

## Synthetic data generation into databases

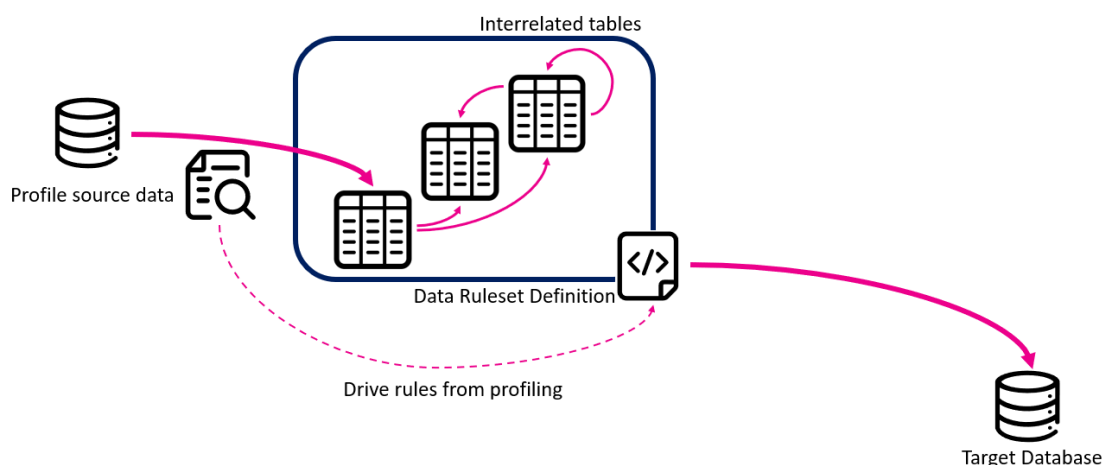
### Introduction

With Curiosity Software, our synthetic data generation capabilities allow you to create data that can be inserted into a database, either directly or as a file to be uploaded through another process. This training will show you how to set up the database synthetic data generation activity and how to configure the ruleset ready for execution using multiple techniques, including our AI capabilities.

The techniques will be more suitable for different scenarios and each technique will briefly describe when they are most useful, for instance applying the default rules will apply the needed synthetic data generation rules for most use cases and are customisable so that they can be adapted to the needs of the user.

### Training overview:

This training course will take you through the journey of creating a database synthetic data generation activity using an existing connection and a profiled definition. We will also introduce you to the main techniques for populating a rule set and show you how to incorporate these into different execution methods such as test data pipelines, API calls and self-service forms.



Here is a visual model of the high-level process of [creating a synthetic data generation rule set >](#)

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

- Create a Database Synthetic Data Generation Activity
- Create templates for synthetic data generation
- Use the AI helper to populate the rule set
- Use Data Painter to populate rule sets
- Incorporate into a test data pipeline
- Use pre and post actions

### Pre-requisites for synthetic data generation training:

- A [Database Connection](#)
- A [Data Definition](#)
- To have completed the [‘Definition’ section of Module 1](#)

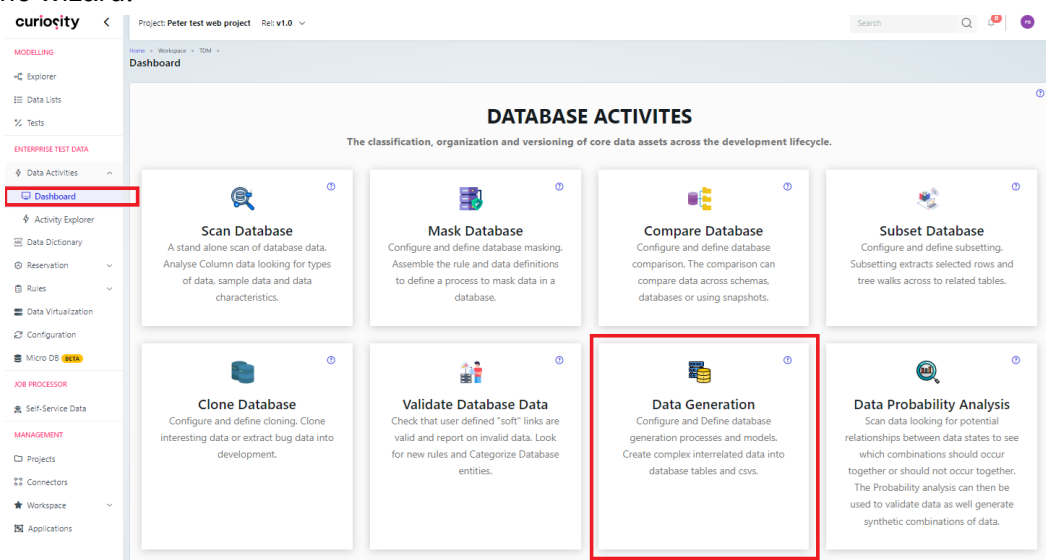
## Section 1 - Set up the data generation activity

The Curiosity Platform allows users to easily set up Data Activities to create re-usable assets such as the synthetic data generation activity. The activity will allow us to create a rule set which will contain the necessary rules to create the synthetic data that is needed.

There is also the option to create a data generation submit form, which allows you to execute the synthetic data generation routine. This can be executed either via the tool in a self-service portal or through other methods such as an API call.

### Step 1 – Create the data generation activity

In the enterprise test data dashboard, you will see the ‘Data Generation’ activity. Click on this to start the wizard.



### Step 2 – Fill out the details tab

This will open a dialogue box for you to start filling out information. First up is the ‘Details’ tab. Enter the **Name** of the activity (1), **Description** of the activity (2), **Application** the activity belongs to (3), and choose the default VIP **server** (4) you want to use to execute the action.

**Data Generation** [X]

● DETAILS    ● DEFINITION    ● RULE SET    ● LOCATION    ● SUMMARY

Name \*  1

Application  3

Description\*  2

Notes

Tags

Server to use  4

### Step 3 – Choose the definition

The next tab is the 'Definition' tab, where you will need to select a **Database Definition** (1). The **Version** (2) and **Connection** (3) will be automatically populated based on the selected definition.

**Data Generation** [X]

DETAILS DEFINITION RULE SET LOCATION SUMMARY

Database Definition \*  
Select a definition 1  
This field is required

Version \*  
Version #2 2

Connection  
Select a connection 3

← Previous Step Cancel Next Step →

### Step 4 - Choose the rule set

Next, you need to select the rule set. There are two options for your rule set selection, or you can skip the activity and complete it later.

#### 1. Existing rule set

If you choose an existing rule set, only the rule sets linked to the definition chosen on the previous page will be displayed.

**Data Generation** [X]

DETAILS DEFINITION RULE SET FORM LOCATION SUMMARY

**Existing Rule Set**  
Choose an existing Rule Set Version for the selected Definition.

**New Rule Set**  
Create a new Rule Set and Version for the selected Definition.

**Skip**  
Skip attaching a Rule Set Version for now - can be done later

Rule Set Version\*

← Previous Step Cancel Next Step →

#### 2. New rule set

To create a new rule set click on the '**New Rule Set**' option. You can configure the rule set using the techniques shared in [Section 3 of this training](#). This action will let you set which tables you are going to generate data for, using the '**New Rule Set**' dialogue box.

The 'Details' tab lets you set the name and description for the rule set.

New Rule Set (Version) ×

**Existing Rule Set**  
A new Rule Set Version under an existing Rule Set, out of the Rule Sets linked to any version of this Data Activity.

**New Rule Set**  
A brand new Rule Set and Rule Set Version.

**i** Details **⚙** Configuration **📄** Tables

Name \*

Description\*

Notes

Tags

**i** Please select a Definition with an up-to-date model, or generate the model for the selected Definition to proceed.

The 'Configuration' tab lets you set whether all tables will be used and whether ID columns are active.

**i** Details **⚙** Configuration **📄** Tables

Definition

Version \*

Choose if the identity columns (PKs) are set to **Active** by default.  
ID columns Active ☐ NO

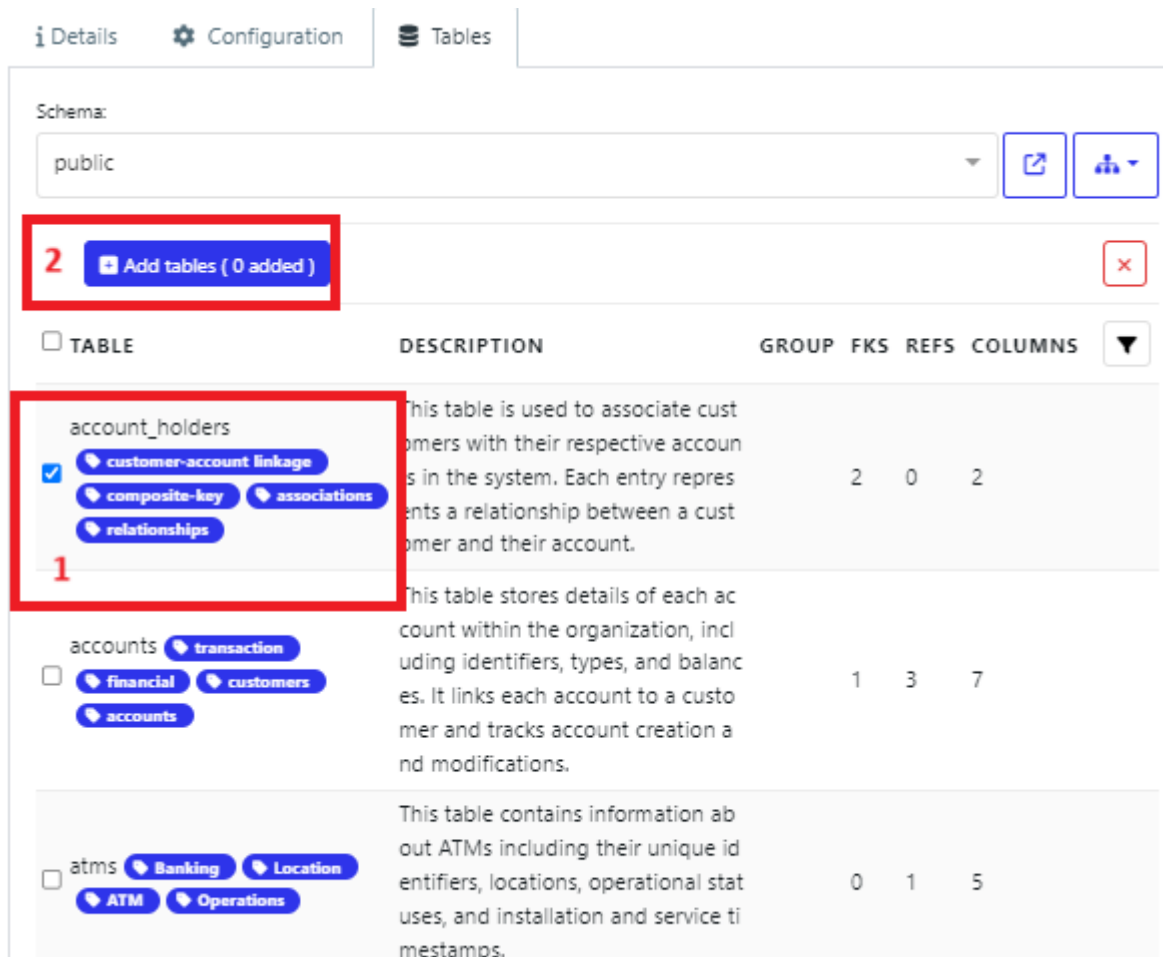
Select the tables you wish to **include** in the application of rules - if **NO**, then all tables will be included.  
Select Tables? ☒ YES ☐ NO

Order Tables by FK ☒ YES ☐ NO

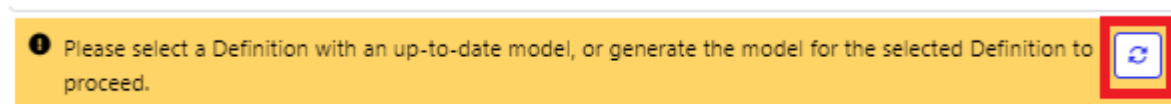
Preview Server

The 'Tables' tab will allow you to select which tables the data generation will be based on.

**Note:** once you have selected one or more tables (1), you need to click the **'Add Tables'** button (2) in order to add them to the rule set.



When you create the rule set, you will need to generate an up-to-date model for the definition if you do not already have one. To do this, click the **'Regenerate'** button that will appear to the right in the yellow dialogue box (shown below). If you already have an up-to-date model, this will not appear.



The job will open in a new browser window and once it is complete, you can save your rule set.

## Step 5 – Create infrastructure and form (Optional)

If you have selected an existing rule set, you will be given the option to also create the infrastructure, the VIP flow, and the submit form. If you have selected the **'New Rule Set'** option, this will not appear.

Once your rule set is complete, you will be able to set up the 'submit form'. By default, '**Create Submit Form**' (1) will be checked. If you uncheck it, then form details part (2) will be removed. The form is necessary to give the user an interface from which to execute the VIP flow, that creates the synthetic data.

If the **Prepare Requisite Infrastructure** box is ticked (3) this will automatically create the VIP flow from the existing ruleset and any other required internal configuration files needed to enable synthetic data creation.

**Data Generation**

DETAILS DEFINITION RULE SET FORM LOCATION SUMMARY

These steps make sense only when using a pre-configured Rule Set. If the Rule Set selected is not configured already, it is best to skip them and do them afterwards.

☒ Prepare Requisite Infrastructure (3)

☒ Create Submit Form (1)

**Select Server** (2)

Server: BIGONE

Type of Submit form to Be Created\*: A Standard Data Generation into a Database

The Name of the Submit Process (Will default to the Name of the Activity if Blank):

The Group to put the new Submit Process in: Data Generation

OR choose an existing Process and Update it:

☒ Add in drop down selections for parameters linked to definitions with enumerations

☐ Include a field to override today's date in the submission form

☒ Create or Update the Parameter List for Modeller

Previous Step Cancel Next Step

**Note:**

If no rule set has been selected, this will not be displayed.

There are more details on setting up a submit form in [Section 4 – Step 3](#) of this training.

## Step 6 – Complete the creation

Once the previous sections have been completed, the next step is to select the location that the Data Activity will be saved in. It defaults to the current project that the user's session is in.

**Data Generation**

DETAILS DEFINITION RULE SET FORM LOCATION SUMMARY

- Projects
  - Peter general cases and test
    - v1.0
      - case tests
      - Components
      - Data Activities
      - Data Sheets
      - Examples
      - Scenarios
      - tests for docs
      - TestIDList.xlsx

← Previous Step Cancel Next Step →

## Step 7 – Finish the data generation activity

The **'summary'** tab provides a list of the actions to be carried out, so you can review them before finishing the wizard.

**Data Generation**

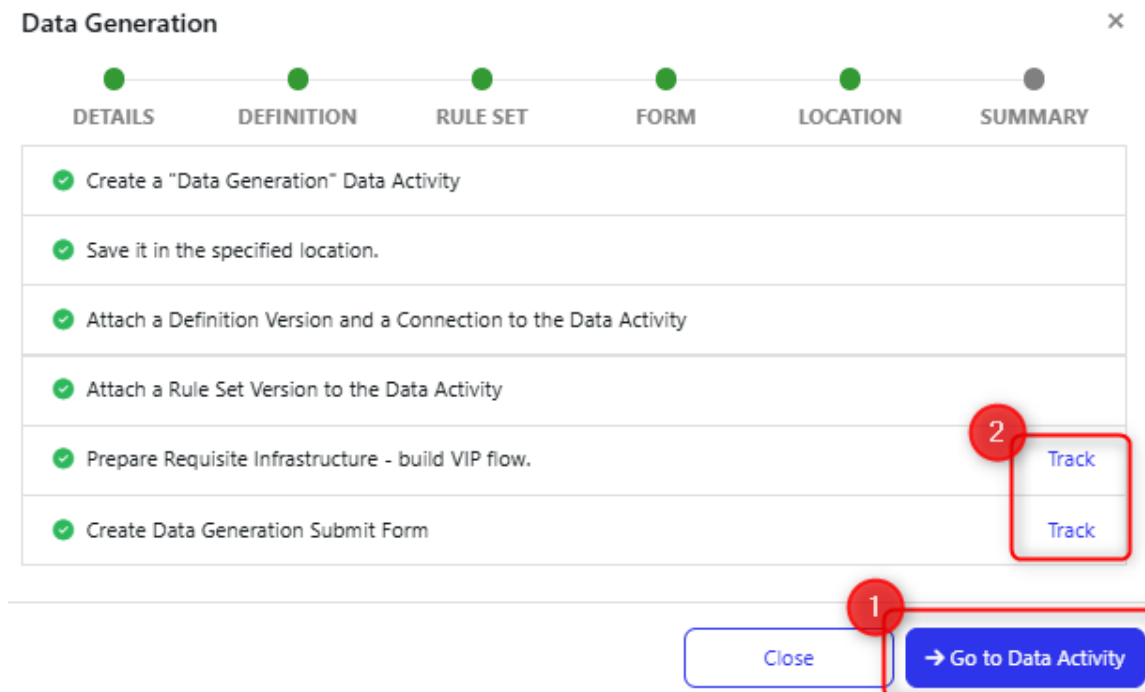
DETAILS DEFINITION RULE SET FORM LOCATION SUMMARY

- Create a "Data Generation" Data Activity
- Save it in the specified location.
- Attach a Definition Version and a Connection to the Data Activity
- Attach a Rule Set Version to the Data Activity
- Prepare Requisite Infrastructure - build VIP flow.
- Create Data Generation Submit Form

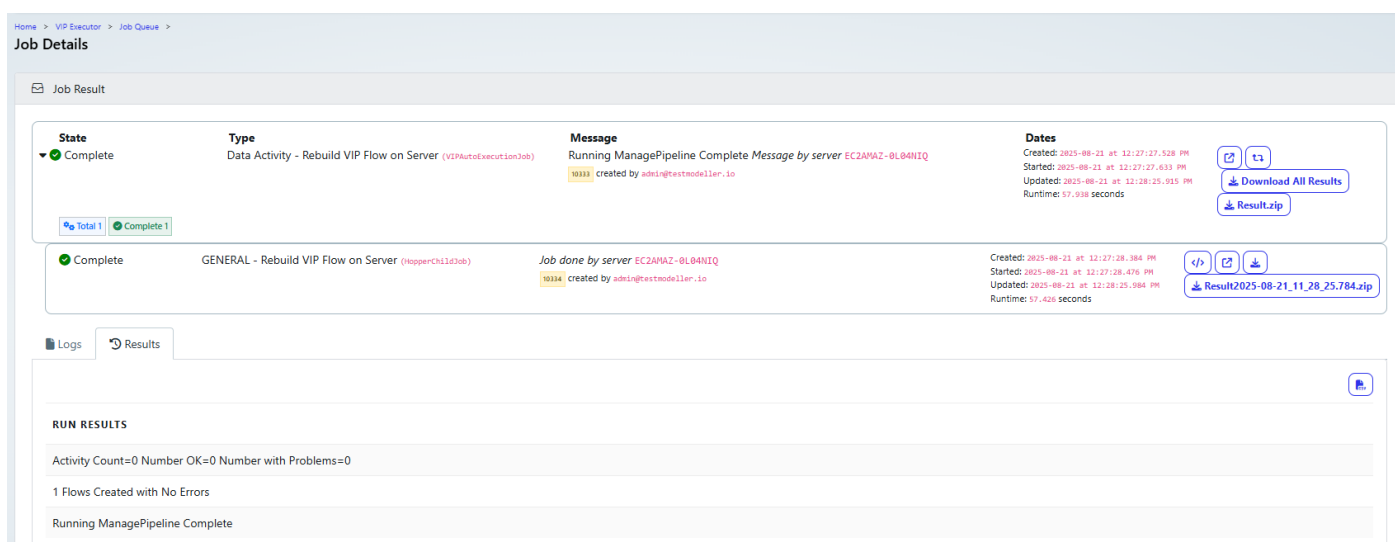
← Previous Step Cancel Finish

Once you click **'Finish'**, you have the option to go directly to the data activity (1) or you can navigate to it in the activity explorer at a later time.

You can also view the jobs that created the VIP Flow and submit form (2), if they were created. By tracking these jobs we can view if there are any compilation issues due to the configuration of the rule set, and see how the job is progressing. Clicking the **‘Track’** button will take you to the job details screen where you can view what is happening and download the logs if needed.



If you have chosen an existing rule set and generated a submit form, the activity is now ready to use. There are further instructions on how to use the activity in [Section 3](#) of this training module.



## Exercise 1

1. Create a new activity using the wizard, choosing to **‘Configure a new rule set’**



## Section 2 – Generation rule set and accelerators

In the generation rule set section, there are various accelerators and helpful bits that allow you to quickly configure the rule sets for your requirements. The generation rule set page allows you to see these assets, automatically create them and manually edit them if required.

There are different techniques in the synthetic data generation activity which allow the rule set and other assets to be populated depending on user requirements. In this section we will take you through some the techniques on offer and how they can assist the user generating rule sets.

### Step 1 – Navigate to the rule set page

To navigate to the rule set page, click on the ruleset name in the activity screen or change the action to modify by the rule set component and then press the play button.

The screenshot shows the Curiosity Enterprise Test Data Platform interface. The left sidebar contains navigation links for MODELLING, ENTERPRISE TEST DATA, and JOB PROCESSOR. The main area displays the 'Data Activity' page for 'Peter test web project'. The 'Components' tab is active, showing a table of components. A red box highlights the 'Employee data > Version #1' component, which is a 'Generation Rule Set Version | 1254'. The 'ACTIONS' column for this component shows a 'Modify' button with a play icon and a red 'X' icon.

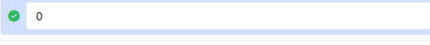
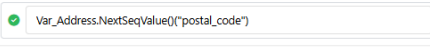
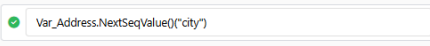
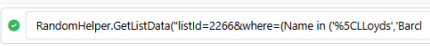
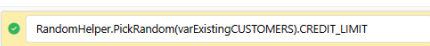
This will open a screen that looks like the below:

The screenshot shows the 'Generation Rule Set' page for 'Example Data Generation'. The page displays the 'Rules' section, which contains a table of rules. The table has columns for #, TABLE, and ACTIVE. The rules listed are: 1. customers, 2. order\_items, and 3. orders. Each rule has a play icon and a red 'X' icon in the ACTIVE column.


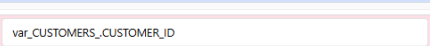
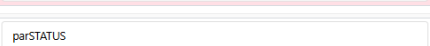
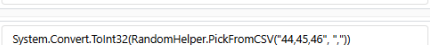
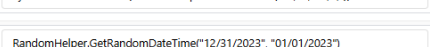
When you open a rule set, the columns will come colour-coded depending on the source of the function.

Colour sample	Description
	Peach – From initialisation
	Light Pink – From pre-processor
	Pale Blue – From validation rules
	Light Yellow – From global defaults
	Cool Gray – From scan defaults
	White – Set manually
	Mint Green – From analysis
	Lavender – From AI
	Aqua Blue – From accelerator

For example, in the below, it is possible to see that most of these rules have been manually edited as they are white, one has come from a global default as it is light yellow, and one has come from a pre-process as it's light pink.

COLUMN	DATA TYPE	NULL	AUTO-INCREMENT	RULES	ACTIVE
1 CUSTOMER_ID	decimal	x	✓	0	
2 NAME	string	x	x	Var_Address.NextSeqValue("postal_code")	
3 ADDRESS	string	✓	x	Var_Address.NextSeqValue("city")	
4 WEBSITE	string	✓	x	RandomHelper.GetListData("listId=2266&where=(Name in ('%5CLLOYDS';Bard	
5 CREDIT_LIMIT	decimal	✓	x	RandomHelper.PickRandom(varExistingCUSTOMERS).CREDIT_LIMIT	




  

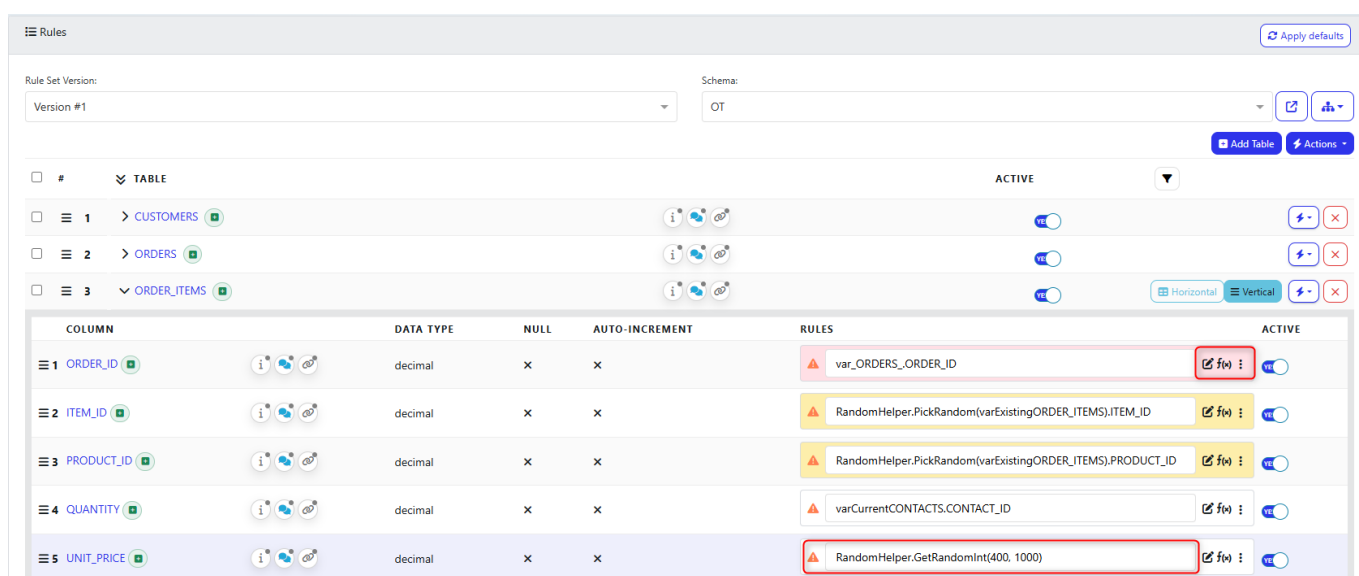
COLUMN	DATA TYPE	NULL	AUTO-INCREMENT	RULES	ACTIVE
1 ORDER_ID	decimal	x	✓	0	
2 CUSTOMER_ID	decimal	x	x	var_CUSTOMERS_CUSTOMER_ID	
3 STATUS	string	x	x	parSTATUS	
4 SALESMAN_ID	decimal	✓	x	System.Convert.ToInt32(RandomHelper.PickFromCSV("44,45,46", ","))	
5 ORDER_DATE	DateTime	x	x	RandomHelper.GetRandomDateTime("12/31/2023", "01/01/2023")	

## Step 2 – Column Actions

On the rule set page, you can view the columns for a table, by clicking the arrow to the left of the table name. You can also view the current rule for the column, the origin of the rule (shown by the colour around it), and also whether the column has been checked to see if it has the correct syntax (shown by a red warning triangle, or a green tick). This allows you to easily see which rules are currently in the rule set and which ones may need changing.

For example, in the view below, the UNIT\_PRICE column is set to “RandomHelper.GetRandomInt(400, 1000)” which will generate a random integer from 400 to 1000. On each column row, there is a set of symbols that will respectively:

-  Open the data painter
-  Insert a function
-  Open a list of accelerators for the column



The screenshot shows the 'Rules' page in the Enterprise Test Data Platform. It displays a table of columns and their associated rules. The columns are ORDER\_ID, ITEM\_ID, PRODUCT\_ID, QUANTITY, and UNIT\_PRICE. The rules are defined using the RandomHelper class. The UNIT\_PRICE rule is highlighted with a red box.

COLUMN	DATA TYPE	NULL	AUTO-INCREMENT	RULES	ACTIVE
ORDER_ID	decimal	x	x	var_ORDERS_ORDER_ID	Active
ITEM_ID	decimal	x	x	RandomHelper.PickRandom(varExistingORDER_ITEMS).ITEM_ID	Active
PRODUCT_ID	decimal	x	x	RandomHelper.PickRandom(varExistingORDER_ITEMS).PRODUCT_ID	Active
QUANTITY	decimal	x	x	varCurrentCONTACTS.CONTACT_ID	Active
UNIT_PRICE	decimal	x	x	RandomHelper.GetRandomInt(400, 1000)	Active

### Accelerators available on a rule set:



- Populate value using AI** – Use our AI helper to suggest a rule for you
- Cast type** – automatically put in logic to fix any data type issues
- (Random) Create Parameter from Table** – select random data from a database column
- (Sequential) Create Parameter from Table** – select sequential data from a database column
- Create Default Rule** – Creates default rule from the expression in the rule set

### Exercise 2

- Navigate to the rule set you have just created and use the ‘Random Create Parameter from Table’ accelerator to get data to be used in synthetic data generation. For example, ‘Product ID’ from the product table.

### Step 3 – User-defined variables

The values can also be set to a variable in the ‘User-defined Variables’ section of the Rule set page. For example, in the below you will see a user-defined variable called ‘LastName’.

User-defined Variables (1) <span>+ Add</span>							
<input type="checkbox"/>	NAME	DESCRIPTION	UI TYPE	GROUP	REQUIRED	DEFAULT	ACTION
<input type="checkbox"/>	LastName <span>STRING</span>	The Last Name is embedded at run time	Text box	Ungrouped		Smith	 

This was added by clicking the ‘+Add’ button and filling in the ‘New User-defined variable’ dialogue, an example of which is below.

#### New User-defined variable ×

Form Parameter ☒

Name\*

Description\*

Type

Form Group

Notes

Help Text

UI Type

Default

☐ Required Parameter

[»Form Field Rules 0](#) [»Advanced](#) [Data Mapping](#)

These user-defined variables can also be linked to a synthetic data generation function edited with VB.Net syntax. In addition, some user-defined variables will be automatically populated by the AI ruleset accelerators and use of the defaults.

### Exercise 3



1. Create a user defined variable that uses one of the synthetic data generation functions

## Additional information

### 1. Pre and post processes







You can see and edit the current pre and post processes that are part of your generation activity. This allows you to do a variety of actions either before or after the generation routine kicks off, for instance to prepare the environment for generation or kicking off a stored procedure.

These processes can be either an expression, custom VIP flow or a link to another activity. When deciding whether to use a pre or post process, consider if the process will need to be done every time a generation routine is kicked off. If so, use a pre or post process, and if not, a test data pipeline will be a better option.

Pre - Post Process (1) <span>+ Add</span>				
NAME	ACTIONS	PARAMETERS	ACTIVE	ACTION
Expression	0	0	<input checked="" type="checkbox"/>	 

### 2. Foreign key rules

It is possible to view any active foreign key or soft key rules that can be used to generate the data. These are mainly detected through our profiling and discovery techniques and can be found in that training.

Foreign Key Rules <span>Actions</span>			
FOREIGN KEY		PARENT	CHILD
FK_ORDER_ITEMS_ORDERS	  	ORDERS	ORDER_ITEMS
FK_ORDERS_CUSTOMERS	  	CUSTOMERS	ORDERS
Showing 2 of 2 foreign key rules			







In this screen you can choose whether the relationship is active and used for generating data by clicking on the **'Active'** toggle.

There is also an **'Actions'** button on the top right-hand corner where you can choose to either activate or de-activate all the rules.

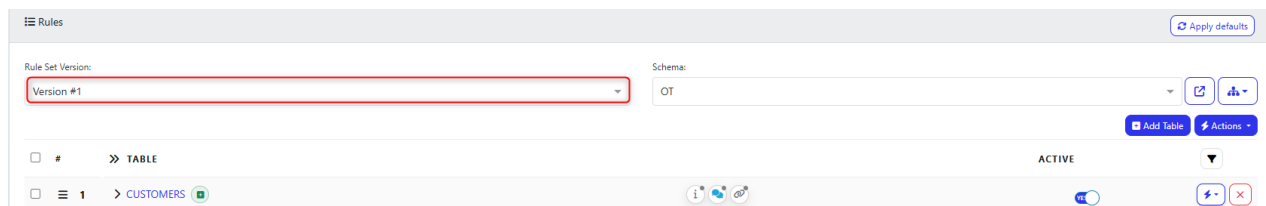
### 3. Manage versions

In this screen users can manage the different and new versions of the activity. Further information can be found in the workplace fundamentals section.

- **New Version** - Creates a new version of the ruleset
- **Clone** – Creates a new version of the activity using the same assets
- **Upgrade** – Creates a new version of the activity using new versions of the assets
- **Compare** – Shows the differences between two different versions

Manage Rule Set Versions (2) <span>New Version</span> <span>Compare</span>				
NAME		DESCRIPTION	#	DEFINITION VERSION
Version #2	  	Upgrade Orders	2	BIGAWS ORACLE OT Version #1
Version #1	  	Used for the find as well	1	BIGAWS ORACLE OT Version #1

You can change the version that the rule set screen views by selecting the drop down as shown below:



#### 4. Data sources

You can also manage different Data sources for the synthetic data generation which can act as a source for data. This can be very useful for getting product information from a master database for instance and is useful for making sure data that is generated is referentially integral across environments.

You can click the **‘+Add’** button to add another data source to the activity. You can also click the edit button to change a current data source.

Data Sources (1) <span>+ Add</span>		
CONNECTION	ALIAS	ACTION
<a href="#">BIGAWS ORACLE CFIRST &gt; XEPDB1 (OT)</a>	BIGAWS__ORACLE__OT	<span>✎</span> <span>✖</span>

#### Next steps:

1. [Check the solution videos for all Exercises in this course >](#)
2. **Start Section 3 of your synthetic data generation training:**
  - [Download Part 2 of your training guide >](#)
  - [Head to the Curiosity Partner Portal >](#)