ONE - Workshop

Data Transformation Plans

Prepared for: v15.4

Prepared by: Ataccama

Dated: November 2024

Contents of the Document

Introduction

Data Transformation Plans are visual tools for defining and executing data processing jobs. After configuring the plan to handle input data, defining the data flow, and specifying the output format, you can run the plan. Each step in the plan involves fundamental algorithms and processes data record-by-record.

There are two main use cases:

- Standalone transformation of catalog items.
- Transformation of data quality monitoring project results.

Data Transformation Plans can include steps like filtering, splitting, joining, transforming, and managing attributes. They also allow for nesting plans through the Embedded Transformation Plan step. After transformation, you can export the data to ONE Data, store it in ONE Object Storage, or use it in another plan. Features like real-time data preview and expression validation are also available.

In this example, We will create a plan that inputs the catalog item Customers, adds attributes, removes other attributes, filters out selected records, and then writes the output to a ONE Data table.

Available Steps

- Catalog Item input
- Input
- ONE Data reader
- File Output
- ONE Data writer
- Output
- Add attributes
- Condition
- Delete attributes

- Filter
- Embedded component
- Join
- Embedded transformation plan
- Split
- Transform data
- Union streams
- Monitoring project post processing input

1 – Transforming a Catalog Item with Data Transformation Plans

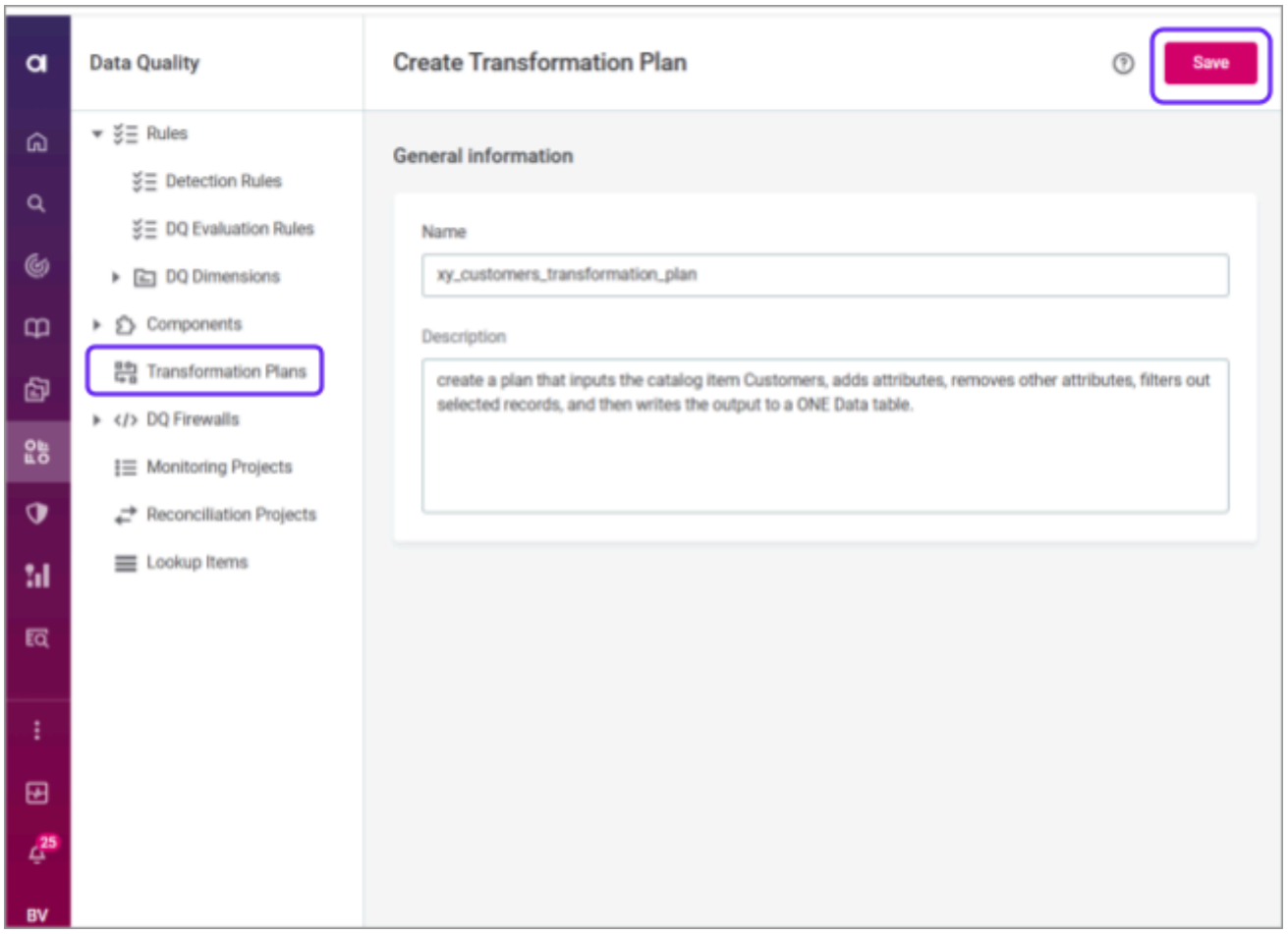
To transform data using Transformation plans:

- Select your input data.
- Build your plans using steps that can, for example, filter, transform, or join the data. (Note: These steps are similar to those in ONE Desktop, which leverage ONE Expressions. Each step represents a data processing job.)
- Draw connections between steps to control the data flow.
- Select your data output.
- Run the transformation to trigger the flow of data through the plan.

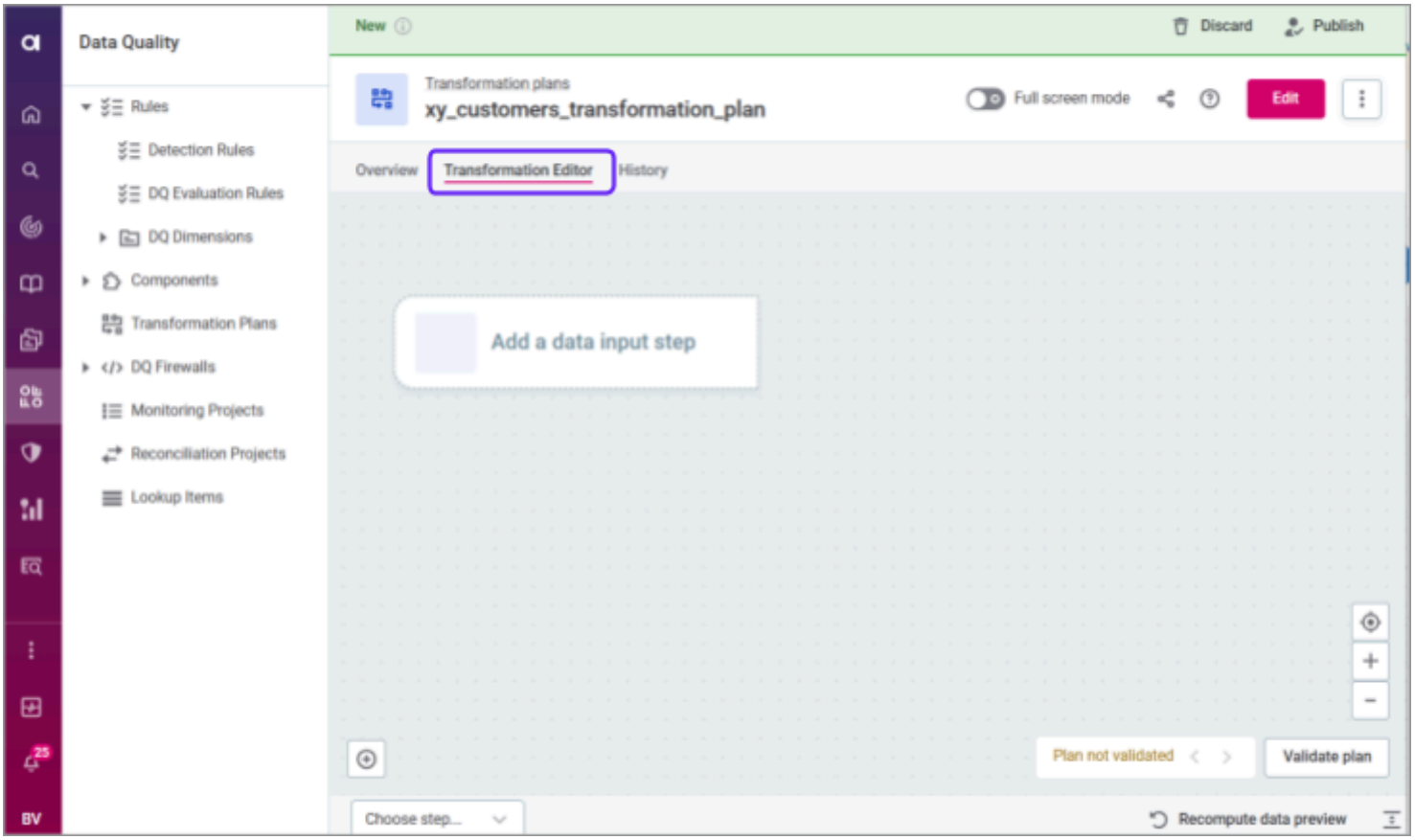
Task 1.1- Create a Transformation Plan

Let's start with creating a new transformation plan and add data input:

- Navigate to the **Data Quality > Transformation Plans** section**.**
- Create a new plan and fill in its **Name** – '**<prefix>_customers_transformation_plan'** and optionally some Description if needed.



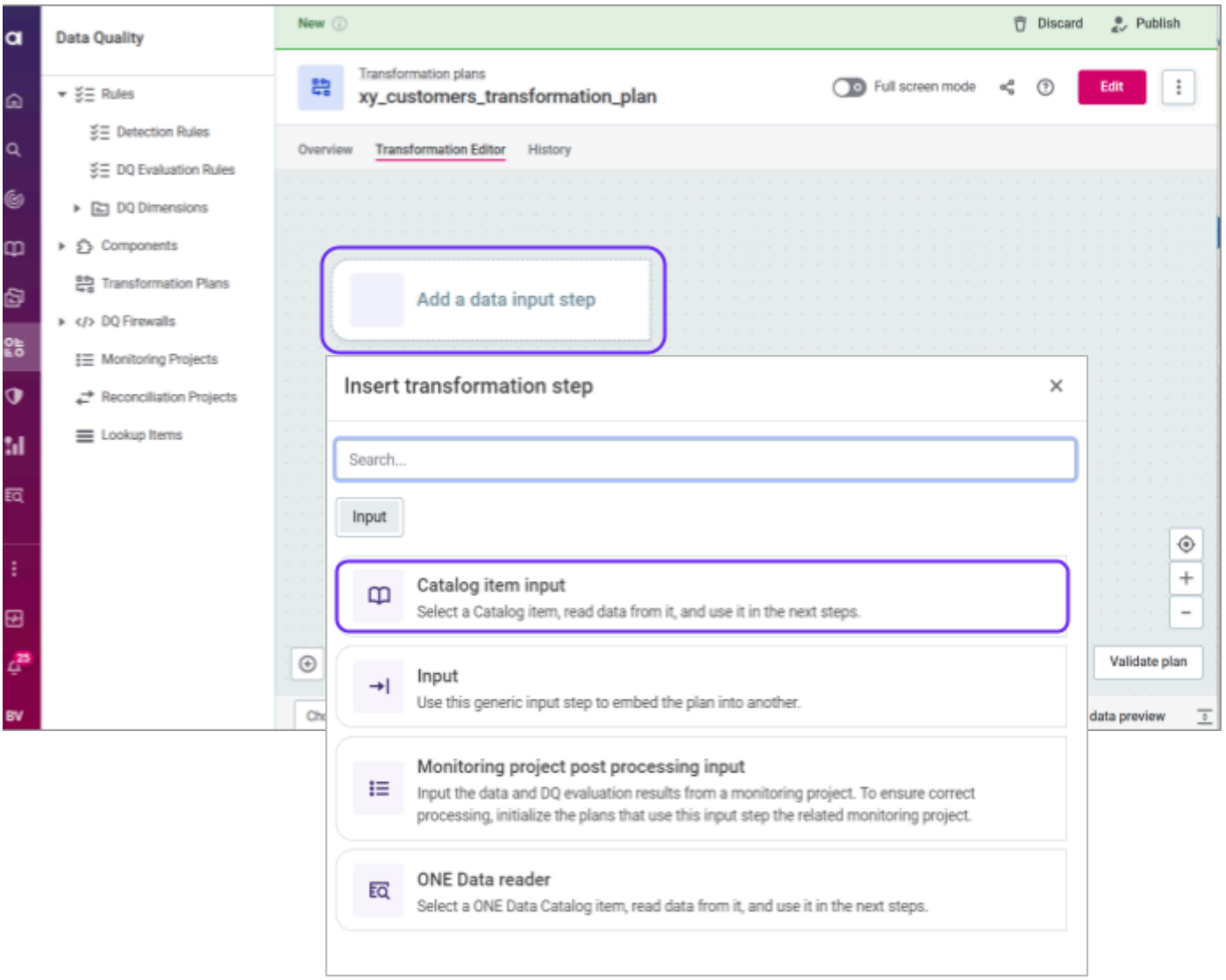
› **Save** your changes and you will be redirected to the **Transformation Editor** tab.



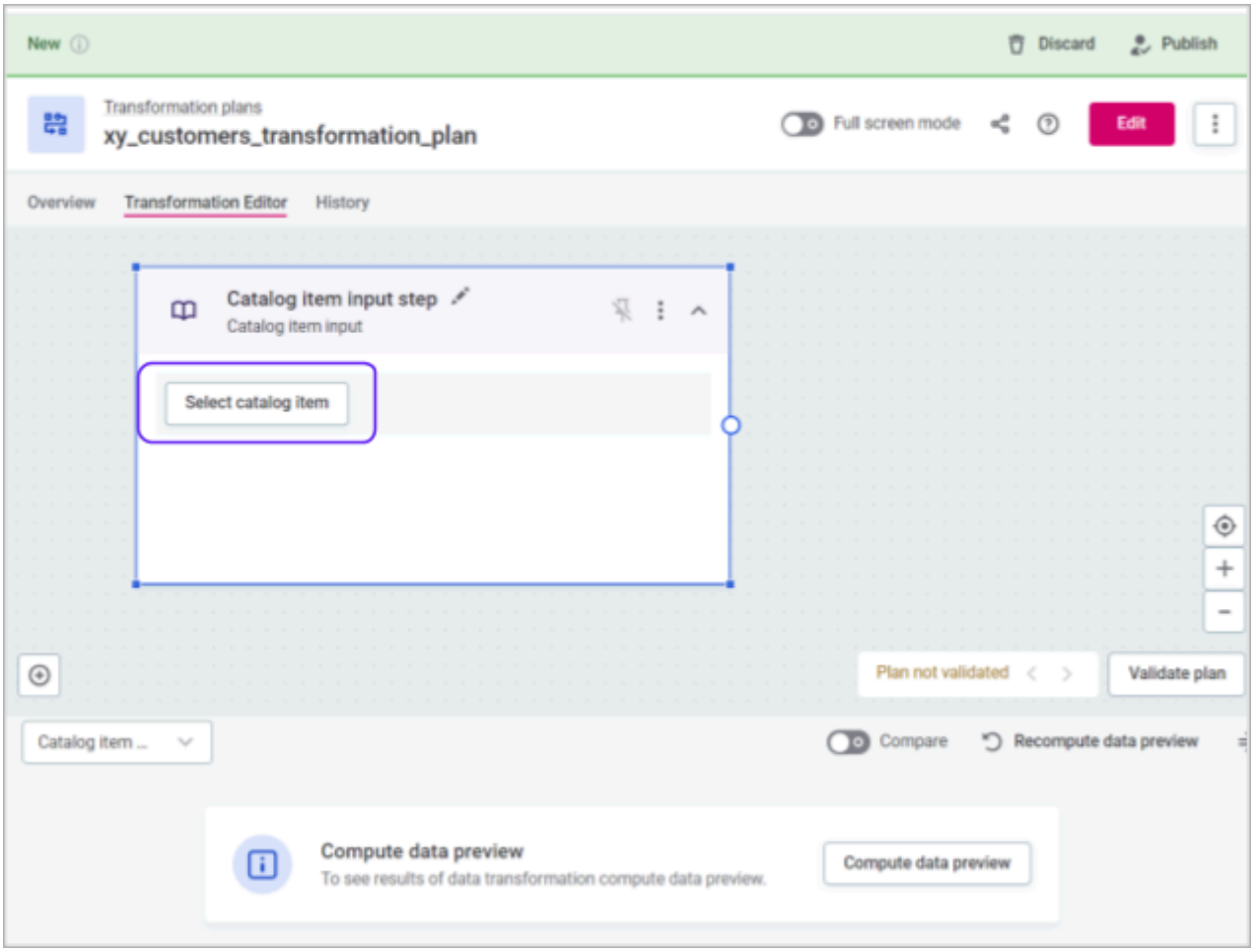
Task 1.2 - Set Up Data Input and output

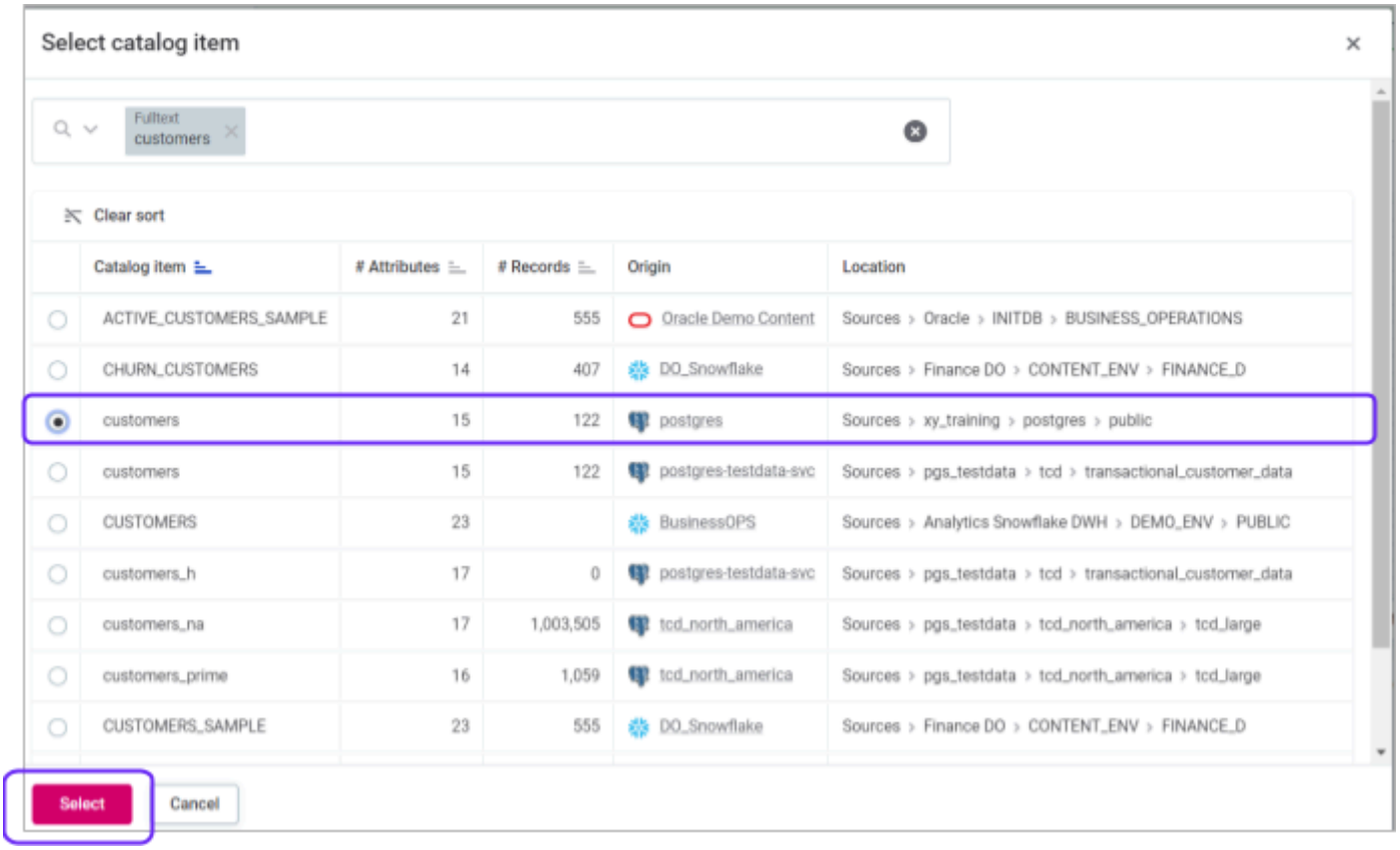
Once the plan is created, we need to add data input:

- In **Transformation Editor** tab and Click on the Add a data input step and add Catalog item input.



- Click anywhere on the blue banner to maximize or minimize a step and access the three dots menu.
- In the **Catalog item**, select the catalog item you would like to use as an input. In our case, this is '**customers**'.


New
Discard
Publish
Transformation plans
xy_customers_transformation_plan
Full screen mode
Edit
Overview
Transformation Editor
History
Catalog item input step
Catalog item input
Select catalog item
Plan not validated
Validate plan
Catalog item ...
Compare
Recompute data preview
Compute data preview
To see results of data transformation compute data preview.
Compute data preview


Select catalog item
Fulltext
customers
Clear sort
Catalog item
Attributes
Records
Origin
Location
ACTIVE_CUSTOMERS_SAMPLE 21 555 Oracle Demo Content Sources › Oracle › INITDB › BUSINESS_OPERATIONS
CHURN_CUSTOMERS 14 407 DO_Snowflake Sources › Finance DO › CONTENT_ENV › FINANCE_D
customers 15 122 postgres Sources › xy_training › postgres › public
customers 15 122 postgres-testdata-svc Sources › pgs_testdata › tcd › transactional_customer_data
CUSTOMERS 23 BusinessOPS Sources › Analytics Snowflake DWH › DEMO_ENV › PUBLIC
customers_h 17 0 postgres-testdata-svc Sources › pgs_testdata › tcd › transactional_customer_data
customers_na 17 1,003,505 tcd_north_america Sources › pgs_testdata › tcd_north_america › tcd_large
customers_prime 16 1,059 tcd_north_america Sources › pgs_testdata › tcd_north_america › tcd_large
CUSTOMERS_SAMPLE 23 555 DO_Snowflake Sources › Finance DO › CONTENT_ENV › FINANCE_D
Select
Cancel

› **To add more steps**, use the plus icon either in the connection or on the step output or select **"Add step."** You can also add steps through the three dots menu, and duplicate or delete them as needed.

NOTE You can utilize keyboard shortcuts (**delete and backspace**) to remove steps. Additionally, on Windows, use the **"Insert"** key to add a new step.

› In this task, we need to add the following steps:

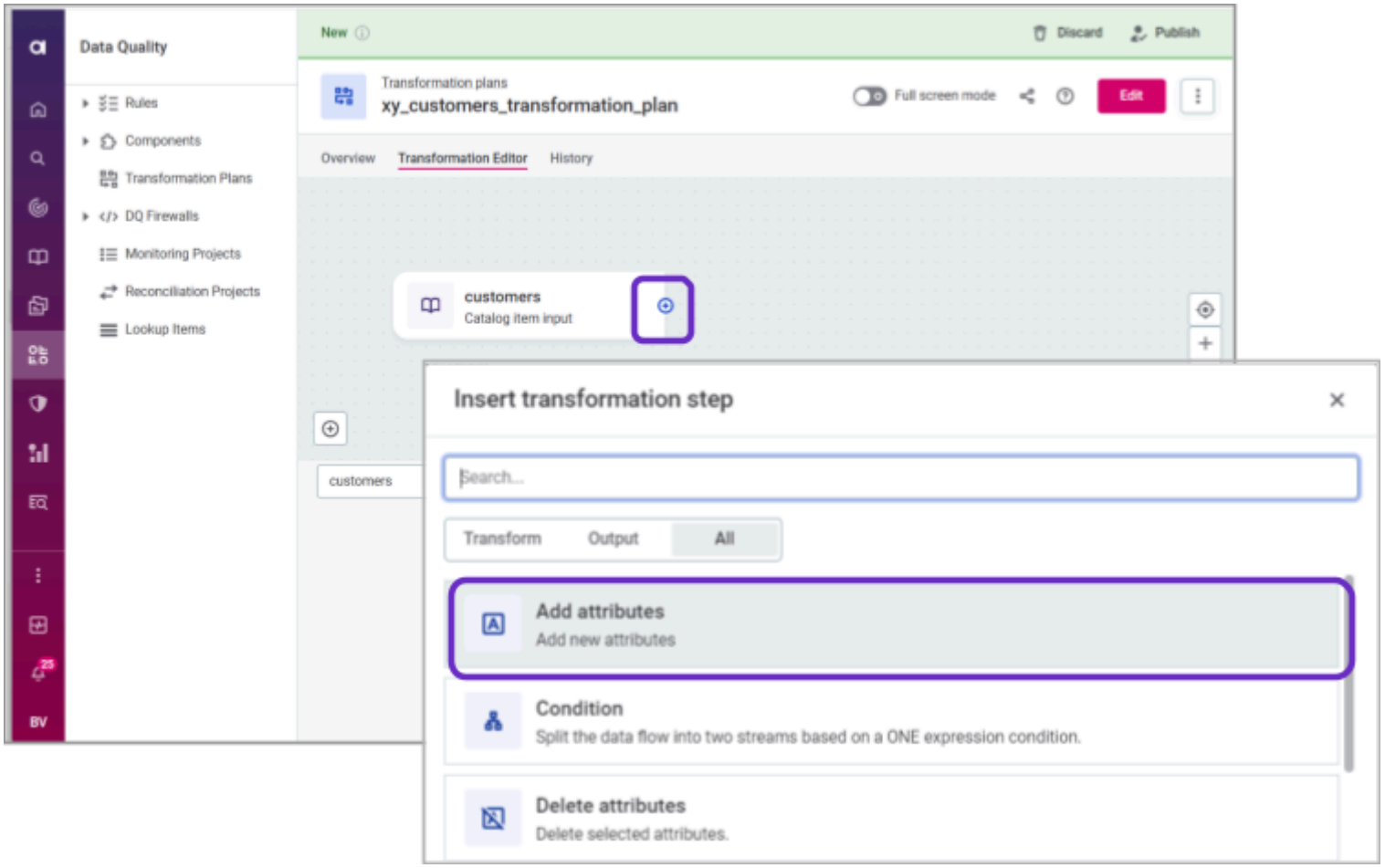
- **Add attributes:** To create a new attribute, **"full_name,"** in the output by concatenating the existing input attributes first name and last name.
- **Delete attributes:** To remove unnecessary attributes, retaining only the key attributes in the output.
- **Filter:** To filter out all records not originating from a specified country; in this instance, everything except those from the United States.

Task 1.3 - Configure transformation steps

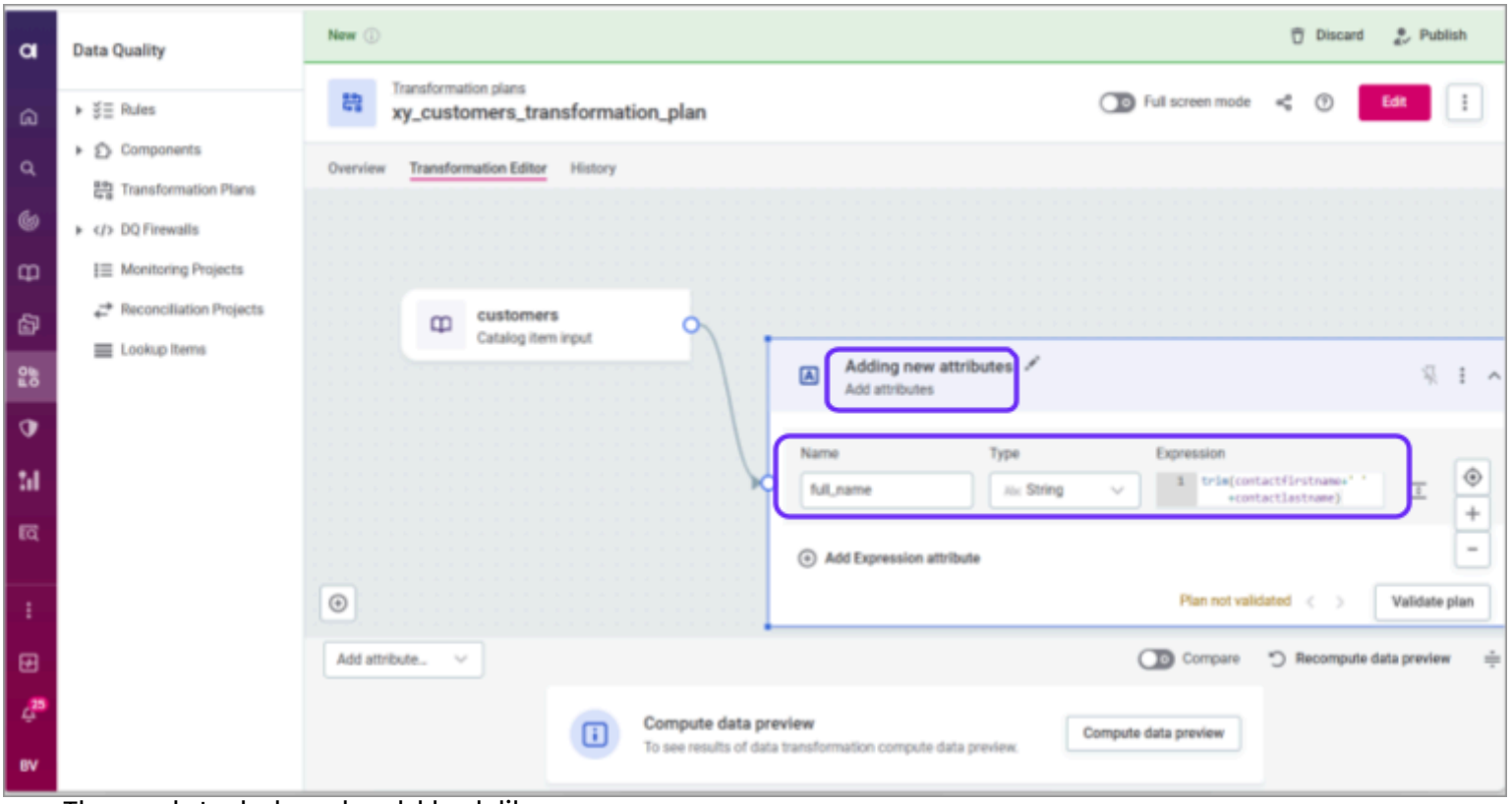
Initially, we aim to generate **a new attribute**, "**full_name**," by concatenating the current input attributes, **"contactfirstname"** and **"contactlastname."**

› To accomplish this, we employ the **"Add attributes"** step with the following configuration:

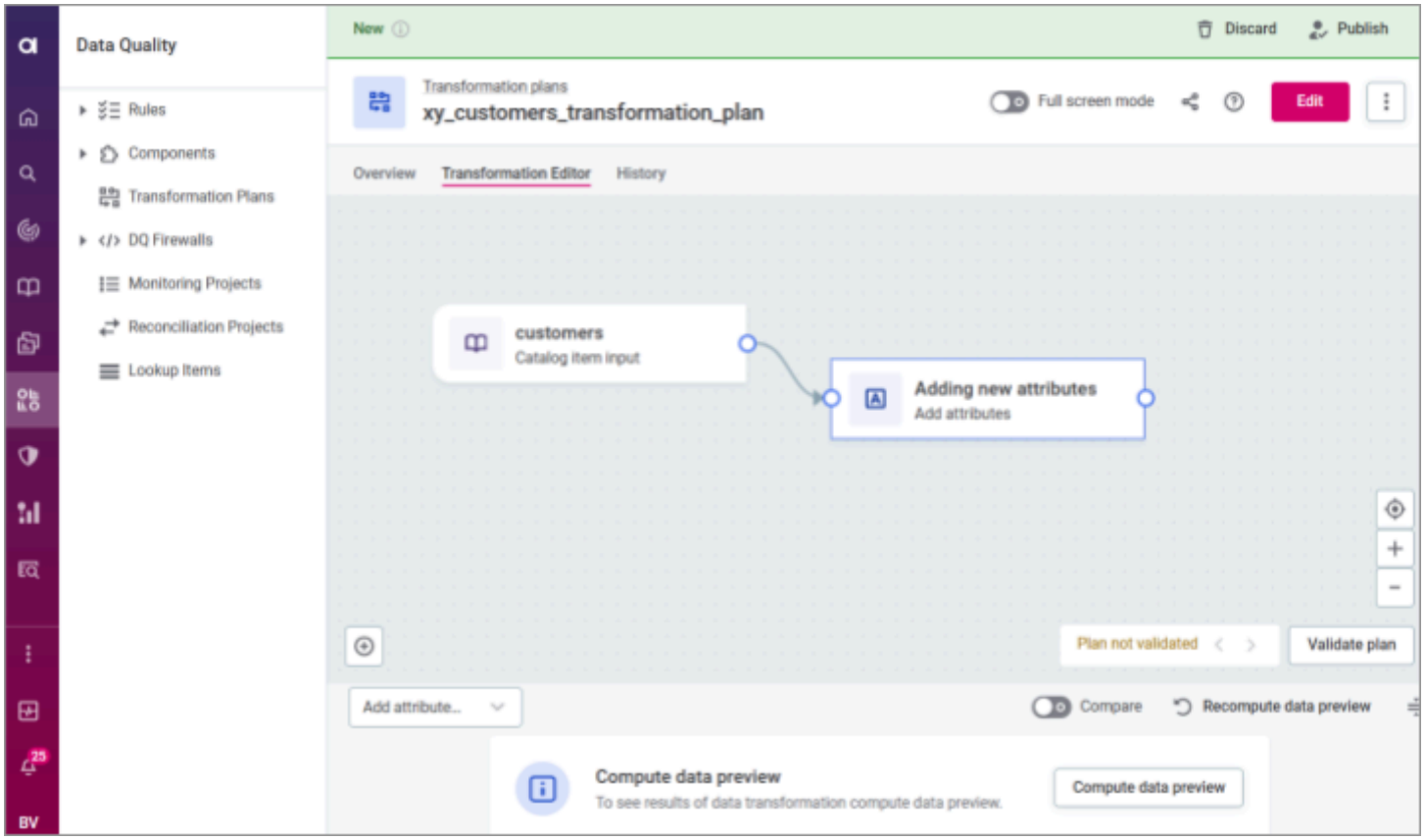
- **Name:** A custom name for the step (optional).
- **Name and Type:** Specifies the attribute name and its data type for the output. In our case, this is **"full_name"** and **"String."**



- **Expression:** Utilizes a **ONE expression** to define the new attribute. In our scenario, the expression is **trim(contactfirstname+' '+contactlastname)**, which also trims the inputs before combining them.



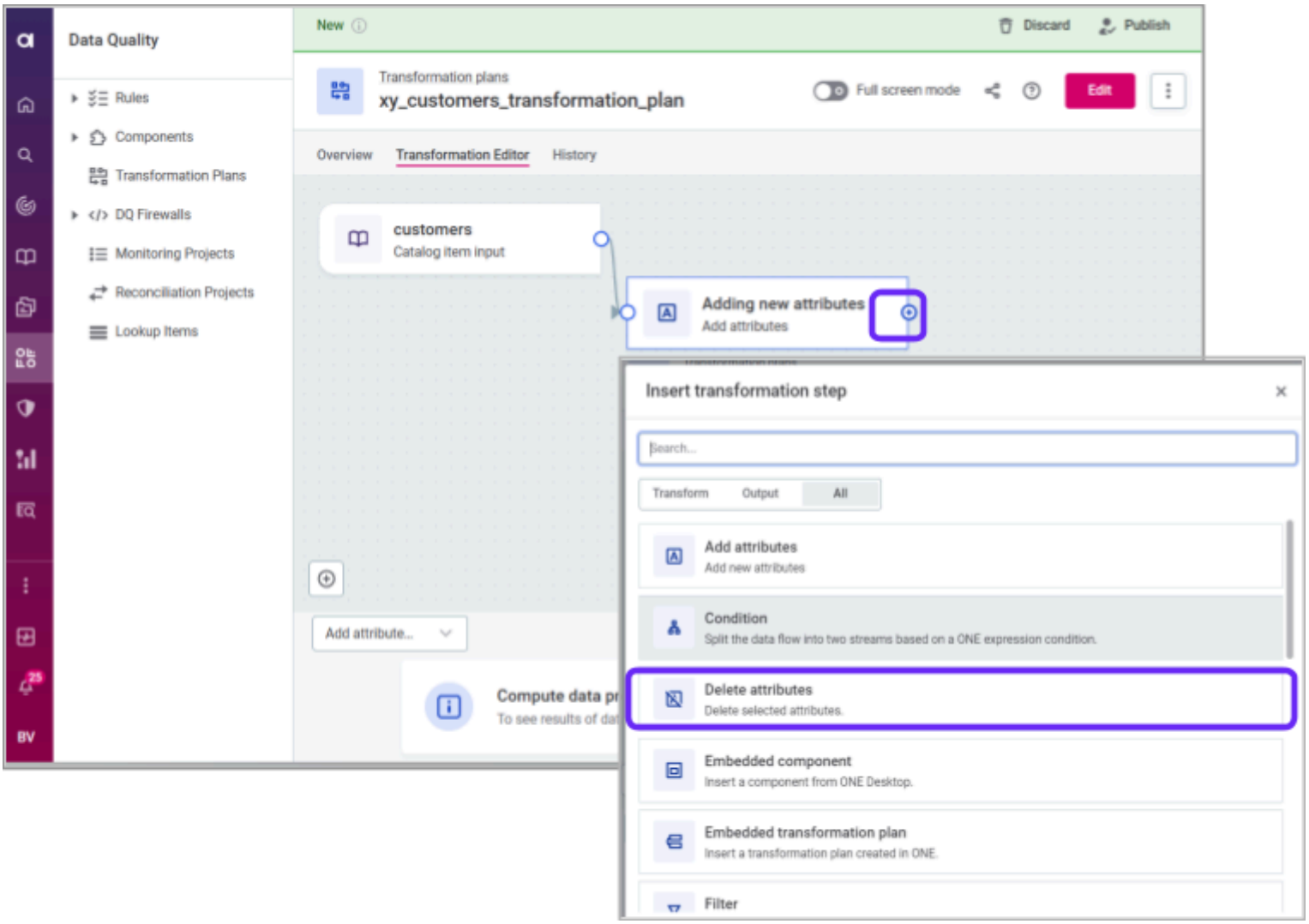
The updated plan should look like:



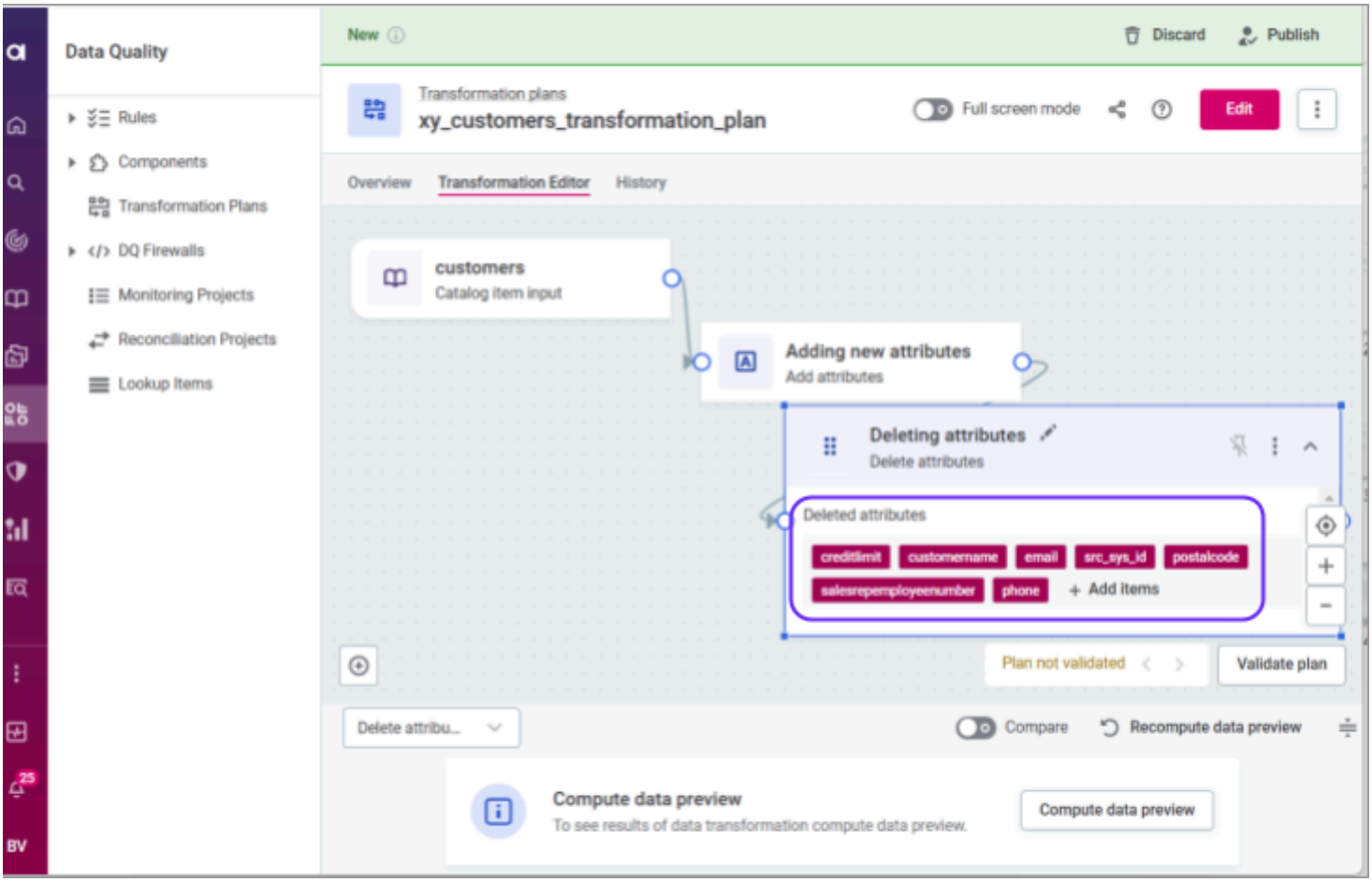
There are some attributes in our catalog item, **customers**, that are unnecessary in the output.

- To eliminate these attributes, we employ the **"Delete attributes"** step with the following configuration:

  Name: A custom name for the step (optional).

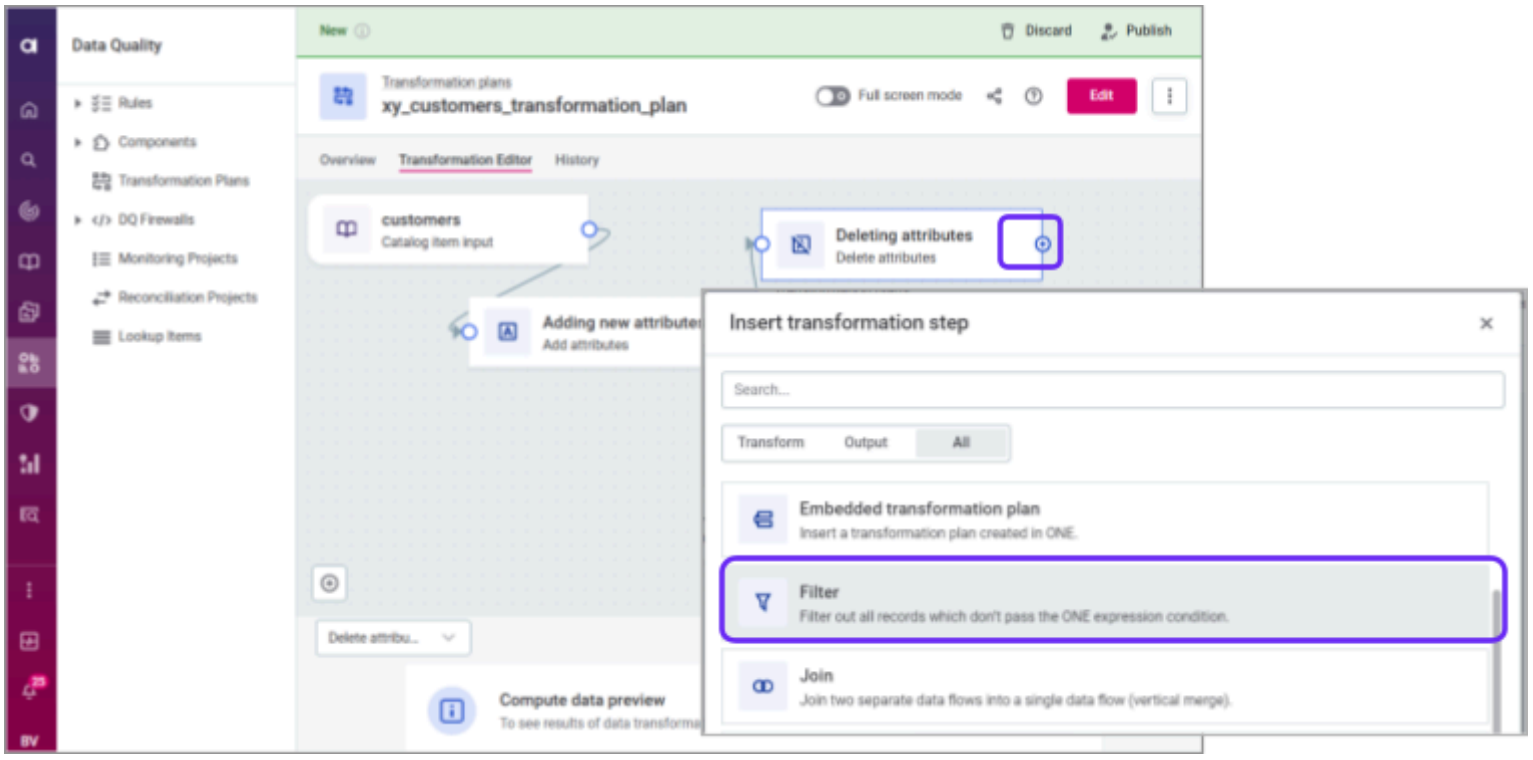


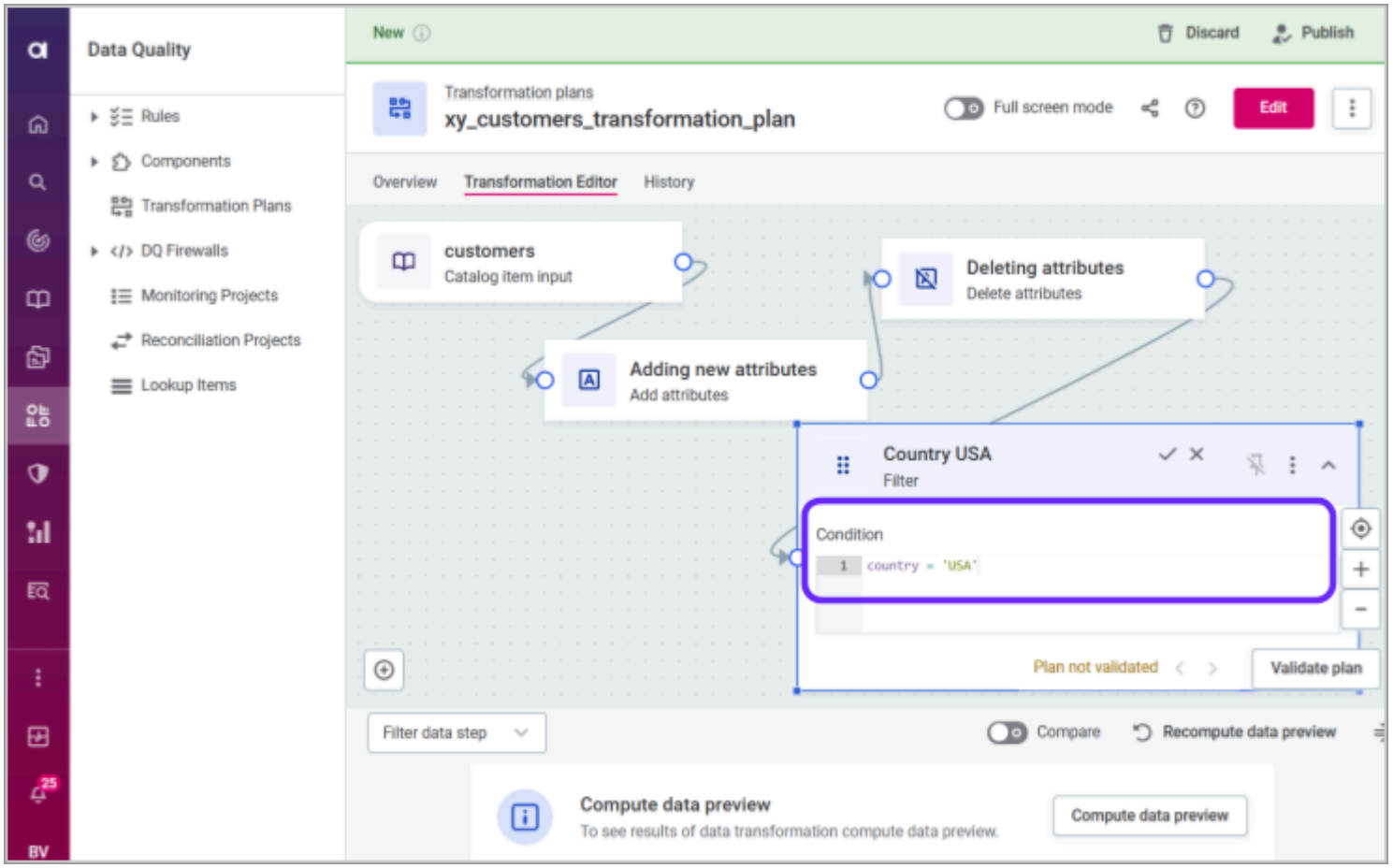
- Use the **+Add option** to select attributes from the available options. In our instance, these include **creditlimit, customername, email, phone, postalcode, salesemployeenumber,** and **src_sys_id**.



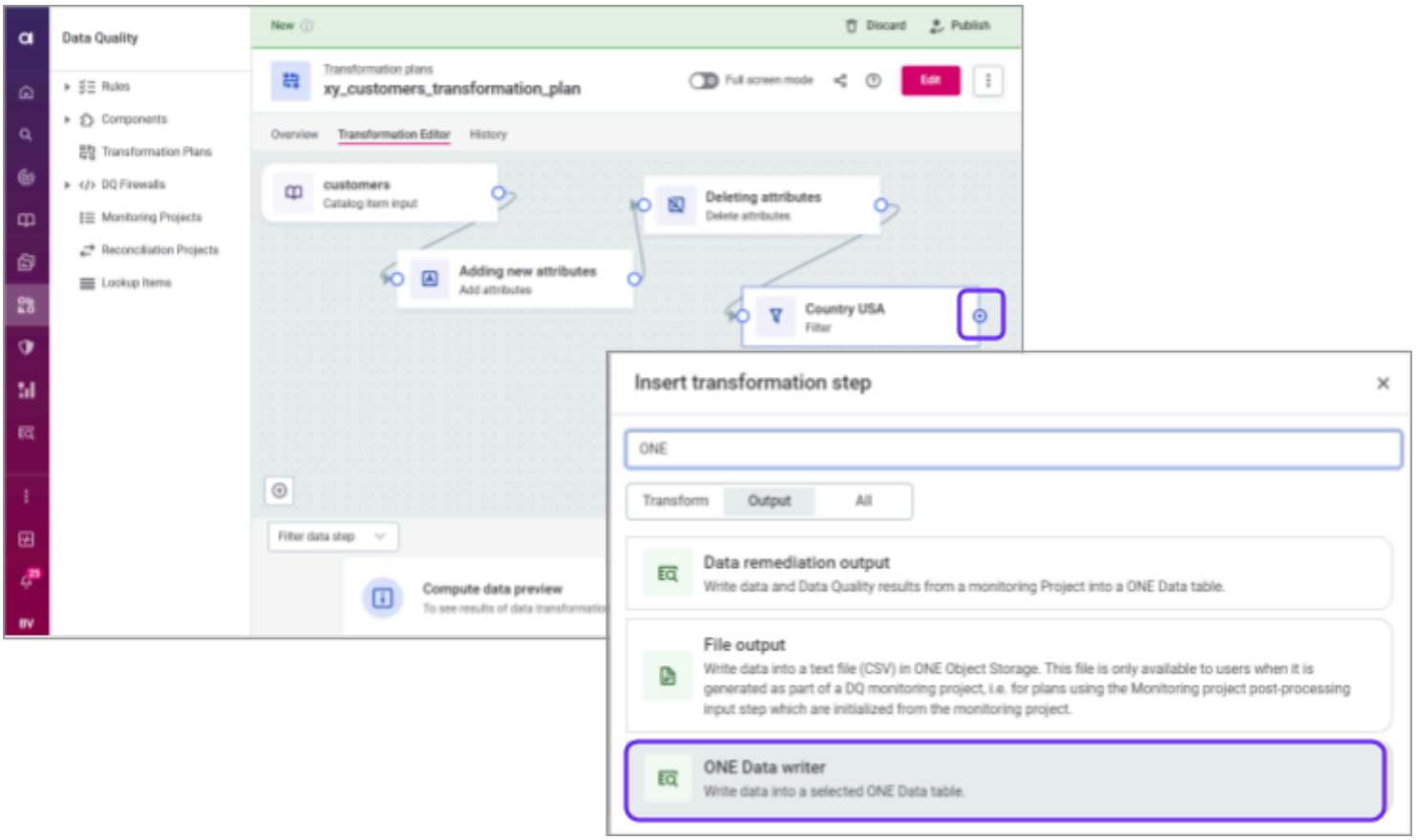
› To filter out all records which are not from the United States, we can use the **Filter step**. It is configured as follows:

- **Name:** A custom name for the step (optional).
- **Condition:** The filter condition is specified using ONE Expressions. In our case, this is going to be **attribute name = 'value',** resulting in **country = 'USA'**.





› Finally, add **data output.** In this example, we are using the **ONE Data writer step**.

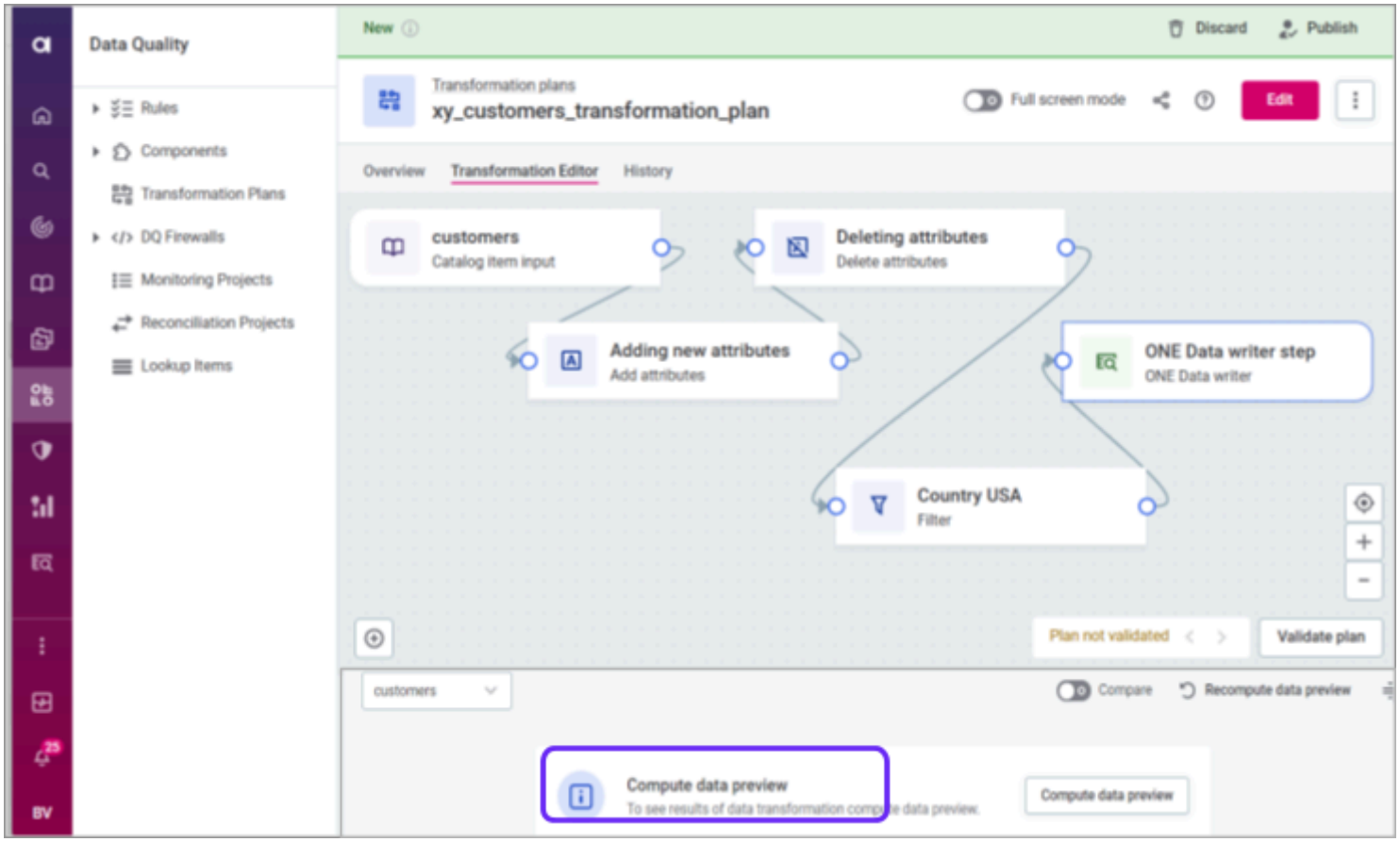


Task 1.4 - Preview Data

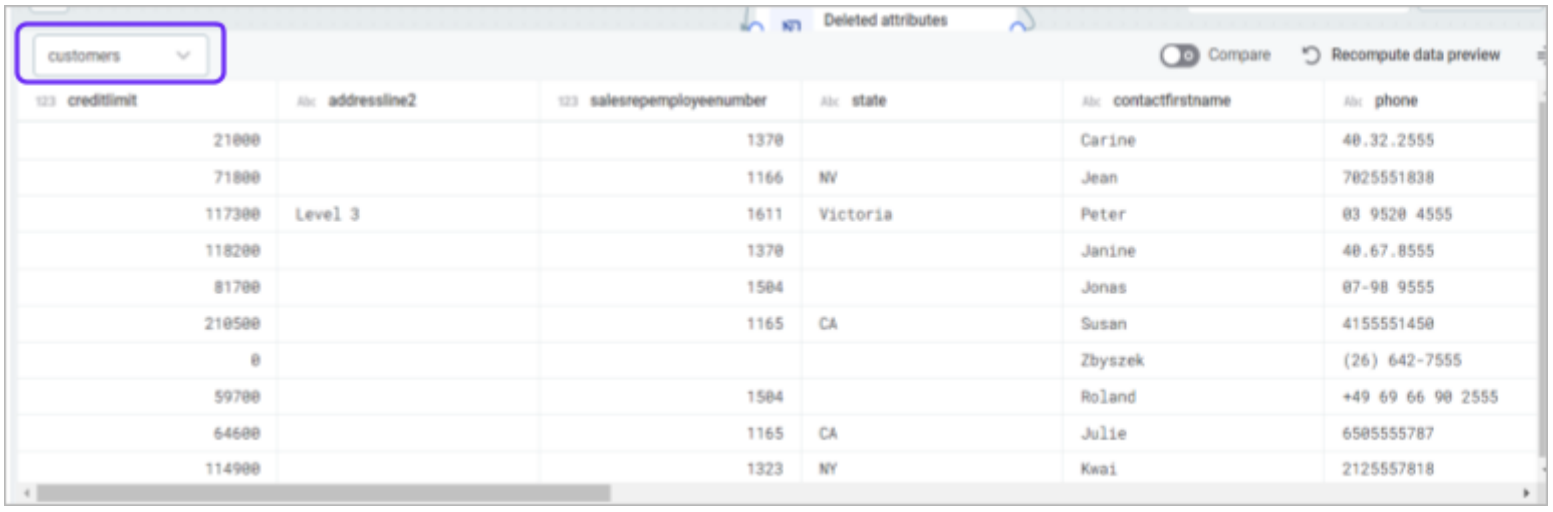
At any point, you can preview the data resulting from the chosen step and confirm that it functions as intended.

Please note that the preview is not automatically recalculated after modifications, so it is necessary to trigger this manually.

- Click on '**Compute data Preview**'. You can also select the step either from the drop-down list or by clicking directly on the step on the canvas.



- Your result page looks like this:



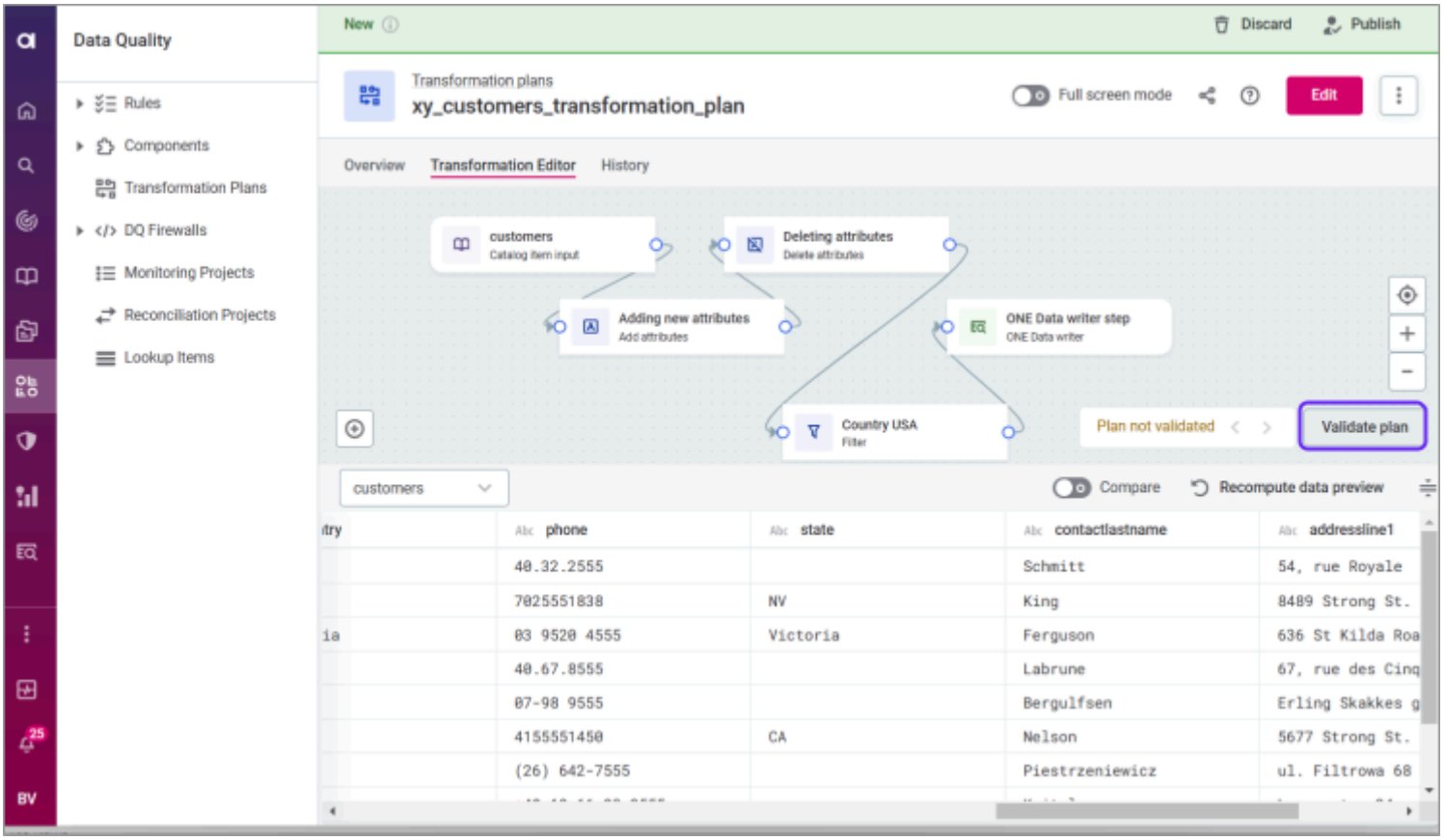
creditlimit	addressline2	salesrepemployeenumber	state	contactfirstname	phone
21000		1370		Carine	40.32.2555
71800		1166	NV	Jean	7025551838
117300	Level 3	1611	Victoria	Peter	03 9520 4555
118200		1370		Janine	40.67.8555
81700		1504		Jonas	07-98 9555
210500		1165	CA	Susan	4155551450
0				Zbyszek	(26) 642-7555
59700		1504		Roland	+49 69 66 90 2555
64600		1165	CA	Julie	6505555787
114900		1323	NY	Kwai	2125557818

NOTE

Preview is not visible for output steps, but you can select the preceding steps and check their outputs instead.

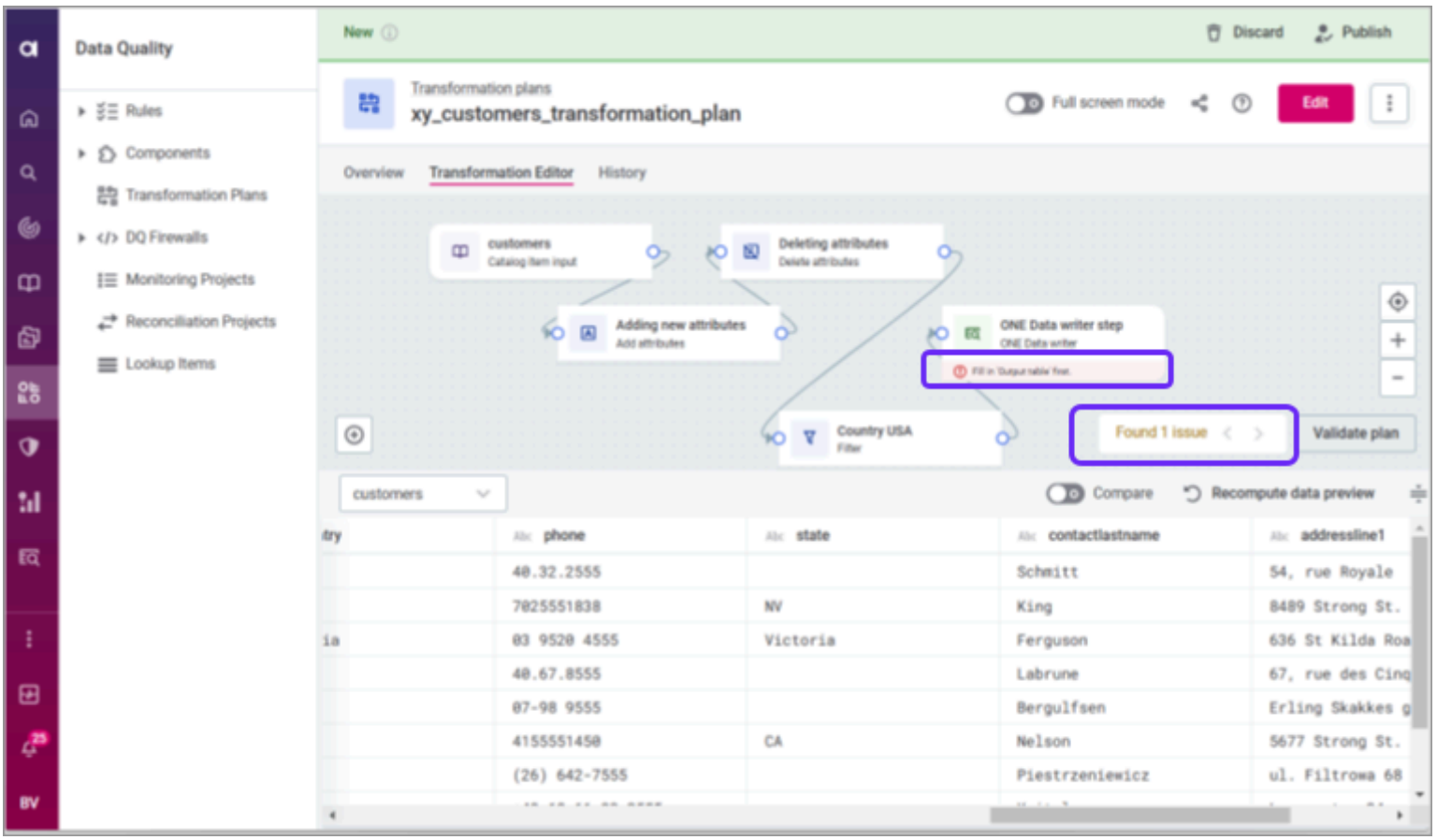
Task 1.5 - Validate Plan

You can also validate the plan during the process by choosing **"Validate plan"** in the bottom right corner of the canvas.



Any errors, such as incorrect expression conditions, will be brought to your attention.

In case validation issues persist, selecting **"Run transformation"** will result in the failure of the data processing job.



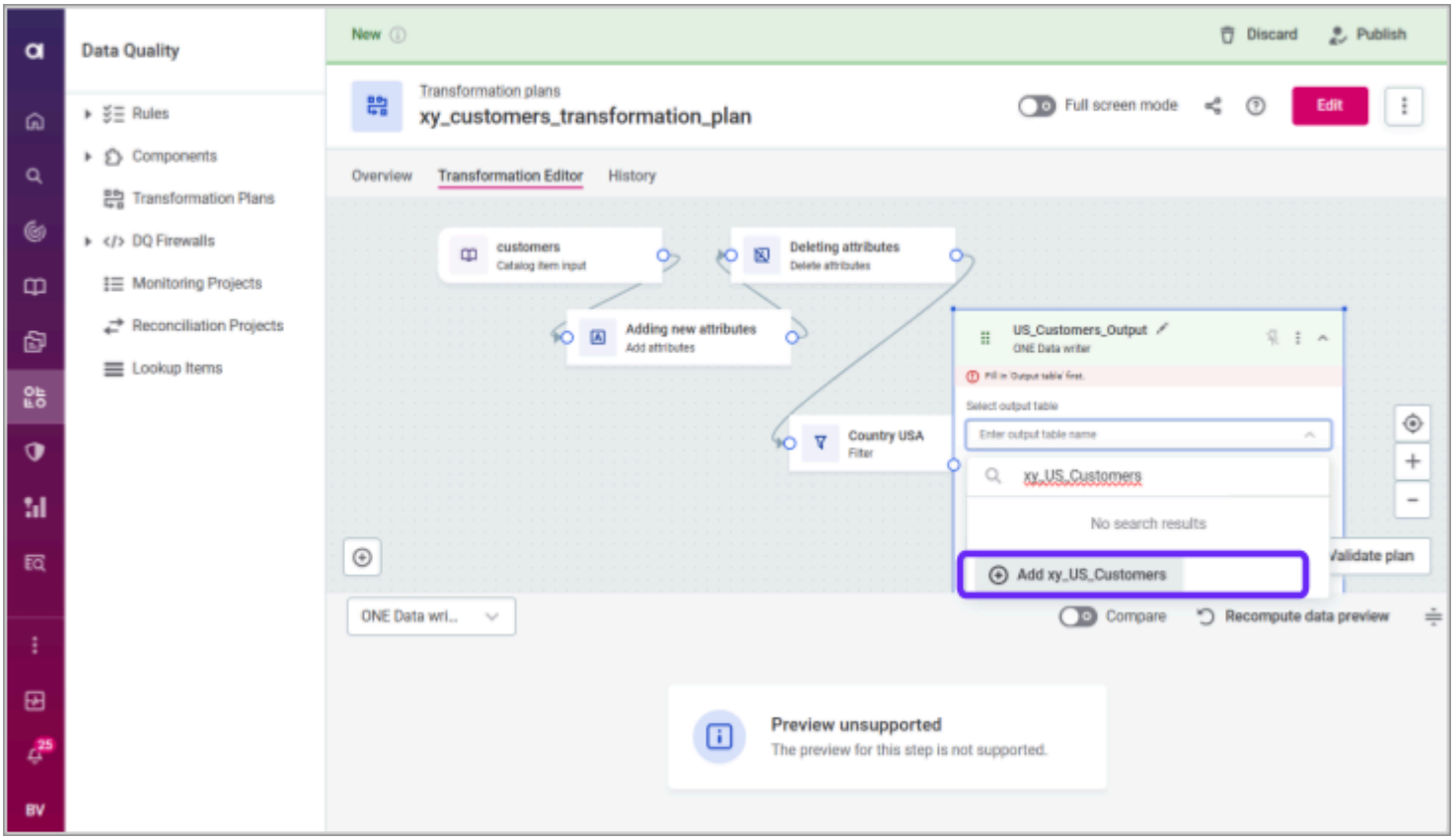
Task 1.6 - Configure Output Step

We want to write the data to a new table in ONE Data, using the ONE Data writer. It is configured as follows:

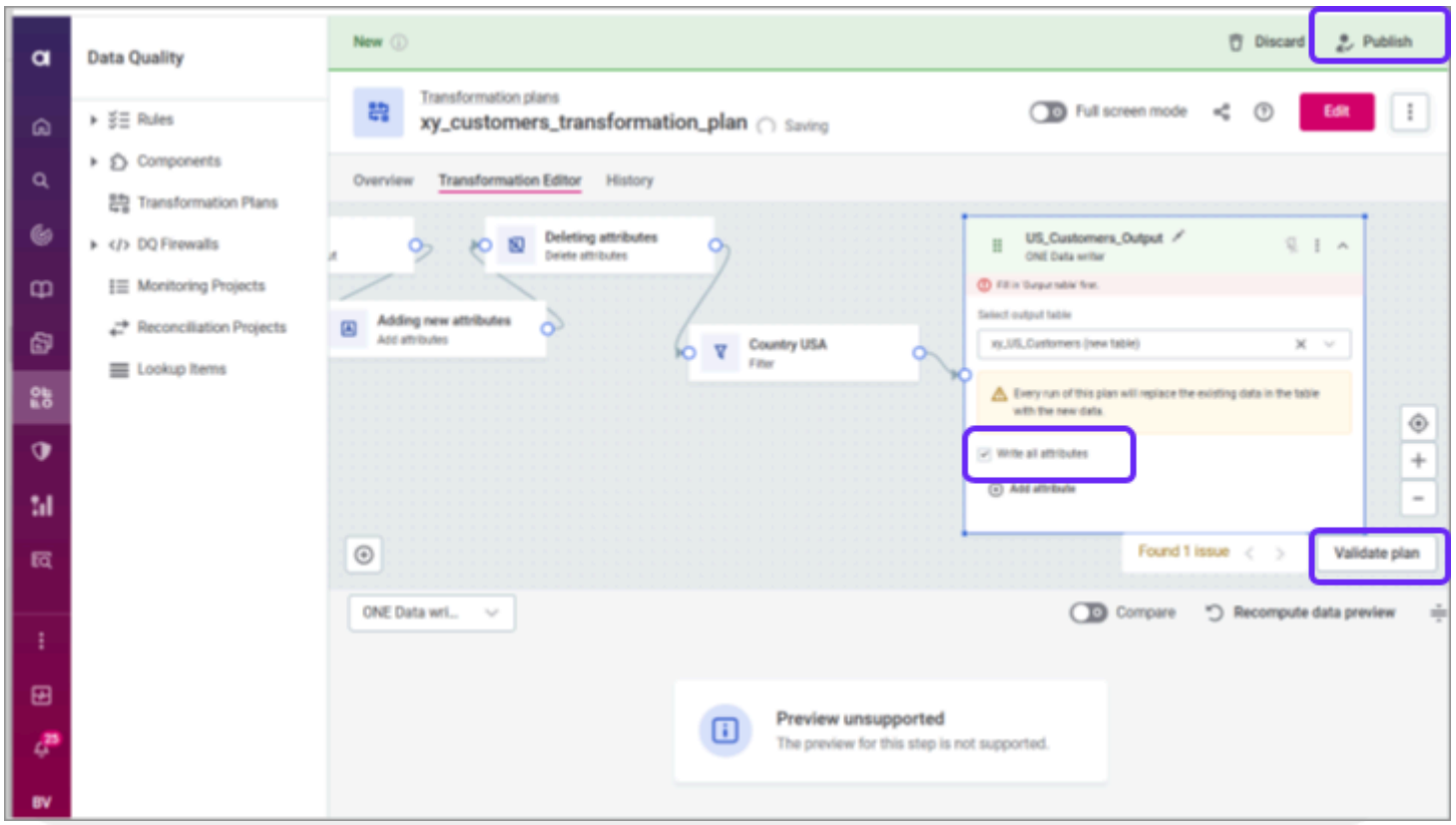
- **Name:** A custom name for the step (optional).
- **File Output Step :** Specify the desired name for the catalog item generated in ONE Data by this step.

Name : **US_Customers_Output**

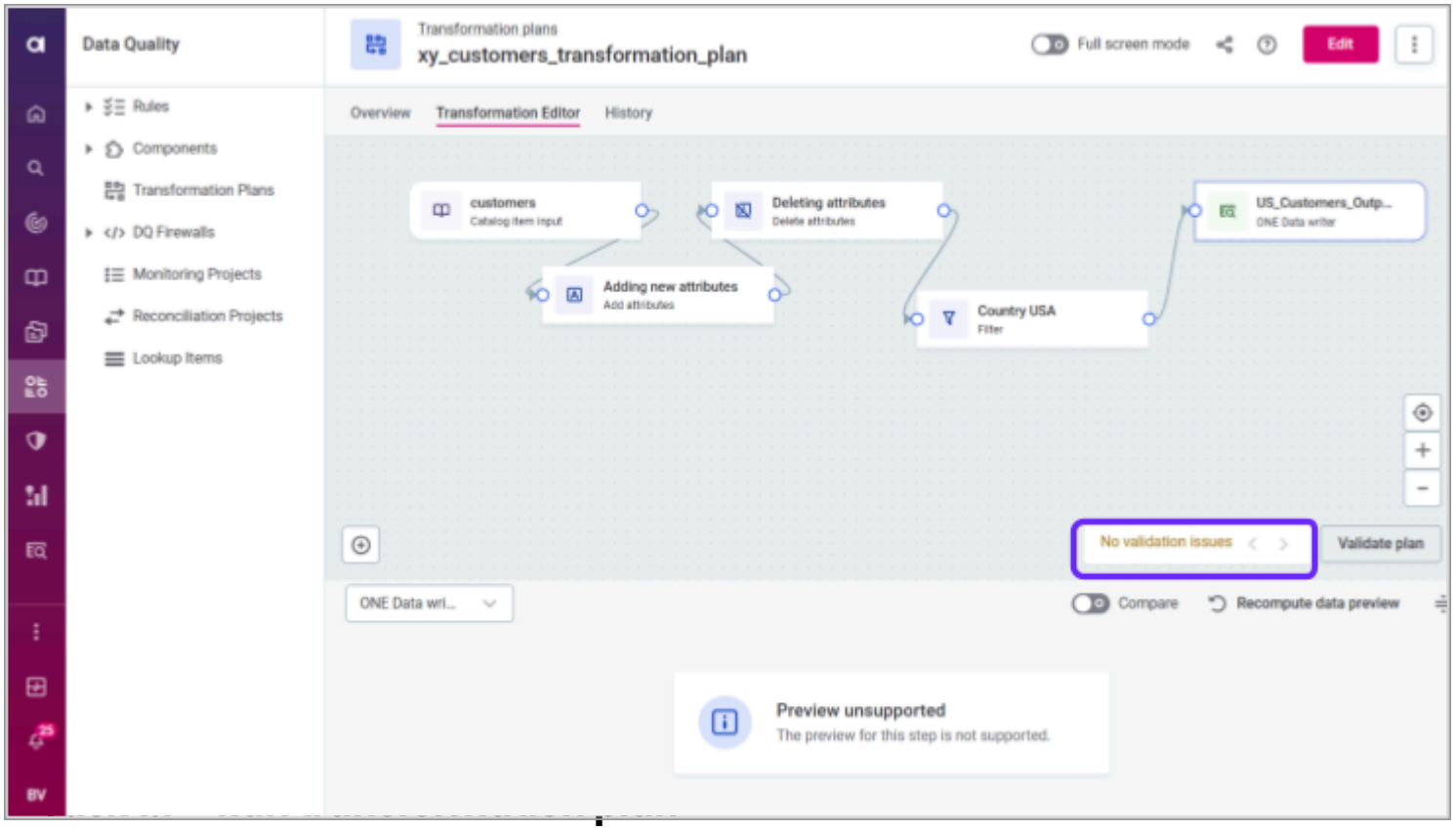
ONE Data table Name : **<name_prefix>_US_Customers**



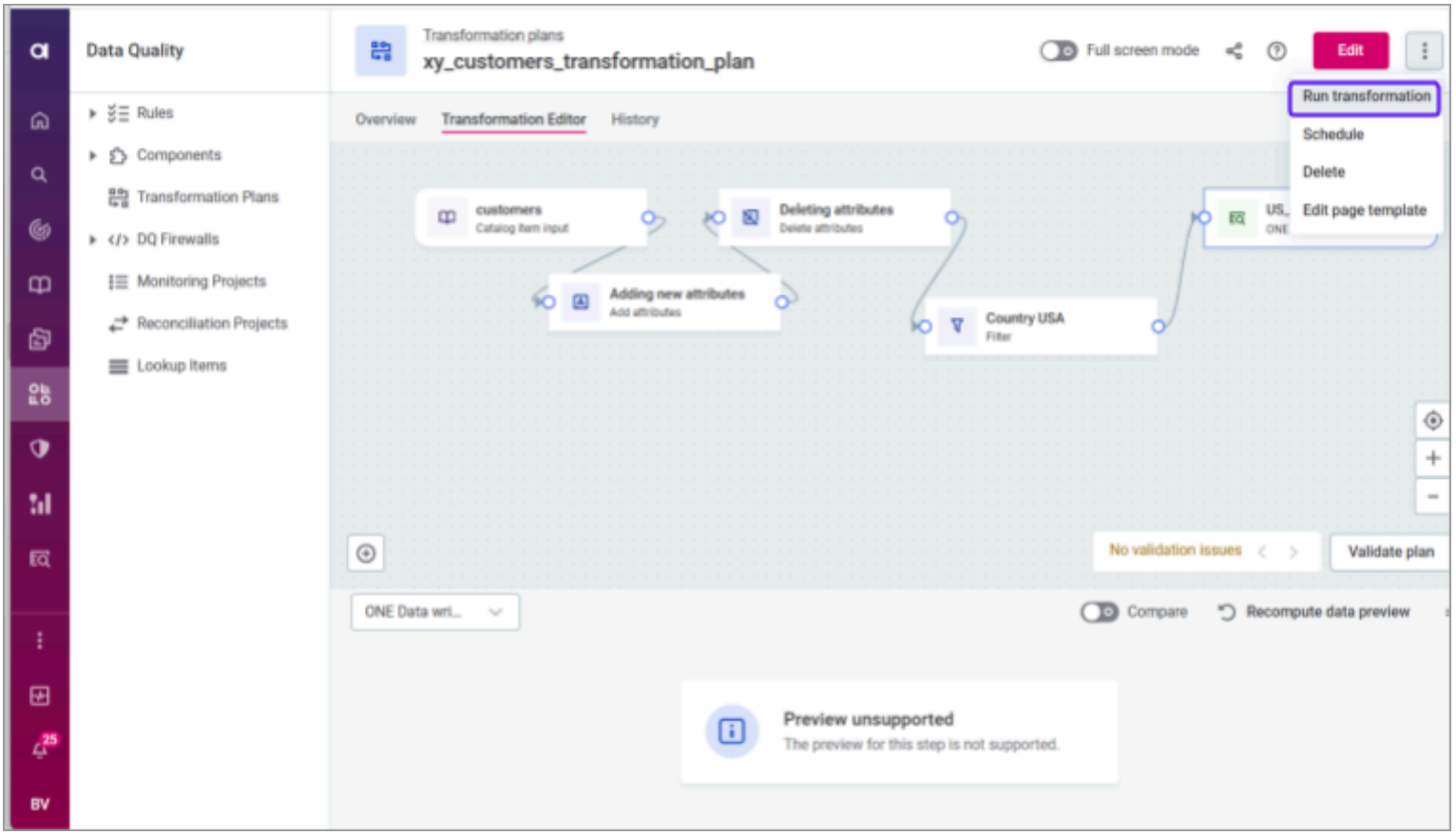
› Select **Write all attributes** to include all attributes which are inputs to the step, without having to add them manually.

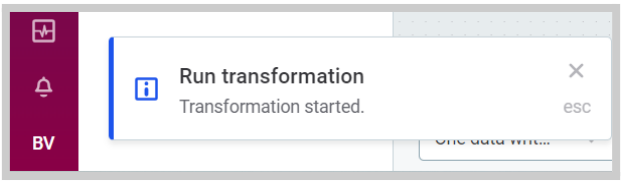


› **Publish** the Changes and Click the **Validate Plan** button again to ensure there are **no validation issues**.



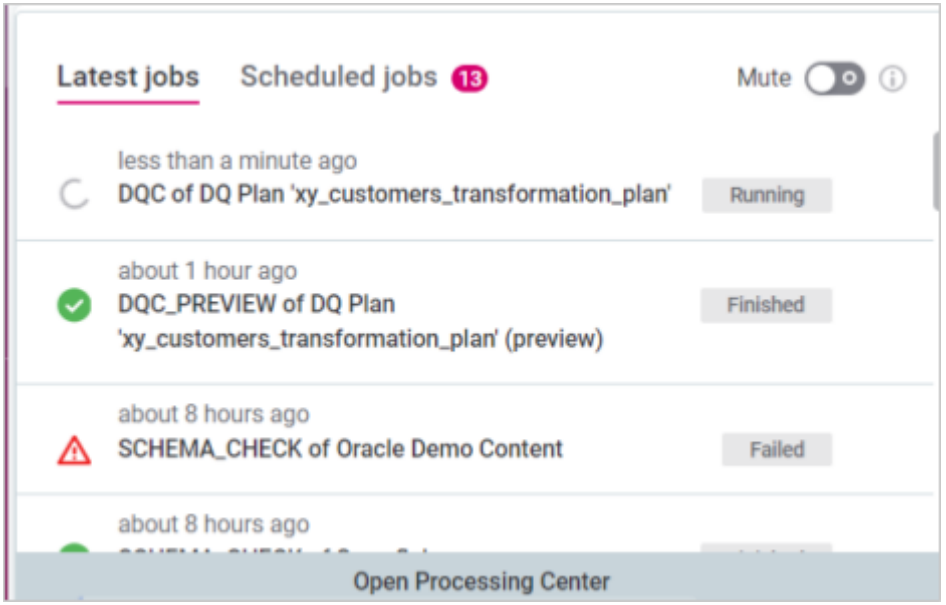
› Navigate to the three dots menu on the top right corner and choose **"Run transformation."**



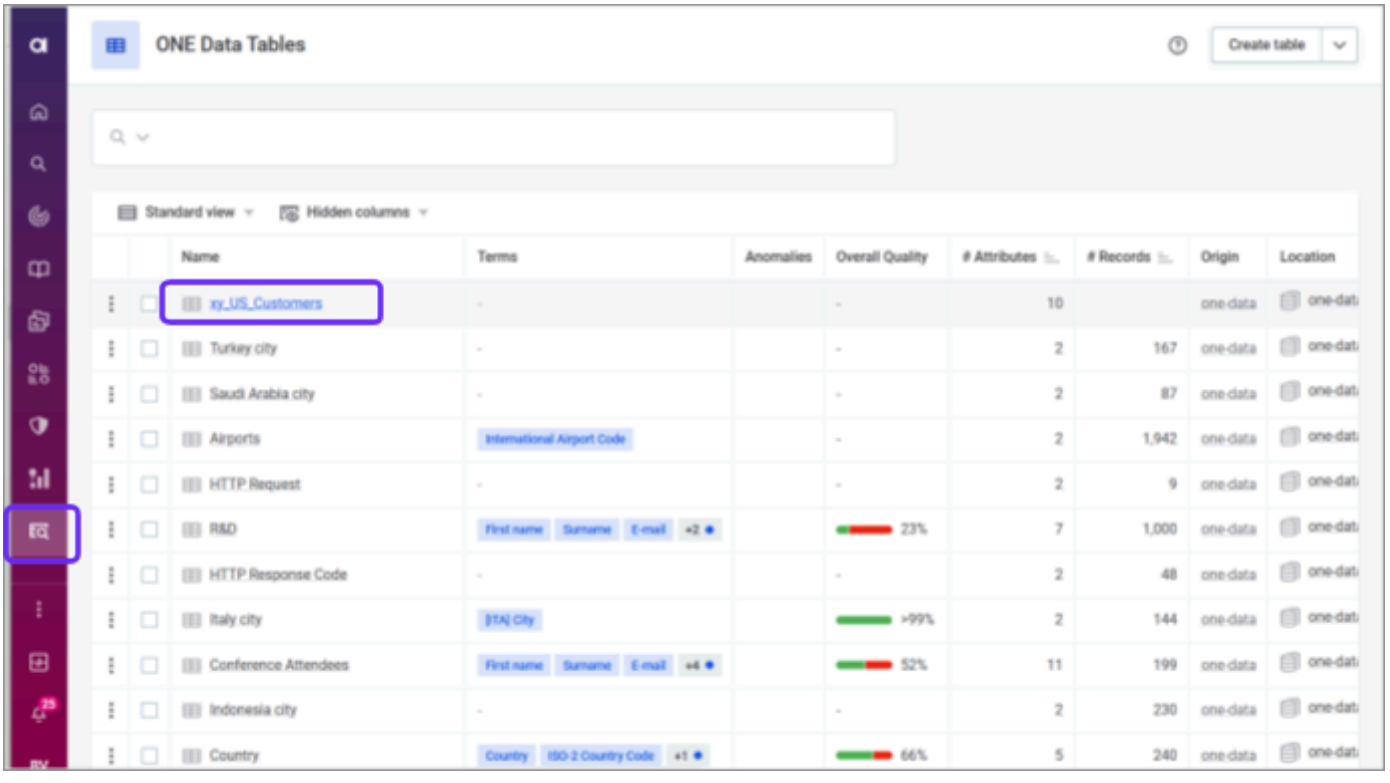


NOTE

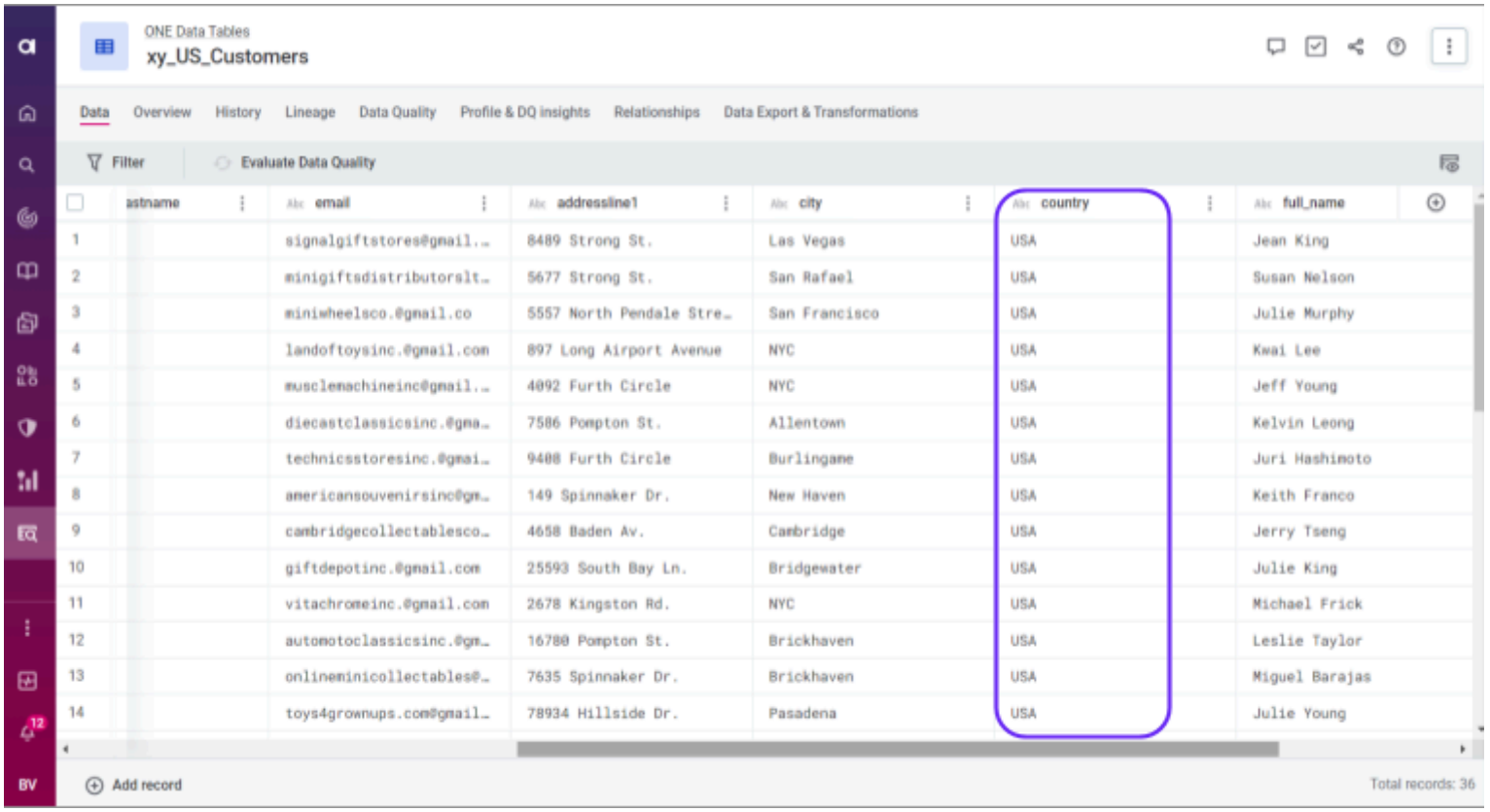
Transformation plan jobs can be found by going to **Processing center › Base jobs › Transformation plan jobs**, or by searching for job Type DQC.



› Navigate to the **ONE Data** section on the left panel and search for the **'<name_prefix>_US_Customers'** table.



› Click on the **'<prefix_name>_US_Customers'** table and verify the results.



Conclusion

We've reached the conclusion of this workshop! Throughout this session, we've successfully created data transformation plans for catalog items.In this way, data transformation plans take precedence over post-processing plans.

Congratulations! You are now ready to embark on the development of your own Data Transformation Plan.