

## Data profiling and discovery with Curiosity Software

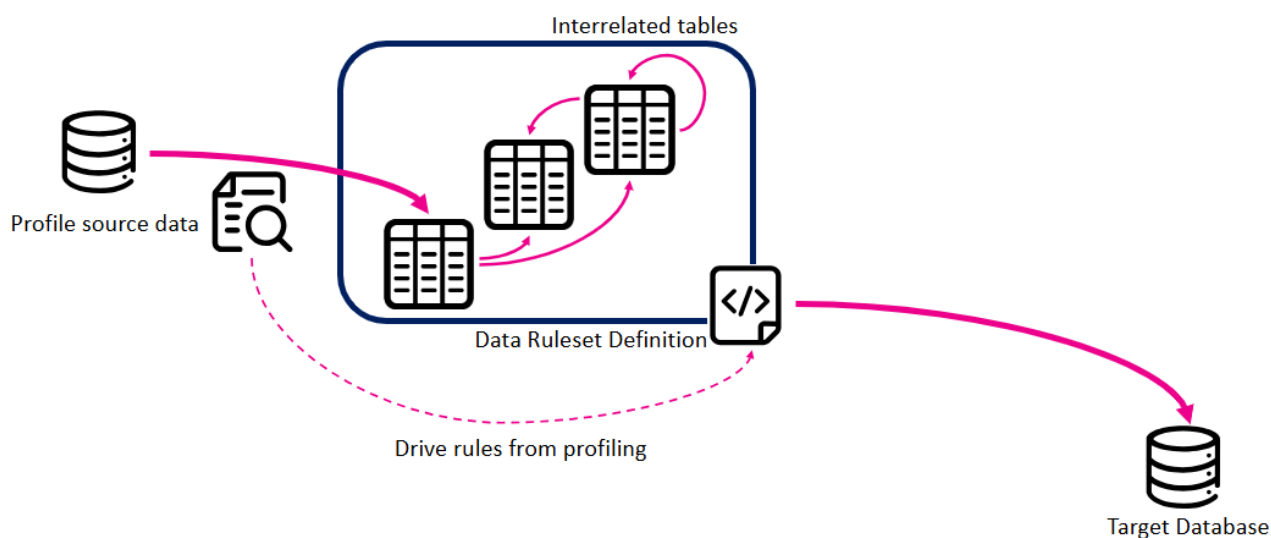
### Introduction:

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

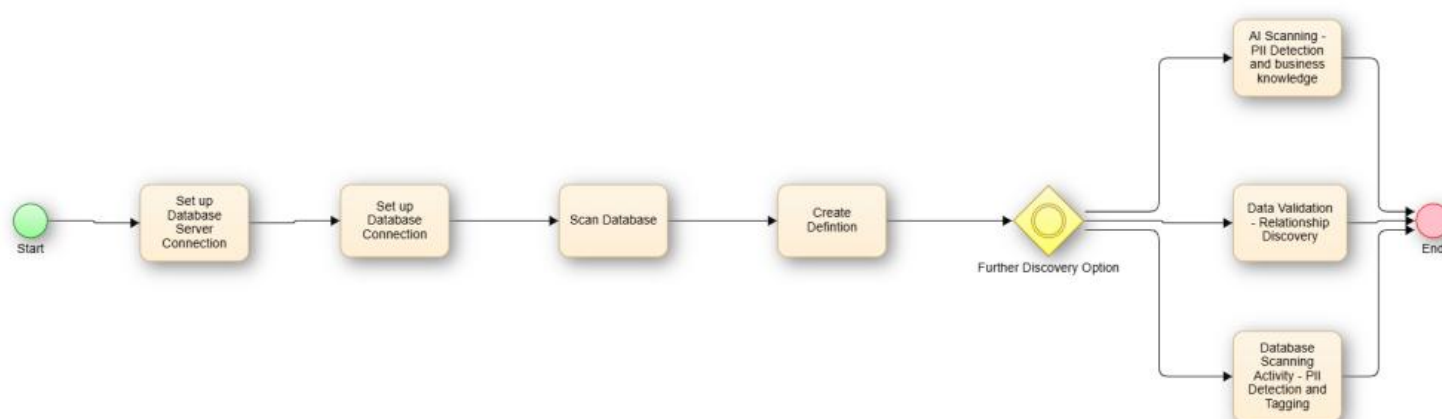
### Key capabilities:

- **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

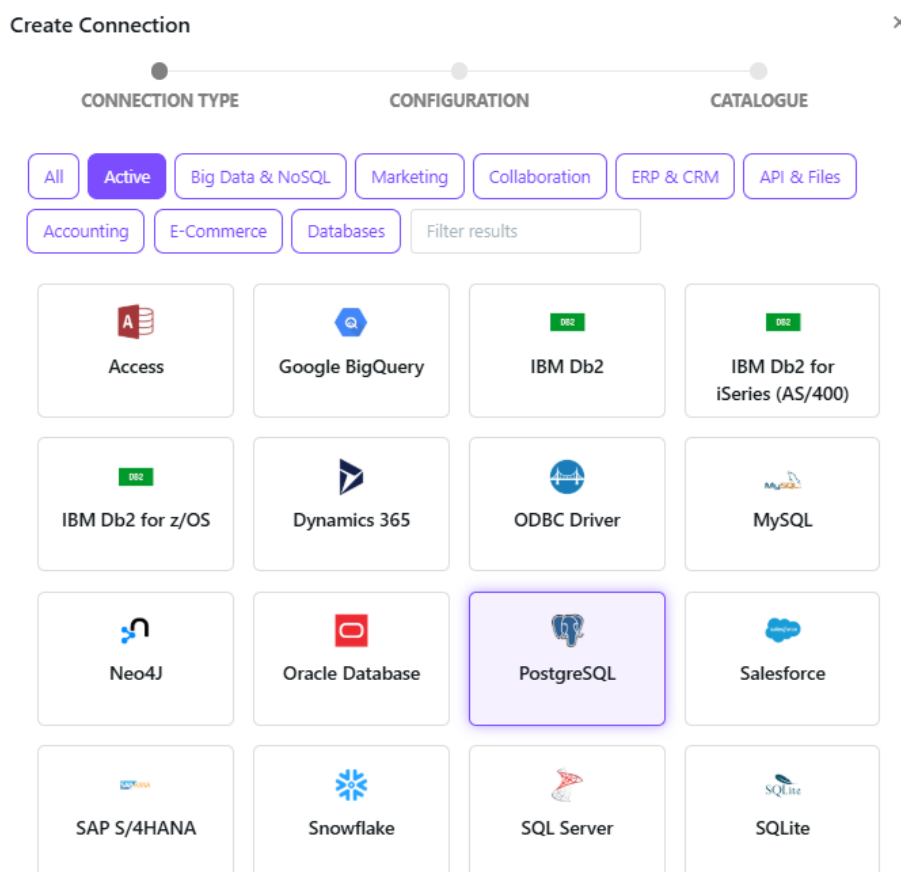
1. To create a new database connection, navigate to the **Data Dictionary** and onto the **Databases** tab.

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.

Create Connection

CONNECTION TYPE CONFIGURATION CATALOGUE

Details Connection Security

Name \*  
Example Postgres connection

Description \*  
Example Postgres connection

Notes  
Notes

Tags  
Add tag

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

Create Connection

CONNECTION TYPE CONFIGURATION CATALOGUE

Details Connection Security

DBMS type  
PostgreSQL

Host \*  
Server host name

Port  
Server port - leave empty for default

Advanced

Provider  
.NET Provider

Linked Server  
EC2AMAZ-0L04NIQ

Advanced Driver Properties

← Previous Step Cancel Next Step →

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

### Create Connection ×

●

●

●

CONNECTION TYPE      CONFIGURATION      CATALOGUE

i Details

⚡ Connection

🔒 Security

Integrated security ☐

User

Username to use

Password

Password to use

← Previous Step

Cancel

Next Step →

## 5. Connect to the database

### Create Connection

CONNECTION TYPE

CONFIGURATION

CATALOGUE

Details

Matching Criteria

Database \*

Database to use

This field is required

Schemas \*

.+

Regular expression for which schemas to include when scanning. Leave as the default ".+" to include all schemas.

Tables \*

.+

Regular expression for which tables to include when scanning. Leave as the default ".+" to include all tables.

Description\*

Example Postgres connection

Notes

Notes

Tags

Add tag

▼ Advanced

Expose in SQL Composer ☐

Expose in SQL Composer Cross-Platform queries ☐

✓ Test Connection (VIP Server)

✓ Test Connection (Native)

← Previous Step

Cancel

Finish

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.

NAME	DESCRIPTION	TYPE	HOST
BIGAWVS SQLSERVER	BIGAWVS SQLSERVER	SQL Server	bigone.testinsights.io
CData Salesforce Source		Salesforce	odbc-host
Lifeline Subset		SQL Server	bigone.testinsights.io

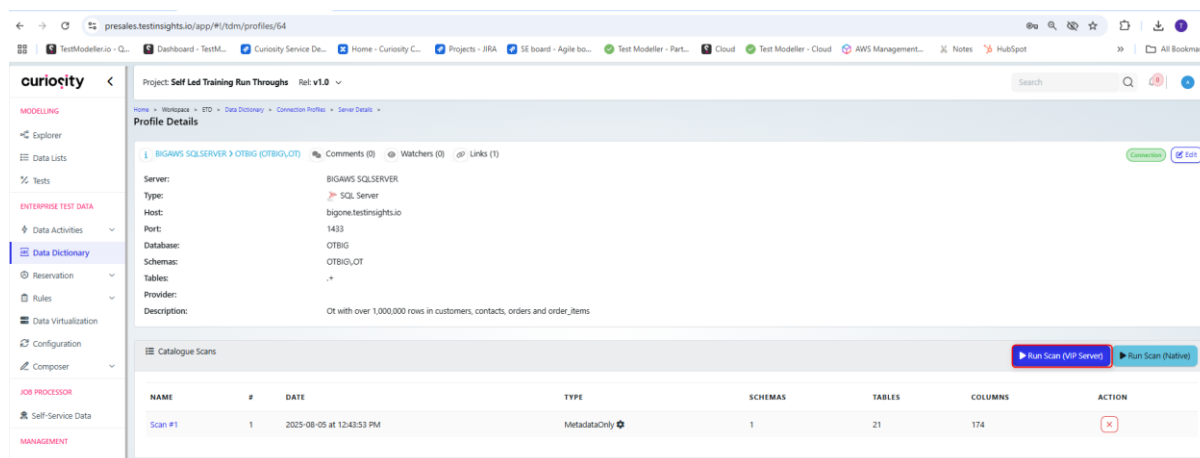
## Exercise 1

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'**

Catalogue Scans							Run Scan (VIP Server)	Run Scan (Native)
NAME	#	DATE	TYPE	SCHEMAS	TABLES	COLUMNS	ACTION	
Scan #1	1	2024-10-18 at 3:36:55 PM	MetadataOnly	4	14	89		

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.

Linked Definition Versions (0)			
User-defined Variables (0)			
Schemas (4)			
NAME	FULL NAME	TABLES	COLUMNS
public	public	7	46
subset	subset	7	43
information_schema	information_schema	0	0
pg_catalog	pg_catalog	0	0



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.

Home > Workspace > TDM > Data Dictionary > Connection Profiles > Profile Details > Scan Details >

### Schema Details

public Comments (0) Watchers (0) Links (0) Scanned Schema Edit

Full name: public  
Counts: 7 tables, 46 columns

TABLE	DESCRIPTION	GROUP	FKS	REFS	COLUMNS	ROWS
> account_holders			2	0	2	
> accounts			1	3	7	
> atms			0	1	5	
> creditcards			2	0	8	
> customers			0	3	7	
> employees			0	0	7	
> transactions			2	0	10	

Showing 7 of 7 tables

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

accounts 1 3 7

Quick Filters

COLUMN	DESCRIPTION	DATA TYPE	NULL	AUTO-INCREMENT
account_id		serial	×	✓
customer_id		int4	✓	×
account_number		varchar(20)	×	×
account_type		varchar(50)	×	×
balance		numeric(15, 2)	×	×
created_at		timestamp	✓	×
updated_at		timestamp	✓	×

**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

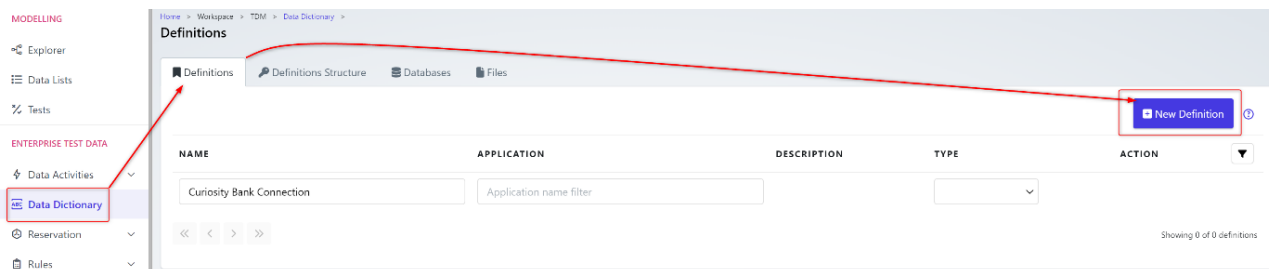
- a) Kick off a scan on the connection you set up in Exercise 1, using:
  1. Run Scan (VIP Server)
  2. Run Scan (Native)

## 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.

### Create Definition

DETAILS

CONNECTION

SUMMARY

Name\*

Curiosity Bank Connection

Application

Base Application

Description\*

Curiosity Bank Connection

**Database**  
Create a definition of a Database structure. This will register the schemas, tables and columns of the selected Database.

**File**  
Create a definition of a File structure. This could be in CSV, JSON, LIST or XML format

**Swagger**  
Create a definition of a Swagger structure. This will register the endpoints, requests and responses of the selected Swagger Specification.

Cancel

Next Step →

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

Create Definition ×

DETAILS CONNECTION SCANS SUMMARY

**Existing Connection**  
Initialize this Database Definition by selecting an existing Connection Profile

**New Connection**  
Initialize this Database Definition by creating a new Connection Profile

**No Connection**  
Initialize this Database Definition by manually adding Database Schemas.

**Skip**  
Skip initialization of this Database Definition for now. It can be initialized later.

Connection Profile\*  
Curiosity Bank Connection × ▼

[← Back](#) [Cancel](#) [Next Step →](#)

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'**.

- Click **'Next Step →'** and **'Finish'**

Create Definition ×

DETAILS CONNECTION SCANS SUMMARY

You can select one or more existing scans to register as Definition Versions.  
You can also choose to trigger a new scan against this Connection Profile.

Scans to use  
× Scan #1 ▼

Trigger new scan ☐ OFF

[← Back](#) [Cancel](#) [Next Step →](#)

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

Create Definition
×

●

DETAILS

●

CONNECTION

●

SCANS

●

SUMMARY

✓

Create a new Definition named **Curiosity Bank Connection**

✓

Link this Definition to the existing **Curiosity Bank Connection** Connection Profile

✓

Use **1 scan** to create a Definition Version

Scan clone job has been created. Click [here](#) to track the progress.

Close

→ Go to Definition

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**.

Manage Versions (1)
Add version

NAME	#	LINKED CATALOGUE SCANS	SCHEMAS	TABLES	COLUMNS	ACTION
Version #1	1	Scan #1 (Curiosity Bank Connection)	4	14	178	<a href="#">✓</a> <a href="#">✗</a>

Table Groups (0)

Cross Definition Fks

Content
Model not generated
Run DDL
Build DDL
Apply defaults

Version: Curiosity Bank Connection Version #1
Schema: public
Add table

TABLE	DESCRIPTION	GROUP	FKS	REFS	COLUMNS	ROWS	
> account_holders			2	0	2		<a href="#">✓</a> <a href="#">✗</a>
> accounts			1	3	7		<a href="#">✓</a> <a href="#">✗</a>
> atlas			0	1	5		<a href="#">✓</a> <a href="#">✗</a>
> creditcards			2	0	8		<a href="#">✓</a> <a href="#">✗</a>
> customers			0	3	7		<a href="#">✓</a> <a href="#">✗</a>
> employees			0	0	7		<a href="#">✓</a> <a href="#">✗</a>
> transactions			2	0	10		<a href="#">✓</a> <a href="#">✗</a>

Showing 7 of 7 tables

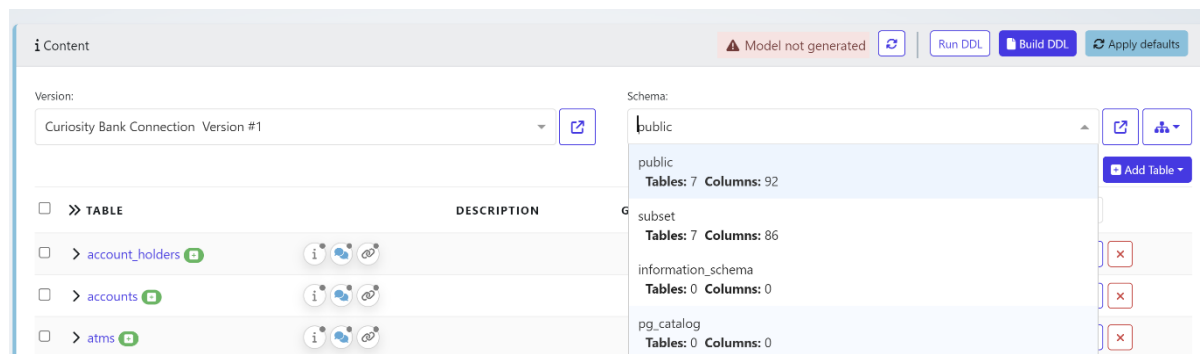
The version and schema within the '**Content**' tab will hold various scanned information. You can choose different versions or schemas from these drop downs.

### Exercise 3

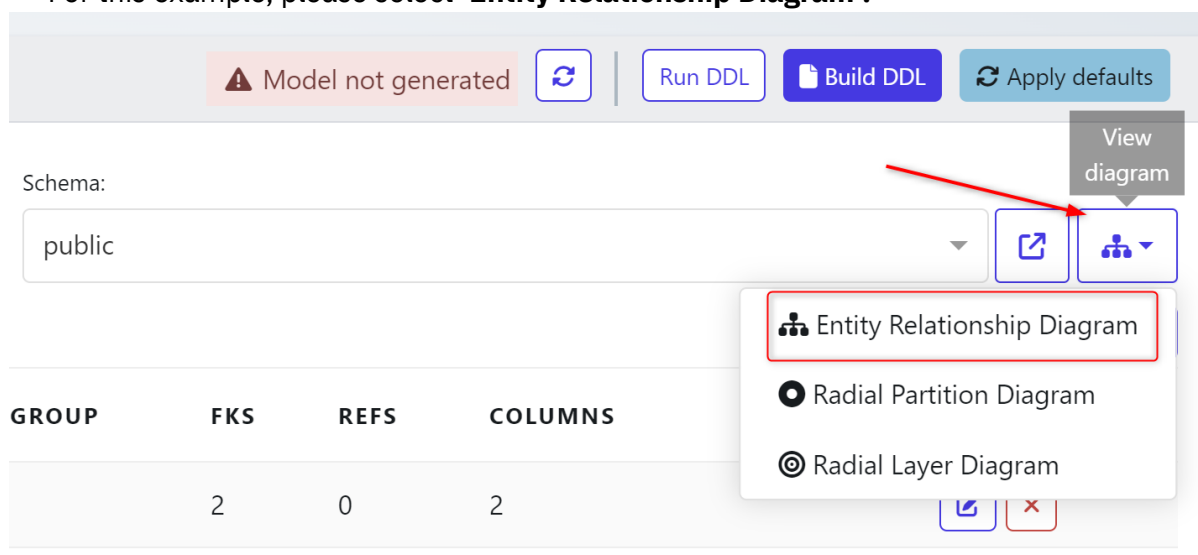
a) Create a new definition based off the scan that you conducted in Exercise 2

## Section 4 - Visualise table relationships

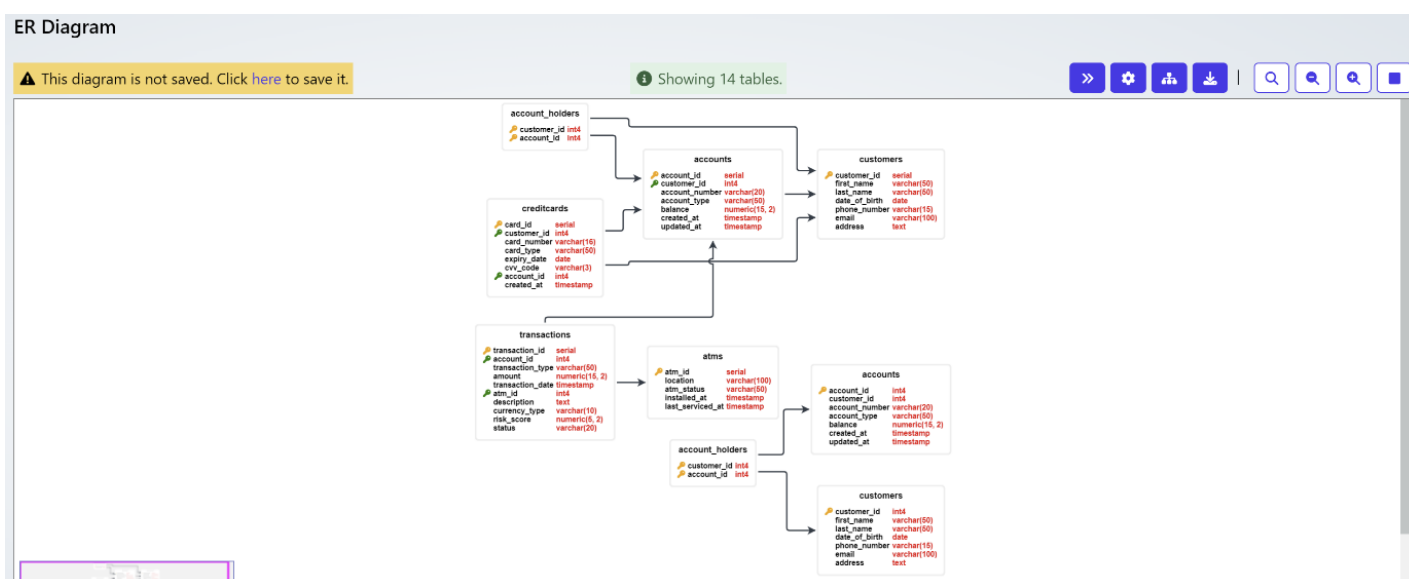
The Curiosity Enterprise Test Data Platform also lets you visualise the table relationships within the definition. This includes foreign key relationships and soft keys uncovered as part of the discovery process.



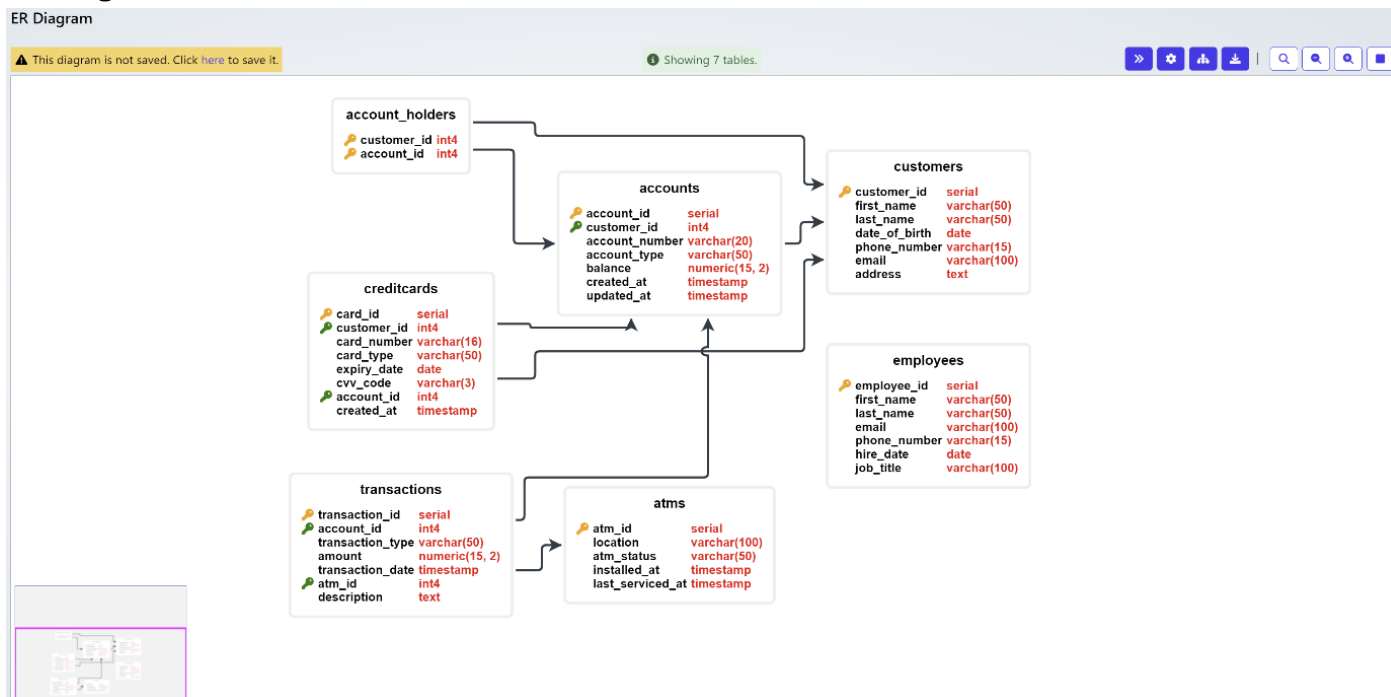
1. Click on the **tree** icon in to view diagram and choose the type of diagram you want to create. For this example, please select '**Entity Relationship Diagram**':



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.

### Settings

×

Show overview ☒ YES ☐

Show column data types ☒ YES ☐

Show FK cardinality ☐ NO ☐

Show groups ☐ NO ☐

Hide unrelated tables ☐ NO ☐

Column visibility  
☒ Show all columns  
☐ Show PKs and FKs only  
☐ Hide all columns

Maximum tables

[Use these settings as defaults](#)

CancelSave

## 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>

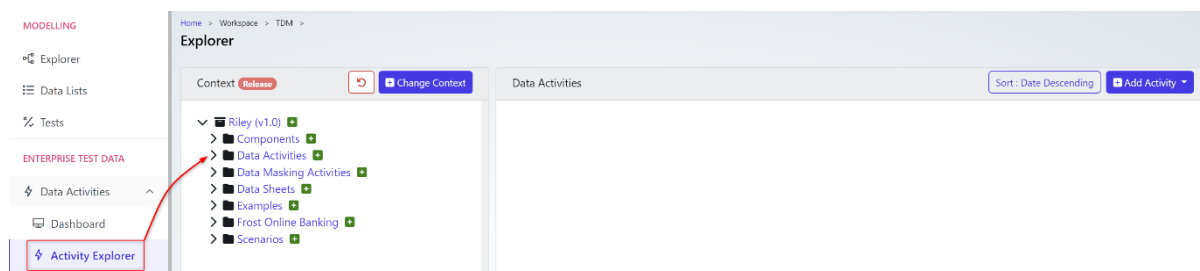
## Section 5 – Set up the database scan activity

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

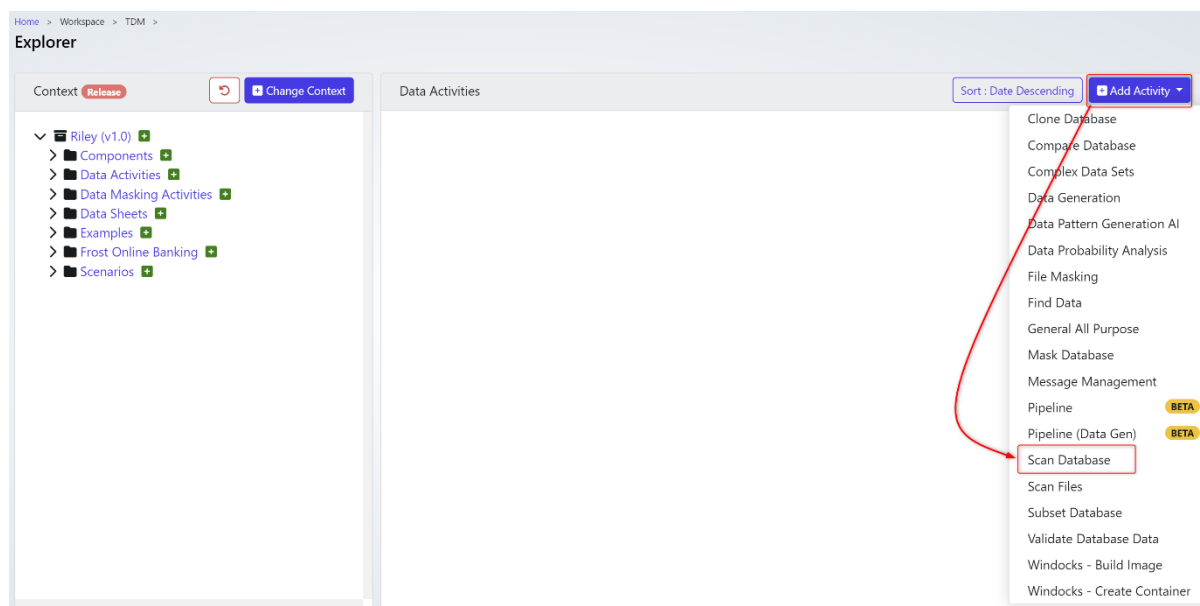
### Set up the database scan activity

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.

Scan Database ×

DETAILS SUMMARY

Name \*  
Curiosity Bank Connection

Application  
Base Application

Description\*  
Curiosity Bank Connection

Notes  
Notes

Tags  
Add tag

Server to use  
BIGONE

Cancel Next Step →

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

Scan Database ×

DETAILS SUMMARY

● Create a "Scan Database" Data Activity

● Save it in the specified location.

← Previous Step Cancel Finish

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

Details Visit the learning portal Version: Version #1

Components Run specs (1) Configuration

No components found

Add Components

Attach Submit Form

Attach Definition Version

Attach Category List

Attach Default Database Connection

Actions

Next Up

Create Data Scanning Submit Form

Create Category List

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

Select Connection ×

Connection Profiles

Curiosity Bank Connection × ▼

Cancel **OK**

7. A connection will display against the data activity.

Details Visit the learning portal Version: Version #1 ↻

Components Run specs (1) Configuration

#	NAME	TYPE	ACTIONS
≡	Curiosity Bank Connection	Connection   275	Modify Connection Profile <span>▼</span> <span>▶</span> <span>✕</span>

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

Add Components

Attach Submit Form

**Attach Definition Version**

Attach Category List

Attach Default Database Connection

Actions

Next Up

Create Data Scanning Submit Form

Create Category List

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

Select Definition Version ×

Definition

Curiosity Bank Connection × ▼

Version

Version #1 ▼

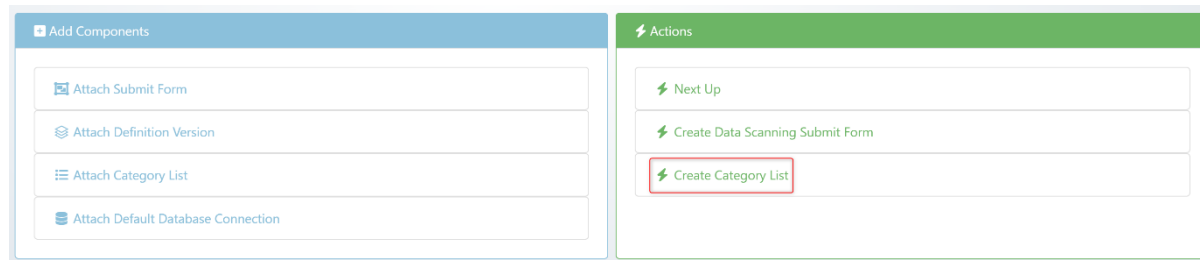
Cancel **OK**

## 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.

Create Scan Rules ×

Name

Example Scan Database Activity

Select Scanning Rules

× Scanning - Seedlist Lookup × Scanning - Columns Regex × Scanning - Data Values Regex × ▾

Application

Default ▾

Cancel OK

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

Details		Visit the learning portal		Version: Version #1			
Components Run specs (0) Configuration							
#	NAME	TYPE	ACTIONS				
≡	Example Scan Database Activity - Seedlist (Category)	Data List   2116	Modify Category List	▾	▶	×	
≡	Example Scan Database Activity - Regex Data (Category)	Data List   2115	Modify Category List	▾	▶	×	
≡	Example Scan Database Activity - Regex Column (Category)	Data List   2114	Modify Category List	▾	▶	×	
≡	BIGAWS Postgres OT otqa1 > Version #1	Definition Version   65					×
≡	Postgres BIGAWS OTDEV2 TEST > OT (otqa1)	Connection   24	Modify Connection Profile	▾	▶	×	

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:

Home > Workspace > UDM > Lists >

### List Details

Scan List Names Execution Processes (0) Comments (0) Watchers (0) Links (0)

Type: GENERAL  
Description: Scan List Names

Contents

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

Enterprise Test Data

Tests

Enterprise Test Data

Data Activities

Data Dictionary

Reservation

Rules

Data Virtualization

Quick Filters

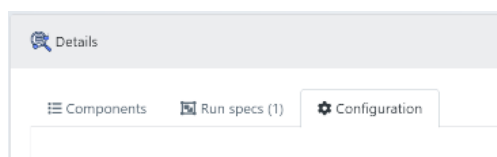
NAME	APPLICATION	DESCRIPTION	CREATED	TYPE	DATA SOURCE	ACTION
Starter	Application name filter					
ListNamesStarterList	Base Application	ListNamesStarterList	2024-10-23 at 10:10 AM	General	Static	
RegexStarterList	Base Application	RegexStarterList	2024-10-23 at 10:11 AM	General	Static	

Showing 7 of 3 items

- 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:

Seedlist-StarterList (General)	Data List   2845	Modify Category List
RegexStarterList (General)	Data List   2844	Modify Category List
Scan List Names (General)	Data List   2840	Modify Category List
Curiosity Bank Connection > Version #1	Definition Version   726	
Curiosity Bank Connection	Connection   375	Modify Connection Profile

- 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’**.

PROPERTY	VALUE
Include view when scanning the database	false
Search for PII data	true
The application containing the seedlists to use in scanning	
Count rows in tables	true
Find the percentage of null column values	true
Count distinct values	true
Clear down categories before scanning	true
Categorize using seedlists	true
Categorize using regular expressions	true
Categorize using column names	true
Number of rows to sample when categorizing data	1000
Minimum number of rows to match before assigning a category	5
Set minimum and maximum values	true
Treat blank columns as nulls when scanning	false
The locale to use in synthetic data generation functions	

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

Edit Config

Locale

English (Default)

Include views

NO

Scan data

YES

Scanning list application Id

Get table counts

YES

Get null percentage

YES

Get distinct count

YES

Clear down masking tags at each run

YES

Categorise using seed lists

YES

Categorise using regular expressions

YES

Categorise using column names

YES

Sample size

1000

List how many to match before assigning tag

5

Get min max values

YES

Blanks as nulls

NO

Cancel

OK

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.

The screenshot shows two side-by-side panels. The left panel, titled 'Add Components', has a blue header and contains four items: 'Attach Submit Form', 'Attach Definition Version', 'Attach Category List', and 'Attach Default Database Connection'. The right panel, titled 'Actions', has a green header and contains three items: 'Next Up', 'Create Data Scanning Submit Form' (which is highlighted with a red border), and 'Create Category List'.

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

### Data Activity - Create Data Scanning Submit Form - Job Parameters

Data Activity - Create Data Scanning Submit Form

The screenshot shows the 'Parameters' tab of a configuration window. At the top, there are two tabs: 'Parameters' (active) and 'Schedule'. Below the tabs, there are three sections. The first section, 'The Name of the Data Scanning Submit Process', has a text input field containing 'Scan Curiosity Bank Database'. The second section, 'The Group to put the new Data Scanning Submit Process in', has a text input field containing 'Scan DB'. The third section, 'OR choose an existing Process and Update it', has a dropdown menu that is currently empty. At the bottom of the window, there are two icons on the left (a code icon and a download icon) and two buttons on the right: 'Cancel' and 'Execute'.

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.

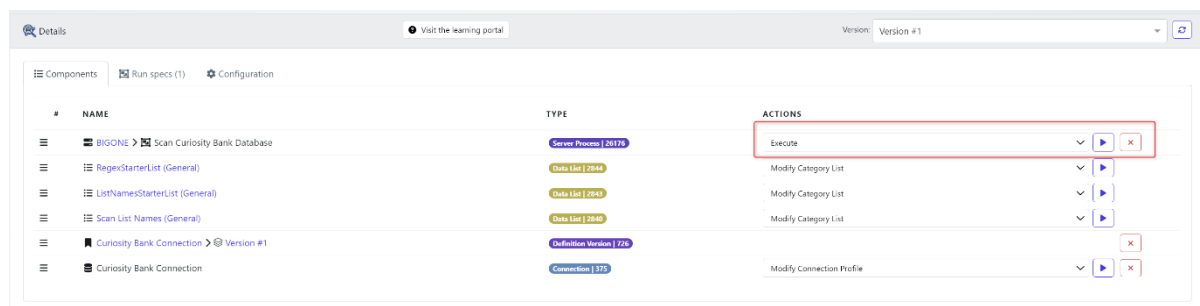
### Exercise 6

- 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.

### Scan Curiosity Bank Database - Job Parameters

Scan Curiosity Bank Database

Parameters

Schedule

Connection ID to scan\*

Curiosity Bank Connection

☒ Run SchemaCrawler to get catalog metadata

Log level used by SchemaCrawler?

INFO

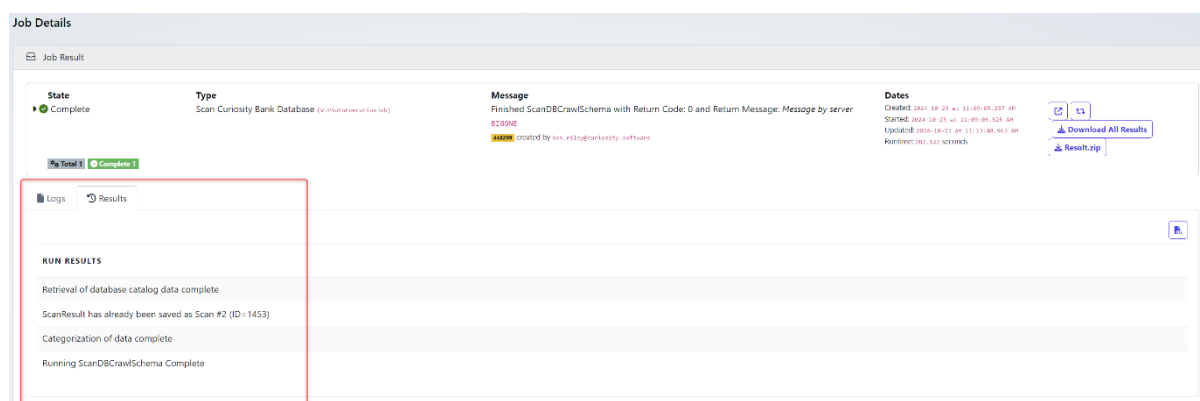
</>

Download

Cancel

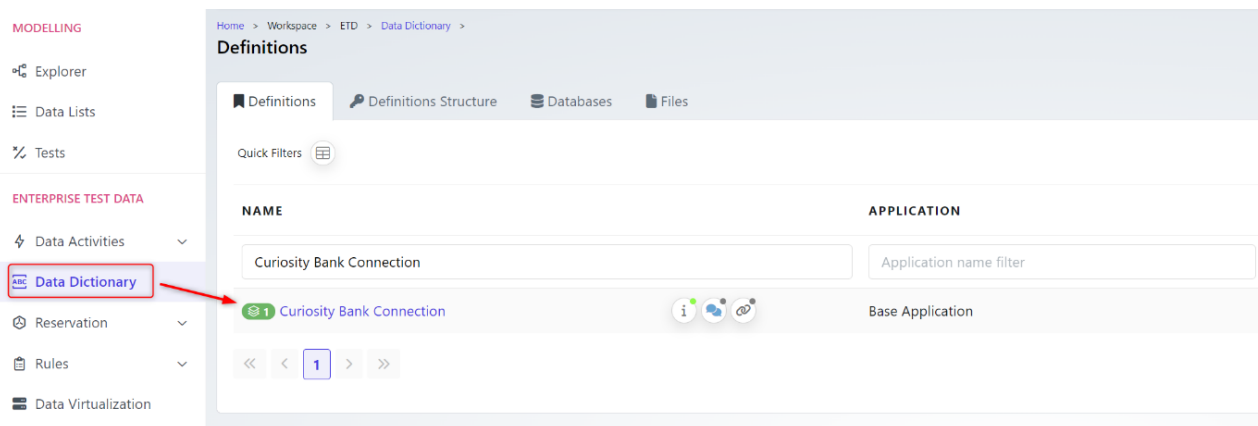
Execute

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



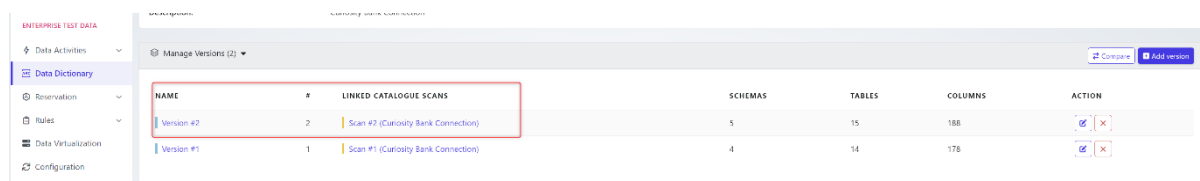
#### 4. Review the results

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

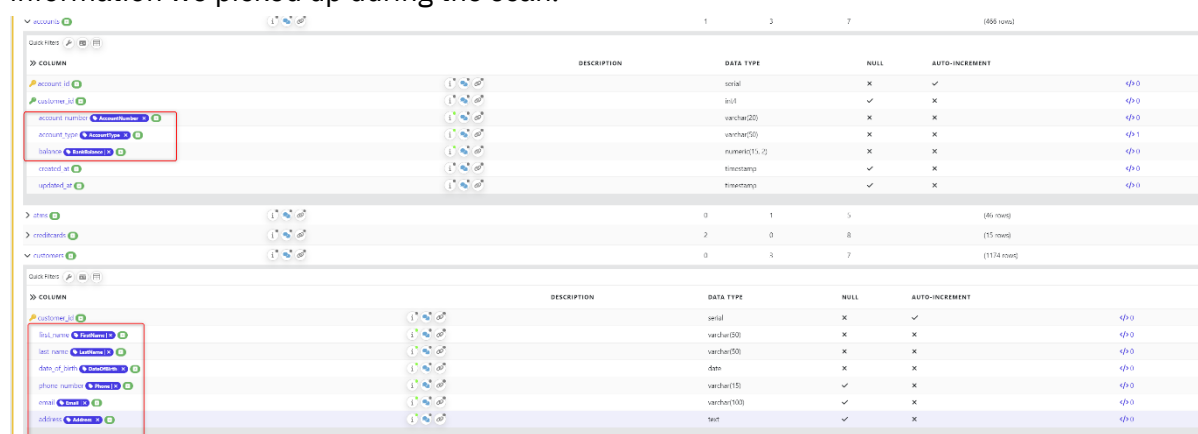


#### 5. The **Scan #2** version will now hold the scanned PII formation. Click on **Scan #2**.

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

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

[Check the solution videos for all Exercises in this course >](#)