MIRACLE
SOFTWARE SYSTEMS

PDF Invoice Data Extraction into Excel File

RPA-UiPath | Digital Summit 2024

Goal

In this hands-on lab, you will learn how to create an automation bot using UiPath Studio Web. The bot will extract specific fields, such as **Invoice Number**, **Date**, and **Due Date**, from a PDF file using **regex patterns**. The extracted information will then be structured into a data table and written into an **Excel file** for easy storage and analysis. By the end of this lab, you will understand how to install and set up UiPath Studio, work with regex for text matching, and automate the process of creating structured outputs from unstructured data.

Pre-Requisites

- Gmail Account

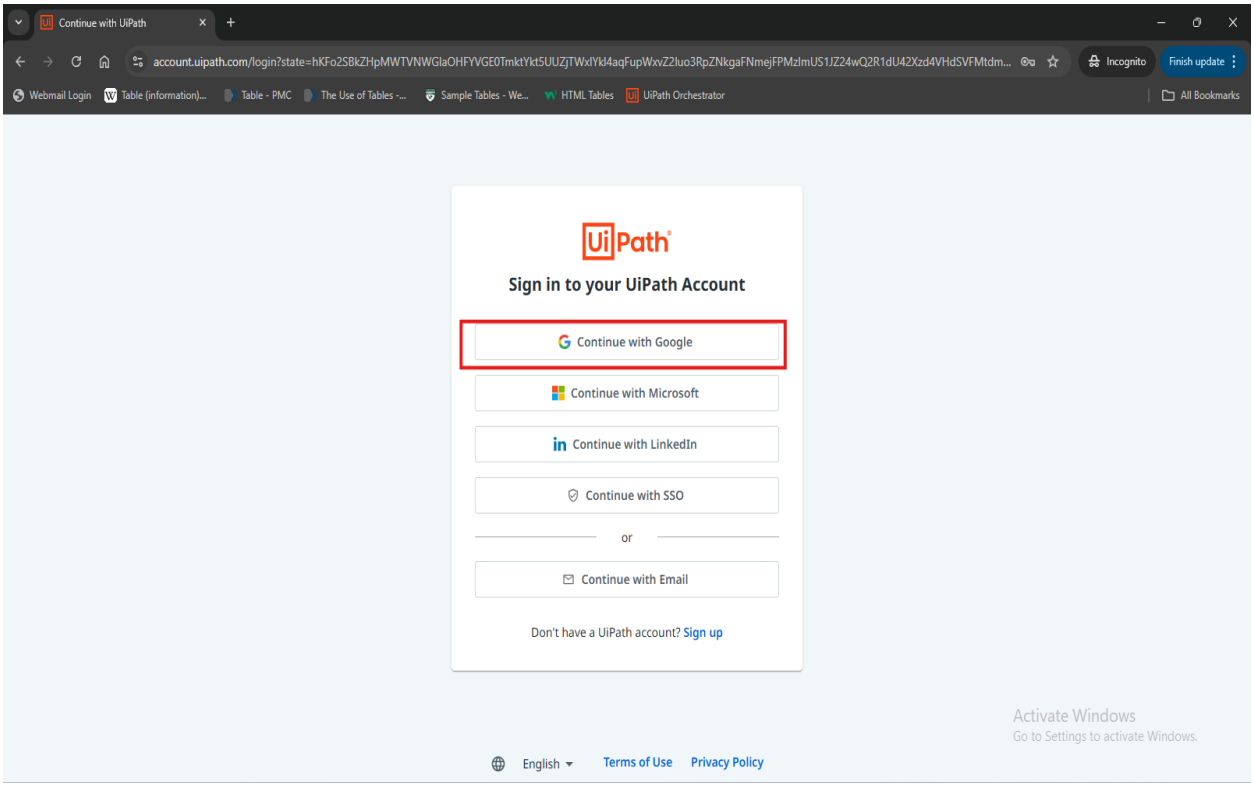
Technology Involved

- UiPath (Studio Web)

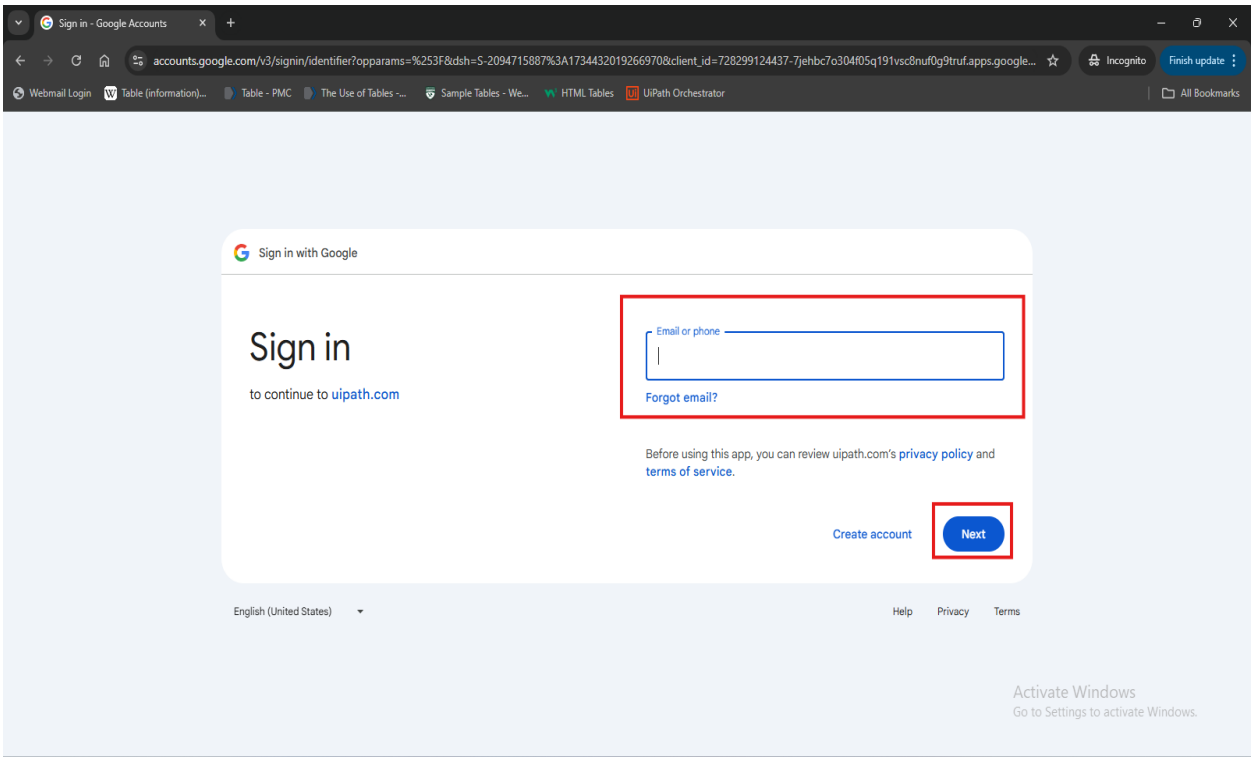
Lab Steps

Let's get started with automation!

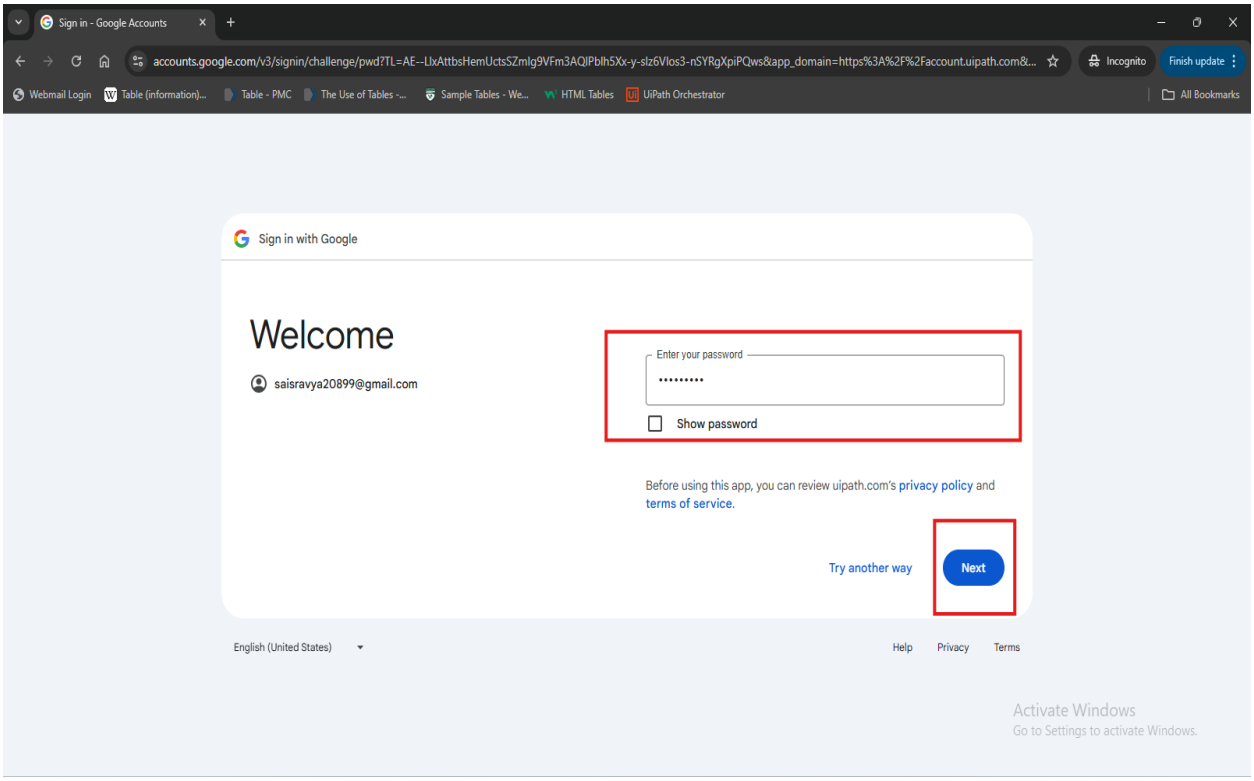
Before you get started, go to https://cloud.uipath.com and click on **Continue with Gmail** to Sign up.



Now enter your **Email** and click on the **Next** button

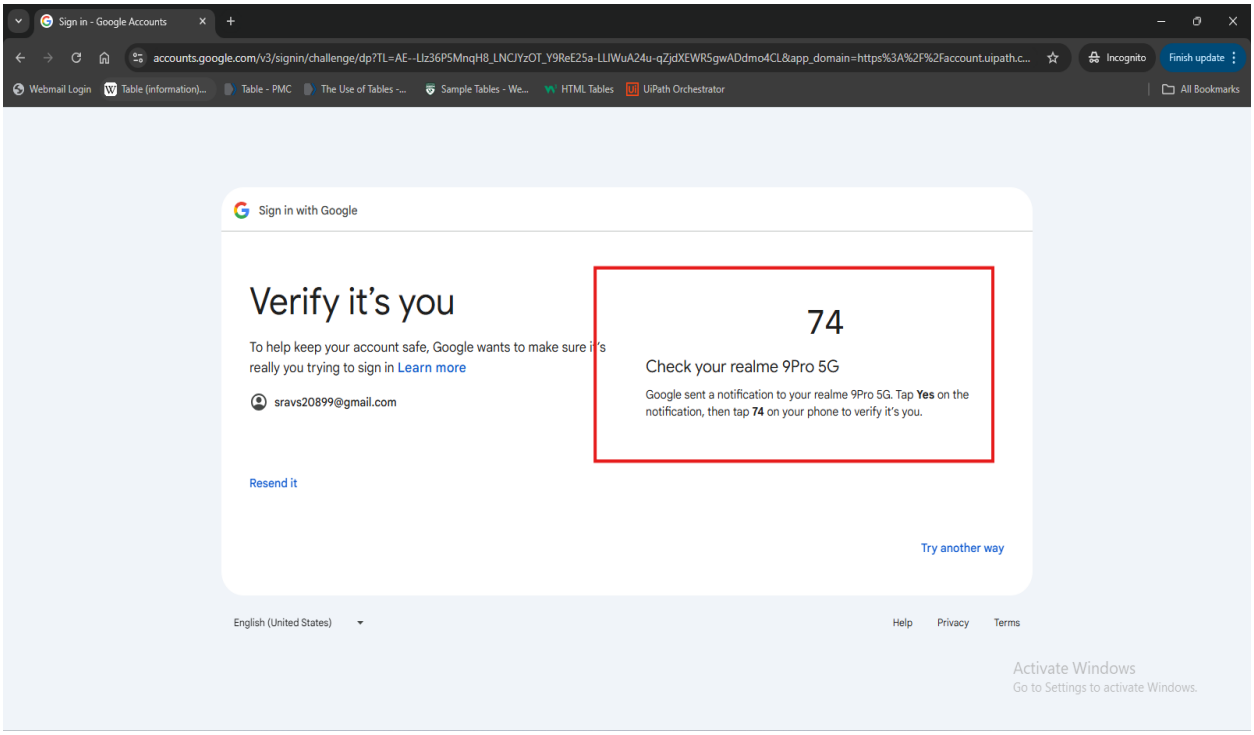


Now enter your account **Password** and click on the **Next** button

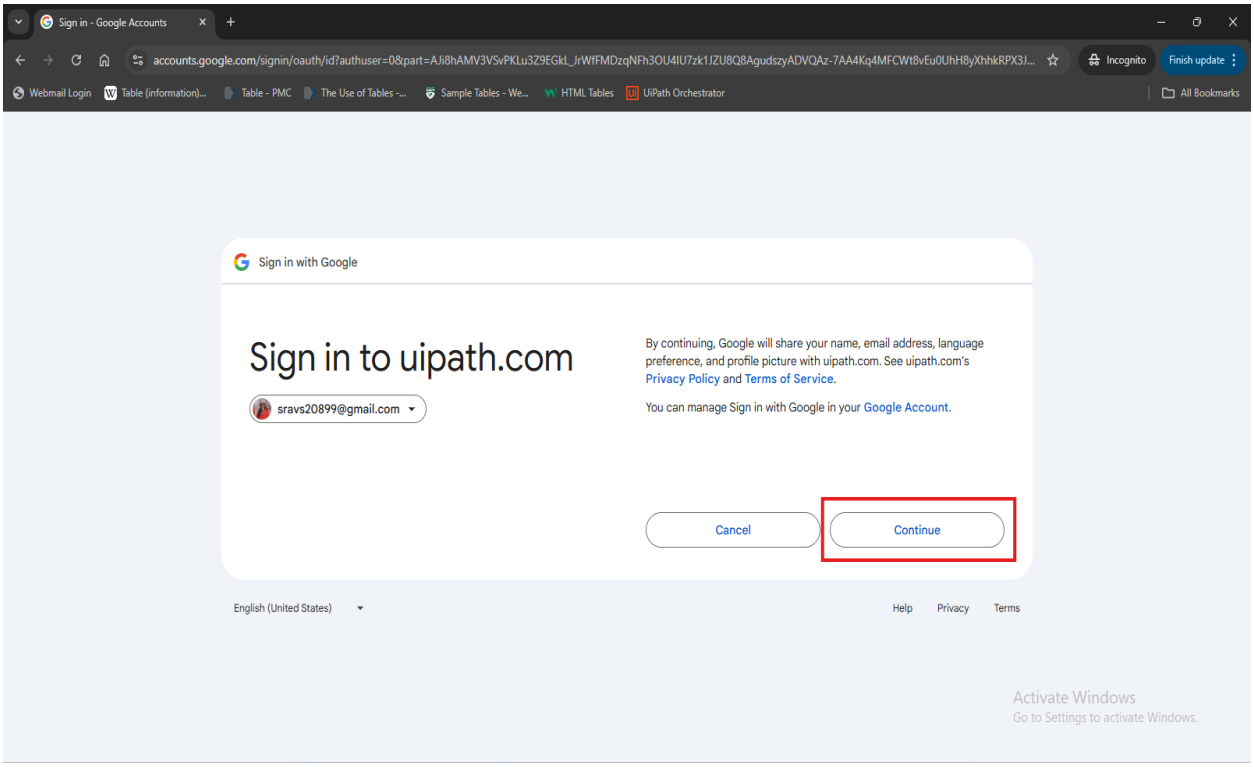


It will ask for account verification and you can follow the steps mentioned there

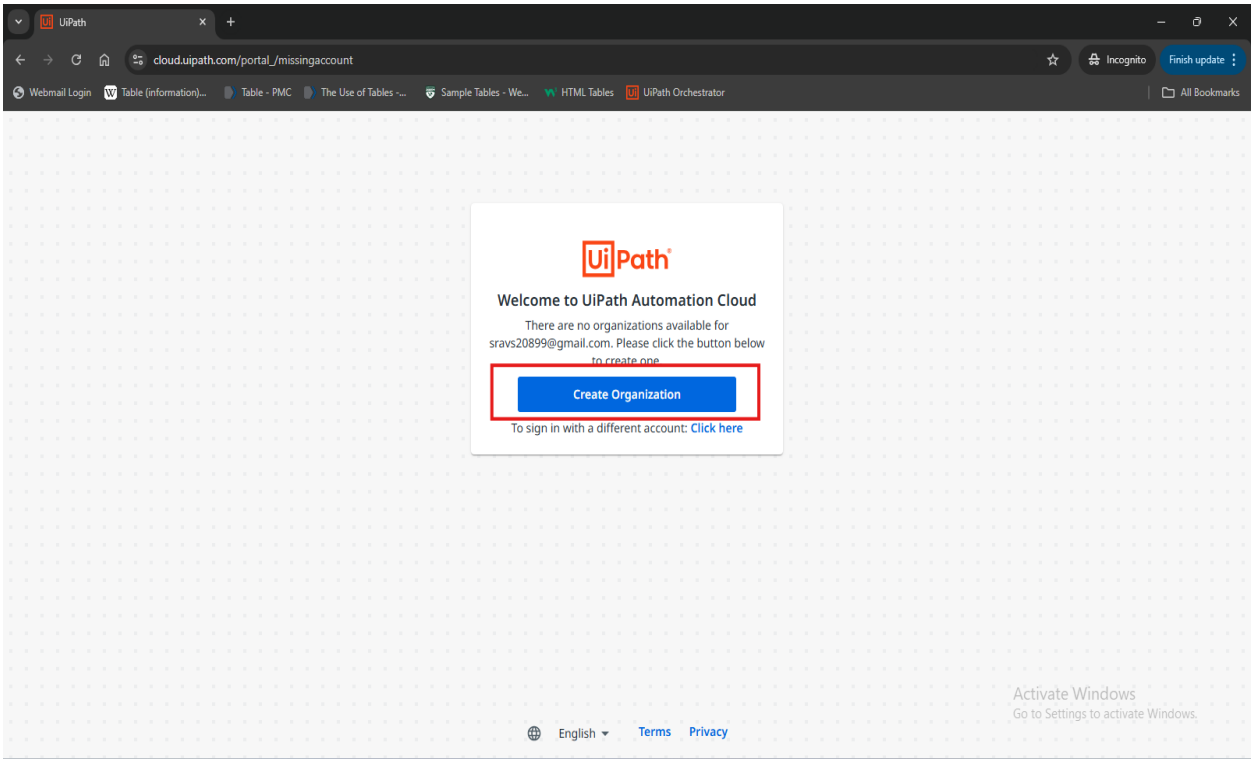
For (Ex: In the below screenshot you can see that it sent a confirmation notification to the given email ID and you should select the given number in your mobile once we receive the notification)



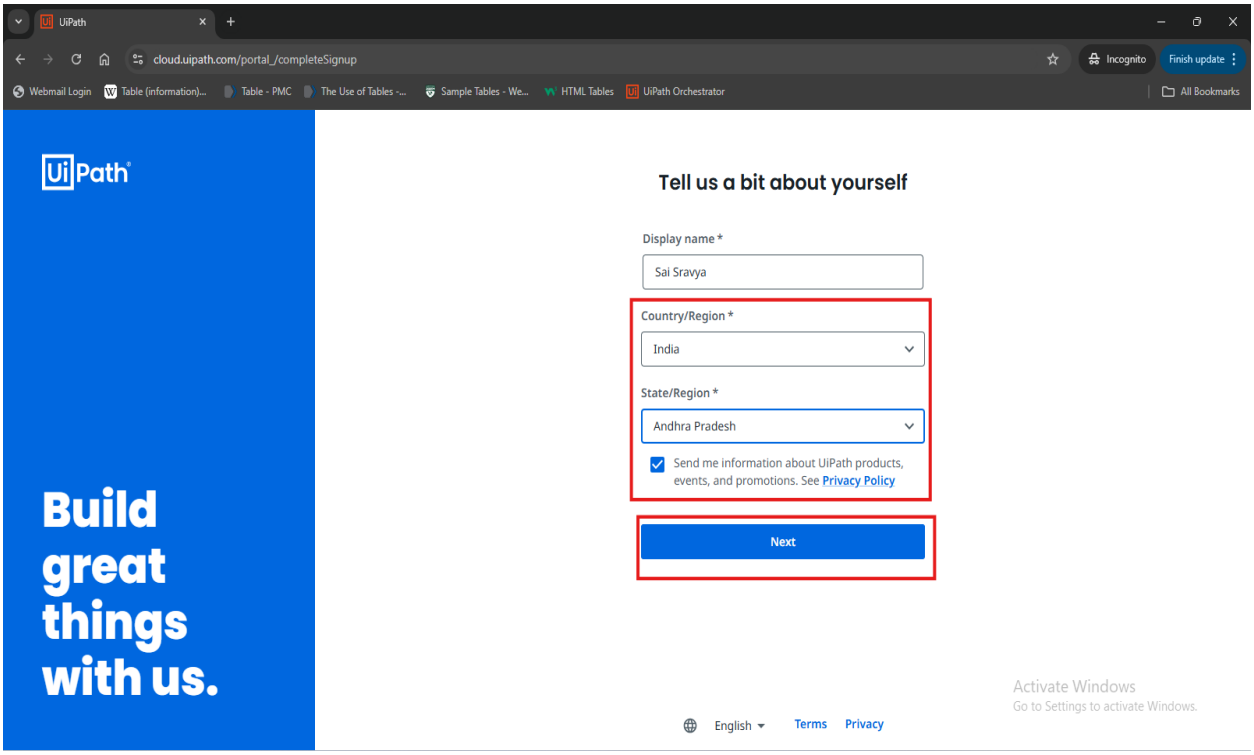
Once the verification is done you will get the below screen to Sign In to UiPath Orchestrator and to do that click on **Continue** button



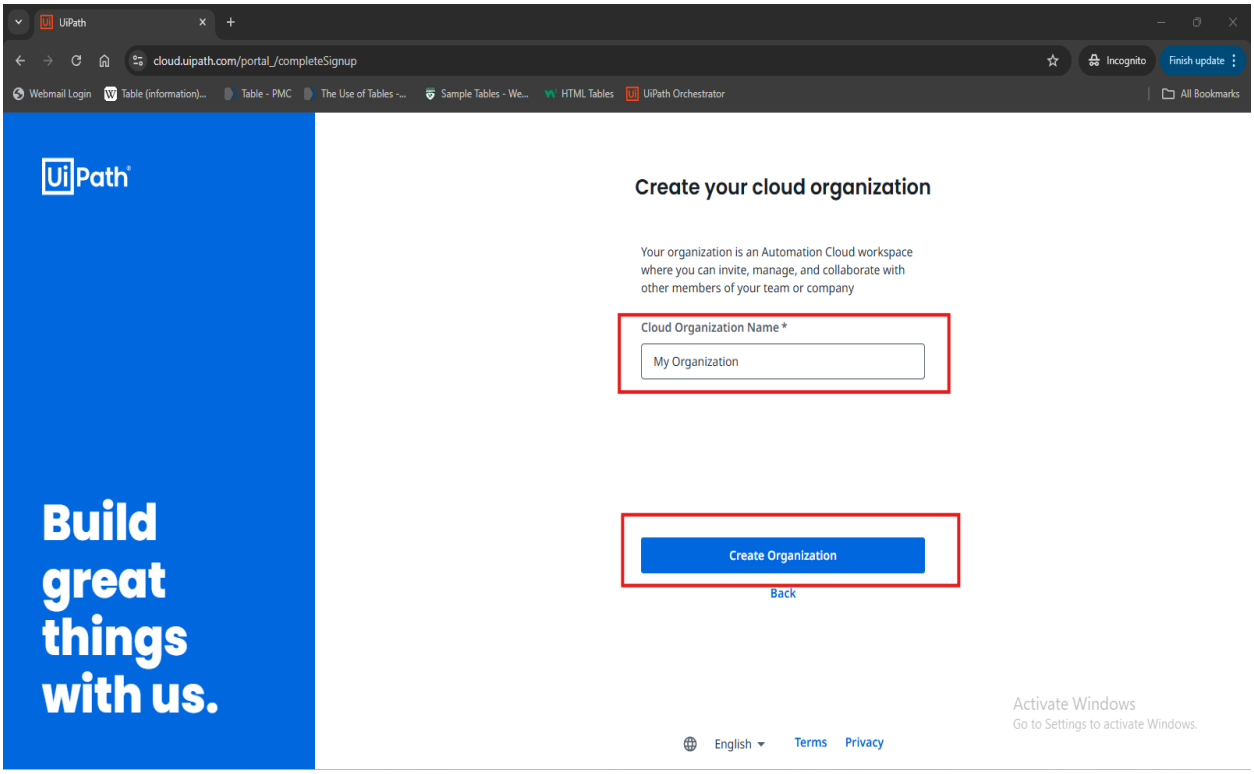
Now we need to create an Organization to start our Automation, for that click on **Create Organization** button as shown below



Enter the required fields as shown below and click on **Next** button



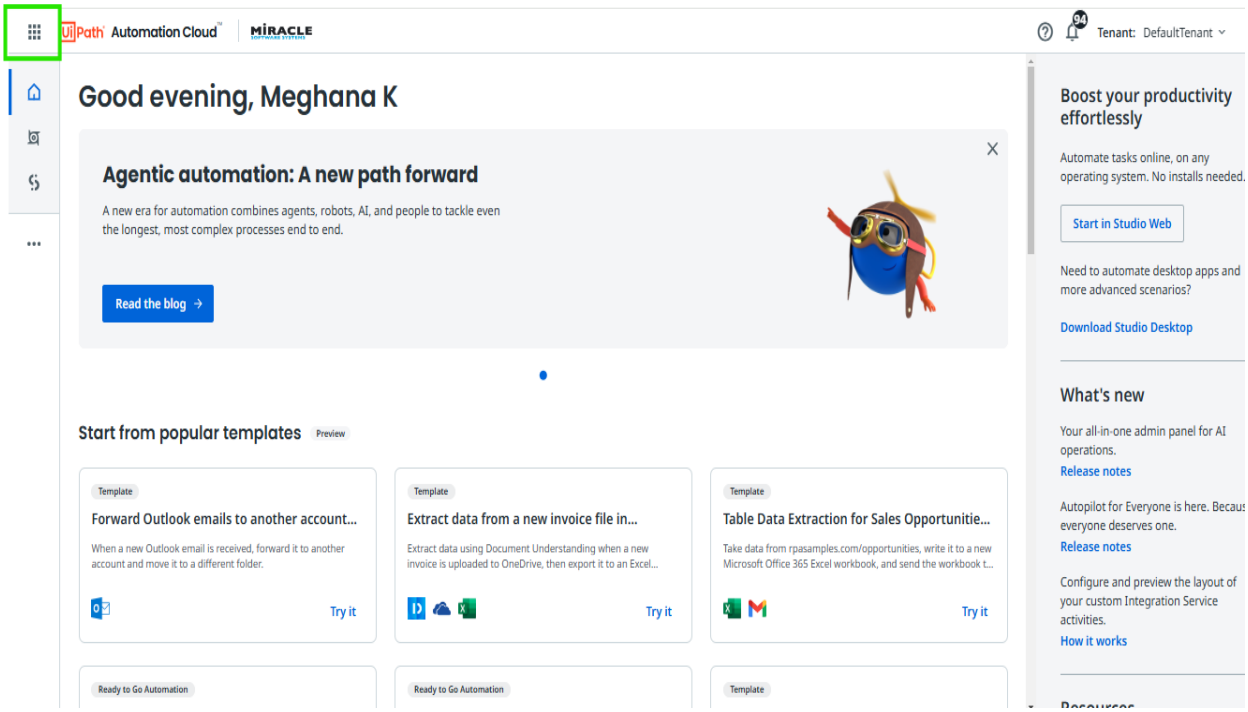
Now give some Name to your Organization and click on the **Create Organization** button



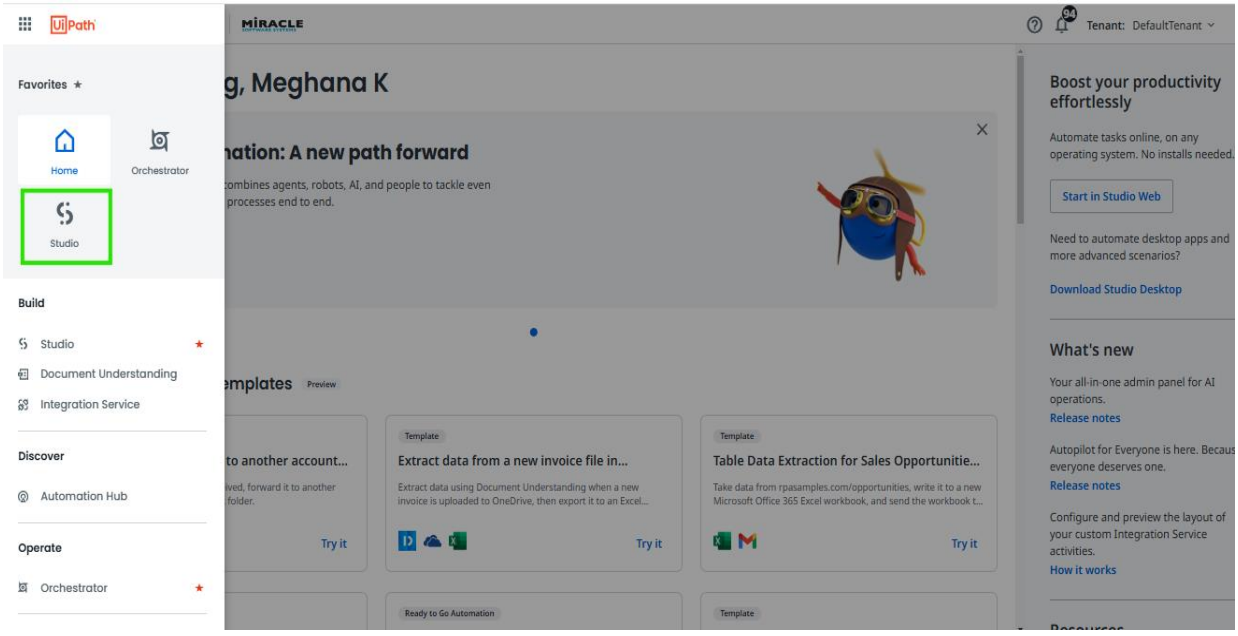
Once all the above steps are done you are set to start your Automation

Step #1 Opening the Studio web

Click on the **Menu** button as shown below.

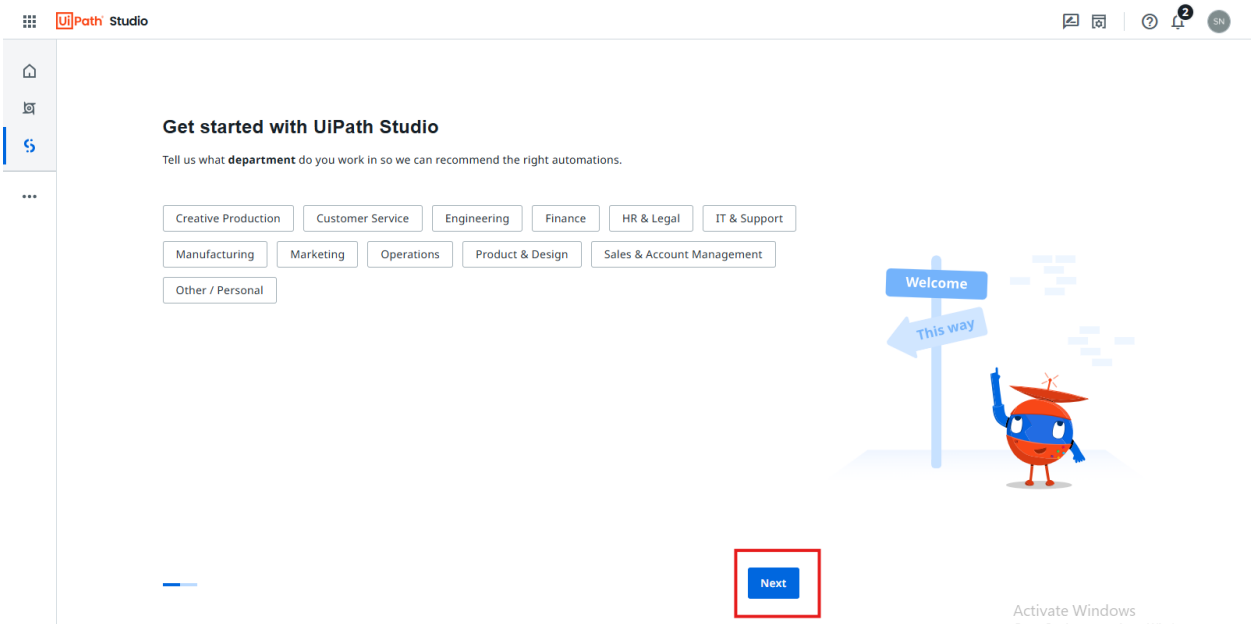


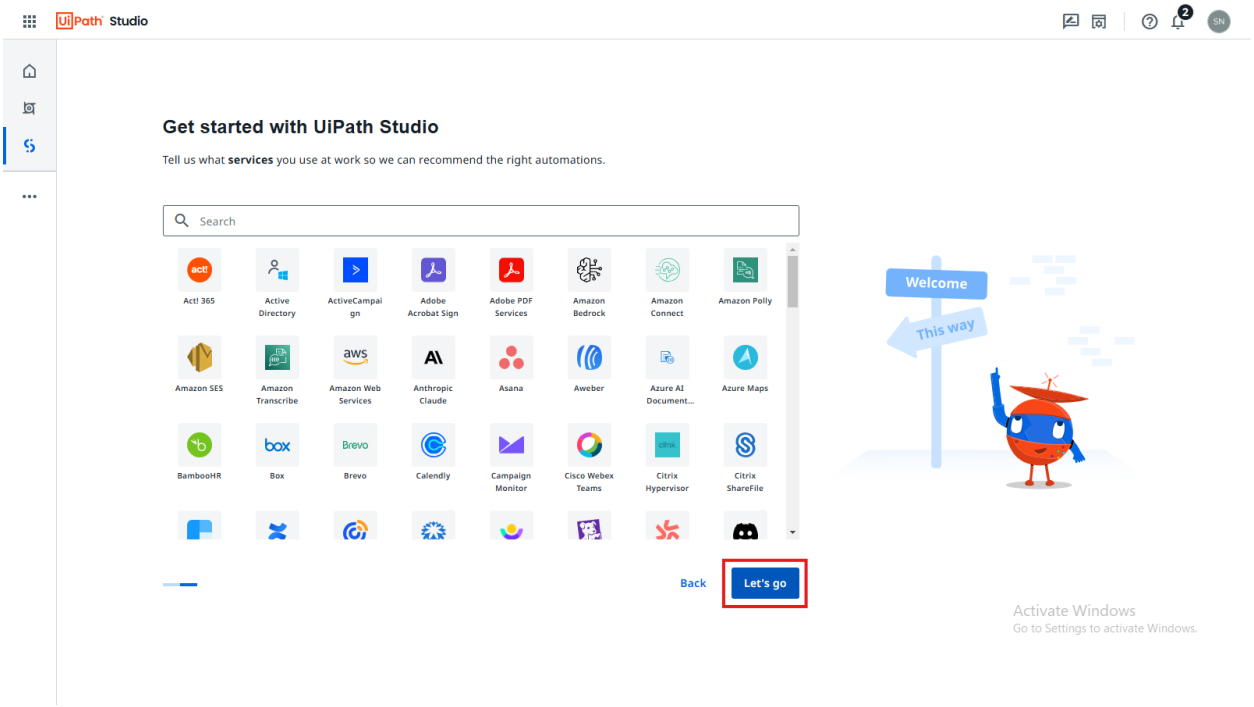
Click on the **Studio** icon.



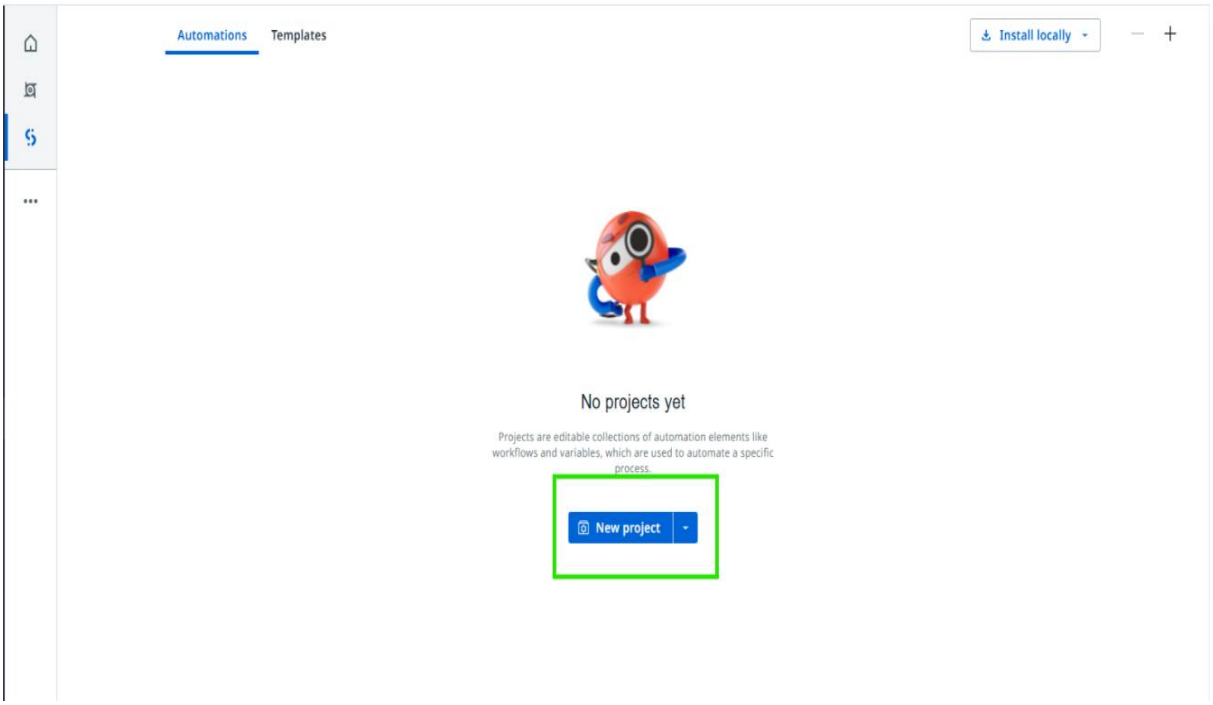
Step #2 Creating the New Project

For this first click on the **Next** button as shown



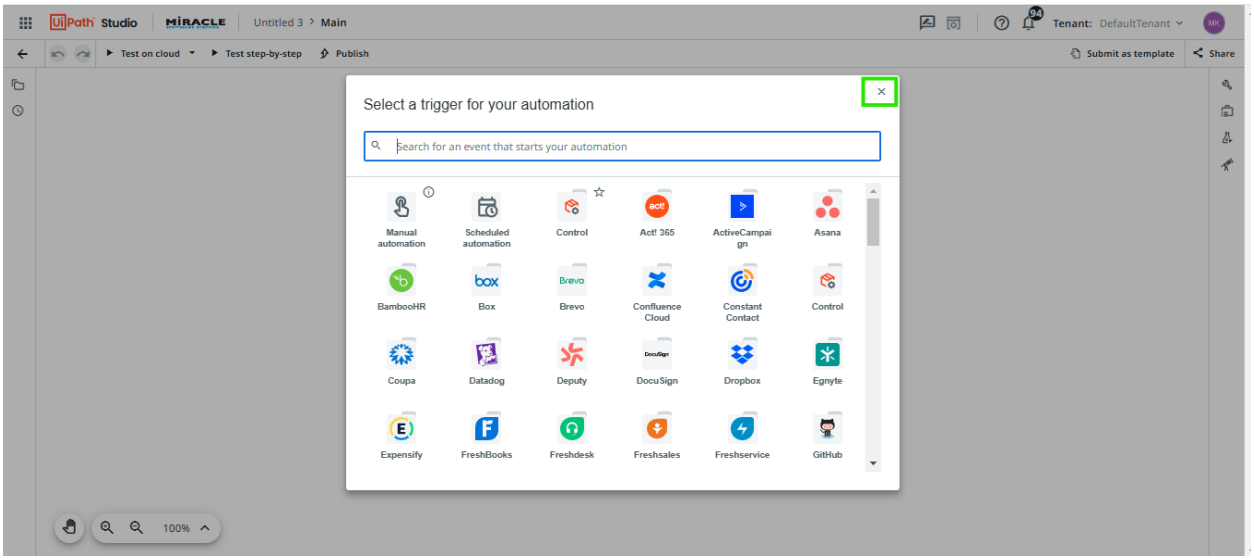


Select the Automation tab and click on New Project.



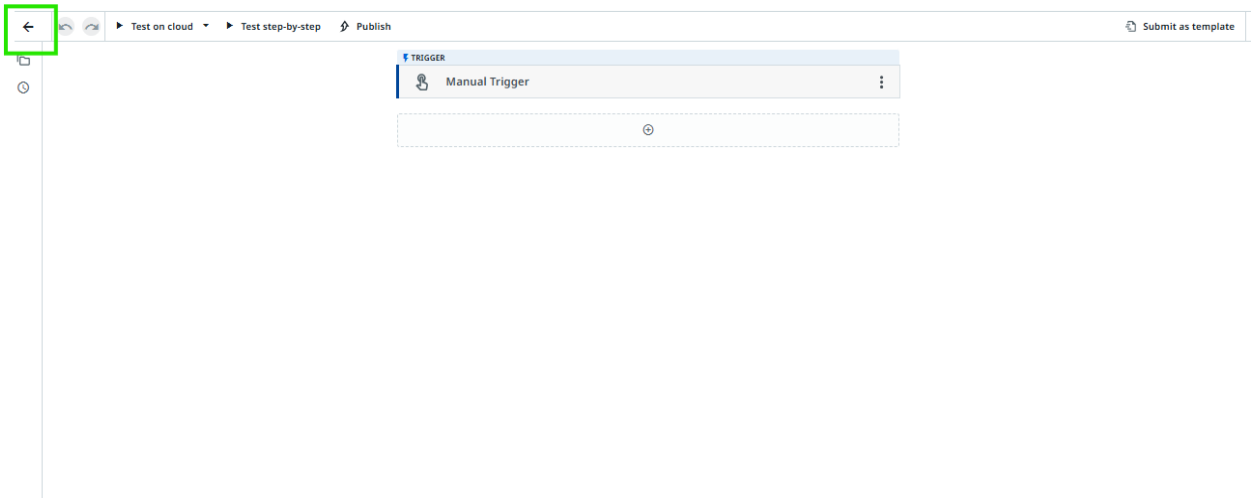
After clicking on the **New project** you will be getting a pop up like below.

You can close this pop up as we won't be using any triggers.

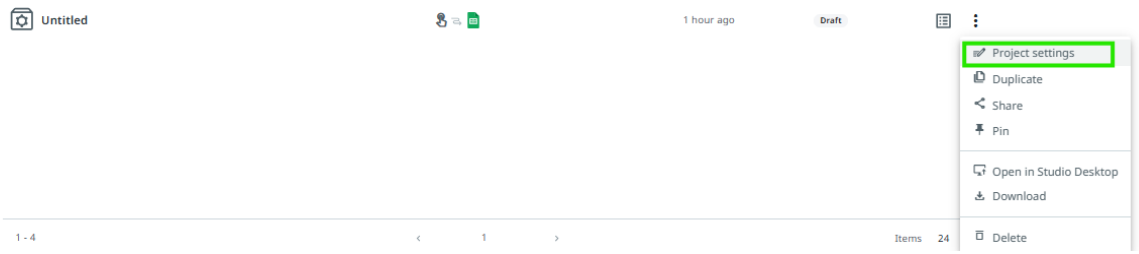


Step #3 Renaming the Project

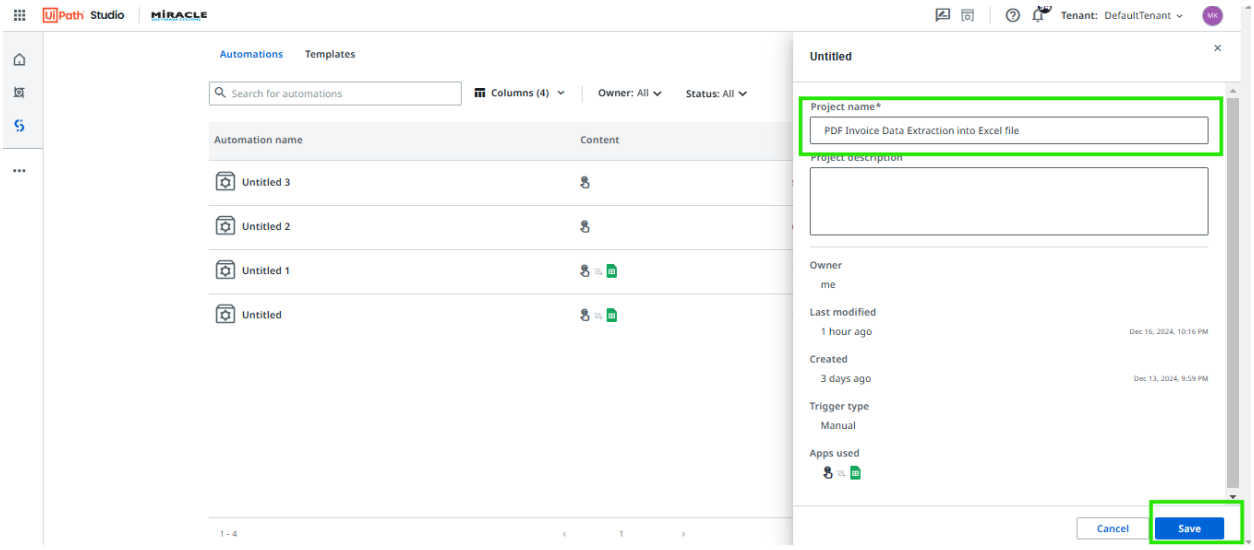
We need to rename our project so go back once, as highlighted in the below.



Go to Project settings by clicking on the 3 dots in the top right corner.



Enter the **Project Name** (Ex: PDF Invoice Data Extraction into Excel) and hit **Save**

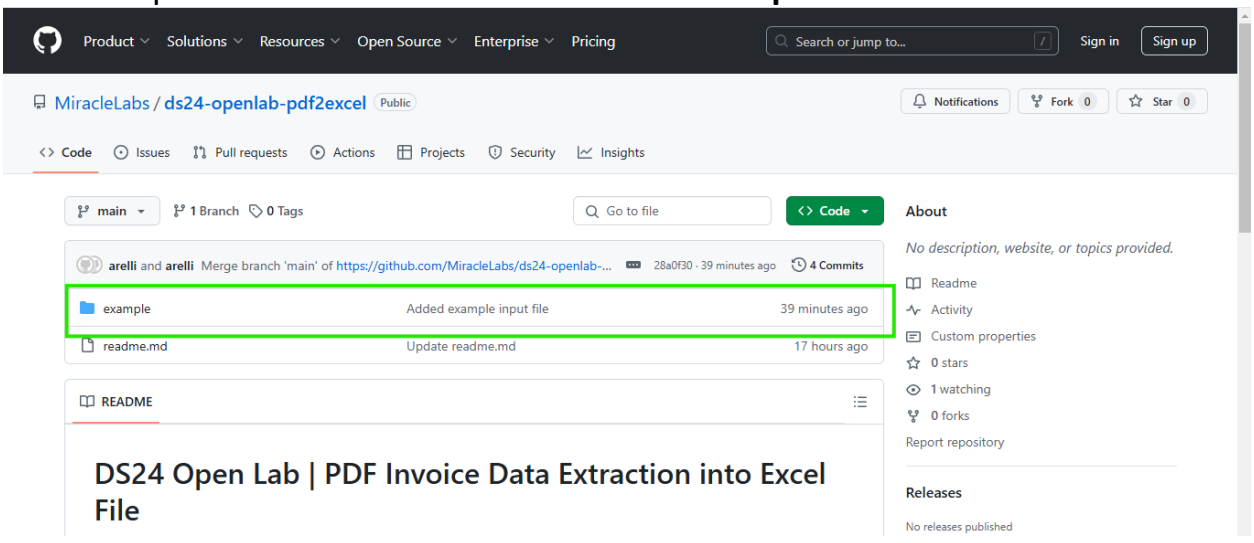


Step #4 Getting input file

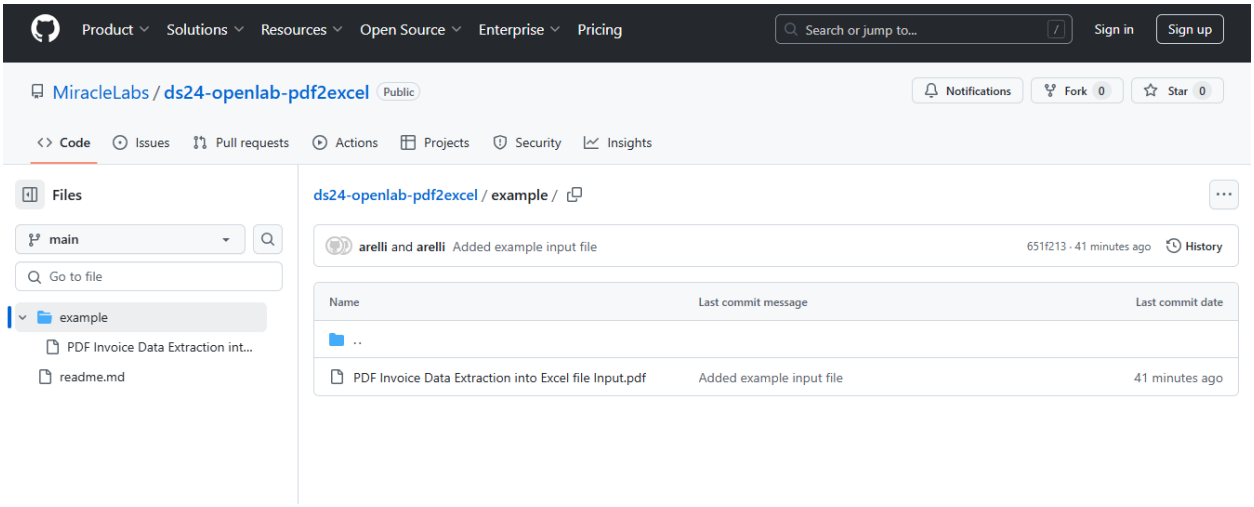
As we are developing the code in studio web, we need only web applications to automate the process.

Before that, we need to get the input file from the GitHub repository from the following link https://github.com/MiracleLabs/ds24-openlab-pdf2excel and then we need to upload the input file to storage buckets in the orchestrator and access it.

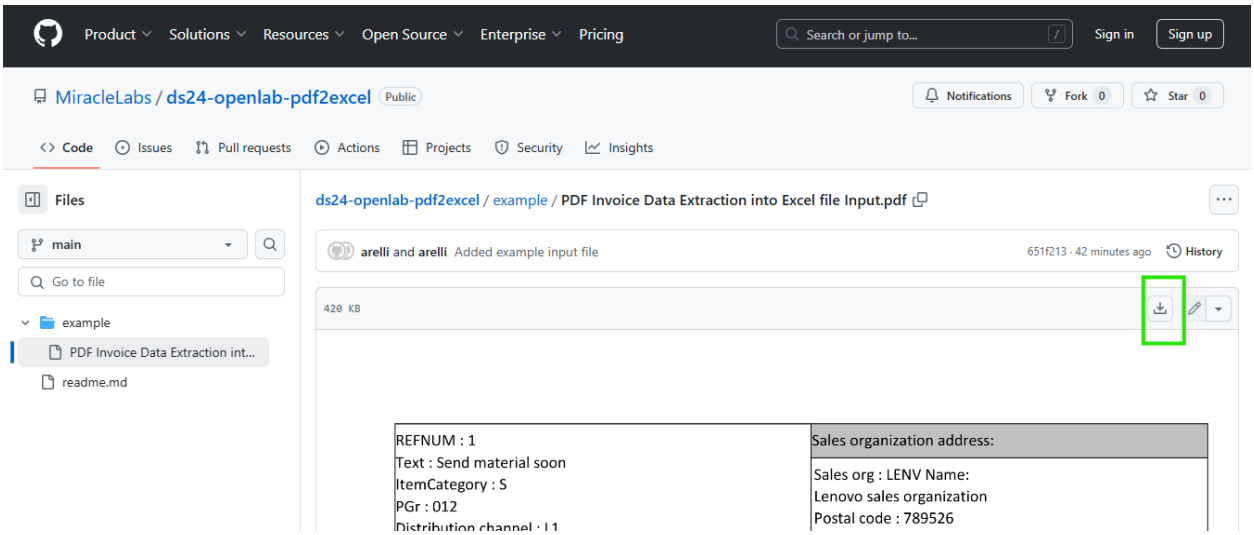
For that open the above link and click on the **Example file**



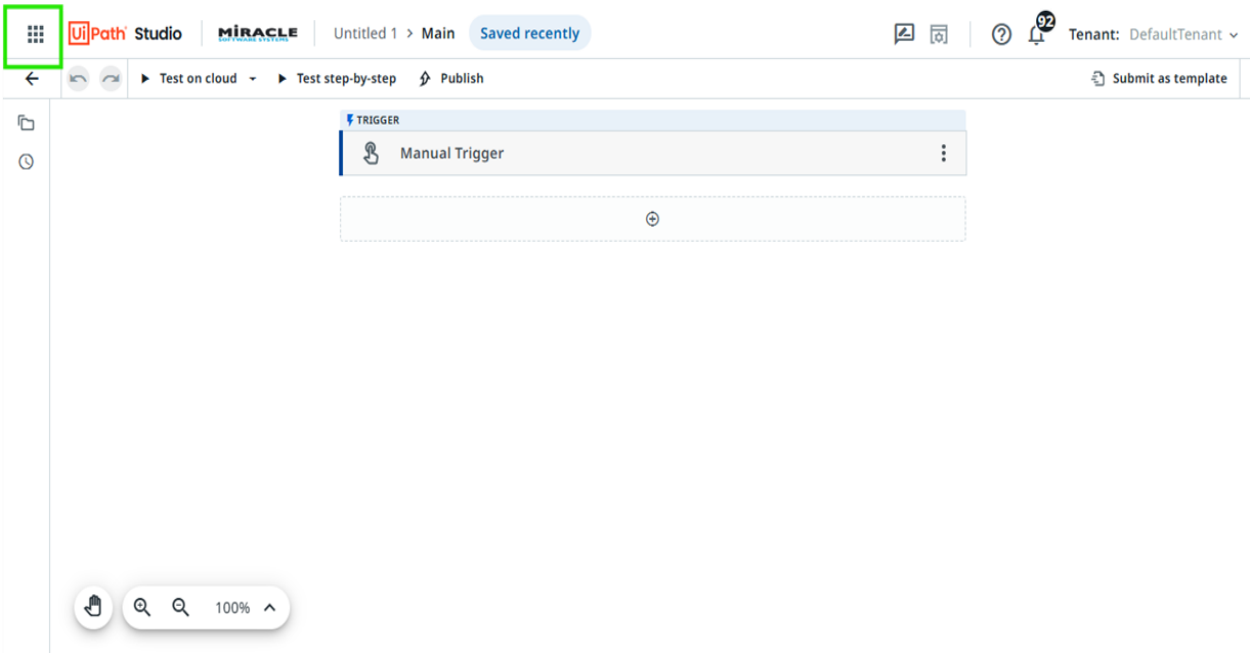
and then Click on PDF input file



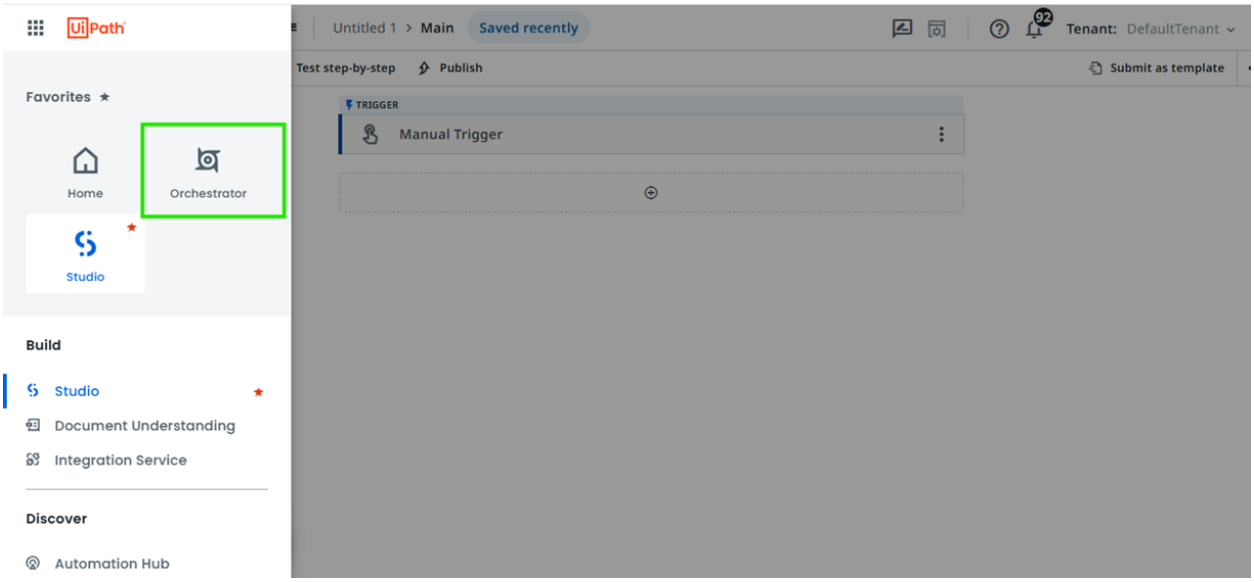
Click on download symbol to download the file



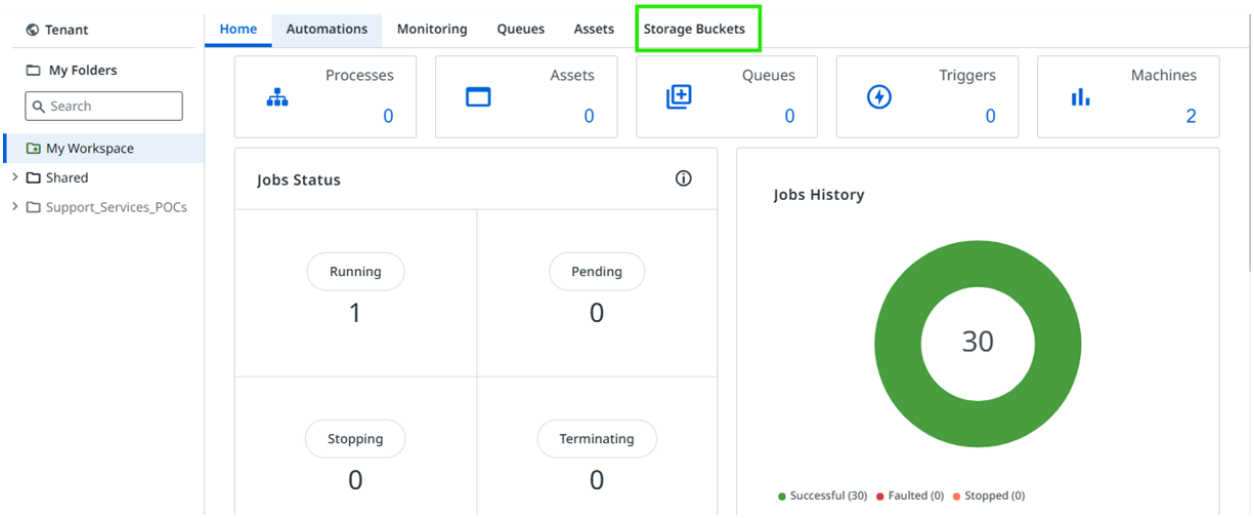
We need to go to Orchestrator, so click the menu as shown below.



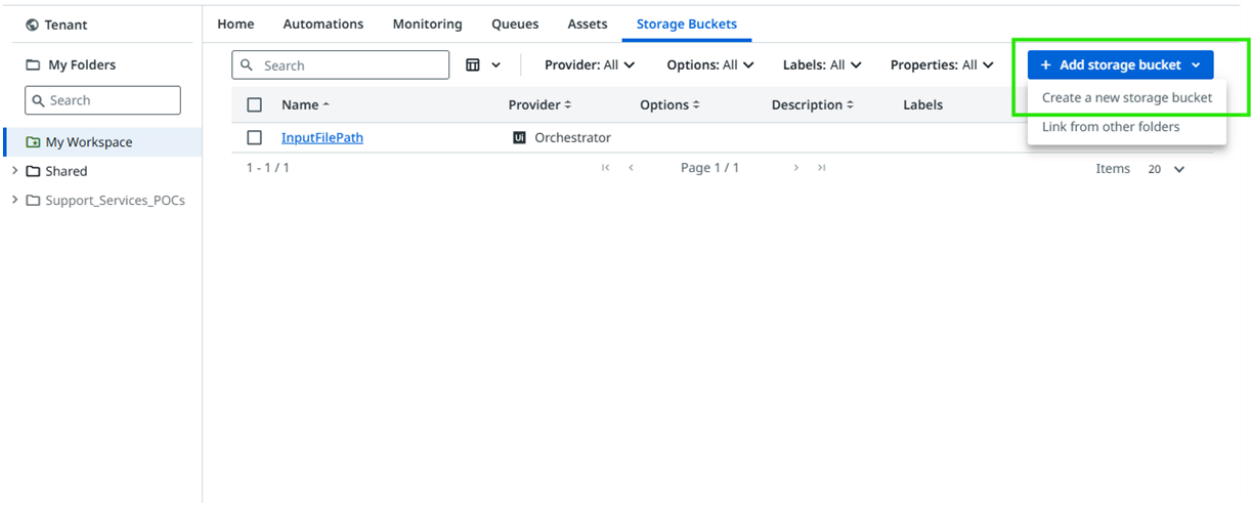
Click on the **Orchestrator** as shown.



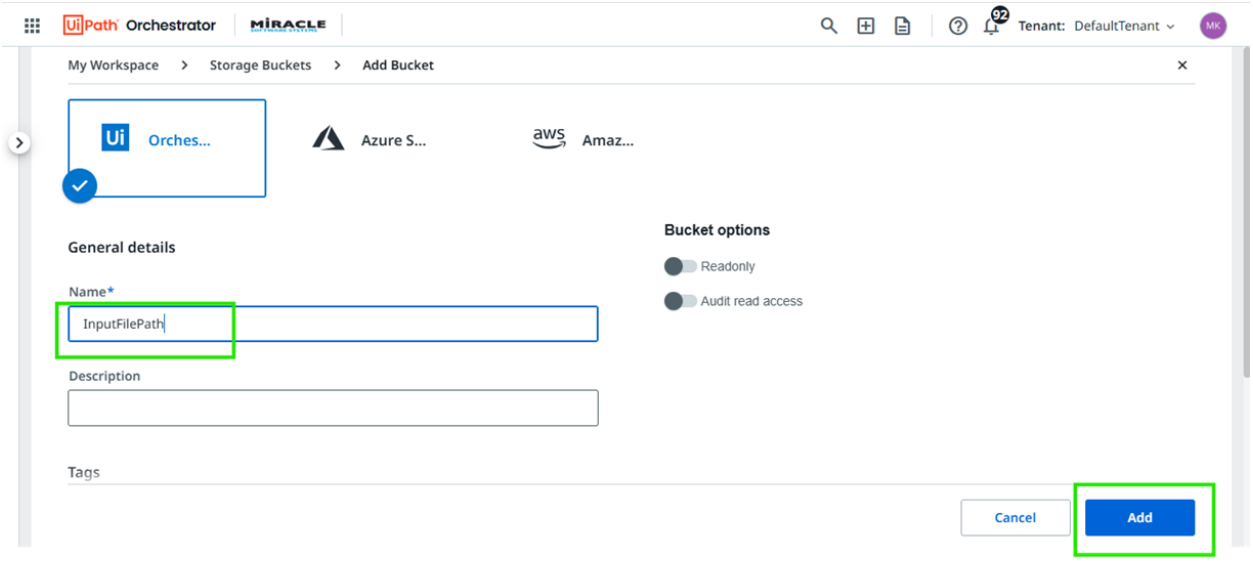
Now go to **Storage Buckets**.



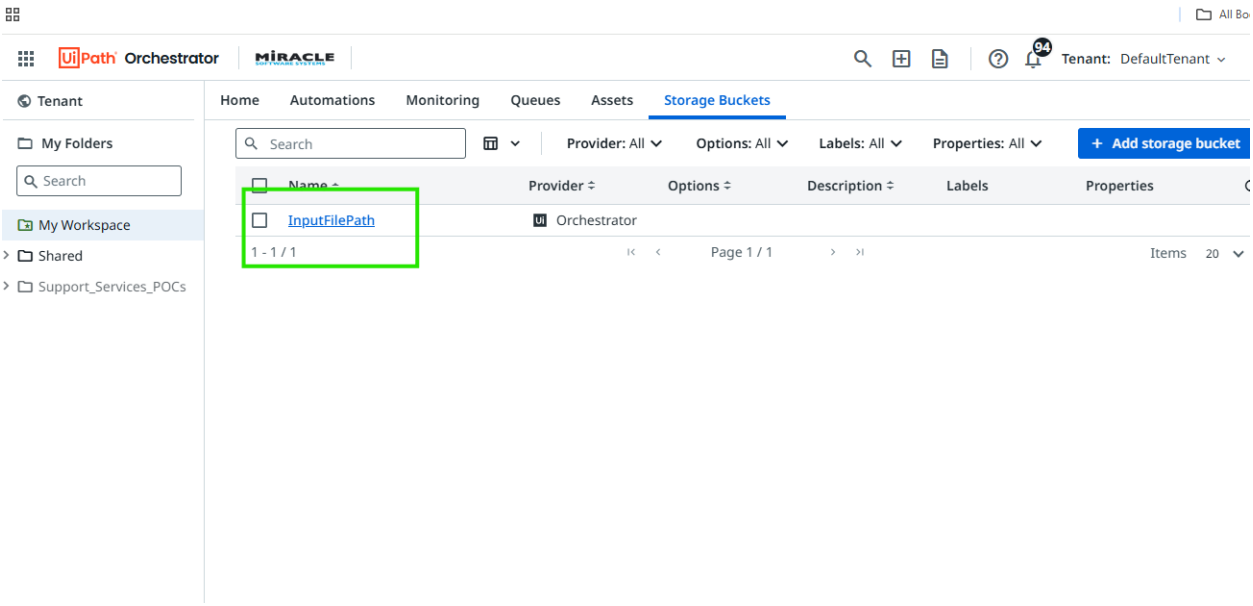
Click on the **Add Storage Button** and create a new storage bucket.



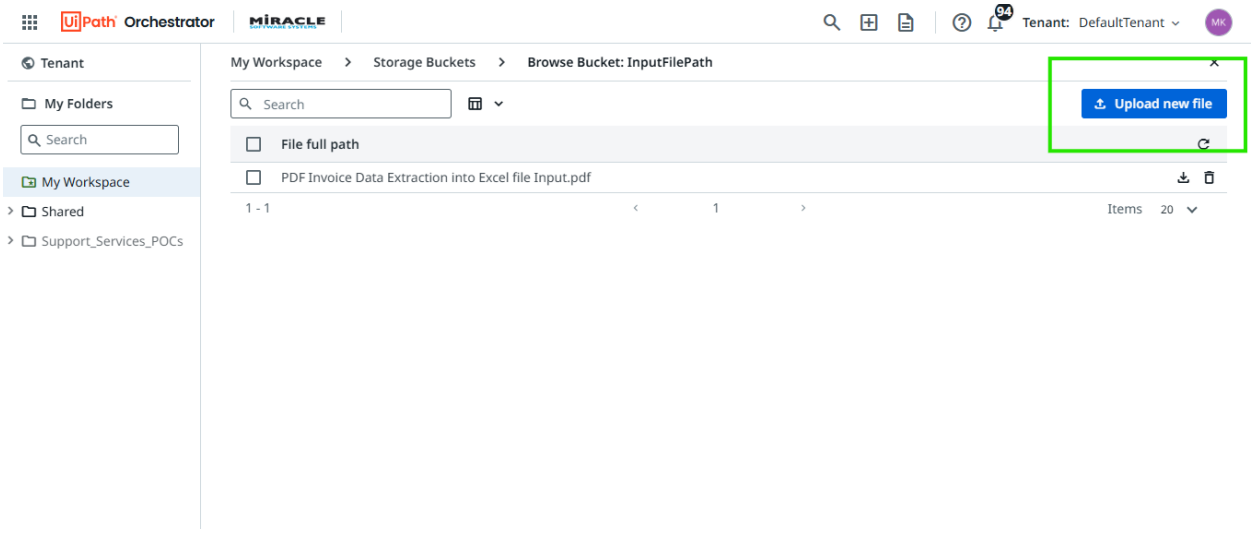
Provide the name for Storage Bucket (ex : InputFilePath)



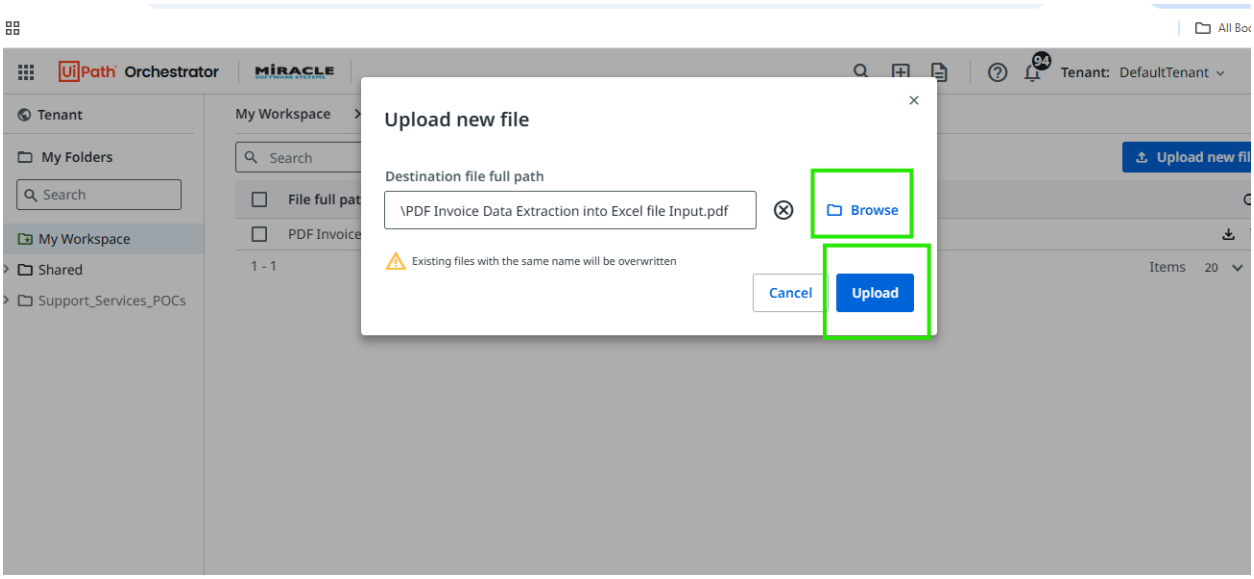
Click on the **Storage Bucket** name (ex: InputFilePath)



Click on **Upload new file** as highlighted below.



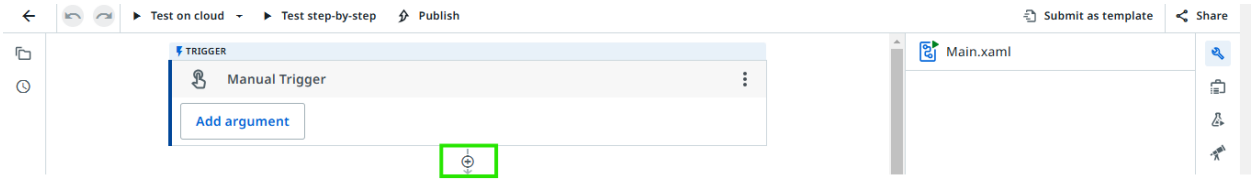
Click on **browse** and upload the input file then click **Upload**.



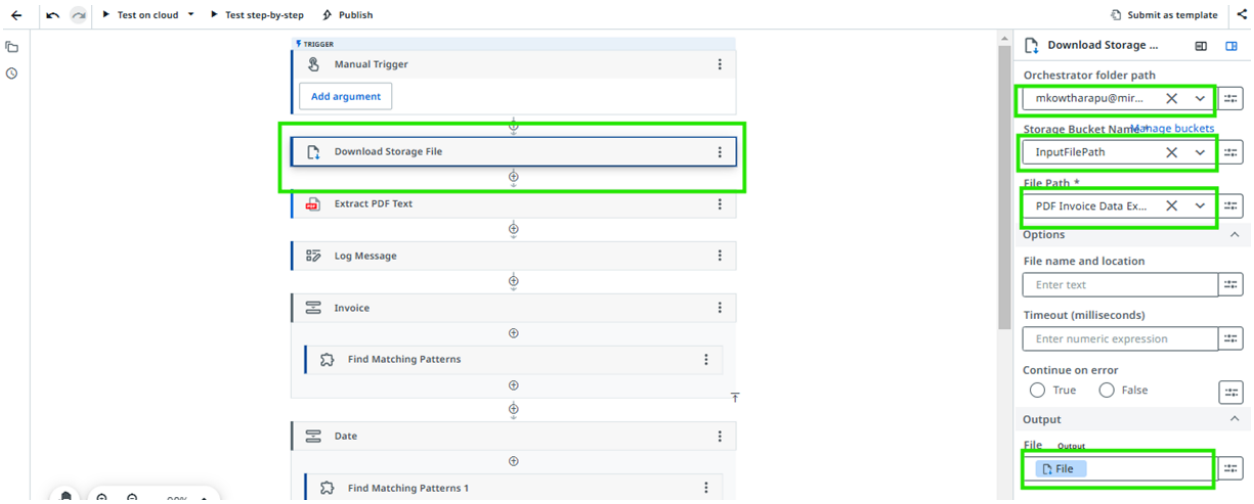
Step #5 Downloading input from Storage Bucket

For this we should get back to Studio Web first

and then click on the plus icon and search for **Download Storage File** activity.



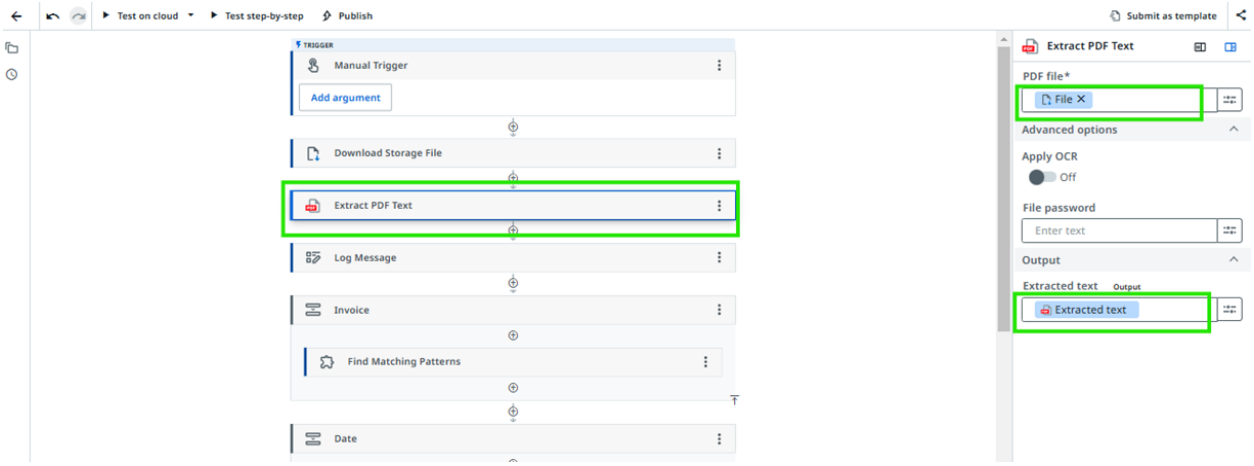
Under the properties panel provide Orchestrator folder path, Storage bucket name, Filepath and by providing all these we will be getting the output (ex: File)



Step #5 Extracting data from PDF

Search for the **Extract PDF Text** activity, select it, and in the properties panel provide the input file which is the output of **Download Storage File** activity.

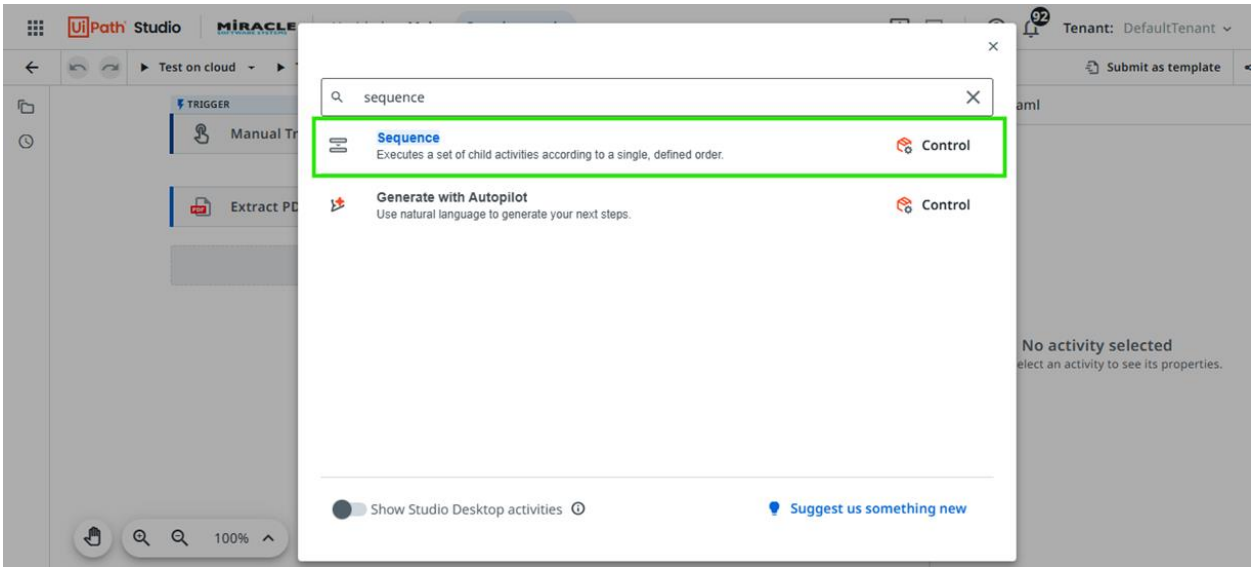
We will be getting an output (ex: Extracted Text).

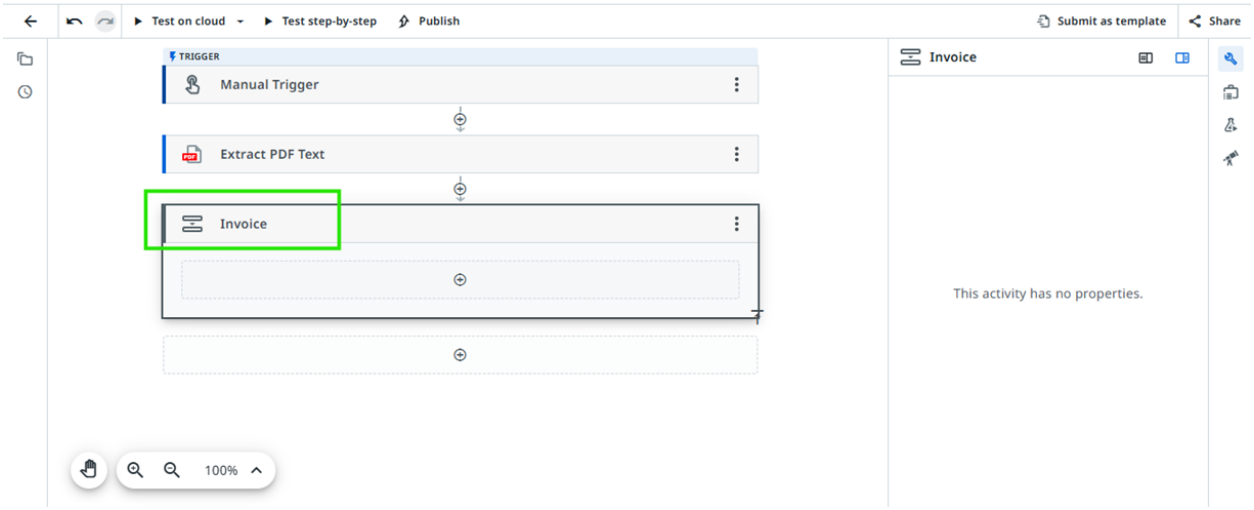


Step #6 Creating Sequence

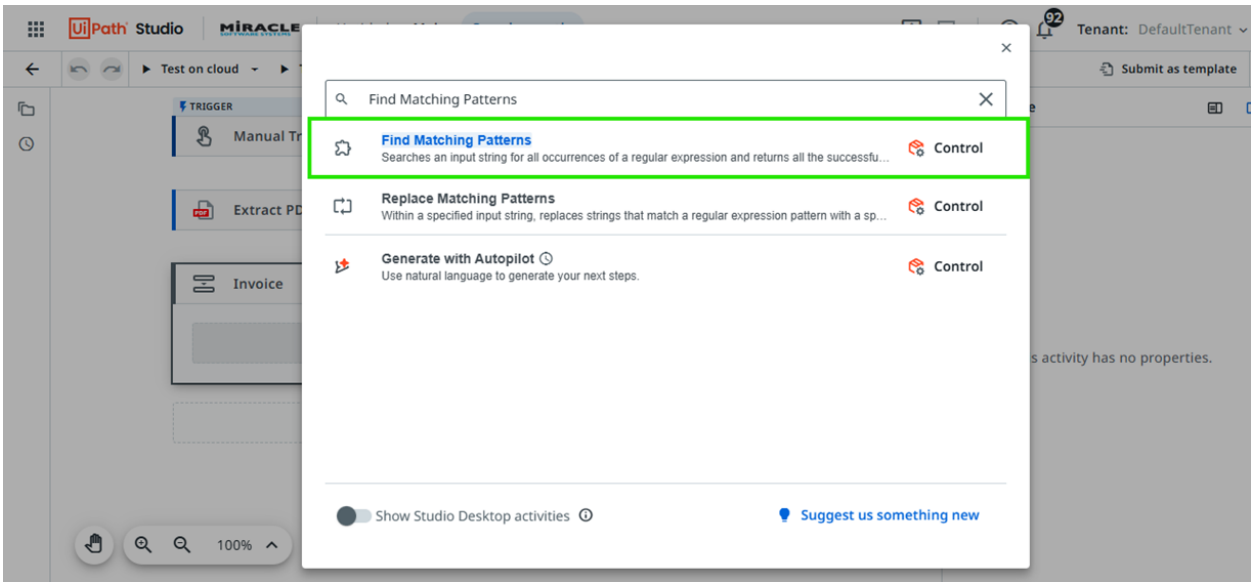
We need to create sequences for each field we are extracting. At first for **Invoice**.

Search for **Sequence**, select it, and rename it.

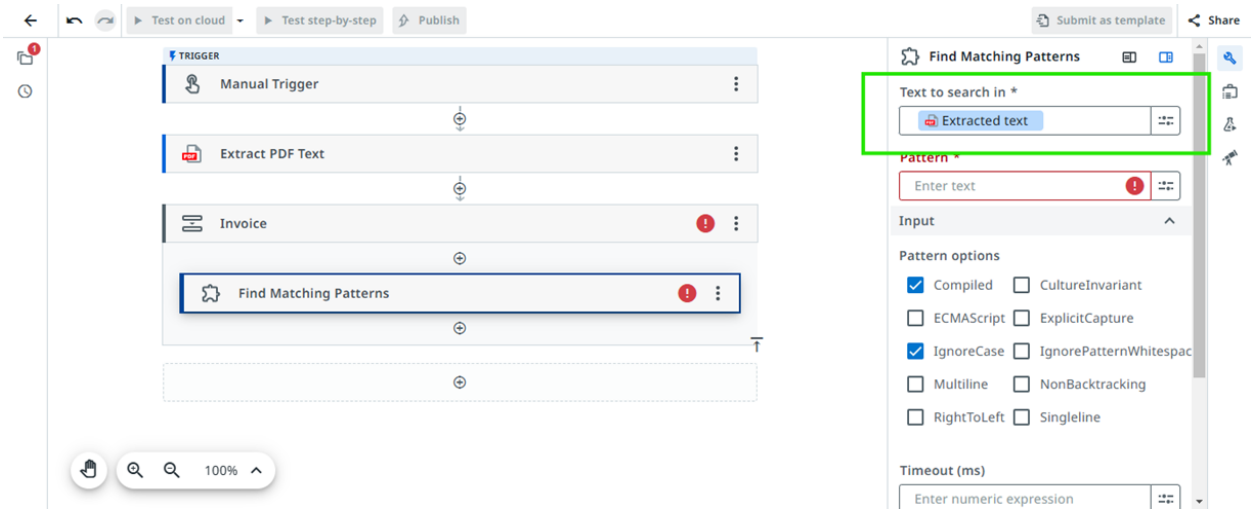




In the invoice sequence search for **Find Matching Patterns** activity and select it.



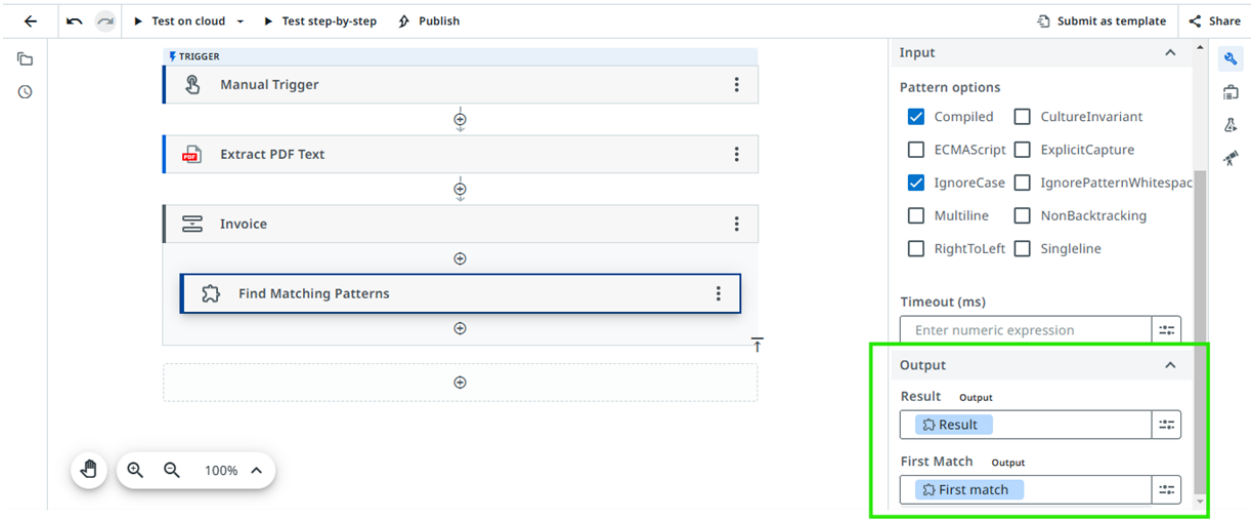
Go to the properties panel and provide the text to search in (ex: extracted text)



Provide the pattern i.e. regex to match with invoice number.

ex: (?<=InvoiceNumber)\d.+

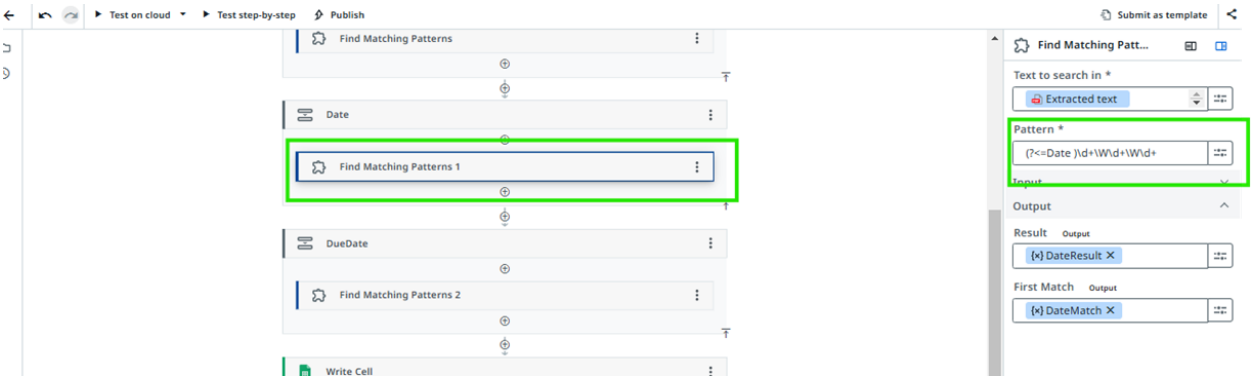
We will be getting the outputs as result and First match as highlighted below.

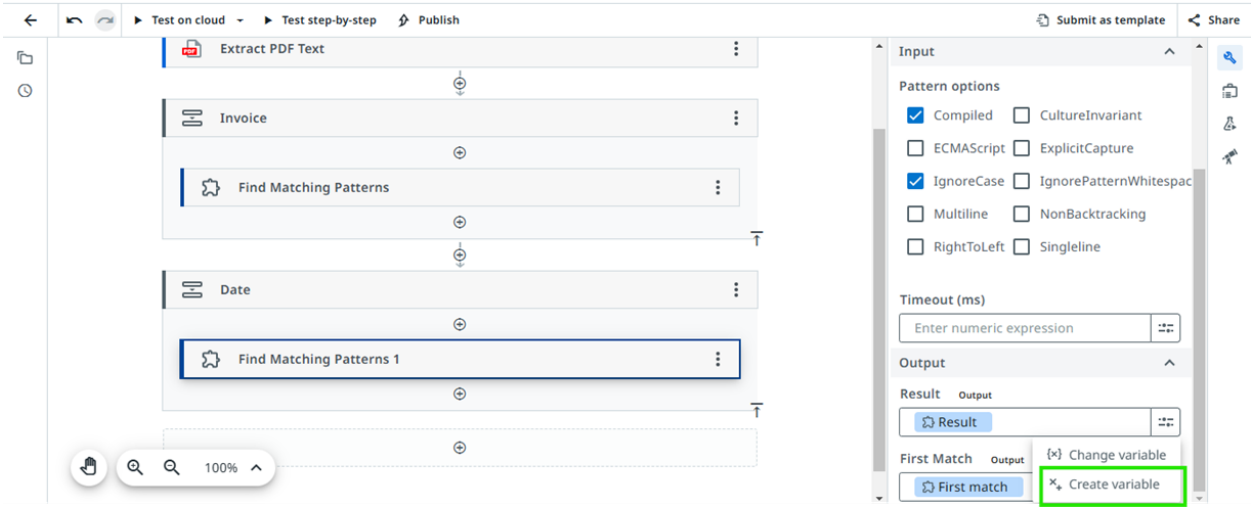


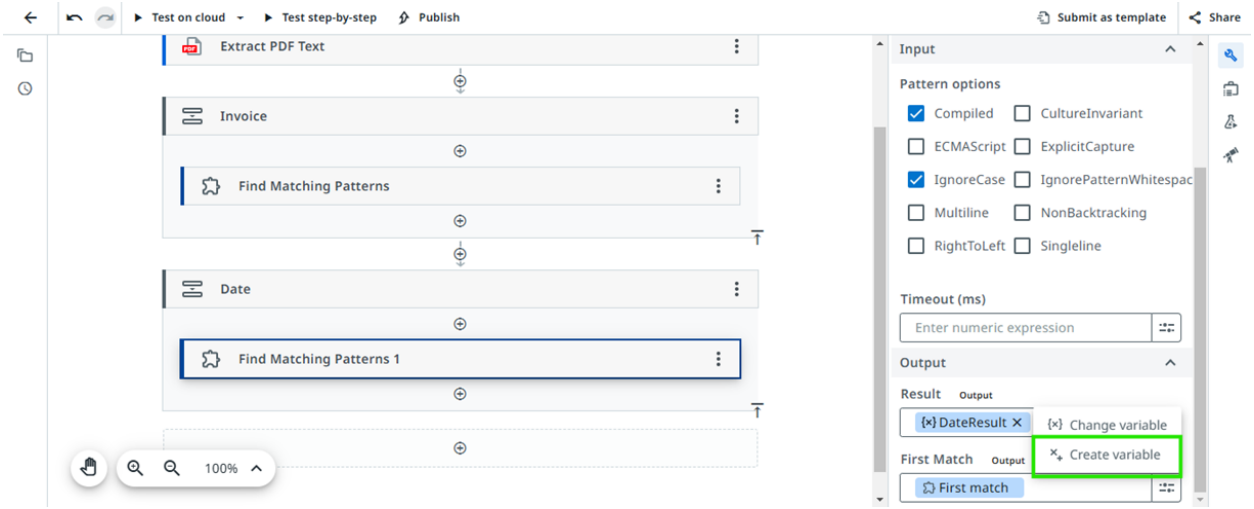
We need to follow the same procedure for extracting Date and DueDate fields

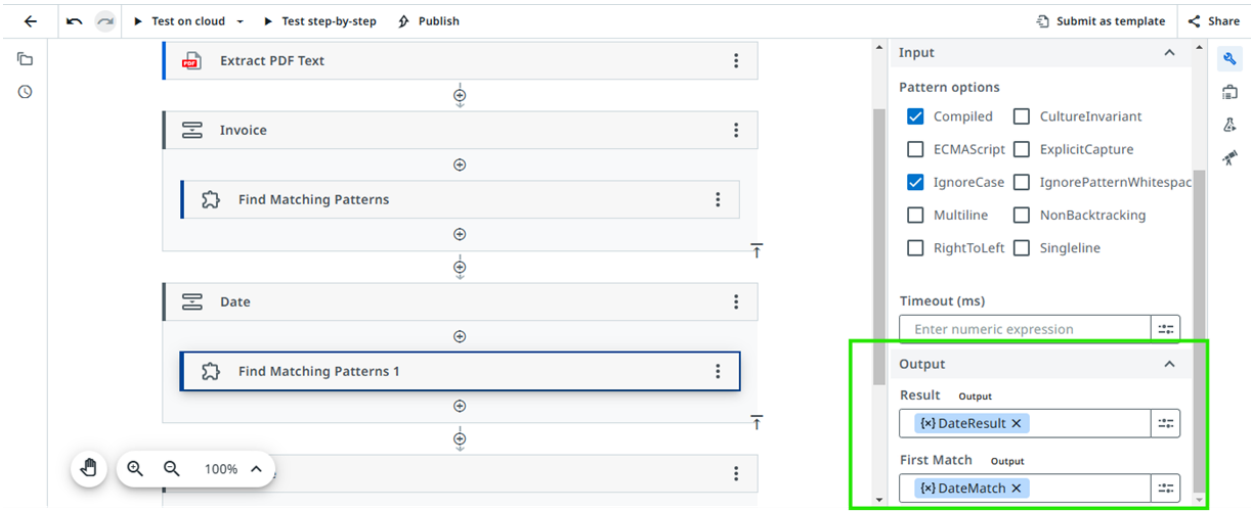
Regex for Date :(?<=Date)\d+\W\d+\W\d+

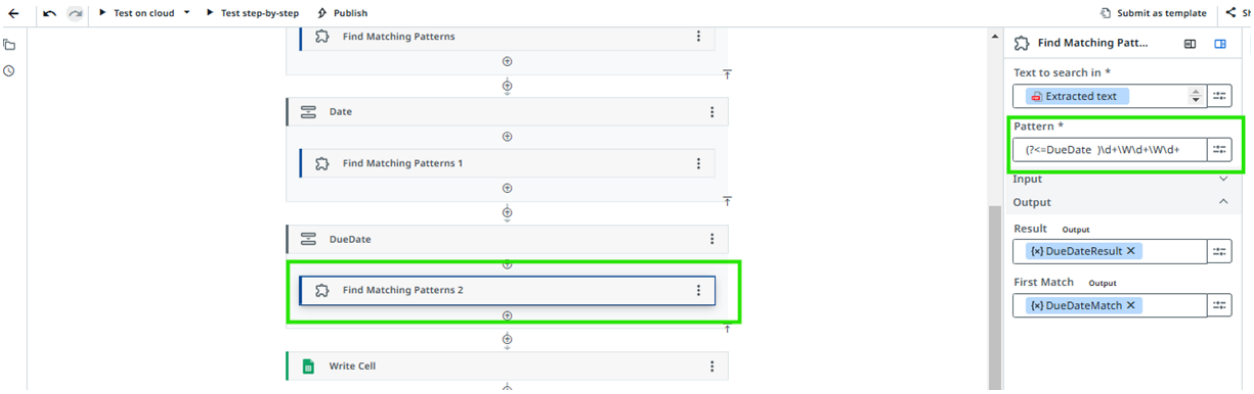
Regex for DueDate : (?<=DueDate)\d+\W\d+\W\d+




Test on cloud
Test step-by-step
Publish
Submit as template
Share
Extract PDF Text
Invoice
Find Matching Patterns
Date
Find Matching Patterns 1
100%
Input
Pattern options
Compiled
CultureInvariant
ECMAScript
ExplicitCapture
IgnoreCase
IgnorePatternWhitespac
Multiline
NonBacktracking
RightToLeft
Singleline
Timeout (ms)
Enter numeric expression
Output
Result
Output
Result
First Match
Output
First match
Change variable
Create variable


Test on cloud
Test step-by-step
Publish
Submit as template
Share
Extract PDF Text
Invoice
Find Matching Patterns
Date
Find Matching Patterns 1
100%
Input
Pattern options
Compiled
CultureInvariant
ECMAScript
ExplicitCapture
IgnoreCase
IgnorePatternWhitespac
Multiline
NonBacktracking
RightToLeft
Singleline
Timeout (ms)
Enter numeric expression
Output
Result
Output
DateResult
Change variable
First Match
Output
Create variable
First match



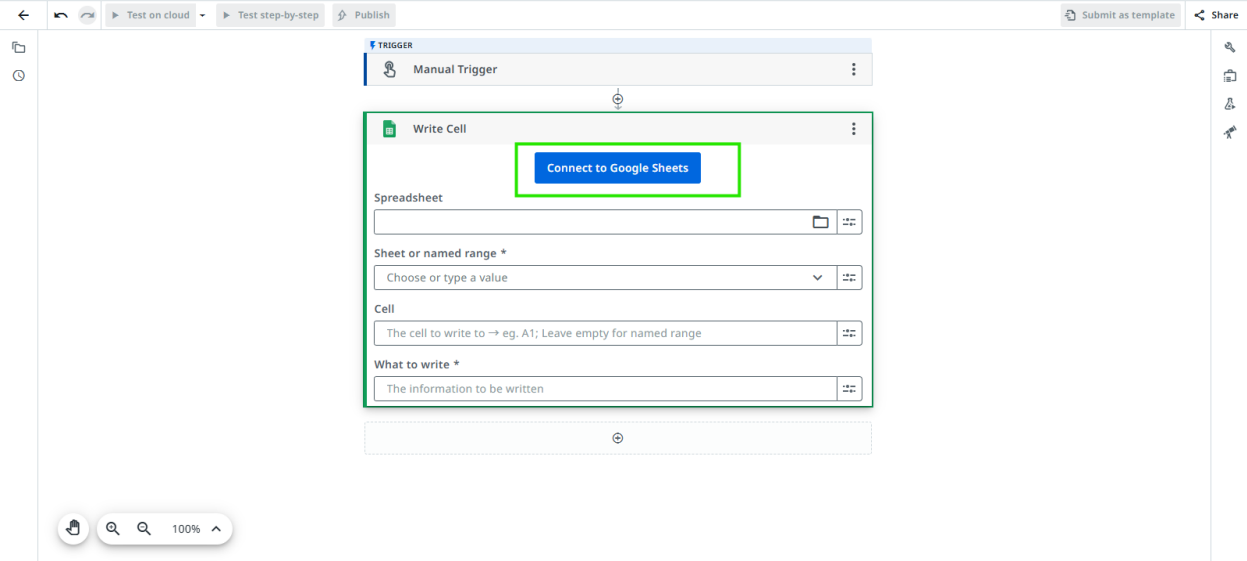


Step #7: Get the Output

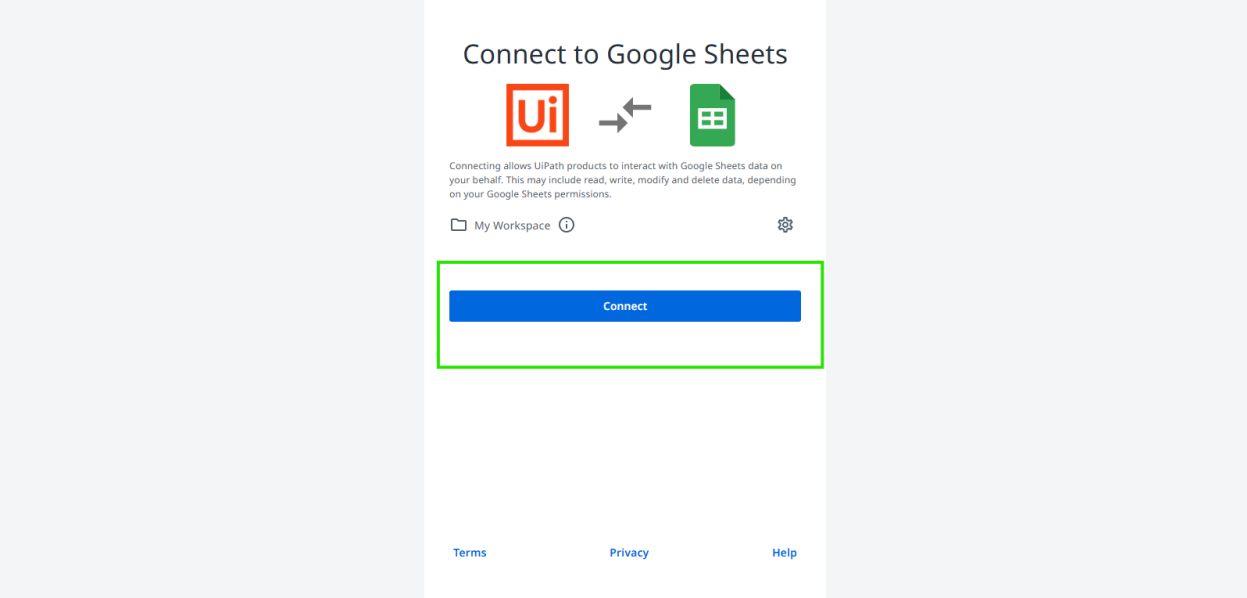
To get the output we need to write the extracted information to google sheet.

Select **write cell** activity and it will ask us to connect to Google Sheets.

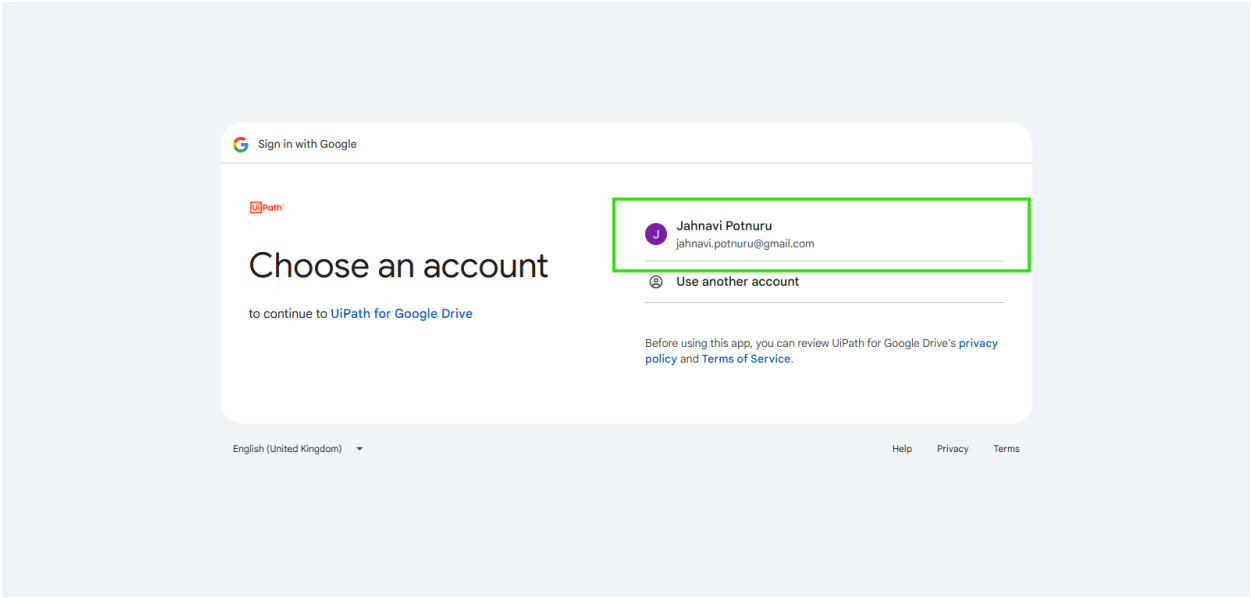
For that, we need to connect this Uipath to Google Sheets by clicking on button as highlighted.



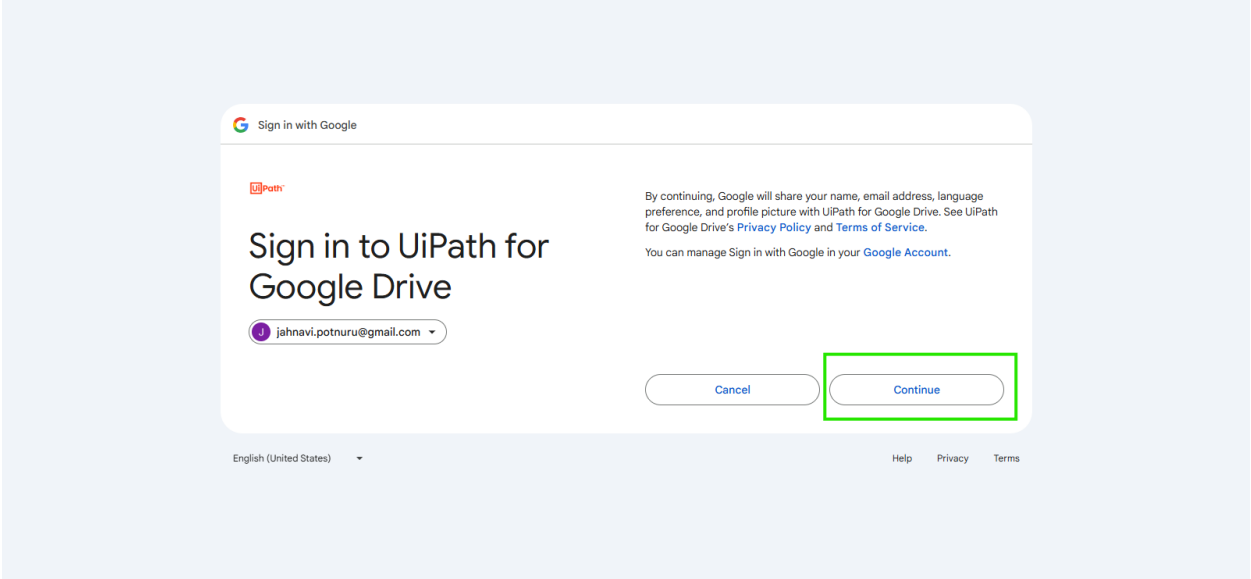
Click on connect.



Choose your Gmail account to connect.



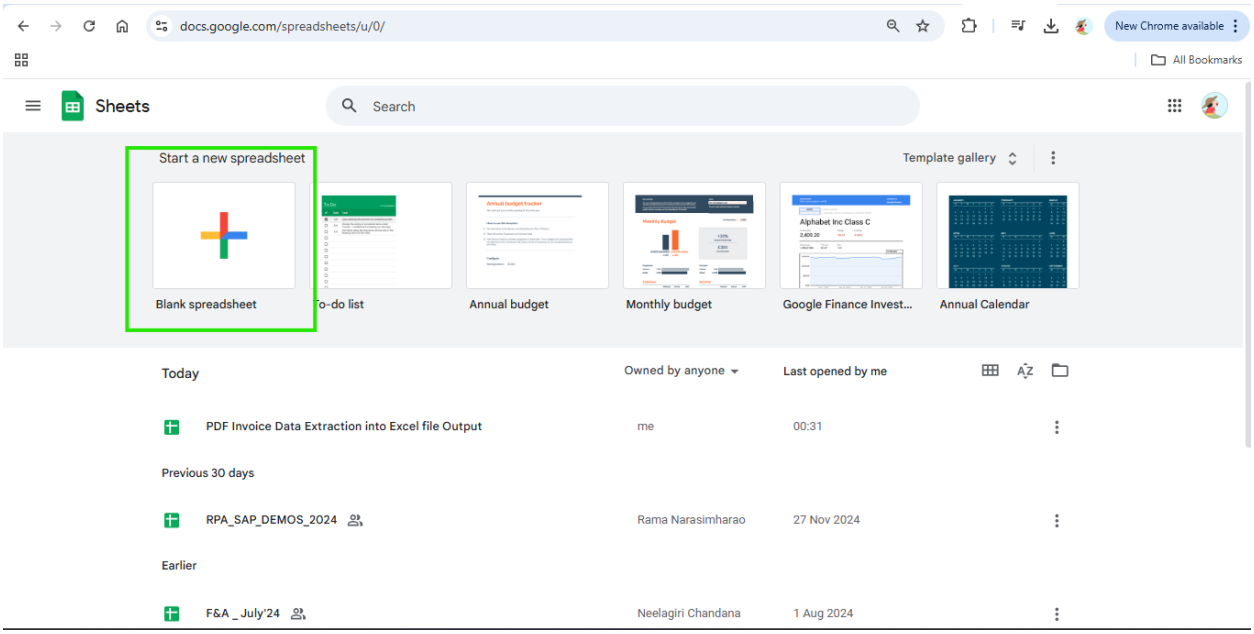
Click on continue.



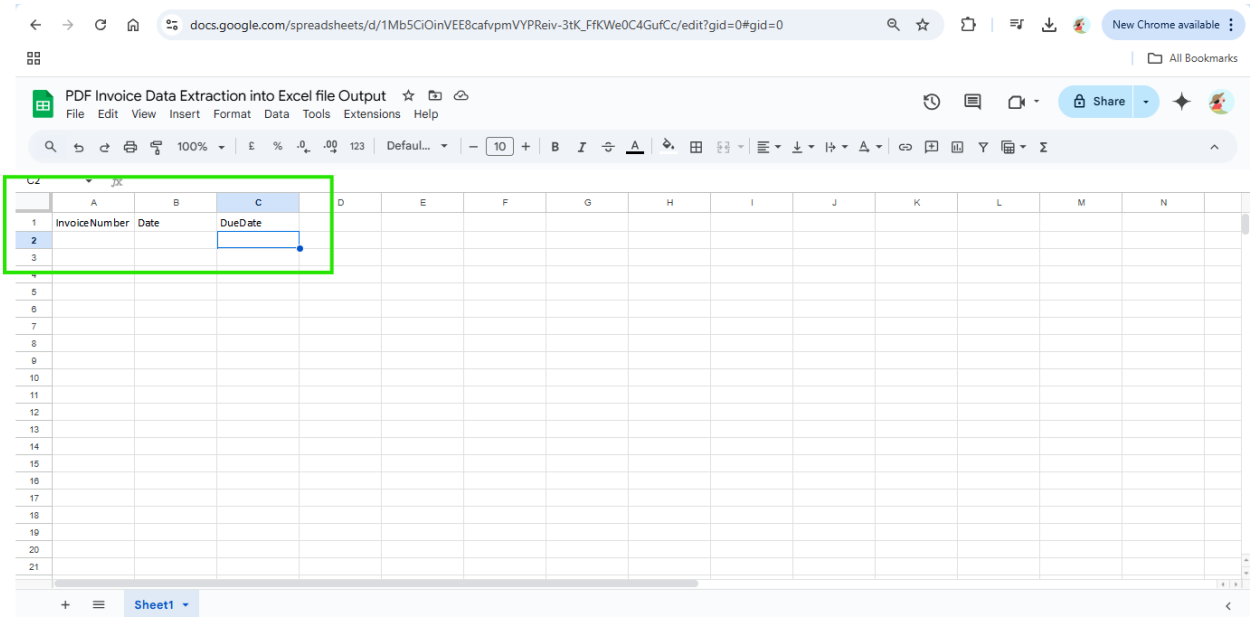
Thus we connected UiPath with google drive.

Step #8 Creating a Spreadsheet

Go to Google Spreadsheets and click on blank process to create a new spreadsheet for our Output.

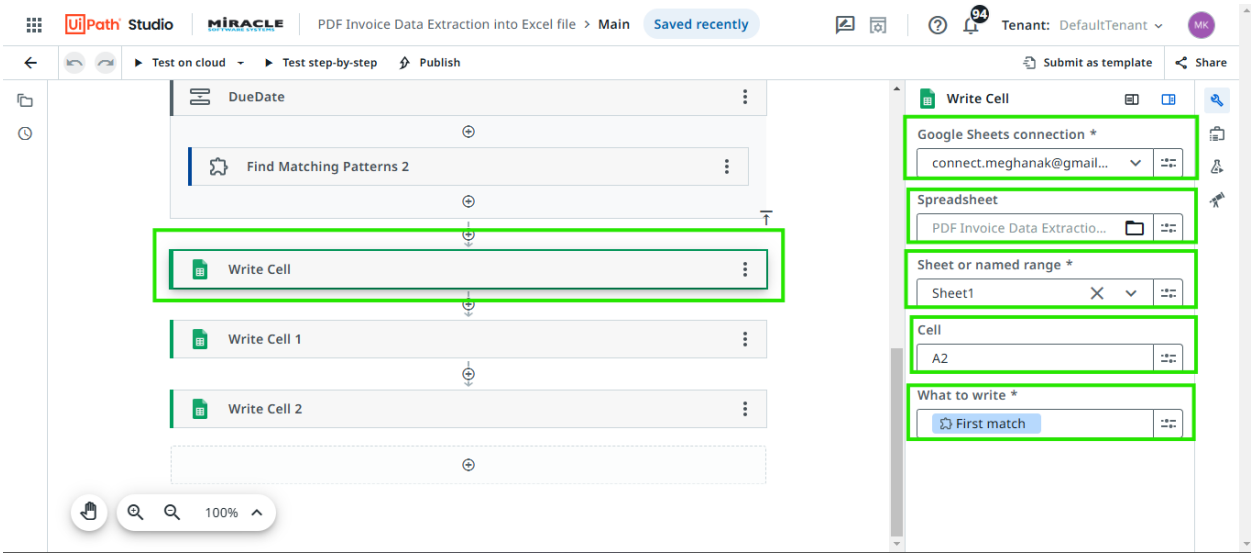


Create three columns i.e. Invoice number, Date and DueDate.



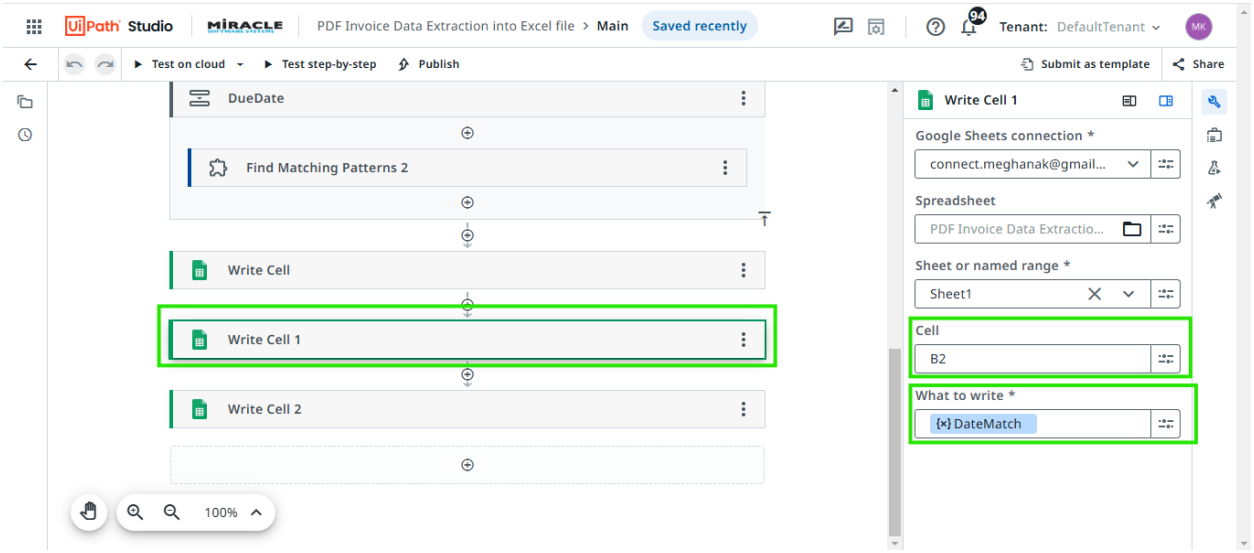
Step #9 Writing to Spreadsheet

Go to Write Cell activity properties and provide the information like the mail which we connected to uipath, Spreadsheet name, sheet name, the cell (ex: A2) which we are writing to and finally provide what we are writing (ex: FirstMatch) ie. the output of the invoice.

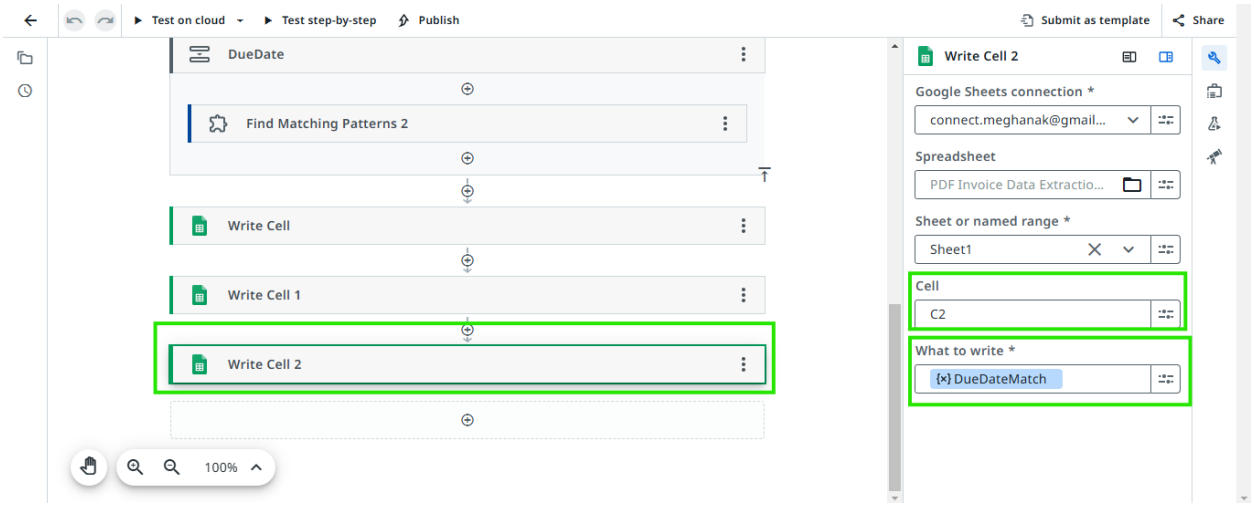


Same procedure for the date and due date too.

Mention cell and what to write properly (ex: for date cell: B2 what to write: DateMatch)

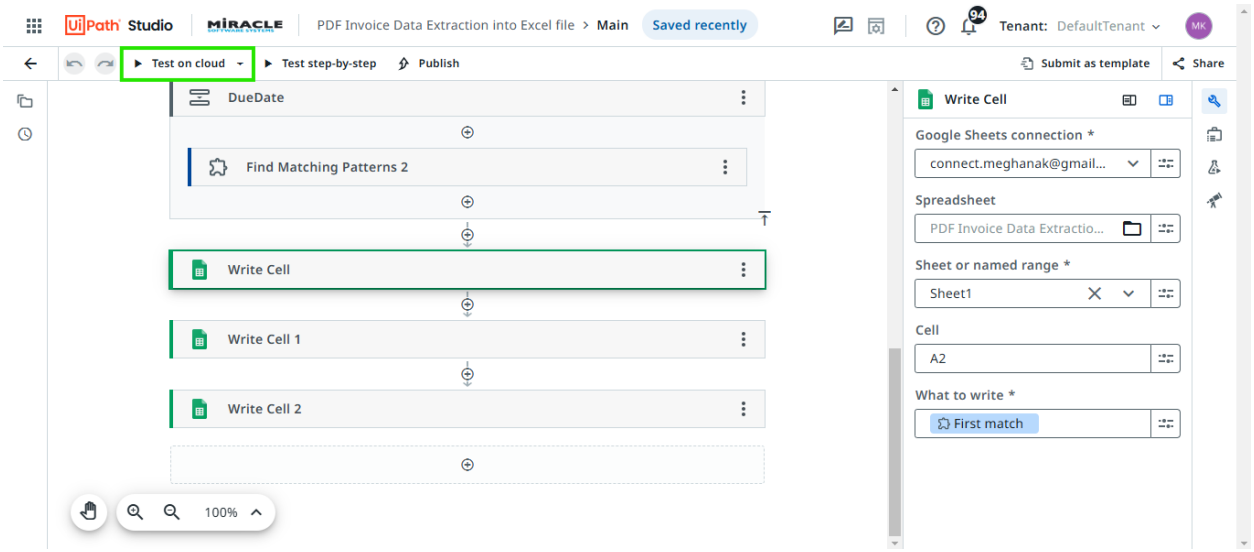


(ex : for due date cell: C2 what to write: DueDateMatch)



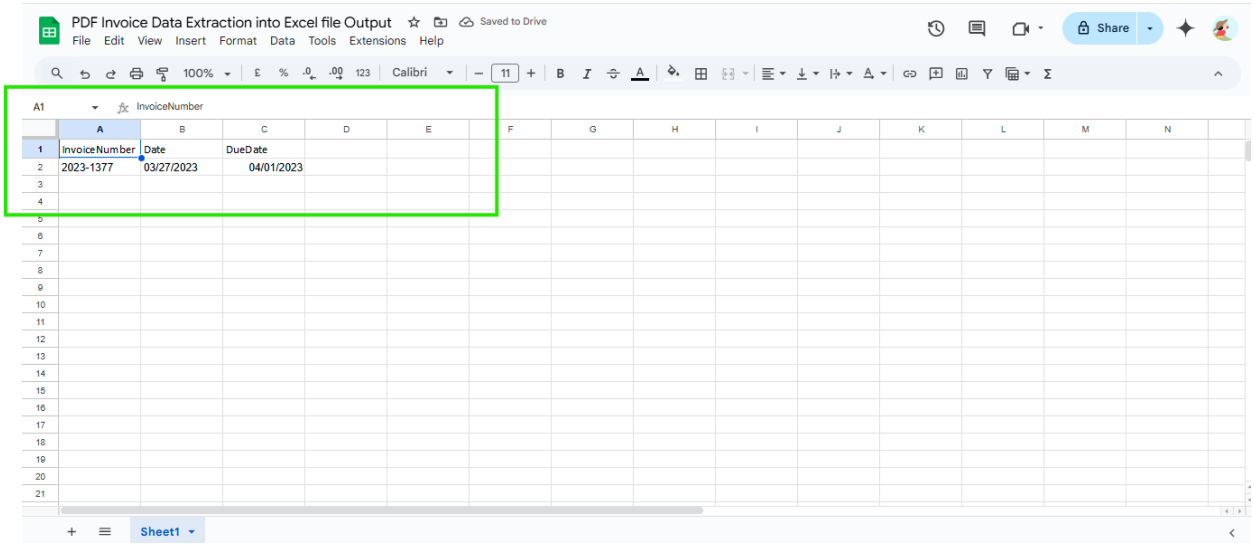
Step #9 Run the code

Click on **Test on Cloud** to run the code.



Step #10 Output

We will be getting the output like this.



Hope you enjoyed this Automation!