Collibra

Collibra Data Governance Center

Data Catalog

Version 5.7

Collibra

Collibra Data Governance Center - Data Catalog

Release date: June 24th, 2022

Revision date: Thu Jun 23, 2022

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 DGC.
Data Sources	Data sources that are used for data source registrations.
Data Dictionary	All data assets in Collibra DGC.
Technology Assets	All technology assets in Collibra DGC.
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.

> Note When you use Edge, Catalog experience has to be enabled.

Page layout

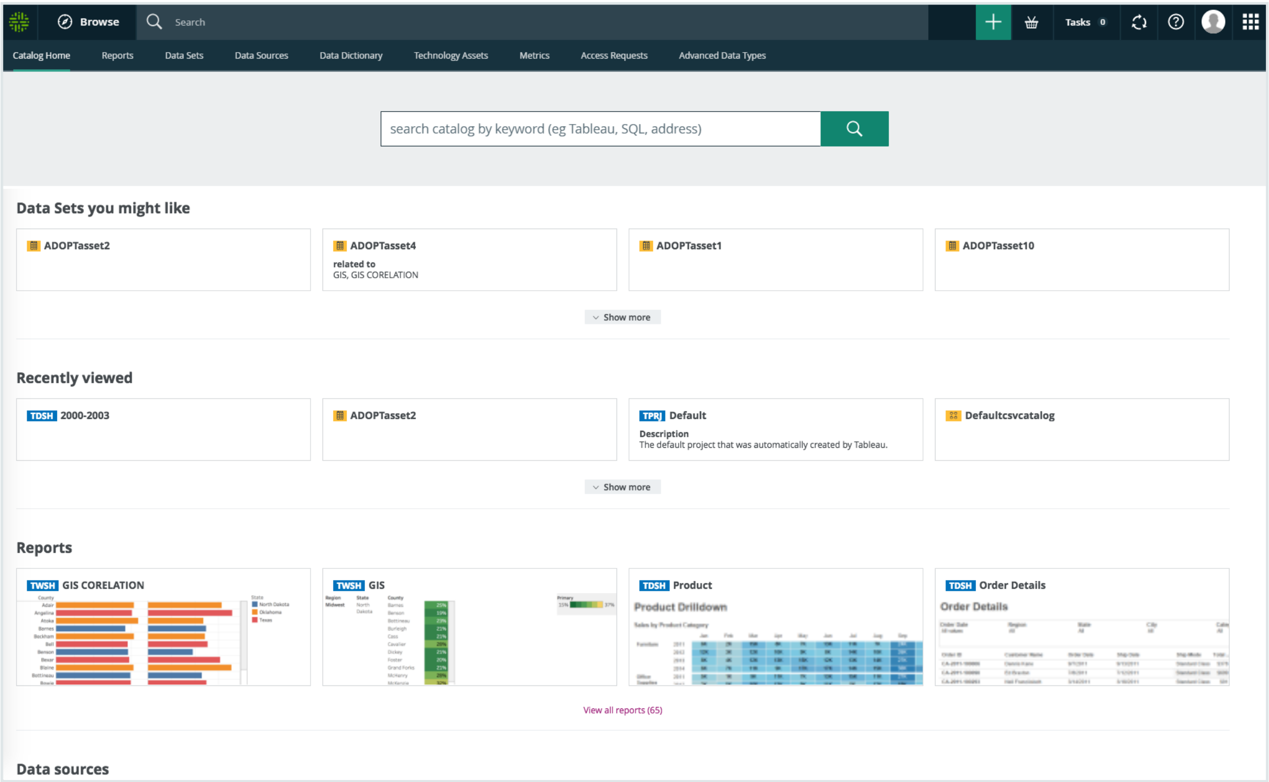
For more information on the Data Catalog asset pages, see the online version of this guide.

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.



search catalog by keyword (eg Tableau, SQL, address)

Data Sets you might like

- ADOPTasset2
- ADOPTasset4 related to GIS, GIS CORELATION
- ADOPTasset1
- ADOPTasset10

Show more

Recently viewed

- TDSH 2000-2003
- ADOPTasset2
- TPRJ Default Description The default project that was automatically created by Tableau.
- Defaultcsvcatalog

Show more

Reports

- TWSH GIS CORELATION
- TWSH GIS
- TDSH Product
- TDSH Order Details

View all reports (65)

Data sources

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.

Element name	Description
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.
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

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 Data Governance Center.

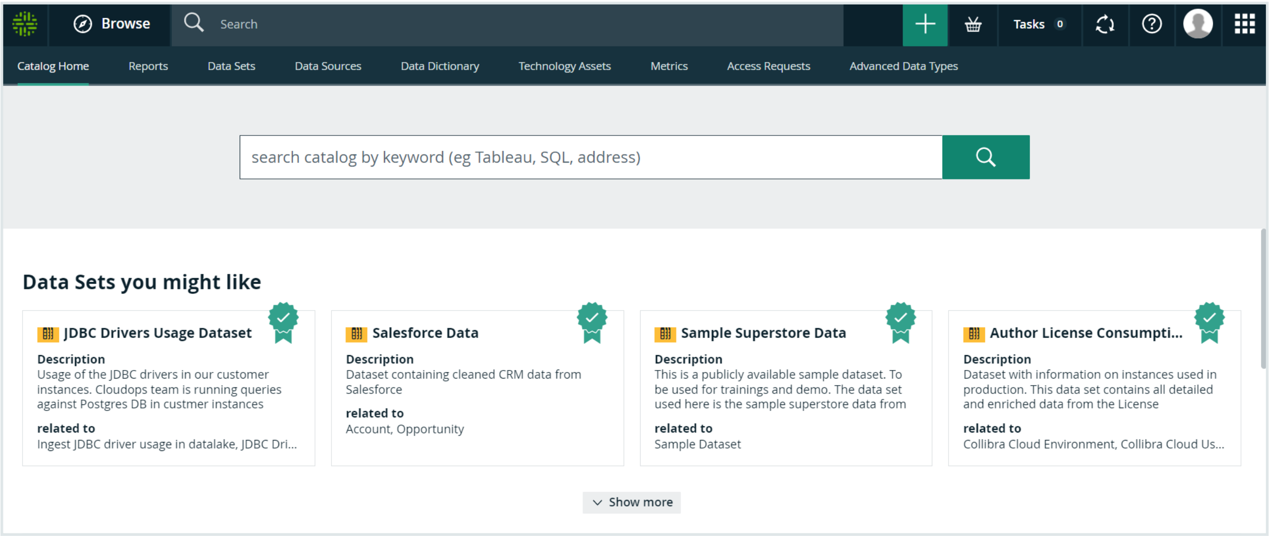
Recommendation of data sets to users

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



Data Sets you might like

JDBC Drivers Usage Dataset
Description
Usage of the JDBC drivers in our customer instances. Cloudops team is running queries against Postgres DB in custmer instances
related to
Ingest JDBC driver usage in datalake, JDBC Dri...

Salesforce Data
Description
Dataset containing cleaned CRM data from Salesforce
related to
Account, Opportunity

Sample Superstore Data
Description
This is a publicly available sample dataset. To be used for trainings and demo. The data set used here is the sample superstore data from
related to
Sample Dataset

Author License Consumpti...
Description
Dataset with information on instances used in production. This data set contains all detailed and enriched data from the License
related to
Collibra Cloud Environment, Collibra Cloud Us...

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 DGC, it only considers 3 parameters:
>
> - Certified
> - Quality
> - Popularity (number of views of the data set asset page)

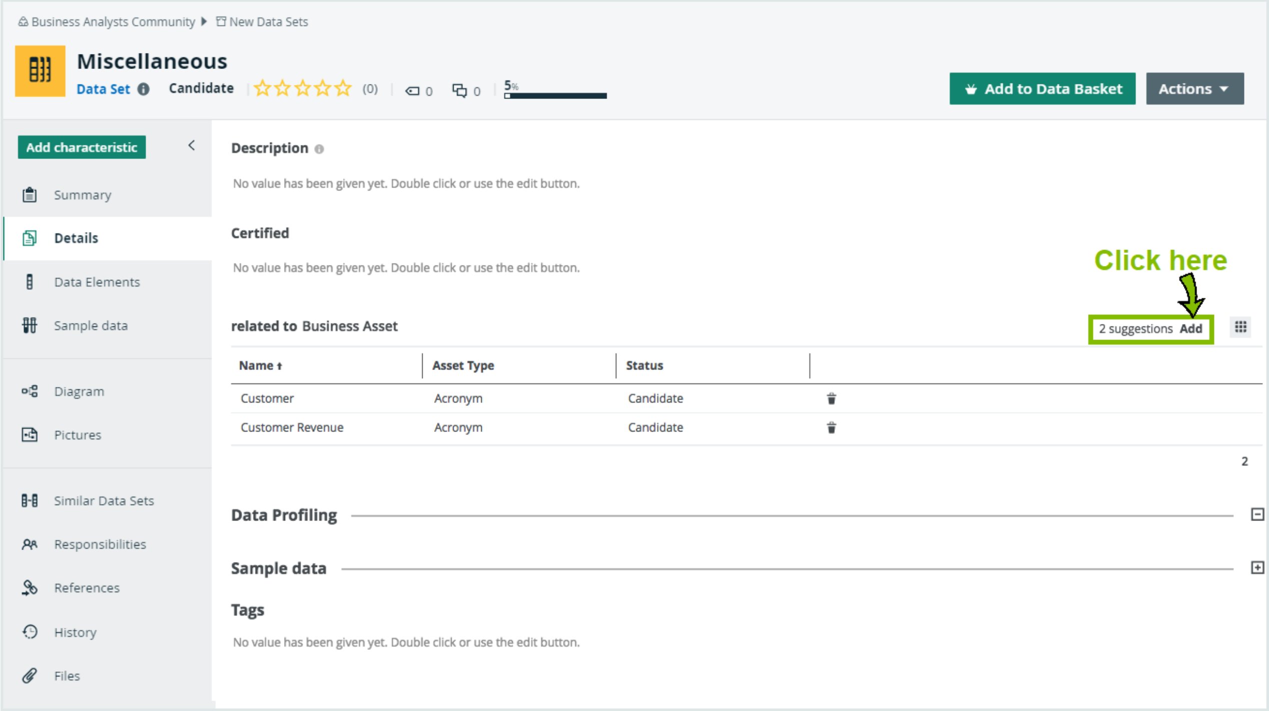
Recommendation of business assets to data sets

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



Business Analysts Community ▸ New Data Sets

Miscellaneous

Data Set · Candidate · ☆☆☆☆☆ (0) · 0 · 0 · 5%

Add to Data Basket · Actions ▾

Description

No value has been given yet. Double click or use the edit button.

Certified

No value has been given yet. Double click or use the edit button.

Click here

related to Business Asset 2 suggestions Add

Name	Asset Type	Status
Customer	Acronym	Candidate
Customer Revenue	Acronym	Candidate

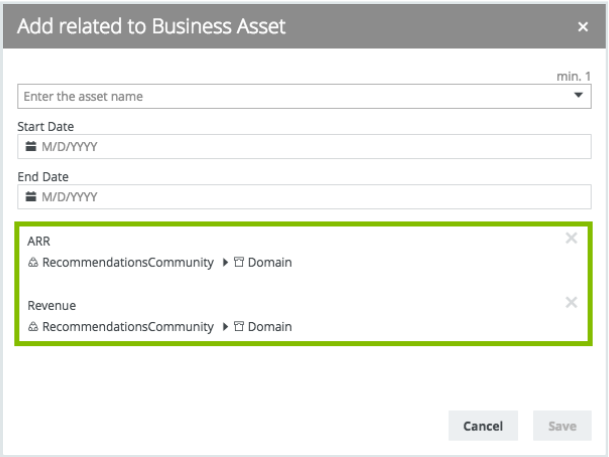
2

Data Profiling

Sample data

Tags

No value has been given yet. Double click or use the edit button.



Add related to Business Asset ×

min. 1

Enter the asset name

Start Date

M/D/YYYY

End Date

M/D/YYYY

ARR
RecommendationsCommunity ▸ Domain

Revenue
RecommendationsCommunity ▸ Domain

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.

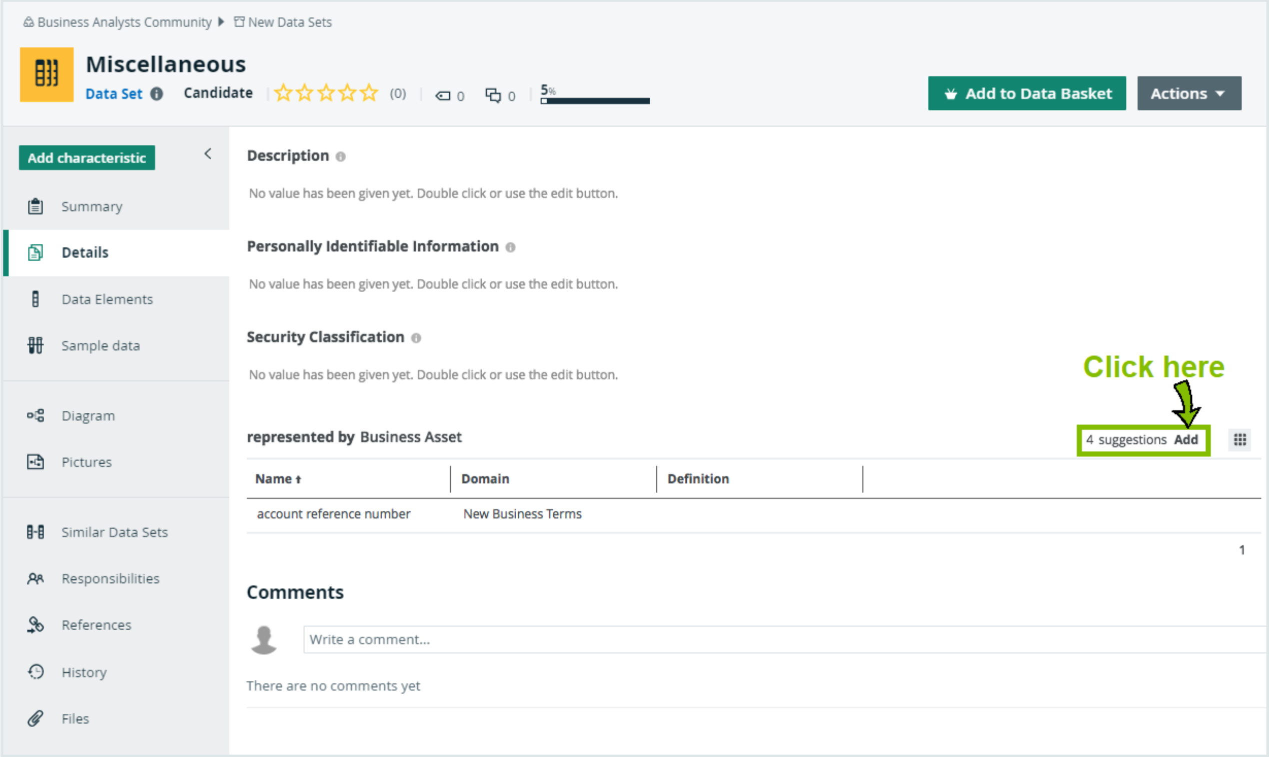
4. Recommend those business assets on the data set page and in the **Add related to** dialog box.

> Note If the recommender does not have enough data, for example if you just started using Collibra DGC, 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 DGC. The recommendations are shown in the section of **data asset represented by business asset** relation.

Example



Description

No value has been given yet. Double click or use the edit button.

Personally Identifiable Information

No value has been given yet. Double click or use the edit button.

Security Classification

No value has been given yet. Double click or use the edit button.

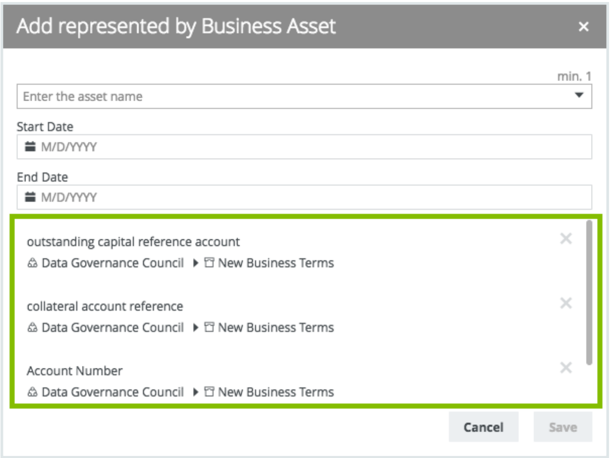
represented by Business Asset

Name	Domain	Definition
account reference number	New Business Terms	

Comments

Write a comment...

There are no comments yet



Add represented by Business Asset

min. 1

Enter the asset name

Start Date

M/D/YYYY

End Date

M/D/YYYY

outstanding capital reference account
Data Governance Council ▸ New Business Terms

collateral account reference
Data Governance Council ▸ New Business Terms

Account Number
Data Governance Council ▸ New Business Terms

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 DGC. 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.

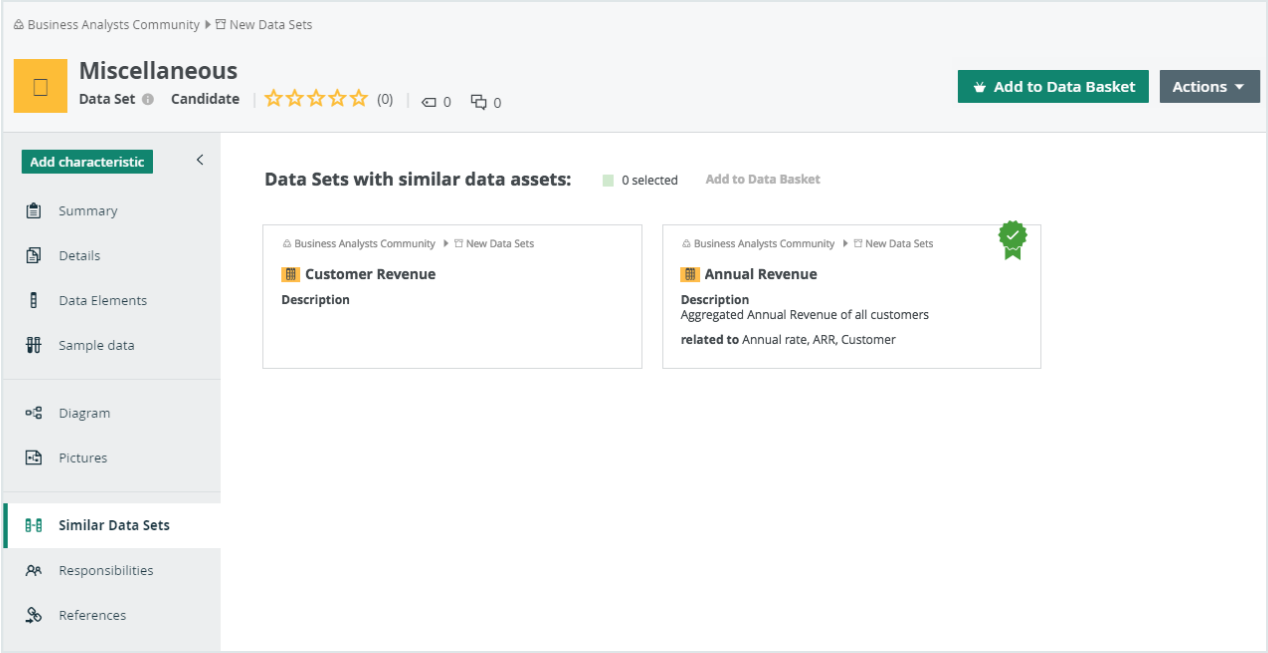
Matchers find similar data sets and schemas based on the name and the attributes.

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 Data Governance Center.

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.



Business Analysts Community ▸ New Data Sets

Miscellaneous

Data Set · Candidate · ☆☆☆☆☆ (0) · 0 · 0

Data Sets with similar data assets: 0 selected Add to Data Basket

Business Analysts Community ▸ New Data Sets
Customer Revenue
Description

Business Analysts Community ▸ New Data Sets
Annual Revenue
Description
Aggregated Annual Revenue of all customers
related to Annual rate, ARR, Customer

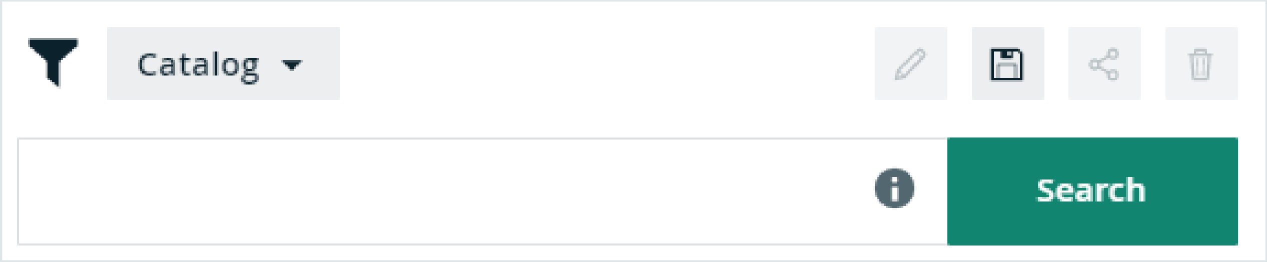
Schema matcher

The schema matcher is currently not used in Collibra DGC.

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 DGC 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.

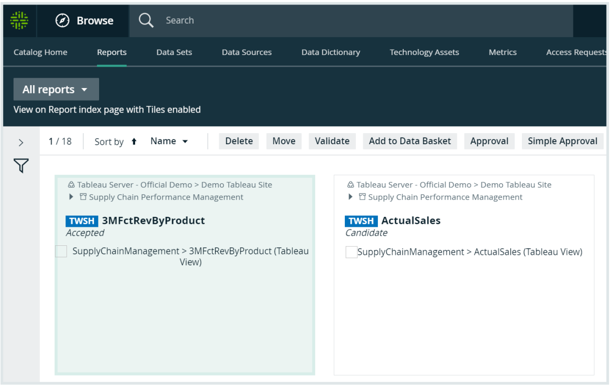
Report views

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.



View on Report index page with Tiles enabled

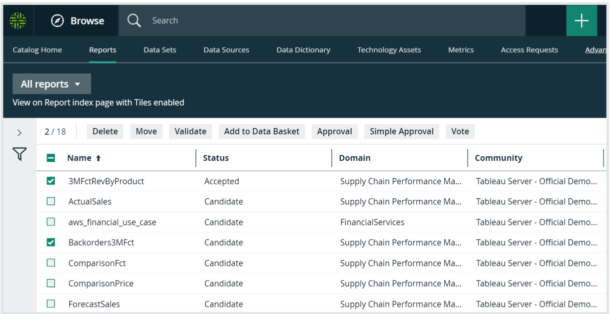
Tableau Server - Official Demo > Demo Tableau Site
Supply Chain Performance Management
TWSH 3MFctRevByProduct
Accepted
SupplyChainManagement > 3MFctRevByProduct (Tableau View)

Tableau Server - Official Demo > Demo Tableau Site
Supply Chain Performance Management
TWSH ActualSales
Candidate
SupplyChainManagement > ActualSales (Tableau View)

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.



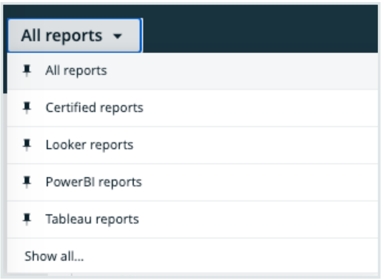
View on Report index page with Tiles enabled

	Name	Status	Domain	Community
☑	3MFctRevByProduct	Accepted	Supply Chain Performance Ma...	Tableau Server - Official Demo...
□	ActualSales	Candidate	Supply Chain Performance Ma...	Tableau Server - Official Demo...
□	aws_financial_use_case	Candidate	FinancialServices	Tableau Server - Official Demo...
☑	Backorders3MFct	Candidate	Supply Chain Performance Ma...	Tableau Server - Official Demo...
□	ComparisonFct	Candidate	Supply Chain Performance Ma...	Tableau Server - Official Demo...
□	ComparisonPrice	Candidate	Supply Chain Performance Ma...	Tableau Server - Official Demo...
□	ForecastSales	Candidate	Supply Chain Performance Ma...	Tableau Server - Official Demo...

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.

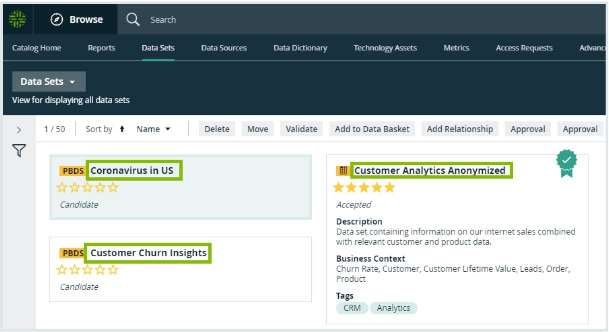
The Catalog **Data Sets** overview page displays existing data sets in a table or as tiles. The page displays the name of the data set, its description, its certification status, and, if there are any, connections to existing business assets in Collibra Data Governance Center.

Data Sets overview page

The Data Sets overview page contains the data sets that are available in Collibra Data Governance Center. 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.



View for displaying all data sets

Coronavirus in US

Customer Churn Insights

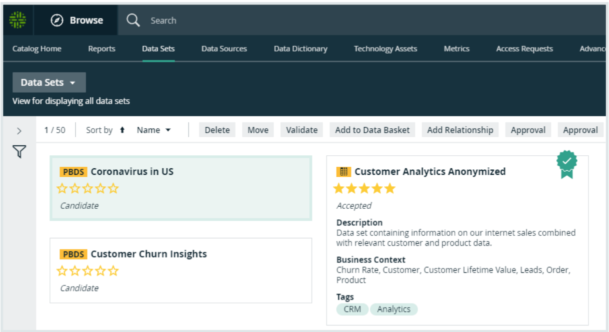
Customer Analytics Anonymized

Description
Data set containing information on our internet sales combined with relevant customer and product data.

Business Context
Churn Rate, Customer, Customer Lifetime Value, Leads, Order, Product

Tags

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



PBDS Coronavirus in US
☆☆☆☆☆
Candidate

PBDS Customer Churn Insights
☆☆☆☆☆
Candidate

Customer Analytics Anonymized
★★★★★
Accepted

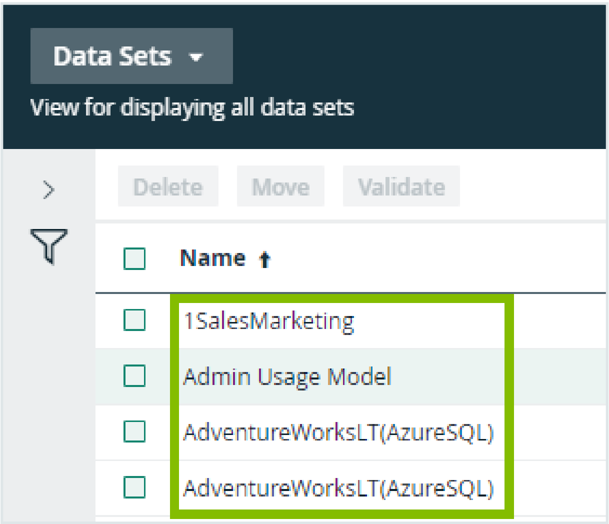
Description
Data set containing information on our internet sales combined with relevant customer and product data.

Business Context
Churn Rate, Customer, Customer Lifetime Value, Leads, Order, Product

Tags
CRM Analytics

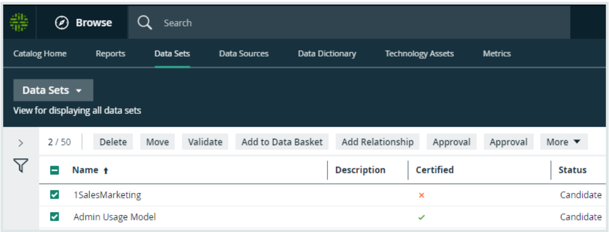
Table display mode

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



□	Name
□	1SalesMarketing
□	Admin Usage Model
□	AdventureWorksLT(AzureSQL)
□	AdventureWorksLT(AzureSQL)

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



☐	Name	Description	Certified	Status
☑	1SalesMarketing		✗	Candidate
☑	Admin Usage Model		✓	Candidate

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 Data Governance Center 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 and reports
- 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. In the main menu, click the **Create** (+) button.
   » 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. In the main menu, click ⠿, then 🛢 **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. In the main menu, click ⠿, then ☰ **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.

Requesting access to data

You can request access to data by adding the relevant data sets or reports to your Data Basket and checking out your Data Basket.

Adding data sets or reports to the Data Basket

You can add data sets or reports to the Data Basket by clicking **Add to Data Basket**. This button appears:

- When you've selected one or more data sets or reports in Catalog.
- On Data Set asset pages and Report asset pages.

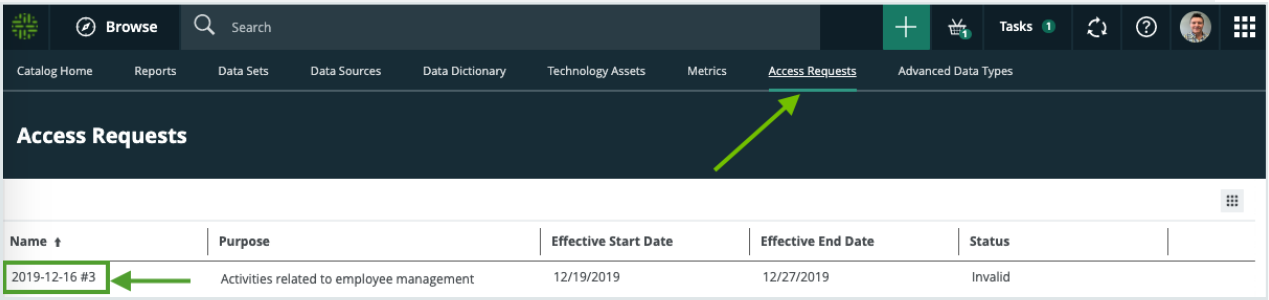
When you click **Add to Data Basket**:

- A Data Usage asset is created.
- All of the data sets and reports you selected are shown in the Data Basket.

Data Usage asset

The Data Usage asset is created in the Data Usages domain. The name of the Data Usage asset is "USER_BASKET_" followed by the UUID of the user.

> Tip The Data Usages domain is a "hidden" domain in the Business Analysts Community. This means it doesn't appear in the Collibra DGC Browser, which helps to avoid it being inadvertently deleted. To view the Data Usages domain, go to the Access Requests page and click the name of a Data Usage asset. The Data Usages domain appears in the breadcrumb, on the Data Usage asset page.



Access Requests

Name ↑	Purpose	Effective Start Date	Effective End Date	Status
2019-12-16 #3	Activities related to employee management	12/19/2019	12/27/2019	Invalid

The Data Usage asset page shows all of the important information related to the access request, including:

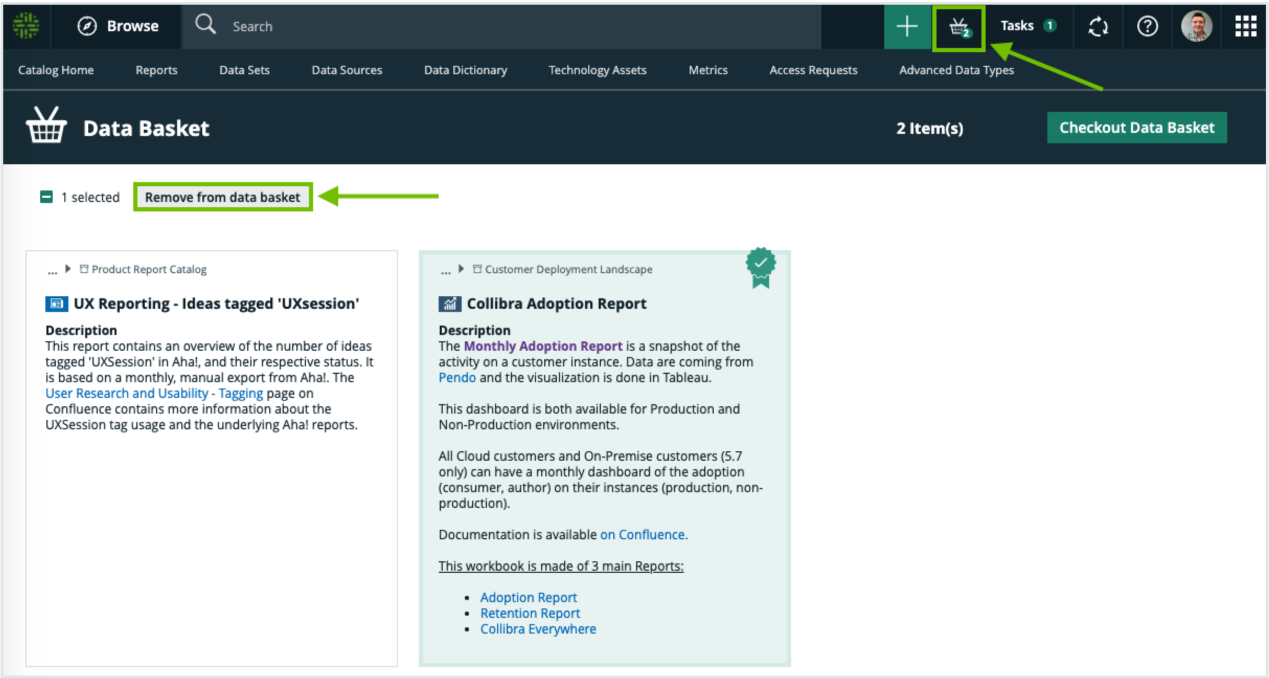
- The data sets or reports to which access is requested.
- The purpose for requesting access to the data.
- The access start date and end date.

Data Basket

The Data Basket is a view that shows all of the data sets or reports you've selected and to which you want access.

To access the Data Basket, click 🧺.

You can remove data sets and reports from the Data Basket by clicking on the relevant tiles, and then clicking **Remove from data basket**.



UX Reporting - Ideas tagged 'UXsession'

Description

This report contains an overview of the number of ideas tagged 'UXSession' in Aha!, and their respective status. It is based on a monthly, manual export from Aha!. The User Research and Usability - Tagging page on Confluence contains more information about the UXSession tag usage and the underlying Aha! reports.

Collibra Adoption Report

Description

The **Monthly Adoption Report** is a snapshot of the activity on a customer instance. Data are coming from Pendo and the visualization is done in Tableau.

This dashboard is both available for Production and Non-Production environments.

All Cloud customers and On-Premise customers (5.7 only) can have a monthly dashboard of the adoption (consumer, author) on their instances (production, non-production).

Documentation is available on Confluence.

This workbook is made of 3 main Reports:

- Adoption Report
- Retention Report
- Collibra Everywhere

Checking out your Data Basket

When you're ready to check out, click **Checkout Data Basket**. This starts the packaged Request Data Sets Access workflow, by which your request is approved or rejected.

When you check out the Data Basket, the Data Usage asset is renamed in the format YYYY-MM-DD #X, where X is a sequential number, for example **2019-12-16 #3**.

All your access requests are shown in Catalog, on the Access Requests page.

Request access to data sets and reports

To use the data referred to in data sets or reports, you need to request access to it.

Steps

1. Add assets to your basket by doing one of the following:

Searching	a. In Collibra Data Governance Center, search for the data set or report you need and click its name. » The asset page of the data set or report opens. b. Click **Add to Data Basket**.
Navigating	a. In the main menu, click ⣿, then 🛢 **Catalog**. » The Catalog Home opens. b. In the submenu, click **Data Sets** or **Reports**. c. If necessary, search for the data sets or reports that you want to access. d. Select the check boxes of the data sets or reports that you want to access. e. Above the tiles, click **Add to Data Basket**

   » A message at the upper-right indicates that the assets have been added to your basket.

2. Open your basket by clicking 🧺 on the main menu bar.
3. Review your basket.

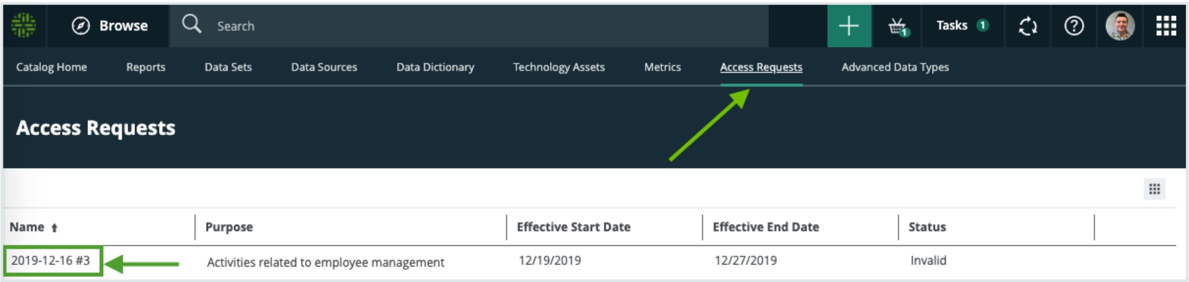
   To remove unnecessary assets, select them and click **Remove from data basket**.
4. Click **Checkout Data Basket**.

   » A dialog box appears.
5. Enter the required information and click **Submit**.
6. If the **Add Purpose to the Data Usage** dialog box appears, start typing and select the Purpose asset that describes the business use for which you are requesting access to the data sets or reports, and then click **Submit**.

   > Note This dialog box only appears if Collibra Data Privacy is installed.

What's next?

- A workflow starts to approve the request and to grant you access to the data.
- A Data Usage asset is created in Collibra Data Governance Center. You can view all your requests and their current status on the Access Requests page. For more information, see Requesting access to data.



Name	Purpose	Effective Start Date	Effective End Date	Status
2019-12-16 #3	Activities related to employee management	12/19/2019	12/27/2019	Invalid

Data Sources page

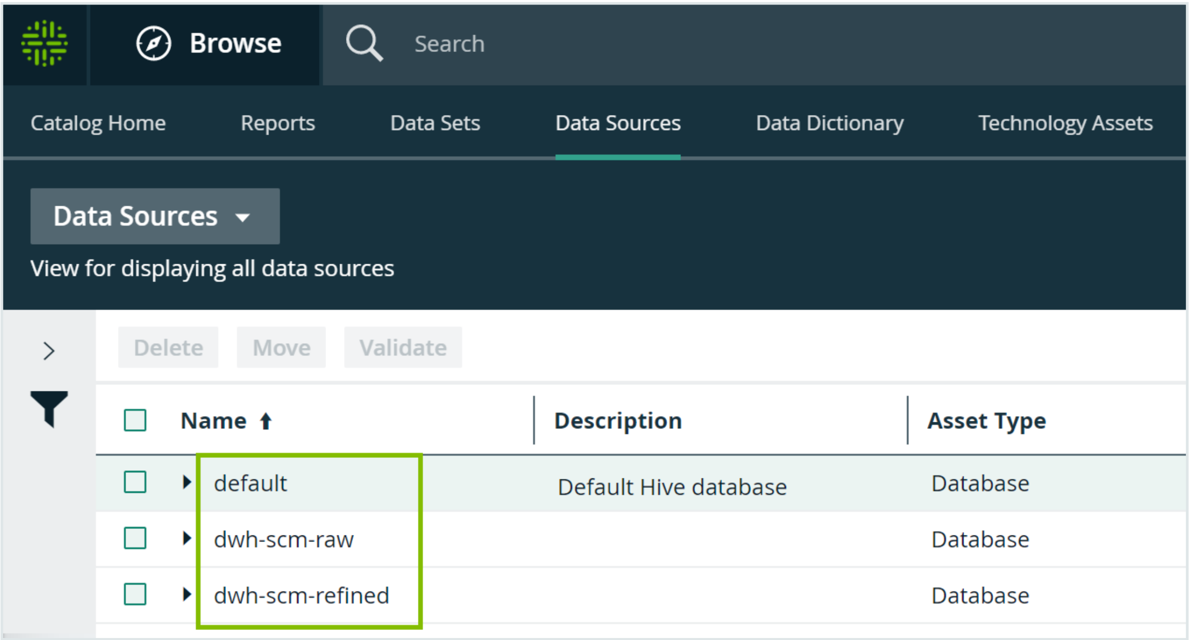
The **Data Sources** page is a view that shows all the data sources (in other words, all assets of asset type **Data Assets**) that were used in the creation of Data Catalog Data Sets.

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.

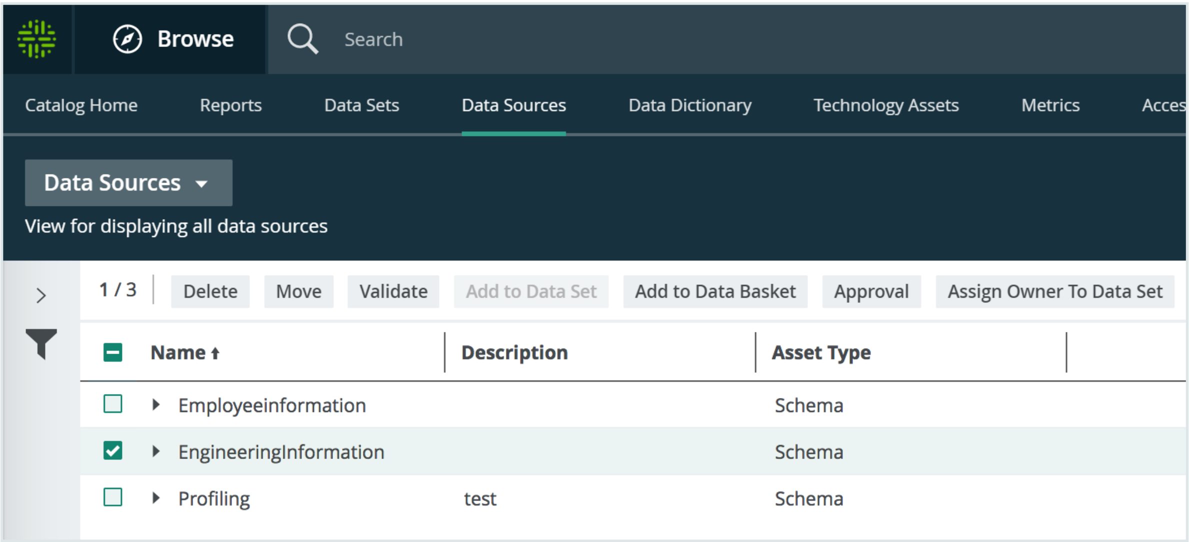
- Click an asset name to open the relevant asset page.



View for displaying all data sources

□	Name ↑	Description	Asset Type
□	default	Default Hive database	Database
□	dwh-scm-raw		Database
□	dwh-scm-refined		Database

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

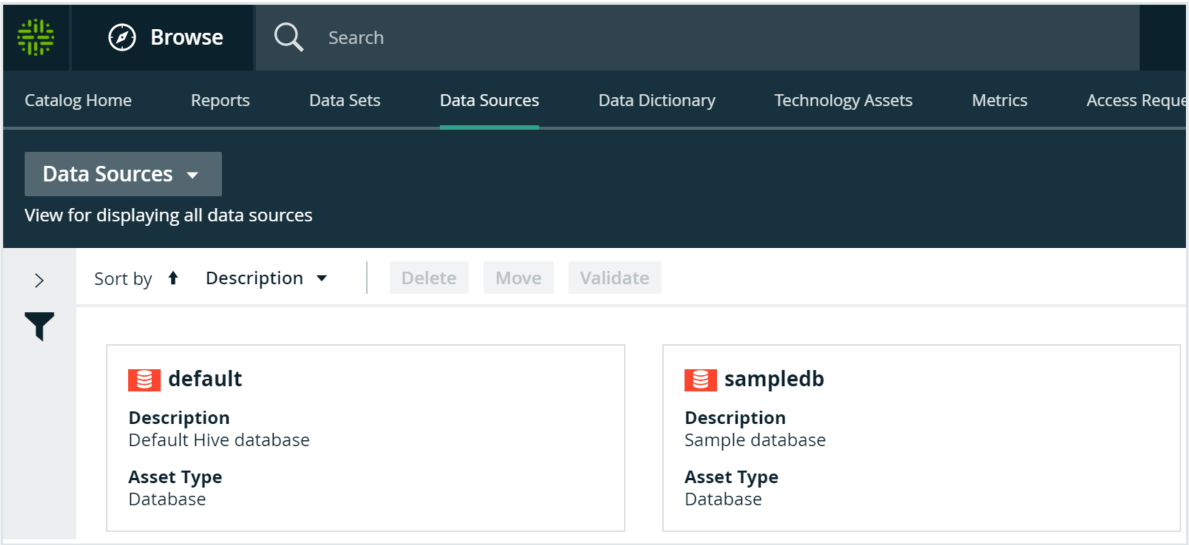


View for displaying all data sources

▬	Name ↑	Description	Asset Type
□	Employeeinformation		Schema
☑	EngineeringInformation		Schema
□	Profiling	test	Schema

Data sources in tile display mode

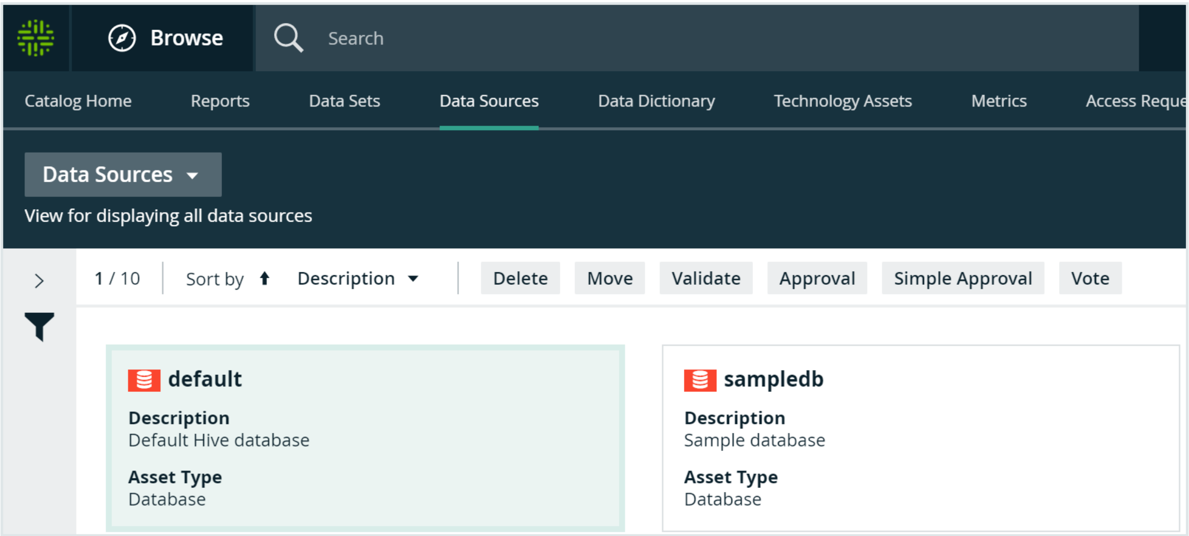
- Click an asset name to open the relevant asset page.



default
Description
Default Hive database
Asset Type
Database

sampledb
Description
Sample database
Asset Type
Database

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



default
Description
Default Hive database
Asset Type
Database

sampledb
Description
Sample database
Asset Type
Database

Data Dictionary page

The **Data Dictionary** page is a view that shows all assets of every data asset type in Collibra Data Governance Center.

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 Data Governance Center.

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

On this page, you can perform the following actions:

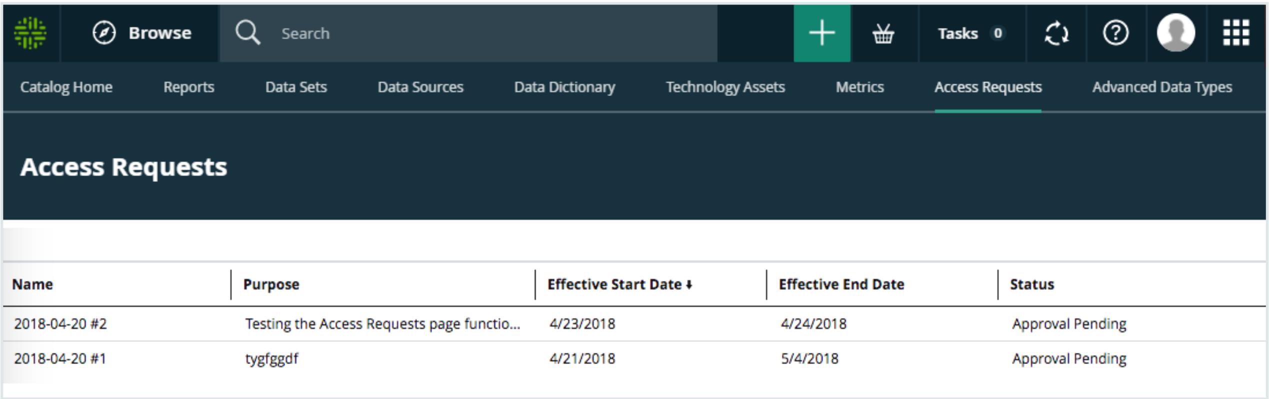
- Create views.
- Filter assets.
- Sort assets by name, description and asset type.
- Delete assets.
- Move assets to another domain.

Access Requests page

If you have requested access to one or more data sets, the Access Requests page allows you to view the status of your requests.

When you request access to a data set:

- an asset of the Data Usage type is created in the Data Usages domain in your community.
- the request Assets Access workflow is started.



Access Requests

Name	Purpose	Effective Start Date ↓	Effective End Date	Status
2018-04-20 #2	Testing the Access Requests page functio...	4/23/2018	4/24/2018	Approval Pending
2018-04-20 #1	tygfggdf	4/21/2018	5/4/2018	Approval Pending

The names of your requests are automatically generated with the date of your request. You can click the request name to open the asset page, which shows all the information relative to your request.

If you've requested access to many data sets, you can sort on any of the columns on the Access Requests page, to help you find a specific access request.

Advanced data types

When you profile data when registering a data source, Collibra Data Governance Center can detect some basic data types, such as numbers and text. Besides these basic data types, you can create your own advanced data types.

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

Data type detection

When you run a data profiling when registering a data source, Collibra Data Governance Center tries to detect the data type of each column.

1. Collibra DGC tries to match the fields of each column with every data type.
2. Collibra DGC remembers the matches for each field, also if a field has multiple matches.
3. Collibra DGC calculates the matching percentage of how many fields of the column match the same data type.
4. Collibra DGC 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 DGC Installation and Configuration Guide.

5. Collibra DGC 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.

Advanced data type management prerequisites

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

- Catalog role
- Advanced Data Type global permission

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. In the main menu, click ▦, then ⛁ **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.
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
<b>Examples</b>
<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>
</table>
</td></tr>
</table>

<table>
<tr><th>Option</th><th>Description</th></tr>
<tr><td></td><td>
<table>
<tr><th>Base data type</th><th>Field name</th><th>Patterns</th></tr>
<tr><td>Date and Time</td><td>Custom Date and Time</td><td><code>MM/dd/yyyy HH:mm:ss</code></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 <b>DateTimeFormatter</b> format. See the official Java documentation.</td></tr>
<tr><td>Time</td><td>Time pattern</td><td>List of time patterns using the <b>DateTimeFormatter</b> format. See the official Java documentation.</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><th>Option</th><th>Description</th></tr>
<tr><td></td><td>
<table>
<tr><th>Base data type</th><th>Field name</th><th>Description</th></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**.

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. In the main menu, click ⠿, then 🛢 **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:
<ul>
<li>Text</li>
<li>Geographical</li>
<li>True/False</li>
<li>Date</li>
<li>Time</li>
<li>Date and Time</li>
<li>Whole Number</li>
<li>Decimal Number</li>
<li>Array</li>
<li>N/A</li>
</ul>
<b>Examples</b>
<table>
<tr><th>Base data type</th><th>Field name</th><th>Patterns</th></tr>
<tr><td>Text</td><td>Email address</td><td><code>[a-z0-9]+[_a-z0-9\.-]*[a-z0-9]+@[a-z0-9-]+(\.[a-z0-9-]+)*(\.[a-z]{2,4})</code></td></tr>
<tr><td>Text</td><td>IP address</td><td><code>\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</code></td></tr>
<tr><td>Date</td><td>Custom Date</td><td><code>yyyy-MM-dd</code></td></tr>
<tr><td>Time</td><td>Custom Time</td><td><code>HH mm</code></td></tr>
</table>
</td></tr>
</table>

<table>
<tr><th>Option</th><th>Description</th></tr>
<tr><td></td><td>
<table>
<tr><th>Base data type</th><th>Field name</th><th>Patterns</th></tr>
<tr><td>Date and Time</td><td>Custom Date and Time</td><td><code>MM/dd/yyyy HH:mm:ss</code></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 <b>DateTimeFormatter</b> format. See the official Java documentation.</td></tr>
<tr><td>Time</td><td>Time pattern</td><td>List of time patterns using the <b>DateTimeFormatter</b> format. See the official Java documentation.</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><th>Option</th><th>Description</th></tr>
<tr><td></td><td>
<table>
<tr><th>Base data type</th><th>Field name</th><th>Description</th></tr>
<tr><td>Date and Time</td><td>Date and Time pattern</td><td>List of date and time patterns using the <b>DateTimeFormatter</b> 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 <b>True</b> value.
◦ List of values that are accepted as <b>False</b> 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**.

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. In the main menu, click ⠿, then 🛢 **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. In 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)**.

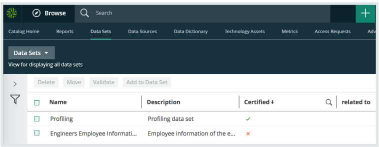
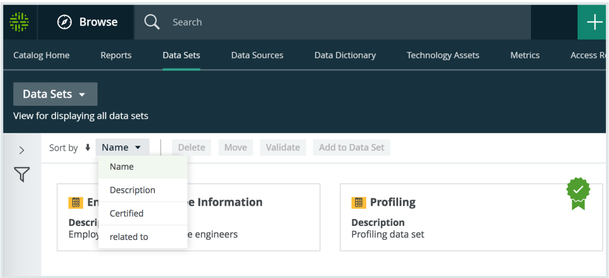
The data type attributes that contain the deleted advanced data type are reset to the base data type that was used for the advanced data type.

Sort Catalog submenu pages

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

Steps

1. In the main menu, click ▦, then 🛢 **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.  Browse Search Catalog Home Reports Data Sets Data Sources Data Dictionary Technology Assets Metrics Access Requests Data Sets View for displaying all data sets Delete Move Validate Add to Data Set Name Description Certified related to Profiling · Profiling data set · ✓ Engineers Employee Informati... · Employee information of the e... · ×	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.  Browse Search Catalog Home Reports Data Sets Data Sources Data Dictionary Technology Assets Metrics Access R Data Sets View for displaying all data sets Sort by Name Delete Move Validate Add to Data Set Name Description Certified related to En e Information Descri Employ e engineers Profiling Description Profiling data set

Chapter 2

Register a data source

Registering a data source makes metadata from that source available in Collibra DGC to create data sets that can then be used for creating reports and analyzing data. Optionally, Collibra DGC can perform data profilingdata profiling on the registered data and extract sample datasample data from it.

> Note If you are using a Collibra Data Intelligence Cloud environment with an on-premises Jobserver, they 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 5.7.13-0

About registering a data source

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

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 Data Governance Center. Collibra DGC 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 DGC 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 DGC 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 DGC to approve the terms and link them to more technical assets.

Naming convention

When you register a data source, Collibra DGC 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.

Supported data sources for data source registration

Collibra Data Governance Center supports several databases to register as a data source. Depending on your data source, you can use Collibra-provided Catalog connector, or your own JDBC driver.

Your own JDBC drivers

For certain data sources, you can use your own JDBC driver. The following table contains the packaged data sources and versions that have been tested.

Data source	Tested versions	Support for profiling and sample data	JDBC driver version
Amazon Redshift	1.0	Yes	v. 1.1.13.1013
Cloudera Hive	5.10 - 5.14	No	Apache driver v. 1.2.1
Hortonworks Hive	2.5, 2.6	No	Apache driver v. 1.2.1
HP Vertica	7.0	Yes	v. 07.01.0200
IBM DB2	10.5	Yes	v. 4.9.78
MySQL	5.6, 5.7	Yes	v. 5.1.38
Oracle	11g, 12c	Yes	v. 12c
PostgreSQL	9.4, 9.5	Yes	v. 9.4.1207
Microsoft SQL Server	2014, 2016	Yes	v. 5.1.38 Note Only Microsoft drivers and drivers available via Collibra Marketplace are supported.

Data source	Tested versions	Support for profiling and sample data	JDBC driver version
Teradata	15.0, 16.20.07.01	Yes	No driver tested
Snowflake			

> Note We cannot guarantee that other data sources or driver versions work correctly. If you use a generic JDBC driver or an unsupported version, data ingestion, data profiling and sample data may not work as expected.

Authentication and permissions

Both ingestion and profiling (including sampling and advanced data type detection) rely on JDBC drivers to operate. Those drivers authenticate to the data sources as a user registered in that data source with specific permissions attached to the user profile in the data source.

To ingest a database without profiling, Data Catalog requires read access to the database metadata: description of schema, tables, columns, including some more complex properties such as the primary and foreign keys.

However, if you enable one or more profiling options, Data Catalog also requires the permission to read the full table. Which permissions are required exactly depends on the data source type, version and configuration. Additionally, they can also differ according to the provider and version of the JDBC driver. Most of the queries required to retrieve the information above are hidden by the driver. As a consequence, Collibra cannot give a exhaustive list.

> Note Collibra DGCsupports several authentication methods, including credentials, NTLM, CyberArk and Kerberos. If you are using a certified Collibra-provided driver on the Collibra Marketplace, you can also authenticate using Windows Authentication.

> Tip If you need more detailed information, we recommend to contact your JDBC driver provider.

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.

For example, you can create a schedule for LDAP synchronizations, Purge cycles or to automatically send emails using cron patterns. You can also use it to create a Cron map for your statistics.

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

> Warning If you create an invalid Cron pattern, Collibra Data Governance Center 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

Position	Field	Mandatory	Allowed values	Allowed special characters	Examples
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.
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.

Position	Field	Mandatory	Allowed values	Allowed special characters	Examples
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.
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.
?	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”.

<table>
<tr><th>Character</th><th>Description</th></tr>
<tr><td>L</td><td>
<p>Has different meaning in each of the two fields in which it is allowed.</p>
<blockquote>Example The value <code>L</code> in the <b>day-of-month field</b> 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.</blockquote>
<p>If you use <code>L</code> in the <b>day-of-week field</b> 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”.</p>
<p>When using the <code>L</code> option, it is important not to specify lists, or ranges of values, because you may get unexpected results.</p>
</td></tr>
<tr><td>W</td><td>
<p>Used to specify the weekday (Monday-Friday) nearest the given day.</p>
<blockquote>Example <code>15W</code> in the value for the day-of-month field, means the nearest weekday to the 15th of the month:</blockquote>
<ul>
<li>If the 15th is a Saturday, the trigger will fire on Friday the 14th.</li>
<li>If the 15th is a Sunday, the trigger will fire on Monday the 16th.</li>
<li>If the 15th is a Tuesday, then it will fire on Tuesday the 15th.</li>
</ul>
<p>However if you specify <code>1W</code> 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.</p>
<blockquote>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"*.</blockquote>
</td></tr>
<tr><td>#</td><td>
<p>Used to specify “the nth” XXX day of the month.</p>
<blockquote>Example <code>6#3</code> 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: <code>2#1</code> is the first Monday of the month and <code>4#5</code> is the fifth Wednesday of the month. Note that if you specify <code>#5</code> and there is not 5 of the given day-of-week in the month, then no firing will occur that month.</blockquote>
</td></tr>
</table>

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.

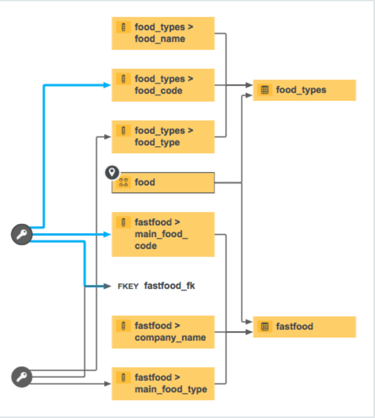
- The table with the primary key is referred to as the referenced table or parent table.
- The table with the foreign key is referred to as the child table.

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_types > food_name food_types > food_code food_types > food_type food fastfood > main_food_code FKEY fastfood_fk fastfood > company_name fastfood > main_food_type food_types fastfood	• **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.

Registering a data source via Jobserver

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

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 DGC. It makes metadata of the data source available in Collibra DGC. Note You can register a data source using a Collibra-provided driver or your own driver.
2	Ingestion	After registering a data source, Collibra DGC creates a **Physical Data Dictionary** domain and new assets of the type Schema, Table and Column, corresponding to the data in your data source.
3	Refresh a data source	Refreshing the schema of a registered data source updates the metadata of the data source in Collibra DGC. 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 typecalled 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. In the main menu, click ⠿, then ⛁ **Catalog**.
   » The Catalog Home opens.
2. In the main menu, click the **Create** (+) button.
   » 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 DGC 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 DGC.

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.
<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 Data Governance Center 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.

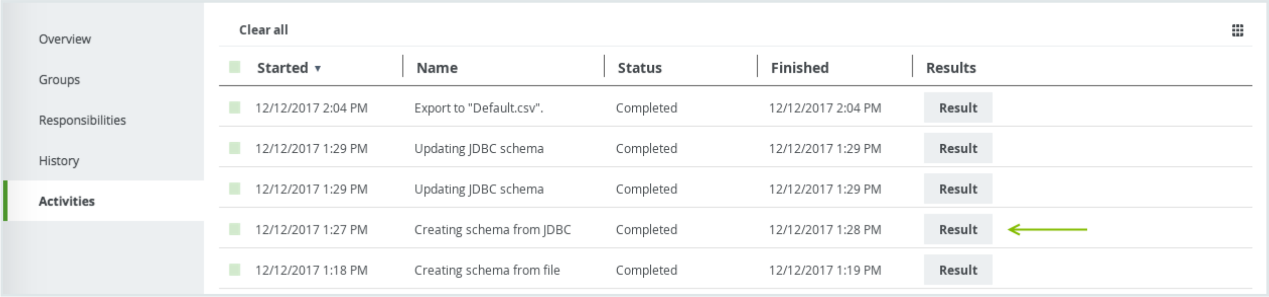
8. Click **Next**.
9. Select the data profiling options.

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 typecalled 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.



□ Started ▾	Name	Status	Finished	Results
□ 12/12/2017 2:04 PM	Export to "Default.csv".	Completed	12/12/2017 2:04 PM	Result
□ 12/12/2017 1:29 PM	Updating JDBC schema	Completed	12/12/2017 1:29 PM	Result
□ 12/12/2017 1:29 PM	Updating JDBC schema	Completed	12/12/2017 1:29 PM	Result
□ 12/12/2017 1:27 PM	Creating schema from JDBC	Completed	12/12/2017 1:28 PM	Result
□ 12/12/2017 1:18 PM	Creating schema from file	Completed	12/12/2017 1:19 PM	Result

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.

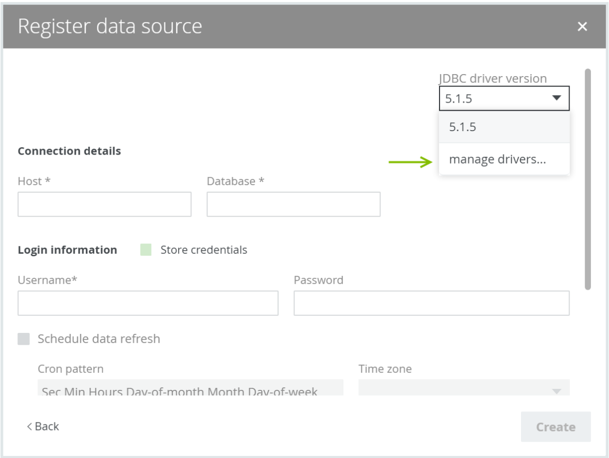
Steps

1. In the main menu, click ⠿, then 🛢 **Catalog**.
   » The Catalog Home opens.
2. In the main menu, click the **Create** (+) button.
   » 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 DGC 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 DGC.

   b. Click **Next**.

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



JDBC driver version: 5.1.5 ▾

- 5.1.5
- manage drivers...

Connection details

Host *

Database *

Login information □ Store credentials

Username*

Password

□ Schedule data refresh

Cron pattern: Sec Min Hours Day-of-month Month Day-of-week

Time zone

5. Do one of the following:
   - 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.
⭱ 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.
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 Data Governance Center.
- 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 5.7.13-0
- 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, consult the Databases supported versions section.
> - For the JDBC connection details of the various databases, consult the JDBC connection details section.

Steps

1. In the main menu, click ⠿, then 🛢 **Catalog**.
   » The Catalog Home opens.
2. In the main menu, click the **Create** (+) button.
   » 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 DGC 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 DGC.

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>&lt;Configuration properties&gt;</td><td>
<p>The connection properties as defined in your JDBC driver.</p>
<blockquote>Note For more information on the connection details of supported data sources, see JDBC connection details.</blockquote>
<p>If you want to use Kerberos authentication, you also need the following connection properties.</p>
<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>
<p>If you want to use NTLM authentication, you also need the following connection properties.</p>
<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>
<p>If you want to use CyberArk authentication, you need the following connection properties.</p>
</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 `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`</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 `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).</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>
</table>

Option	Description
Schedule data refresh	Enable or disable a schedule to automatically refresh the data registration.
Cron pattern	Schedule of the data refresh as a Quartz Cron pattern. Warning If you create an invalid Cron pattern, Collibra Data Governance Center 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.

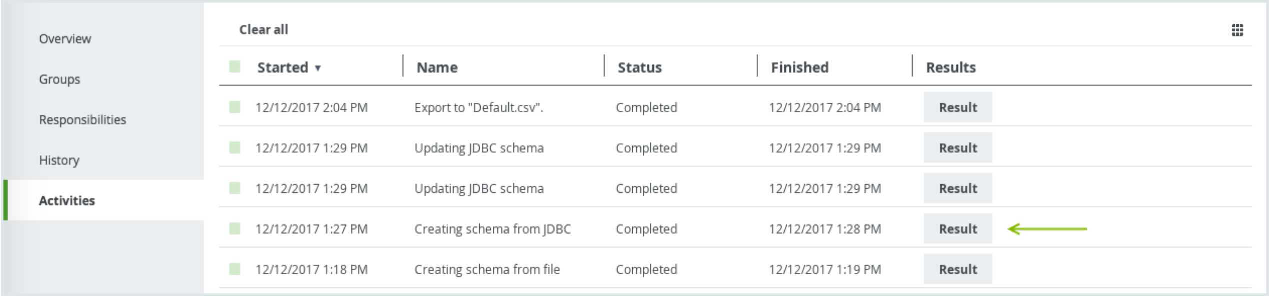
Option	Description
Store Data Profile	Option to perform data profiling on the registered data.
Detect advanced data types	Option to detect advanced data types in the data source.
Store Sample Data	Option to extract sample data from the registered data.

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 typecalled 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.



□	Started ▾	Name	Status	Finished	Results
□	12/12/2017 2:04 PM	Export to "Default.csv".	Completed	12/12/2017 2:04 PM	Result
□	12/12/2017 1:29 PM	Updating JDBC schema	Completed	12/12/2017 1:29 PM	Result
□	12/12/2017 1:29 PM	Updating JDBC schema	Completed	12/12/2017 1:29 PM	Result
□	12/12/2017 1:27 PM	Creating schema from JDBC	Completed	12/12/2017 1:28 PM	Result
□	12/12/2017 1:18 PM	Creating schema from file	Completed	12/12/2017 1:19 PM	Result

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, they 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 5.7.13-0

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 Data Governance Center.
- You have a resource role with the following resource permissions:
  - Asset > add
  - Attribute > add
  - Domain > add
  - Attachment > add

Steps

1. In the main menu, click ⠿, then 🛢 **Catalog**.

   » The Catalog Home opens.

   Or open any asset of the type Schema, Data Set, Table, Column or Tableau Server.
2. In the main menu, click the **Create** (+) button.

   » 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 DGC 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 DGC.

6. Click **Next**.
7. Select the data profiling options.

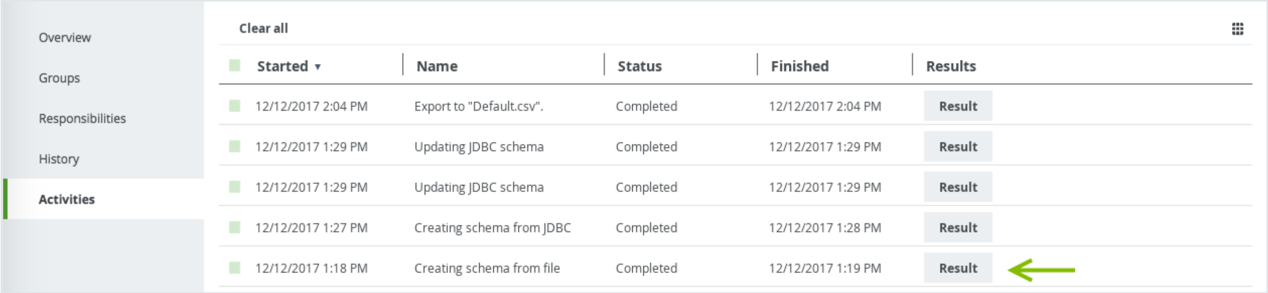
Option	Description
Store Data Profile	Option to perform data profiling on the registered data.
Detect advanced data types	Option to detect advanced data types in the data source.
Store Sample Data	Option to extract sample data from the registered data.

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 typecalled 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.



Started ▾	Name	Status	Finished	Results
12/12/2017 2:04 PM	Export to "Default.csv".	Completed	12/12/2017 2:04 PM	Result
12/12/2017 1:29 PM	Updating JDBC schema	Completed	12/12/2017 1:29 PM	Result
12/12/2017 1:29 PM	Updating JDBC schema	Completed	12/12/2017 1:29 PM	Result
12/12/2017 1:27 PM	Creating schema from JDBC	Completed	12/12/2017 1:28 PM	Result
12/12/2017 1:18 PM	Creating schema from file	Completed	12/12/2017 1:19 PM	Result

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 DGC 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, they 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 5.7.13-0

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 Data Governance Center.
- You have a resource role with the following resource permissions:
  - Asset > add
  - Attribute > add
  - Domain > add
  - Attachment > add

Steps

1. In the main menu, click ▦, then 🛢 **Catalog**.
   » The Catalog Home opens.

   Or open any asset of the type Schema, Data Set, Table, Column or Tableau Server.
2. In the main menu, click the **Create** (+) button.
   » 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 DGC 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 DGC.

6. Click **Next**.

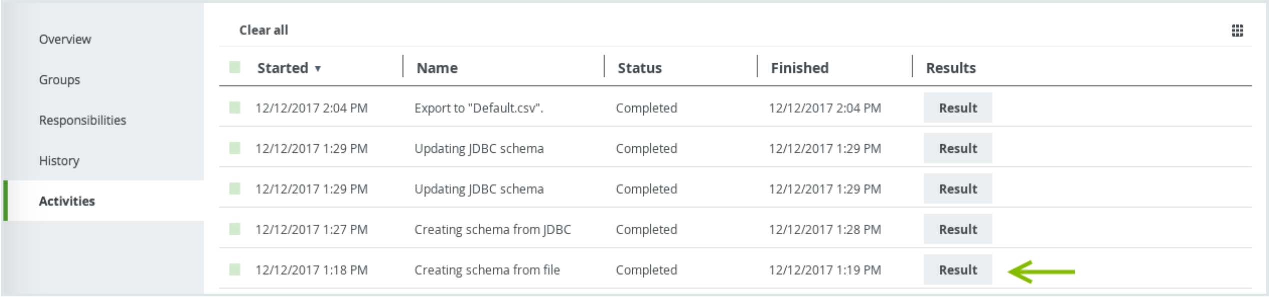
7. Select the data profiling options.

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 typecalled 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.



Started ▾	Name	Status	Finished	Results
12/12/2017 2:04 PM	Export to "Default.csv".	Completed	12/12/2017 2:04 PM	Result
12/12/2017 1:29 PM	Updating JDBC schema	Completed	12/12/2017 1:29 PM	Result
12/12/2017 1:29 PM	Updating JDBC schema	Completed	12/12/2017 1:29 PM	Result
12/12/2017 1:27 PM	Creating schema from JDBC	Completed	12/12/2017 1:28 PM	Result
12/12/2017 1:18 PM	Creating schema from file	Completed	12/12/2017 1:19 PM	Result

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 DGC 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.

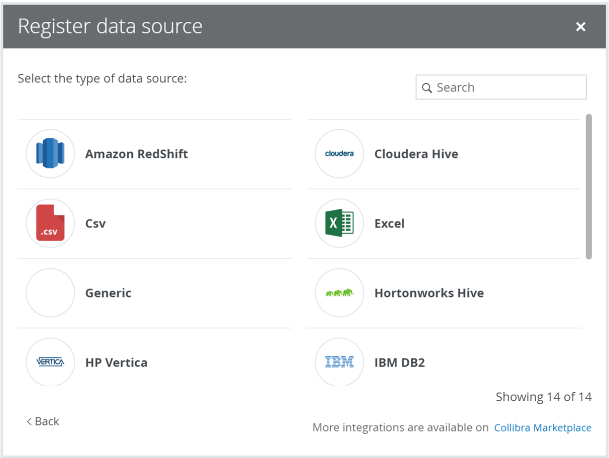
This operation should only be executed by your database administrator.

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 Data Governance Center.
- You have a resource role with the following resource permissions on the **Schema** community:
  - Asset > add
  - Attribute > add
  - Domain > add
  - Attachment > add

Steps

1. In the main menu, click ▦, then 🛢 **Catalog**.
   » The Catalog Home opens.
2. In the main menu, click the **Create** (+) button.
   » 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.



Register data source ×

Select the type of data source:

Search

Amazon RedShift

Cloudera Hive

Csv

Excel

Generic

Hortonworks Hive

HP Vertica

IBM DB2

Showing 14 of 14

‹ Back

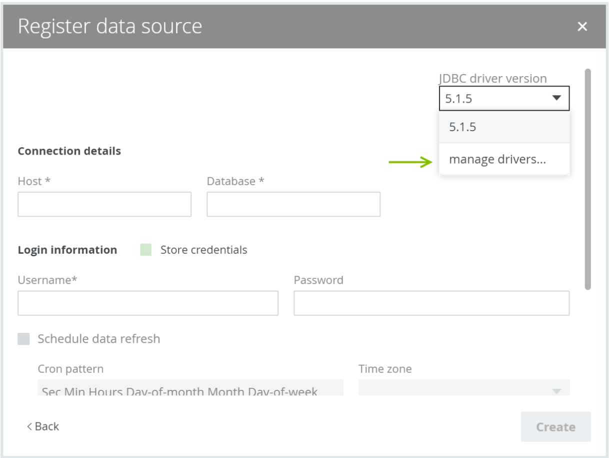
More integrations are available on Collibra Marketplace

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 DGC 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 DGC.

   b. Click **Next**.

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



6. Do one of the following:
   - 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.
⤒ 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**.
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
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
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

Label	Property	Mandatory
CyberArk query	cyberarkQuery	Yes

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.

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.

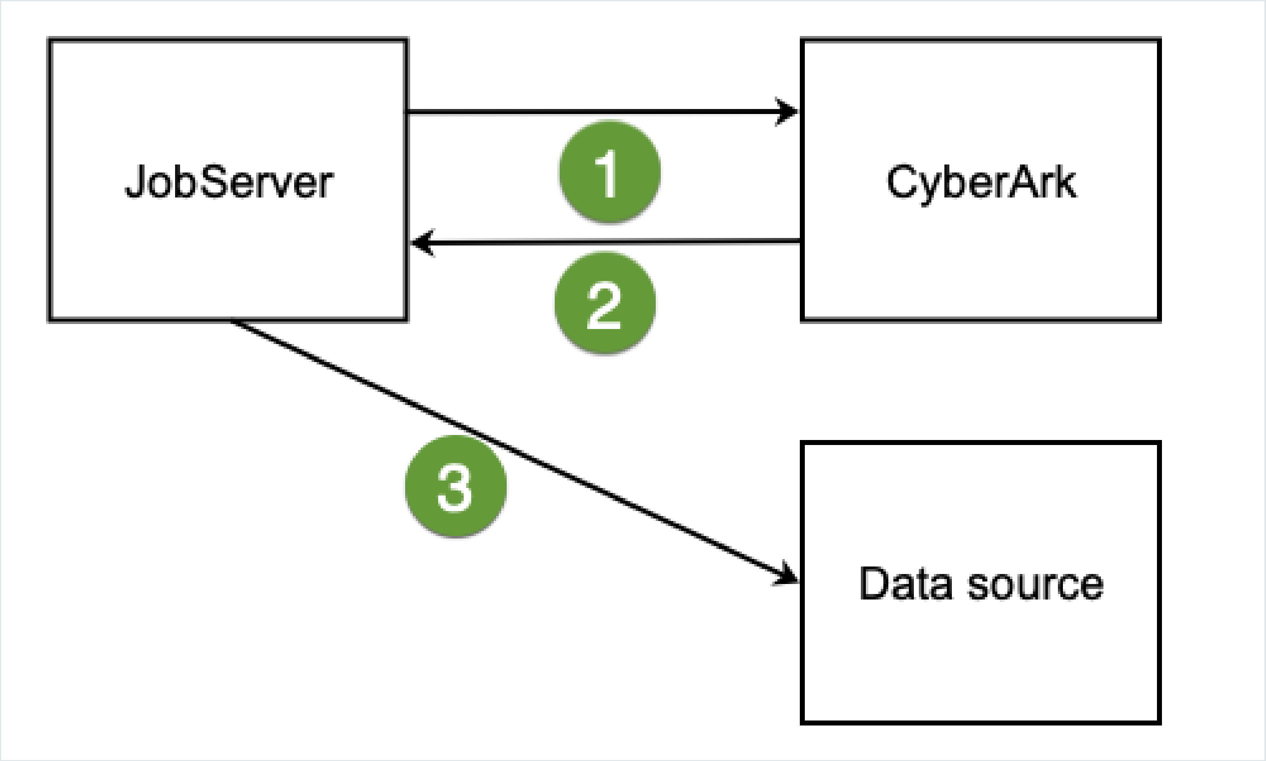
Setting up CyberArk authentication

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 DGC 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

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

Label	Property	Description	Mandatory
CyberArk query	cyberarkQuery	The CyberArk query. This query should be provided by your network or system administrator.	Yes

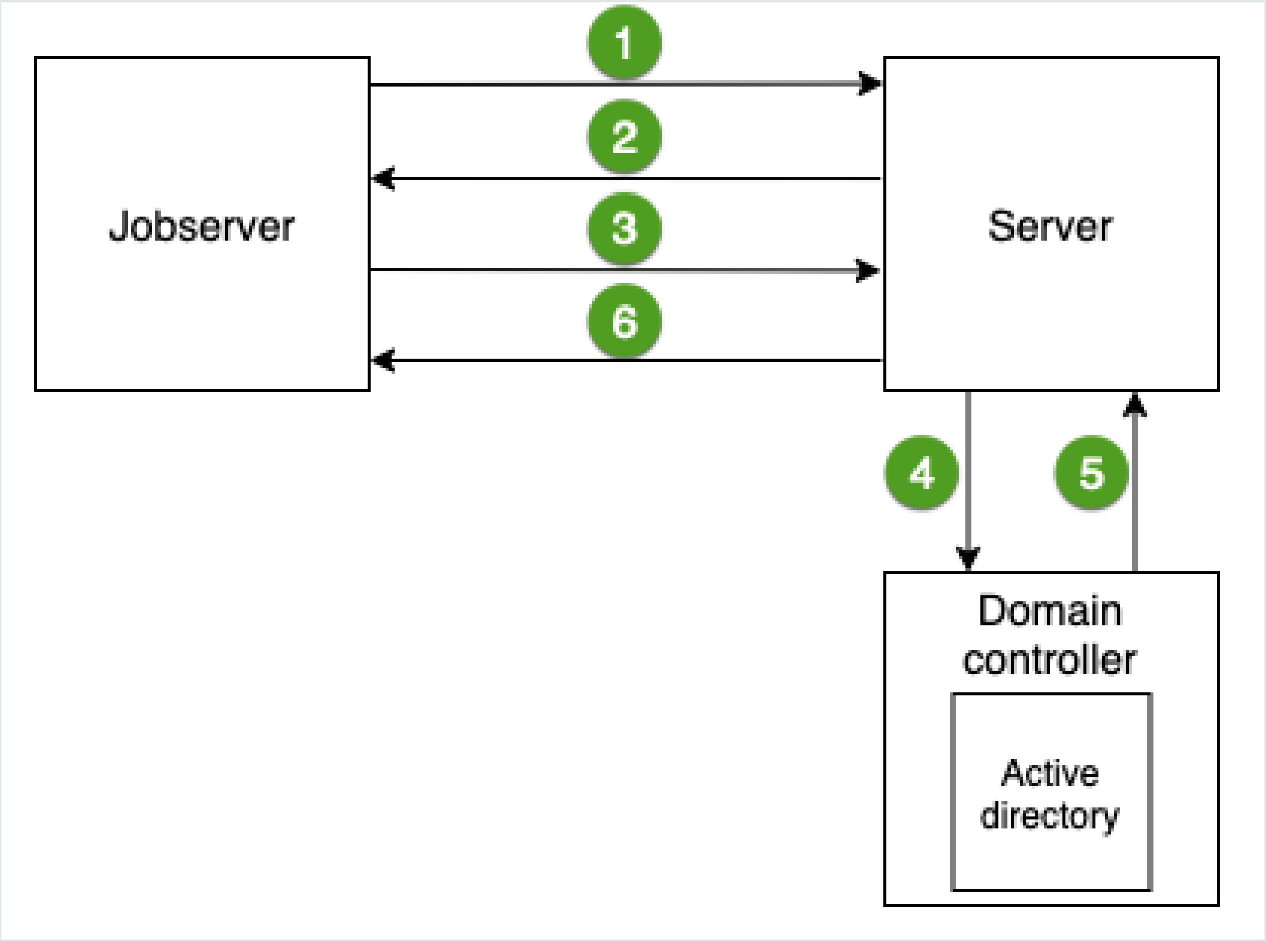
NTLM authentication

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

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>
  }
```

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*.

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. In 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.

About refreshing a schema

Refreshing a schema is the process of updating the metadata of a registered data source in Collibra Data Governance Center.

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.

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 Data Governance Center.
- 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 5.7.13-0
- 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. In the main menu, click ⠿, then ⛁ **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 Data Governance Center 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.

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

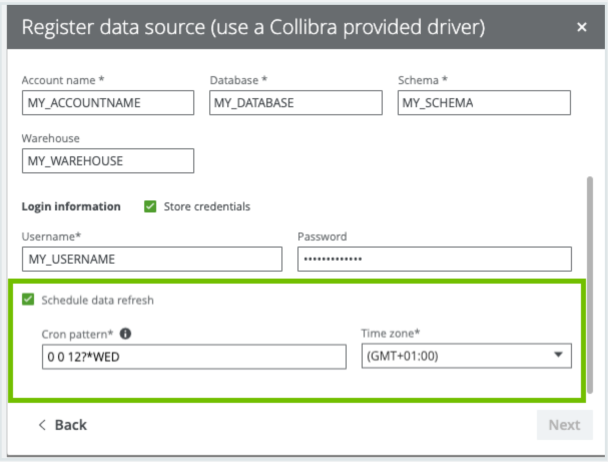
  - Domain: add
  - Attachment: add

> 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 Data Governance Center, 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.



Register data source (use a Collibra provided driver)

Account name *: MY_ACCOUNTNAME

Database *: MY_DATABASE

Schema *: MY_SCHEMA

Warehouse: MY_WAREHOUSE

Login information ☑ Store credentials

Username*: MY_USERNAME

Password: ••••••••••••••

☑ Schedule data refresh

Cron pattern*: 0 0 12?*WED

Time zone*: (GMT+01:00)

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. In the main menu, click ⊞, then 🛢 **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 Data Governance Center 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**.

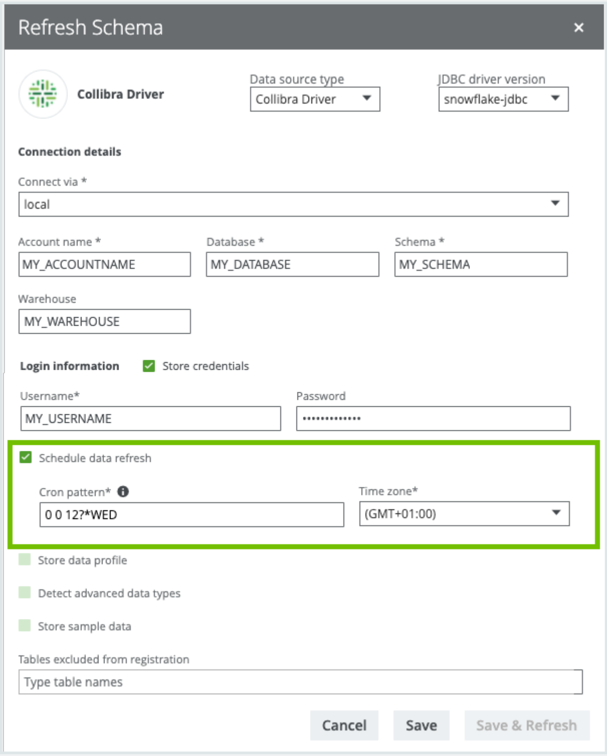
3. In the **Login information** section, check **Store credentials** and enter the username and password you use to access your data source.
   » Your credentials are used to automatically connect to your data source and refresh the metadata in Collibra Data Governance Center.
4. Select **Schedule data 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 Data Governance Center 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.



Data source type: Collibra Driver

JDBC driver version: snowflake-jdbc

Connection details

Connect via *: local

Account name *: MY_ACCOUNTNAME

Database *: MY_DATABASE

Schema *: MY_SCHEMA

Warehouse: MY_WAREHOUSE

Login information ☑ Store credentials

Username*: MY_USERNAME

Password: •••••••••••••

☑ Schedule data refresh

Cron pattern*: 0 0 12?*WED

Time zone*: (GMT+01:00)

□ Store data profile

□ Detect advanced data types

□ Store sample data

Tables excluded from registration: Type table names

Sample data

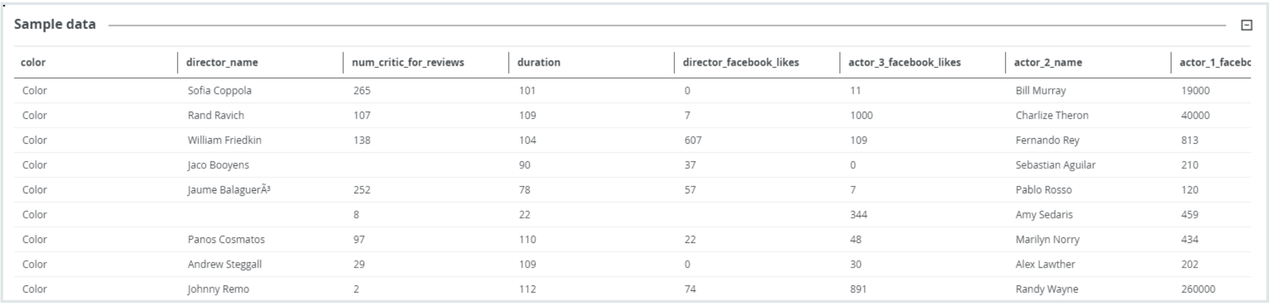
Sample data is a set of representative, random data collected from a registered data source and shown in Data Catalog. This sample data is also used for data classification via the Data Classification Platform.

Sample data when using Jobserver

When working via Jobserver, you can create sample data by:

- registering a data source and choosing to store sample data.
- importing sample data via the Catalog API.

You can find the sample data by opening a table of the registered data source and clicking **Sample data** in the tab pane.



Sample data

color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebc
Color	Sofia Coppola	265	101	0	11	Bill Murray	19000
Color	Rand Ravich	107	109	7	1000	Charlize Theron	40000
Color	William Friedkin	138	104	607	109	Fernando Rey	813
Color	Jaco Booyens		90	37	0	Sebastian Aguilar	210
Color	Jaume BalaguerÃ³	252	78	57	7	Pablo Rosso	120
Color		8	22		344	Amy Sedaris	459
Color	Panos Cosmatos	97	110	22	48	Marilyn Norry	434
Color	Andrew Steggall	29	109	0	30	Alex Lawther	202
Color	Johnny Remo	2	112	74	891	Randy Wayne	260000

To delete sample data, refresh the related schema without storing the sample data. This action will remove all previously stored sample data of all tables in the schema.

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 data by:

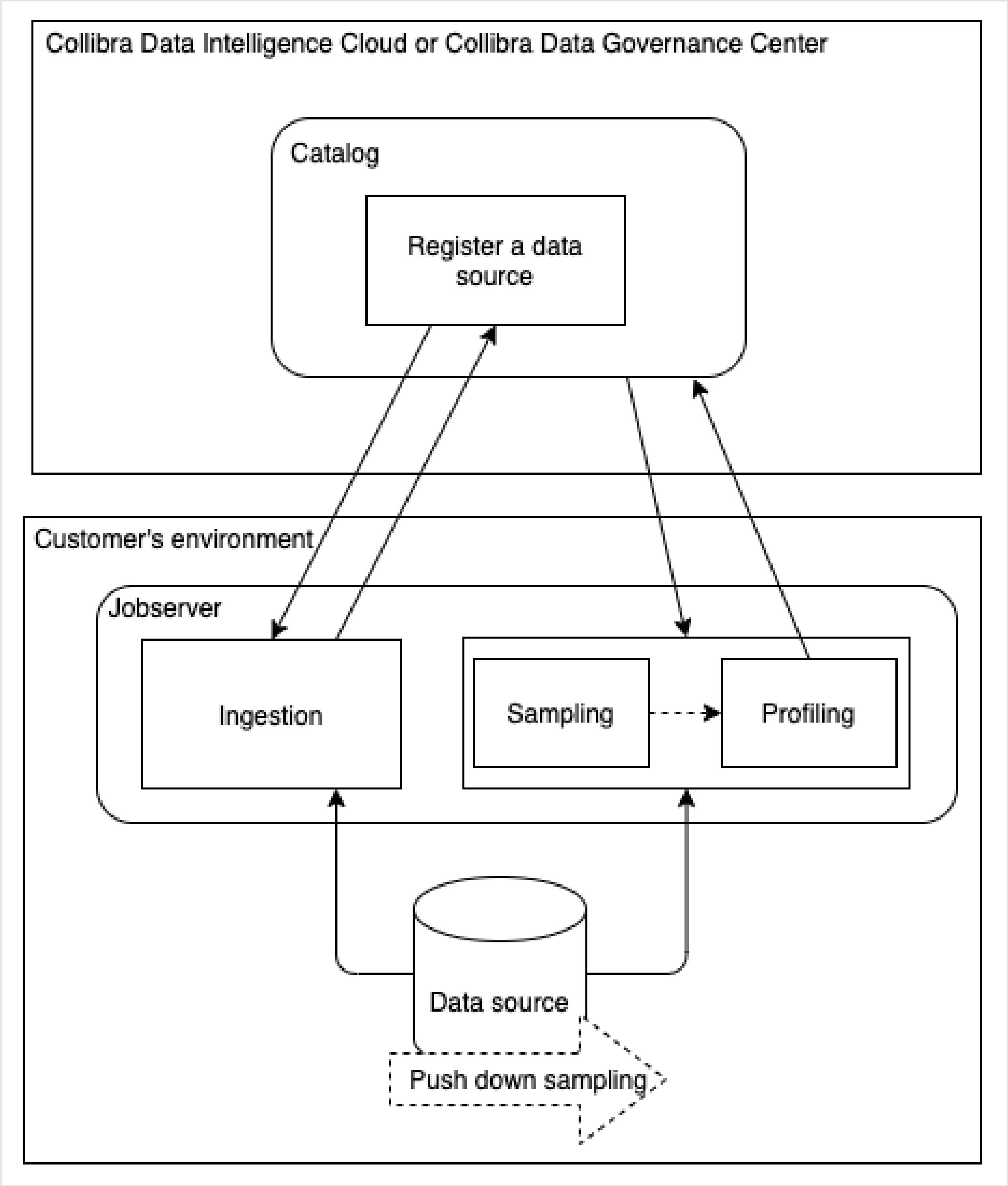
- Registering a data source and choosing to create profiling data.
- Importing profiling data via the Catalog API.

You can find the profiling information on the asset page of a table or a column, by clicking **Data Profiling** in the tab pane.

Profiling process

When you register a data source, Data Catalog triggers the ingestion process via Jobserver. By default, the complete data set is transferred to the Jobserver, which then creates a sample based on your data source. Jobserver profiles the sample and sends the result to Data Catalog.

You enable the **Anonymize data** option to hash or remove profiling information that can be considered sensitive.



Profiling sample

To create a data profile, Data Catalog uses a representative sample of the data.

> Note   This profiling sample is not the same as the sample available in **Sample data**.

If you register a data source via Jobserver, the profiling sample is created when you register the data source.

- If you use Jobserver without push down sampling, the complete data set is transferred to the Jobserver, which then creates the profiling sample based on your data source. The sample 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 profiling sample and sends it to Data Catalog.
  The data source creates the sample from randomly selected data and transfers it to the Jobserver. If the cache storage is reached, the process stops. Because the data source already created the sample randomly, the omitted data can be ignored without lowering the representativeness of the sample.

  > Warning   Push down sampling is only available for specific data sources.

Using push down sampling or partial scan

Push down sampling means that the task of creating the data sample is delegated to the data source itself. In Edge, push down sampling is called partial scan.

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

Push down sampling drastically increases the performance of sampling.

Enable push down sampling

Push down sampling is not used by default. 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 sample of 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 as a sample. 2. Select **Store Data Profile** and, optionally, **Store Sample Data** to profile via Jobserver.

Data anonymization via Jobserver

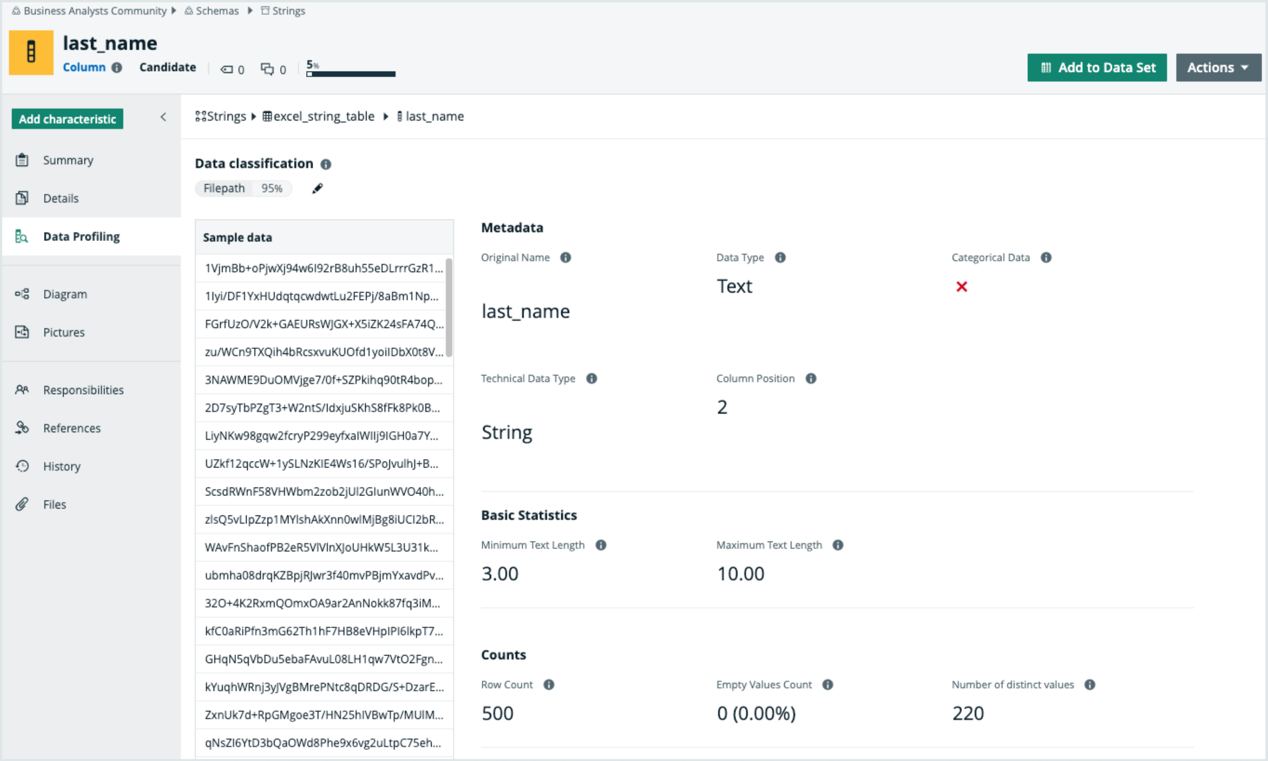
To ensure that sensitive data is not stored in the cloud, you can enable the Anonymize data option in Collibra Console.

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.

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 has been wrongfully anonymized. To solve this, you can manually modify the column's data type and profile again.

Example You enabled the Anonymize data option in Collibra Console and 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:



Data classification

Filepath 95%

Sample data
1VjmBb+oPjwXj94w6I92rB8uh55eDLrrrGzR1...
1lyi/DF1YxHUdqtqcwdwtLu2FEPj/8aBm1Np...
FGrfUzO/V2k+GAEURsWJGX+X5iZK24sFA74Q...
zu/WCn9TXQih4bRcsxvuKUOfd1yoilDbX0t8V...
3NAWME9DuOMVjge7/0f+SZPkihq90tR4bop...
2D7syTbPZgT3+W2ntS/IdxjuSKhS8fFk8Pk0B...
LiyNKw98gqw2fcryP299eyfxalWIlj9IGH0a7Y...
UZkf12qccW+1ySLNzKIE4Ws16/SPoJvulhJ+B...
ScsdRWnF58VHWbm2zob2jUl2GlunWVO40h...
zlsQ5vLIpZzp1MYlshAkXnn0wlMjBg8iUCI2bR...
WAvFnShaofPB2eR5VlVlnXJoUHkW5L3U31k...
ubmha08drqKZBpjRJwr3f40mvPBjmYxavdPv...
32O+4K2RxmQOmxOA9ar2AnNokk87fq3iM...
kfC0aRiPfn3mG62Th1hF7HB8eVHpIPI6lkpT7...
GHqN5qVbDu5ebaFAvuL08LH1qw7VtO2Fgn...
kYuqhWRnj3yJVgBMrePNtc8qDRDG/S+DzarE...
ZxnUk7d+RpGMgoe3T/HN25hlVBwTp/MUlM...
qNsZI6YtD3bQaOWd8Phe9x6vg2uLtpC75eh...

Metadata

Original Name: last_name

Data Type: Text

Categorical Data: ✗

Technical Data Type: String

Column Position: 2

Basic Statistics

Minimum Text Length: 3.00

Maximum Text Length: 10.00

Counts

Row Count: 500

Empty Values Count: 0 (0.00%)

Number of distinct values: 220

Note Jobserver does not automatically anonymize your data. 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.

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.

Data profiling of a table

The **Data Profiling** section of a registered table displays the properties of each registered column.

The following list contains the default displayed columns in this table:

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

For more information about these columns and columns that you can add, see Data profiling information.

You can customize the table by clicking on 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 data • data distribution: numerical data and date and time data See also Data profiling charts.

> Note If you use the Jobserver to register a data source and you have enabled the Anonymize data option in Collibra Console,Collibra anonymizes data in Column assets that have data type Text and Geo. If you use Edge to register a data source, your data is automatically anonymized.

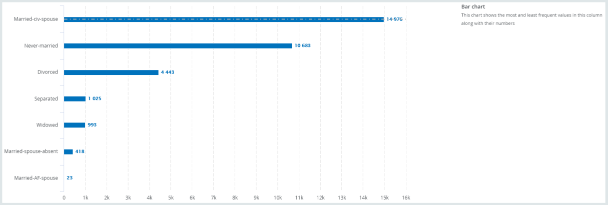
Data profiling charts

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

> Tip In each chart, you can zoom in by selecting the area of your preference. Click the **Reset zoom** button to return to the original chart view.

Bar chart

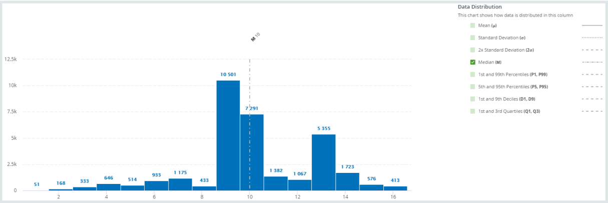
A bar chart is created when the data type is text. It displays the most and least frequent values of a column along with their number of occurrences.



Data distribution

The data distribution chart is created when the data type is numerical. It displays how the data is distributed.

In this chart, you can add extra information such as the mean, standard deviation and so on, by selecting the option at the right of the graph.



Data profiling information

If you create a data profile of registered data, data profiling information is generated.

The shown information depends on the profile options that you selected when you registered the data source and the profiling method, either via Jobserver or via Edge, that you used.

Column	Profiling option (Job-server)	Description
Original Name	No	Column name of the registered table.

Column	Profiling option (Job-server)	Description
Data Type	Store Data Profile If you want to have Advanced Data Type detected, select Detect advanced data types	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.
Row Count	Store Data Profile	The number of rows in the source.
Empty Values Count	Store Data Profile	The number of rows that are empty.
Number of distinct values	Store Data Profile	The number of unique values in the column.
Chart	Store Data Profile	This column displays whether a chart was generated (📊)or not (no icon available). If you hover over the icon, you see a preview of the chart. The chart type varies per data type. There are three charts available: • Frequency chart • Histogram that shows distribution • Probability distribution curve

Column	Profiling option (Job-server)	Description
Frequency	Store Data Profile	A bar chart showing frequency data.
Distribution - Histogram	Store Data Profile	A histogram showing the representation of the distribution of numerical data.
Distribution - Probability distribution curve	Store Data Profile	A curve showing the representation of the probability distribution of numerical data.
Technical Data Type	No	Data type of the column as defined in the source. This value can differ from the **Data Type** value.
Descriptive statistics (decile, percentile, quartiles)	Store Data Profile	The value of the calculated statistic of the registered data.
Categorical Data	Store Data Profile	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.
Categories	Store Data Profile	List of detected categories. This column has only values if the data is categorical.
Char octet Length	No	Maximum number of bytes in a character type's column.
Column Position	No	The index of the column in the source table.

Column	Profiling option (Jobserver)	Description
Is Auto Incremented	No	Indication whether the data in the column is auto-incremented or not.
Is Generated	No	Indication whether the data in the column is generated or not.
Is Nullable	No	Indication whether the column can store NULL values or not.
Is Primary Key	No	Indication whether the column is a primary key or not.
Maximum Text Length	Store Data Profile	The length of the longest text value in the column, including white spaces.
Maximum Value	Store Data Profile	The maximum value in the column.
Mean	Store Data Profile	The mean of all the values in the column, excluding empty rows.
Median	Store Data Profile	The median value of the column.
Minimum Text Length	Store Data Profile	The length of the shortest text value in the column.
Minimum Value	Store Data Profile	The minimum value in the column.
Mode	Store Data Profile	The value with the highest frequency for categorical data.

Column	Profiling option (Jobserver)	Description
Number Of Fractional Digits	No	The number of fractional digits.
Original Column Name	No	The column name as defined in the source.
Primary Key Name	No	The name of the primary key composed by the column.
Size	No	The size of the column in the table.
Standard Deviation	Store Data Profile	The statistical standard deviation of numeric values.
Variance	Store Data Profile	The statistical variance of numeric values.
Sample	Store Sample Data	A random sample of the data set that represents the entire data set.

Data profiling results

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 Data Governance Center.
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.

Modify the column data type of registered data

When Collibra Data Governance Center 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 Data type detection.

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

> 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, your data is 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. In the main menu, click ⠿, then 🛢 **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.

Chapter 5

Automatic Data Classification

When you register a data source in Collibra Data Governance Center, the process doesn't stop at ingestion. In order to unlock the full potential of Collibra DGC, the data needs to be contextualized: it needs to be classified and connected to other nodes in the Data Intelligence knowledge graph. Automatic Data Classification adds context to your data.

In the following sections, you will learn more about Collibra's Automatic Data Classification feature.

About Automatic Data Classification

In Collibra Data Governance Center, Automatic Data Classification is a feature that 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. In other words, data classification automatically (with no human input) assigns “class” values to individual columns of data to identify what kind of data is contained in that column. Examples of different data classes are “name”, “address”, “phone number” and “web browser”.

Why automatic data classification?

When you have ingested data in Data Catalog, the data classification process automatically identifies data structures within the data. As such, it takes less time to learn what kind of data you have ingested.

You can provide feedback by accepting or rejecting the suggested data classes. As a self-learning platform, the Data Classification Platform learns from the feedback, to improve the quality of future predictions.

Required permissions for Automatic Data Classification

The following table shows the required roles and permissions to use the Automatic Data Classification feature.

Action	Global Role	Global Permission	Resource Permission (*)	Required for classification via
Classify column	Catalog	Catalog	Column asset type's attributes (Asset > Attribute): • Add • Remove • Update Column asset type's data (Asset > Data): • View Samples	Data Classification Platform

Action	Global Role	Global Permission	Resource Permission (*)	Required for classification via
Classify table	Catalog	Catalog	Table asset type's attributes (Asset > Attribute): • Add • Remove • Update And the resource permissions to classify a column.	Data Classification Platform

Accept or reject a classification	Catalog	Catalog	Column asset type's attributes (Asset > Attribute):
			• Update
			Column asset type's data (Asset > Data):
			• View Samples
				Data Classification Platform
				and
				Edge

Add a user-defined classification	Catalog	Catalog > Advanced Data Type: • Add	Column asset type's attributes (Asset > Attribute): • Add • Update	Data Classification Platform and Edge

(*) As a user, you need a role that has the resource permission.

Packaged data classes for Automatic Data Classification

The following table shows the data classes that can be detected for columns by the Automatic Data Classification feature.

> Note This list can evolve over time. When you create a user-defined data class and the number of data samples exceeds a certain threshold, we will add this data class to our system.

Data Class shown in Collibra Data Governance Center	Content	Examples
City	Cities	• New York • Los Angeles • Chicago • Houston
Country	Countries	• Belgium • Lesotho • Dominica • Nigeria
Country code	Countries (short/-code)	• USA • ws • CAF • GIN
Credit card number	Credit card number	• 5602223068893246 • 1234-1234-1234-1234 • 3711-123456-12345 • 4123 5123 6123 7123

Data Class shown in Collibra Data Governance Center	Content	Examples
Currency code	Currency code	• zar • ARS • GBP • kes
Date	Date (only)	• 24 January 2004 • 11/21/1974 • 07-Nov-1982 • 11-08-22
Date time	Datetime	• 2018-08-29 20:25:25.0 • 2018-02-05 11:27:10.562 • 2017-10-10 05:34:16.216 • 2017-07-20 09:03:24.0
Education level	Education (level)	• Doctorate • post-secondary • Doctoral • Upper Secondary School
Email	Email	• pdawidas@storify.com • bmcentagartcf@china.com.cn • vgooms6x@barnesandnoble.com • dclatworthy9e@prweb.com
Employment status	Employment status	• Freelance • employed part time • office holder • Homemaker

Data Class shown in Collibra Data Governance Center	Content	Examples
Ethnicity	Race	• Hispanic • Latino • White • Asian
Filepath	Filepath	• E:\x9xOL\VB2ER_2E\ • F:LI\r_dWjux_\ • /u_2/tlk4q2/TwaYgn08A/GU/d-fp/z2vHk5iOW/Ael/M_ • wUmxr/ • BaG_8xxK_m/o1dq4luQ7A/z/kCQXGu.bin
First name	First Name	• Natasha • Manan • Rob • Wojciech
Full name	Full name (name + last name)	• lukas yang • Lukas, Yang • Amelia, Dalton • Dickens, Charles
Gender	Gender	• M • Male • woman • F
IBAN	IBAN - International Bank Account Number	• FO07 4910 6564 9863 03 • FR29 5218 3745 58B7 GH7N FYGZ Q50 • PS74 TSHR P22C D1DE 5OEB CRUG JRFW W • MK66 115I FYVV SOVS Y00

Data Class shown in Collibra Data Governance Center	Content	Examples
Internet domain	Web/internet domain	• slashdot.org • usa.gov • time.com • illinois.edu
IP address	IP address	• 80.206.17.108 • 3a6c:bb28:701a:5aaa:825c:4112:51ea:fadf • 255.139.66.168 • 241.65.195.63
ISBN	ISBN - International Standard Book Number (numeric commercial book identifier)	• 717393709-4 • 106115687-7 • 740540459-6 • 839089904-3
Language	Language	• Deccan • Kazakh • Zulu • Greek
Language code	Language code	• yor • HAU • CE • PS

Data Class shown in Collibra Data Governance Center	Content	Examples
Last name	Surnames / last name	• Burke • Lenaghan • Balmori • Balog
MAC Address	MAC address	• 4E-A0-23-78-53-50 • DE:D3:44:A7:7E:13 • a4-53-08-93-70-a4 • 83:4f-ca:43:93:32
Marital status	Marital status	• unmarried • Married • not-in-family • other-relative
Month	Month	• Mar • September • January • December
NDC Code	FDA NDC code - Food and Drug Administration's National Drug Code	• 55154-5876 • 68927-3491 • 58118-0623 • 55154-3939
Occupation	Occupation	• proofer • transit coach operator • forging machine tender • sports worker

Data Class shown in Collibra Data Governance Center	Content	Examples
Personal Email	Email	- f0ETKExihcHK@comcast.fr, - Ffz0Asl0To@comcast.com.br - jVgNF9v.ranlu@msn.com - u.L79@verizon.net
Phone number	Phone number	- 532-555-0185 - +1 212 555 3000 - 829-394-8017 - 973-491-8723
Religion	Religion	- Buddhist - Confucian - Protestant Anabaptist - Protestant Adventist
Routing Number (ABA)	Routing Number	- 058327451 - 675702815 - 805759224 - 305532637
SSN	SSN - Social security number	- 559-03-4491 - 284-34-1408 - 499-81-8467 - 576-17-9443
Street address	Address (first line) Street + number	- 4 Orinda Way - 61 Broadway

Data Class shown in Collibra Data Governance Center	Content	Examples
Time	Time	- 8:52 AM - 7:36 PM - 06:52 - 17:08:15
Title	Honorific	- Honorable - Rev. - Mr - Ms
UK Drivers License Number	Drivers License	- ENArq262033Xj32333 - ABzPt058106IA18871 - wklrS604032zb31785 - smeeI761300Rc02703
UK National Health Service (NHS) Number	Health Service	- 375 251 3810 - 537 649 5407 - 784 382 2399 - 534 293 9797
URL	URL	- www.sohu.com - http://www.googleweblight.com - https://twitter.com - ftp://mydomena.org/folder1 - http://www.-goolgle.com/search?query=my+query

Data Class shown in Collibra Data Governance Center	Content	Examples
US Adoption Taxpayer Identification Number (ATIN)	Tax Identifier	• 944-93-7219 • 930-93-3562 • 942-93-6471 • 932-93-3182
US Drivers License Number	Drivers License	• QP080580F • W5060999229 • Xm939887D • 70kQF62641
US Employer ID	Employer	• 41-0506939 • 91-0675223 • 43-2942382 • 77-4827140
US Individual Taxpayer Identification Number (ITIN)	Tax Identifier	• 915-78-5757 • 937-83-1696 • 929-75-9337 • 966-88-3886
US License Plate Number	License Plate	• 0HB8609 • 0qM6428 • 0VS0864 • 0lq7470

Data Class shown in Collibra Data Governance Center	Content	Examples
US State	US States	• Illinois • Indiana • Iowa • Kansas
US State code	US state code	• il • WI • ut • MT
UUID	GUID/UUID	• 0ee585a5-6bd3-4fde-9383-827095ed08f3 • 00000000-0000-0000-0000-000000031108 • 0a4281c9-0b6c-4095-b1b6-d8b417cfa952 • ffe27556-7c0d-4007-95c4-306633af3f14
Vehicle Identification Number (VIN)	Vehicle	• 4JGDF7DE1EA269698 • WDAPF3CC1B9465179 • WDBAB33A8EA076439
Web browser	Web browser	• Mozilla • Netscape • Chrome
Weekday	Weekday	• Wednesday • Fri • Wed • Tue

Calculation components for Automatic Data Classification

The following components are used to calculate data classes via Data Classification Platform or Edge:

Component	Purpose
Neural network	A machine learning tool that is continuously trained to identify linguistic patterns. Training data has been collected to have an initial set of patterns.
Regex matcher	A wide range of regular expressions to identify matching patterns. When the matched types in a column exceeds a certain threshold, the result is used in the final calculation of the data class.
Dictionary search	The classification is based on a dictionary attack. Multiple data classes only have a limited number of possible values, for example countries, country codes, currencies and days of week. These are all stored in a dictionary. The sample data is matched against these dictionaries.
Aggregator	The aggregator gathers the responses from the neural network, regex matcher and dictionary search and creates a final response based on underlying algorithms.

How does retraining work?

Currently, data classification on Edge does not retrain the classification model to improve future classification predictions.

In Data Classification Platform, the calculations are all based on the data samples received by the Data Classification Platform. Every time you accept a predicted data class, the sample data used to calculate that data class is added to the Data Classification Platform, to improve future data class predictions. See also Feedback on Automatic Data Classification.

> Example
> Assume you have a single column, C, containing sample data [a,b,c,d]. You classify this column, and the classification algorithm returns class x with confidence 70%. If you accept this class, then future columns containing the values [a,b,c,d] will be slightly more likely to be classified as x. In the future, a column with the same sample data may be classified as x with confidence 71%. The same can be said for a rejection of the above classification, with future results returning a confidence of, for example, 65%.

> Note In reality, changes will be more discrete and take more than one accepted or rejected data class to become effective.

Automatic Data Classification via the 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 Data Classification Platform predicts the data classes of selected columns and sends them back to Collibra Data Governance Center, where you confirm or reject the suggested data classes. The Data Classification Platform uses your feedback to retrain the platform and improve future data classifications.

> Warning If you want to use the Data Classification Platform, request it via your Collibra contact or create a support ticket. See also Data Classification Platform set-up.

Limitations

- Automatic data classification via the Data Classification Platform is a cloud service. Only if your on-premises environment can reach the cloud service, you can use data classification.

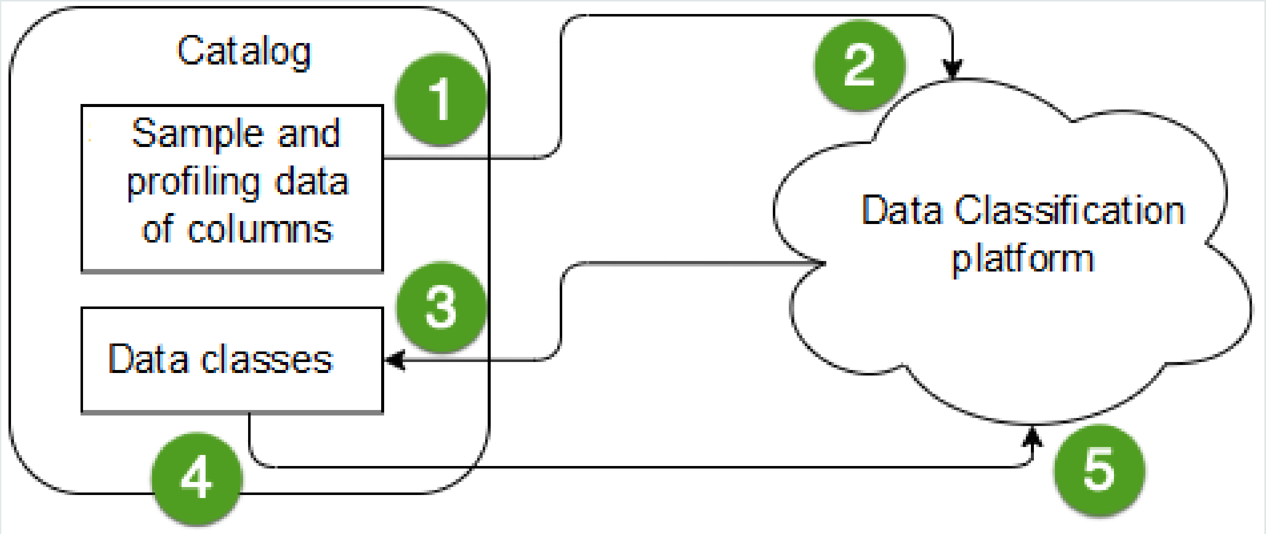
- Out-of-the-box, automatic data classification can predict several data classes. However, you can also create user-defined data classes to increase its prediction quality.
- The only supported language for data classes is English.
- Automatic Data Classification needs sample data and profiling data 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.

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

Automatic data classification flow via Data Classification Platform

In the following schema, you can see the different steps of an automatic data classification flow.


Catalog
Sample and profiling data of columns
Data classes
1
2
3
4
5
Data Classification platform

Step	Description
1	You select the columns that you want to classify and send their sample and profiling data to the Data Classification Platform. See Classify columns in a table
2	The Data Classification Platform predicts the data classes of the columns.
3	The Data Classification Platform sends the data classes to Collibra DGC.
4	You provide feedback by accepting or rejecting the predicted data class of each column or by adding your own new classes. The 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 Data Classification Platform . The Data Classification Platform stores your selections, along with the associated sample data, to retrain the classification model and improve future classification predictions.

Data Classification Platform set-up

If you want to start using the Data Classification Platform, request it via your Collibra contact or create a support ticket.

Requirements

You can only use automatic data classification if you comply with the following requirements.

- 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.

> Note Be aware that after you accept the predicted data classes, all sample data and profiling data is stored on the Data Classification Platform.

Location

We highly recommend to use a Data Classification Platform running in the same region as your Collibra DGC environment.

Currently, Collibra can provide Collibra Data Governance Center environments in Amazon AWS® regions in the following locations:

- United States
- European Union
- United Kingdom
- Canada
- Australia

Classify columns in a table

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

You can classify columns via a:

- 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 Automatic Data Classification for the DGC service.
- 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.
- Catalog experience is enabled in the DGC service configuration.

Via the Database asset page

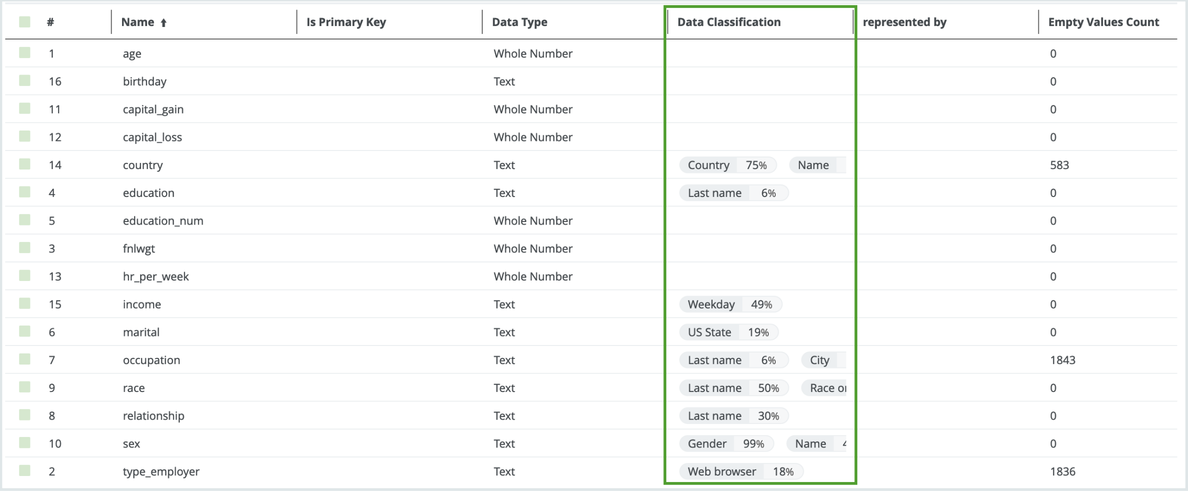
1. Open the Database asset that contains the tables and columns in the schema you want to classify.
   a. In the main menu, click ⠿, then 🛢 **Catalog**.

      » 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**.

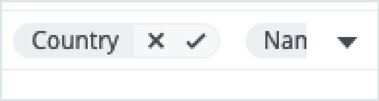
   » You can follow the status of the classification in Activities.
3. Open the database asset with the classified columns.
4. Add the Data Classification column to the table.

   » 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 4		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.

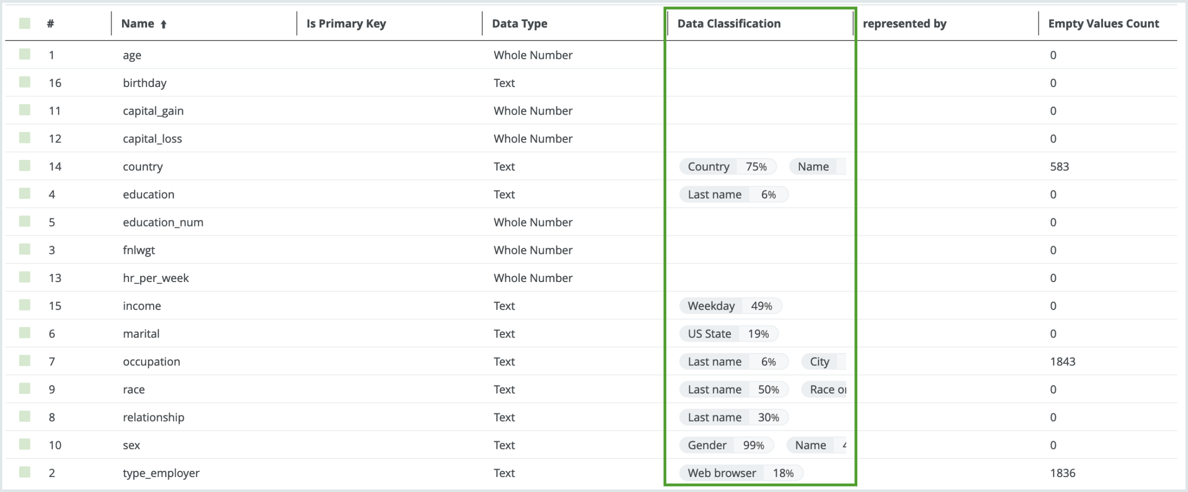
Via the Schema asset page

1. Open the Schema asset that contains the tables and columns that you want to classify.
   a. In the main menu, click ⠿, then 🛢 **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.



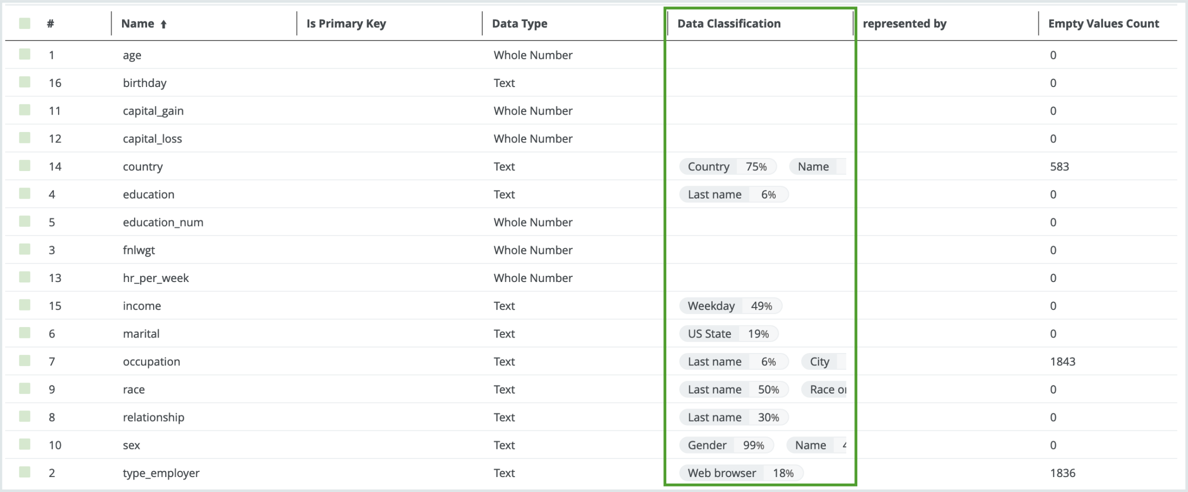
□	#	Name ↑	Is Primary Key	Data Type	Data Classification	represented by	Empty Values Count
□	1	age		Whole Number			0
□	16	birthday		Text			0
□	11	capital_gain		Whole Number			0
□	12	capital_loss		Whole Number			0
□	14	country		Text	Country 75% Name		583
□	4	education		Text	Last name 6%		0
□	5	education_num		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 4		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 4		0
□	2	type_employer		Text	Web browser 18%		1836

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

Feedback on Automatic Data Classification

Each time Collibra DGC predicts data classes for a column, you get the opportunity to send feedback by accepting or rejecting the data class, or by adding a user-defined data class.

To improve future predictions, it is really important to send this feedback.

> Note When using Edge, the feedback and user-defined classes are only stored, and not used to retrain the classification model.

Sending feedback

Sending feedback is the act of accepting or rejecting the data classes that are predicted.

- Reject data class: The data class is removed from the column.
  The Data Classification Platform classification model no longer uses the sample data.

- Accept data class: The data class is added to the column.
  The sample data is permanently added to the Data Classification Platform classification model to improve future data class predictions.

For the Data Classification Platform, accepting a data class is more valuable than rejecting, but in general, we recommend that you always send feedback for every prediction. Without your feedback, the classification model cannot be retrained.

Creating user-defined classes

When columns cannot be classified, you can create your own data classes.

- Avoid duplications. Always check the list of proposed classes before creating a new data class.
- Avoid vague data classes.
- Avoid mixed data classes and accept the best applicable one.

The Data Classification Platform uses this new information to retrain the platform and improve the predictions in the future.

Create a user-defined data class

If the Automatic Data Classification process cannot detect a data class in a column, you can classify the column yourself. If you are using the Data Classification Platform. your data class will be sent to the Data Classification Platform to improve its future predictive capabilities.

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, including these options:
  - Store Data Profile.
  - Store Sample Data. The more sample data (Data profiling section), the better the data class prediction.

> Note In order to improve Automatic Data Classification all Sample data and Profiling data is stored in the cloud.

- Data Catalog experience is enabled in the DGC service configuration.
  - » This will give you access to the improved Schema asset page.
- Catalog experience is enabled in the DGC service configuration.

Create a user-defined data class via the Table asset page

1. Find the table that contains the columns to classify.
2. At the bottom of the **Columns** section, click **See all**.
3. If not yet available, add the Data Classification column to the table.
4. In the row that you want to classify, double-click the **Data Classification** column.
5. Click the **Select** field.

   The list with existing data classes appears.
6. In the **Select** field, enter the 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.

7. Press `Escape` and click ✓.
   - » The new data class is automatically accepted.

Create a user-defined data class via the Column asset page

1. Find the column you want to classify.
2. In the tab pane, click **Data Profiling**.
3. In the **Data classification** section, click ✎.
4. Enter the new data class name and press `Enter`.

   You can add more data classes if applicable.

5. Click **Save**.
    - » The new data class is automatically accepted.

Chapter 6

Data Classification Dashboard

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

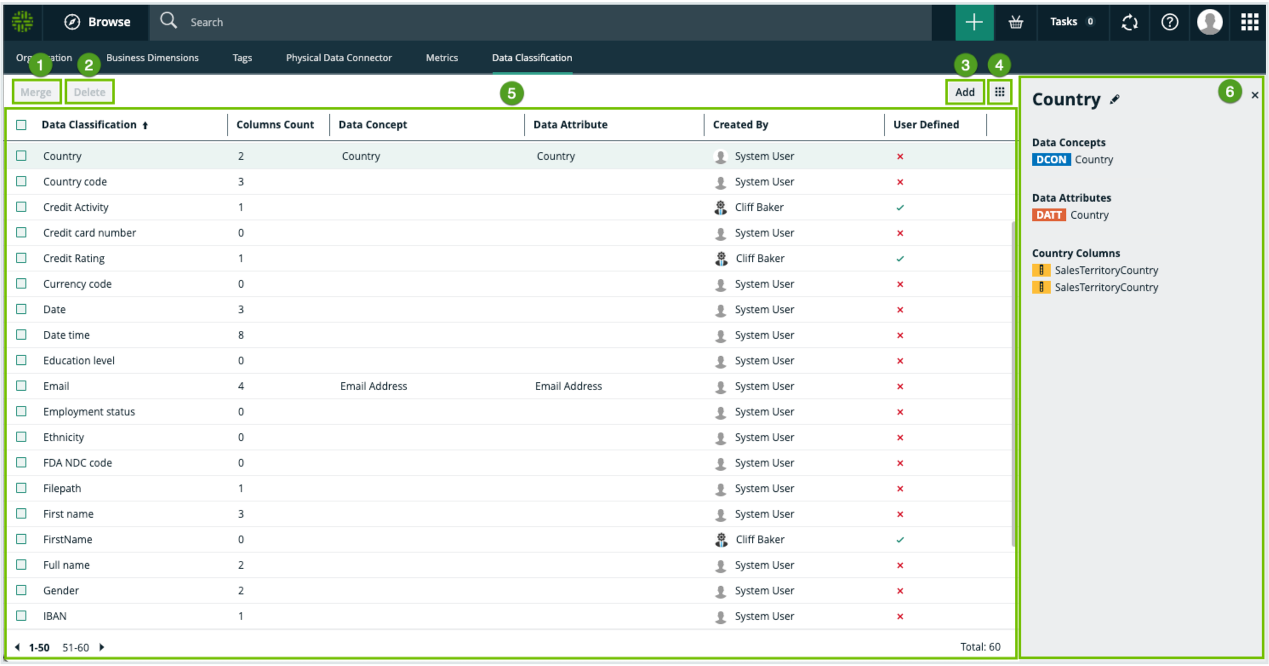
You can use the Data Classification Dashboard to:

- 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 DGC 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 Data Classification platform.

You access the dashboard via the **Data Classification** subpage in the Stewardship application.



☐	Data Classification ↑	Columns Count	Data Concept	Data Attribute	Created By	User Defined
☐	Country	2	Country	Country	System User	✗
☐	Country code	3			System User	✗
☐	Credit Activity	1			Cliff Baker	✓
☐	Credit card number	0			System User	✗
☐	Credit Rating	1			Cliff Baker	✓
☐	Currency code	0			System User	✗
☐	Date	3			System User	✗
☐	Date time	8			System User	✗
☐	Education level	0			System User	✗
☐	Email	4	Email Address	Email Address	System User	✗
☐	Employment status	0			System User	✗
☐	Ethnicity	0			System User	✗
☐	FDA NDC code	0			System User	✗
☐	Filepath	1			System User	✗
☐	First name	3			System User	✗
☐	FirstName	0			Cliff Baker	✓
☐	Full name	2			System User	✗
☐	Gender	2			System User	✗
☐	IBAN	1			System User	✗

◂ 1-50 51-60 ▸ Total: 60

Country

Data Concepts
DCON Country

Data Attributes
DATT Country

Country Columns
SalesTerritoryCountry
SalesTerritoryCountry

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 gives you a clear overview of the data class information of the selected data class.

View data class information

You can view data class information on the Classification Dashboard

Prerequisites

- You have configured Automatic Data Classification in Collibra Console.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

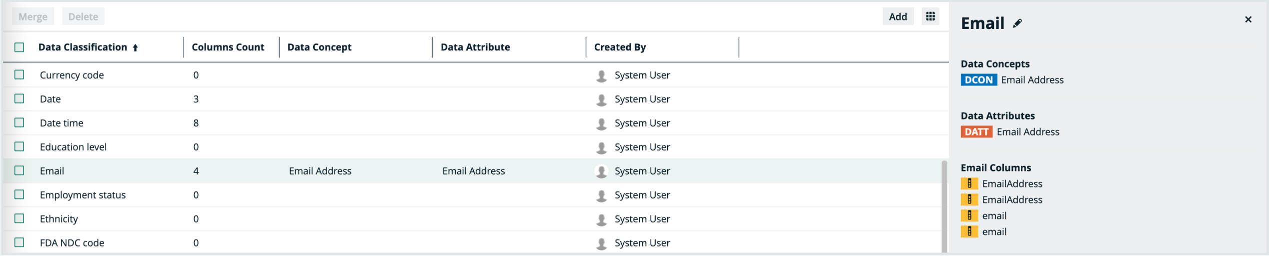
Steps

1. In 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.

Data Class side pane

The Data Class side pane gives you a clear overview of related 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.



☐ Data Classification ↑	Columns Count	Data Concept	Data Attribute	Created By
☐ Currency code	0			System User
☐ Date	3			System User
☐ Date time	8			System User
☐ Education level	0			System User
☐ Email	4	Email Address	Email Address	System User
☐ Employment status	0			System User
☐ Ethnicity	0			System User
☐ FDA NDC code	0			System User

Email ✎ ×

Data Concepts
DCON Email Address

Data Attributes
DATT Email Address

Email Columns
EmailAddress
EmailAddress
email
email

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 ✎ .

Attribute	Description
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 DGC contains a large number of packaged data classes, but if a certain data class is not available, you can add your own.

> Tip You can also manually create data classes for a specific Column asset.

Prerequisites

- You have configured Automatic Data Classification in Collibra Console.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

Steps

1. In 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.
5. Click **Create**.
   - » The classes are added and automatically sent to the Data Classification Platform.
6. Optionally you can link the new classes to a Data Concept or Data Attribute asset.
   1. In the **Data Concept** column, click ✎.
   2. Click in the **Select** field.
      - » The list with existing Data Concept assets appears.
   3. Select one or more Data Concept assets from the drop-down list and click ✓.
   4. 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 DGC 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 in Collibra Console.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

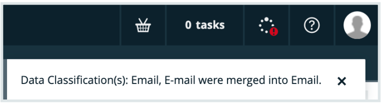
Steps

1. In 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**.



Data Classification(s): Email, E-mail were merged into Email. ×

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 in Collibra Console.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

Steps

1. In 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.

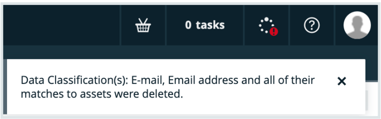
Prerequisites

- You have configured Automatic Data Classification in Collibra Console.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

Steps

1. In 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**.

   Data Classification(s): E-mail, Email address and all of their matches to assets were deleted.

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 in Collibra Console.
- You have the necessary permissions to classify tables and columns.
- You have registered a data source.

Steps

1. In 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

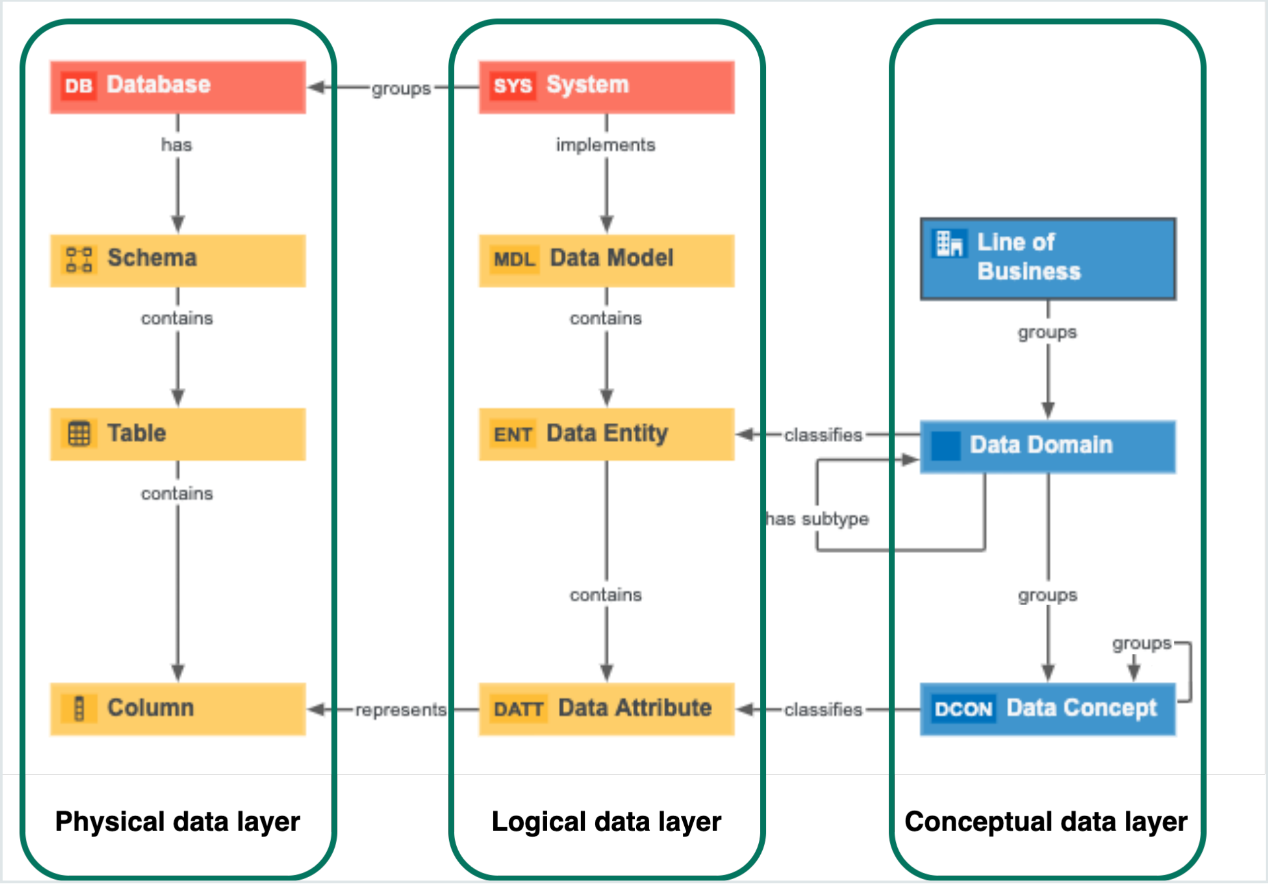
The Guided Data Stewardship operating model defines the structure of the information in Catalog. For this reason, the Guided Data Stewardship operating model is sometimes also referred to as the Data Catalog operating model.

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

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, or 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.

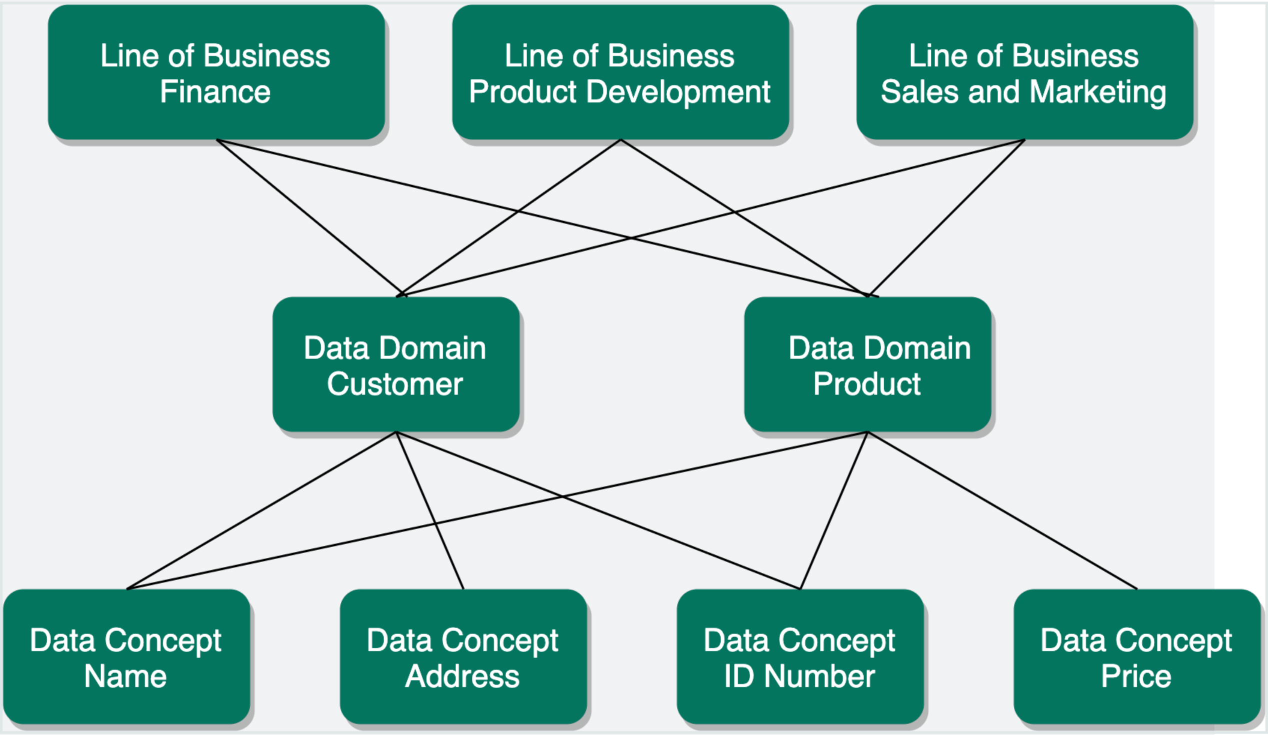
This graph-based approach, based on many-to-many relationships, makes the conceptual data layer more concise and flexible.

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

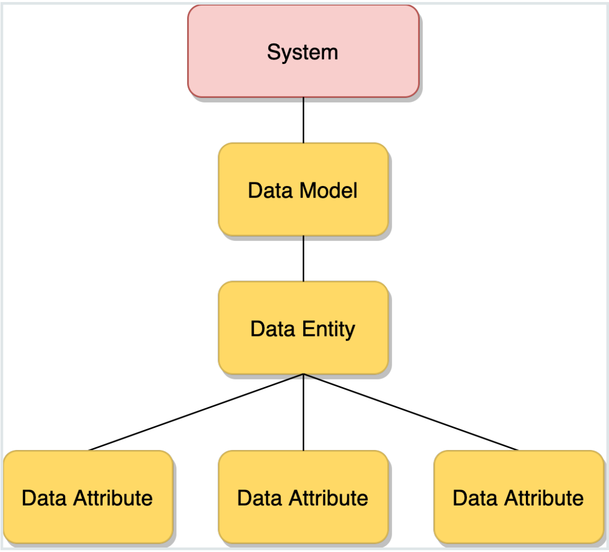
The logical data layer defines data structures within an organization's systems, whereas the conceptual data layer represents context-independent data structures within an organization.

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.

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 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 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.

Related to...	Via the relation type...	Description
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.

Related to...	Via the relation type...	Description
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.

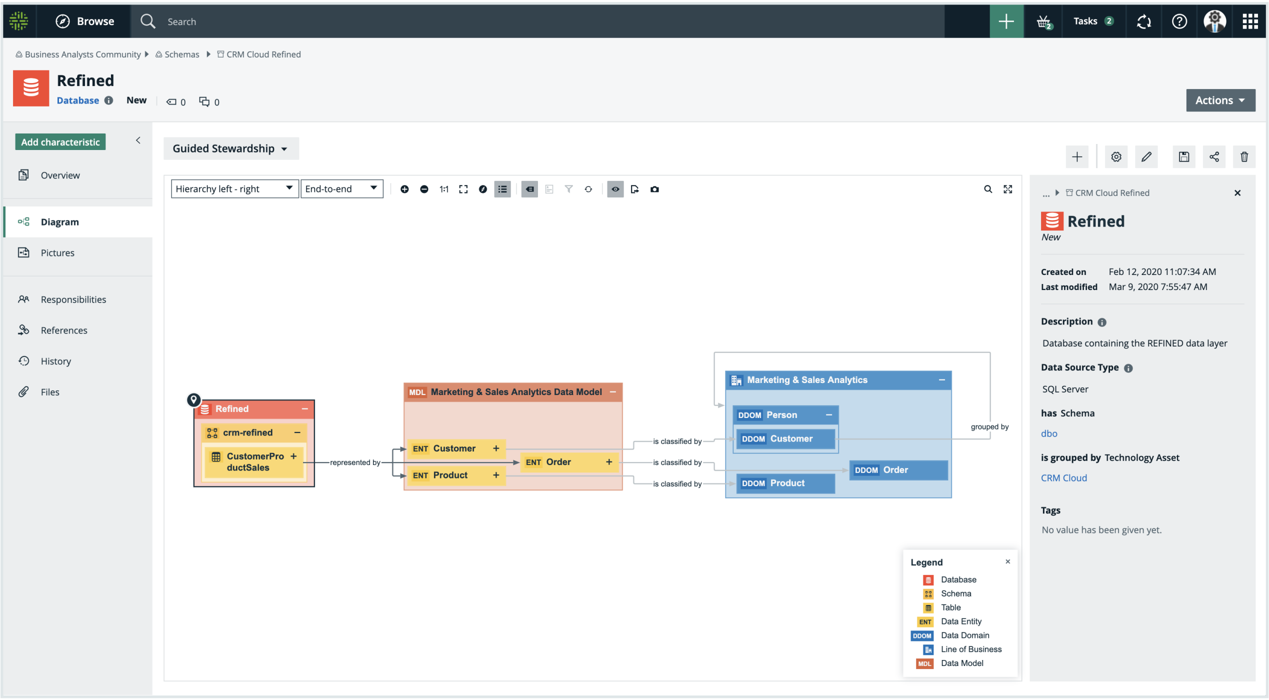
Related to...	Via the relation type...	Description
Database assets	System groups / is grouped by Database	One-to-many relation, whereby: • A System asset can group many Database assets. • A Database asset can be grouped by only one System asset.

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.



CRM Cloud Refined

Refined

New

Created on Feb 12, 2020 11:07:34 AM
Last modified Mar 9, 2020 7:55:47 AM

Description

Database containing the REFINED data layer

Data Source Type

SQL Server

has Schema

dbo

is grouped by Technology Asset

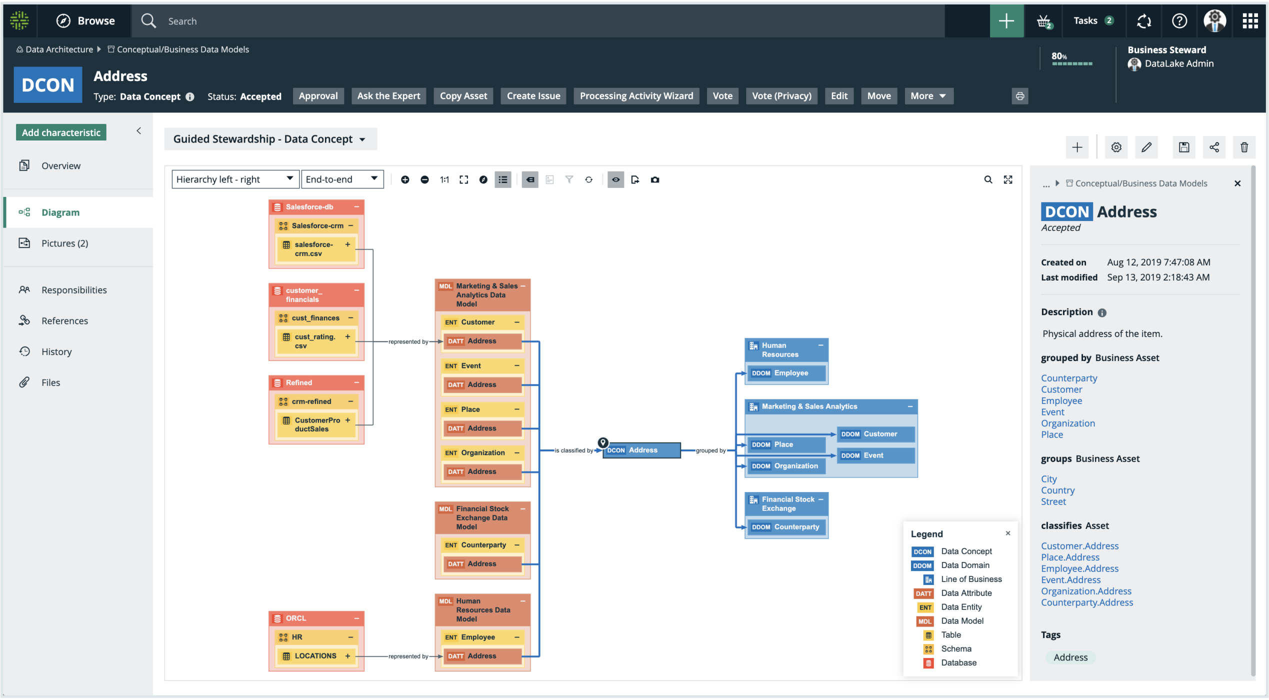
CRM Cloud

Tags

No value has been given yet.

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.



Type: Data Concept Status: Accepted

Conceptual/Business Data Models

DCON Address

Accepted

Created on Aug 12, 2019 7:47:08 AM
Last modified Sep 13, 2019 2:18:43 AM

Description

Physical address of the item.

grouped by Business Asset

Counterparty
Customer
Employee
Event
Organization
Place

groups Business Asset

City
Country
Street

classifies Asset

Customer.Address
Place.Address
Employee.Address
Event.Address
Organization.Address
Counterparty.Address

Tags

Address

For more information, see Diagram views.

Chapter 8

Physical Data Connector

The Physical Data Connector shows a high-level overview of database information on which you can filter.

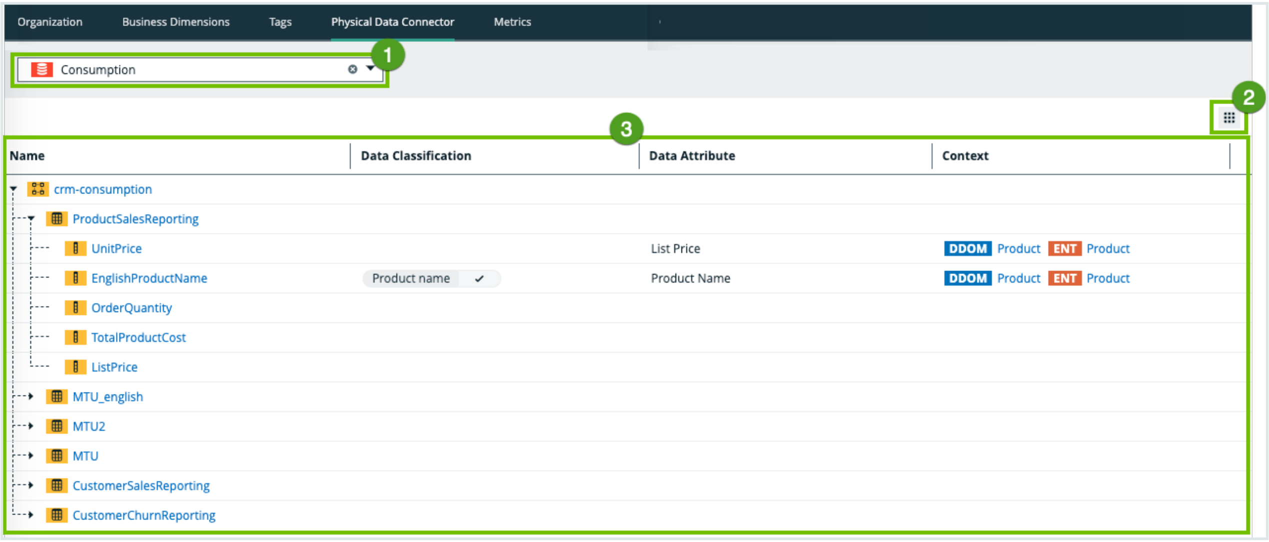
You can use the Physical Data Connector to:

- 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.

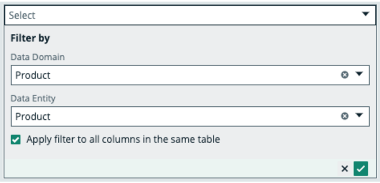
You access the Physical Data Connector via the Physical Data Connector subpage on the Stewardship tab.



Name	Data Classification	Data Attribute	Context
crm-consumption			
ProductSalesReporting			
UnitPrice		List Price	DDOM Product ENT Product
EnglishProductName	Product name ✓	Product Name	DDOM Product ENT Product
OrderQuantity			
TotalProductCost			
ListPrice			
MTU_english			
MTU2			
MTU			
CustomerSalesReporting			
CustomerChurnReporting			

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

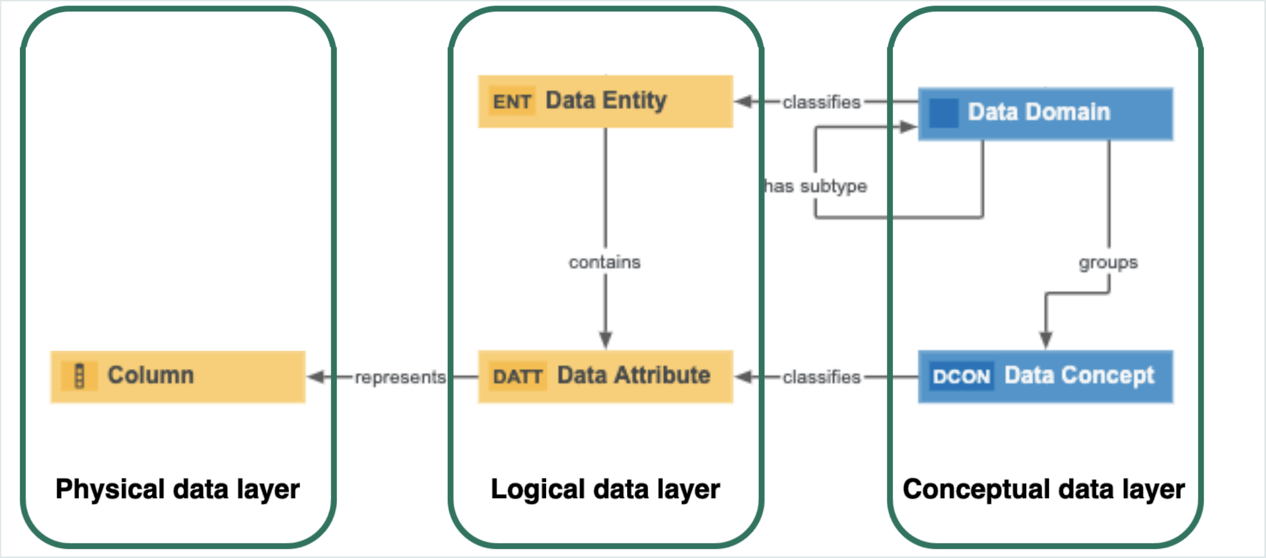
No.	Name	Description
	Data Attribute	The Data Attribute asset linked to the Column asset via relation type "Data Attribute represents / represented by Column". When you filter on a Data Domain or Data Entity, the other drop-down lists dynamically update to only show content that relates to your filter. 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.  Select Filter by Data Domain: Product Data Entity: Product ☑ Apply filter to all columns 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



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. In 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. In 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 Amazon S3

Amazon S3 is an online object storage service hosted by Amazon. For more information about Amazon S3, see the Amazon S3 documentation.

In Collibra Data Governance Center, you can synchronize with Amazon S3 in multiple ways.

Synchronization method	Advantages and disadvantages	More information
S3 file system integration	The resulting assets represent the folder structure by means of S3 Bucket, Directory, File, Table and Column assets. You can't profile and classify columns and tables.	Amazon S3 file system integration

Synchronization method	Advantages and disadvantages	More information	
Catalog connector	You can profile and classify the columns and tables in your S3 buckets. The folder structure of your S3 bucket isn't represented in Data Catalog.	Jobserver	Edge
		Register an Amazon S3 data source using the AWS Glue Catalog connector	1. Set up an Edge site. 2. Create a JDBC connection to your Amazon S3 data source by means the AWS Glue Catalog connector. 3. Add the following capabilities to the Edge site: Catalog JDBC ingestion and JDBC Profiling. 4. Register the Amazon S3 data source via Edge. 5. Synchronize your Amazon S3 data source.

About the Amazon S3 file system integration

The Amazon S3 file system integration allows registration of Amazon S3 as a data source in Collibra DGC and synchronization of data in Amazon S3.
After synchronization, the files and directories of Amazon S3 are represented in Collibra DGC by specific asset types, retaining the original names.

> Note
> - Only some file types are fully supported.
> - You can restrict the AWS regions to which CollibraData Catalog is allowed to connect. This step is recommended for efficient synchronization.
> - When you use this method, you cannot profile or classify data. See Working with Amazon S3.

Amazon S3 integration workflow

Step	What?	Description
1	Register an Amazon S3 file system as a data source	Creates an initial structure of a Storage Catalog domain and S3 File System asset in the selected parent community.
2	Connect to Amazon S3	Sets up the connection to Amazon S3.
3	Create crawlers	Creates crawlers to find and ingest the data of Amazon S3.
4	Synchronize Amazon S3	Runs the crawlers to ingest the data of Amazon S3.

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

Collibra DGC 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 DGC does everything for you. For more information about AWS Glue, see the AWS Glue documentation.

> Note Collibra DGC 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 DGC 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 DGC needs programmatic access to Amazon S3 and AWS Glue by means of a user. The following policies and permissions are required:

- Policies:
    - 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": "*"
        }
    ]
}
```

- Permissions:
    - In Collibra Data Intelligence Cloud 2020.11 and newer and Collibra Data Governance Center 5.7.7 and newer, the programmatic user needs the following permissions:

```
{
    "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": "*"
        }
    ]
}
```

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.

Amazon S3 asset and domain types

The Amazon S3 file system integration of Collibra Data Governance Center 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
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

Asset type	Description	Domain type
Technology Asset ▸ File Container	An asset type that represents 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. Examples: businessGlossary.xls, dataDictionary05220.csv, datacatalogv25.txt	• 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 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 Data Governance Center. 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

Register an Amazon S3 file system

You can register an Amazon S3 file system in Data Catalog.

The newly created S3 file system does not automatically connect to Amazon S3. You connect manually in the S3 File System asset that is created during the registration of the S3 file system.

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. In the main menu, click ⠿, then 🛢 **Catalog**.
   » The Catalog Home opens.
2. In the main menu, click the **Create** (+) button.
   » 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.

Field	Description
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.
   » An 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.

Connect to Amazon S3

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 cre-

ate menu of Collibra Data Governance Center.

- 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.

Connect a file system asset to Amazon S3 via Edge

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, for example, if you want to use another capability than the one you originally selected or if you want to switch to Jobserver.

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 an Edge capability with the S3 synchronization capability template.
- You have registered an Amazon S3 file system.
- You have a global role with the View Edge connections and capabilities global permission.
- 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. In the right corner, select **Edge**.
5. Select an Edge capability.
6. Click **Save**.

What's next?

You can now create crawlers.

Configure role-based Amazon S3 access control

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.

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 CollibraData Catalog is allowed to connect to synchronize Amazon S3.

> 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.

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. In the **Register data source** section, enter the required information:
3.

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 DGC to make a connection to any AWS region, leave the field empty.

4. Click the green **Save all** button.

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 Data Governance Center. 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.

> Note
>
> - By default, AWS Glue allows up to 25 crawlers per account. For more information, see the AWS Glue documentation. This has consequences for Collibra DGC:
>   - If you created crawlers in AWS Glue directly, Collibra DGC can create less crawlers for synchronization.
>   - Because Collibra DGC 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 DGC 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 DGC, 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.
- 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 Data Governance Center.
- You have connected an S3 File System asset to Amazon S3.
- You have a global role with the View Edge connections and capabilities global permission.
- 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 DGC. 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 DGC>`. ◦ 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 DGC 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.

Field	Description
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.
- 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 Data Governance Center.
- You have connected an S3 File System asset to Amazon S3.
- You have created one or more crawlers.

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 DGC. 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 DGC>`. ◦ 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 DGC 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.

Field	Description
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.
- 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 Data Governance Center.
- You have connected an S3 File System asset to Amazon S3.
- You have created one or more crawlers.

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**.

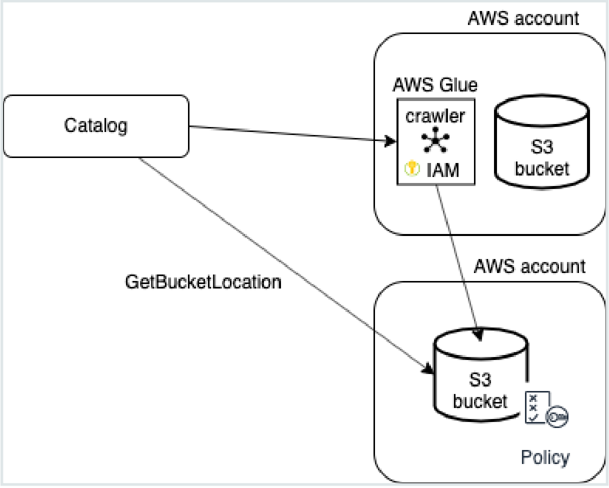
Cross-account crawling

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:::*"
        ]
    }
  ]
}
```

About synchronizing Amazon S3

Synchronizing Amazon S3 is the process of ingesting metadata from a selected Amazon S3 repository and making the data available in Collibra Data Governance Center.

When you synchronize Amazon S3, the content of your Amazon S3 repository is analyzed and represented in Collibra DGC by means of assets and their characteristics. Technically, the synchronization happens in several steps:

1. Collibra DGC creates crawlers in AWS Glue, based on the crawlers defined in Collibra DGC.
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 DGC ingests those databases and creates assets, attributes and relations as required to match the metadata.
5. The AWS Glue crawlers are deleted.

Starting the synchronization

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.

Synchronization results

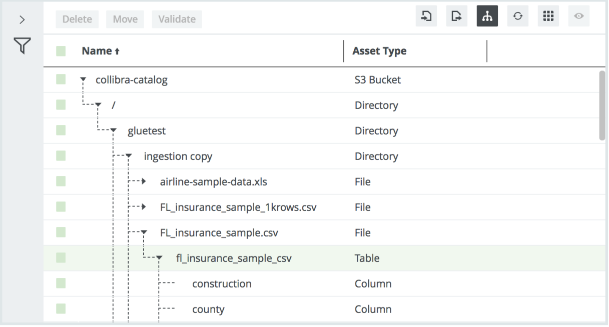
After synchronization, the resulting assets are in the domain that was specified in the crawler.

> 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. S3 Bucket contains Directory
2. Directory contains Directory
3. Directory 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.



Name ↑	Asset Type
collibra-catalog	S3 Bucket
/	Directory
gluetest	Directory
ingestion copy	Directory
airline-sample-data.xls	File
FL_insurance_sample_1krows.csv	File
FL_insurance_sample.csv	File
fl_insurance_sample_csv	Table
construction	Column
county	Column

> 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.

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 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 Data Governance Center.
- 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 connected an S3 File System asset to Amazon S3.
- You have created one or more crawlers.
- You have a global role with the View Edge connections and capabilities global permission.
- 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. 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.
   » The **Synchronization schedule** section displays the time of the last synchronization.

> 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.

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 S3 synchronization schedule

To keep the content of Collibra Data Governance Center 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 a global role with the View Edge connections and capabilities global permission.
- 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 Data Governance Center.
- You have connected an S3 File System asset to Amazon S3.
- 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 the schemas automatically, for example daily, weekly or based on a 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 the schemas, 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 the schemas , 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 the schemas automatically, for example 14:00. 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 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 Data Governance Center.
- You have connected an S3 File System asset to Amazon S3.
- You have created one or more crawlers.
- You have added a synchronization schedule.

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 the schemas automatically, for example daily, weekly or based on a 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 the schemas, 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 the schemas , 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 the schemas automatically, for example 14:00. 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 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 Data Governance Center.
- You have connected an S3 File System asset to Amazon S3.
- You have created one or more crawlers.
- You have added a synchronization schedule.

Steps

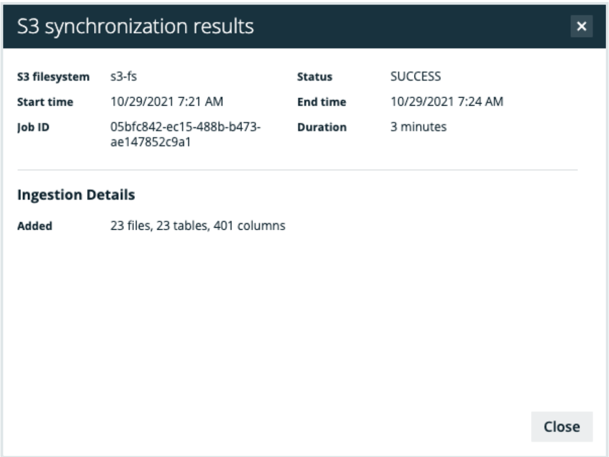
1. Open an S3 File System asset page.
2. In the tab pane, click ⚙ **Configuration**.
3. In the **Synchronization schedule** section, click **Remove Schedule**.

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 Data Governance Center

Steps

1. Open the Activities list.
2. In the row containing the S3 synchronization job, click **Result**.
   » The **S3 synchronization results** dialog box appears.



S3 synchronization results

S3 filesystem	s3-fs	Status	SUCCESS
Start time	10/29/2021 7:21 AM	End time	10/29/2021 7:24 AM
Job ID	05bfc842-ec15-488b-b473-ae147852c9a1	Duration	3 minutes

Ingestion Details

Added 23 files, 23 tables, 401 columns

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.

Delete an S3 File System asset from Collibra DGC

You can delete an S3 File System asset from Collibra Data Governance Center.

Note

- The crawlers of the S3 File System asset are deleted.
- The assets that were created by synchronizing 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. In 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 for the S3 file system integration

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 DGC 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 DGC 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 DGC.	• 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.
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 failed and AWS logs show an **Internal server error**

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.

No assets created after 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 were ingested

Jobs in Collibra DGC 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 DGC and the Jobserver.

> Note When you start synchronization, the crawlers are created in AWS Glue. Once the crawlers are created, they are executed. If Collibra DGC 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.

File size (or other property) is not filled 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.

A file is wrongly considered as 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)

The only work-around that Collibra found, is to 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 number of crawlers in AWS Glue (25 by default - can be expanded on request).

My 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.

Synchronizing an S3 File System fails with a **relationMaxLimitReachedTarget** message in 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 thanone 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.

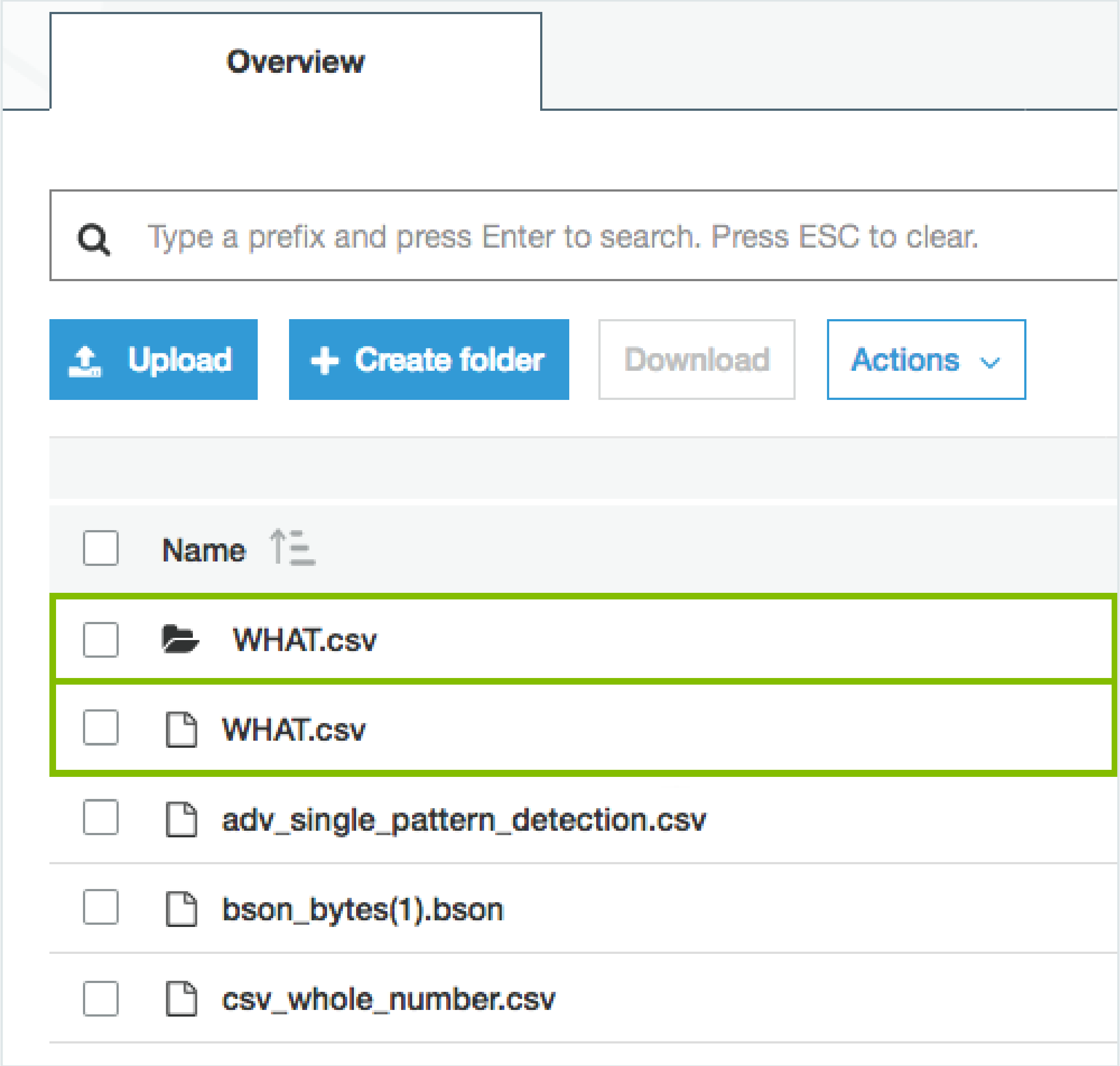
> Example
> You work in domain called Amazon, which contains a Directory asset called Main. The Main Directory asset has a child asset of the File type, called Names.
>
> You move the Main Directory asset to another domain called Local.
>
> When you synchronize again, Data Catalog first recreates the 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.

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 re-synchronise as soon as possible.

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.



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 DGC. 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 DGC, several attribute types will be changed to the plain text kind to avoid this issue.

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.

Some of the folders and files in Amazon S3 are not visible in Collibra DGC

You may notice that the content of your Amazon S3 does not always match the content in Collibra DGC. Some folders from Amazon S3 may not appear in Collibra DGC and some files are merged or split into different assets. This is not a bug in Collibra DGC. 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 DGC 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 DGC are created.

See the AWS Glue documentation for more information about folders and tables in Amazon S3 and what happens when a crawler runs.

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.
```

You must grant the AWS programmatic user the following permissions to synchronize Amazon S3 :

```
{
    "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": "*"
        }
    ]
}
```

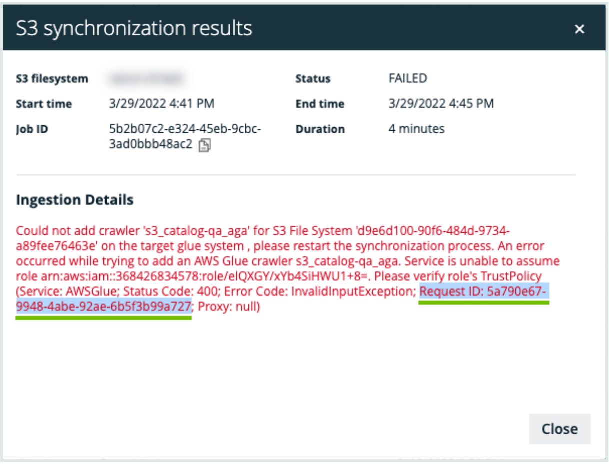
For more information about AWS requirements, see the Amazon S3 file system section.

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.

Where do I find the **Request ID** for AWS troubleshooting?

When an S3 synchronization fails, you can find a detailed error message, including the **Request ID**, in the S3 synchronization results.



S3 synchronization results

S3 filesystem	[redacted]	**Status**	FAILED
Start time	3/29/2022 4:41 PM	**End time**	3/29/2022 4:45 PM
Job ID	5b2b07c2-e324-45eb-9cbc-3ad0bbb48ac2	**Duration**	4 minutes

Ingestion Details

Could not add crawler 's3_catalog-qa_aga' for S3 File System 'd9e6d100-90f6-484d-9734-a89fee76463e' on the target glue system , please restart the synchronization process. An error occurred while trying to add an AWS Glue crawler s3_catalog-qa_aga. Service is unable to assume role arn:aws:iam::368426834578:role/elQXGY/xYb4SiHWU1+8=. Please verify role's TrustPolicy (Service: AWSGlue; Status Code: 400; Error Code: InvalidInputException; Request ID: 5a790e67-9948-4abe-92ae-6b5f3b99a727; Proxy: null)

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.

Chapter 10

Working with Tableau

Tableau is business intelligence software that helps people see and understand their data. Integrating Tableau in Collibra Data Governance Center enables you to see metadata from Tableau Server and Tableau Online in CollibraData Catalog.

In this section, we describe how you can ingest Tableau metadata in CollibraData Catalog and synchronize the metadata via the Data Catalog user interface.

> Tip We have made available a new Tableau integration method that entails use of the lineage harvester, a standalone Java application. We recommend that you use the new method, which has numerous advantages; however, the following is true of the new integration method:
>
> - It is a cloud-only feature.
> - It is currently available only to customers who do not need to migrate existing Tableau assets to the new operating model.
> - The new Tableau operating model is only available in Collibra DGC versions 2021.10 and newer.
>
> We will soon make available a migration tool for those who would like to benefit from the new method, but need to migrate existing Tableau assets.

> Important
> - The development of new features and improvements to the Tableau integration via the Data Catalog user interface is discontinued.
> - The two Tableau integration methods—Tableau integration via the Data Catalog and the new integration method via lineage harvester—coexist, and you are free to use the method of your choosing.

About the Tableau integration

Tableau integration means registering Tableau as a system in Collibra Data Governance Center and ingesting, or synchronizing, the Tableau metadata. After synchronization, metadata from Tableau Server or Tableau Online are represented in Collibra DGC by specific asset and domain types, retaining their original names.

Tableau ingestion

The table below shows the steps required for ingesting Tableau metadata.

Step	What?	Description
1	Register Tableau as a system.	Creates an initial structure of a community, BI Catalog domain and Tableau Server asset in the selected parent community.
2	Connect to Tableau Server or Tableau Online.	Connects to Tableau server or Tableau Online.
3	Synchronize Tableau Server or Tableau Online.	Ingests the metadata from Tableau.

Step	What?	Description
4	Stitch Tableau logical data layer and physical data layer.	Optionally, stitch Tableau assets to assets of registered data sources in Data Catalog.

Authentication

Data Catalog uses Tableau's REST API to get metadata information and follows Tableau's requirements regarding authentication methods. As a consequence, you need a Tableau user with access to the relevant Tableau sites.

For more information, see the Tableau documentation.

Password encryption

Collibra's integration of Tableau 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 Tableau.

Limitations

Collibra does not support the following Tableau features:

- Gziped encoding in REST results from Tableau.
- Tableau data sources that are created using Custom SQL.
- Tableau data sources that are created using Multiple tables union.

Supported Tableau Server versions

Collibra Data Intelligence Cloud supports the following Tableau Server versions:

- 10.4
- 10.5

- 2018.x
- 2019.x
- 2020.1
- 2020.2
- 2020.3
- 2020.4
- 2021.1
- 2021.2
- 2021.3
- 2021.4

> Note Depending on your Tableau version, Data Catalog uses different APIs to integrate Tableau. You need different Tableau permissions according to the Tableau version that you want to integrate.

Tableau terminology

Before you start using Tableau to ingest data, read more about the Tableau terminology and how it maps with the Collibra Data Governance Center terminology.

Tableau term	Description	Collibra DGC equivalent
Site	A site is a stand-alone collection of content, such as projects, workbooks and users. Each site has its own URL and its own set of users.	Subcommunity and Tableau Site asset
Project	A project organizes related content resources. Content resources are workbooks, views and data sources.	Tableau Project asset
Workbook	A workbook is a collection of views.	Tableau Workbook asset
View	A view is a way to represent data.	Tableau View asset

Tableau term	Description	Collibra DGC equivalent
Story	A story contains a sequence of worksheets or dashboards that work together to convey information.	Tableau Story asset
Dashboard	A dashboard is a collection of views from multiple worksheets.	Tableau Dashboard asset
Worksheet	A worksheet contains a single view, along with shelves, legends, and the Data pane.	Tableau Worksheet asset
Tableau data sources	Tableau Data Sources consist of metadata that describe the connection information, information about how to access or refresh the data and customizations.	Tableau Data Source asset
Dimension	Dimensions contain qualitative values (such as names, dates, or geographical data).	Tableau Report Attribute asset
Measure	Measures contain numeric, quantitative values that you can measure.	Tableau Report Attribute asset
Tableau data attribute	Tableau Data Attributes define a property of a Tableau data entity.	Column asset
Tableau data entity	Tableau Data Entities are an abstraction of the physical implementation of database tables, used for Tableau report creation.	Schema asset and Table asset

Tableau term	Description	Collibra DGC equivalent
Tableau data model	Tableau Data Models are an abstraction for the physical implementation of databases, schemas, files, etc., used for Tableau report creation.	Database asset

Tableau asset and domain types

The Tableau integration of Collibra Data Governance Center uses a specific subset of asset types and domain types. All of these come out of the box with your software.

The following table contains the asset an domain types that are used for the Tableau integration. Above each asset type you can see the parent asset types in the breadcrumbs.

Asset type	Description	Domain type
Business Asset ▸ Business Dimension ▸ BI Folder ▸ Tableau Project	Collection of Tableau workbooks and data sources.	BI Catalog
Business Asset ▸ Business Dimension ▸ BI Folder ▸ Tableau Site	Collection of content (workbooks, data sources, users, …) that's walled off from any other content on that instance of Tableau Server.	BI Catalog

Asset type	Description	Domain type
Business Asset ▸ Report ▸ BI Report ▸ Tableau View ▸ Tableau Dashboard	A collection of several worksheets and supporting information, shown on a single screen, so that you can simultaneously compare and monitor a variety of data.	BI Catalog
Business Asset ▸ Report ▸ BI Report ▸ Tableau View ▸ Tableau Worksheet	A worksheet is a single sheet on which you can build views of your data.	BI Catalog
Business Asset ▸ Report ▸ BI Report ▸ Tableau Workbook	Collection of sheets. A sheet can be a worksheet, a dashboard or a story.	BI Catalog
Data Asset ▸ Data Element ▸ Data Attribute ▸ BI Data Attribute ▸ Tableau Data Attribute	A specification that defines a property of a Tableau data entity. Examples: CustomerBirthDate, EmployeeFirstName.	BI Catalog
Data Asset ▸ Data Element ▸ Report Attribute ▸ BI Report Attribute ▸ Tableau Report Attribute	An atomic unit of data that represents a Tableau report. Examples: ExpenseAmount, RiskAmount	BI Catalog

Asset type	Description	Domain type
Data Asset ▸ Data Structure ▸ Data Entity ▸ BI Data Entity ▸ Tableau Data Entity	An abstraction from the physical implementation of database tables, used for Tableau report creation.	BI Catalog
Data Asset ▸ Data Structure ▸ Data Model ▸ BI Data Model ▸ Tableau Data Model	An abstraction from the physical implementation of database, schema, file, etc., used for Tableau report creation.	BI Catalog
Technology Asset ▸ Server ▸ BI Server ▸ Tableau Server	A visual analytics platform for creating interactive dashboards and rich visualisations	BI Catalog
Technology Asset ▸ System ▸ BI Data Source ▸ Tableau Data Source	The link between Tableau and an external system. A Tableau data source contains the information to connect to external data, table names, the table relationships, and any customizations that you make.	BI Catalog

Note The BI Data Catalog domain type was formerly known as the Tableau Data Catalog domain type.

Tableau business logic

Tableau business users work with Tableau projects, workbooks and worksheets to make business decisions. Collibra's Tableau integration offers business users several advantages:

- Easily find certified Tableau content.
- Shop for Tableau reports.
- Trace Tableau data to its source.
- Find where content is stored in Tableau.

Tableau asset pages

Tableau metadata is represented by assets of various types. Depending on the Tableau asset type, the asset page shows different information ingested from Tableau. You can find a specific Tableau asset page using Data Catalog search or via the Data Catalog BI domains in which you ingested the Tableau metadata.

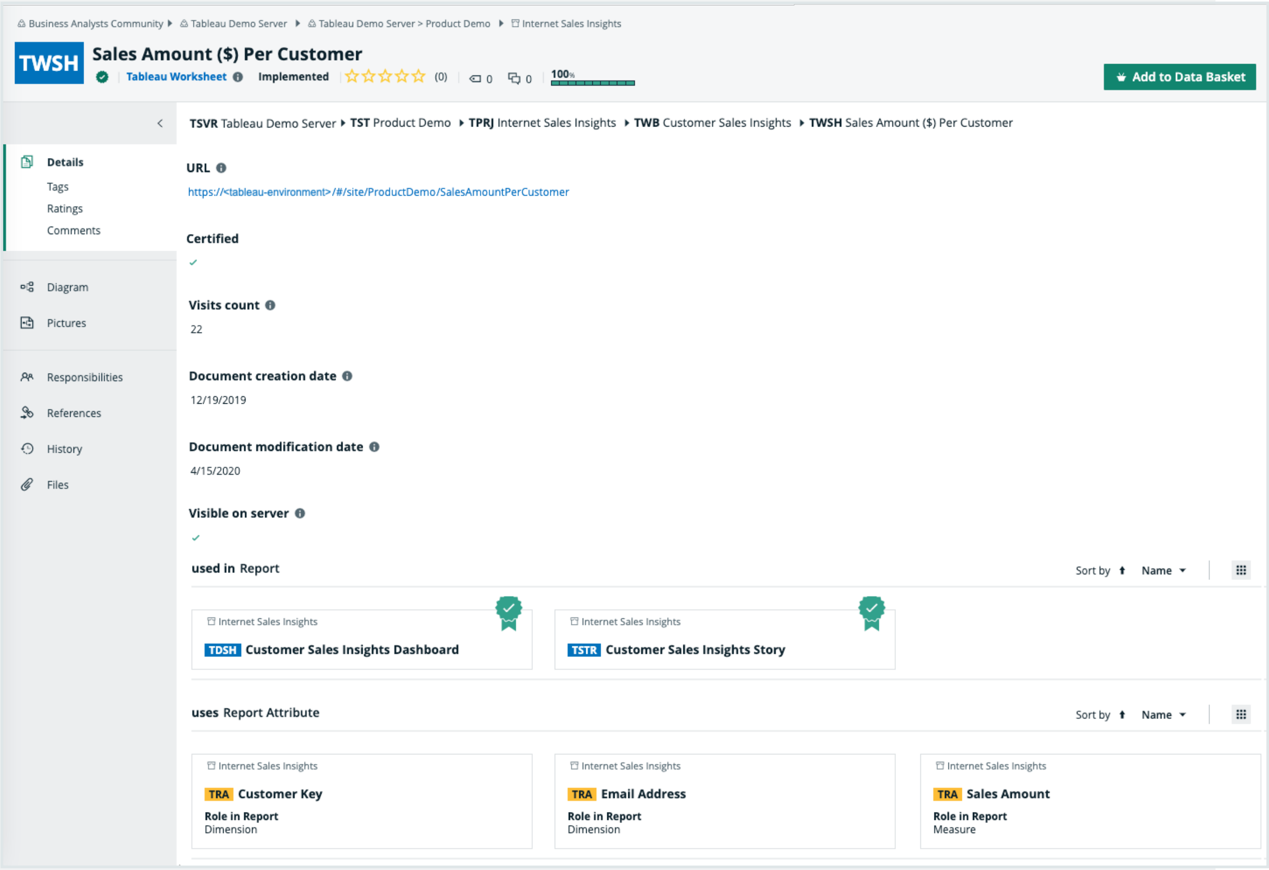
Details

Asset pages show attributes and relations to other assets. This information is synchronized from Tableau. However, you can add additional characteristics, tags or comments.

If you want access to one or more Tableau assets, you can add them to your Data Basket and check out the Data Basket. You can request access to assets of the following types:

- Tableau Workbook
- Tableau Worksheet
- Tableau Dashboard
- Tableau Story

Example The following Tableau Worksheet asset shows in which Tableau Dashboard and Tableau Story it is used and which Tableau Report Attribute it uses. This asset and the related reports are certified, indicating that the data is considered reliable.



Business Analysts Community ▸ Tableau Demo Server ▸ Tableau Demo Server > Product Demo ▸ Internet Sales Insights

TWSH **Sales Amount ($) Per Customer**

Tableau Worksheet · Implemented · ☆☆☆☆☆ (0) · 0 · 0 · 100%

TSVR Tableau Demo Server ▸ **TST** Product Demo ▸ **TPRJ** Internet Sales Insights ▸ **TWB** Customer Sales Insights ▸ **TWSH** Sales Amount ($) Per Customer

URL
https://<tableau-environment>/#/site/ProductDemo/SalesAmountPerCustomer

Certified
✓

Visits count
22

Document creation date
12/19/2019

Document modification date
4/15/2020

Visible on server
✓

used in Report

Internet Sales Insights
TDSH **Customer Sales Insights Dashboard**

Internet Sales Insights
TSTR **Customer Sales Insights Story**

uses Report Attribute

Internet Sales Insights
TRA **Customer Key**
Role in Report
Dimension

Internet Sales Insights
TRA **Email Address**
Role in Report
Dimension

Internet Sales Insights
TRA **Sales Amount**
Role in Report
Measure

Diagrams

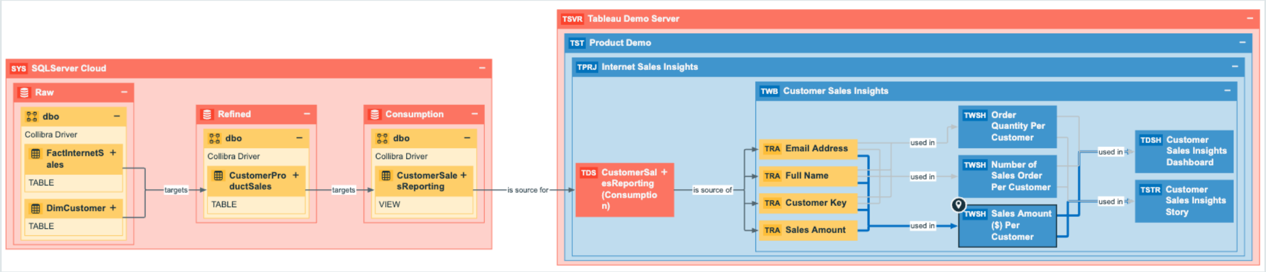
Diagrams is a feature to show and interact with assets based on their relations in an easy-to-read diagram. The diagram helps you to quickly see how assets are related. As such, the diagram can show a high-level presentation of a Tableau Workbook. If the Tableau assets are stitched to registered assets in Data Catalog, you can also see the stitching results in the diagram. This enables you, for example, to see:

- In which Tableau Project the Tableau Workbook is stored.
- In which Tableau Site the Tableau Project is stored.
- Which Tableau Data Source is the source of the Tableau Report Attributes in the

Tableau Workbook.

- Which Table assets are the source for the Tableau Data Source asset via stitching.

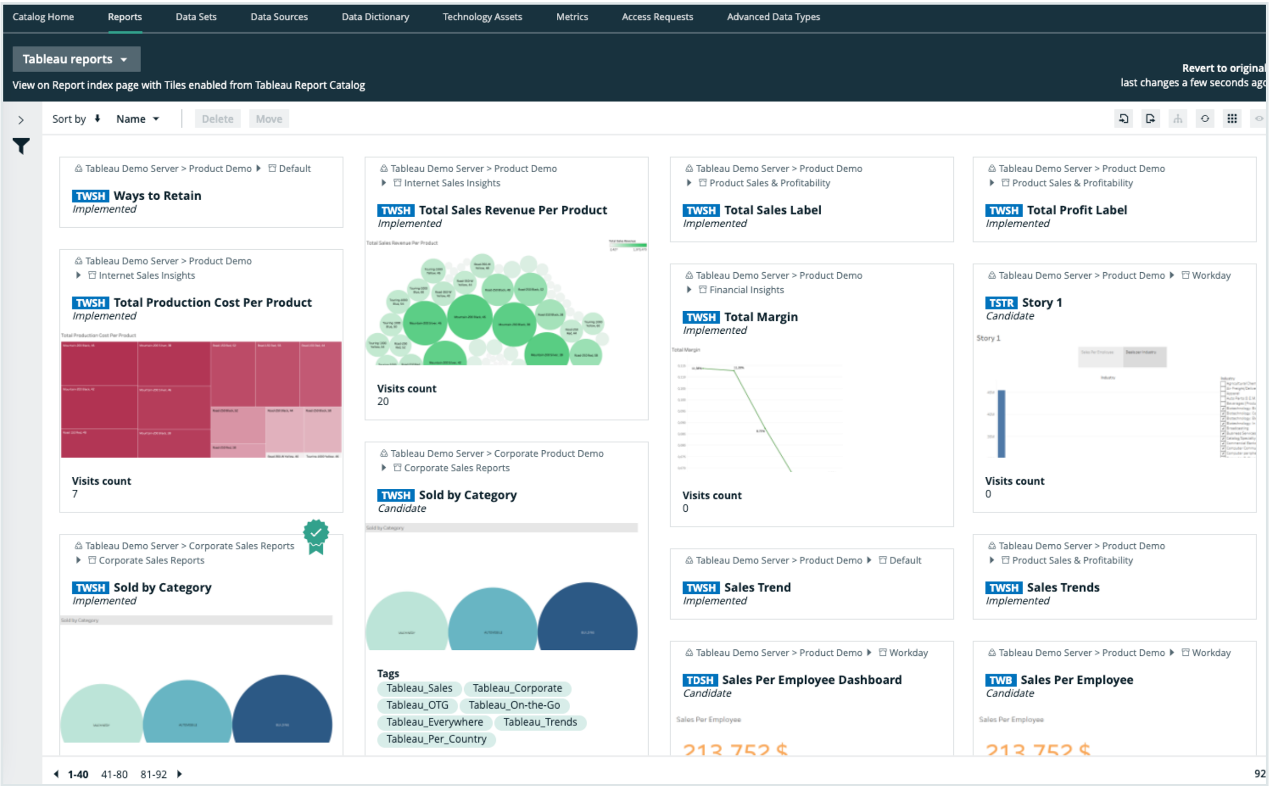
> Example The following diagram shows the *Customer Sales Insights* Tableau Workbook, which is stored in the *Internet Sales Insights* Tableau Project. The Tableau Workbook contains Tableau Report Attributes that have the *CustomerSalesReporting* Tableau Data Source as source. This Tableau Data Source is stitched to the *CustomerSalesReporting* Table asset in the *SQL Server Cloud* data source.



Report views

The Tableau integration feature enables you to find all ingested Tableau Workbook assets and children of this asset type in a single location.

In the **Reports** tab page in Data Catalog you can see an overview of all BI Report assets and their children. Optionally, you can create a view with a filter to only show Tableau assets. This is useful if you quickly want to find a specific report or if you want to know which reports are certified.


Catalog Home
Reports
Data Sets
Data Sources
Data Dictionary
Technology Assets
Metrics
Access Requests
Advanced Data Types
Tableau reports
View on Report index page with Tiles enabled from Tableau Report Catalog
Revert to original
last changes a few seconds ag
Sort by
Name
Delete
Move
Tableau Demo Server > Product Demo ▸ Default
TWSH Ways to Retain
Implemented
Tableau Demo Server > Product Demo ▸ Internet Sales Insights
TWSH Total Sales Revenue Per Product
Implemented
Visits count
20
Tableau Demo Server > Product Demo ▸ Product Sales & Profitability
TWSH Total Sales Label
Implemented
Tableau Demo Server > Product Demo ▸ Product Sales & Profitability
TWSH Total Profit Label
Implemented
Tableau Demo Server > Product Demo ▸ Internet Sales Insights
TWSH Total Production Cost Per Product
Implemented
Visits count
7
Tableau Demo Server > Product Demo ▸ Financial Insights
TWSH Total Margin
Implemented
Visits count
0
Tableau Demo Server > Product Demo ▸ Workday
TSTR Story 1
Candidate
Story 1
Visits count
0
Tableau Demo Server > Corporate Product Demo ▸ Corporate Sales Reports
TWSH Sold by Category
Candidate
Tags
Tableau_Sales
Tableau_Corporate
Tableau_OTG
Tableau_On-the-Go
Tableau_Everywhere
Tableau_Trends
Tableau_Per_Country
Tableau Demo Server > Corporate Sales Reports ▸ Corporate Sales Reports
TWSH Sold by Category
Implemented
Tableau Demo Server > Product Demo ▸ Default
TWSH Sales Trend
Implemented
Tableau Demo Server > Product Demo ▸ Product Sales & Profitability
TWSH Sales Trends
Implemented
Tableau Demo Server > Product Demo ▸ Workday
TDSH Sales Per Employee Dashboard
Candidate
Sales Per Employee
213,752 $
Tableau Demo Server > Product Demo ▸ Workday
TWB Sales Per Employee
Candidate
Sales Per Employee
213,752 $
1-40
41-80
81-92
92

Register a Tableau server

Before you can synchronize Tableau, you have to register a Tableau server to create an initial structure of a community, meaning a BI Catalog domain and a Tableau Server asset in a selected parent community in Collibra Data Governance Center.

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 global role that has the Manage all resources global permission.
- You have a role with the following resource permissions on the Tableau community you create when you register a Tableau server:
  - Asset: add
  - Attribute: add

  - Domain: add
  - Attachment: add
- You have enabled the Tableau metadata API in Collibra Console and in Tableau if you use Tableau 2020.2 or newer.

Steps

1. In the main menu, click ⠿, then 🛢 **Catalog**.
   » The Catalog Home opens.
2. In the main menu, click the **Create** (+) button.
3. In the **Create** dialog box, click **Register system**.
4. In the **Register system** dialog box, click **Tableau Server**.
5. In the **Register Tableau server** dialog box, enter the required information.

Field	Description
Community	The name of the parent community in which the initial Tableau structure will be created.
Tableau server name	The name of the Tableau server. The name that you fill in here will be the name of the subcommunity, the domain in this subcommunity and the Tableau Server asset.
Description	A description to provide extra information about the Tableau server. This content is used as the description of the Tableau Server asset.
Owner	The owner of the data in the created community. By default, your user is selected.

6. Click **Register**.
   » A Tableau Server asset is created.
   » A Tableau Catalog domain is created.
   » The configuration page of the Tableau Server asset is automatically opened.

What's next?

You can now connect to Tableau Server or Tableau Online.

Connect to Tableau

To retrieve data from Tableau, you have to connect to Tableau via a Tableau Server asset in your Collibra Data Intelligence Cloud environment.

> Tip You have to register the Tableau Server asset before you can connect to it.

You can edit the connection settings at any time, for example, if you want to use another user than the one you originally used.

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 Tableau.
- If you connect to Tableau Online, you have a Tableau user with at least Viewer rights.
- If you connect to Tableau Server, you have a Tableau user with access to at least one site.
- You have the necessary Tableau permissions.

Steps

1. Open a Tableau Server 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	Required
On-premises | Online	The Tableau product that you use.	✓ Yes
Tableau URL or endpoint	The URL of your Tableau Server or Tableau Online. Example http://my-tableau.collibra.com	✓ Yes
Site ID	The ID of a Tableau site. ◦ If you don't enter a site ID, your Tableau user must have access to the Default site. ◦ If you enter a site ID, your Tableau user must have access to that site. Note If you connect to Tableau Server, the site ID does not determine which sites you can synchronize from that server. It is used to validate the permissions of the Tableau user. Eeven if you enter one site ID, you can still synchronize the other sites from Tableau Server. Tip You can find the site ID in the URL of the Tableau site. The site ID is the string between `/site/` and `/projects/`. In the following URL, the site ID is `collibra`. `https://example.collibra.online.tableau.com/#/site/collibra/projects`	✕ No for Tableau Server ✓ Yes for Tableau Online

Field	Description	Required
Token Name/Username	For Tableau Online with multi-factor authentication, the Personal Access Token (PAT) name of the Tableau user. Otherwise, the username of the Tableau user.	✓ Yes
Token Secret/Password	For Tableau Online with multi-factor authentication, the Personal Access Token secret of the Tableau user. Otherwise, the password of the Tableau user.	✓ Yes

5. Click **Save**.
   » The connection is verified. If successful, you can see the list of available sites in Tableau.

What's next?

You can now synchronize one or more sites.

About synchronizing Tableau

Synchronizing Tableau is the process of ingesting metadata from a selected Tableau Server or Tableau Online and making the data available in Collibra Data Governance Center.

In this section, you can find the relevant actions and permissions to successfully synchronize Tableau.

For complete information on synchronizing Tableau, see the Collibra Data Intelligence Cloud User Guide.

Synchronizing Tableau

Synchronizing Tableau is the process of ingesting metadata from a selected Tableau Server or Tableau Online and making the data available in Collibra Data Governance

Center.

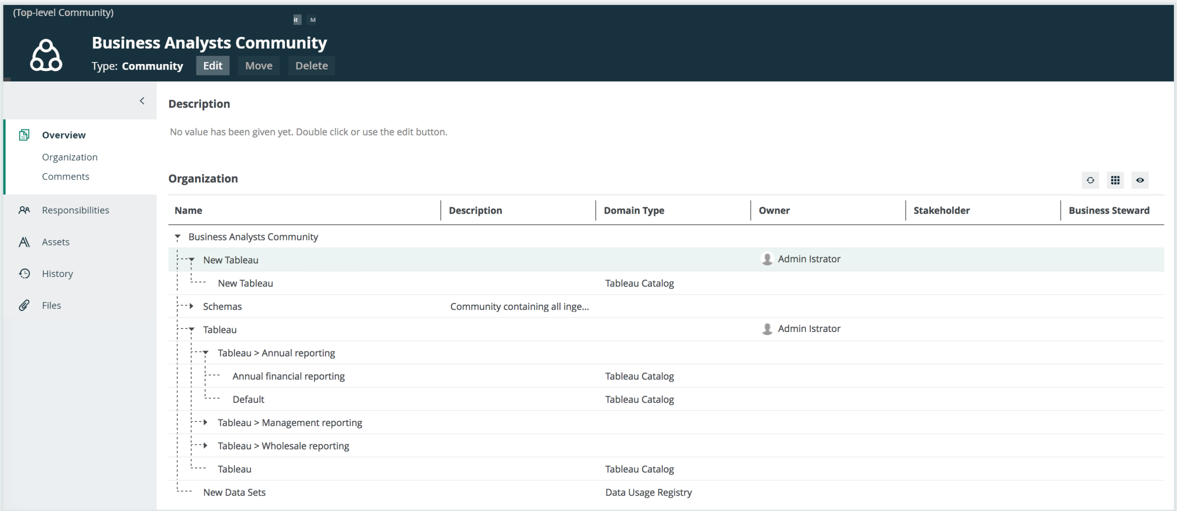
Synchronization includes the following actions:

- For each Tableau site, a subcommunity is created in the community that was created during the registration of Tableau Server or Tableau Online.
- For each Tableau project, a Tableau Catalog domain is created in the community.
- In each Catalog BI domain, a Tableau Site asset is created, with the same name as the site.
- In each Catalog BI domain, the relevant assets are created, depending on the Tableau user's permissions.

> Note
>
> - Relations that were created between Tableau assets and other assets via a relation type in the Tableau operating model, are deleted after synchronization.
> - Currently, we only support published Tableau data sources with an extract or a live connection. For more information, see the Tableau documentation.

Example

The following image shows an example structure after synchronizing Tableau.



Description

No value has been given yet. Double click or use the edit button.

Organization

Name	Description	Domain Type	Owner	Stakeholder	Business Steward
Business Analysts Community					
New Tableau			Admin Istrator		
New Tableau		Tableau Catalog			
Schemas	Community containing all inge...				
Tableau			Admin Istrator		
Tableau > Annual reporting					
Annual financial reporting		Tableau Catalog			
Default		Tableau Catalog			
Tableau > Management reporting					
Tableau > Wholesale reporting					
Tableau		Tableau Catalog			
New Data Sets		Data Usage Registry			

Starting synchronization

You can synchronize manually, or you can automate the process by adding a synchronization schedule via a cron expression.

You can only synchronize one Tableau Server asset 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 running and a new synchronization of the same Tableau Server asset is triggered (manually or automatically), the running synchronization continues and the new synchronization request is ignored.

> Note If you have stitched Tableau's logical data layer to Data Catalog physical data layer, you have to restitch to make sure that all relations are up-to-date.

Synchronization errors

In the following situations, nothing is synchronized and no subcommunities, domains or assets are created:

- If the job fails to start due to connection problems.
- If the job fails in the middle of the procedure.
- If the job is canceled.

For more information about Tableau synchronization issues, see the troubleshooting section.

> Warning If you upgrade to Tableau version 2020.2 or newer, but previously synchronized an older Tableau version via the REST API and XML mapping, you have to prepare the migration procedure to prevent losing manually added relations, attributes, tags, comments and stitching results.

Limitations and considerations

Collibra does not support the following Tableau features:

- Gziped encoding in REST results from Tableau.
- Tableau data sources that are created using Custom SQL.

- Tableau data sources that are created using Multiple tables union.

Collibra does support Tableau data sources that are created using:

- Cross-database joins
- Multiple tables join
- Relationships
- Single table

For more information, see the Tableau documentation.

Naming convention

When you synchronize Tableau, Collibra DGC follows a strict naming convention for the names of the new assets. Each asset has a display name and full name. The full name represents the asset path from asset to the database in which it is located. You can freely edit the display name. However, you should never edit the full name, because Data Catalog may need it to synchronize and stitch data sources. This can cause unexpected results and break the synchronization process.

> Warning   Editing full name of the Tableau Server or Tableau Online assets may lead to errors during the synchronization process.

Synchronized Tableau data

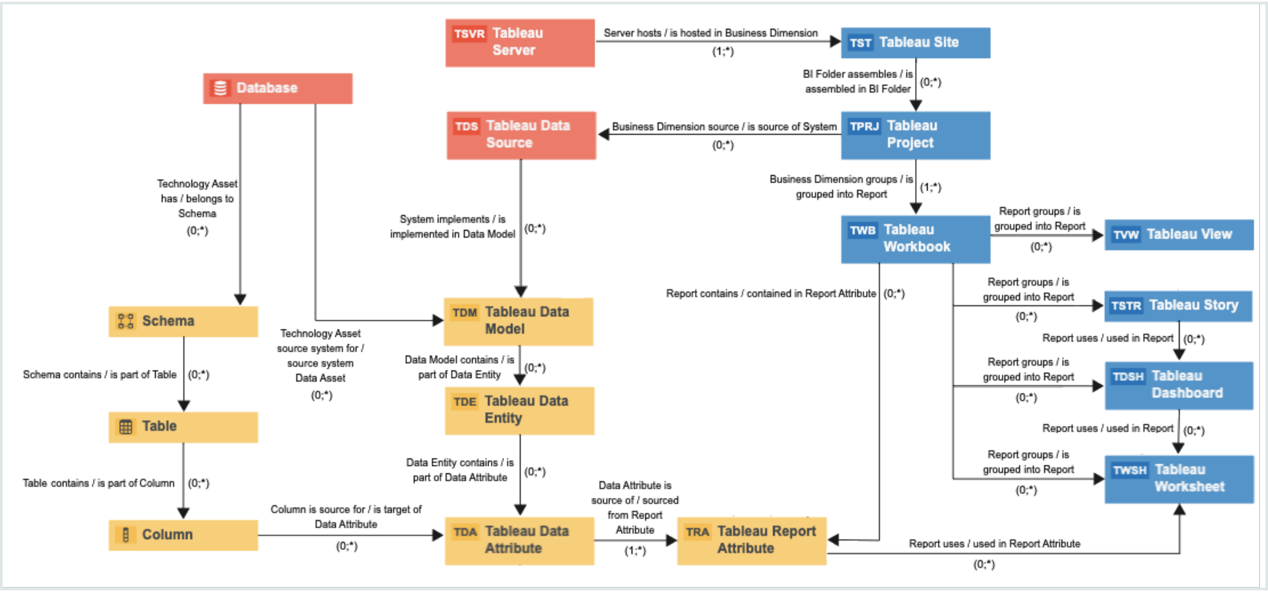
Synchronizing Tableau data means ingesting metadata from Tableau to your Collibra Data Intelligence Cloud environment. The metadata is represented as assets of specific types and their characteristics.

Note

- The assets have the same names as their counterparts in Tableau.
- Some asset types are only created if the Tableau user specified in the connection settings has specific permissions.
- There might be differences between the hierarchy of assets in Data Catalog and in Tableau. For example, Tableau, shows the relation from a parent project to a child project. In Data Catalog, this relation does not exist. Instead, all projects are shown on the Tableau Site asset page and the hierarchy of projects is shown in the Full name of the Tableau Project asset and the name of its domain.
- If the Tableau data has tags, they are also added to the corresponding assets in Collibra DGC with the prefix *Tableau_*.
- Relations that were created between Tableau assets and other assets via a relation type in the Tableau operating model, are deleted after synchronization.

Tableau operating model

The following image shows the relations between Tableau asset types and the cardinality of the relation types in the assets' assignment.


TSVR Tableau Server
Server hosts / is hosted in Business Dimension
(1;*)
TST Tableau Site
Database
BI Folder assembles / is assembled in BI Folder
(0;*)
TDS Tableau Data Source
Business Dimension source / is source of System
(0;*)
TPRJ Tableau Project
Technology Asset has / belongs to Schema
(0;*)
Business Dimension groups / is grouped into Report
(1;*)
System implements / is implemented in Data Model
(0;*)
Report groups / is grouped into Report
TWB Tableau Workbook
(0;*)
TVW Tableau View
Report contains / contained in Report Attribute
(0;*)
Report groups / is grouped into Report
(0;*)
TSTR Tableau Story
Schema
TDM Tableau Data Model
Technology Asset source system for / source system Data Asset
(0;*)
Report uses / used in Report
(0;*)
Data Model contains / is part of Data Entity
(0;*)
Report groups / is grouped into Report
(0;*)
TDSH Tableau Dashboard
Schema contains / is part of Table
(0;*)
TDE Tableau Data Entity
Table
Report uses / used in Report
(0;*)
Data Entity contains / is part of Data Attribute
(0;*)
Report groups / is grouped into Report
(0;*)
TWSH Tableau Worksheet
Table contains / is part of Column
(0;*)
Data Attribute is source of / sourced from Report Attribute
(1;*)
Column is source for / is target of Data Attribute
(0;*)
Column
TDA Tableau Data Attribute
TRA Tableau Report Attribute
Report uses / used in Report Attribute
(0;*)

Synchronized metadata per asset type

This table shows the metadata for each Tableau asset type.

Asset type	Synchronized metadata
Tableau Server	• URL • Server hosts / is hosted in Business Dimension
Tableau Site	• URL: The link to the data in Tableau • Original name: The name of the data as used in Tableau • BI Folder assembles / Is assembled in BI Folder • Server hosts / is hosted in Business Dimension
Tableau Project	• Description • Original name: The name of the project in Tableau • Business Dimension groups / grouped into Report • Business Dimension source is / source of System • BI Folder assembles / is assembled in BI Folder • Business Asset groups / is grouped by Business Asset

Asset type	Synchronized metadata
Tableau Workbook	• Certified • Original name: The name of the workbook in Tableau. • Report image: The image of the report. Note ◦ Images are not downloaded or stored in Data Catalog. Instead, Data Catalog stores a link to the image. Every time you open the asset page, the image is fetched from Tableau. If the images do not render correctly, see the Troubleshooting section. ◦ You can also exclude images from synchronization in the **Tableau sites** section on the **Configuration** page of the Tableau Server asset. • Document size • Document creation date • Document modification date • Report groups / is grouped into Report • Report is grouped in / groups Business Dimension • Report related to / is impacted by Business Asset • Report Attribute contained in / contains in Report • Technology Asset is source for / sourced from Business Asset

Asset type	Synchronized metadata
Tableau View	• URL: The link to the data in Tableau • Certified • Original name: The name of the view in Tableau. • Report image: The image of the report Note ◦ Images are not downloaded or stored in Data Catalog. Instead, Data Catalog stores a link to the image. Every time you open the asset page, the image is fetched from Tableau. If the images do not render correctly, see the Troubleshooting section. ◦ You can also exclude images from synchronization in the **Tableau sites** section on the **Configuration** page of the Tableau Server asset. • Visits count: The number of times that the view has been visited in Tableau • Document creation date • Document modification date • Visible on server • Tags • Report groups /is grouped into Report • Report relates / is impacted by Business Asset Note Assets of this type are only created if the Tableau user does not have the Download/Save As permission on the workbook.

Asset type	Synchronized metadata
Tableau Story	- URL: The link to the data in Tableau - Certified - Original name: The name of story in Tableau. - Report image: The image of the report. Note ◦ Images are not downloaded or stored in Data Catalog. Instead, Data Catalog stores a link to the image. Every time you open the asset page, the image is fetched from Tableau. If the images do not render correctly, see the Troubleshooting section. ◦ You can also exclude images from synchronization in the **Tableau sites** section on the **Configuration** page of the Tableau Server asset. - Visits count: The number of times that the view has been visited in Tableau. - Document creation date - Document modification date - Visible on server - Tags - Report groups /is grouped into Report - Report related to / is impacted by Business Asset - Report uses / used in Report Note Assets of this type are only created if the Tableau user has the Download/Save As permission on the workbook.

Asset type	Synchronized metadata
Tableau Dashboard	• URL: The link to the data in Tableau • Certified • Original name: The name of story in Tableau. • Report image: The image of the report. Note ◦ Images are not downloaded or stored in Data Catalog. Instead, Data Catalog stores a link to the image. Every time you open the asset page, the image is fetched from Tableau. If the images do not render correctly, see the Troubleshooting section. ◦ You can also exclude images from synchronization in the **Tableau sites** section on the **Configuration** page of the Tableau Server asset. • Visits count: The number of times that the view has been visited in Tableau. • Document creation date • Document modification date • Visible on server • Tags • Report groups /is grouped into Report • Report related to / is impacted by Business Asset • Report uses / used in Report Note Assets of this type are only created if the Tableau user has the Download/Save As permission on the workbook.

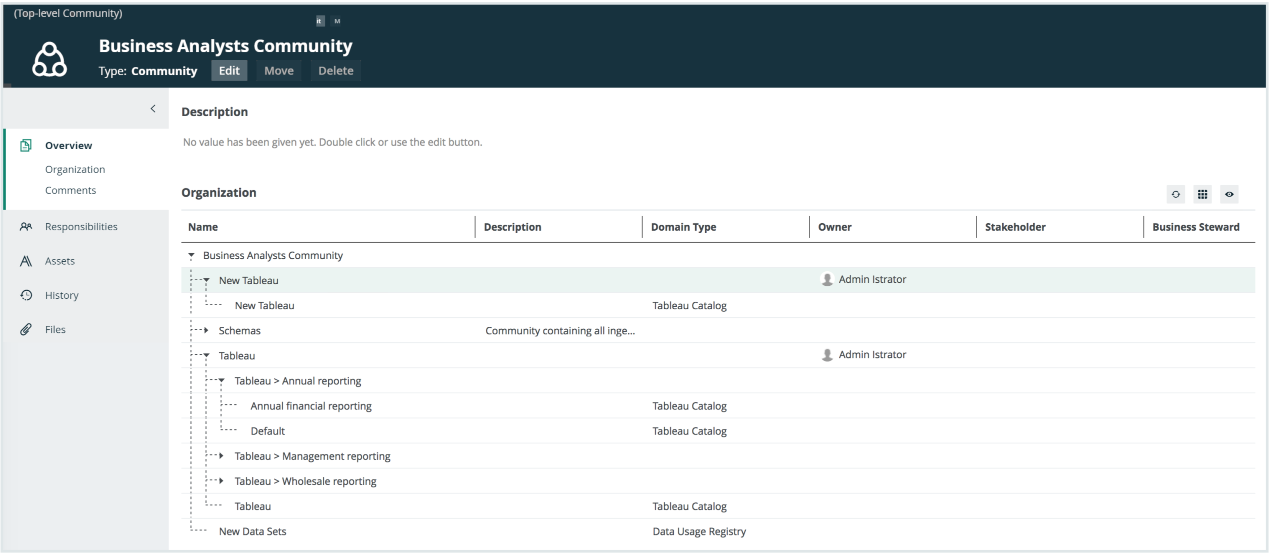
Asset type	Synchronized metadata
Tableau Worksheet	• URL: The link to the data in Tableau • Certified • Original name: The name of the data as used in Tableau • Report image: The image of the report. Note ◦ Images are not downloaded or stored in Data Catalog. Instead, Data Catalog stores a link to the image. Every time you open the asset page, the image is fetched from Tableau. If the images do not render correctly, see the Troubleshooting section. ◦ You can also exclude images from synchronization in the **Tableau sites** section on the **Configuration** page of the Tableau Server asset. • Visits count: The number of times that the view has been visited in Tableau. • Document creation date • Document modification date • Visible on server • Tags • Report uses / used in Report Attribute • Report groups / is grouped into Report • Report related to / impacted by Business Asset Note Assets of this type are only created if the Tableau user has the Download/Save As permission on the workbook.

Asset type	Synchronized metadata
Tableau Report Attribute	• Description • Original Name: The name of the attribute as used in Tableau • Technical Data Type • Role in Report • Calculation Rule: Formula used in measure • Report Attribute contained in / contains in Report • Report Attribute is source for / is target of Report Attribute • Report Attribute sourced from / is source of Data Attribute • Report uses / used in Report Attribute Note • Assets of this type are only created if the Tableau user has the Download/Save As permission on the workbook. • These are only the report attributes that are used in Tableau Worksheet of the Tableau Workbook.
Tableau Data Attribute	• Original Name: The name of the attribute as used in Tableau • Technical Data Type: The Data Type of a data asset as it is declared by the data source. • Report Attribute sourced from / is source of Data Attribute Note Assets of this type are only created if the Tableau user has the Download/Save As permission on the data source.
Tableau Data Entity	• Data Entity contains / is part of Data Attribute • Data Entity is part of / contains Data Model Note Assets of this type are only created if the Tableau user has the Download/Save As permission on the data source.

Asset type	Synchronized metadata
Tableau Data Model	• Data Source Type • Location • Data Entity is part of / contains Data Model • System implements / is implemented in Data Model Note Assets of this type are only created if the Tableau user has the Download/Save As permission on the data source.
Tableau Data Source (Published only)	• Certified • Original name: The name of the data as used in Tableau • Document creation date • Document modification date • Business Dimension sources / is source of System • System implements / is implemented in Data Model • Technology Asset implements /is implemented in Data Asset Note Currently, we only support published data sources with an extract or a live connection. For more information, see the Tableau documentation.

Examples of synchronized metadata

The following image shows an example structure after synchronizing Tableau.



(Top-level Community)

Business Analysts Community

Type: **Community**

Description

No value has been given yet. Double click or use the edit button.

Organization

Name	Description	Domain Type	Owner	Stakeholder	Business Steward
Business Analysts Community					
New Tableau			Admin Istrator		
New Tableau		Tableau Catalog			
Schemas	Community containing all inge...				
Tableau			Admin Istrator		
Tableau > Annual reporting					
Annual financial reporting		Tableau Catalog			
Default		Tableau Catalog			
Tableau > Management reporting					
Tableau > Wholesale reporting					
Tableau		Tableau Catalog			
New Data Sets		Data Usage Registry			

The following image shows an example of a diagram of a Tableau server.

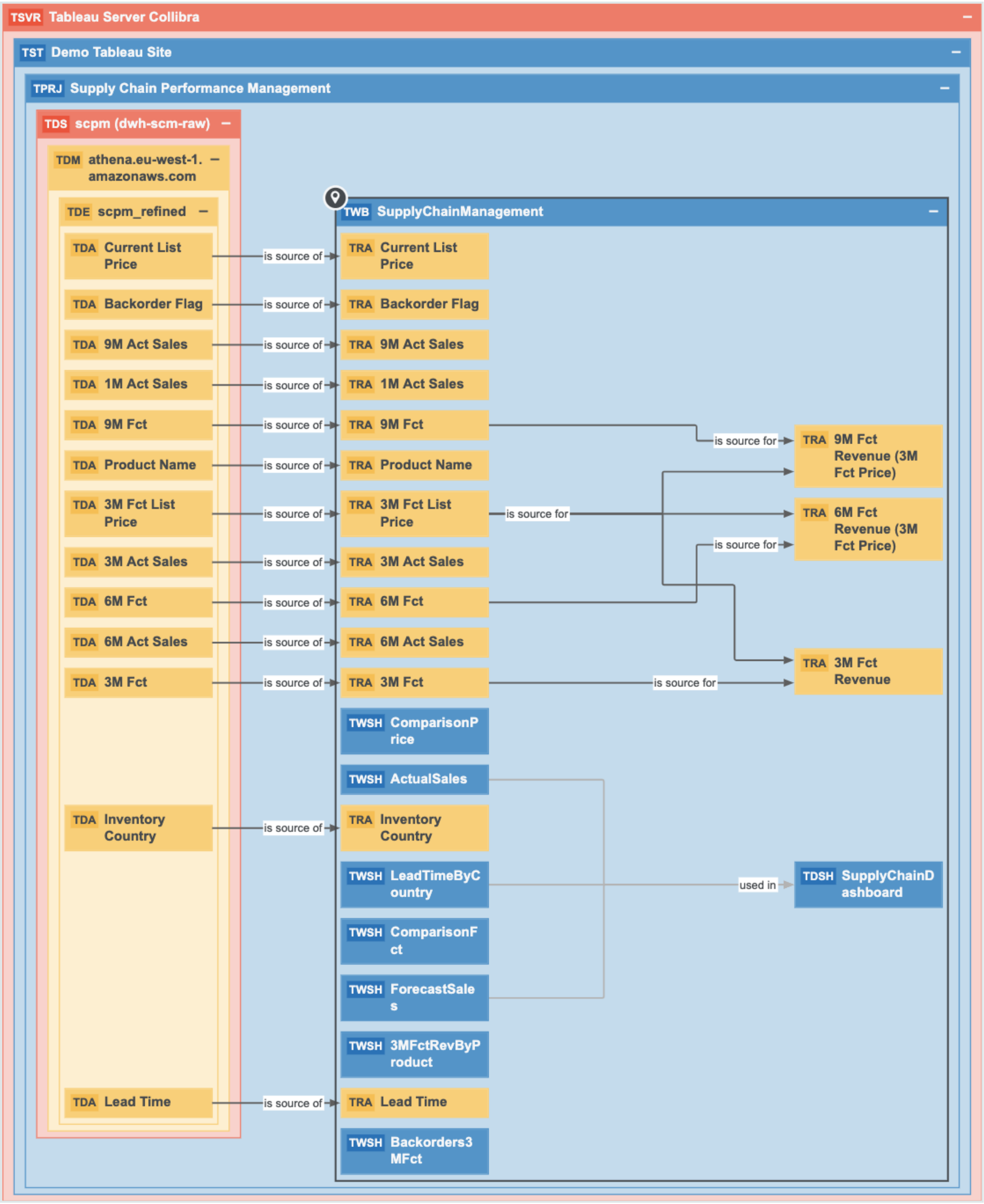

TSVR Tableau Server Collibra
TST Demo Tableau Site
TPRJ Supply Chain Performance Management
TDS scpm (dwh-scm-raw)
TDM athena.eu-west-1.amazonaws.com
TDE scpm_refined
TWB SupplyChainManagement
TDA Current List Price
is source of
TRA Current List Price
TDA Backorder Flag
is source of
TRA Backorder Flag
TDA 9M Act Sales
is source of
TRA 9M Act Sales
TDA 1M Act Sales
is source of
TRA 1M Act Sales
TDA 9M Fct
is source of
TRA 9M Fct
is source for
TRA 9M Fct Revenue (3M Fct Price)
TDA Product Name
is source of
TRA Product Name
TDA 3M Fct List Price
is source of
TRA 3M Fct List Price
is source for
TRA 6M Fct Revenue (3M Fct Price)
is source for
TDA 3M Act Sales
is source of
TRA 3M Act Sales
TDA 6M Fct
is source of
TRA 6M Fct
TDA 6M Act Sales
is source of
TRA 6M Act Sales
TRA 3M Fct Revenue
TDA 3M Fct
is source of
TRA 3M Fct
is source for
TWSH ComparisonPrice
TWSH ActualSales
TDA Inventory Country
is source of
TRA Inventory Country
TWSH LeadTimeByCountry
used in
TDSH SupplyChainDashboard
TWSH ComparisonFct
TWSH ForecastSales
TWSH 3MFctRevByProduct
TDA Lead Time
is source of
TRA Lead Time
TWSH Backorders3MFct

Tableau permissions and ingestion results

When you synchronize Tableau, you need certain permissions to access the data in Tableau. The extent of your permissions dictates the scope of the ingestion results

The following table shows the minimum role and permissions requirements for successful synchronization and the scope of the ingestion results in Data Catalog.

Tableau version	Tableau site role	Minimum required permissions			Result in Data Catalog
		Project	Workbook	Data Source	
Older than 2020.2	Viewer	View	View	View	Tableau Workbooks and Tableau Data Sources are not parsed. Resulting asset types: • Tableau Server • Tableau Site • Tableau Project • Tableau Data Source • Tableau Workbook • Tableau View

Tableau version	Tableau site role	Minimum required permissions			Result in Data Catalog
		Project	Workbook	Data Source	
Older than 2020.2	Explorer Note If your Tableau version is older than 2018.1, the Tableau site role is Interactor.	View	View	View, Download/Save As	Tableau Data Sources are parsed. Resulting asset types: • Tableau Server • Tableau Site • Tableau Project • Tableau Workbook • Tableau View • Tableau Data Source • Tableau Data Model • Tableau Data Entity • Tableau Data Attribute

Tableau version	Tableau site role	Minimum required permissions			Result in Data Catalog
		Project	Workbook	Data Source	
Older than 2020.2	Explorer Note If your Tableau version is older than 2018.1, the Tableau site role is Interactor.	View	View, Download/Save As	View	Tableau Report Attributes are synchronized and Tableau Workbooks are parsed. Resulting asset types: • Tableau Server • Tableau Site • Tableau Project • Tableau Data Source • Tableau Workbook • Tableau Story • Tableau Dashboard • Tableau Worksheet

Tableau version	Tableau site role	Minimum required permissions			Result in Data Catalog
		Project	Workbook	Data Source	
Older than 2020.2	Explorer Note If your Tableau version is older than 2018.1, the Tableau site role is Interactor.	View	View, Download/Save As	View, Download/Save As	Tableau Report Attributes are synchronized, and Tableau Data Sources and Tableau Workbooks are parsed. Resulting asset types: - Tableau Server - Tableau Site - Tableau Project - Tableau Data Source - Tableau Data Model - Tableau Data Entity - Tableau Data Attribute - Tableau Workbook - Tableau Story

Tableau version	Tableau site role	Minimum required permissions			Result in Data Catalog
		Project	Workbook	Data Source	
					• Tableau Dashboard • Tableau Worksheet

Tableau version	Tableau site role	Minimum required permissions			Result in Data Catalog
		Project	Workbook	Data Source	
2020.2 and newer	Viewer or Explorer	View	View	View	If you enabled the metadata API, Data Catalog creates new assets according to your content in Tableau without accessing metadata in Tableau databases and tables. Resulting asset types: • Tableau Server • Tableau Site • Tableau Project • Tableau Data Source • Tableau Workbook • Tableau Story • Tableau Dashboard

Tableau version	Tableau site role	Minimum required permissions			Result in Data Catalog
		Project	Workbook	Data Source	
					- Tableau Worksheet If you did not enable the Tableau metadata API, Tableau reports and data sources are ingested into Catalog, but with a limited scope. Resulting asset types: - Tableau Server - Tableau Site - Tableau Project - Tableau Data Source - Tableau Workbook - Tableau View

Tableau version	Tableau site role	Minimum required permissions			Result in Data Catalog
		Project	Workbook	Data Source	
2020.2 and newer	Tableau Server Administrator or Site Administrator	View	View	View	If the metadata API is enabled, Data Catalog creates new assets according to your content in Tableau using metadata in Tableau databases and tables. Resulting asset types: • Tableau Server • Tableau Site • Tableau Project • Tableau Data Source • Tableau Report Attribute • Tableau Data Model

Tableau version	Tableau site role	Minimum required permissions			Result in Data Catalog
		Project	Workbook	Data Source	
					• Tableau Data Entity • Tableau Data Attribute • Tableau Workbook • Tableau Story • Tableau Dashboard • Tableau Worksheet If you did not enable the Tableau metadata API, Tableau reports and data sources are ingested into Catalog, but with a limited scope. Resulting asset types: • Tableau Server • Tableau Site • Tableau

Tableau version	Tableau site role	Minimum required permissions			Result in Data Catalog
		Project	Workbook	Data Source	
					Project • Tableau Data Source • Tableau Workbook • Tableau View

Warning We do not support a full ingestion of Tableau Server or Tableau Online version 2020.2 or newer if the metadata API is disabled. If you try to synchronize a Tableau Server or Tableau Online asset after a Tableau upgrade to 2020.2 or newer without the metadata API, the synchronization result in Data Catalog will fail. This prevents data loss of manually added relations and attributes.

Tip For more information about Tableau permissions, site roles and licenses, see the Tableau Online Help.

Tableau data structure

You can only synchronize Tableau elements if the Tableau user specified in the connection settings has permissions to access them. If you have permissions to access a Tableau element, but not its parent elements, the parent elements are skipped when synchronizing Tableau and do not appear in Data Catalog.

This happens in the following situations:

- The Tableau user has permissions to access a Tableau workbook, but not its parent, the Tableau project.
- The Tableau user has permissions to access a Tableau view, but not its parent, the Tableau workbook.

- The Tableau user has permissions to access a Tableau view, but not its parent, the Tableau project.

Metadata API

If you register a Tableau Server or Tableau Online version 2020.2 or newer, Data Catalog requires the metadata API to synchronize Tableau assets.

Tableau metadata consists of information about Tableau content and assets. Data Catalog creates GraphQL queries to collect metadata from Tableau Online or Tableau Server. If the metadata API is enabled in Tableau and in Collibra Console, Collibra Data Governance Center uses this metadata to create new assets in Data Catalog.

Upgrading Tableau to 2020.2 or newer

If you have previously ingested and synchronized a version of Tableau older than 2020.2 and have since upgraded to version 2020.2 or newer, you have to enable the metadata API in Tableau and in Collibra Console. If you synchronize using the metadata API, Data Catalog removes all Tableau assets created via XML mapping and creates new ones using the metadata API. This means that all manually added relations, attributes, tags, comments and stitching results will be lost.

> Tip We highly recommend to contact your Collibra Customer Success Manager before you synchronize a Tableau Server or Tableau Online asset after upgrading to Tableau version 2020.2 or newer.

Parsing Tableau metadata

Parsing Tableau metadata is an automated procedure that allows the metadata to be captured and identified in Data Catalog at a more granular level. Typically, the result is that you have more assets of different types in Data Catalog, which leads to more complete information and better lineage diagrams.

Parsing takes place automatically during Tableau synchronization, depending on the Tableau permissions of the Tableau user who launched the synchronization process.

Parsing Tableau workbooks

Without parsing, Tableau Workbooks contain Tableau Views, without further details. However, if your Tableau user has the Download/Save As permission for the Workbook, the Tableau workbook is parsed. As a consequence, there is no Tableau View asset, but there is at least one Worksheet asset, and, if they exist on Tableau: Tableau Story assets and Tableau Dashboard assets.

Without Parsing	With Parsing
• Tableau Workbook • Tableau View	• Tableau Workbook • Tableau Story • Tableau Dashboard • Tableau Worksheet

Parsing Tableau Data Source

Without parsing, Tableau Data Sources do not contain further information about the data source. However, if your Tableau user has the Download/Save As permission for the Data Source , the Tableau Data Source is parsed. As a consequence, there is at least one Tableau Data Model asset and one or more Tableau Data Entity assets and Tableau Data Attribute assets. These assets are required for Tableau stitching.

Without parsing	With parsing
• Tableau Data Source	• Tableau Data Source ◦ Tableau Data Model ◦ Tableau Data Entity ◦ Tableau Data Attribute

Working with Tableau APIs

When you register or synchronize a Tableau Server, Data Catalog uses the Tableau APIs to ingest the Tableau metadata. Data Catalog uses different APIs depending on your version of Tableau. This happens automatically and should have little impact on the resulting assets. However, if you synchronize Tableau 2020.2 or newer, you must perform a few extra actions.

Tableau versions

The following table shows which APIs Data Catalog uses to register or synchronize a Tableau Server.

Tableau versions using the REST API and XML parsing	Tableau versions using the REST API in combination with the GraphQL metadata API
• 10.4 • 10.5 • 2018.x • 2019.x • 2020.1	• 2020.2 • 2020. 3 • 2020.4

> Warning If you upgrade to Tableau version 2020.2 or newer, but previously synchronized an older Tableau version via the REST API and XML mapping, you have to prepare the migration procedure to prevent losing manually added relations, attributes, tags, comments and stitching results.

Differences between the metadata API and XML parsing via REST API

The following table shows the differences and similarities between the metadata API and the REST API with XML parsing.

Part of synchronization process	REST API and XML parsing	Metadata API
API	Data Catalog connects to Tableau via the REST API and uses custom parsing mechanisms. The result is XML data.	Data Catalog connects to Tableau via the REST API and the metadata API. The result is GraphQL data. Note We highly recommend that you synchronize Tableau after working hours. This is necessary to make sure that no Tableau data is added, changed, renamed or deleted on Tableau's side during the synchronization process. If there are any inconsistencies between the Tableau data collected via the REST API and Tableau data collected via the GraphQL metadata API, the corresponding Tableau assets are not synchronized in Data Catalog.

<table>
<tr><th>Part of synchronization process</th><th>REST API and XML parsing</th><th>Metadata API</th></tr>
<tr><td>Settings</td><td>You don't need change any settings to start Tableau synchronization.</td><td>You have to enable the Tableau metadata API in Collibra Console before you can ingest or synchronize.
Note Also make sure that the Tableau metadata API is enabled in Tableau.</td></tr>
<tr><td>Relevant asset types</td><td colspan="2">The resulting Tableau assets that are created after registering or synchronizing a Tableau Server are similar and mainly depend on the permissions of your Tableau user.</td></tr>
<tr><td>Performance</td><td colspan="2">Performance results are similar.</td></tr>
<tr><td>Collibra Data Governance Center permissions</td><td colspan="2">The required permissions are the same: a resource role with the Configure external system resource permission.</td></tr>
<tr><td>Stitching</td><td colspan="2">Stitching works the same.</td></tr>
</table>

Migration procedure

When you synchronize a Tableau Server for the first time after you upgraded to Tableau 2020.2 or newer, Data Catalog tries to match your Tableau assets that were previously ingested via the Tableau REST API in Data Catalog to their counterparts in Tableau. If the asset names match, Data Catalog changes the full name of the Tableau assets without removing manually added data and stitching results.

To make sure Collibra Data Intelligence Cloud is able to match your Tableau assets in Data Catalog to their counterparts in Tableau, you must prepare the migration procedure.

> Tip We highly recommend that you create a backup of your Collibra environment before synchronizing a Tableau Server asset after you upgraded to Tableau 2020.2 or newer. We also recommend that you synchronize the first time after working hours.

Prepare migration after upgrading to Tableau 2020.2 or newer

If you upgraded to Tableau version 2020.2 or newer, but previously synchronized an older version via XML mapping, Data Catalog changes the full names of your Tableau assets to match them to their counterparts in Tableau. This is necessary to prevent losing manually added relations, attributes, tags, comments and stitching results.

You only have to follow these steps once after your upgrade to Tableau 2020.2 or newer. After that, you can follow the default synchronization process.

> Note Collibra Data Intelligence Cloud can only migrate your assets if:
>
> - All Tableau Report Attribute assets have the same name as their counterparts in Tableau.
> - Each Tableau Report Attribute asset name is unique within the same Tableau workbook.

> Tip If you never manually changed the name of the assets in Data Catalog, they should automatically be the same as their counterparts in Tableau.

Prerequisites

- You have registered Tableau.
- You have connected a Tableau Server asset to a Tableau Server or Tableau Online.
- You have a resource role with the Configure external system resource permission, for example Owner.

- You have a resource role with the Asset > Update resource permission.
- You have a global role with the Catalog global permission, for example Catalog Author.
- You have previously ingested Tableau 2020.1 or older and have since upgraded to Tableau 2020.2 or newer.
- Your Tableau user has the right permissions to synchronize Tableau 2020.2 or newer.
- You have enabled the Tableau metadata API in Tableau.

Steps

1. Match the names of all Tableau Report Attributes assets of a Tableau Workbook with their counterparts in Tableau.
   a. Open a Tableau Report Attribute asset page.
   b. In the resource toolbar, click **Edit**.
      » The **Edit <asset name>** dialog box appears.
   c. Change the name of the asset to the exact name used in Tableau.
   d. Click **Save**.

   > Tip We highly recommend that you also match the display names of Tableau Data Attribute assets, Tableau Data Entity assets and Tableau Data Model assets. While Data Catalog automatically tries to match these assets to their counterparts in Tableau based on the Tableau Report Attribute asset, making sure the Tableau assets have the same name helps to prevent issues. Unless you manually changed their names in Data Catalog, the names should already be the same as their counterparts in Tableau.

2. Optionally, create a backup of your Collibra environment.

   > Note We highly recommend that you create a backup before you synchronize a Tableau Server to prevent losing data in Data Catalog if something goes wrong during the migration process.

3. Enable the Tableau metadata API in Collibra Console.

4. Synchronize a Tableau Server asset after working hours.

   > Note We highly recommend that you synchronize the first time after upgrading to Tableau 2020.2 or newer after working hours. This is necessary to make sure that no Tableau data is added, changed, renamed or deleted on Tableau's side during the synchronization process.

   a. Open a Tableau Server asset page.
   b. In the tab pane, click ⚙ **Configuration**.
   c. In the **Tableau sites** section, do the following:
      i. Select one or more sites.
      ii. Enable or disable report images as required for each site.

      > Note Images are never downloaded or stored in Data Catalog. Depending on the Report image setting, Data Catalog either ignores images completely or stores a link to the image on Tableau and loads that image when you open the relevant asset page.

   d. In the **Tableau sites** section, click **Synchronize now**.

» The synchronization job appears in the **Activities** list as a bulk synchronization.

» The full names of the Tableau assets are updated to include the GraphQL ID.

» The log files show a summary of the migration process.

> Example
>
> ```
> "Summary of tableau xml to graphql data migration for
> site with id <Tableau-site-ID> and name <Tableau-site-
> name> executed on server"
> ```

The log files also show how many Tableau assets were found in Data Catalog and how many were migrated to match their counterparts in Tableau.

Example

```
"Found 50 existing xml assets to migrate.";
"Migrated 48 assets.";
```

If some Tableau assets could not be migrated, Collibra Data Intelligence Cloud recreates the Tableau asset so that it matches in Tableau. The log file shows how many and which assets were recreated in Data Catalog.

Example

```
"It was impossible to migrate 2 assets. These assets
were re-created based on graphql data.";
   List of assets that were not migrated:
       ID: xxxxxxxx-xxxx-xxxx-xxx, Fullname: Tableau-
migration > tableaumigration.xxxxx > [tableau-
migration-asset-name-1] (Tableau Report Attribute).",
       ID: xxxxxxxx-xxxx-xxxx-xxx, Fullname: Tableau-
migration > tableaumigration.xxxxx > [tableau-
migration-asset-name-2] (Tableau Report Attribute)."
```

Synchronize Tableau site manually

You can manually start a synchronization job of a Tableau Server asset. This can be useful if you don't want to wait for the scheduled job to synchronize your Tableau sites.

Warning You can choose which sites to synchronize after successfully connecting to Tableau. Select the same or more sites when you synchronize again. If you only synchronize some of the Tableau sites, Data Catalog deletes all other Tableau sites and their content from Collibra Data Governance Center.

Tip You can also add a synchronization schedule to synchronize automatically.

Prerequisites

- You have registered Tableau.
- You have connected a Tableau Server asset to a Tableau Server or Tableau Online.
- 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 Tableau community you create when you register a Tableau server:
    - Asset: add
    - Attribute: add
    - Domain: add
    - Attachment: add
- If you want to stitch Tableau's logical data layer to Data Catalog's physical data layer, the Tableau user must have the Download/Save As permission on the data source.
- You have enabled the Tableau metadata API in Collibra Console and in Tableau if you use Tableau 2020.2 or newer.

> Warning   If you upgrade to Tableau version 2020.2 or newer, but previously synchronized an older Tableau version via the REST API and XML mapping, you have to prepare the migration procedure to prevent losing manually added relations, attributes, tags, comments and stitching results.

Steps

1. Open a Tableau Server asset page.
2. In the tab pane, click ⚙ **Configuration**.

3. In the **Tableau sites** section, do the following:
   a. Select one or more sites.
   b. Enable or disable report images as required for each site.

   > Note Images are never downloaded or stored in Data Catalog. Depending on the Report image setting, Data Catalog either ignores images completely or stores a link to the image on Tableau and loads that image when you open the relevant asset page.

4. In the **Tableau sites** section, click **Synchronize now**.
   » The synchronization job appears in the **Activities** list as a bulk synchronization.

   > Note We highly recommend that you synchronize a Tableau Server version 2020.02 and newer after working hours. This is necessary to make sure that no Tableau data is added, changed, renamed or deleted on Tableau's side during the synchronization process. If there are any inconsistencies between the Tableau data collected via the REST API and Tableau data collected via the GraphQL metadata API, the corresponding Tableau assets are not synchronized in Data Catalog.

> Tip If your Tableau synchronization fails, go to the troubleshooting section to find a solution.

What's next?

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 you have stitched Tableau's logical data layer to Data Catalog's physical data layer, you have to restitch to make sure that all relations are up to date.

Add a Tableau synchronization schedule

To keep the content of Collibra Data Governance Center synchronized with your Tableau Server or Tableau online, you can synchronize manually or create a schedule to automatically do this with a fixed interval.

> Note
> - You can only create one synchronization schedule.
> - If you have stitched Tableau's logical data layer to Data Catalog physical data layer, you have to restitch after each synchronization to make sure that all relations are up to date.
> - We highly recommend that you synchronize a Tableau Server version 2020.02 and newer after working hours. This is necessary to make sure that no Tableau data is added, changed, renamed or deleted on Tableau's side during the synchronization process. If there are any inconsistencies between the Tableau data collected via the REST API and Tableau data collected via the GraphQL metadata API, the corresponding Tableau assets are not synchronized 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 a role with the following resource permissions on the Tableau community you create when you register a Tableau server:
  - Asset: add
  - Attribute: add
  - Domain: add
  - Attachment: add
- You have registered Tableau.
- You have connected a Tableau Server asset to a Tableau Server or Tableau Online.
- You have enabled the Tableau metadata API in Collibra Console and in Tableau if you use Tableau 2020.2 or newer.

> Warning If you upgrade to Tableau version 2020.2 or newer, but previously synchronized an older Tableau version via the REST API and XML mapping, you have to prepare the migration procedure to prevent losing manually added relations, attributes, tags, comments and stitching results.

Steps

1. Open a Tableau Server 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 the schemas automatically, for example daily, weekly or based on a 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 the schemas, 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 the schemas , 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 the schemas automatically, for example 14:00. 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**.

> Tip If your Tableau synchronization fails, go to the troubleshooting section to find a solution.

Edit a Tableau synchronization schedule

You can edit the synchronization schedule of a Tableau Server 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, for example Owner.
- You have a global role with the Catalog global permission, for example Catalog Author.
- You have registered Tableau.
- You have connected a Tableau Server asset to a Tableau Server or Tableau Online.
- You have added a synchronization schedule.

Steps

1. Open a Tableau Server 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 the schemas automatically, for example daily, weekly or based on a 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 the schemas, 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 the schemas , 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 the schemas automatically, for example 14:00. 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 a Tableau synchronization schedule

You can remove a synchronization schedule from a Tableau Server asset to stop automatically synchronizing Tableau.

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 Tableau.
- You have connected a Tableau Server asset to a Tableau Server or Tableau Online.
- You have added a synchronization schedule.

Steps

1. Open a Tableau Server asset page.
2. In the tab pane, click ⚙ **Configuration**.
3. In the **Synchronization schedule** section, click **Remove Schedule**.

Delete a Tableau site from Collibra DGC

You can delete a Tableau site and all of its contents from the Tableau site synchronization. Collibra Data Governance Center then deletes the community related to the Tableau site, including the domains and assets that it contains.

> Note The **Tableau sites section** on a Tableau Server asset page shows all sites that exist in Tableau. If you want to remove Tableau sites from this list, you must remove them in Tableau.

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.
- You have connected a Tableau Server asset to a Tableau Server or Tableau Online.
- You have registered Tableau.
- You have synchronized Tableau at least once.

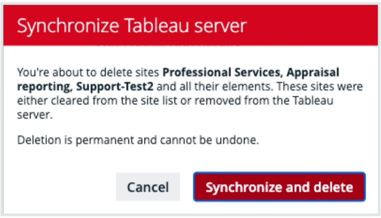
Steps

1. Open a Tableau Server asset page.
2. In the tab pane, click ⚙ **Configuration**.
3. In the **Tableau sites** section, clear the sites that you want to delete from Data Catalog.

   > Tip Only select the Tableau sites that you would like to keep. If you want to delete all Tableau sites from Data Catalog, clear all checkboxes.

4. In the **Tableau sites** section, click **Synchronize now**.

   » The **Synchronize Tableau server** dialog box appears.

   You're about to delete sites **Professional Services, Appraisal reporting, Support-Test2** and all their elements. These sites were either cleared from the site list or removed from the Tableau server.

   Deletion is permanent and cannot be undone.

5. Click **Synchronize and delete**.

   » The synchronization job appears in the **Activities** list as a bulk synchronization.

   After the synchronization, the cleared sites are deleted.

What's next?

If you deleted the wrong Tableau site or you want to reintroduce it, you can select that Tableau site and synchronize it again.

Tableau stitching

Stitching is a process that creates relations between assets representing the same data source: the data source of a Tableau report and the Data Catalog database. This allows you to clearly represent the lineage from the data source to the Tableau reports where it is used. As a consequence, you can easily perform impact analyses. For example, you can quickly see which reports will be affected if you refresh a table of your database, or which reports will be impacted if you drop one column from the table.

About Tableau stitching

Before you can perform stitching, you have to ingest a Tableau report –including its data source– and register that data source separately in Data Catalog. The same data is then represented by Tableau assets as well as by regular Data Catalog assets such as Schema, Table and Column assets. Tableau stitching is based on the matching of the full name of Tableau Data Attribute assets and Column assets of registered data sources in Data Catalog. Follow the steps in the table below to enable Collibra Data Governance Center to automatically create relations between Tableau assets and assets of a registered data source in Data Catalog.

> Note
> - You can only perform stitching if the Tableau report is based on a database. Stitching Tableau reports based on files such as CSV is not supported.
> - Tableau stitching is based on full names and is case-sensitive. As a consequence, we recommend that you do not manually edit any asset names of data sources or Tableau assets. See the Tableau naming convention for more information.

Tableau stitching steps

To use Tableau stitching, you have to prepare the assets representing the data source in Tableau's logical data layer and in Data Catalog's physical data layer:

Step	What	Simplified instructions
1	Prepare the Tableau logical data layer.	1. Register Tableau Server or Tableau Online. 2. Connect to Tableau Server or Tableau Online. 3. Synchronize Tableau sites.
2	Prepare the physical data layer.	1. Register a database as data source. 2. Create a Database asset with the same name as the data source. 3. Create a relation between the Database asset and the Schema asset using the Technology Asset has / belongs to Schema relation type.

Step	What	Simplified instructions
3	Stitch Tableau logical data layer and physical data layer.	1. On the Tableau Data Model asset page, click **Stitch with data source**.
4	View stitching results.	1. Open the asset page of the Tableau Server asset. 2. In the tab pane, click ⚬⁞ **Diagram**. 3. In the **Explore** drop-down list, select **Data Catalog Lineage 5.7**.

> Note
> - If there were changes in Tableau or the data source, you have to do the following:
>   a. Synchronize Tableau. This can be done manually or automatically, by means of a synchronization schedule.
>   b. Refresh the schema of your data source. This can be done manually or automatically, by scheduling it during data source registration.
>   c. Restitch Tableau's logical data layer or Data Catalog's physical data layer. This has to be done manually.
> - You can also remove stitching.

Data layers

Tableau's logical data layer

We call the data source in Tableau the logical data layer, because it consists of Tableau metadata, rather than the physical data. It is created when you synchronize a Tableau server. It contains Tableau report metadata, including the data source.

> Note
>
> - You can combine different data sources in one Tableau data source by using different methods, for example, **Join** or **Union**.
> - If you combine physical data sources in the Tableau data source with the **Join** method, the Tableau logical data layer is created in Data Catalog. For more information about the **Join** method, see Join Your Data.
> - If you combine physical data sources in the Tableau data source with other methods, for example, **Union**, the Tableau logical data layer is not created in Data Catalog.

Data Catalog's physical data layer

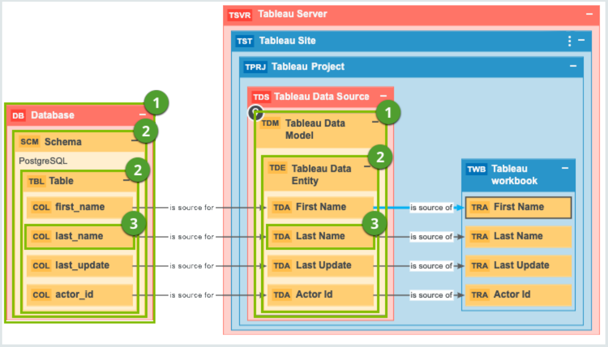
We call the data source in Data Catalog the physical data layer, which contains the physical tables and columns. It is created when you register a database as a data source. It contains the physical data of the data source.

Stitching results

Each element is represented twice in Collibra DGC: once in Tableau's logical data layer and once in Data Catalog's physical data layer.

The corresponding assets are linked by relations:

- A relation of the type "Technology Asset source system for / source system Data Asset" type between the Database asset and the Tableau Data Model asset.
- Relations of the type "Data Element targets / sources Data Element" type between the Column assets and the Data Attribute assets, based on the full names of the assets.



Number	Data Catalog's physical data layer	Tableau's logical data layer	Description
1	Database (DB)	Tableau Data Model (TDM)	An abstraction from the physical implementation of database, schema, file, etc., used for Tableau report creation.
2	Schema (SCM) and Table (TBL)	Tableau Data Entity (TDE)	An abstraction from the physical implementation of database tables, used for Tableau report creation.
3	Column (COL)	Tableau Data Attribute (TDA)	A specification that defines a property of a Tableau data entity. Examples: CustomerBirthDate, EmployeeFirstName.

Naming convention

When you ingest a data source in Tableau, Tableau automatically creates names for the data source, data model, data elements and data attributes. When you create the logical data layer by synchronizing Tableau, Data Catalog uses the names in Tableau to create the corresponding Tableau assets. As a result, in Data Catalog, Tableau assets have as a full name the same name as the original data source names in Tableau.

When you create the physical data layer by registering the data source directly in Data Catalog, you enter the names of the Schema and Database assets manually. To make stitching work, we highly recommend to use the same name as the original data source to which the Tableau assets correspond as well:

- The name of the Schema asset should match a part of the Tableau Data Entity asset's full name. For example, *database-name > schema-name*.
- The name of the Database asset should match a part of the Tableau Data Model asset's full name.

The full name of the asset should match the asset path from the asset to the database it belongs to. For example, the full name of a Column asset would be *database>schema>table>column name*.

> Warning   Editing full name of the Tableau Server or Tableau Online assets may lead to errors during the synchronization process.

Prepare the Tableau logical data layer

Before you can perform stitching, you have to prepare Tableau's logical data layer and Data Catalog's physical data layer. In this section, we describe how to prepare the logical data layer.

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.
- The Tableau user has the Download/Save As permission on the data source.

Steps

1. Register Tableau Server or Tableau Online.
2. Connect to Tableau Server or Tableau Online.
3. Synchronize Tableau sites.
   » After synchronization, the assets of the following asset types are created in Data

Catalog:

- Tableau Data Model
- Tableau Data Entity
- Tableau Data Attribute

What's next?

If you haven't done so yet, prepare the Data Catalog physical data layer.

After both the logical data layer and the physical data layer are created, you can stitch them.

Prepare the Data Catalog physical data layer for Tableau stitching

Before you can perform stitching, you have to prepare Tableau's logical data layer and Data Catalog's physical data layer. In this section, we describe how to prepare the physical data layer.

Prerequisites

- You have a with the Catalog , 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

Steps

1. Register a database as data source.

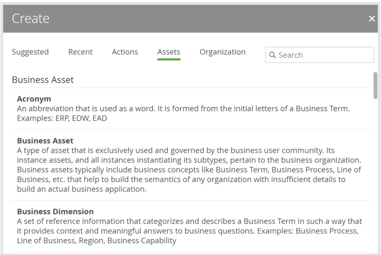
   » After registration, the assets of the following asset types are created in Data Catalog:

   - Schema
   - Table
   - Column

2. Create a Database asset.

> Tip We strongly recommend to use the name as your original data source, so that the name of the Database asset matches Tableau's naming convention.

1. Open Catalog.
2. In the main menu, click the **Create** (+) button.
3. Click the **Assets** tab.



Business Asset

Acronym
An abbreviation that is used as a word. It is formed from the initial letters of a Business Term. Examples: ERP, EDW, EAD

Business Asset
A type of asset that is exclusively used and governed by the business user community. Its instance assets, and all instances instantiating its subtypes, pertain to the business organization. Business assets typically include business concepts like Business Term, Business Process, Line of Business, etc. that help to build the semantics of any organization with insufficient details to build an actual business application.

Business Dimension
A set of reference information that categorizes and describes a Business Term in such a way that it provides context and meaningful answers to business questions. Examples: Business Process, Line of Business, Region, Business Capability

4. Click Database.
   » The **Create Asset** dialog box appears.
5. Enter the required information.

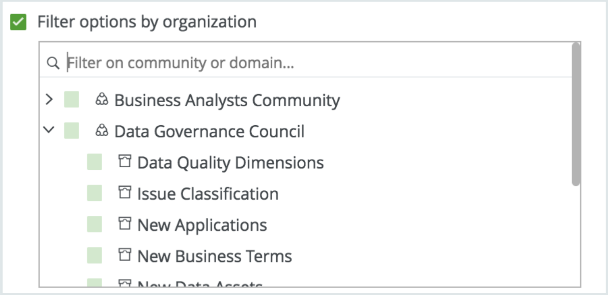
Field	Description
Type	The asset type of the asset that you are creating, in this case Database.
Domain	The domain to which the new asset will belong. You can only create a asset type in any domain of a domain type that is assigned to a Database asset type.
Name	The name of the Database asset. This has to match the name of the Tableau Data Model. Tip You can create multiple assets in one go. To do this, press `Enter` after typing a value and then type the next. Depending on the settings, asset names may have to be unique in their domain. If you type a name that already exists, it will appear in strike-through style.

6. Click **Create**.
   - » A message at the top-right of your screen confirms that one or more assets are created.

3. Create a relation between the Database asset and the Schema asset using the Technology Asset has / belongs to Schema relation type.
   a. In the tab pane, click **Add Characteristic**.
      - » The **Add a characteristic** dialog box appears.

   b. Click **Relations**.
   c. Search for and click **has schema**.
      - » The **Add has schema** dialog box appears.

d. Enter the required information.

Option	Description
Assets	The name of the schema.
Filter suggested assets by organization	Option to filter the suggestions based on selected communities and domains. If this option is selected, the organization tree appears. You can then filter and select domains and communities.  ☑ Filter options by organization Filter on community or domain... Business Analysts Community Data Governance Council Data Quality Dimensions Issue Classification New Applications New Business Terms New Data Assets
Start date	Optionally enter the date on which the relation between the assets becomes applicable. Leave this field empty to create a permanent relation.
End date	Optionally enter the date on which the relation between the assets is no longer applicable. Leave this field empty to create a permanent relation.

e. Click **Save**.

4. Check that the following relations are created for all Column assets that you want to stitch to Tableau assets:
   - Schema contains / is part of Table
   - Column is part of / contains Table

What's next?

If you haven't done so yet, prepare the Tableau logical data layer.

After both the logical data layer and the physical data layer are prepared, you can stitch them.

Supported data sources for Tableau stitching

You can stitch Tableau's logical data layer and Data Catalog's physical data layer for several data sources. The following table contains the packaged data sources and the driver versions that have been tested for Tableau stitching. We cannot guarantee that stitching works as expected for other data sources or versions.

Data source	Tested versions for Tableau stitching
Amazon Redshift	1.0.124969
HP Vertica	7.1.1-0
IBM DB2	This data source is not supported by Tableau.
MySQL	Tableau stitching is not possible because this data source has no schema.
Oracle	11.2.0.4.0
PostgreSQL	9.5.1
Microsoft SQL Server	2014 (12.0.4422.0)
Snowflake	Snowflake editions supported by Tableau

> Note Currently, we only support published Tableau data sources with an extract or a live connection. For more information, see the Tableau documentation.

Stitch the Tableau logical data layer and the Data Catalog physical data layer

You can stitch Tableau's logical data layer and Data Catalog's physical data layer to represent the lineage from the data source to the Tableau reports.

Prerequisites

- You have prepared Tableau's logical data layer.
- You have prepared Data Catalog's physical data layer.

Steps

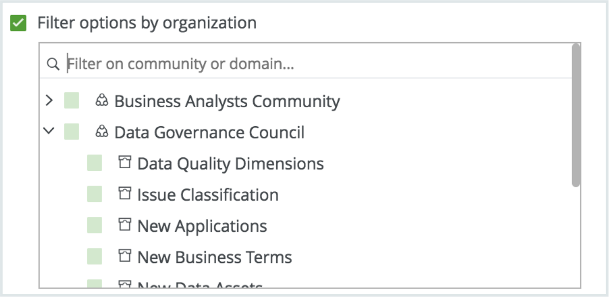
1. Open the Tableau Data Model asset page.

   > Tip You can use the Search to quickly find the relevant asset.

2. In the upper-right corner, click **Stitch**.

   » The **Stitch with data source** dialog box appears.

3. Enter the required information.

Field	Description
Data Source	The Database asset that you want to stitch to this Data Model asset.
Filter suggested assets by organization	Option to filter the suggestions based on selected communities and domains. If this option is selected, the organization tree appears. You can then filter and select domains and communities.  ☑ Filter options by organization Filter on community or domain... Business Analysts Community Data Governance Council Data Quality Dimensions Issue Classification New Applications New Business Terms New Data Assets

4. Click **Stitch**.

> Note
> If a relation exists between the Tableau Data Model and the corresponding Database asset, of the "Technology Asset source system for / source system Data Asset" type, stitching happens immediately after clicking **Stitch**, without showing the dialog box.
>
> This occurs if you created the relation manually, or if you restitch.

What's next?

Stitching is performed, creating relations between assets of Data Catalog's physical data layer and those of Tableau's logical data layer.

More precisely, these relations are created:

- A relation of the type "Technology Asset source system for / source system Data Asset" type between the Database asset and the Tableau Data Model asset.

- Relations of the type "Data Element targets / sources Data Element" type between the Column assets and the Data Attribute assets, based on the full names of the assets.

> Tip You can view the stitching result as a diagram.

Restitch the Tableau logical data layer and the Data Catalog physical data layer

After you completed stitching, there might be changes in Tableau or in the data source. For example, Tableau may have a new report and the data source may have a new column. To make sure that the lineage diagrams are also updated, you can restitch the data layers.

Prerequisites

- You have previously stitched Tableau's logical data layer and Data Catalog's physical data layer.
- 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 resource role with the Attribute > Add resource permission.

Steps

1. Ensure that Tableau's logical data layer is synchronized.
2. Ensure that Data Catalog's physical data layer is refreshed.
3. Open the Tableau Data Model asset page.

   > Tip You can use the Search to quickly find the relevant asset.

4. In the upper-right corner, click **Stitch**.

View stitching results

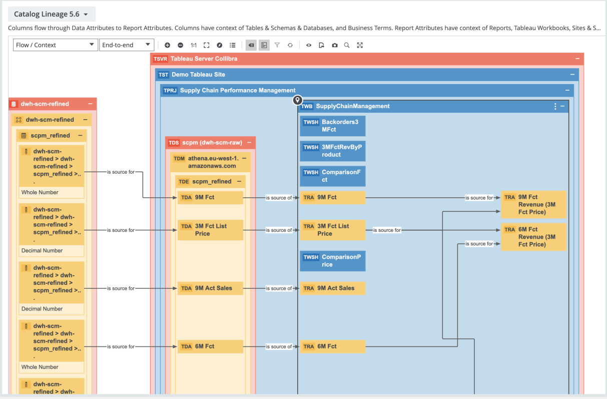
When stitching is complete, you can view the end-to-end lineage between the database and the Tableau report.

Prerequisites

- You have prepared Tableau's logical data layer.
- You have prepared Data Catalog's physical data layer.
- You have stitched the logical data layer and the physical data layer.

Steps

1. Open the Tableau Server asset page.
2. In the tab pane, click **Diagram**.
3. In the view selector, select **Data Catalog Lineage 5.7**.

> Example
>
> 
>
> Catalog Lineage 5.6
>
> Columns flow through Data Attributes to Report Attributes. Columns have context of Tables & Schemas & Databases, and Business Terms. Report Attributes have context of Reports, Tableau Workbooks, Sites & S...
>
> Flow / Context
>
> End-to-end
>
>

Remove stitching between the Tableau logical data layer and the Data Catalog physical data layer

You can remove stitching to remove the relations between the logical data layer in Tableau and the physical data layer in Data Catalog.

More precisely, the following relations are removed:

- A relation of the type "Technology Asset source system for / source system Data Asset" type between the Database asset and the Tableau Data Model asset.
- Relations of the type "Data Element targets / sources Data Element" type between the Column assets and the Data Attribute assets, based on the full names of the assets.

Prerequisites

- You have a resource role with the Configure external system resource permission, for example Owner.
- You have a resource role with the Attribute > Remove resource permission.
- You have a global role with the Catalog global permission, for example Catalog Author.
- You have stitched Tableau's logical data layer and Data Catalog's physical data layer.

Steps

1. Open the Tableau Data Model asset page.

   > Tip   You can use the Search to quickly find the relevant asset.

2. Click **Actions** → **Remove stitching**.

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

Tableau provisioning

With Data Catalog, you can create data sets and convert them to the Tableau format. This enables you to use Collibra DGC-managed data in Tableau.

The Tableau provisioning file

A Tableau provisioning file is a packaged data source file with the extension TDSX. The packaged data source file is a ZIP file that contains a data source file and any local file data sources. You can import it in Tableau to, for example, analyze the data. It has the extension TDSX.

You can create a Tableau provisioning file from any data set in Data Catalog.

The file contains the following information:

- A TDS file: This is an XML file that contains the data source definition.
- The actual ingested files, if the data set contains data from Excel or CSV data sources.

```
<?xml version="1.0" encoding="UTF-8"?>
<datasource xmlns:user-
r="http://www.tableausoftware.com/xml/user"
formatted-name="<name of your data set>" inline="true" ver-
sion="10.0">
   <connection class="federated">
       <named-connections>
           <named-connection caption="public" name="<connection-ID>">
               <connection authentication="username-password" class-
s="<data-source-type>" dbname="<database-ID>" port="" schem-
a="public" server="<hostname:port>"/>
           </named-connection>
       </named-connections>
       <relation connection="<relation-ID>" name="<name-of-rela-
tion>" table="[public].[<name-of-relation>]" type="table"/>
   </connection>
</datasource>
```

Required JDBC driver information for Tableau provisioning

To create a Tableau provisioning file from a data set, the JDBC driver of its data source needs the following properties:

Data source	Required connection properties
Amazon Redshift	• host • port • database • schema
HP Vertica	• host • port • database • schema
MySQL	• host • port • database
Oracle	• host • port • database • schema
PostgreSQL	• host • port • database • schema
SQL Server	• host • port • database • schema

For more information, see the JDBC configuration details of the various databases.

Create Tableau provisioning file

In Data Catalog, you can create Tableau provisioning files from data sets.

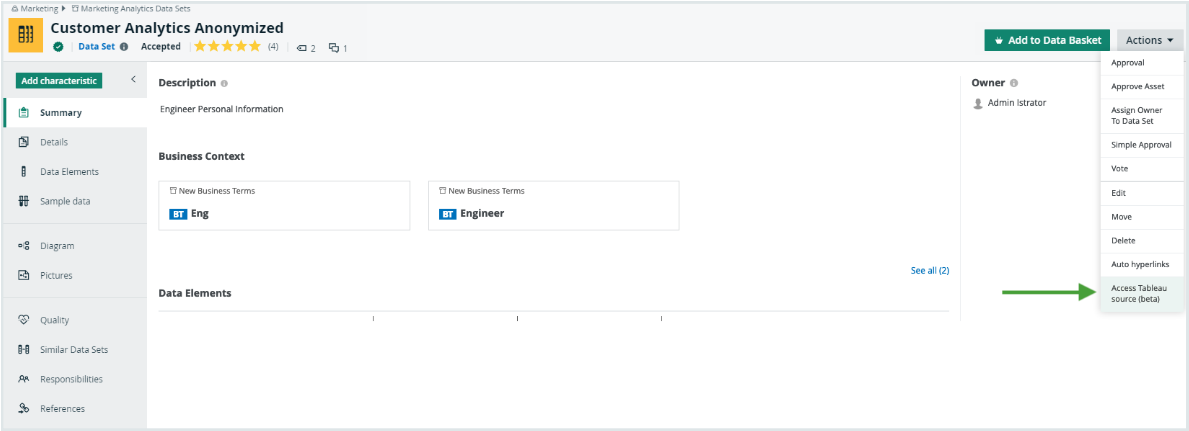
> Tip If your data set's origin is a relational database, you need the credentials to connect to that database. Make sure the JDBC driver has all the required information in the correct format before you create the provisioning file.

Prerequisites

- You have a resource role with the Access data resource permission, for example Data Analyst Level 2.
- You have a global role with the Catalog global permission, for example Catalog Author.
- You have enabled Tableau provisioning in Collibra Console.

Steps

1. In the main menu, click ⠿, then 🛢 **Catalog**.
   » The Catalog Home opens.
2. In the submenu, click **Data Sets**
3. Click the data set that you want to use in Tableau.
4. Above the table, to the right, click **Actions** → **Access Tableau source (beta)**.



Engineer Personal Information

Admin Istrator

» The Tableau provisioning file in TDSX format is downloaded.

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

What's next?

You can now import the TDSX file in Tableau.

Troubleshooting

The following table contains the most common issues that you can encounter while ingesting or synchronizing Tableau.

Issue	Solution
Tableau images are not fetched correctly	Synchronizing Tableau data means ingesting metadata from Tableau to your Collibra Data Intelligence Cloud environment. The metadata is represented as assets of specific types and their characteristics. Images such as report thumbnails are not downloaded and stored in Data Catalog. Instead, Data Catalog stores a link to the image. Every time you open the asset page, the image is fetched from Tableau. Images are not fetched correctly if there is a problem with this link. A common issue is caused by the base URL parameter, which is part of the link. If the base URL is not set correctly in Collibra Console, the links to the Tableau images are broken. To fix this issue, edit the base URL in Collibra Console.

Issue	Solution
When you synchronize a Tableau Server 2020.2, some Tableau data is skipped.	In most cases, this occurs when people are actively using Tableau while Data Catalog is synchronizing the Tableau Server. The technical reason is that the APIs collect Tableau data at different times. If users make changes in Tableau, the data that is collected by the APIs may be inconsistent. When that happens, the corresponding assets are not synchronized in Data Catalog. We highly recommend that you synchronize Tableau after working hours. This reduces the chance that Tableau data is added, changed, renamed or deleted on Tableau's side during the synchronization process.
Tableau synchronization fails with error message `Duplicate key`.	The Tableau synchronization fails with the `Duplicate key` error when you have multiple views with the same name in the same workbook. To solve this problem, we highly recommend to give each view in Tableau a unique name before you synchronize the Tableau Server in Data Catalog.

Catalog workflows

To keep the information flows that are shipped with the Catalog product configurable, a part of the functionality is achieved through workflows. You can configure the packaged workflows, but they are designed to work together: if you decide to change one of the workflows, verify the other Catalog workflows, since they may depend on one another.

> Tip For more information about workflows, see the Collibra Developer portal.

Name	Description
Assessments	This process notifies the Business Steward (by default) that an Assessment Review asset is ready for review and prompts the Business Steward to approve or reject the asset.
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.

Name	Description
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 Data Governance Center. 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 shopping cart. All data owners have to approve the request before you can access the assets. More information 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. Note This workflow replaces the Request Data Sets Access workflow. If you restore a 5.4.x backup or older, the Requests Data Sets Access will overwrite the packaged Request Assets Access workflow. You have to deploy the Requests Assets Access workflow again and apply all possible customizations to the new workflow. You can also manually request access to data sets access to data sets and reports.
Simple Approval	The Simple Approval workflow is a single-step process that allows you to approve an asset in Collibra Data Governance Center.
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.

Chapter 12

Catalog Troubleshooting

If you are experiencing general issues with the Data Catalog feature, consult the articles in this section.

If you have issues with ingesting a BI source or with Collibra Data Lineage, please visit their individual troubleshooting sections:

- Tableau troubleshooting
- Power BI troubleshooting
- Looker troubleshooting
- Collibra Data Lineage troubleshooting

What's the difference between Data Catalog and Collibra Connect?

Data Catalog and Collibra Connect have many overlapping features. Which of them is more suited for your situation, depends on a number of factors.

In a nutshell, you use Data Catalog for ingesting metadata from popular database types via a predefined ingestion logic, which is ideal for business users. You can then see the metadata in the form of assets and characteristics. You use Collibra Connect to read and write metadata in any API-supported system and provide the metadata to Collibra Data Governance Center. Collibra Connect has more flexibility with regard to ingestion, but requires technical skills.

	Data Catalog	Collibra Connect
Definition	The Collibra Data Catalog is an application that helps the business data analyst to discover, describe, assemble and govern data sets, in order to improve trust in analytics based on those data sets.	Collibra Connect is an integration platform that enables integrations between Collibra DGC and other third-party products, such as Informatica, Salesforce.com and JIRA.
Purpose	Data Catalog can ingest and represent metadata of specific data sources as assets and characteristics, including diagrams.	Collibra Connect is meant as an advanced interface between Collibra DGC and data sources of any third-party vendors.
Processes	• Metadata ingestion • Profiling and data type detection • Read only	• Bidirectional synchronization of metadata • No profiling • Read and write
Integrations	• JDBC-supported databases such as PostgreSQL and IBM DB2. • File-based databases in Excel and CSV. • External systems such as Tableau and Amazon S3.	Any system with: • API support • Structured metadata format such as XML and JSON
Ingestion	Predefined metamodel and ingestion logic	Flexible and configurable metamodel and ingestion logic
Usability	• Usable via Collibra DGC • Business user friendly	• Configuration via IDE • Requires development skills to set up

	Data Catalog	Collibra Connect
More information	- The Data Catalog: What it is, Why you Need it, and How to Make it Successful - The Data Catalog section of the Collibra DGC user guide.	- Introduction to Collibra Connect - The Collibra Connect user guide.

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, see Environment log settings for DGC services and 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.dgc.catalog.service.schema.impl` ◦ `com.collibra.dgc.catalog.service.impl` ◦ `com.collibra.jobserver.client` ◦ `com.collibra.dgc.catalog.service.datausage.impl` ◦ `com.collibra.catalog.core.service.datausage.impl` ◦ `com.collibra.catalog.core.service.schema.impl` ◦ `com.collibra.catalog.core.service.schema.impl.ingestion` ◦ `com.collibra.catalog.core.service.schema.impl.profiling` ◦ `com.collibra.catalog.core.service.schema.impl.report` ◦ `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.

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.
1. Open a terminal session on the server that hosts the jobserver.
2. Open the file **/opt/collibra/spark-defaults.conf** and do the following.
   a. Add the following line to the configuration file:

   ```
   spark.driver.maxResultSize = 1536m
   ```

   b. Save and close the file.
3. Open the **/opt/collibra/spark-jobserver/conf/log4j-server.properties** file and do the following.
   a. In the `Root logger option` section, update the properties to match this section:

   ```
   # Root logger option
   log4j.rootLogger=INFO,LOGFILE
   log4j.appender.LOGFILE=org.apache.log4j.RollingFileAppender
   log4j.appender.LOGFILE.File=${LOG_DIR}/spark-job-server.log
   log4j.appender.LOGFILE.layout=org.apache.log4j.PatternLayout
   log4j.appender.LOGFILE.layout.ConversionPattern=%d{yyyy-MM-dd HH:mm:ss.SSS} %-5p [%t] %c{3} - %m%n
   log4j.appender.LOGFILE.maxFileSize=100MB
   log4j.appender.LOGFILE.maxBackupIndex=30
   log4j.logger.org.apache.spark=WARN
   log4j.logger.spark.jobserver.context=WARN
   log4j.logger.akka=WARN
   log4j.logger.com.collibra.jobserver.job=DEBUG
   log4j.logger.com.collibra.catalog.profilers=DEBUG
   log4j.logger.com.collibra.catalog.profilers.Pass1TableProfiler$=INFO
   log4j.logger.com.collibra.catalog.ingestion=DEBUG
   log4j.logger.com.collibra.jdbc=DEBUG
   ```

   b. Save and close the file.
4. Start the environment again.

Ingestion out-of-memory error

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.

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.

Different versions for Collibra DGC and Jobserver

You can install the services of a Collibra Data Governance Center 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] - Package com.collibra.jobserver.dto.catalog.ingestion, different version detected: client uses version 1.2.4-SNAPSHOT, server uses version 1.2.2-SNAPSHOT
  ```
- Collibra DGC 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 DGC services with the same installer.

Resolve schema refresh conflicts via Jobserver

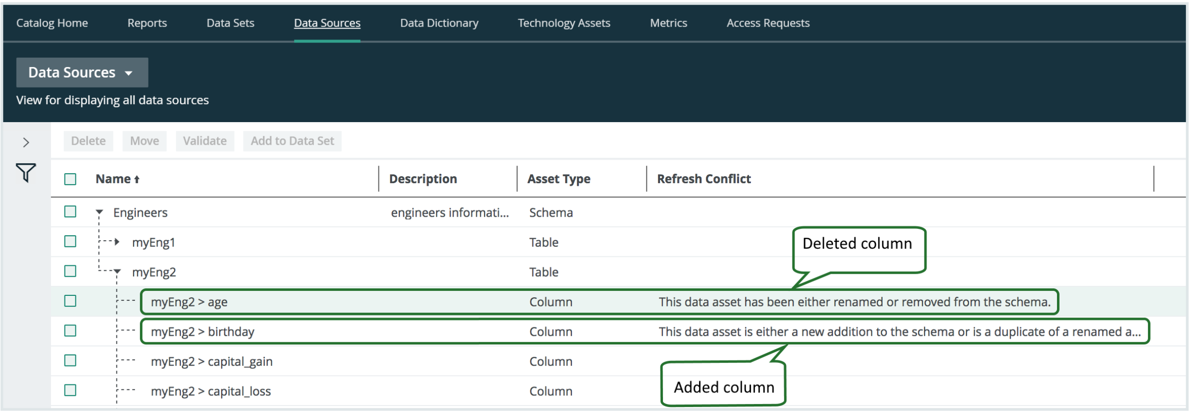
If you refresh a schema via Jobserver, the ingestion process detects differences between the original schema, already in Collibra Data Governance Center, and the updated schema.

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 DGC. 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.

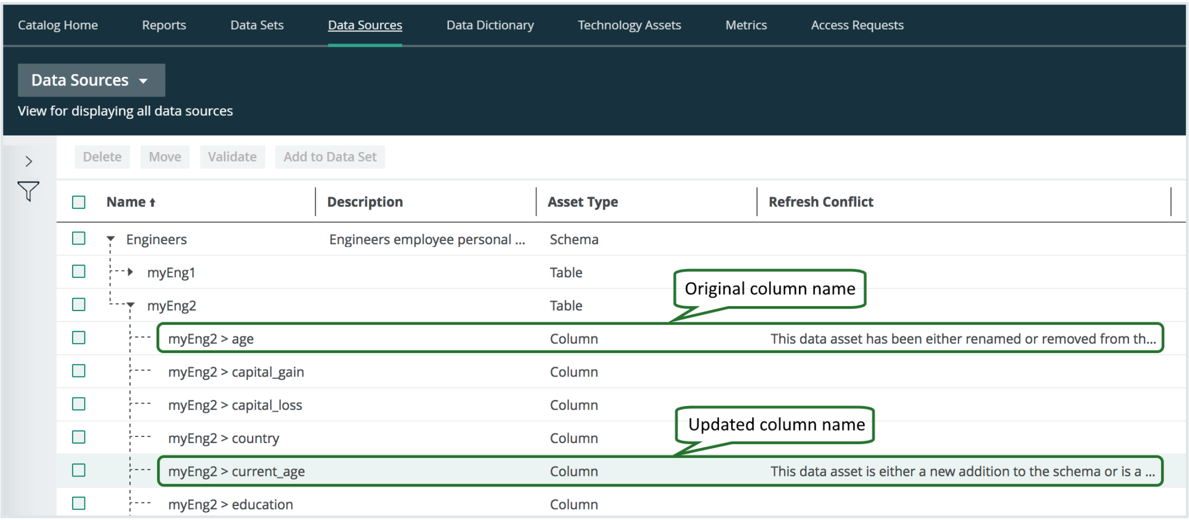


☐	Name ↑	Description	Asset Type	Refresh Conflict
☐	▾ Engineers	engineers informati...	Schema	
☐	▸ myEng1		Table	
☐	▾ myEng2		Table	
☐	myEng2 > age		Column	This data asset has been either renamed or removed from the schema.
☐	myEng2 > birthday		Column	This data asset is either a new addition to the schema or is a duplicate of a renamed a...
☐	myEng2 > capital_gain		Column	
☐	myEng2 > capital_loss		Column	

Deleted column

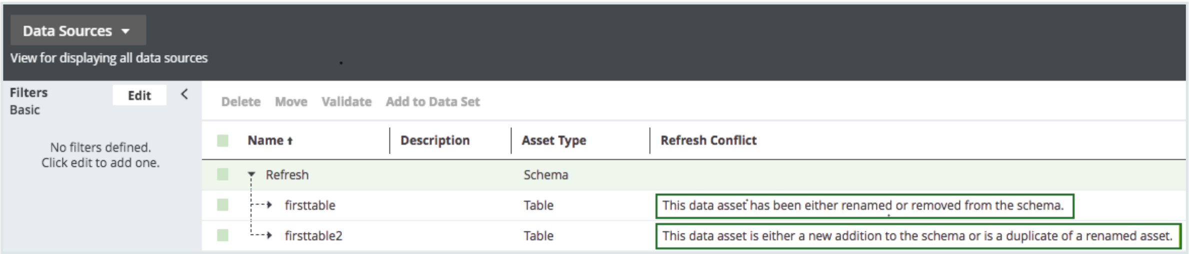
Added column

- 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.



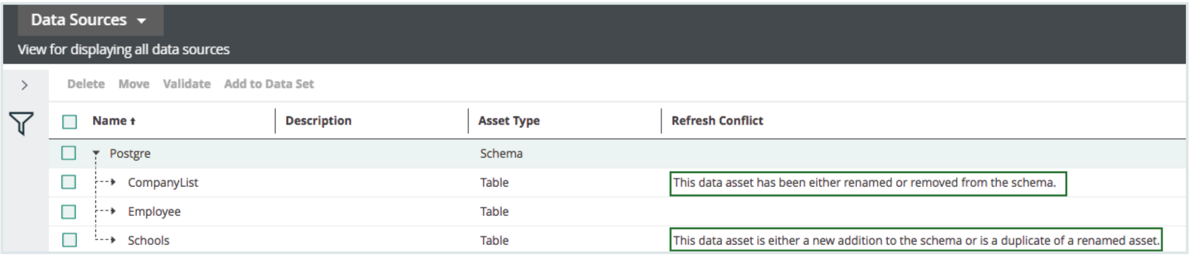
□	Name ↑	Description	Asset Type	Refresh Conflict
□	Engineers	Engineers employee personal ...	Schema	
□	myEng1		Table	
□	myEng2		Table	
□	myEng2 > age		Column	This data asset has been either renamed or removed from th...
□	myEng2 > capital_gain		Column	
□	myEng2 > capital_loss		Column	
□	myEng2 > country		Column	
□	myEng2 > current_age		Column	This data asset is either a new addition to the schema or is a ...
□	myEng2 > education		Column	

- 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.



	Name ↑	Description	Asset Type	Refresh Conflict
	Refresh		Schema	
	firsttable		Table	This data asset has been either renamed or removed from the schema.
	firsttable2		Table	This data asset is either a new addition to the schema or is a duplicate of a renamed asset.

- 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.



Data Sources ▾

View for displaying all data sources

Delete Move Validate Add to Data Set

□	Name ↑	Description	Asset Type	Refresh Conflict
□	Postgre		Schema	
□	CompanyList		Table	This data asset has been either renamed or removed from the schema.
□	Employee		Table	
□	Schools		Table	This data asset is either a new addition to the schema or is a duplicate of a renamed asset.

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.

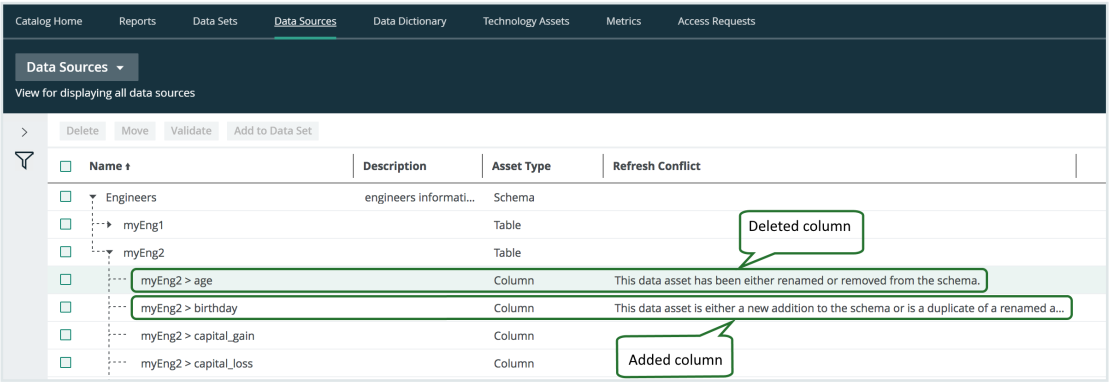
In the following example, the ingested schema has a column **age** and in the updated schema, the column **age** is removed and a column **birthday** is added.

To resolve such a refresh conflict, follow these steps:

1. Look up the data source with the search function or as follows:
   a. In the main menu, click ▦, then 🛢 **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.



Data Sources ▾

View for displaying all data sources

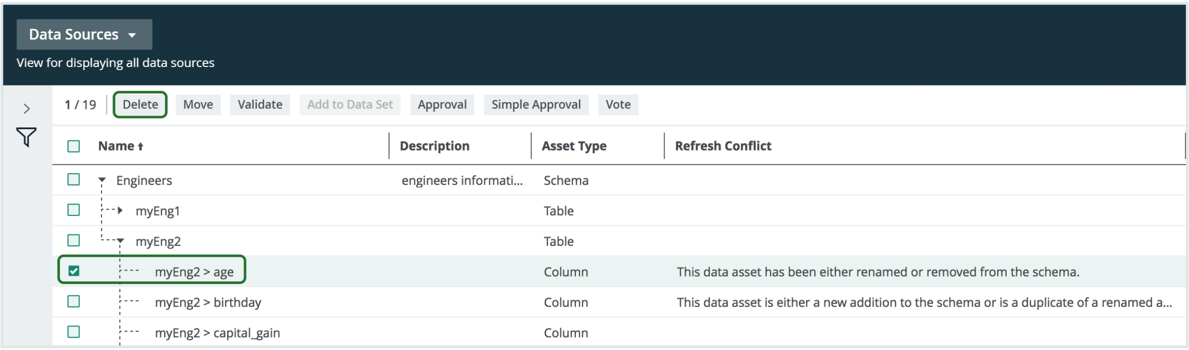
Delete Move Validate Add to Data Set

□	Name ↑	Description	Asset Type	Refresh Conflict
□	Engineers	engineers informati...	Schema	
□	myEng1		Table	
□	myEng2		Table	
□	myEng2 > age		Column	This data asset has been either renamed or removed from the schema.
□	myEng2 > birthday		Column	This data asset is either a new addition to the schema or is a duplicate of a renamed a...
□	myEng2 > capital_gain		Column	
□	myEng2 > capital_loss		Column	

Deleted column

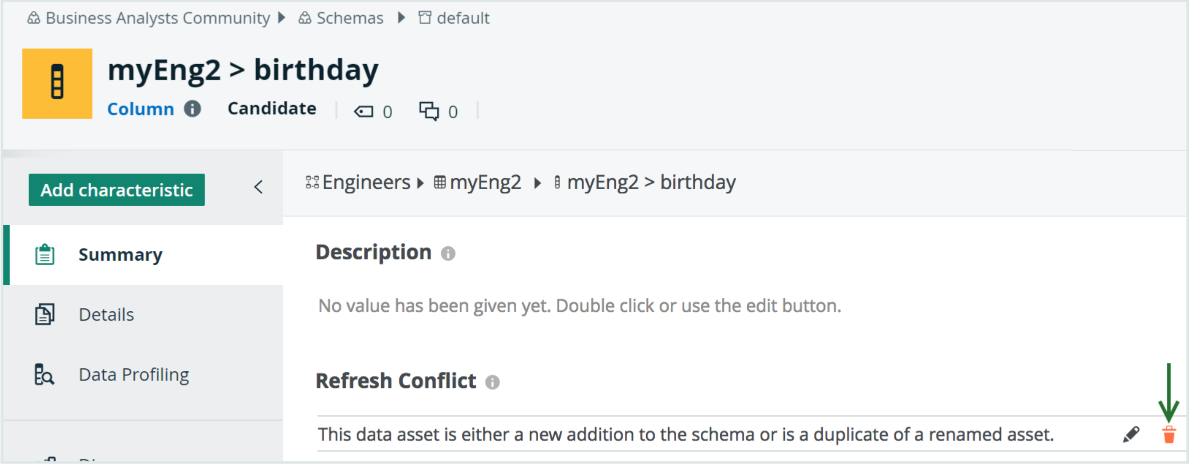
Added column

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**.



	Name ↑	Description	Asset Type	Refresh Conflict
□	Engineers	engineers informati...	Schema	
□	myEng1		Table	
□	myEng2		Table	
☑	myEng2 > age		Column	This data asset has been either renamed or removed from the schema.
□	myEng2 > birthday		Column	This data asset is either a new addition to the schema or is a duplicate of a renamed a...
□	myEng2 > capital_gain		Column	

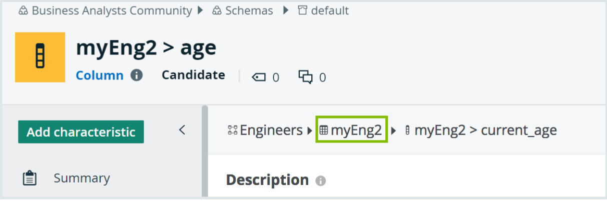
4. Click **Yes** to confirm the deletion of the column.
5. Click the name of the added column name.
   » The column asset page appear.
6. In the **Refresh Conflict** section of the column asset page, hover over the message and click 🗑 on the right-hand side.



No value has been given yet. Double click or use the edit button.

This data asset is either a new addition to the schema or is a duplicate of a renamed asset.

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.'


Business Analysts Community ▸ Schemas ▸ default
myEng2 > age
Column
Candidate
0
0
Add characteristic
Engineers ▸ myEng2 ▸ myEng2 > current_age
Summary
Description

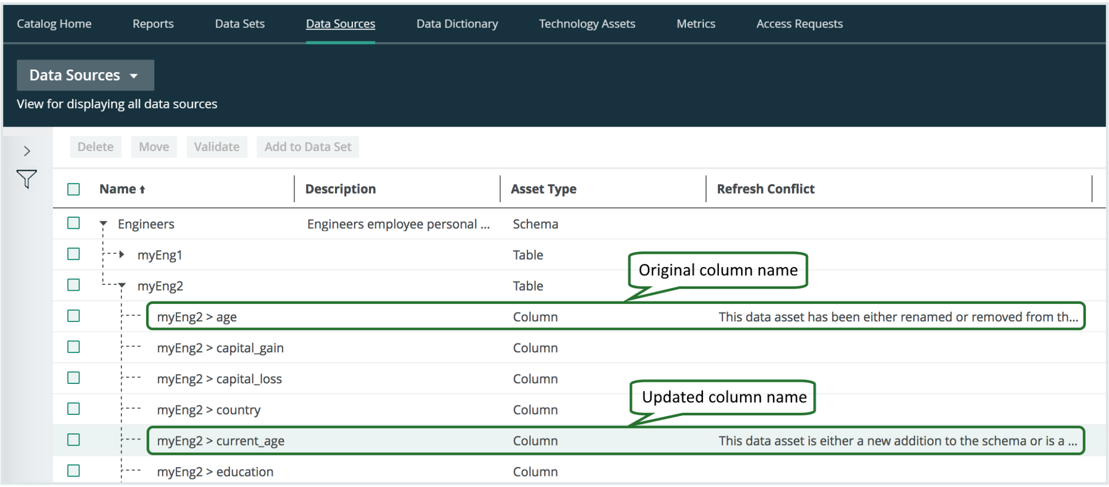
9. Repeat steps 5 to 8 for all other added columns.

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. In the main menu, click ⠿, then ⛁ **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.

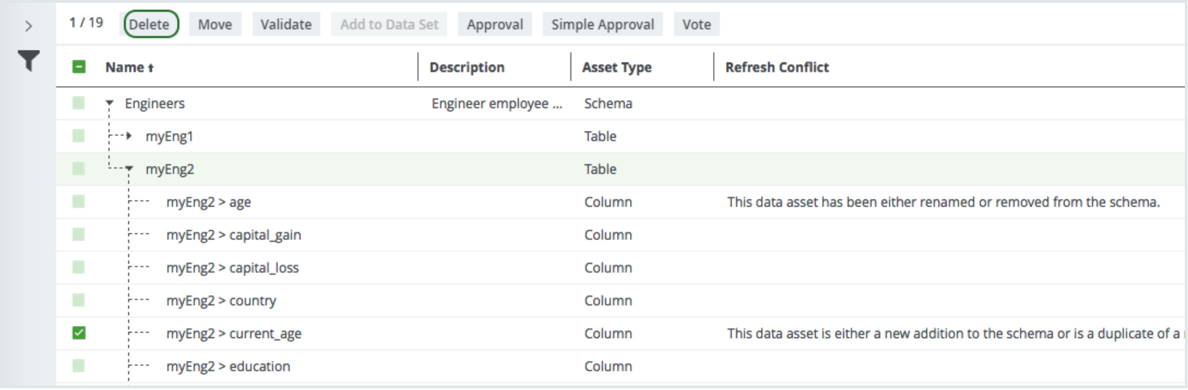


View for displaying all data sources

□	Name ↑	Description	Asset Type	Refresh Conflict
□	▾ Engineers	Engineers employee personal ...	Schema	
□	▸ myEng1		Table	
□	▾ myEng2		Table	
□	myEng2 > age		Column	This data asset has been either renamed or removed from th...
□	myEng2 > capital_gain		Column	
□	myEng2 > capital_loss		Column	
□	myEng2 > country		Column	
□	myEng2 > current_age		Column	This data asset is either a new addition to the schema or is a ...
□	myEng2 > education		Column	

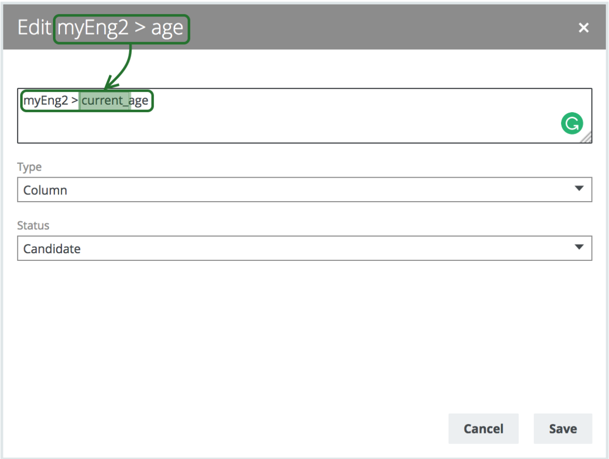
2. Select the updated column name and click **Delete** above the table.

   If necessary, select all column assets that are removed from the data source.



	Name	Description	Asset Type	Refresh Conflict
☐	Engineers	Engineer employee ...	Schema	
☐	myEng1		Table	
☐	myEng2		Table	
☐	myEng2 > age		Column	This data asset has been either renamed or removed from the schema.
☐	myEng2 > capital_gain		Column	
☐	myEng2 > capital_loss		Column	
☐	myEng2 > country		Column	
☑	myEng2 > current_age		Column	This data asset is either a new addition to the schema or is a duplicate of a
☐	myEng2 > education		Column	

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. In the resource toolbar, click **Actions** > **Edit**.
   » The **Edit <asset name>** dialog box appears.
6. Change the name to the new ingested name.



Edit myEng2 > age

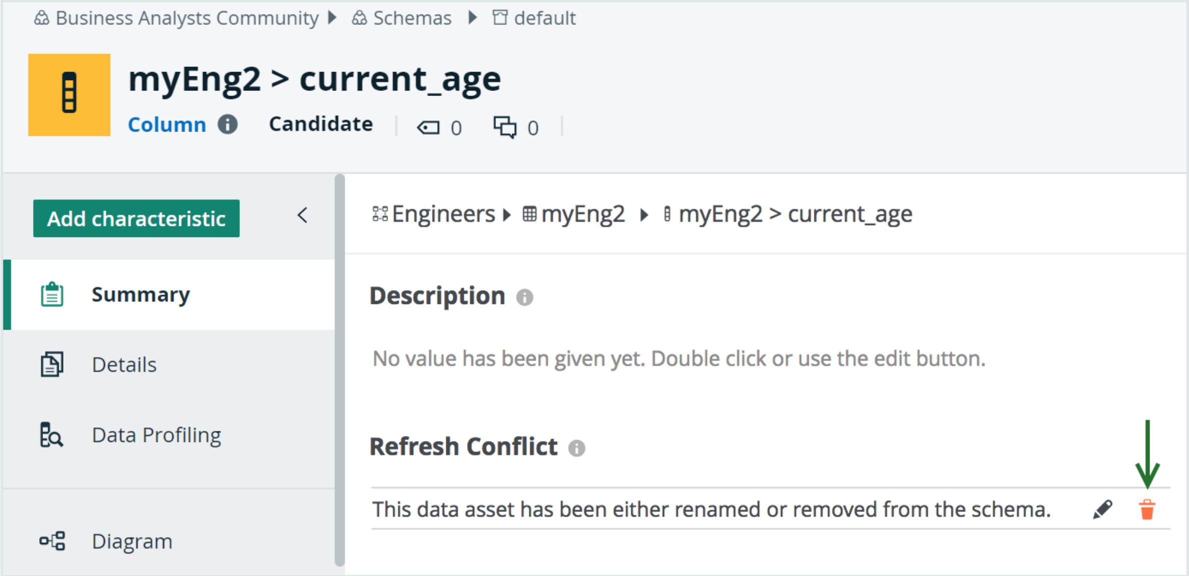
myEng2 > current_age

Type: Column

Status: Candidate

7. Click **Save**.
8. Refresh the page.
9. Leave the column asset page open.

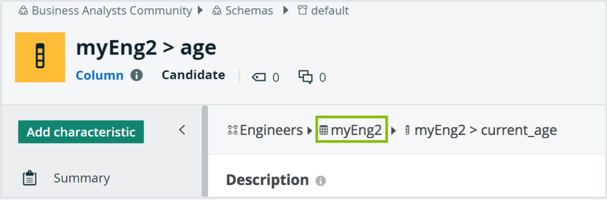
10. In the **Refresh Conflict** section of the column asset page, hover over the message and click 🗑 on the right-hand side.



No value has been given yet. Double click or use the edit button.

This data asset has been either renamed or removed from the schema.

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.

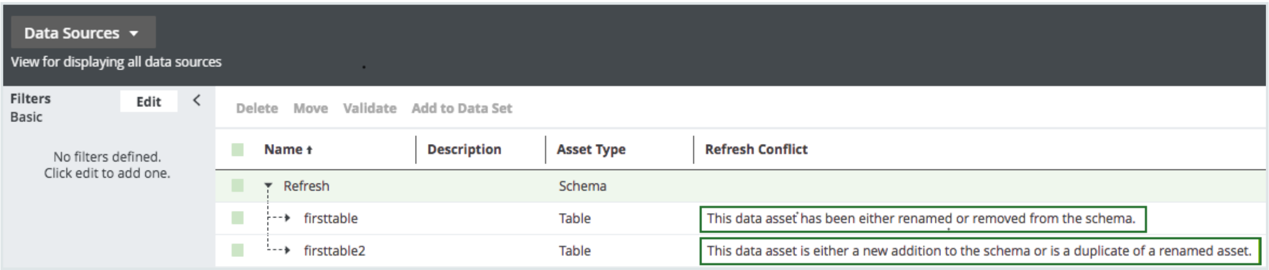
What's next?

You can now safely refresh the schema with the new data source; however, keep in mind this may take some time.

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:



Name	Description	Asset Type	Refresh Conflict
Refresh		Schema	
firsttable		Table	This data asset has been either renamed or removed from the schema.
firsttable2		Table	This data asset is either a new addition to the schema or is a duplicate of a renamed asset.

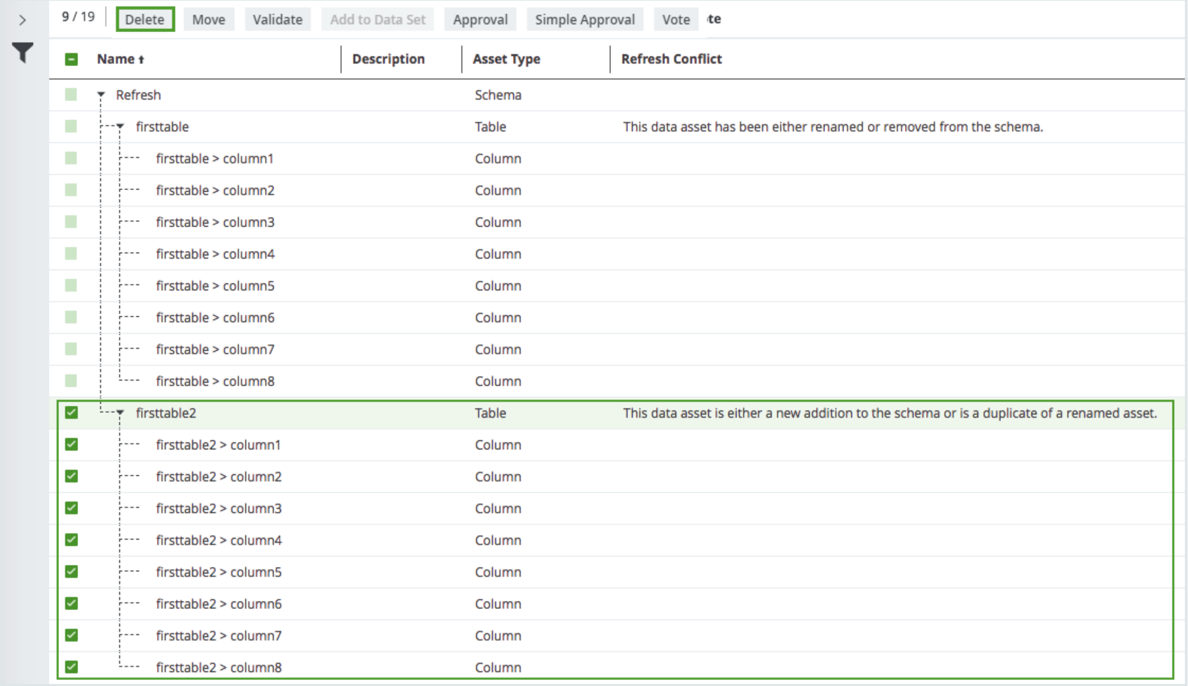
You have to manually resolve the conflicts before you continue. It is not possible to refresh a schema when there are conflicts.

> Note You have to add the **Refresh Conflict** column to the table if it is not there already.

Steps

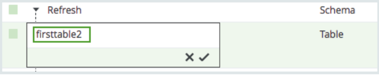
1. In the main menu, click ⠿, then 🛢 **Catalog**.
   - » The Catalog Home opens.
   - » The Catalog Home appears
2. In the submenu, click **Data Sources**.
3. Expand the tables to see all the columns that are contained in them.
4. Select the renamed table and all its contained columns, in this example, **firsttable2**.

5. Above the table, click **Delete**.



☐	Name	Description	Asset Type	Refresh Conflict
☐	Refresh		Schema	
☐	firsttable		Table	This data asset has been either renamed or removed from the schema.
☐	firsttable > column1		Column	
☐	firsttable > column2		Column	
☐	firsttable > column3		Column	
☐	firsttable > column4		Column	
☐	firsttable > column5		Column	
☐	firsttable > column6		Column	
☐	firsttable > column7		Column	
☐	firsttable > column8		Column	
☑	firsttable2		Table	This data asset is either a new addition to the schema or is a duplicate of a renamed asset.
☑	firsttable2 > column1		Column	
☑	firsttable2 > column2		Column	
☑	firsttable2 > column3		Column	
☑	firsttable2 > column4		Column	
☑	firsttable2 > column5		Column	
☑	firsttable2 > column6		Column	
☑	firsttable2 > column7		Column	
☑	firsttable2 > column8		Column	

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.

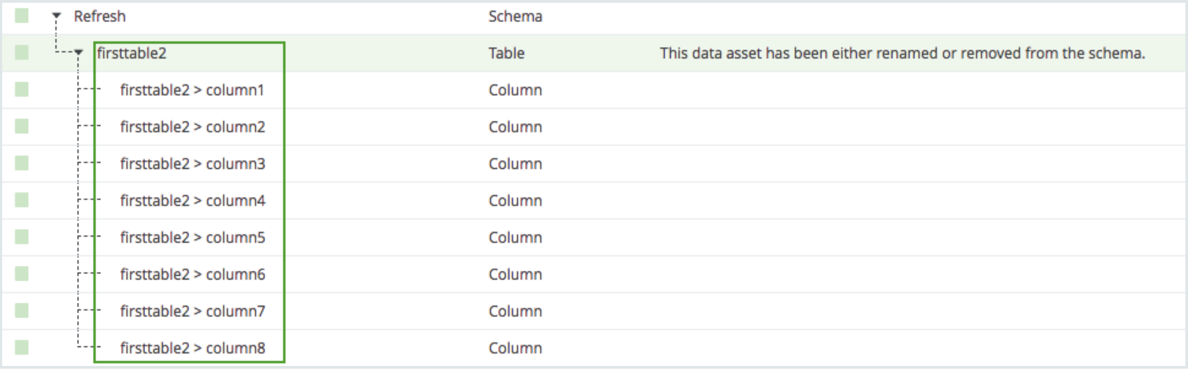


☐	Refresh	Schema
☐	firsttable2 ✕ ✓	Table

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.

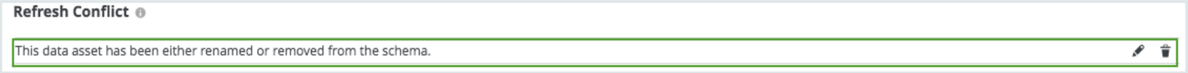
    The column name is a concatenation of the table name and the original column name and so you just have to replace the table part of the name with the new table name. For example, to rename the column name **firsttable > column1** to **firsttable2 > column1**, you just have to change **firsttable** to **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.



□	Refresh	Schema	
□	firsttable2	Table	This data asset has been either renamed or removed from the schema.
□	firsttable2 > column1	Column	
□	firsttable2 > column2	Column	
□	firsttable2 > column3	Column	
□	firsttable2 > column4	Column	
□	firsttable2 > column5	Column	
□	firsttable2 > column6	Column	
□	firsttable2 > column7	Column	
□	firsttable2 > column8	Column	

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.



Refresh Conflict

This data asset has been either renamed or removed from the schema.

14. Click **Yes** to confirm the deletion of the Refresh Conflict attribute.

What's next?

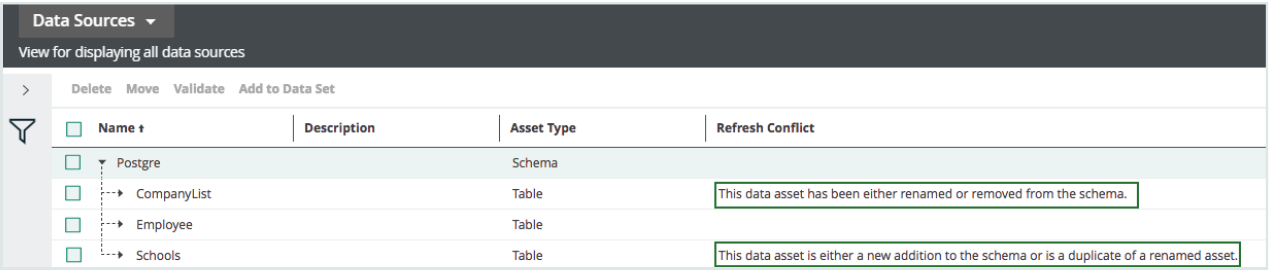
You can now safely refresh the schema with the data source.

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:



Data Sources ▾
View for displaying all data sources
Delete Move Validate Add to Data Set

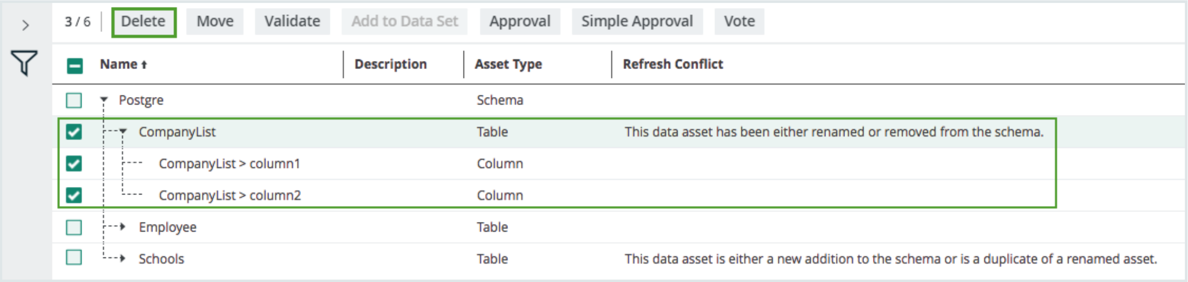
	Name	Description	Asset Type	Refresh Conflict
□	Postgre		Schema	
□	CompanyList		Table	This data asset has been either renamed or removed from the schema.
□	Employee		Table	
□	Schools		Table	This data asset is either a new addition to the schema or is a duplicate of a renamed asset.

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

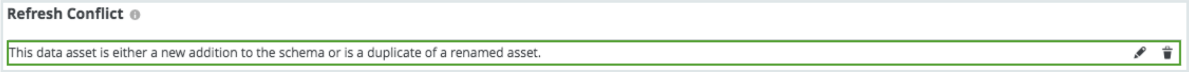
1. In the main menu, click ▦, then 🛢 **Catalog**.
   - » The Catalog Home opens.
   - » The Catalog Home appears.
2. In the submenu, click **Data Sources**.
3. Select the deleted table and all its contained columns, in this example, **CompanyList**.



3 / 6 Delete Move Validate Add to Data Set Approval Simple Approval Vote

	Name	Description	Asset Type	Refresh Conflict
□	Postgre		Schema	
☑	CompanyList		Table	This data asset has been either renamed or removed from the schema.
☑	CompanyList > column1		Column	
☑	CompanyList > column2		Column	
□	Employee		Table	
□	Schools		Table	This data asset is either a new addition to the schema or is a duplicate of a renamed asset.

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.



Refresh Conflict

This data asset is either a new addition to the schema or is a duplicate of a renamed asset.

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.

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.

Jobserver troubleshooting

This is a list of known issues in versions older than Collibra DGC 5.7.13.

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 DGC.
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

Jobserver jobs

To ingest data in Collibra Data Governance Center, 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 DGC 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.