Data profiling and discovery with Curiosity Software

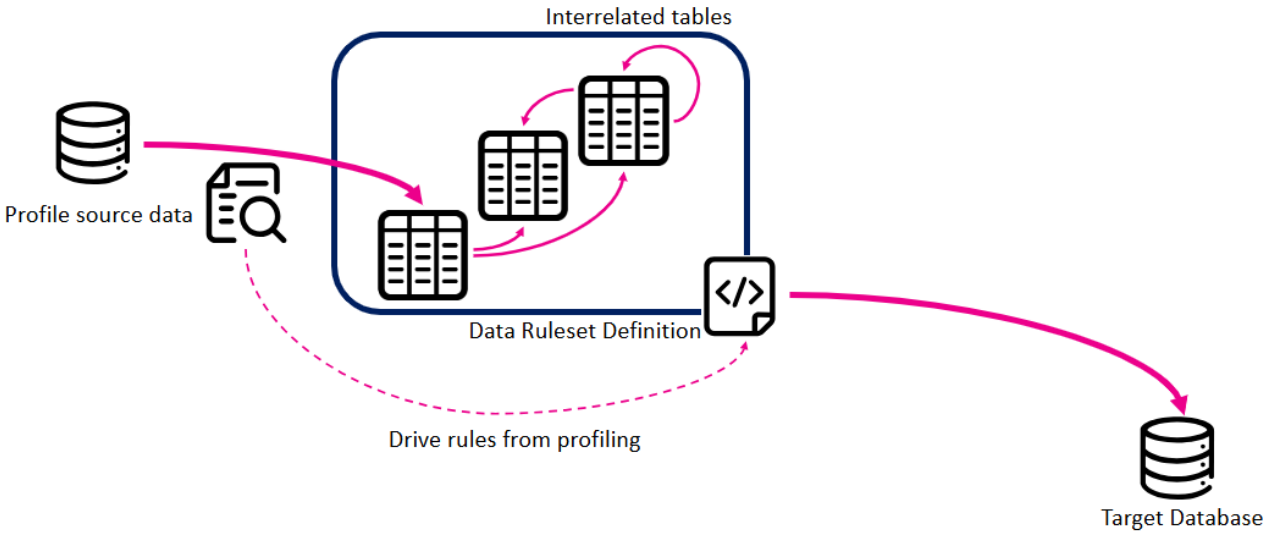
Introduction:

The Curiosity Platform provides an extensive corporate dictionary allowing users to understand all aspects of their data throughout their data ecosystem. Definitions are used to track and store detailed information about the databases or files we work with and are at the heart of the corporate dictionary Curiosity provides. In addition, the platform has extensive abilities to profile your data and conduct deep scanning of your data to populate the corporate dictionary with a rich set of data.

In this module, we'll walk through the process of setting up the corporate dictionary and how to set up the scanning to populate it with rich, meaningful data which can then be leveraged in your data activities and test data pipelines.

Training overview:

This training course will take you through how to set up a connection to your data source and how to scan your data. In addition, it will also take you through how to set a profiling activity and how to leverage our AI technology to take your profiling to the next level.



By the end of this self-led training, you will be able to do the following:

- Create a connection to database server
- Scan a database
- Create a definition from the scan
- Set up a Scan Data Activity
- Use AI to profile and set it up as a regular task

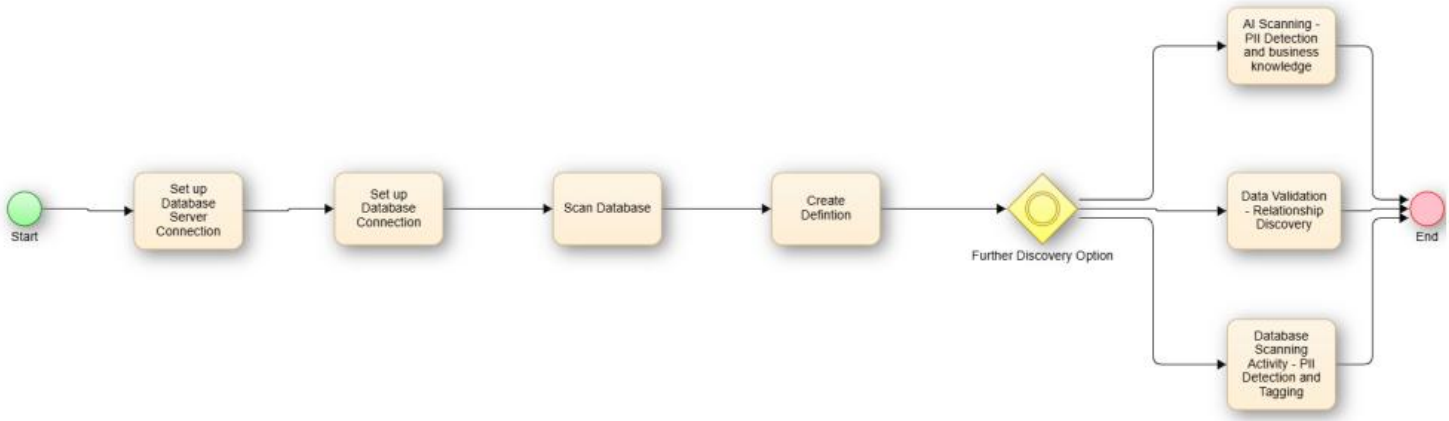
Key capabilities:

- **Catalogue**: Automatically map, catalogue, and visualise all data assets across your organisation.
- **Sensitive data identification:** Detect and classify sensitive data such as Personally Identifiable Information and Protected Health Information. Understand where your most critical data is stored and how it's being used to ensure compliance with regulations.
- **Relationship Mapping:** Uncover how your data assets are interconnected. Automatically map relationships between tables, fields, and systems, providing a complete picture of how data flows through your organisation.
- **Data gap analysis:** Identify gaps and missing data elements that could hinder your analytics, testing, or compliance strategies. Our platform provides detailed reports, and synthetic data generation capabilities, to ensure that every scenario is accounted for.

Pre-requisites for profiling and discovery training:

- Access to the Curiosity Platform
- Connectivity between Curiosity and your data source

You can follow this high-level process diagram for setting up a connection and database scan activity:

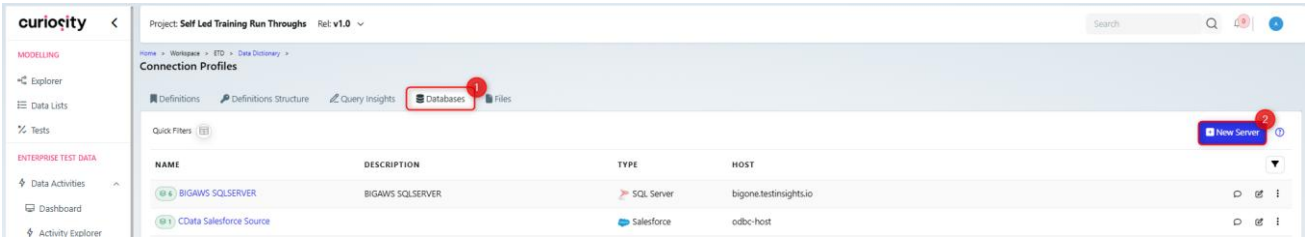


Section 1 – Set up server and database connections

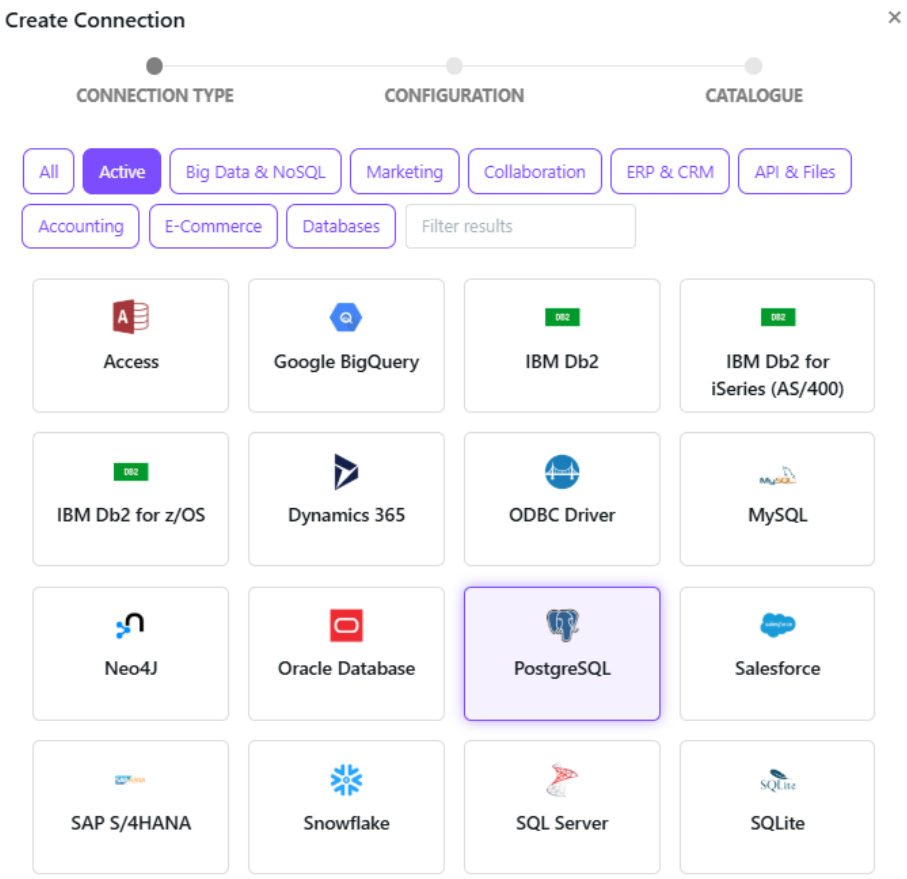
1. To create a new database connection, navigate to the **Data Dictionary** and onto the **Databases** tab.

2. Click on **+New Server**

When you click on the Data Dictionary it takes you first to the definitions tab, therefore we recommend double checking you have clicked on the databases tab.



3. Choose the type of database you wish to connect to:

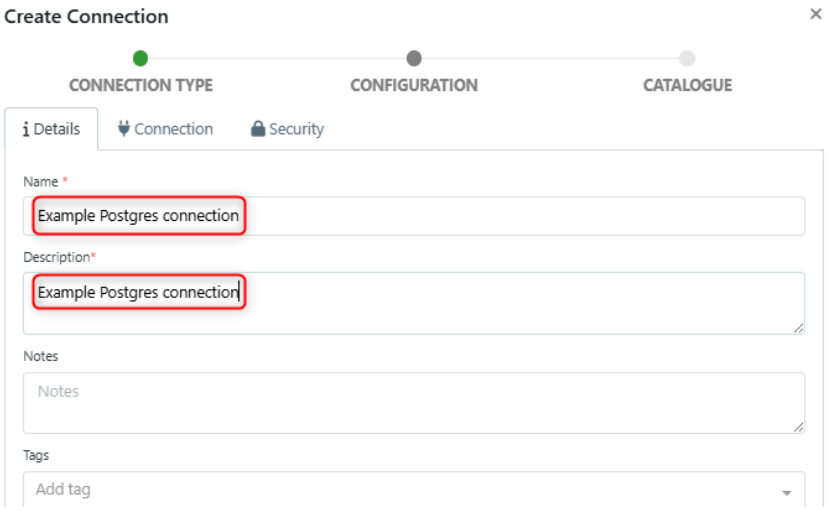


This will open the 'Create Connection' form.

4. Create the server connection

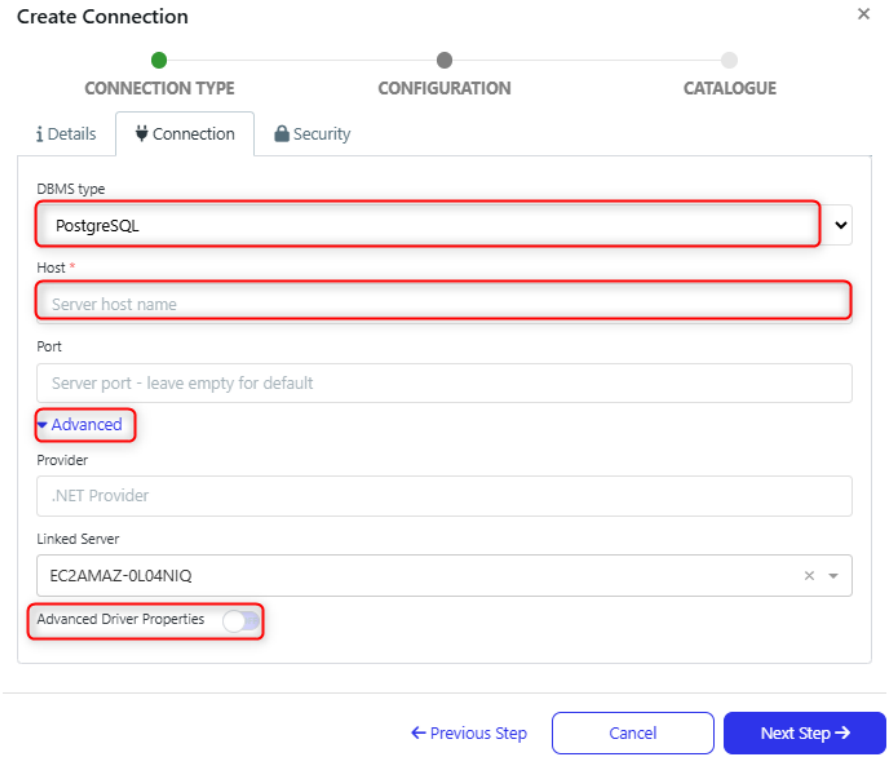
   The **Details** tab contains the high-level information about the connection.

   In the details tab the name and description fields are mandatory. We recommend having naming standards so that you can easily identify which definition you would like to use.



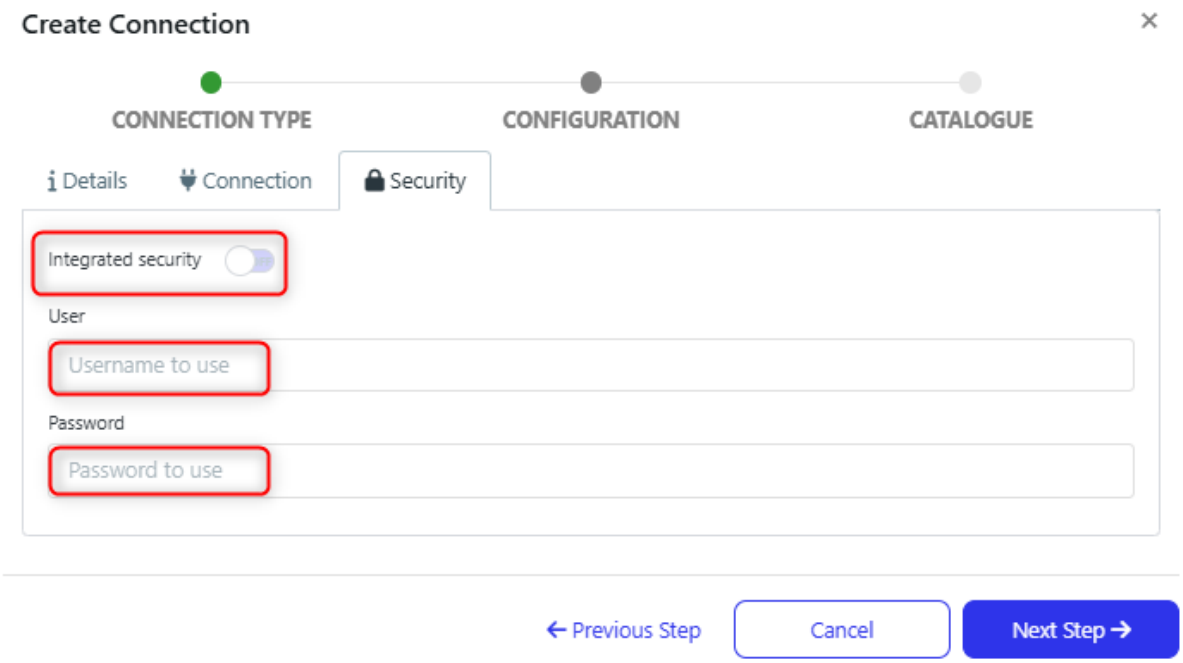
The **Connection** tab has all the relevant connection details that need to be completed.

- **DBMS Type** - Drop down of supported database types this will be auto filled by your previous selection
- **Host** – Enter the host name of the database server
- **Port –** Enter the port number to connect over (if left it will use the default port)
- **Linked Server –** Select the VIP server that we used as default to execute jobs against the connection
- **Advanced Driver Properties –** Configure custom drivers not supplied by the Curiosity Platform

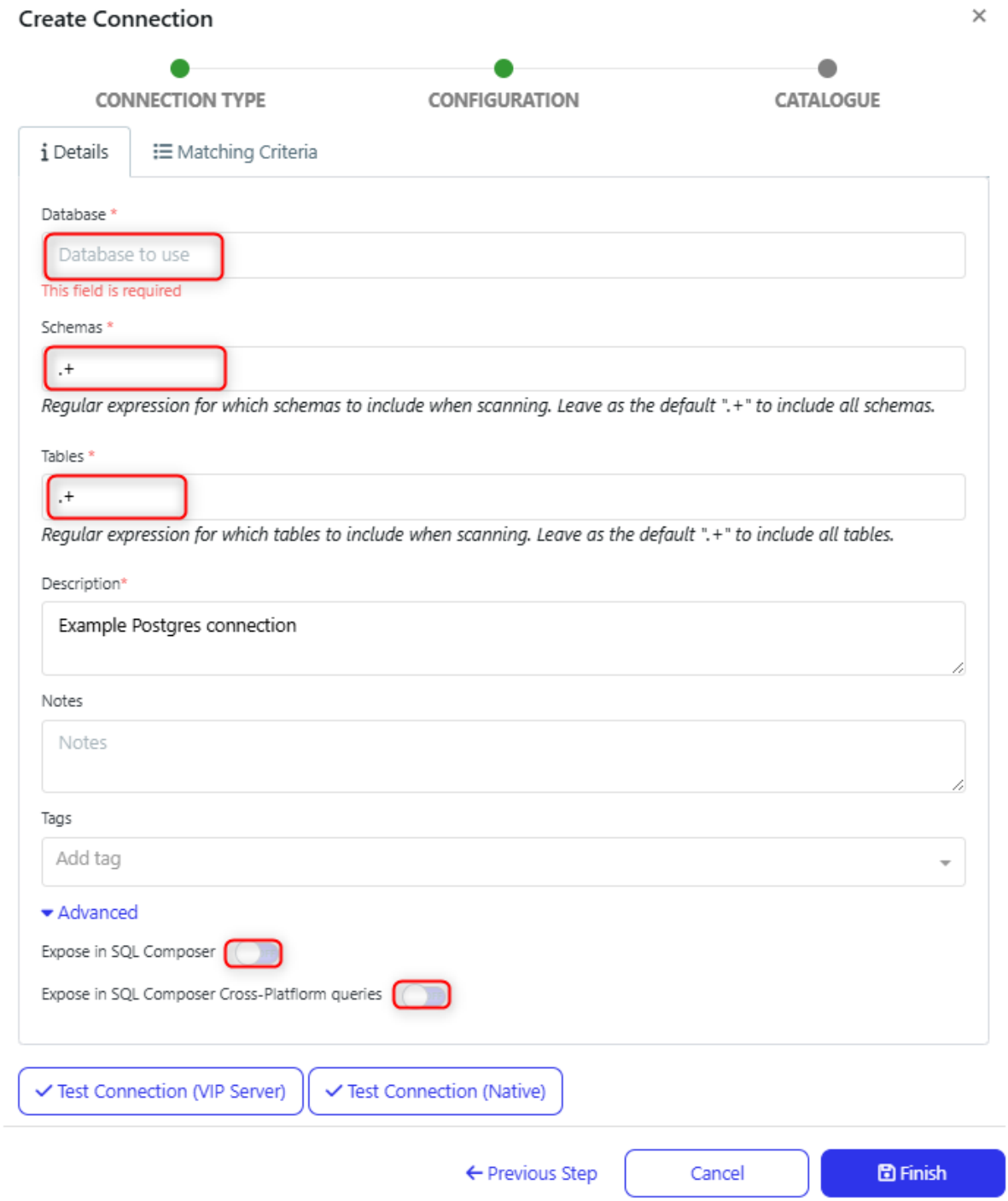


The **Security** tab holds the **Username** & **Password** details to connect to the database with.

- **User –** The Username of the login details
- **Password –** The Password of the user
- **Integrated Security –** Use your organisation's access controls to login rather than a username/password combination



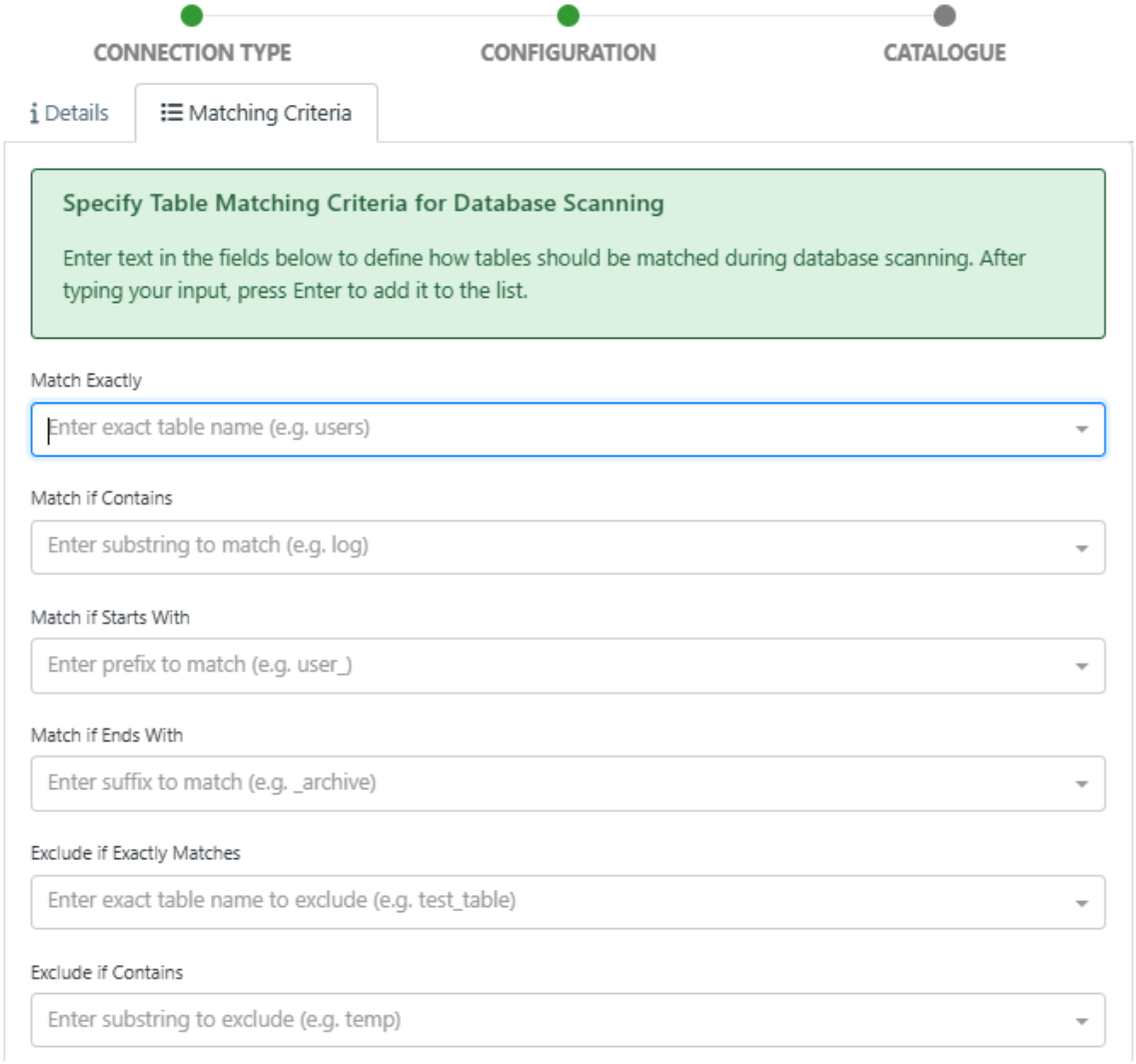
5. **Connect to the database**



1. **Database** – Enter the database name
2. **Schemas** – You can apply regular expression here to choose which schemas to include
3. **Tables** – You can apply regular expression here to choose which tables to include
4. **Expose in SQL Composer** – This exposes the connection in Query Composer
5. **Expose in SQL Composer Cross-Platform queries** – This exposes the connection to be used in the cross-platform queries in Query Composer

The '**Test Connection VIP Server'** and **'Native'** buttons function the same as when we edit the connection and are described in more detail in the below section.

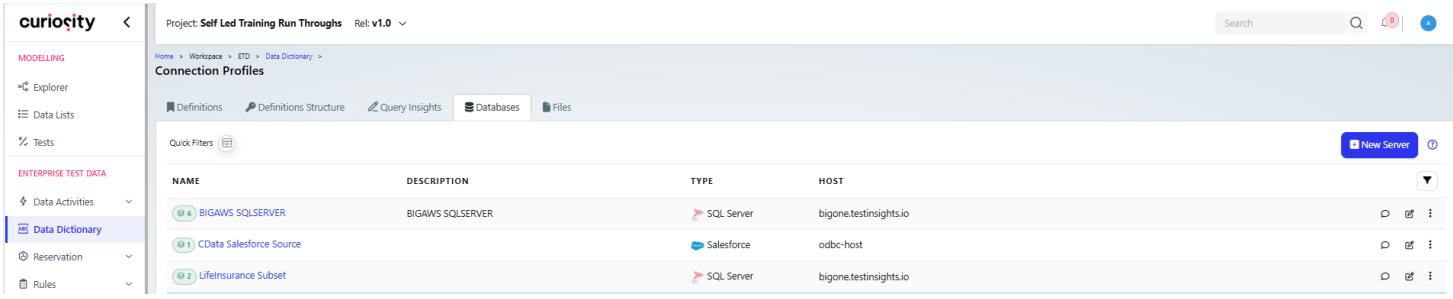
6. Match the criteria



In the 'Matching Criteria' tab, we can filter the names of tables that will be brought into the data dictionary using a variety of filtering criteria as shown above.

When finished click '**OK**' to continue and create the connection.

Within the Data Catalogue you will now have a database server connection ready to use.



Exercise 1

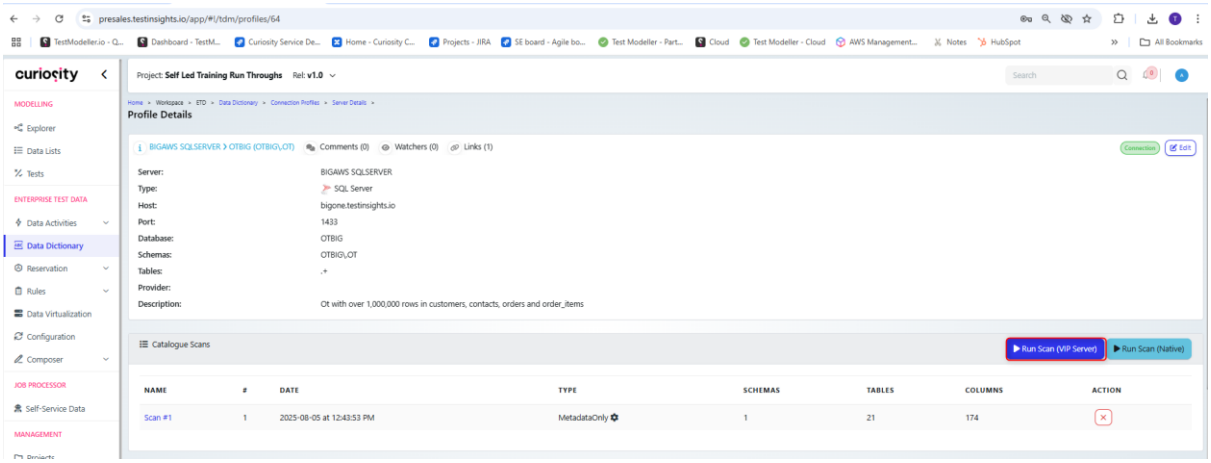
1. Set up a database server connection. If a connection already exists – you only need one server connection to the database you want to connect to for your organisation – then you can select this connection and proceed to step 2.

2. Set up a database connection

Section 2 - Scan a database

This process will scan and store a version of the database metadata within the Curiosity Platform's catalogue. It will collect statistical properties of the data, data types and much more, all of which can be used in your data activities and test data pipelines.

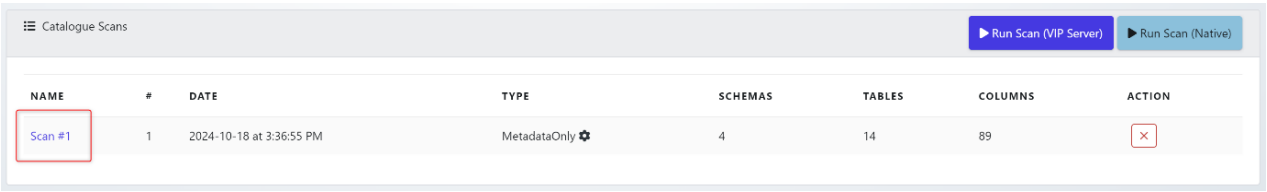
1. To scan the database, navigate to the **Data Dictionary → Databases** and click on the newly setup database connection.
2. Click on **Run Scan (VIP Server)**



This will open up a form asking you to select a process, select the **'Get Schema Metadata'** option. After this you can also choose whether to scan tables and views.

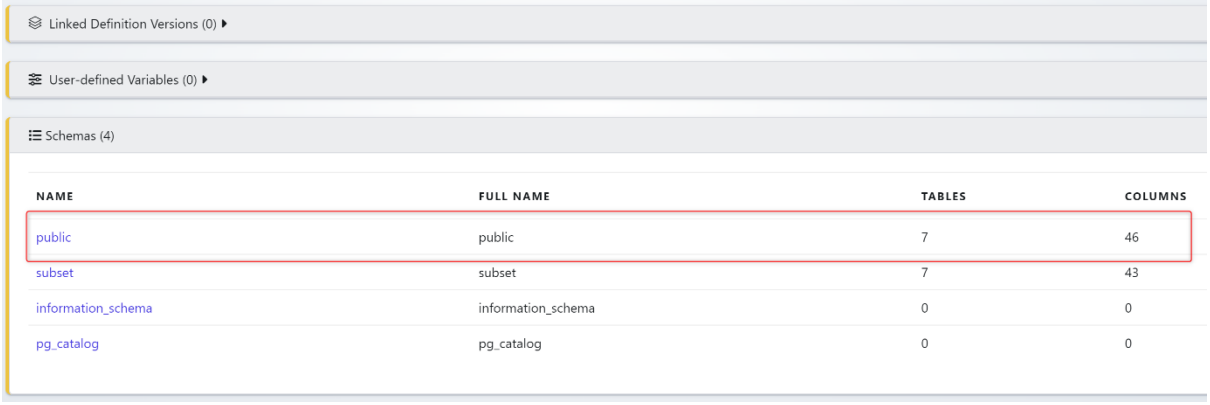
When this job completes, you will have a scanned database to review. It will show schemas, tables and columns.

3. To view the scan details, click on the '**Scan #1**'



If the database is updated, you can scan multiple times. You will then have multiple versions of scans.

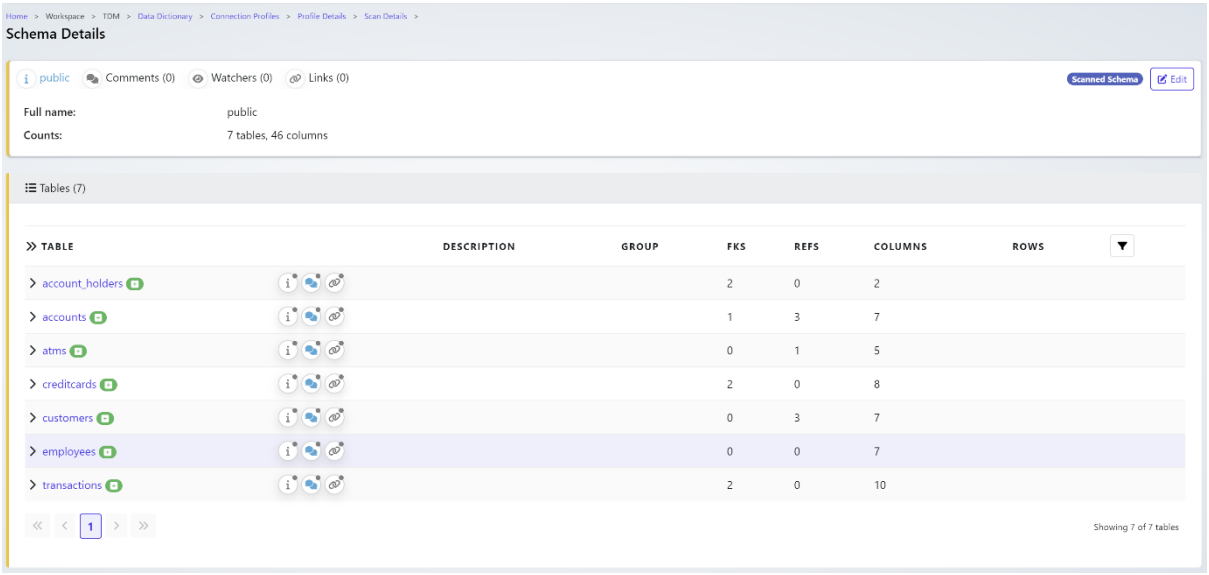
The available schemas and some associated information will be presented.



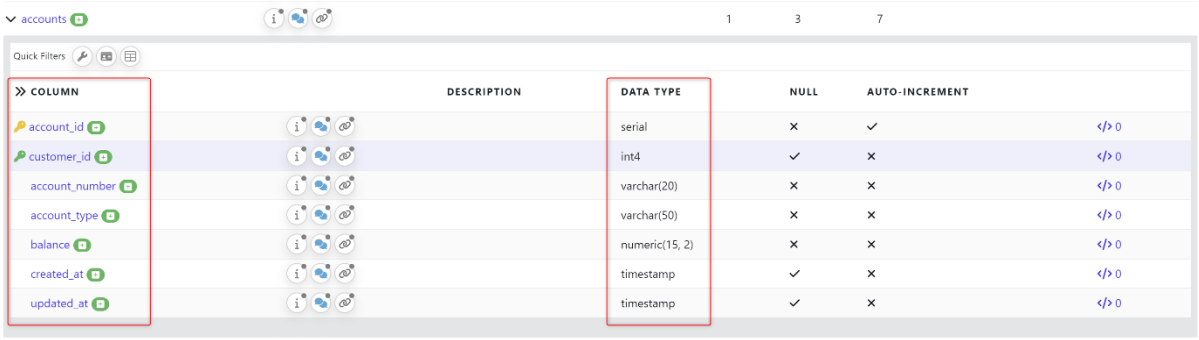
In this case, we'd like to see the **public** schema in more detail.

4. Click on **'public'** to learn more.

The **schema details** with the column details, foreign keys and references are now displayed.



Clicking on any table will show further details on each table.



Column information & data types will often start to drive the decisions made in terms of masking or data generation routines. You can also click on the column to view statistical information about the data.

Exercise 2

a) Kick off a scan on the connection you set up in Exercise 1, using:

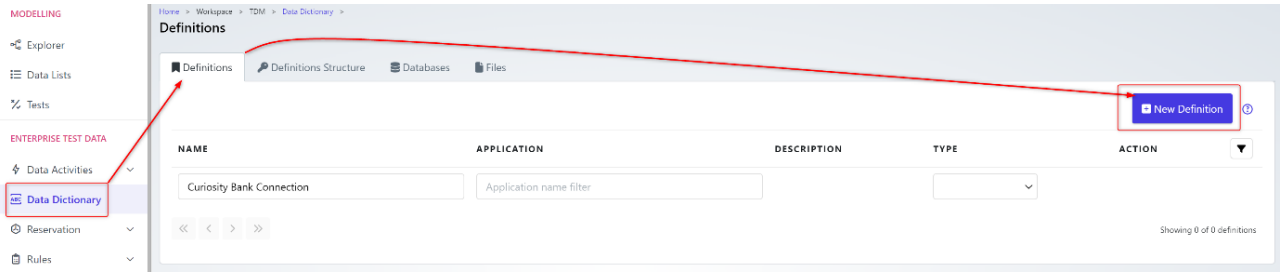
1. Run Scan (VIP Server)
2. Run Scan (Native)

Section 3 - Create a definition

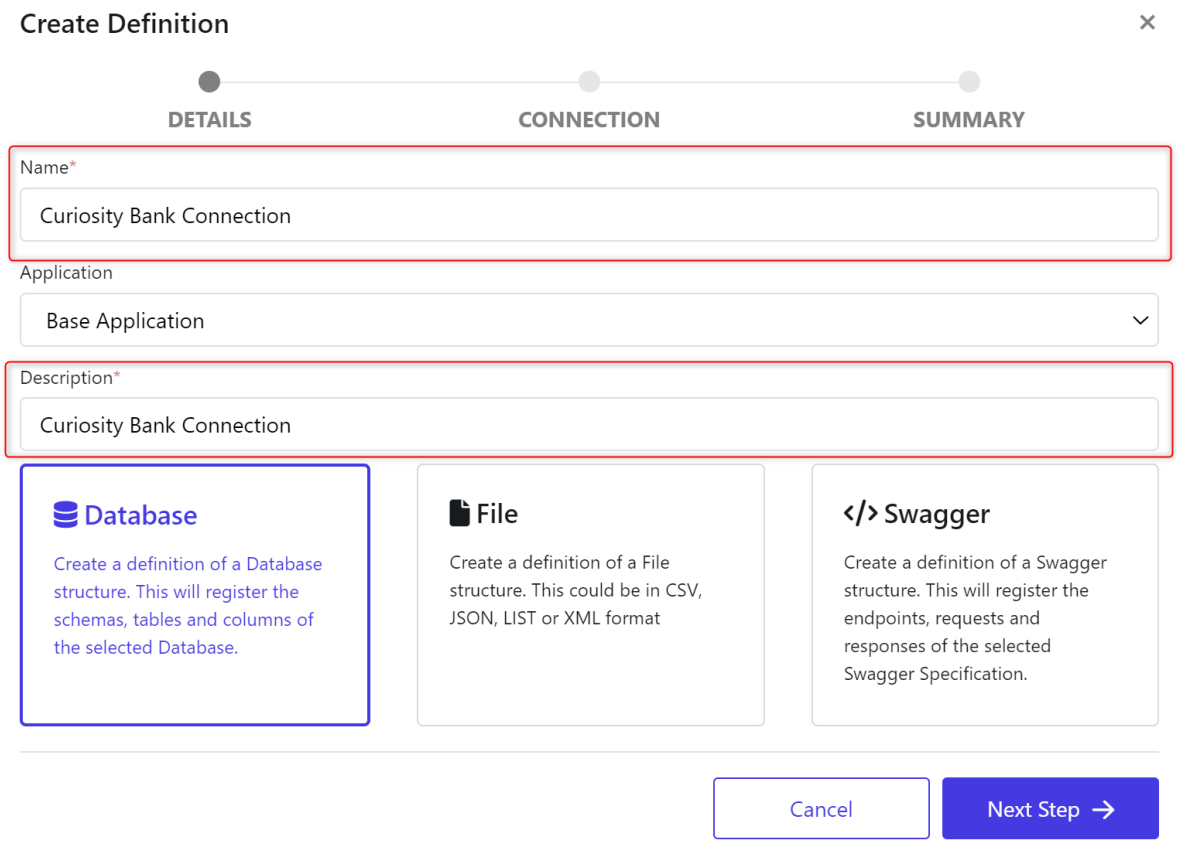
Definitions are used to track and store detailed information about the databases or files we work with. In this section, we'll walk through the process of connecting to a pre-configured database connection. We'll explore how to scan the database, capture its details, and leverage the insights gained to make informed decisions about the next steps in managing and optimising your data.

To create a new definition:

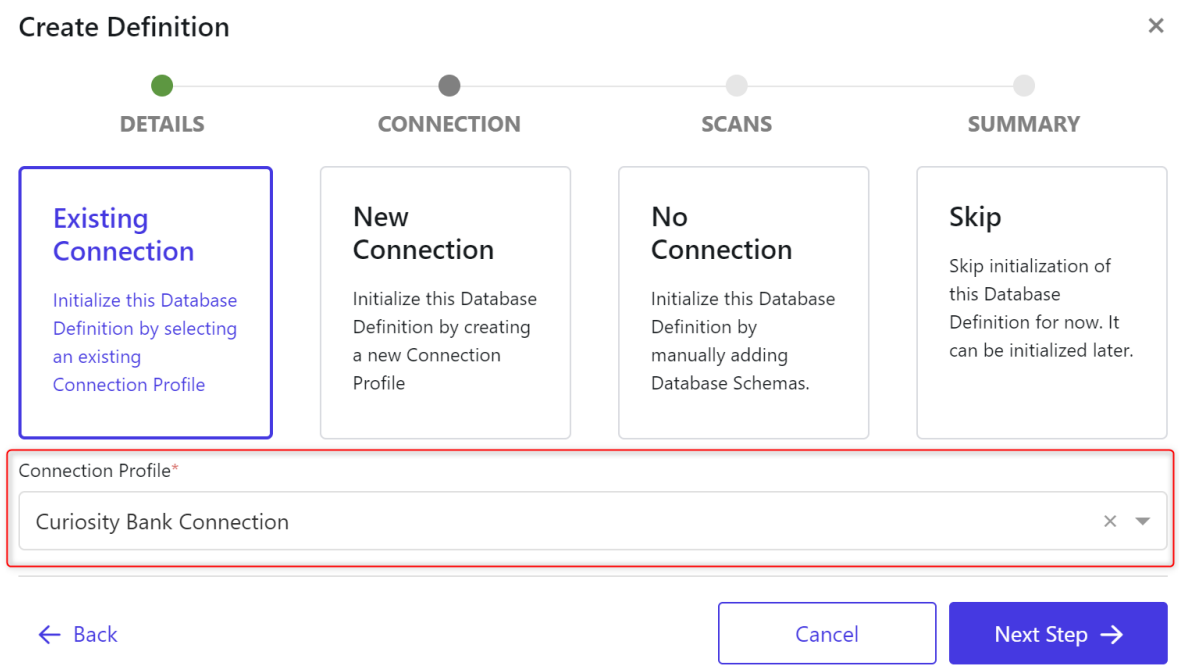
1. Navigate to the Data Dictionary and click **+New Definition**



2. Provide a **Name** & **Description** and choose **Database**. Click **'Next Step'** when ready.

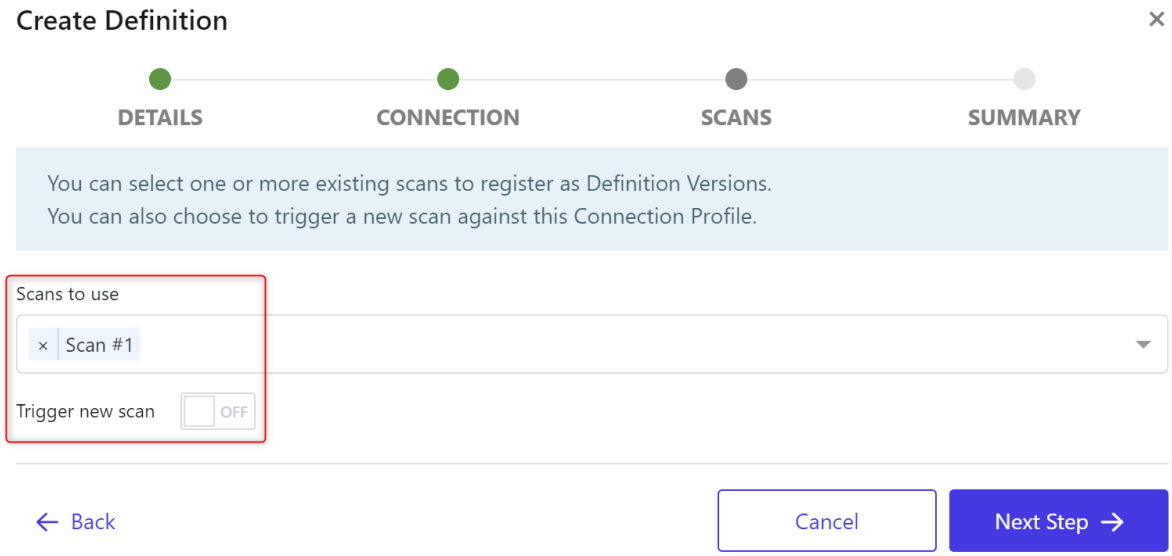


3. Pick '**Existing Connection**' and choose the previously set-up connection profile.

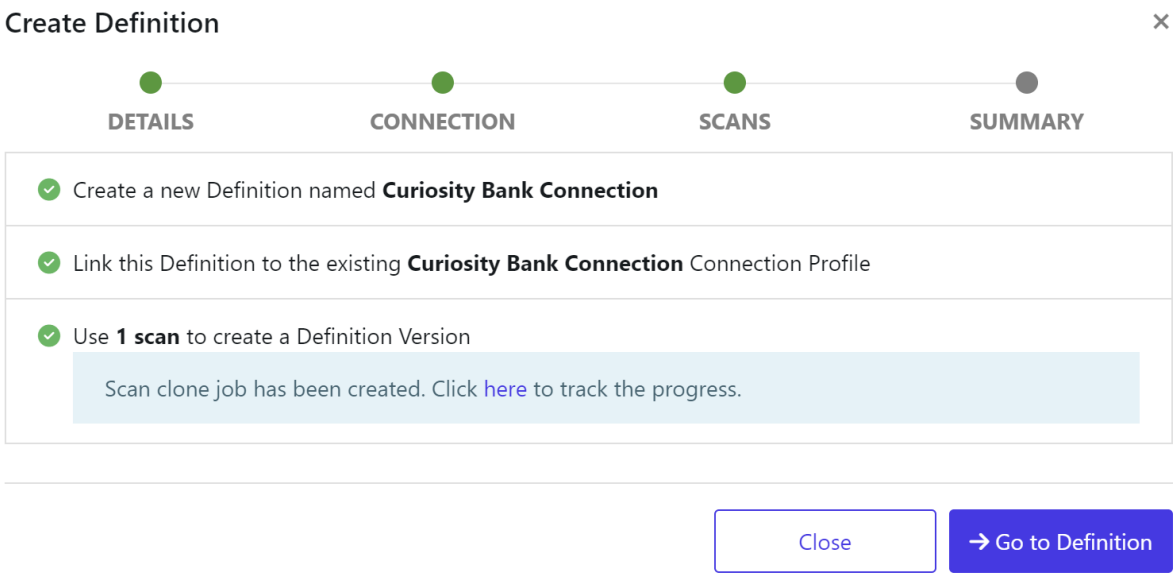


If the database connection has a scan already it will show here. If no scan is available or you are unsure, click the toggle for **'Trigger new scan'**.

4. Click '**Next Step** →' and '**Finish**'

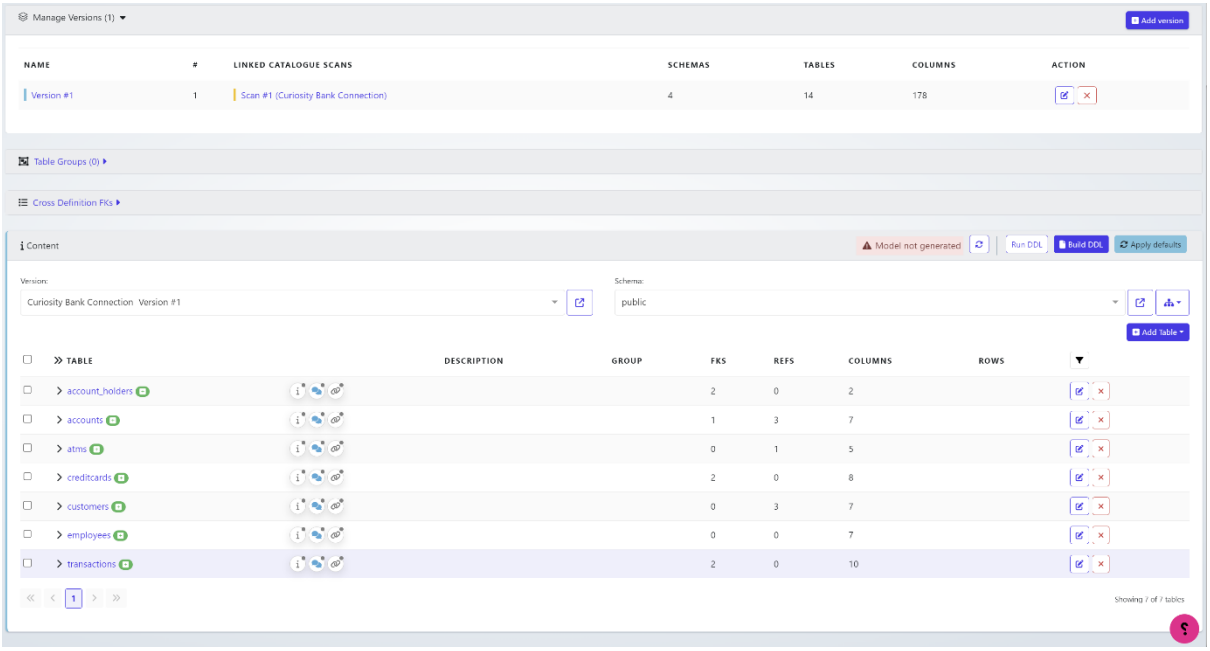


Once completed you will have a new **Definition** linked to a **Connection Profile** with a completed **Scan**.



5. When finished click **'Go to Definition'**.

You will see the current version, table details and additional actions you can now perform against the **Definition**.



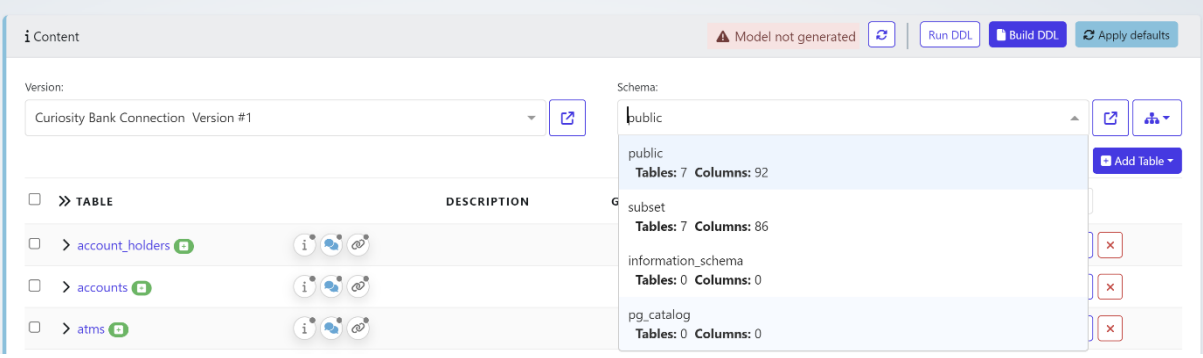
The version and schema within the **'Content'** tab will hold various scanned information. You can choose different versions or schemas from these drop downs.

Exercise 3

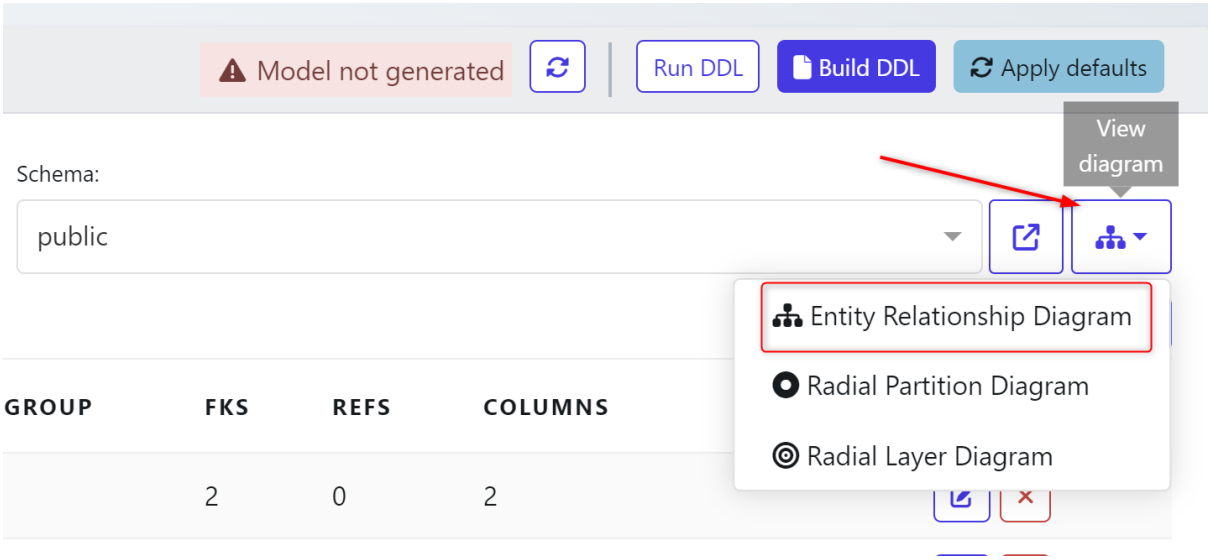
a) Create a new definition based off the scan that you conducted in Exercise 2

Section 4 - Visualise table relationships

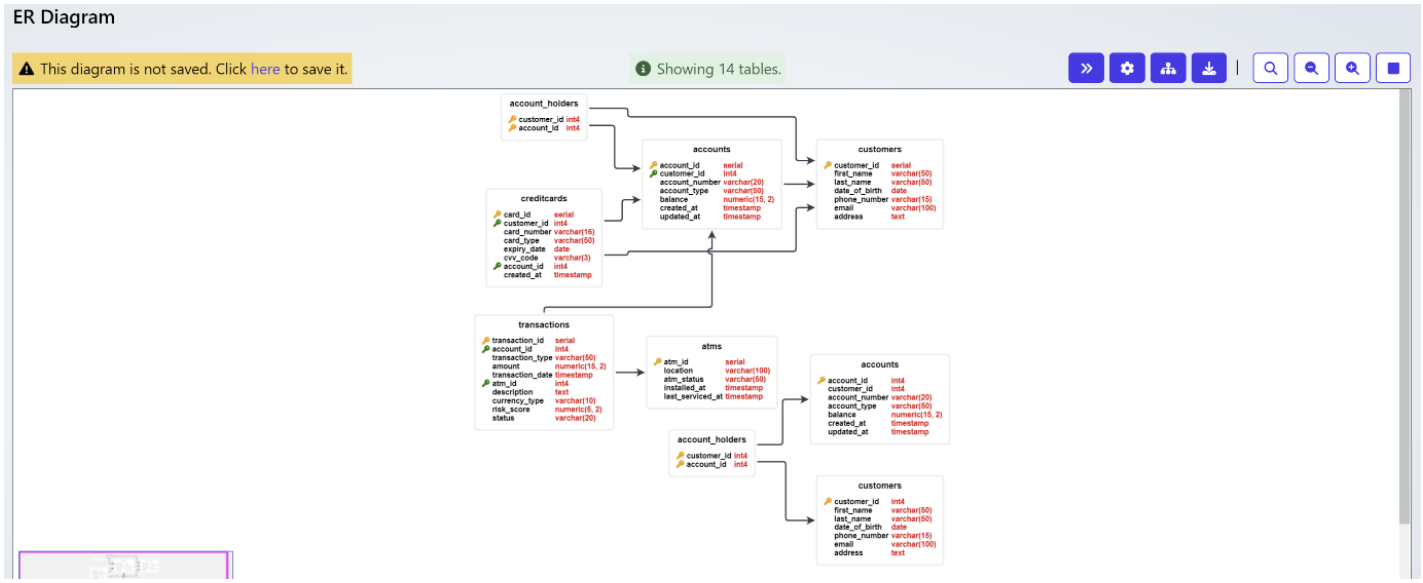
The Curiosity Enterprise Test Data Platform also lets you visualise the table relationships within the definition. This includes foreign key relationships and soft keys uncovered as part of the discovery process.



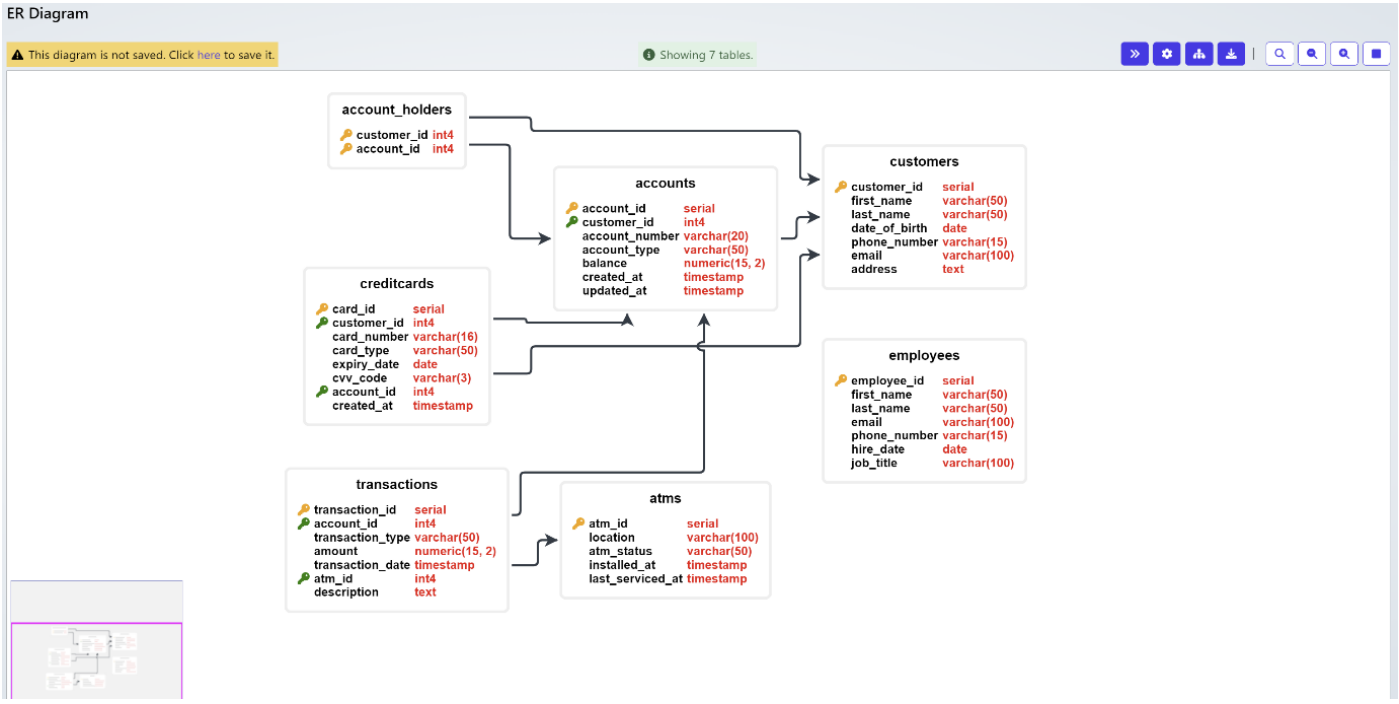
1. Click on the **tree** icon in to view diagram and choose the type of diagram you want to create. For this example, please select **'Entity Relationship Diagram':**



2. The diagram will then be generated.



This will present the existing found **relationships** that need to be considered and be used to maintain referential integrity in our data activities and test data pipelines. Here's a closer look at the diagram:



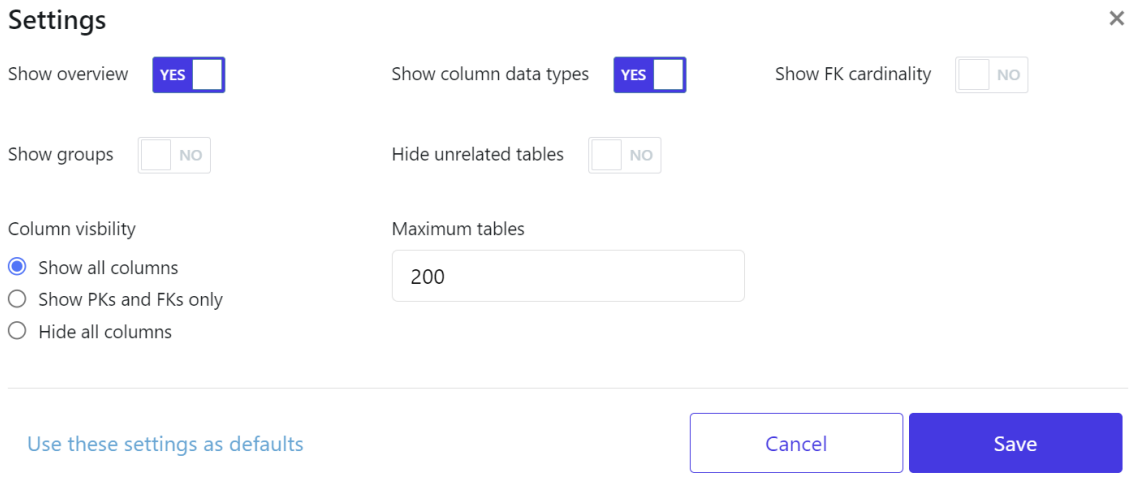
In the top right, additional actions are available:

These include, from left to right:

- Toggle collapse / Expand the image
- Settings
- Relayout
- Download as .png
- Search
- Zoom in / Zoom out
- Fit to screen

Settings

Some databases & files have large structures and table amounts/column amounts. **Settings** allow you to customise what is displayed and how much based on various configurations. The number of tables displayed and the columns you wish to see, should be configured here for an easier viewing experience when looking at large databases.



Exercise 4

1. From the definition you created in Exercise 3, try to create one of each type of the following diagrams:
    a. Entity relationship
    b. Radial Partition
    c. Radial Layer

Further learning:

For further information on the Data Dictionary and Definition structures, visit our Knowledge Base to see the documentation: https://knowledge.curiositysoftware.ie/docs/data-dictionary-definition-structure

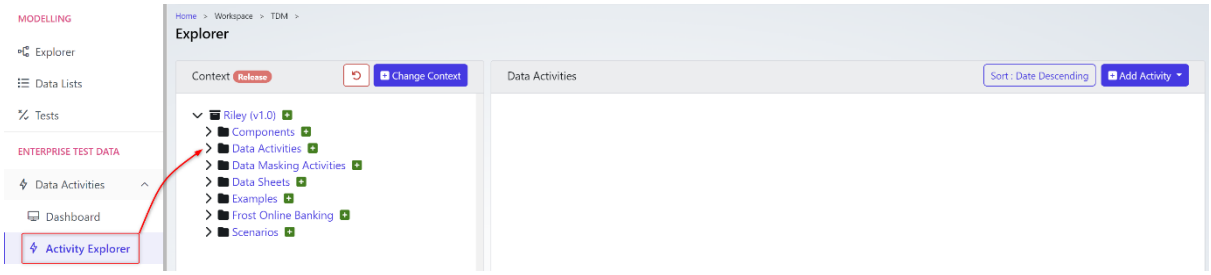
Section 5 – Set up the database scan activity

The Curiosity platform enables deep scanning of your data to help you:

- Detect and classify sensitive information, including Personally Identifiable Information (PII) and Protected Health Information (PHI)
- Gain visibility into where your most critical data is stored and how it's being used, ensuring compliance with regulatory requirements

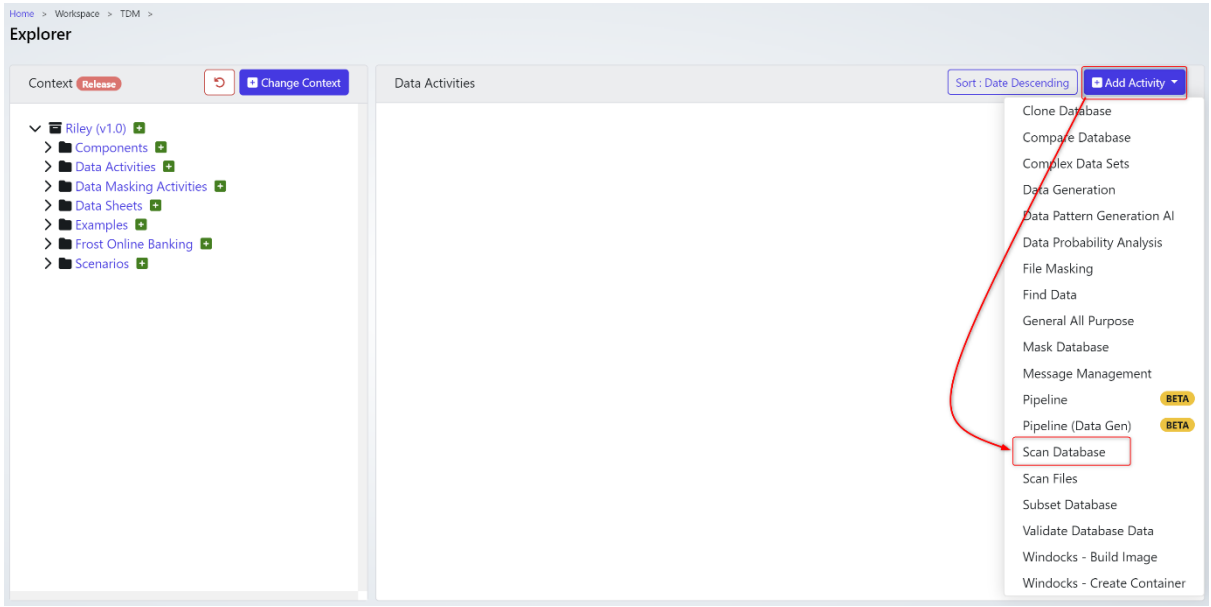
Set up the database scan activity

1. From the side ribbon navigate to the '**Data Activities**' → '**Activity Explorer**' and choose an Explorer folder. This will set the context to save the Data Profiling Activity within.

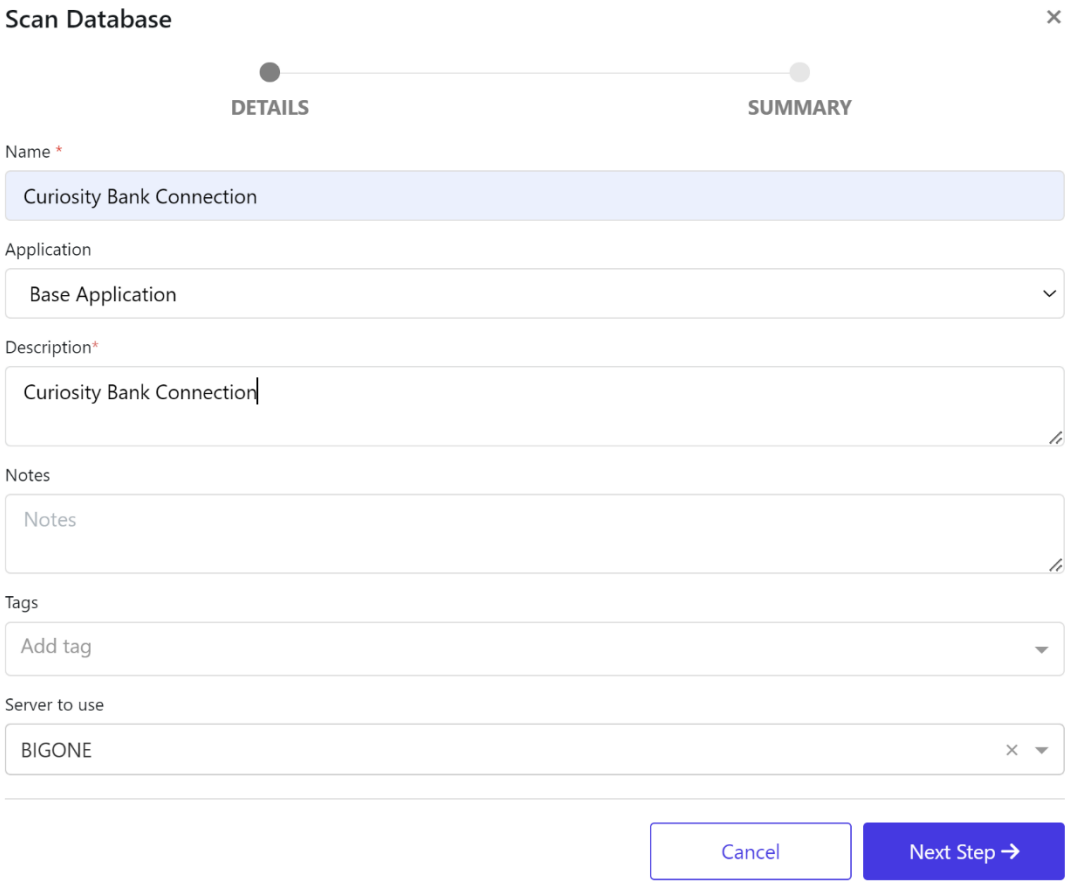


2. Click **+Add Activity** and choose **Scan Database**.

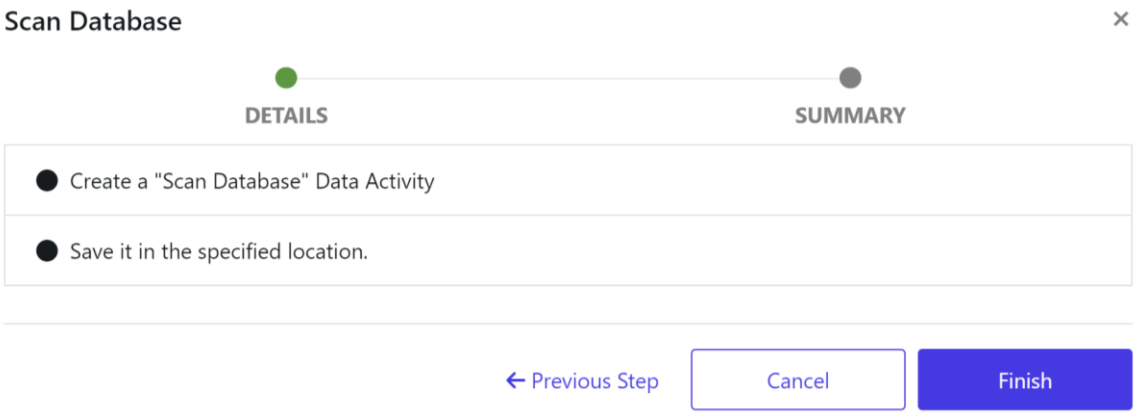
Here, you will build this data activity and configure it to find sensitive data.



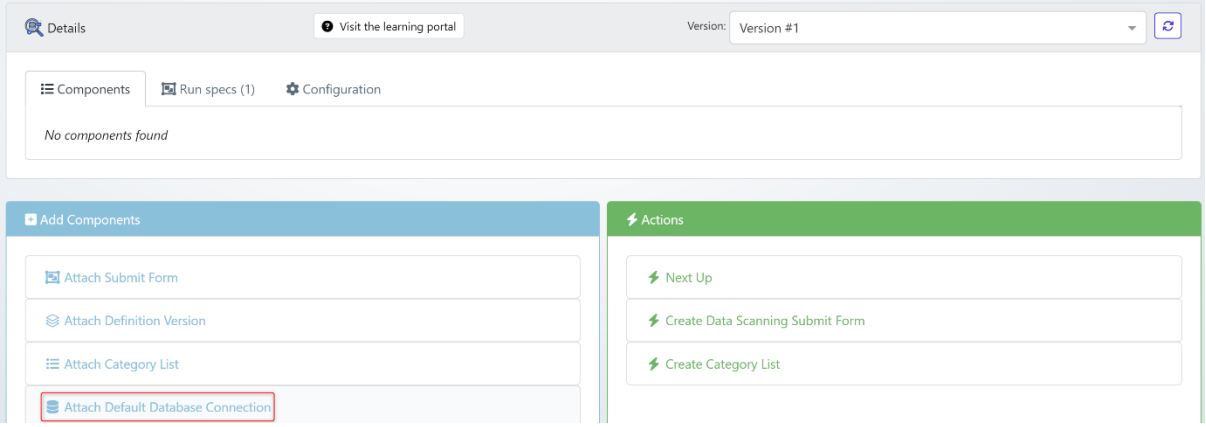
3. Enter a **Name, Description** and **Server**. Click '**Next Step**' when you've done this.



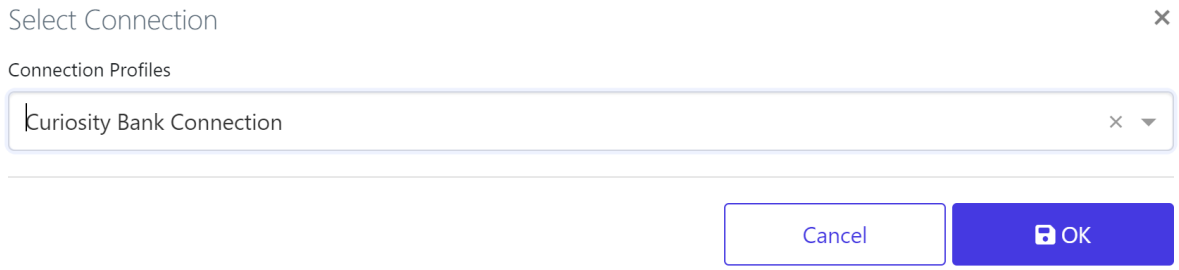
4. When ready, click '**Finish**' and then **'Go to Data Activity'**.



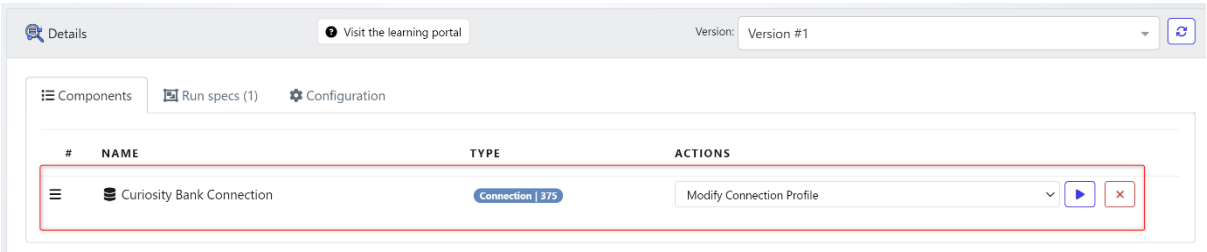
5. First, attach a **Default Database Connection** to the activity. Choose the previously defined connection you set up in Exercise 1.



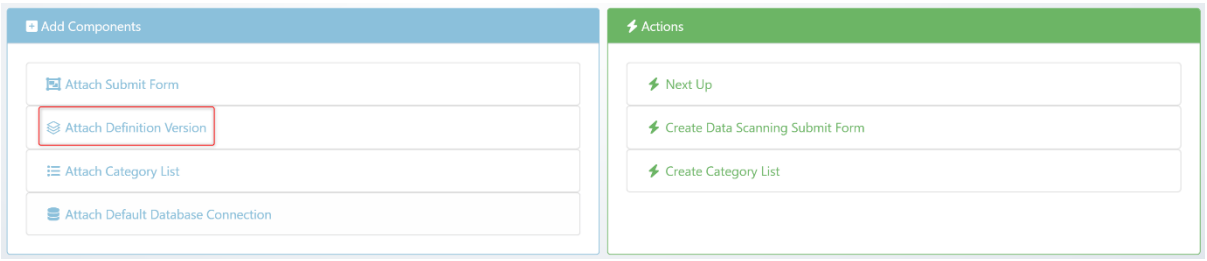
6. Click **'OK'** when you have selected a profile.



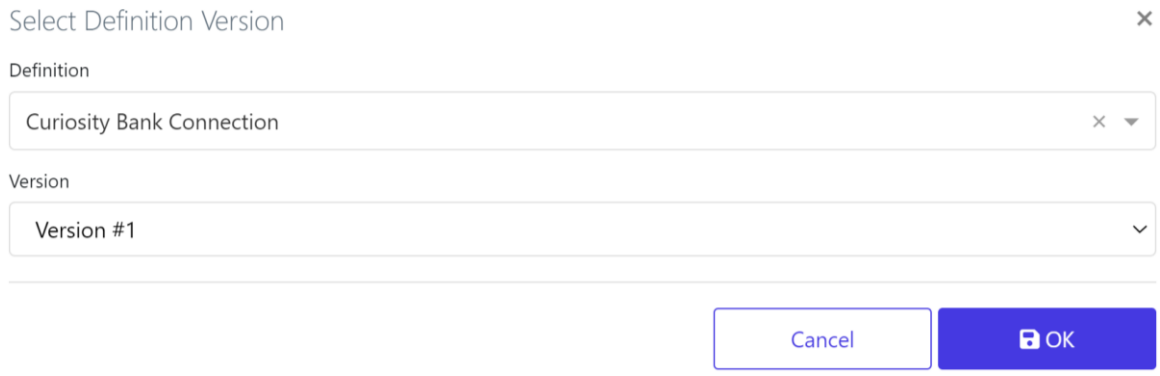
7. A connection will display against the data activity.



8. Next, you need to attach a '**Definition Version**'



9. Select a definition version to scan. Click **OK** when ready.

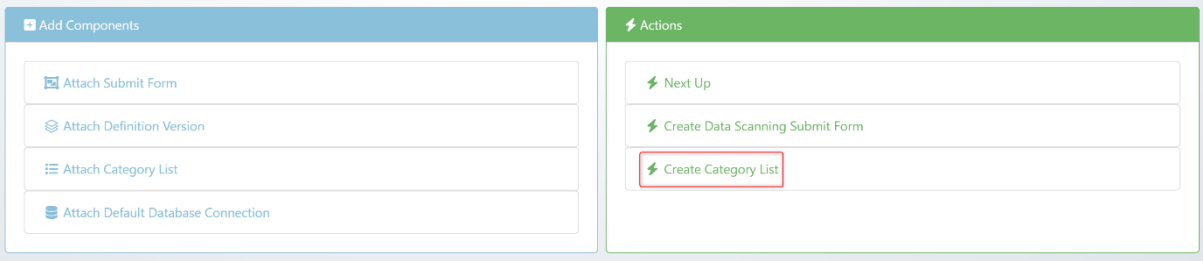


Exercise 5

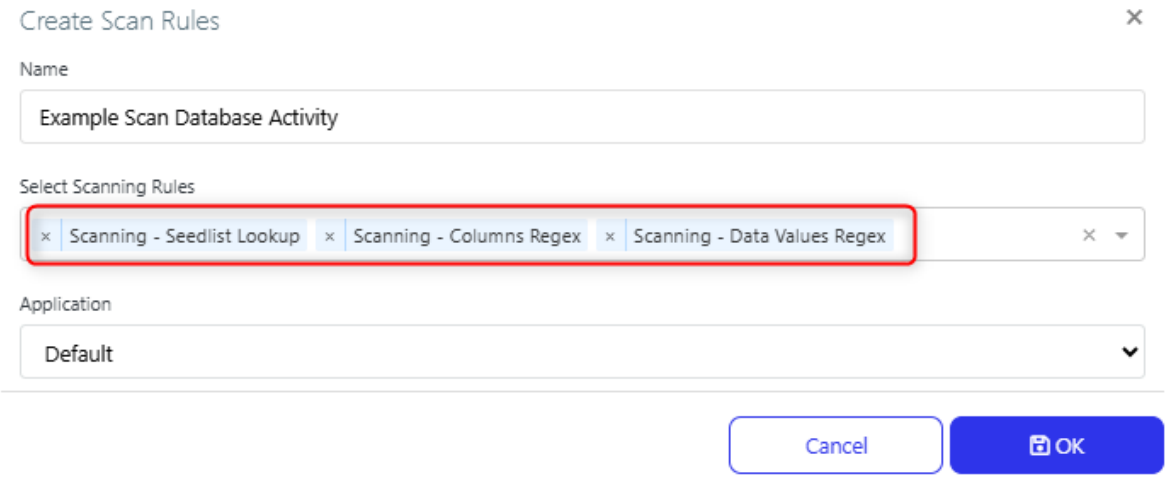
a) Create a data scan activity and then attach a connection and a definition

Section 6 - Create scanning rules

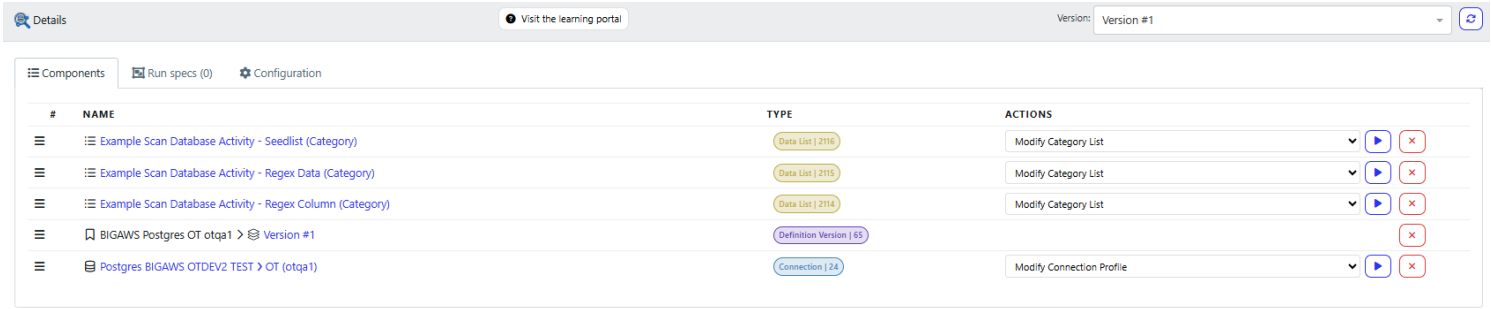
1. You now need to build the scanning rules that the profiling will work with. To do this click on **'Create Category List'.**



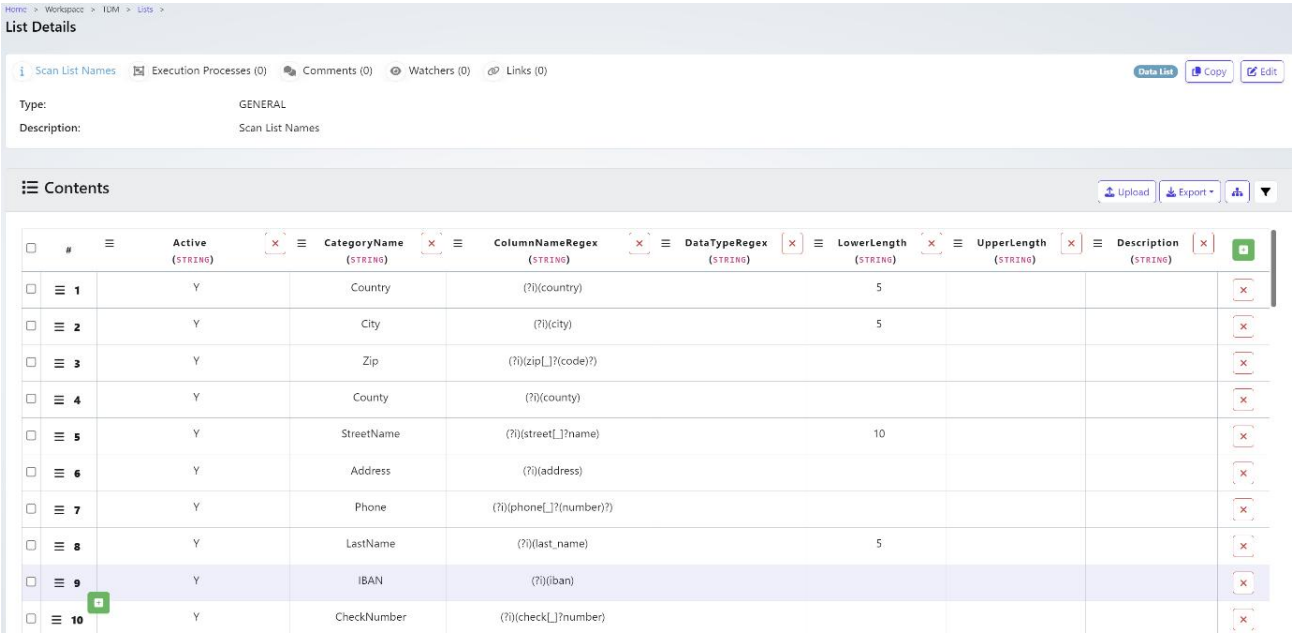
By default, the Curiosity Platform comes with a set of rules that will be used as a starting point. By attaching these to the activity they are exposed, allowing customisation if desired.



2. Click **'OK'** when you're ready, and the data lists will be created against the activity. Here's an example:



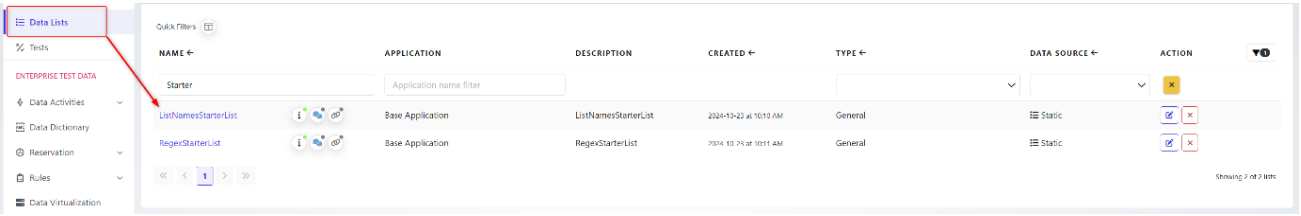
If you click on one of the lists you can view the categories of data we will scan for. Below is an example of what you'll see when you click:



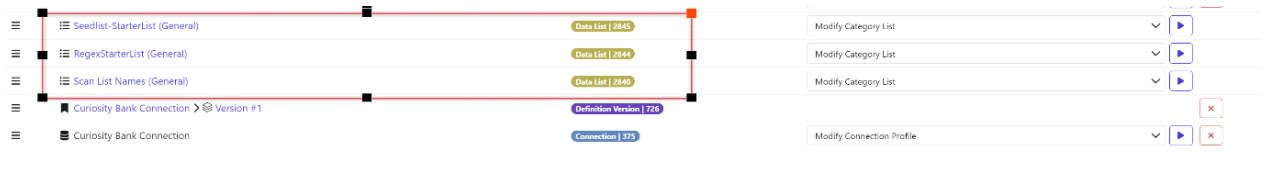
#	Active (STRING)	CategoryName (STRING)	ColumnNameRegex (STRING)	DataTypeRegex (STRING)	LowerLength (STRING)	UpperLength (STRING)	Description (STRING)
1	Y	Country	(?i)(country)		5		
2	Y	City	(?i)(city)		5		
3	Y	Zip	(?i)(zip[_]?(code)?)				
4	Y	County	(?i)(county)				
5	Y	StreetName	(?i)(street[_]?name)		10		
6	Y	Address	(?i)(address)				
7	Y	Phone	(?i)(phone[_]?(number)?)				
8	Y	LastName	(?i)(last_name)		5		
9	Y	IBAN	(?i)(iban)				
10	Y	CheckNumber	(?i)(check[_]?number)				

Each list will provide a different type of search type, from RegEx patterns analysis to specific data.

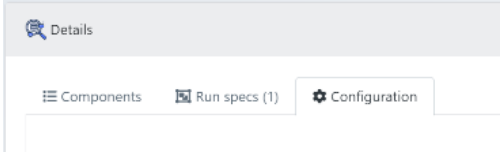
'Data Lists' are searchable and editable from the left-hand menu.



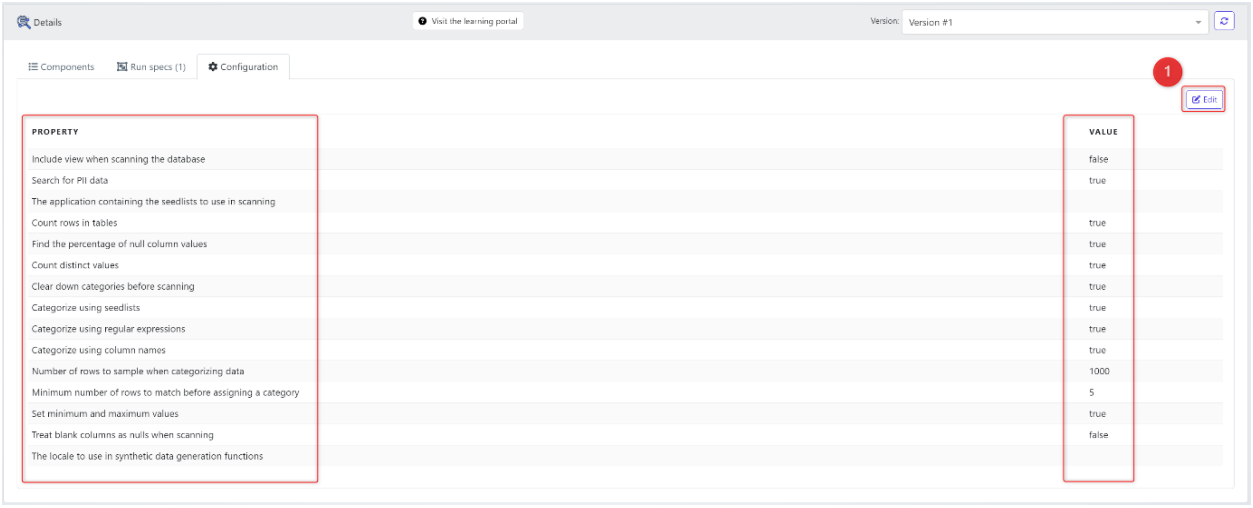
3. Click on one of the lists to view and alter the types of records that are being searched. Below you can see the regular expression being used:



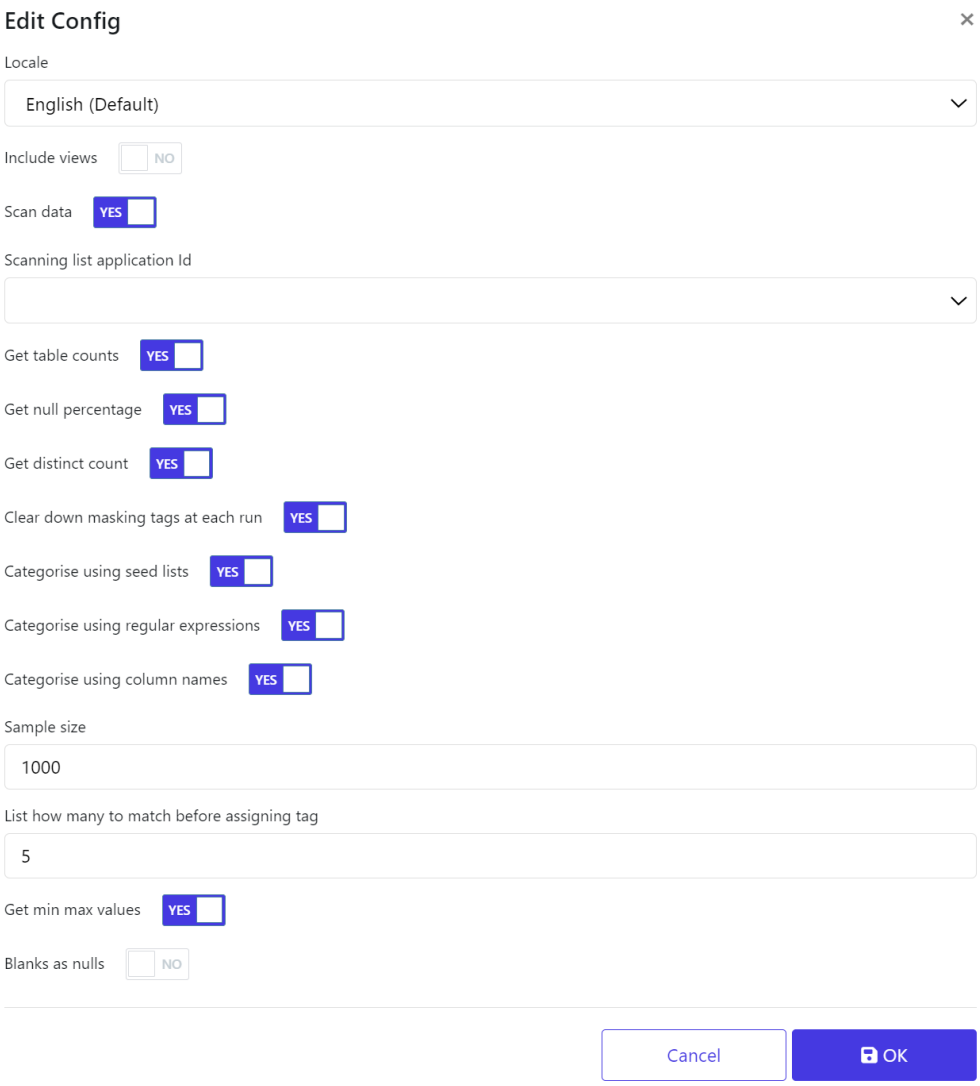
4. Now it's time to customise the property values. To do this, first click on the **'Configuration'** tab.



5. The **'Property'** column will show some of the things you can customise. For example: including views to be scanned, counting rows in tables or finding distinct values. The default parameters are optimally configured, but alter them as your requirements need. Click **'Edit'.**

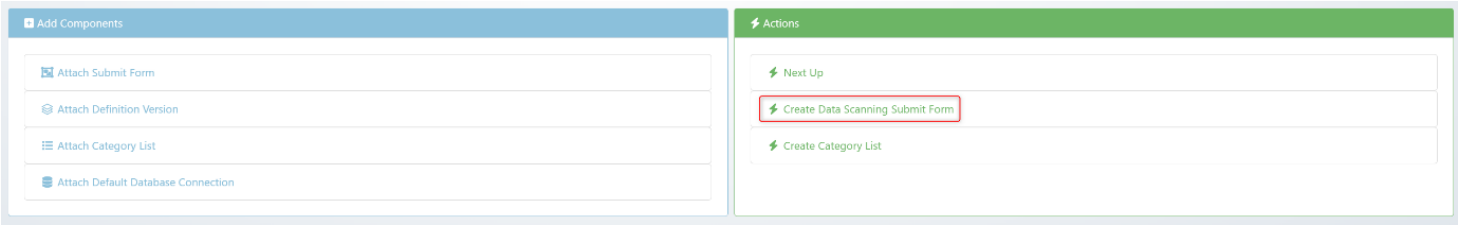


You can also toggle values on and off. Click '**OK**' when finished.

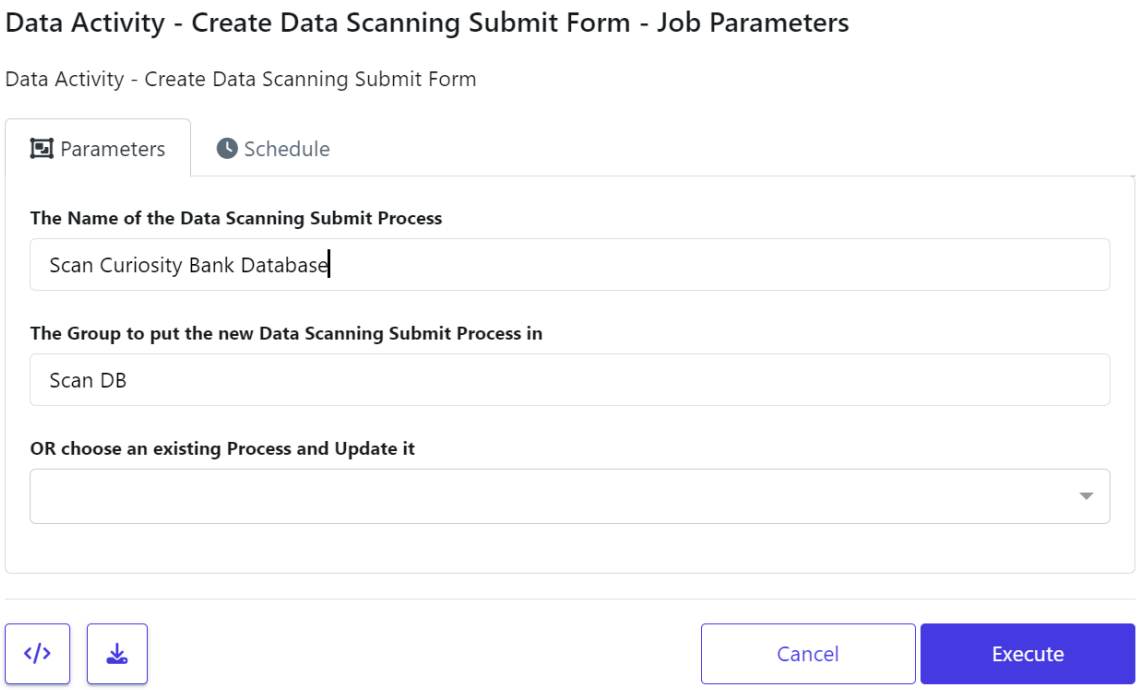


6. From the Data Activity you will now create a Data Scanning Submit Form, which will let you run the job.

   Click on the '**Data Scanning Submit Form'** action.



7. The form requires a **Name** & **Group**



8. The group can be an existing group from the '**Self-service Data'** page or a new group.

   If you are updating an existing process, pick it from the bottom drop down list.

   Click **'Execute'** when ready.

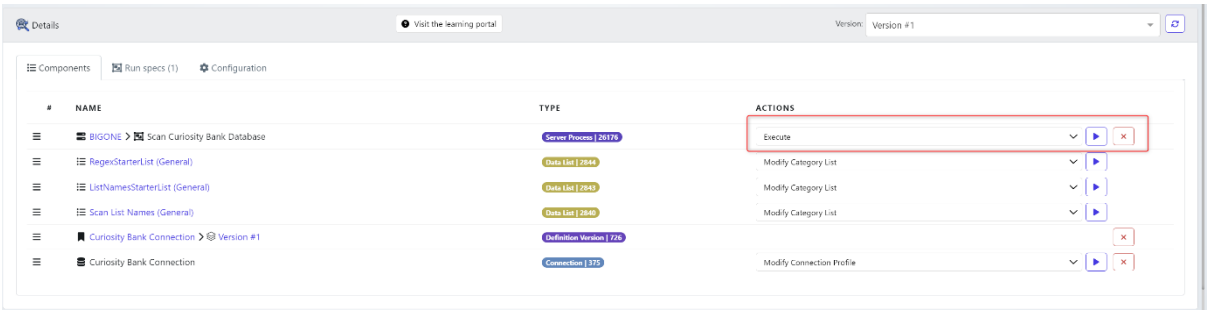
Exercise 6

a) Create the starter lists
b) For the Regex Data list try adding an additional rule
c) For the Regex Column list try adding an additional rule

Section 7 - Run the data profiling activity

You are now configured to run the job. This can be done manually, or scheduled as part of a routine or via an API.

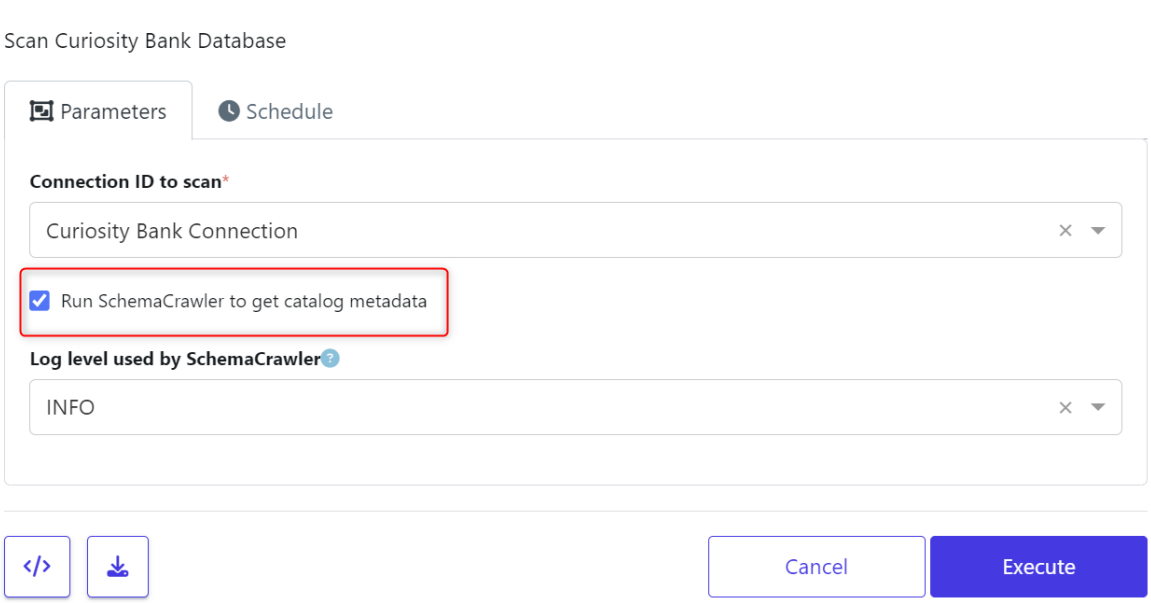
1. Click the **'Play'** icon to run the routine.



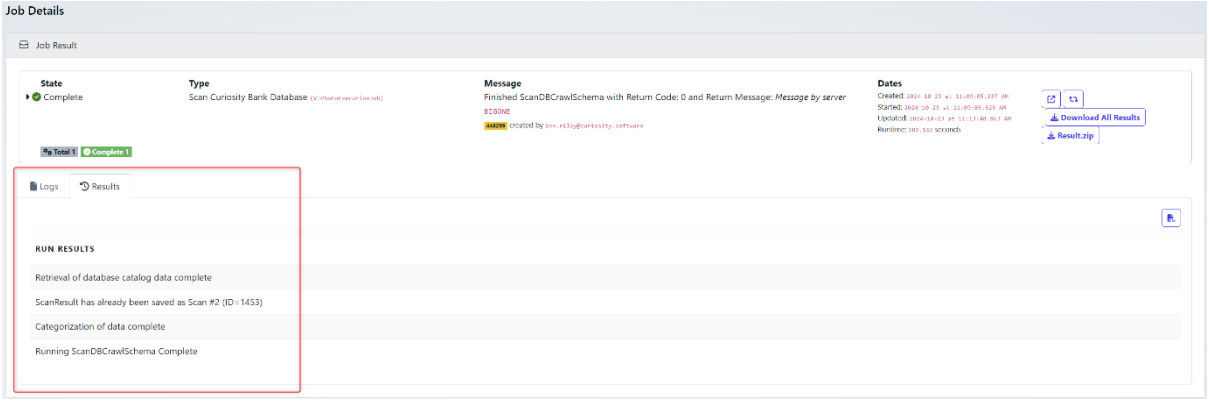
2. When ready, you can run this job.

   Below, we've selected the option to run the schema crawler to gather further metadata.

   Click **'Execute'** if you're happy the **'Connection ID to scan'** is correct.

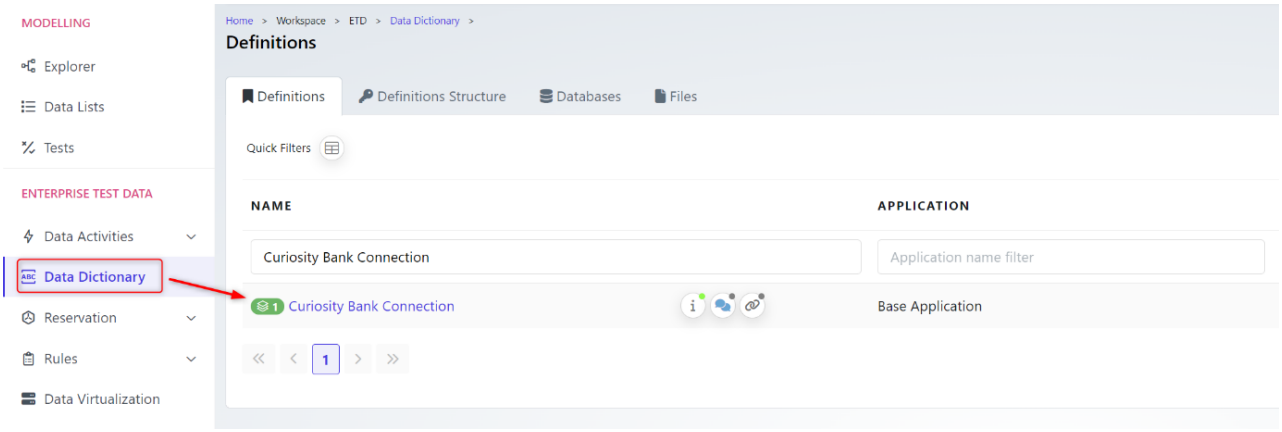


3. Follow the **Job** to check it's finished and the results have been collected.



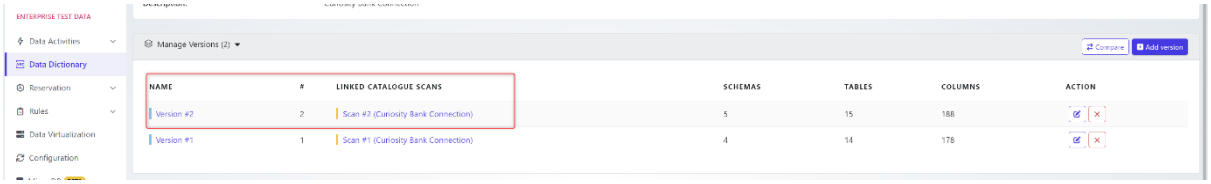
4. Review the results

When the job completes a new scan will be available against the '**Data Definition'**. Navigate to the Data Definition and observe the new tags that have been added as part of the scan

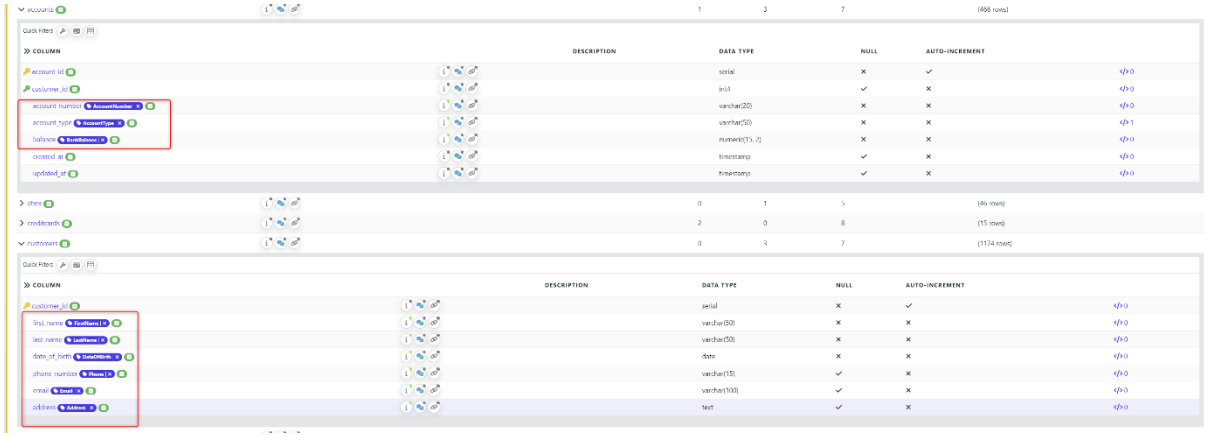


5. The **Scan #2** version will now hold the scanned PII formation. Click on **Scan #2.**

The scan will hold the information for PII we are looking for. The Version #2 will hold the schema details.



The scanned tables and columns will now have the Tags assigned to them based on the information we picked up during the scan.



Exercise 7

a) Create a Data Scanning Submit Form
b) Execute the scan and check the results for the definition

Check the solution videos for all Exercises in this course >

