



# ONE – Workshop

## Catalog & Profiling

Prepared for: v15.4

Prepared by: Ataccama

Dated: November 2024

## Contents of the Document

Introduction	3
Tasks	3
1. Creating a Data Source	3
2. Creating Catalog Items through profiling a Data Source	7
3. Reviewing new Catalog Items	10
4. Creating a SQL Catalog Item	15
Conclusion	19

# Introduction

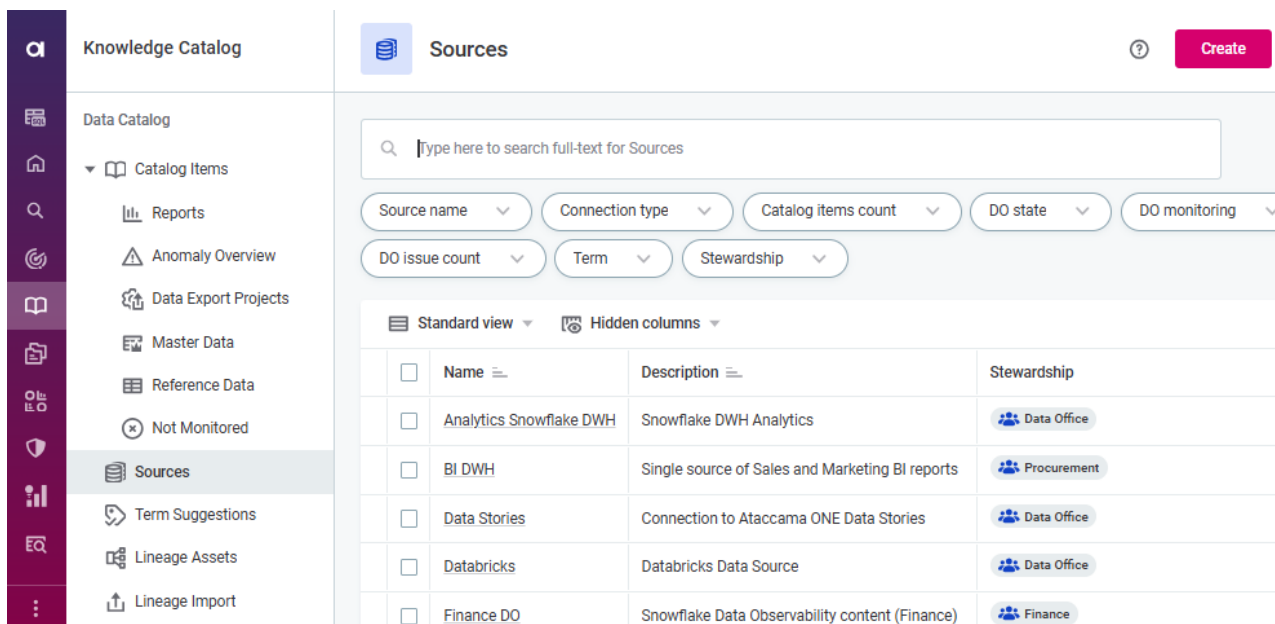
In this workshop, we will learn how to create catalog items from our data sets using two different methods. Additionally, we will explore the various types of information that become available after data profiling.

## Tasks

### 1. Creating a Data Source

As the first step to create catalog items, we need to add the data source whose datasets we want to use in the ONE. For this training, we will add a PostgreSQL database specifically prepared for this purpose.

- › Click on the **Knowledge Catalog** in the purple panel on the left side.
- › Click on **Sources** under the **Data Catalog** white panel.
- › Click on the **Create** button in the upper right corner.
- › Provide a name for your source, (e.g., "**prefix\_training**") and a description if needed:



The screenshot shows the 'Knowledge Catalog' interface. On the left is a purple sidebar with navigation icons. The main area is titled 'Sources' and features a search bar, several filter buttons (Source name, Connection type, Catalog items count, DO state, DO monitoring, DO issue count, Term, Stewardship), and a table of existing sources. The table has columns for Name, Description, and Stewardship. A 'Create' button is in the top right corner.

	Name	Description	Stewardship
<input type="checkbox"/>	Analytics Snowflake DWH	Snowflake DWH Analytics	Data Office
<input type="checkbox"/>	BI DWH	Single source of Sales and Marketing BI reports	Procurement
<input type="checkbox"/>	Data Stories	Connection to Ataccama ONE Data Stories	Data Office
<input type="checkbox"/>	Databricks	Databricks Data Source	Data Office
<input type="checkbox"/>	Finance.DO	Snowflake Data Observability content (Finance)	Finance



NOTE

**IMPORTANT:** When creating an object in a shared training environment, it is highly recommended to prefix or suffix the object's name with your initials or a unique number to distinguish it from others. For example: **BV\_training** or **00\_training**

**Knowledge Catalog**

**Create Source** ? Save and publish

**Data Catalog**

- ▼ **Catalog Items**
  - Reports
  - Anomaly Overview
  - Data Export Project...
  - Master Data

**General information**

A business perspective of the data source. It can refer to business purpose, department, location or to a particular user group. It should capture the organization of data sources in your company.

Name

xy\_training

Description

- › Next, click the **Add Connection**; as we will be using a Postgres database for the training, select "**PostgreSQL**" as the **Connection type**.

**Sources**

**xy\_training** Share ?

**Overview** **Data Observability** **Connections** **Catalog Items** **History** **Relationships**

**No Connection Details**

Represents a physical connection to the data source e.g. a database or a filesystem from which you can populate your Data catalog. A data source can consist of more connections.

Add Connection

- › Fill in the **Name** of the connection that will appear in your connection list.
- › Fill in the connection string details; for training purposes, we will use the following (or use the ones that were provided to you by your trainer):

**Name:** *postgres*

**JDBC:** *<jdbc:postgresql://db-training.czhmef985xre.eu-central-1.rds.amazonaws.com:5432/postgres>*

**Credentials name:** *training*

**Username:** *training*

**Password:** *AtaccamaONE*

Add Connection?Save

General information

Connection type

PostgreSQL

Name \*

postgres

Description

training database

Add Connection?Save

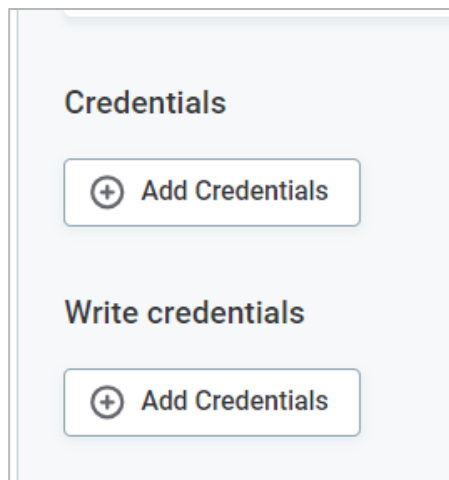
JDBC \*

jdbc:postgresql://db-training.czhmef985xre.eu-central-1.rds.amazonaws.com:5432/postgres

Spark processing

☐ Spark enabled i

› Click the **Add Credentials**; select the credential type **Username and Password** for this particular data source and enter the credential details that are provided to you in the previous page or those provided by your trainer.

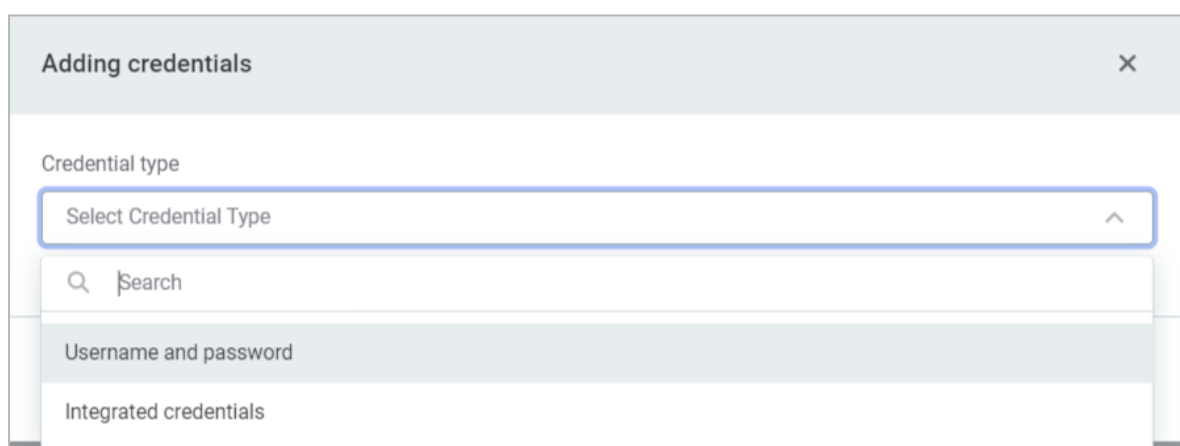


**Credentials**

+ Add Credentials

**Write credentials**

+ Add Credentials



Adding credentials

Credential type

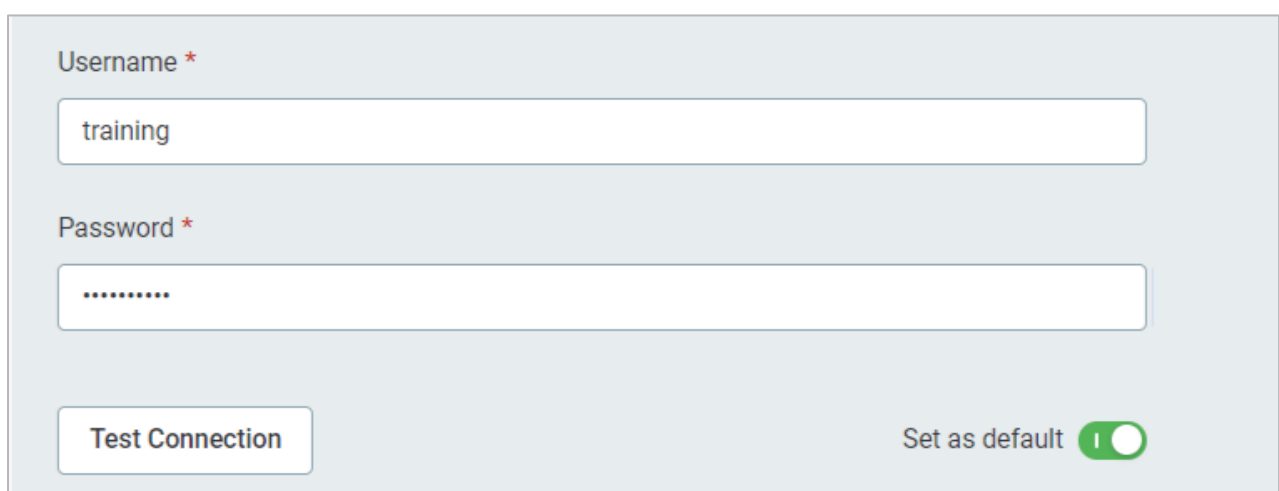
Select Credential Type

Search

Username and password

Integrated credentials

- › To verify the connection, click the **Test Connection** button; if successful, a checkmark will appear next to the **Test Connection** label.
- › Set the credentials as **default** by activating the slider element.



Username \*

training

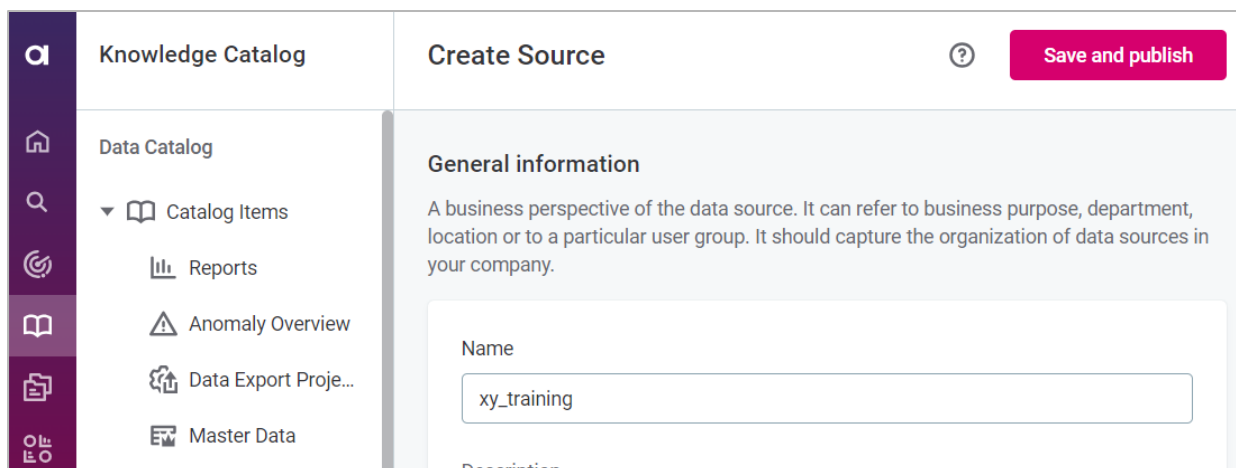
Password \*

.....

Test Connection

Set as default ☒

- › In the top right corner click the **Save and publish** button.

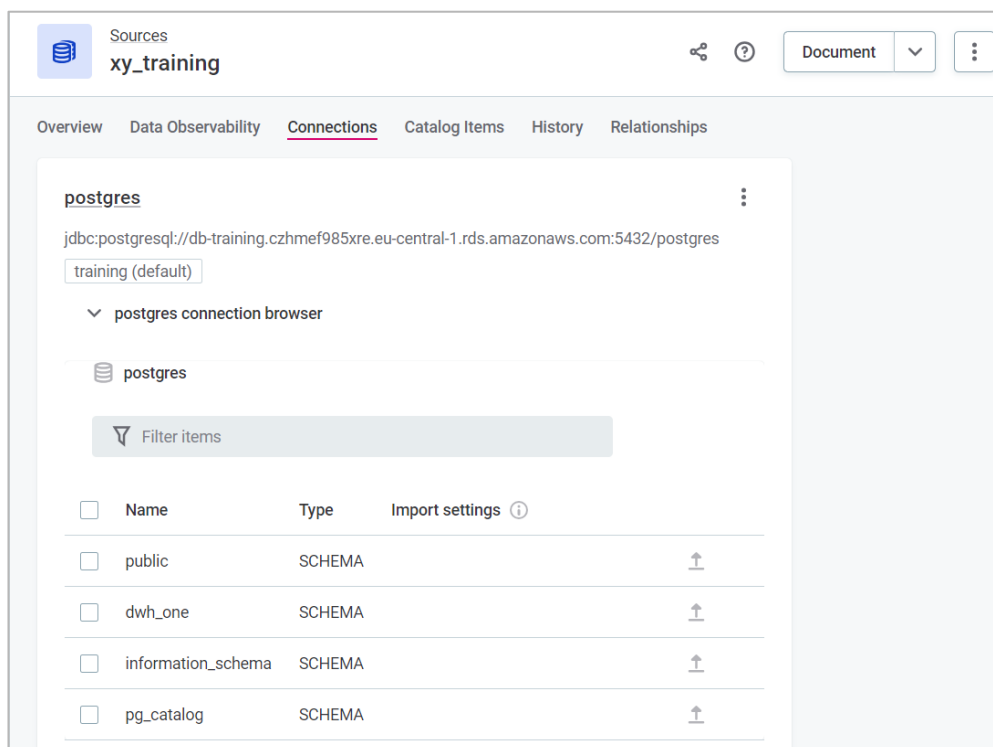


All set. Your data source is now properly identified. You can proceed to start profiling a data set.

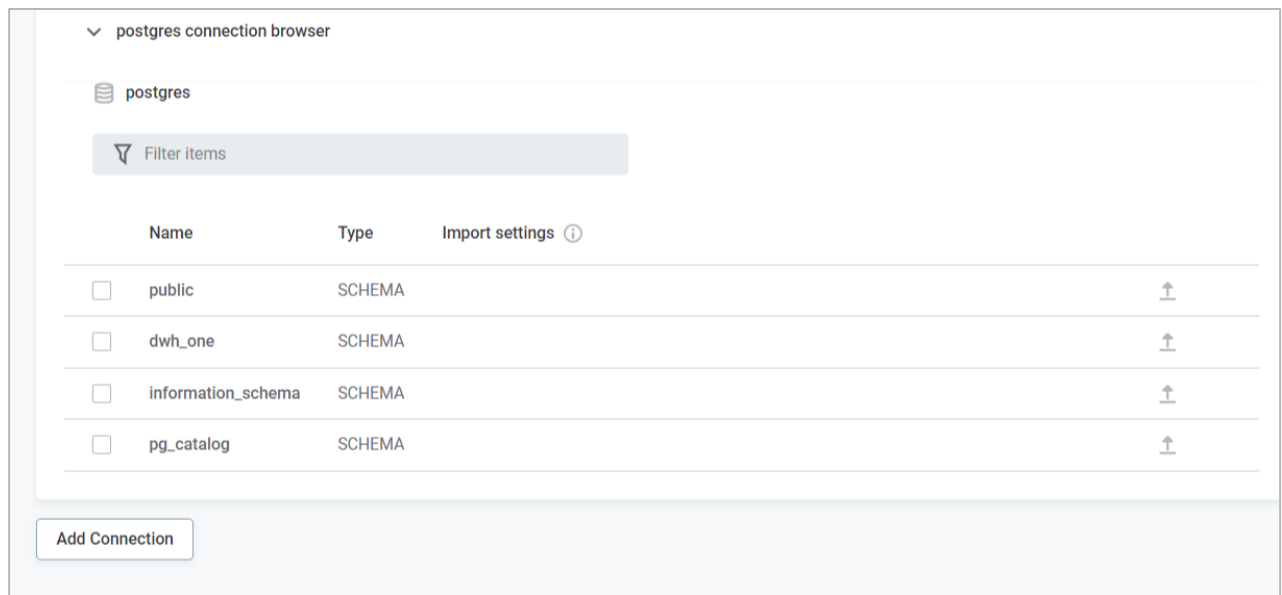
## 2. Creating Catalog Items through profiling a Data Source

In this task, you will browse your newly added data source, pick some tables, and profile them.

- › Select the new **'training'** source in the **Knowledge Catalog > Data Catalog > Sources** section.
- › Click on the **Connections** tab, then click on the **'connection\_name connection browser'** (ex: **postgres connection browser**) button of the newly added connection.



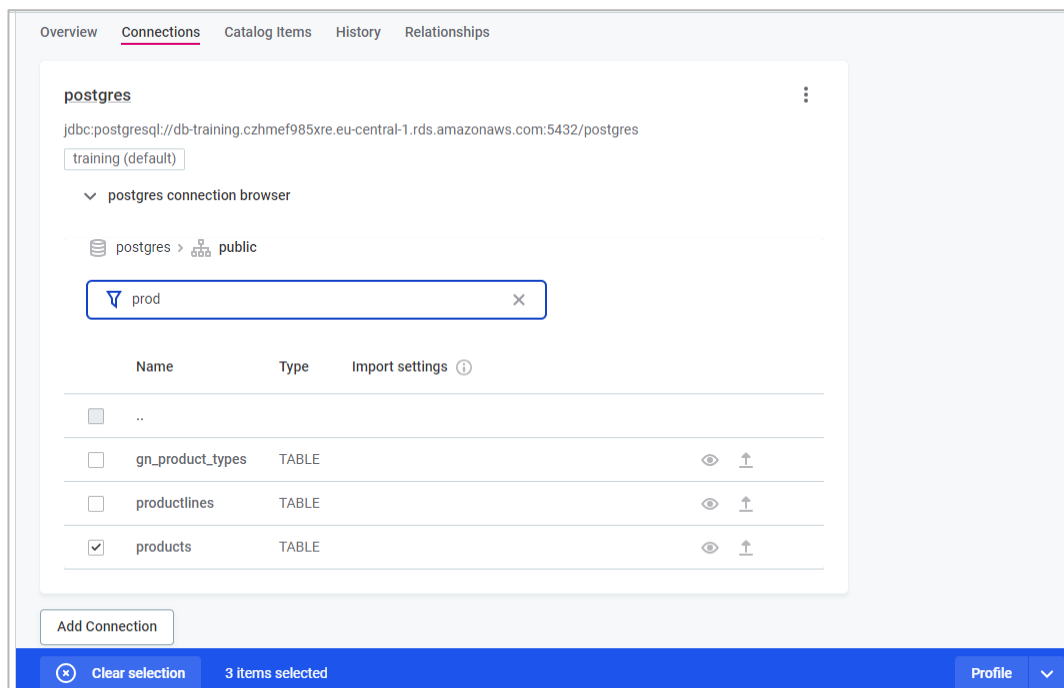
- › Next, click on the default **public** schema to open it. Be sure to click on the label and not the checkbox.



› Add the following tables in the selected schema by ticking their box selectors:

- **customers**
- **orders**
- **products**

› Click the **'Profile'** button or select the Profile option from the drop-down list to start analyzing all your data.

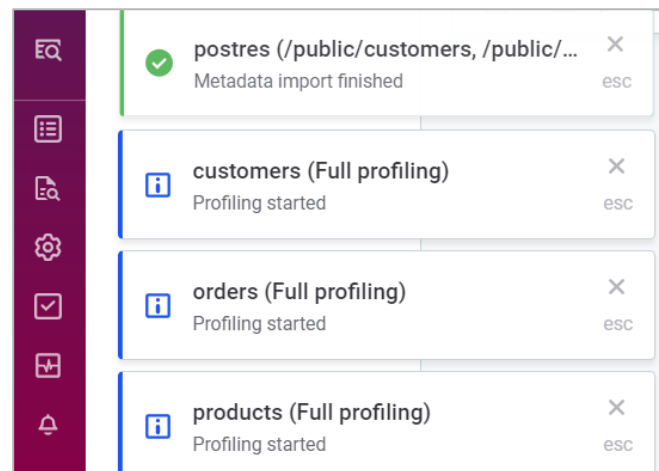


NOTE

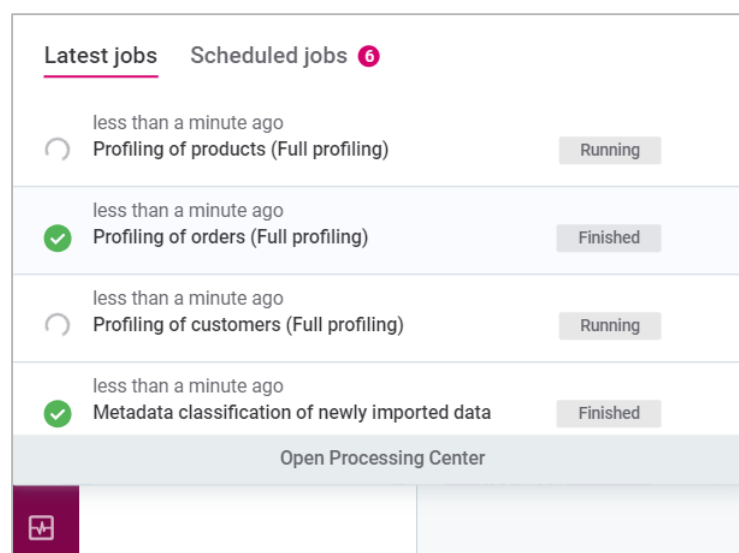
Each table from your data source that you selected is profiled and appeared as a new catalog item. Profiling results will be stored for these items as metadata. You can profile all assets in the data source by clicking the Document option in the upper right corner.



Once the profiling is started, you will receive notifications in the bottom left corner informing you about the started events:



You can also monitor the progress of the profiling jobs in the **Processing center**.



CAUTION

*By default, a newly added catalog item is automatically visible and can be modified by selected user roles, e.g. **MMM\_admin** or **MMM\_data-manager**.*

### 3. Reviewing new Catalog Items

Once the items are profiled successfully, you can find them under both the **Catalog Items** tab of the relevant source and the main repository of the Catalog Items (Navigate to **Data Catalog > Catalog Items** list and find the items you have just created).



When dealing with high numbers of catalog items, navigation and finding the right item can be tricky; you can use **filters** to find those you want; in this case, using the Data Source filter can be a good choice.



Let's apply a filter that will only show Catalog Items with less than 150 records that are from the 'training' source we created earlier.

› Review the number of profiled rows per table. You can also preview all attributes, location or the data's origin on the right-hand side panel:

› Click the name of a Catalog item (e.g. **customers**) to open the view with all tabs. Here, you can browse different types of information that Ataccama collects about the selected catalog item:

- The **Overview tab** is where you can see highlights of the Catalog item.

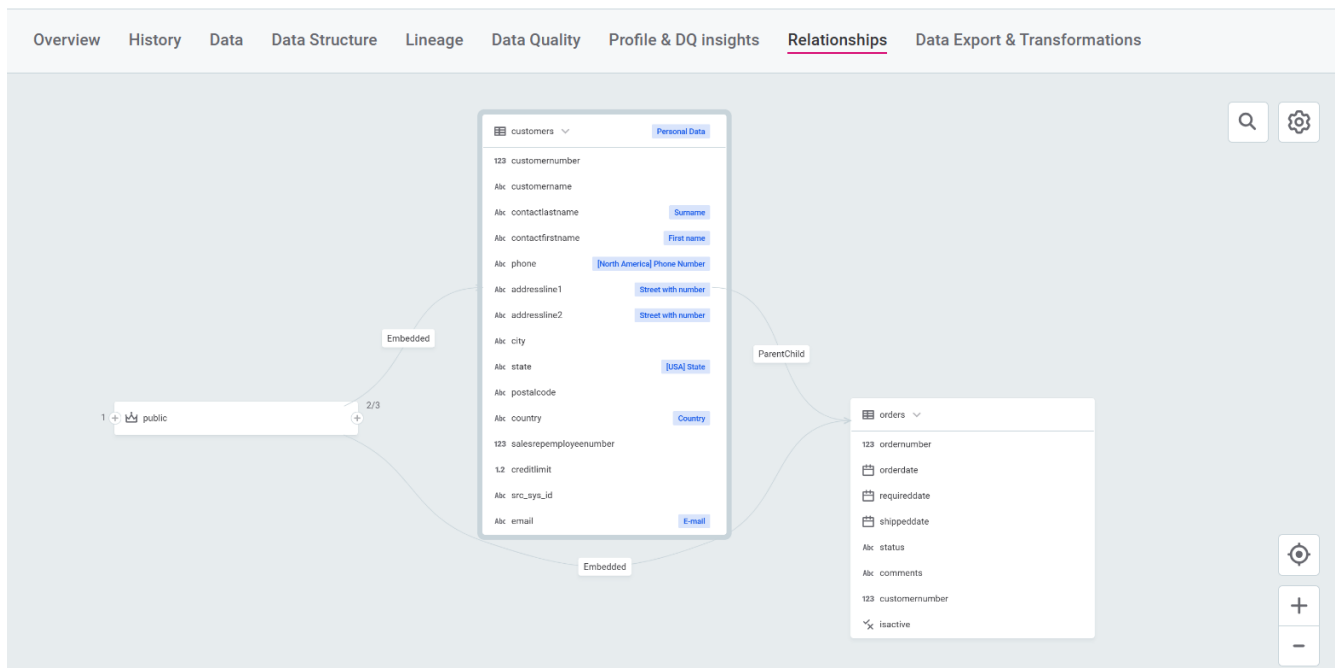
The screenshot displays the 'Knowledge Catalog' interface for a table named 'customers'. The left sidebar shows navigation options like 'Data Catalog', 'Catalog Items', 'Reports', 'Anomaly Overview', 'Data Export Projects', 'Master Data', 'Reference Data', 'Not Monitored', 'Sources', 'Term Suggestions', 'Lineage Assets', and 'Lineage Import'. The main panel has tabs for 'Overview', 'History', 'Data', 'Data Structure', 'Lineage', 'Data Quality', 'Profile & DQ insights', 'Relationships', and 'Data Export & Transformations'. The 'Profile & DQ insights' tab is active, showing a list of attributes with their respective data quality scores (e.g., 'Customer code' at 100%, 'Customer ID' at 91%, 'customername' at 99%, 'contactlastname' at 74%, 'contactfirstname' at 68%, 'phone' at 100%, 'addressline1' at 100%, 'addressline2' at 74%, 'city' at 100%, 'state' at 100%, and 'postalcode' at 100%). A 'Data Quality' section at the bottom indicates that data quality has not been evaluated. On the right, there are sections for 'Glossary terms', 'Summary' (including source, origin, table type, number of records, and data slices), and 'Stewardship'.

- On the **Profile & DQ Insights** tab, you can access profiling results and various analysis outputs.

This screenshot provides a detailed view of the 'Profile & DQ insights' for the 'customers' table. It shows a table with columns for 'Name', 'Anomalies', 'Glossary Terms', 'DQ insights', 'Top 3 values', and 'Masks'. The table lists various attributes such as 'customernumber', 'customername', 'contactlastname', 'contactfirstname', 'phone', 'addressline1', 'addressline2', and 'city'. For each attribute, it displays data quality metrics (e.g., NULL percentage, DQ insights bar chart) and top values (e.g., '103', '112', '114' for 'customernumber'). A 'Filter Attribute profiles' button and a 'Check for Rule suggestions' button are also visible at the top of the table.

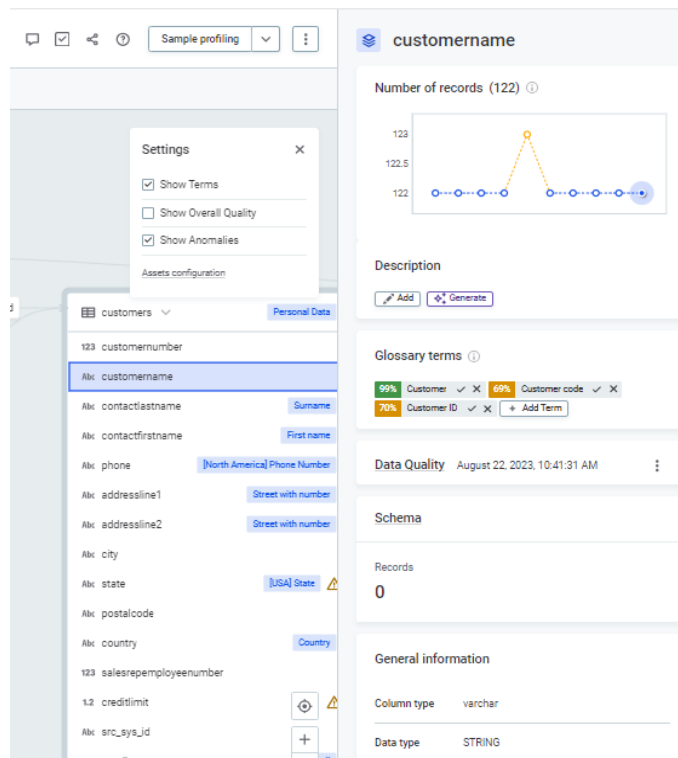
- On the **Relationships** tab, you can review the relationship of the selected item with other items as well as the data source. If set, a **Parent-Child** relationship between items from the terms perspective would be visible here as well. You can search through the relationships

graph or set which assets you want to see. Both these actions can be done using the **Search** and **Settings** buttons respectively in the upper right corner.

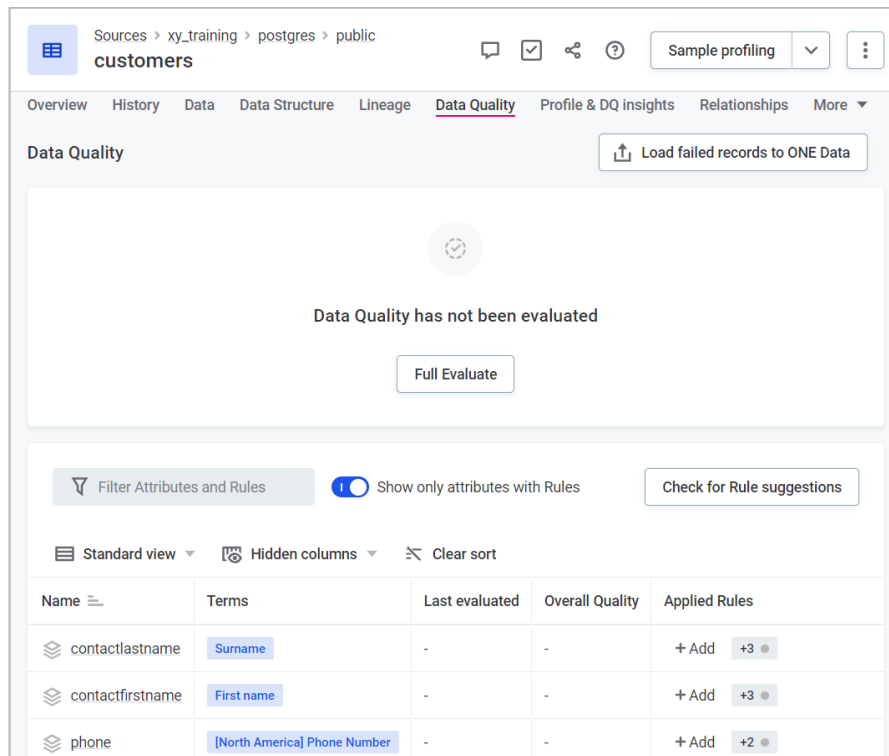


**CAUTION**

You may not see Data Quality results at the attribute level unless the selected attribute has Data Quality rules or Glossary terms associated with it. This will be explained further in the Glossary Terms section later.



- › Now, switch to the **Data Quality** tab and press the **Evaluate** button if available.



Sources > xy\_training > postgres > public  
**customers**

Overview History Data Data Structure Lineage **Data Quality** Profile & DQ insights Relationships More ▾

Data Quality Load failed records to ONE Data


Data Quality has not been evaluated

Full Evaluate

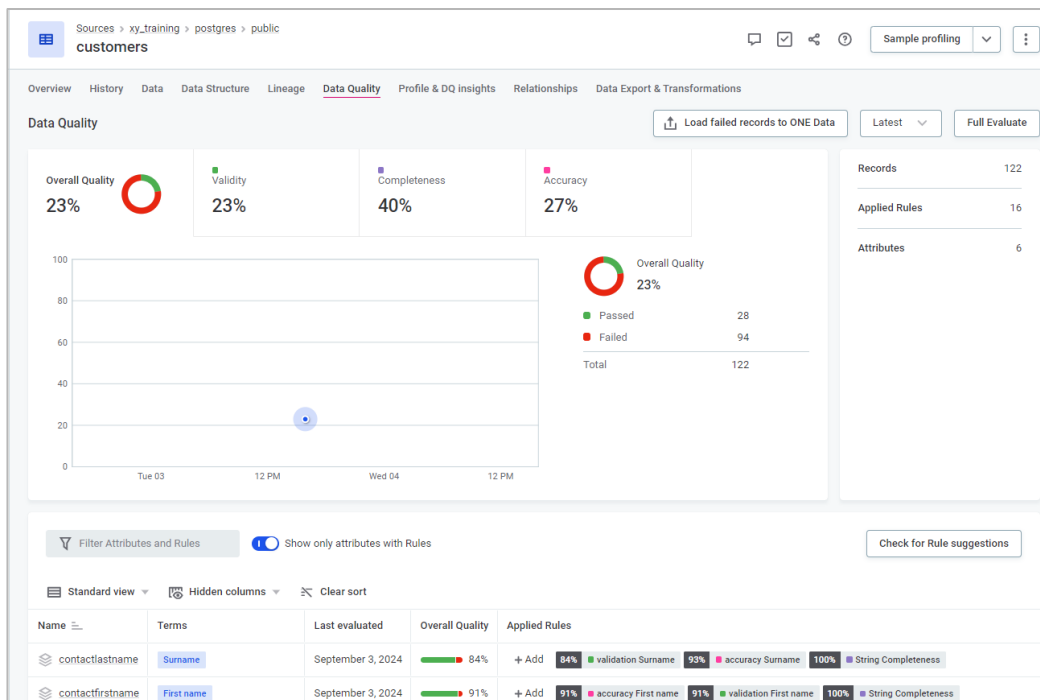
Filter Attributes and Rules Show only attributes with Rules Check for Rule suggestions

Standard view Hidden columns Clear sort

Name	Terms	Last evaluated	Overall Quality	Applied Rules
contactlastname	Surname	-	-	+ Add +3
contactfirstname	First name	-	-	+ Add +3
phone	[North America] Phone Number	-	-	+ Add +2

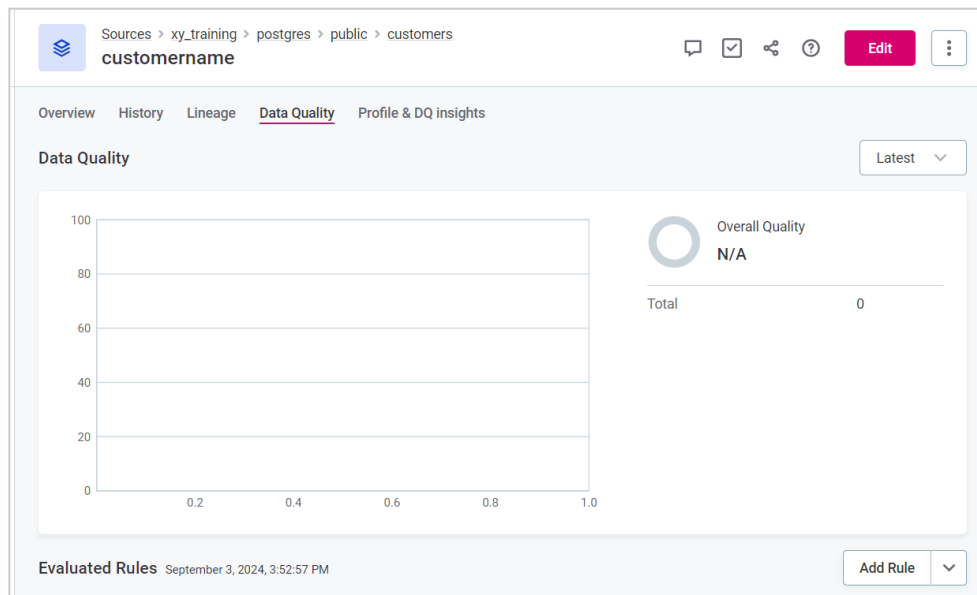
**CAUTION**  The Data Quality “Evaluate” button will appear only for the items with at least one rule on one of their attributes, directly or through applied glossary terms.

- › Refresh the application and re-check the **Data Quality** in the Data Quality tab or the overview tab of the catalog item to see the results such as **overall quality, Quality** per **Dimension, number of Applied Rules** as well as participating **attributes**.



› Go to the overview tab of the item and click on any attribute to explore the metadata at attribute level (e.g. the **contactfirstname** attribute of the **customers** catalog item.)

**CAUTION** You may not see Data Quality results at the attribute level unless the selected attribute has Data Quality rules or Glossary terms associated with it. This will be explained further in the Glossary Terms section later.



## 4. Creating a SQL Catalog Item

In the final task, a new Catalog Item will be created using a different method—by applying a SQL query directly to a data source table. For example, this query could limit the result to a specific range of values and selected attributes.

In this case, we will apply the query to the "orders" table to retrieve only the "order number" and "comments" of the canceled orders.

› While on the **Catalog Items** page, click the three dots menu on the up right hand corner and choose the **'Create SQL Catalog Item'** option.

	Name	Description	Terms	Anomalies	Overall Quality	# Attributes	# Records
	customers		Country First name Surname +5		-	15	1
	products		-		-	9	1
	orders		-		-	8	3
	KS_Orders_Cancelled		-		-	2	
	products		-		-	9	1
	customers		Country First name Surname +5		-	15	1

› Choose the source of your Catalog Item's data (**'postgres'**) and write the SQL query in the window.

You can use generative AI to assist in writing your SQL query.

Select the necessary catalog items. In the **AI prompt**, describe your use case and then enter. If the prompt is successful, the generated query can be seen under SQL query or use the below query.

### SQL query:

***select ordernumber, comments from public.orders where status='Cancelled'***

› Press the **'Run Query'** to see if your results are displayed properly.

## Create SQL Catalog Item

1. Select source
2. Transform using SQL
3. Review catalog item

AI prompt

Used Catalog Items

orders

+ catalog.catalogItemsSelections.addItems

Describe the data you need

SQL query

```
1 select ordernumber, comments from public.orders where status='Cancelled'
```

Run query

Used Catalog Items

Query preview

123	ordernumber	Abc	comments
	10167		Customer called to cancel. The warehouse was notified in time and the order didn't ship. They ha
	10179		Customer cancelled due to urgent budgeting issues. Must be cautious when dealing with them in the
	10248		Order was mistakenly placed. The warehouse noticed the lack of documentation.
	10253		Customer disputed the order and we agreed to cancel it. We must be more cautions with this custom
	10260		Customer heard complaints from their customers and called to cancel this order. Will notify the

Back to connection selection
Continue

- Press **Continue** to store the new item in a designated location. Fill in the folder name (e.g. **'SQL\_items'**):

## Create SQL Catalog Item

1. Select source
2. Transform using SQL
3. Review catalog item

Location

First, you should select a location for the catalog item, then name it.
Folder \*

Search

No folders are available. Please create a new one

Create new folder

Create new Folder

General information

Name

SQL\_Items

Description

Orders SQL Catalog Items

Stewardship

Assigning Stewardship group when creating an asset is best practice. You can transfer Stewardship to another group later.
Owner

Back to transformation

Create

Save

Cancel

- Fill in the **Name** of the new Catalog Item (e.g. **'prefix\_Orders Cancelled'**) and optionally provide values to the other fields. Press the **'Create'** to complete the task.
- Profile your SQL Catalog Item



# Create SQL Catalog Item

1. Select source
2. Transform using SQL
3. Review catalog item

## Location

First, you should select a location for the catalog item, then name it.

Folder \*

SQL\_Items

## Catalog item summary

Name \*

xy\_Orders\_Cancelled

Description

B
I
</>
:≡
:≡
H1
H2
🔗
/
↶
✓
Fix grammar
▼

Purpose

+ Add Purpose

Back to transformation

Create

› Go back to the **Data Catalog** > **Catalog Items** or your (e.g. **training Source**) > **Catalog Items** to locate your SQL catalog item (e.g. **'prefix\_Orders Cancelled'**); notice the item icon as well as its source properties.

Catalog Items

Published
Unpublished
All

Type here to search full-text for Catalog items

Terms
Data Quality
Data Source
xy\_training
Location
Number of Attributes

Number of Records
Profile Date
Anomalies
Suggested terms
Stewardship

Standard view
Hidden columns

	Name	Description	Terms	Anomalies	Overall Quality
⋮	xy_Orders_Cancelled		-		-
⋮	products		-		-
⋮	customers		Country First name Surname +5		2
⋮	orders		-		-

xy\_Orders\_Cancelled

### Attributes

123
ordernumber
⋮

Abc
comments
⋮

### Glossary terms

No terms assigned + Add Term

### Summary

Description

Add Generate

Source
xy\_training > SQL\_Items

Origin
postgres

› Hit the Full profiling button to further explore your new **prefix\_Orders Cancelled** SQL Catalog item.

Sources > xy\_training > SQL\_Items

xy\_Orders\_Cancelled

Full profiling

Profile the full dataset

Overview

History

Data

Data Structure

Lineage

Data Quality

Profile & DQ inside

More

Attributes

Add Attribute

123

ordernumber

Abc

comments

Glossary terms

+ Add Term

Data Quality

Data Quality has not been evaluated yet

To evaluate Data Quality you need to either add Rules or add Terms that are connected with Rules to attributes of your interest.

Summary

Description

Add

Generate

Source

xy\_training > SQL\_Items

Origin

postgres

Number of records

-

Data Slices

0

View

Stewardship

Edit

# Conclusion

We've successfully completed the workshop!

During the session, we created catalog items by profiling selected datasets from a data source that we added to the ONE. We also reviewed the profiling results and explored the collected information during this process. Additionally, we created a SQL catalog item by applying a specific query to our data source.