ONE - Workshop

Catalog & Profiling

Prepared for: v15.4

Prepared by: Ataccama

Dated: November 2024

Contents of the Document

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

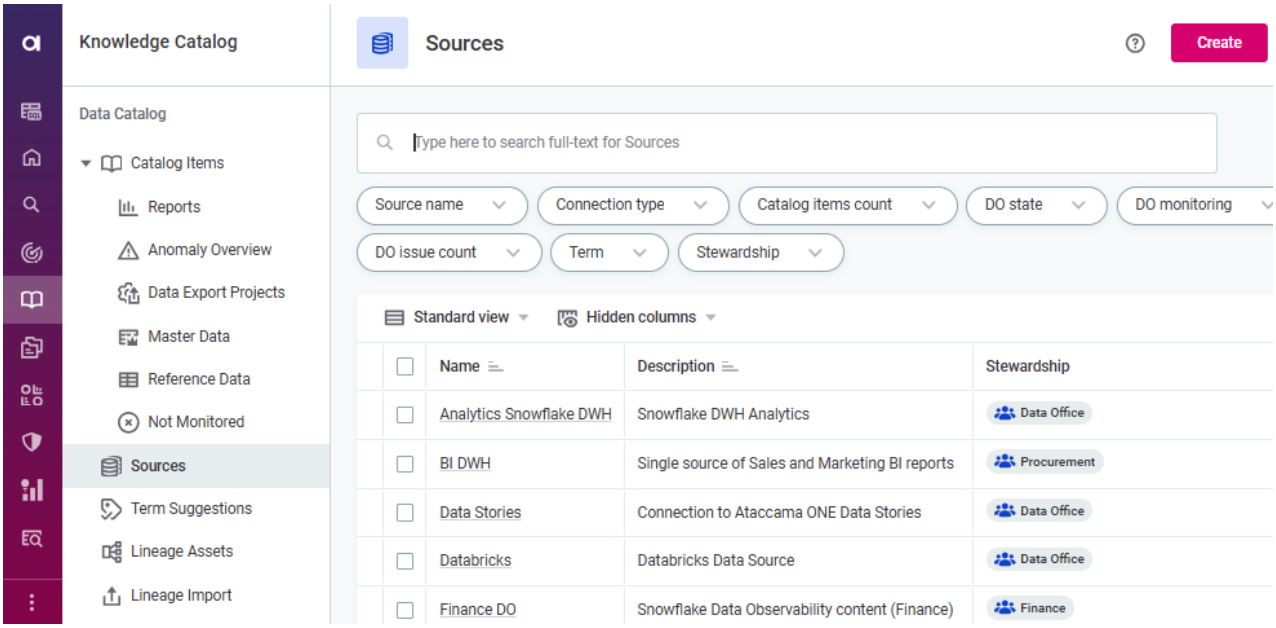
As the first step to create catalog items, we need to add the data source whose datasets we want to use in the ONE. For this training, we will add a PostgreSQL database specifically prepared for this purpose.

›Click on the ***Knowledge Catalog*** in the purple panel on the left side.

›Click on **Sources** under the ***Data Catalog*** white panel.

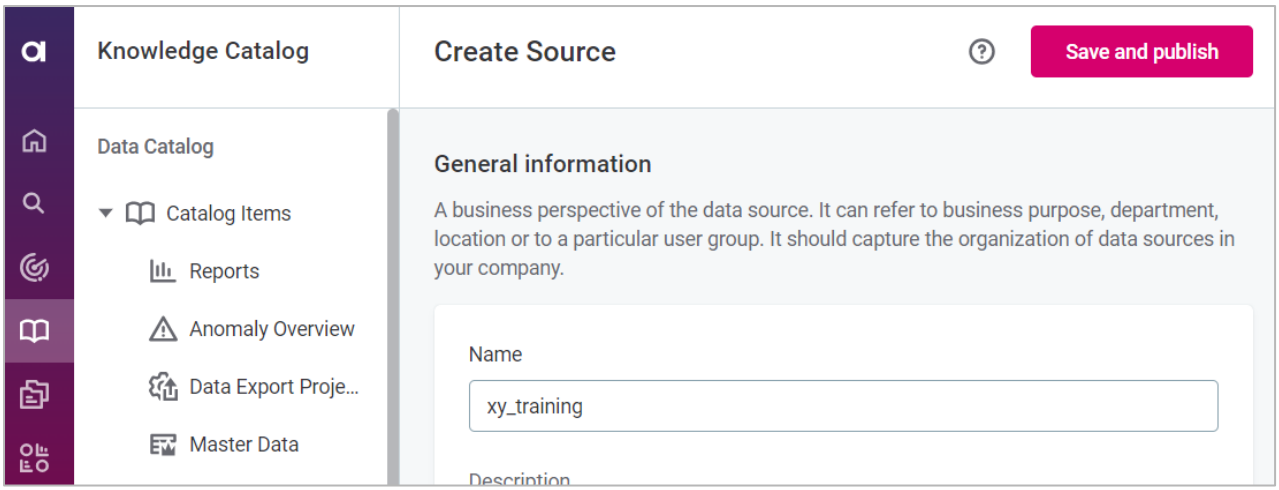
›Click on the **Create** button in the upper right corner.

›Provide a name for your source, (e.g., "***prefix_training***") and a description if needed:

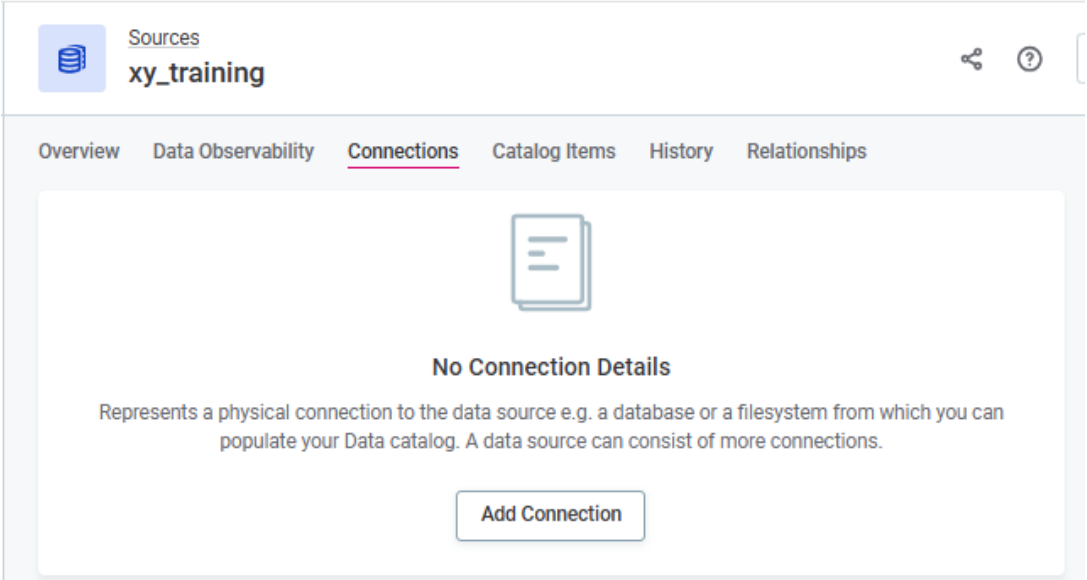


NOTE

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*



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



› 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):

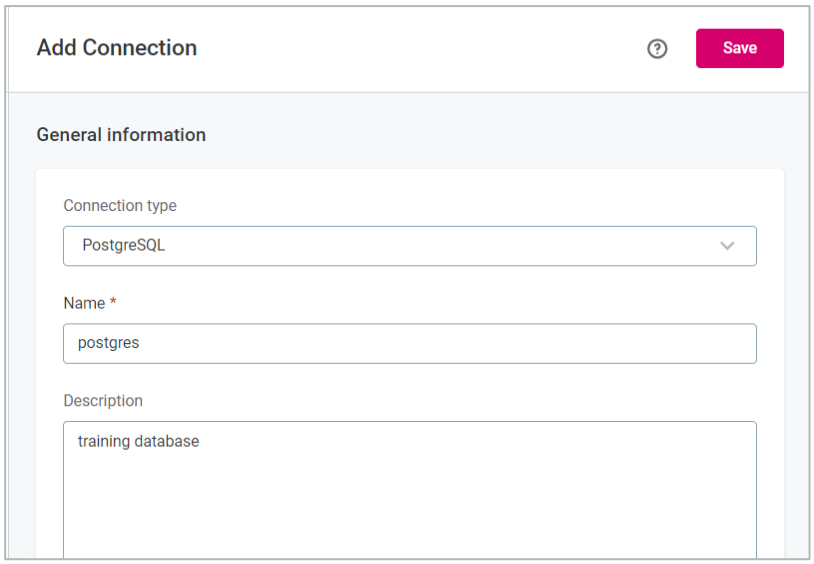
Name: *postgres*

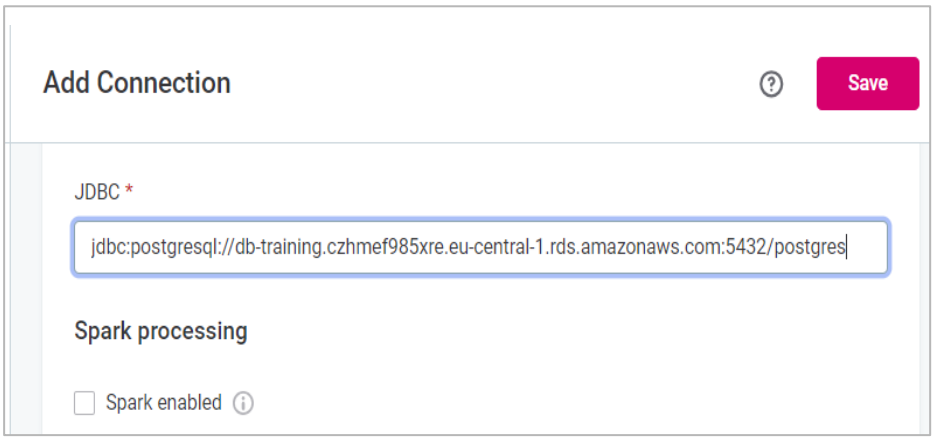
JDBC: *jdbc:postgresql://db-training.czhmef985xre.eu-central-1.rds.amazonaws.com:5432/postgres*

Credentials name: *training*

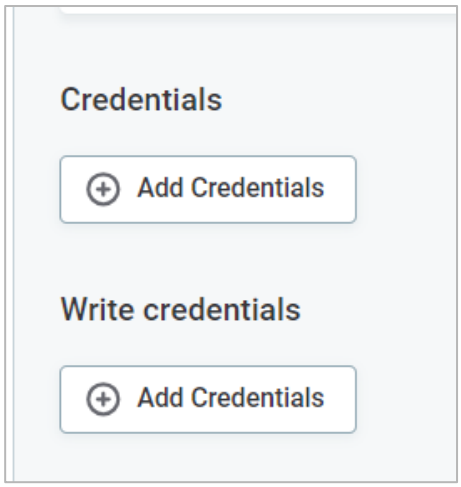
Username: *training*

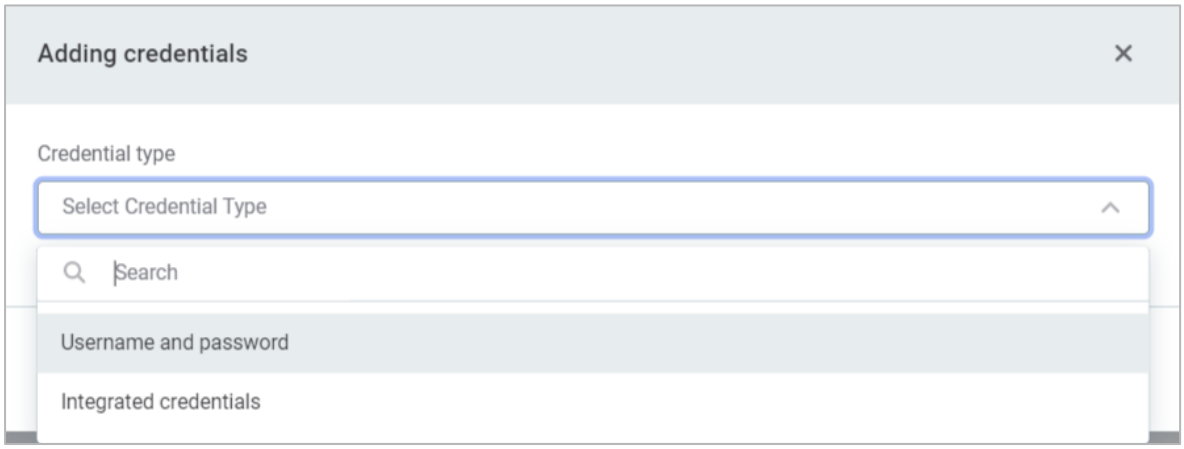
Password: *AtaccamaONE*





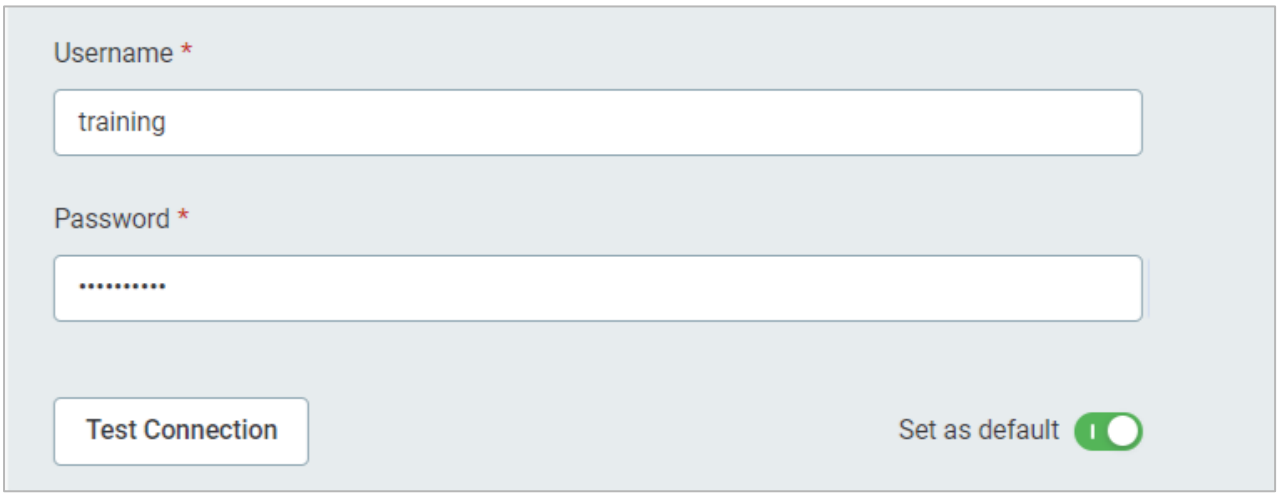
› Click the **Add Credentials**; select the credential type **Username and Password** for this particular data source and enter the credential details that are provided to you in the previous page or those provided by your trainer.



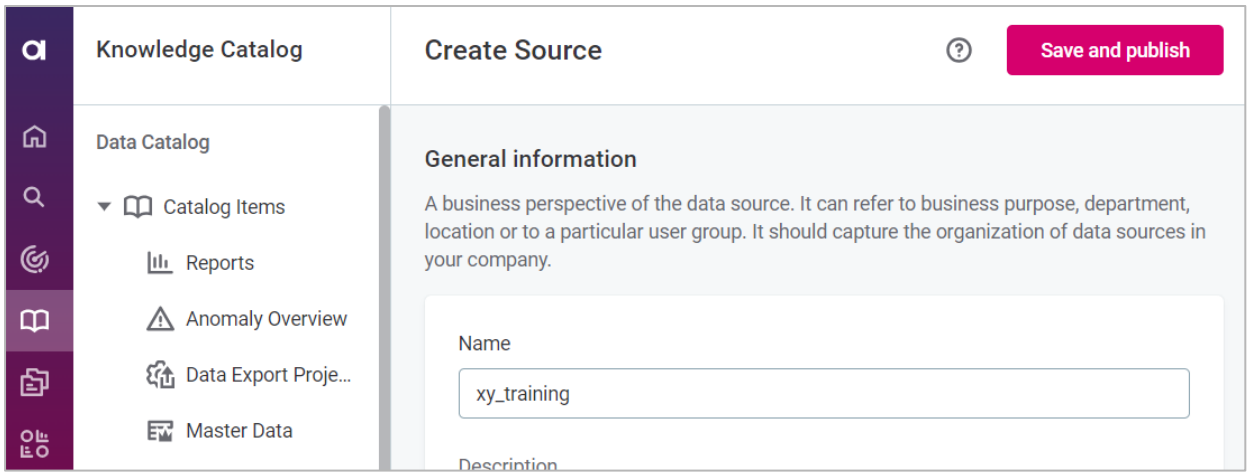


› To verify the connection, click the **Test Connection** button; if successful, a checkmark will appear next to the **Test Connection** label.

› Set the credentials as **default** by activating the slider element.



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



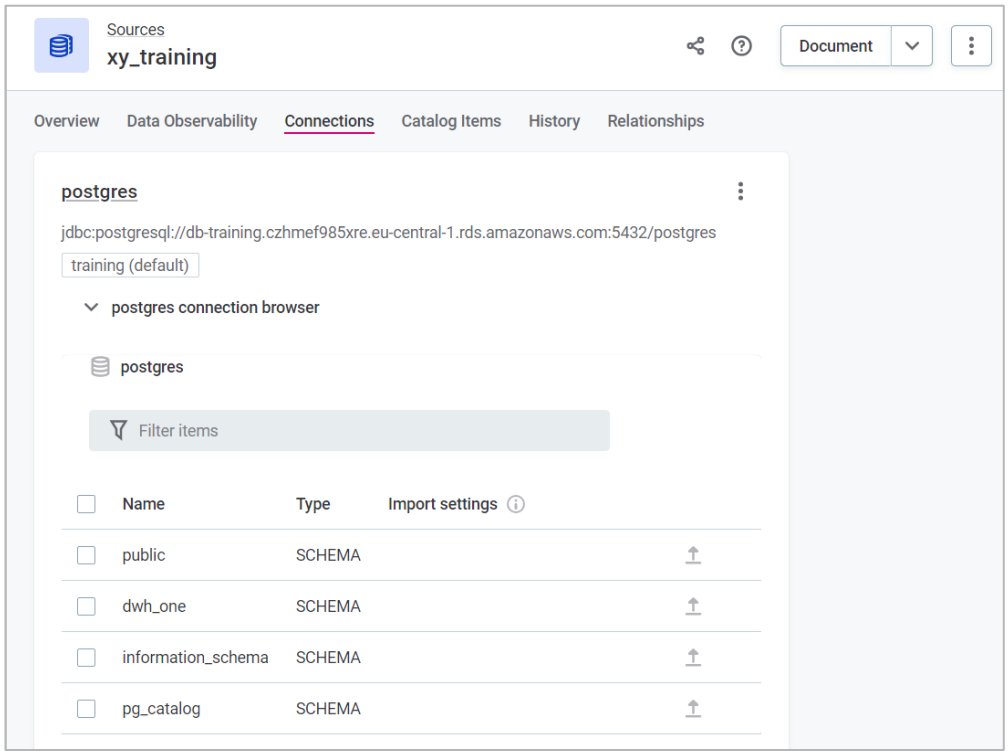
All set. Your data source is now properly identified. You can proceed to start profiling a data set.

2. Creating Catalog Items through profiling a Data Source

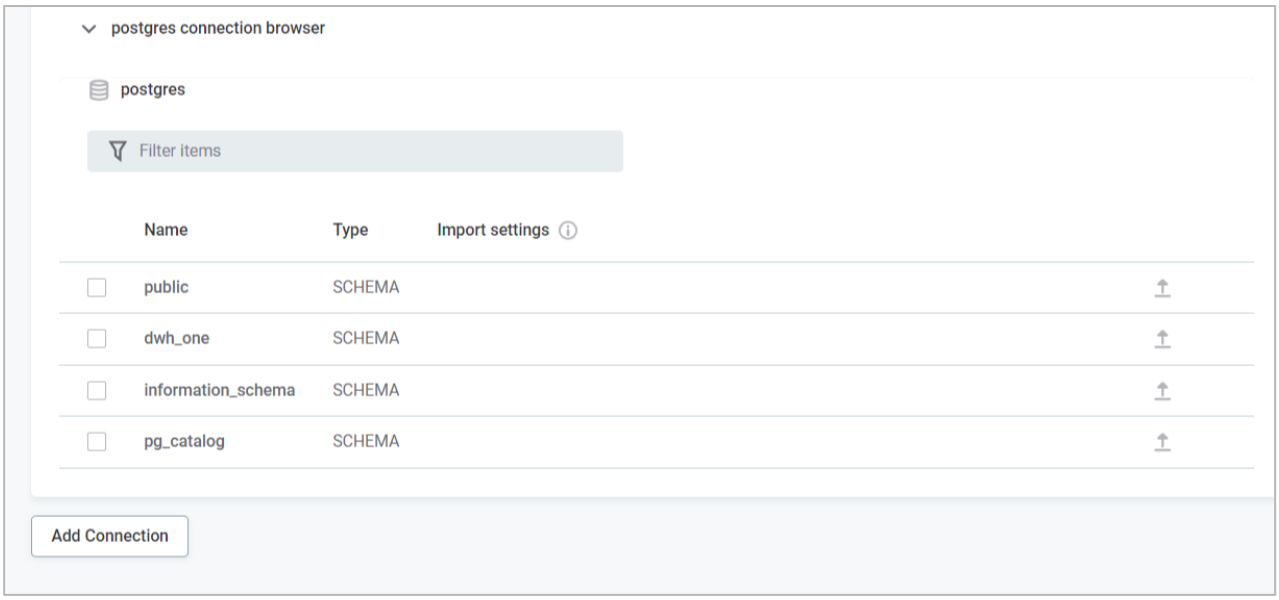
In this task, you will browse your newly added data source, pick some tables, and profile them.

› 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**.**



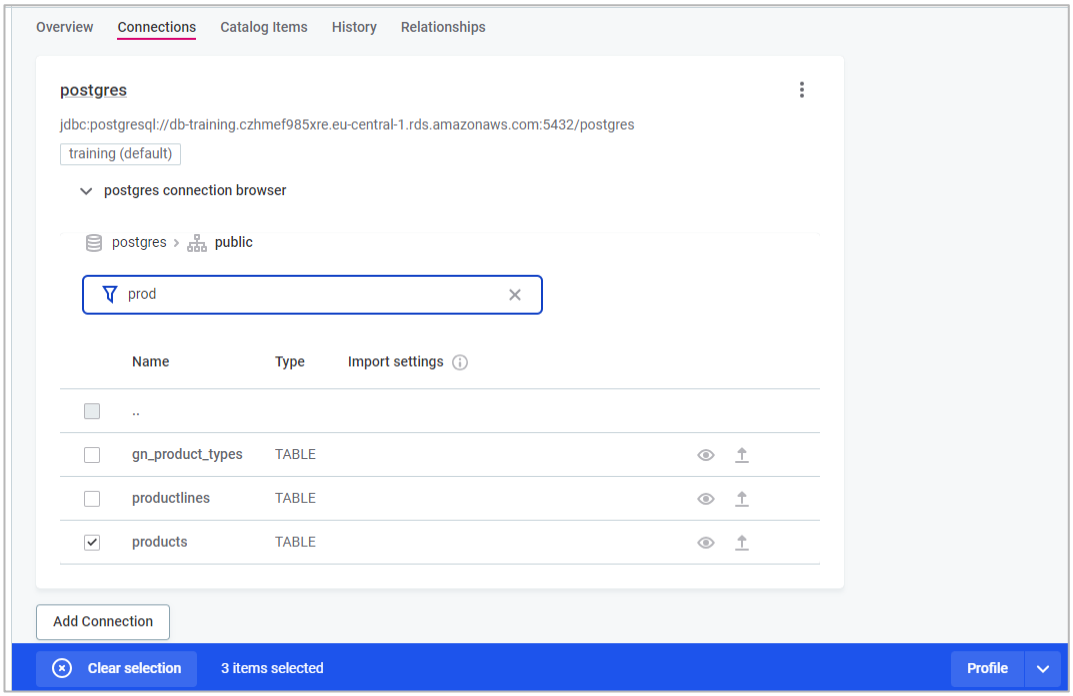
› Next, click on the default **public** schema to open it. Be sure to click on the label and not the checkbox.



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

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

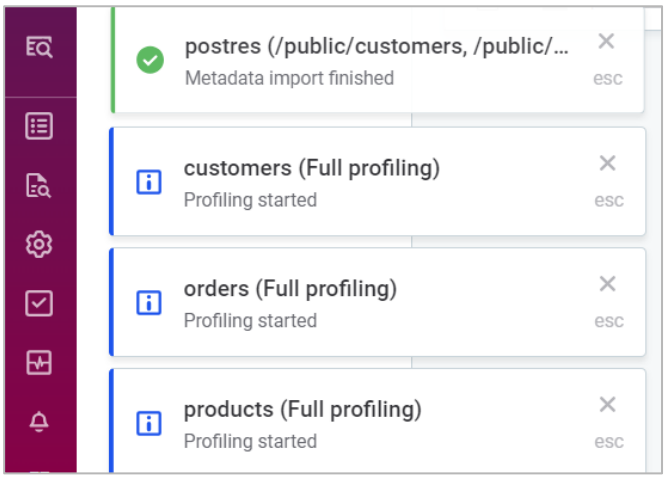
› Click the **'Profile'** button or select the Profile option from the drop-down list to start analyzing all your data.



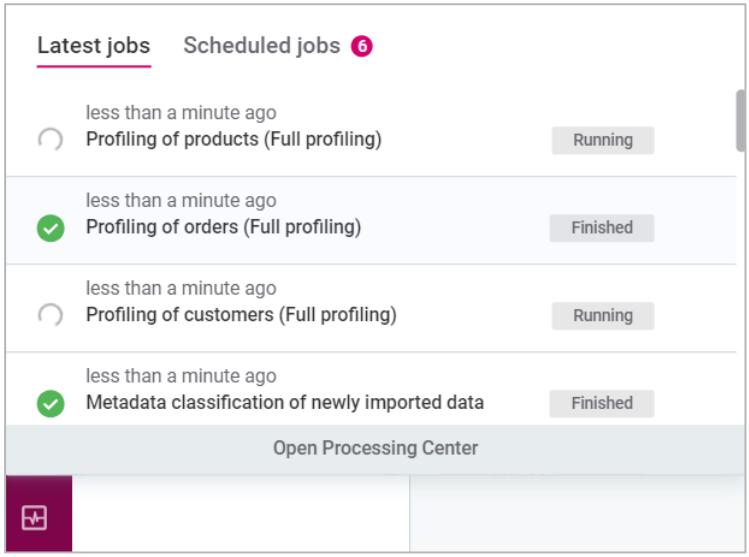
NOTE

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).

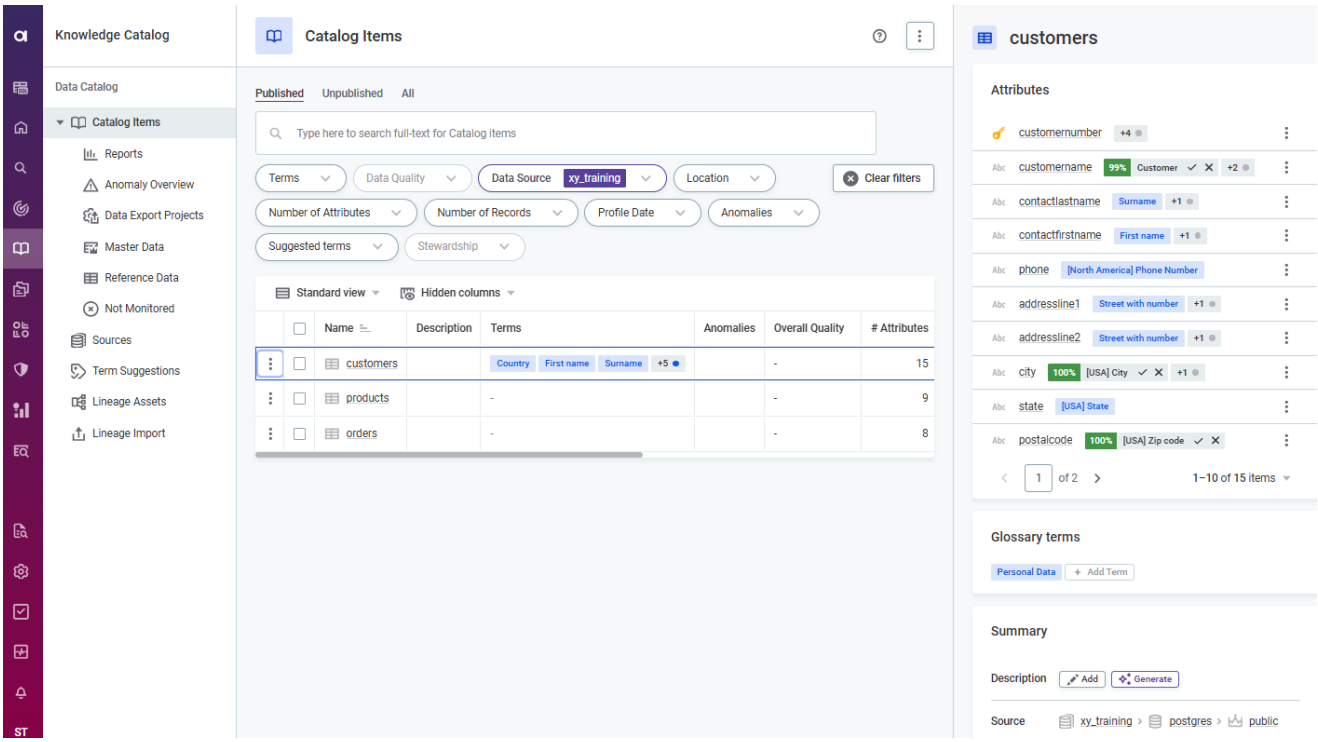
HINT

*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 aood choice.*

TASK

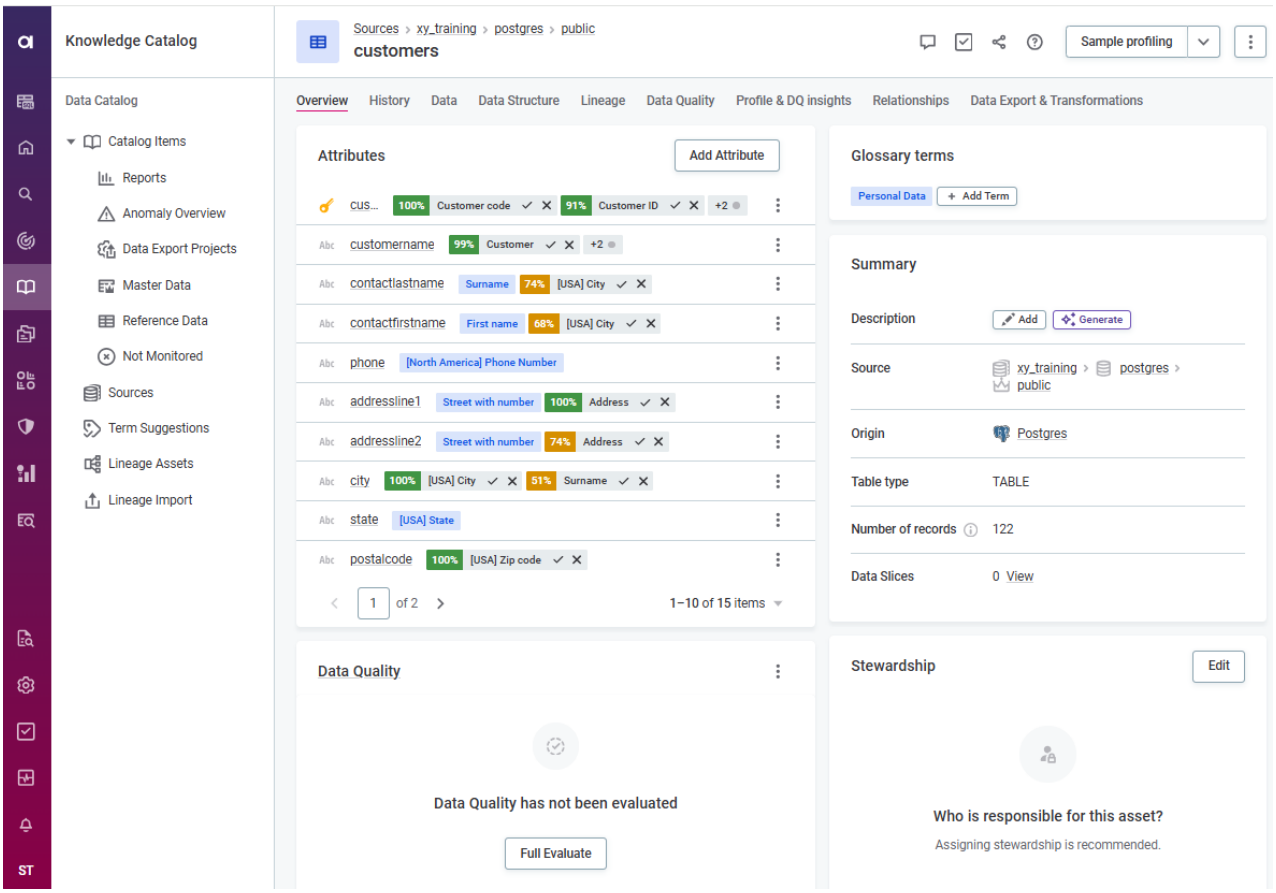
Let's apply a filter that will only show Catalog Items with less than 150 records that are from the 'training' source we created earlier.

› 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:

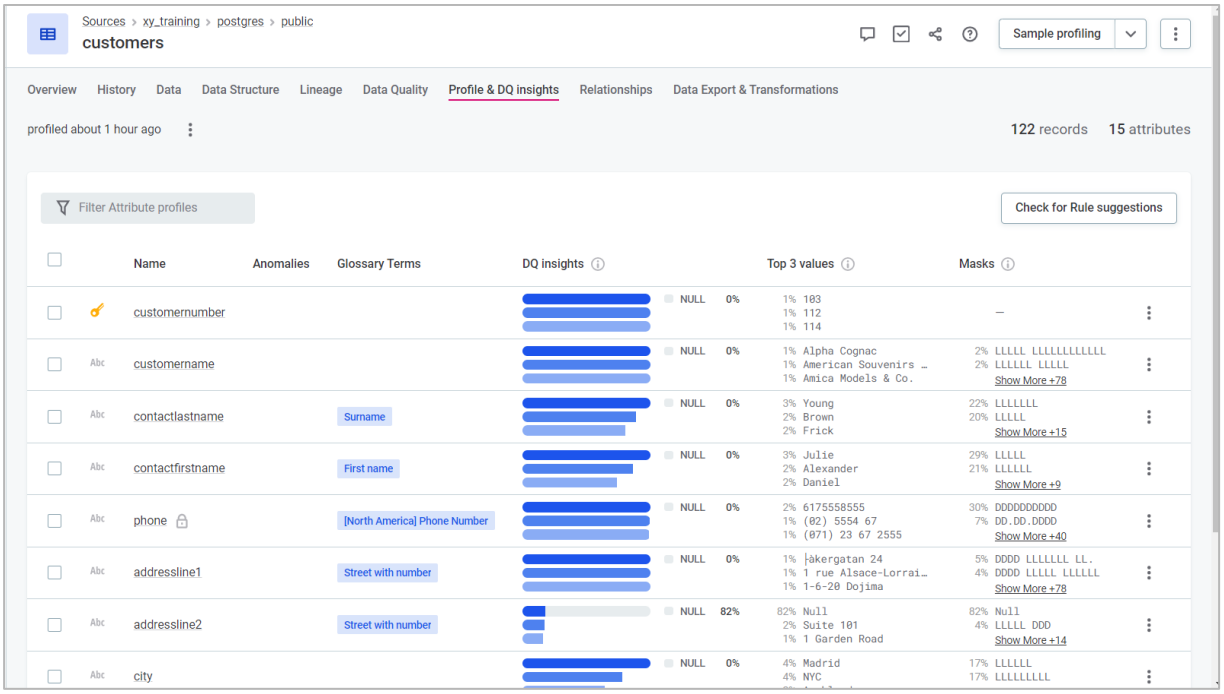


› Click the name of a Catalog item **(**e.g. **customers)** to open the view with all tabs. Here, you can browse different types of information that Ataccama collects about the selected catalog item:

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

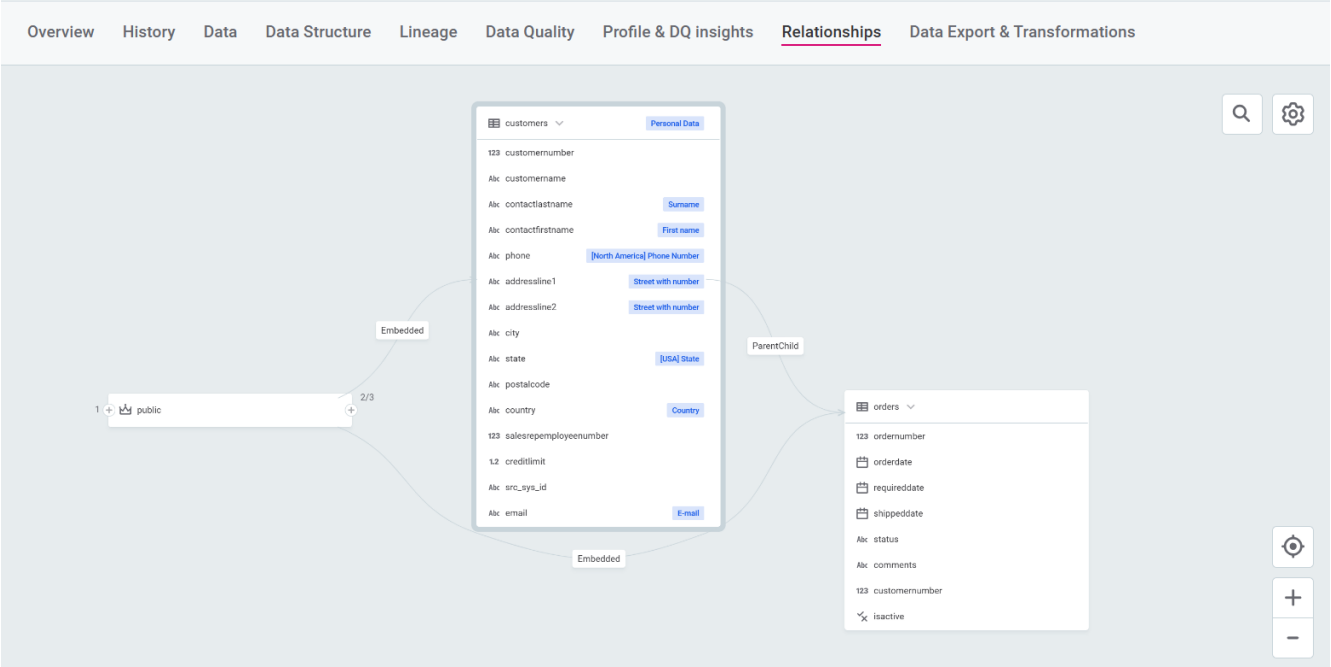


- On the **Profile & DQ Insights** tab, you can access profiling results and various analysis outputs.



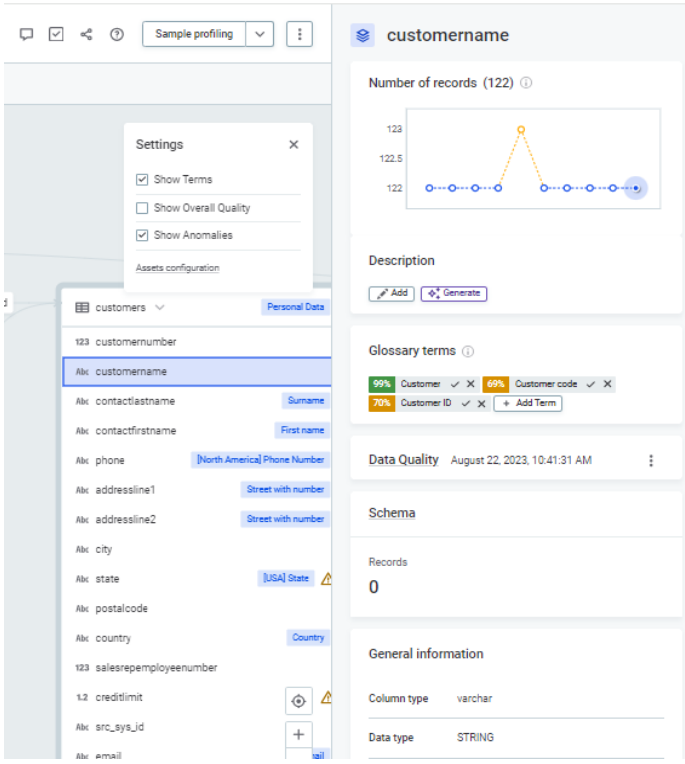
- On the **Relationships** tab, you can review the relationship of the selected item with other items as well as the data source. If set, a **Parent-Child** relationship between items from the terms perspective would be visible here as well. You can search through the relationships

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.

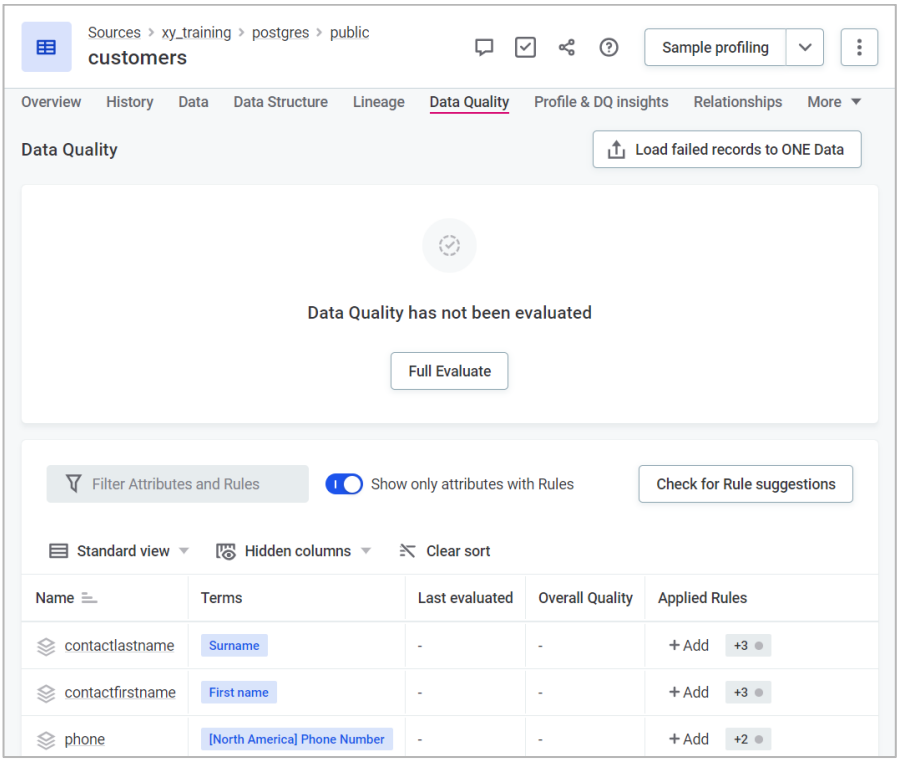


CAUTION

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.

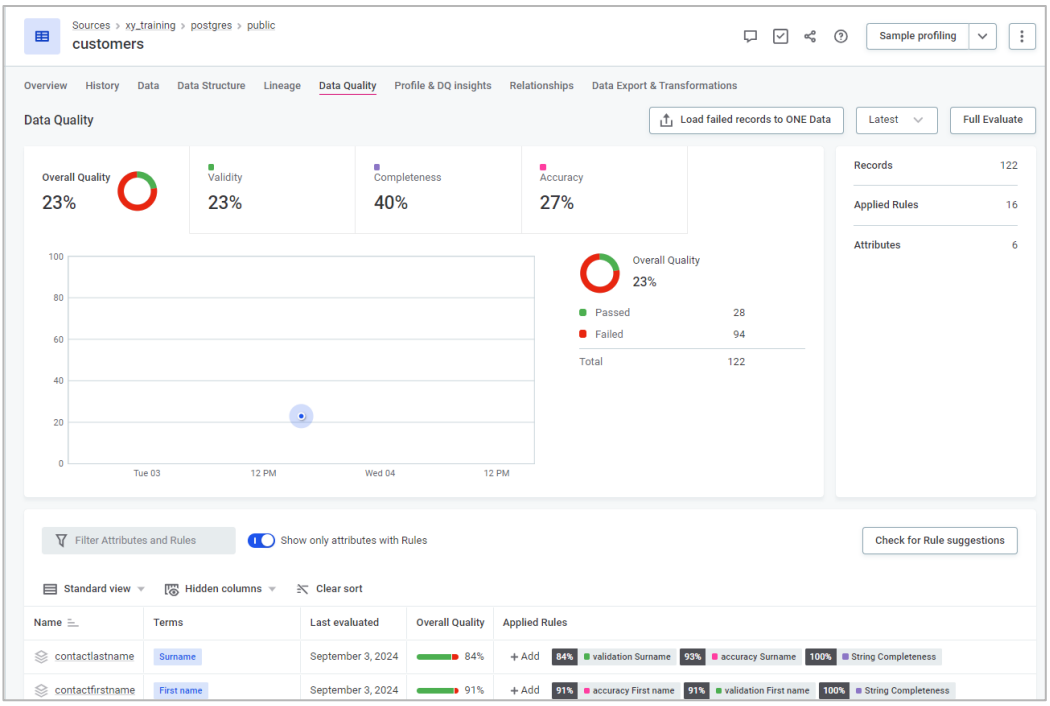


› Now, switch to the **Data Quality** tab and press the **Evaluate** button if available.



> **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.*

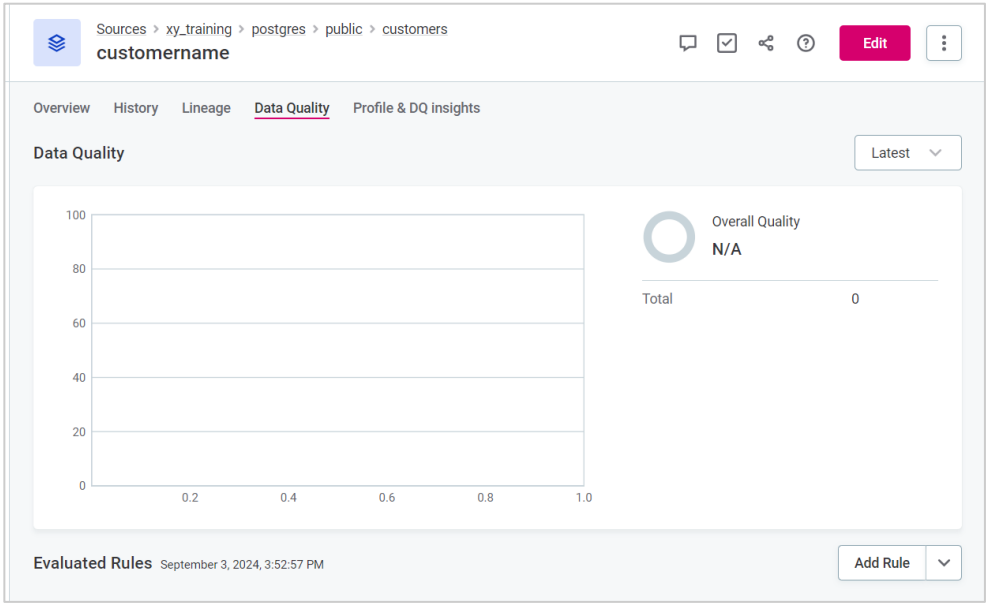
› Refresh the application and re-check the **Data Quality** in the Data Quality tab or the overview tab of the catalog item to see the results such as **overall quality, Quality** per **Dimension, number of Applied Rules** as well as participating **attributes.**



› 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.

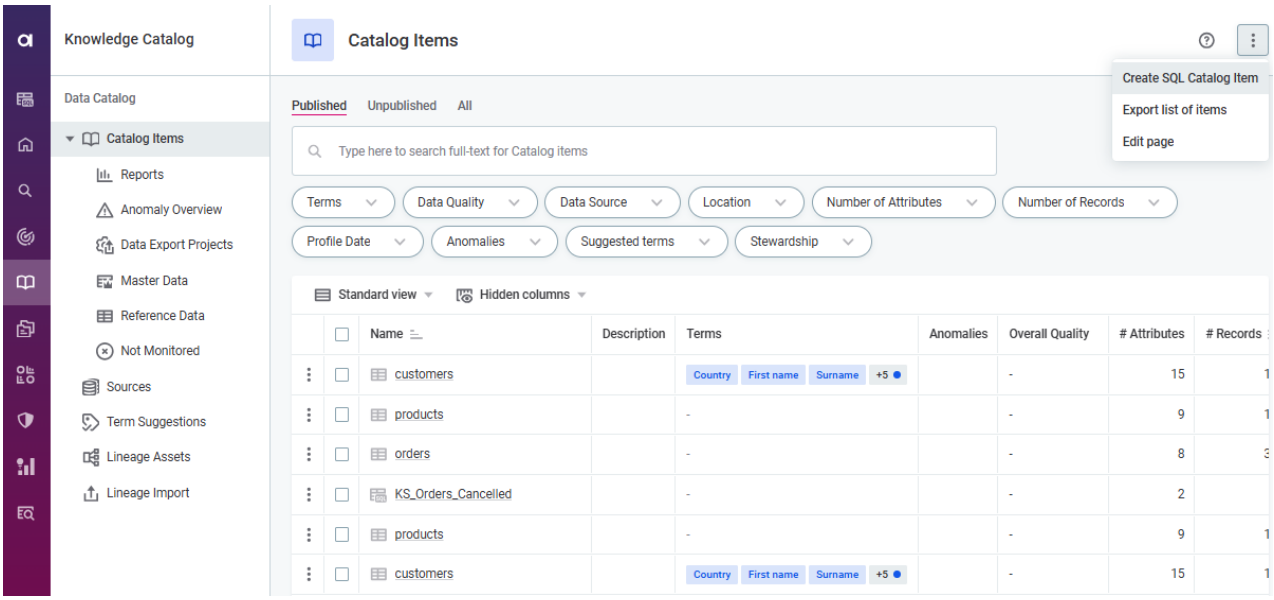


4. Creating a SQL Catalog Item

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.



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

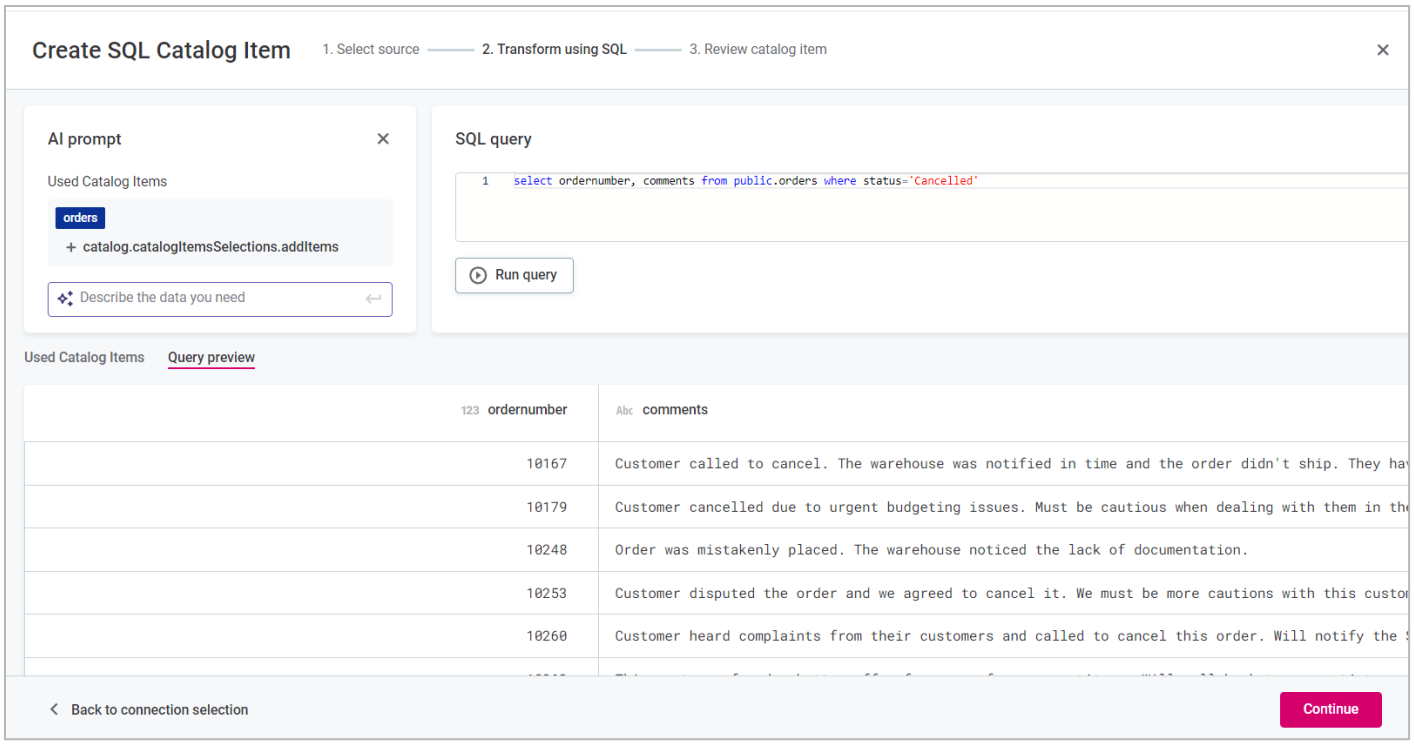
You can use generative AI to assist in writing your SQL query.

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.

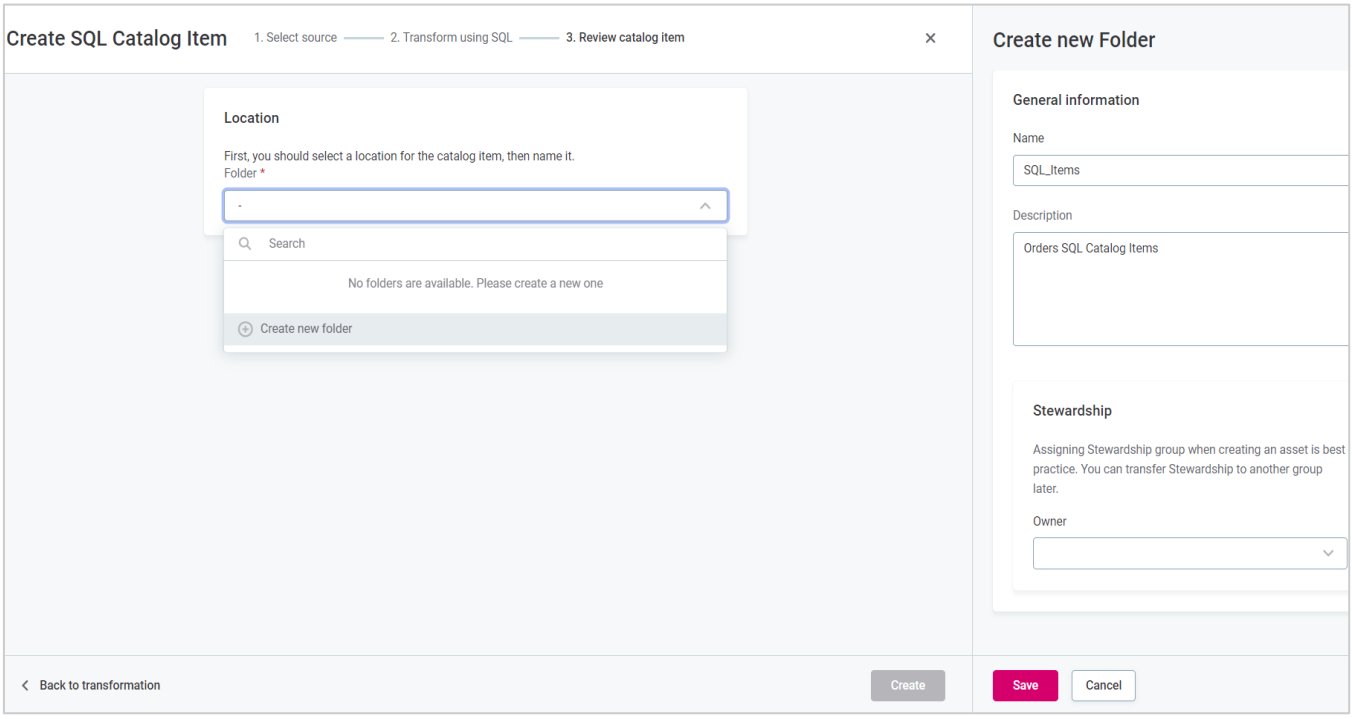
SQL query:

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

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

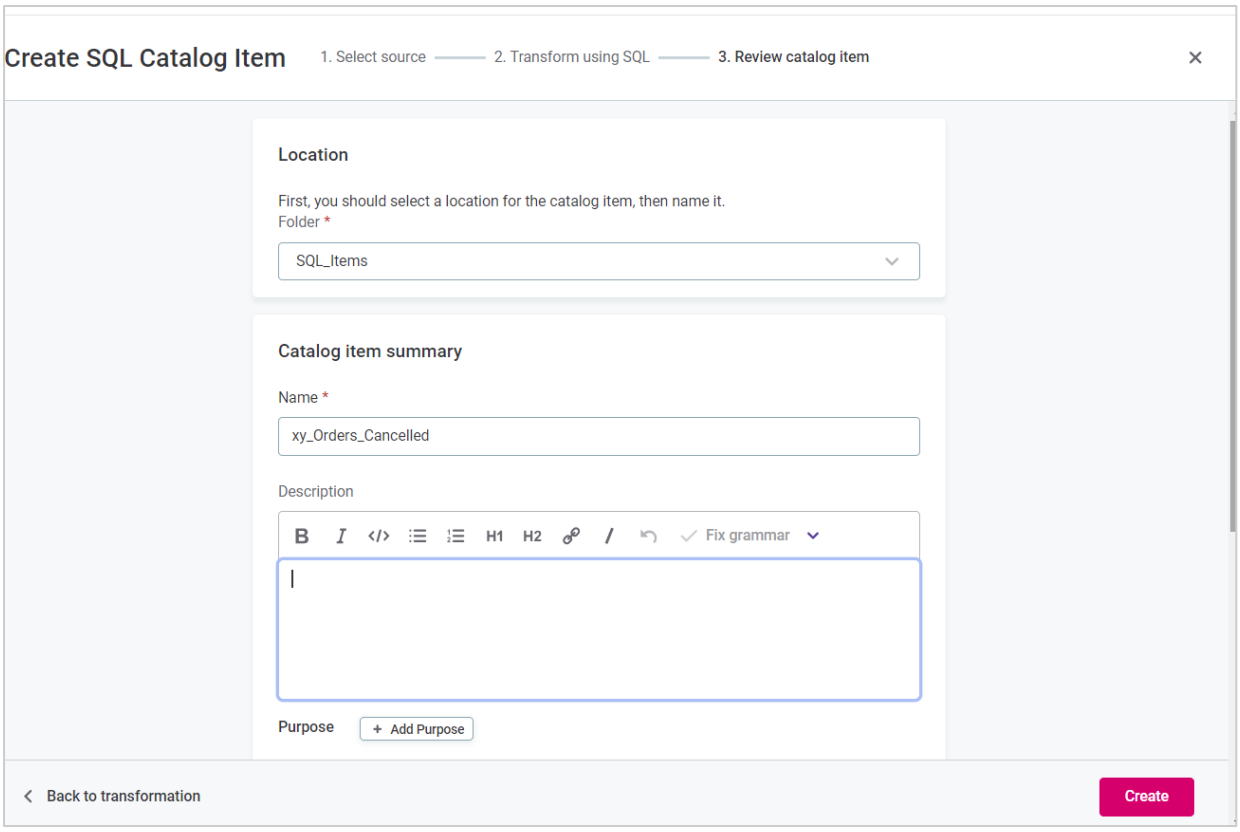


› Press **Continue** to store the new item in a designated location. Fill in the folder name (e.g. **'SQL_items'**):

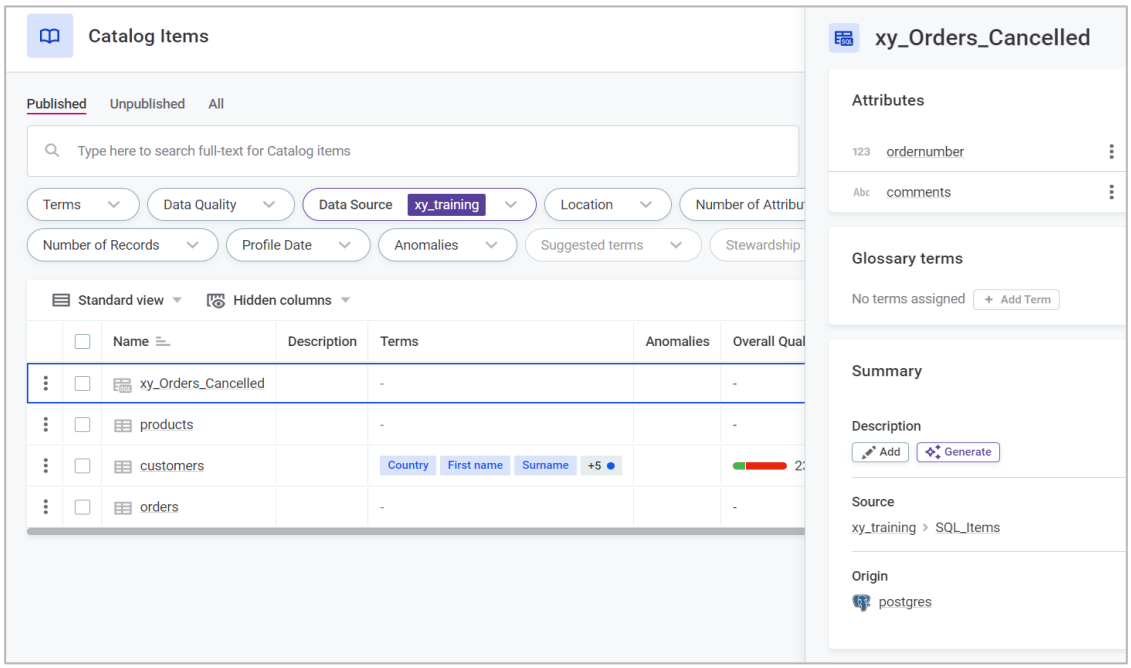


› Fill in the **Name** of the new Catalog Item *(e.g. **'prefix_Orders Cancelled'**)* and optionally provide values to the other fields. Press the **'Create'** to complete the task.

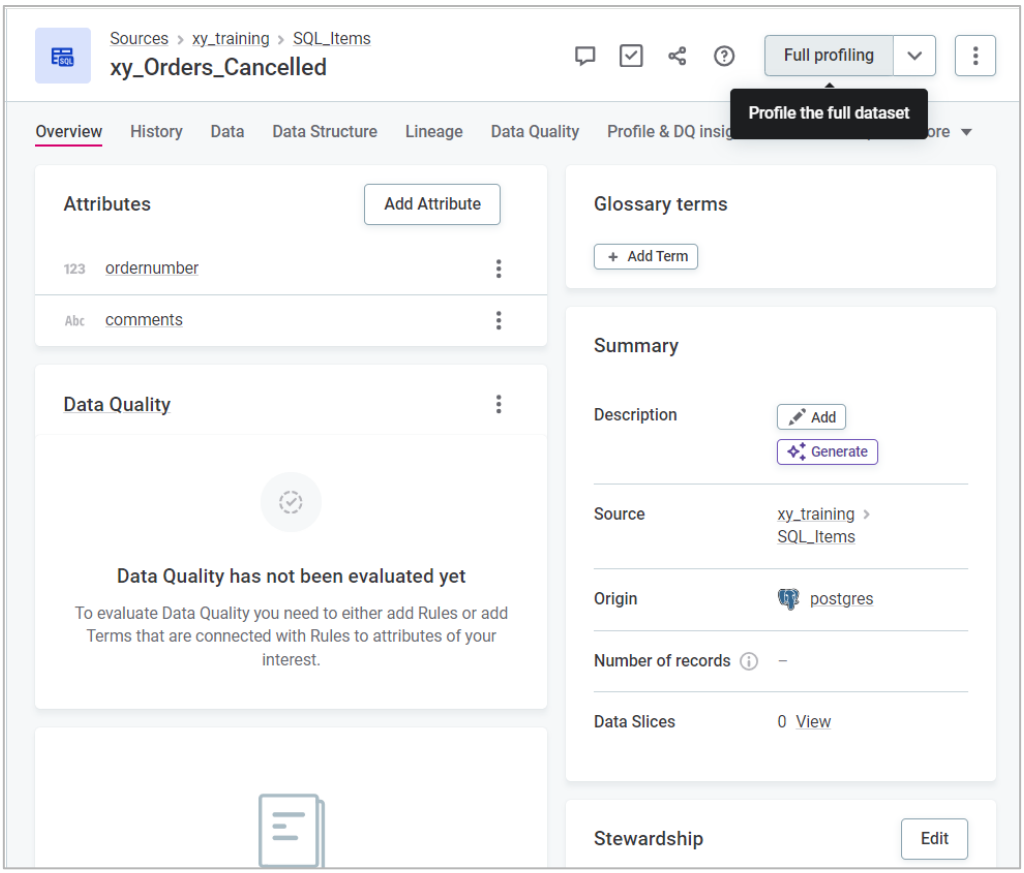
› Profile your SQL Catalog Item



› 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.



› 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
Attributes
Add Attribute
ordernumber
comments
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.
Glossary terms
+ Add Term
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.