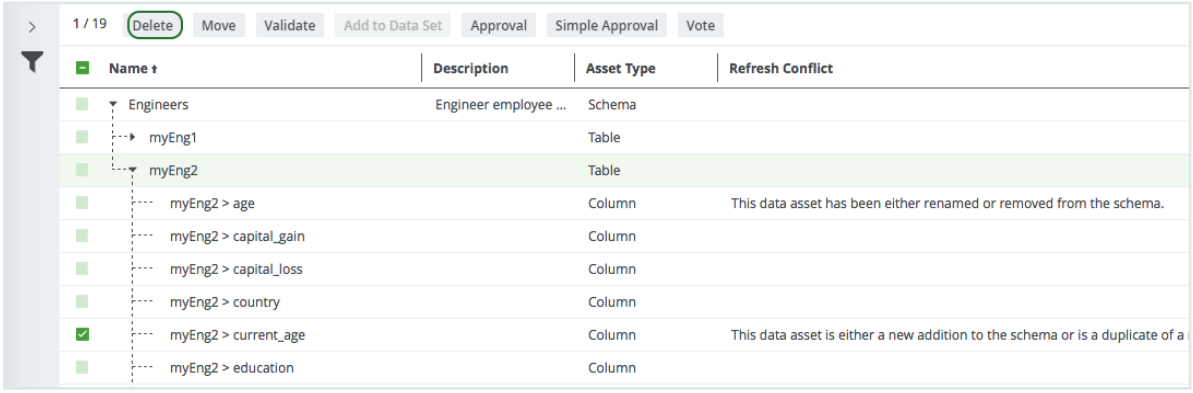
contains a column **age**, and in the updated schema, the column name has become **current_age**.

To resolve a refresh conflict due to a column rename, follow these steps:

1. Look up the new column with the search function or as follows:
   a. On the main menu, click ⠿, and then click **Catalog**.
      » The Catalog Home opens.
   b. In the submenu, click **Data Sources**.
   c. Optionally, add the **Refresh Conflict** column to the table.
   d. In the table, expand the relevant schema and table to find the columns with refresh conflicts.

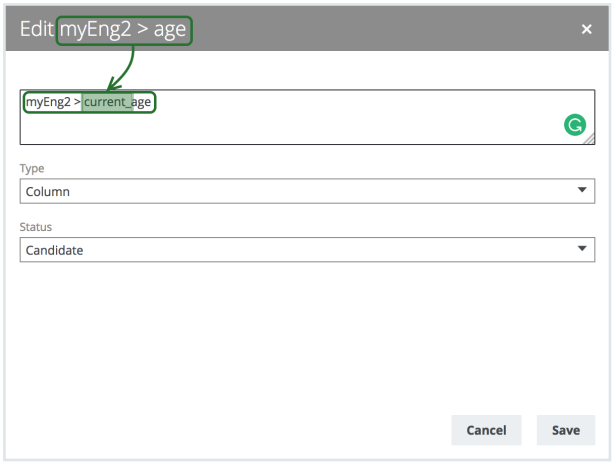


2. Select the updated column name and click **Delete** above the table.
   If necessary, select all column assets that are removed from the data source.

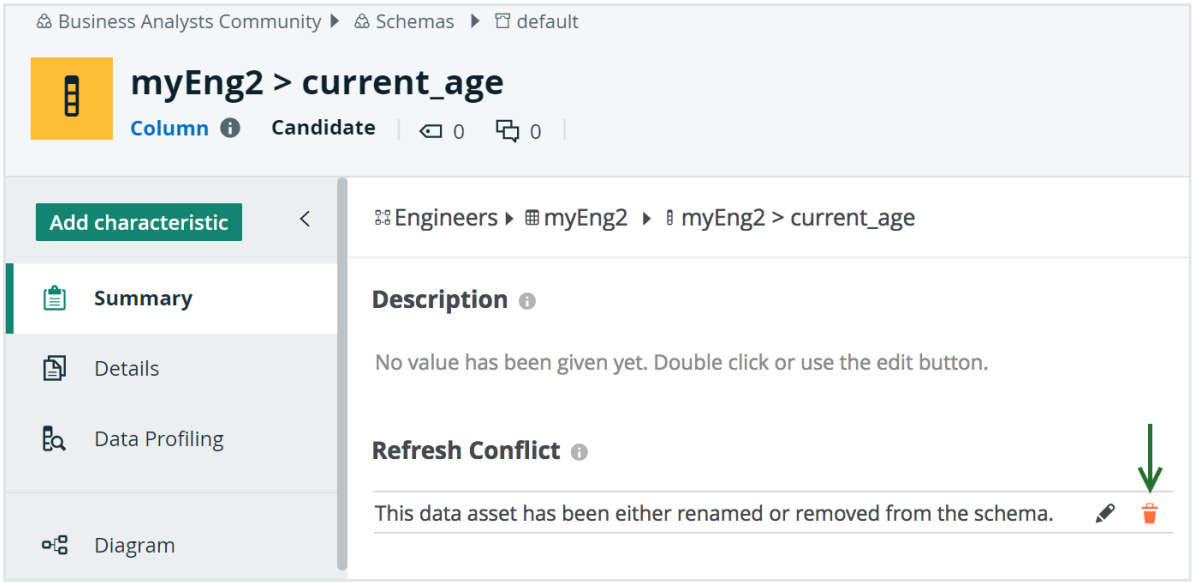


3. Click **Yes** to confirm the deletion of the column asset(s).
4. Click the name of the original column name.
   » The column asset page appears.
5. Click **Actions** > **Edit**.
   » The **Edit <asset name>** dialog box appears.

6. Change the name to the new ingested name.

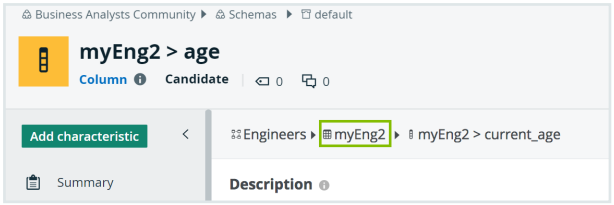


7. Click **Save**.
8. Refresh the page.
9. Leave the column asset page open.
10. In the **Refresh Conflict** section of the column asset page, hover over the message and click on the right-hand side.



11. Click **Yes** to confirm the deletion of the attribute.
12. Click the browser's **Back** button to return to the **Data Sources** view of the schema. You can also click on the breadcrumb, as shown in the following image, to open the

table asset page of the ingested schema.



13. If necessary, repeat steps 4 to 12 for other renamed column assets.

What's next?

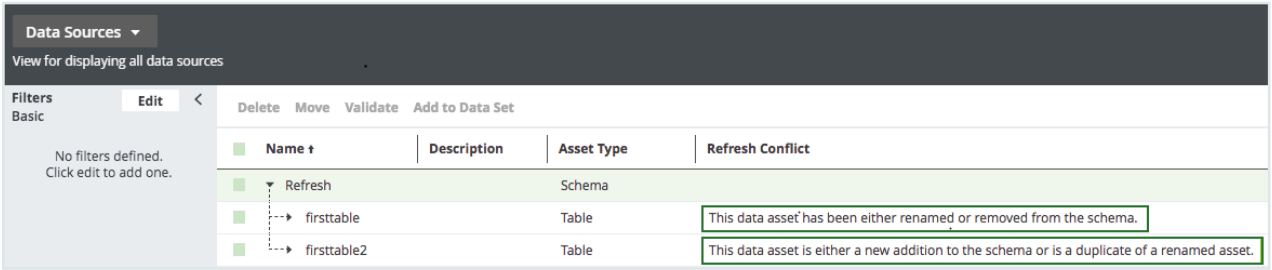
You can now safely refresh the schema with the new data source; however, keep in mind this may take some time.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Resolve a schema refresh conflict for a renamed table

If you refresh a schema where the data source contains a table that has been renamed, the ingestion process detects a conflict.

In the following example, the original schema **Refresh** contains the table **firsttable**. This table has been renamed to **firsttable2**. After refreshing the schema, refresh conflicts appear, as shown in the following image:

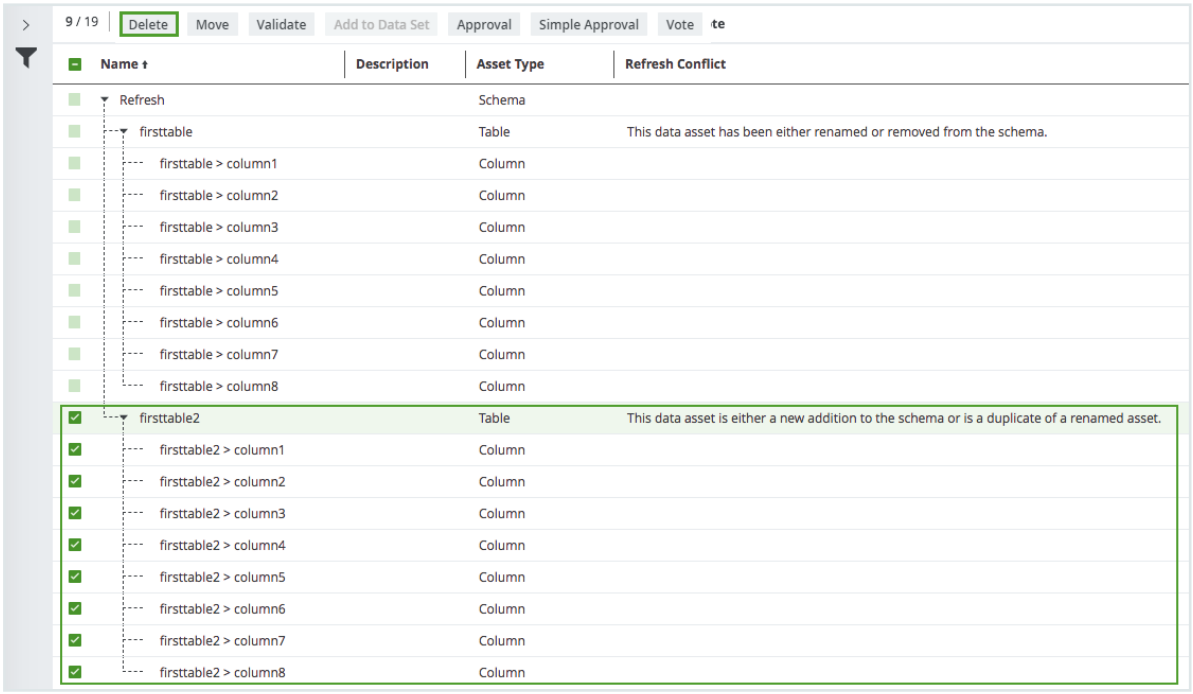


You have to manually resolve the conflicts before you continue. It is not possible to refresh a schema when there are conflicts.

> Note You have to add the **Refresh Conflict** column to the table if it is not there already.

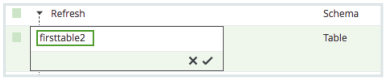
Steps

1. On the main menu, click ⋮⋮⋮, and then click **Catalog**.
   » The Catalog Home opens.
   » The Catalog Home appears
2. In the submenu, click **Data Sources**.
3. Expand the tables to see all the columns that are contained in them.
4. Select the renamed table and all its contained columns, in this example, **firsttable2**.
5. Above the table, click **Delete**.



6. Click **Yes** to confirm the deletion.
7. Hover over the original table, in this example, **firsttable**, and click ✎ to the right of the table name.

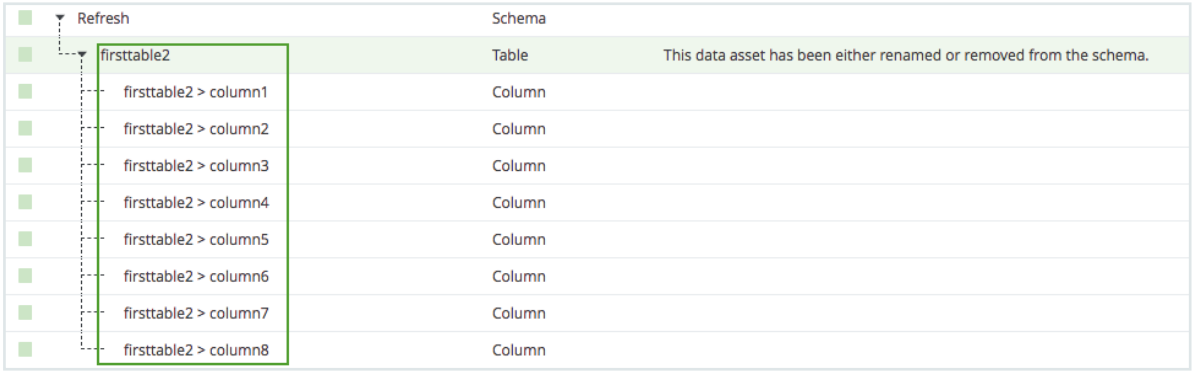
8. Change the name to the new ingested table name, in this example, **firsttable2**, and click ✓ to apply the change.



9. Hover over a column contained in the table you just renamed and click ✎ to the right of the column name.
10. Rename the column by replacing the table part of the name with that of the renamed table and click ✓ to apply the change.

    The column name is a concatenation of the table name and the original column name and so you just have to replace the table part of the name with the new table name. For example, to rename the column name **firsttable > column1** to **firsttable2 > column1**, you just have to change **firsttable** to **firsttable2** so that the column name becomes **firsttable2 > column1**.
11. Repeat this action for all the columns in the renamed table.

    Now, you only see the new ingested table, **firsttable2**, and the columns contained in the table.



12. Click the name of the renamed table.

    » The table asset page appears.
13. In the **Refresh Conflict** section, hover over the refresh conflict message and click 🗑 on the right-hand side.

14. Click **Yes** to confirm the deletion of the Refresh Conflict attribute.

What's next?

You can now safely refresh the schema with the data source.

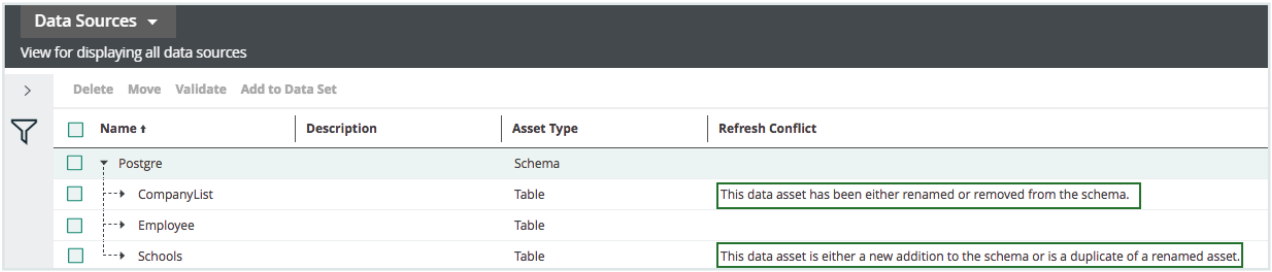
Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Resolve a schema refresh conflict when tables are added and deleted at the same time

When you refresh a schema, the ingestion process detects conflicts if the data source has the following changes at the same time:

- A table has been removed.
- A table has been added.

In the following example, the original schema **Postgre** contains the table **Employee** and the table **CompanyList**. A new table **Schools** has been added to the schema and the table **CompanyList** has been deleted. After refreshing the schema, refresh conflicts appear for the added table and the deleted table, as shown in the following image:



You have to manually resolve the conflicts before you continue. It is not possible to refresh a schema when there are conflicts.

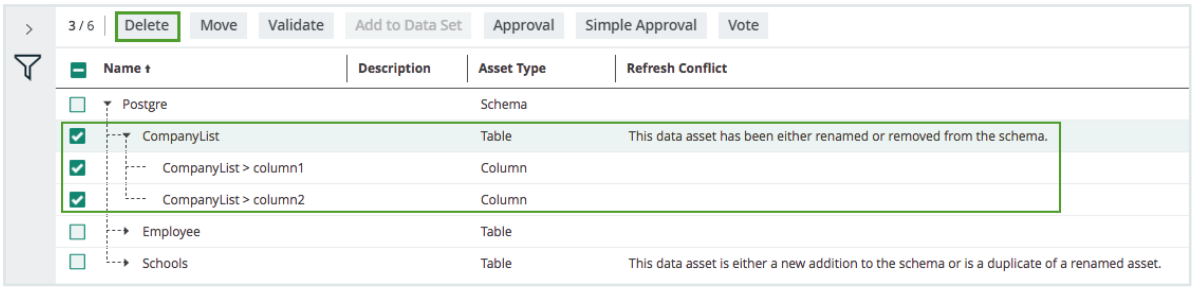
Note You have to add the **Refresh Conflict** column to the table if it is not there already.

Steps

1. On the main menu, click , and then click **Catalog**.
   » The Catalog Home opens.

» The Catalog Home appears.

2. In the submenu, click **Data Sources**.
3. Select the deleted table and all its contained columns, in this example, **CompanyList**.



4. Above the table, click **Delete**.
5. Click **Yes** to confirm the deletion.
6. Click the name of the added table, in this example, **Schools**.

» The table asset page appears.

7. In the **Refresh Conflict**section, hover over the refresh conflict message and click on the right-hand side.

8. Click **Yes** to confirm the deletion of the Refresh Conflict attribute.

What's next?

You can now safely refresh the schema with the data source.

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments. For more information, go to Announcements.

Advanced data type detection is slow

Advanced data type (ADT) detection is the process that compares each value in the database with each pattern in the ADT definition list.

The following non-exhaustive list contains the factors that affect the detection time:

- The higher the number of ADTs in Catalog, the longer the detection time.
- The higher the number of patterns in each ADT, the longer the detection time.
  For example, a text ADT can contain one or more regular expressions. The more regular expressions that you add to this ADT, the longer the detection time will take.

Tip As a general rule, try to limit both the number of ADTs and the number of patterns per ADT.

Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Jobserver troubleshooting

This is a list of known issues in versions older than Collibra 2023.11.

Problem	Solution
One or more of the following error messages appear: • `context JS-<context ID> not found` in the Jobserver node in DGC logs. • `manager_start - /opt/collibra/spark-job-server/bin/manager_start.sh: line 73: <process id> killed` in the Jobserver server logs. • Spark context logs are interrupted during Spark processing. • `It is not possible to allocate enough memory` in the Spark process or other process on the same machine.	If the Spark context crashes or is unresponsive, it can be related to a memory shortage. Make sure that you have enough memory. • In 5.7, a Jobserver node should have 64GB RAM, 16 CPUs and 500GB SSD. • In 5.7.1, the Spark context process configuration for each Jobserver requires you to change the lower the heap memory to 40GB and replace the -XX:+UseG1GC option by -XX:+UseParallelGC.
An ingestion job keeps on running due to lingering Spark Context.	Restart the Jobserver, then restart Collibra.

Problem	Solution
Communication failure occurs between Jobserver and Spark Context when profiling large tables.	The following relevant parameters can be edited in the Jobserver configuration file to decrease the chance that this problem occurs: • `acceptable-heartbeat-pause` should be 600s. • `heartbeat-interval` should be 300s. • `threshold` should be 12.0

> Warning We have announced the end of life of Jobserver and all related Jobserver integrations for **September 30, 2024**, with the exception of Public Sector customers using GovCloud or on-prem environments.
> For information on registering a data source via Edge, go to Registering and synchronizing a data source via Edge.

Jobserver jobs

To ingest data in Collibra Platform Self-Hosted, you have to register a data source. During the ingestion, you can include to run a data profiling, data sampling and to detect advanced data types in the data.

The DGC service is responsible for the ingestions, the Jobserver is responsible for the data profiling, data sampling and advanced data type detection.

The following table shows how many jobs it takes to complete a task. The jobs are executed sequentially.

Task	Number of jobs
Data profiling	4 jobs per table
Data sampling	2 jobs per table
Advanced data type detection	1 job per table
Data ingestion	0 job

If you have to troubleshoot Jobserver jobs, you need the following log files when you create a diagnostic file.

- Collibra logs
- Jobserver logs: You have to enable the ingestion and profiling logs.
- Spark logs: You have enable to the Spark logs. When you create a diagnostic file, these are included with the Jobserver logs.

Removing outdated drivers during upgrade to 2022.11

Why remove outdated drivers?

When you register a data source via Jobserver for PostgreSQL, Oracle or SQL Server and select **Register data source [use your own driver]** in the **Create** dialog box, Collibra automatically provides the driver configuration with the PostgreSQL, Oracle, or SQL Server driver JAR file. These JAR files, however, are not maintained and need to be removed.

During the 2022.11 upgrade, we removed the JAR files from the out-of-the-box driver configurations for PostgreSQL, Oracle, and SQL Server from your environment.

- If you did not use the out-of-the-box driver configuration in any Jobserver connection, the out-of-the-box driver configuration is removed and you don't need to do anything.
- If you used the out-of-the-box driver configuration, the configuration remains available but the driver JAR file is removed. You have to update the driver with a new driver JAR file after the upgrade. You can find the driver JAR files on the official downloads pages of the data source vendors.

The dgc.log file contains a section about the removal of the these driver JAR files.

After the 2022.11 release, you can still register a data source via **Register data source [use your own driver]** for PostgreSQL, Oracle and SQL Server, but Collibra will no longer provide the outdated driver configuration and JAR file.

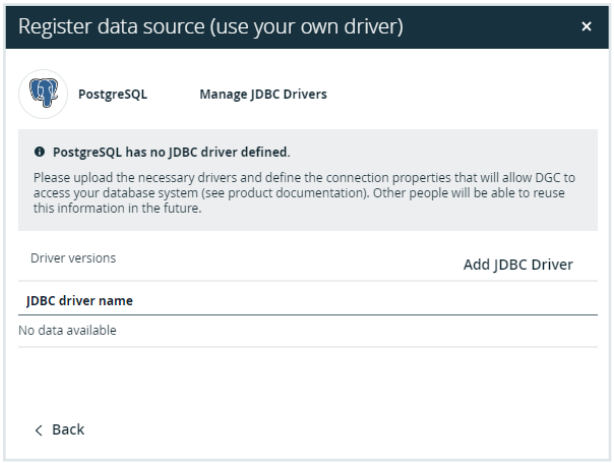
How can you see the impact on your environment?

You can check the impact:

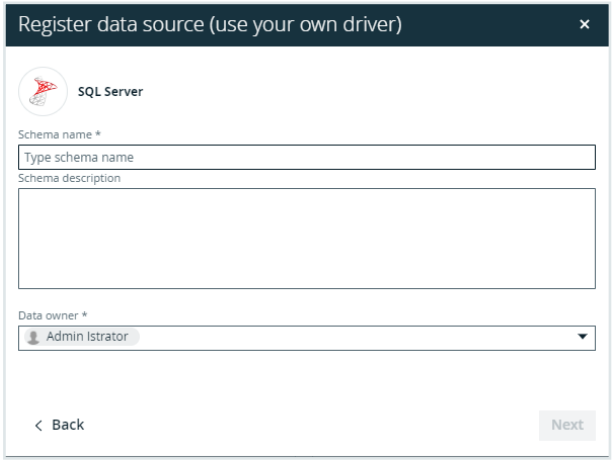
- Via the Register data source dialog box
- In the dgc.log file

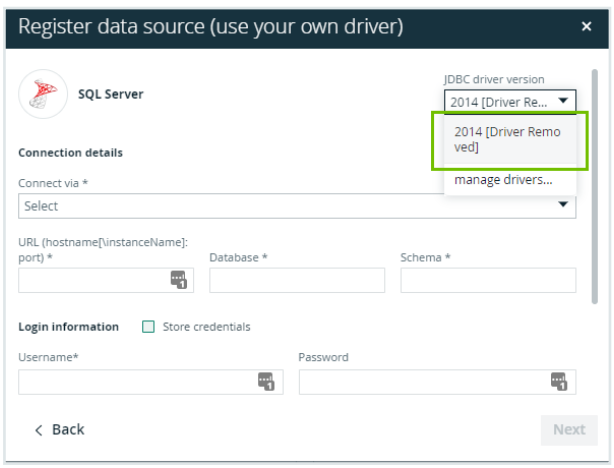
Check the impact via the **Register data source**dialog box

- **You don't have to do anything** if you register a data source via Jobserver for PostgreSQL, Oracle or SQL Server, select **Register data source [use your own driver]** in the **Create** dialog box, and you don't receive an out-of-the-box configuration.



- You have to update the driver if you register a data source via Jobserver for PostgreSQL, Oracle or SQL Server, select **Register data source [use your own driver]** in the **Create** dialog box, provide a name, click Next and see a driver with **[Driver Removed]** in its name.

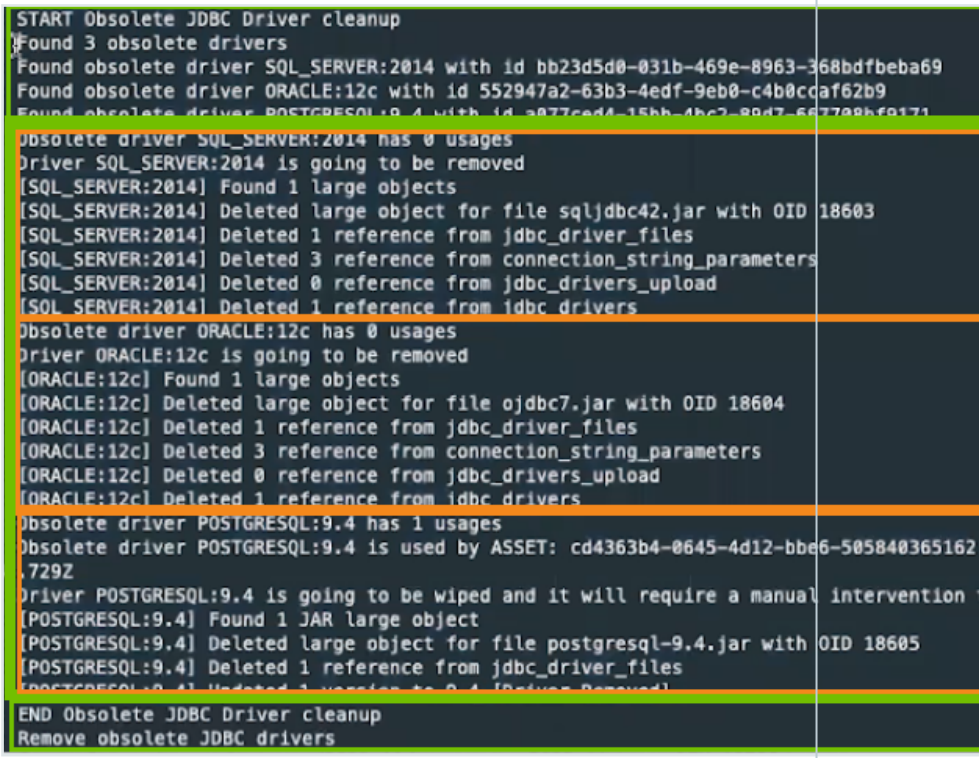




Check the impact via the dgc.log file

After the 2022.11 upgrade, the dgc.log file contains a section about the removal of the driver JAR files. Based on this section, you can find out whether you have to do anything.

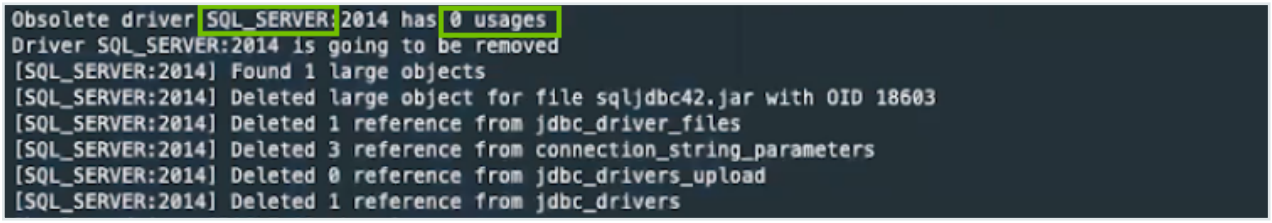
1. Download the dgc.log file.
   a. Open Collibra Console with a user profile that has at least the **ADMIN** role.
      » Collibra Console opens with the **Infrastructure** page.
   b. Expand an environment and select **Data Governance Center service**.
   c. Click **Logs**.
      » The **Logs** tab page opens.
   d. For dgc.log, click the Download icon.
      » The zipped log file is downloaded to your machine.
2. Open the dgc.log file.
   The dgc.log file contains a section about the removal of the drivers which consists of multiple subsections.

Nr	Section	Example
1	`Start obsolete JDBC driver cleanup.` This is an introduction to the removal of the drivers for PostgreSQL, Oracle and SQL Server.	
2	A subsection for each obsolete driver. This is where Collibra checks if the driver has been used in any Jobserver connection. The possible results are: ◦ `Obsolete driver has 0 usages` ◦ `Obsolete driver has a number of usages`	

Nr	Section	Example
3	`End obsolete JDBC driver cleanup`	

Obsolete driver has 0 usages

If the message for the data source subsection indicates there are zero usages, it means the out-of-the-box driver configuration was not used to register a data source with your own driver. In that case,Collibra removes the out-of-the-box driver configuration and **you don't have to do anything**.
In this example, the SQL-server out-of-the-box driver configuration driver was not used to register a data source with your own driver. You don't have to do anything for Microsoft SQL Server in the upgraded environment.



Obsolete driver has a number of usages

If the message in the data source section indicates there are usages:

- Collibra does the following:
  - Adds the schema asset IDs connected to the old driver in the dgc.log file.
  - Deletes the old JAR file from the driver.
  - Renames the driver to name + [Driver Removed].
- You **need to update the driver**. For information, go to Update a driver after the 2022.11 upgrade.

In this example, the Postgresql driver was used once. The dgc.log file provides the schema asset ID. You have to update the Postgresql driver in the upgraded environment.

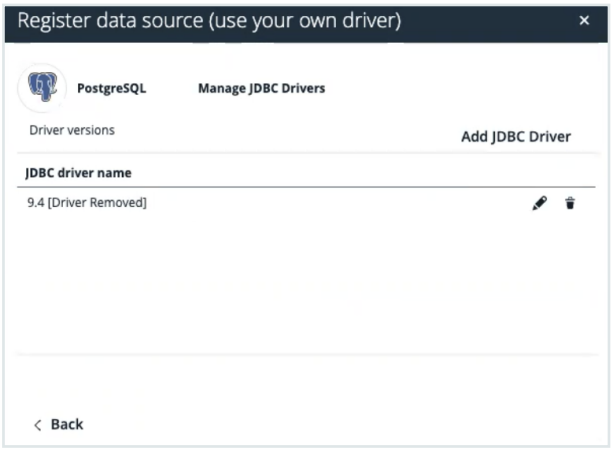
```
Obsolete driver POSTGRESQL 9.4 has 1 usages
Obsolete driver POSTGRESQL:9.4 is used by ASSET: cd4363b4-0645-4d12-bbe6-505840365162 CREATED BY 000000(
1.729Z
Driver POSTGRESQL:9.4 is going to be wiped and it will require a manual intervention to fix its usages
[POSTGRESQL:9.4] Found 1 JAR large object
[POSTGRESQL:9.4] Deleted large object for file postgresql-9.4.jar with OID 18605
[POSTGRESQL:9.4] Deleted 1 reference from jdbc_driver_files
[POSTGRESQL:9.4] Updated 1 version to 9.4 [Driver Removed]
```

Update a driver after the 2022.11 upgrade

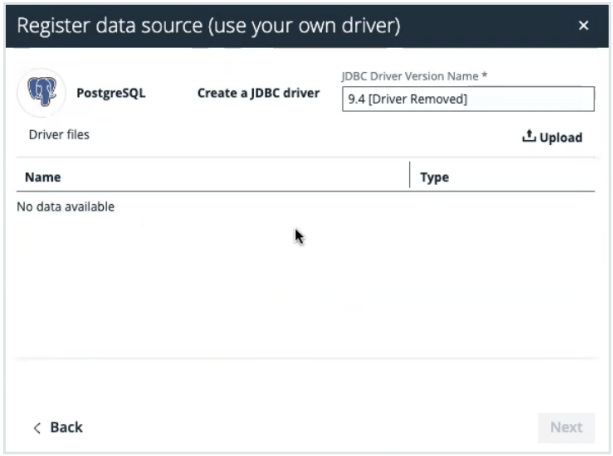
When you upgrade to Collibra 2022.11, you may need to replace PostgreSQL, Oracle, and SQL Server drivers.

Steps

1. Check and identify which drivers you need to update.
2. For each driver you need to update, do the following:
   a. Download the latest JAR file for the driver.
      You can download the driver JAR files from the official downloads pages of the data source vendors.
   b. Navigate to one of the schemas that used the old driver.
      The log file contains the IDs of the affected schemas.
   c. In the Schema asset, go to **Actions** → **Refresh**.
      » The **Refresh Schema** dialog box appears.
   d. In **JDBC driver version**, click **Manage drivers**.



e. Click ✎ for the driver with **[Driver Removed]** in its name.



f. Click **Upload** and upload the new JAR file.

g. Change the name of the driver to the driver name.

> Tip It is useful to mention the latest version number in the name.

h. Click **Next**.
You do not need to update the properties.

i. Click **Update**.

j. Click **Save and refresh**.
» The schema is refreshed and uses the new driver.
» If other schemas used this driver configuration too, they are updated automatically.