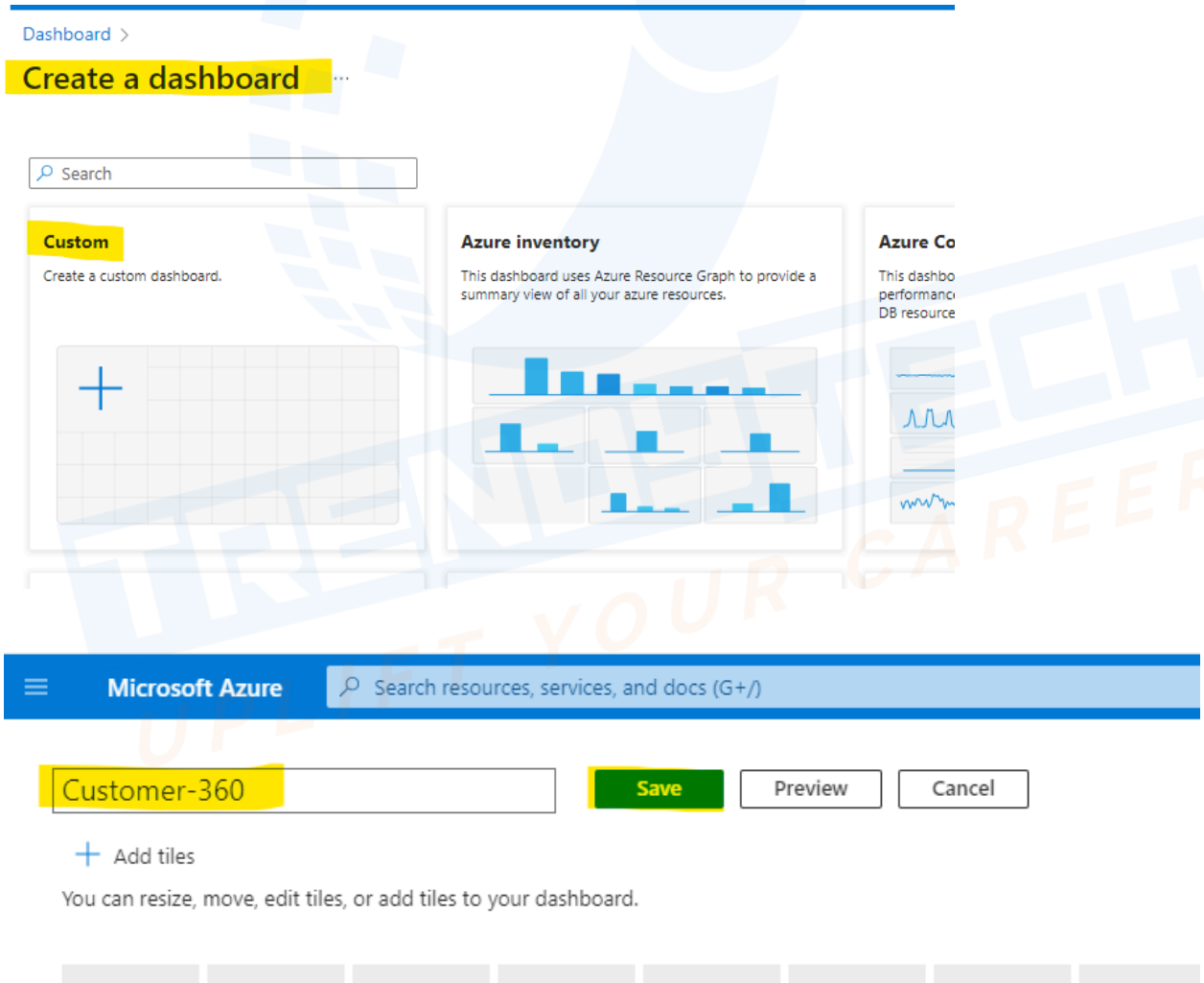
Use-case3: Ingesting the data from Source(Blob Storage) updated by third party vendors regularly, transferring and loading the data to Sink(ADLS GEN2)

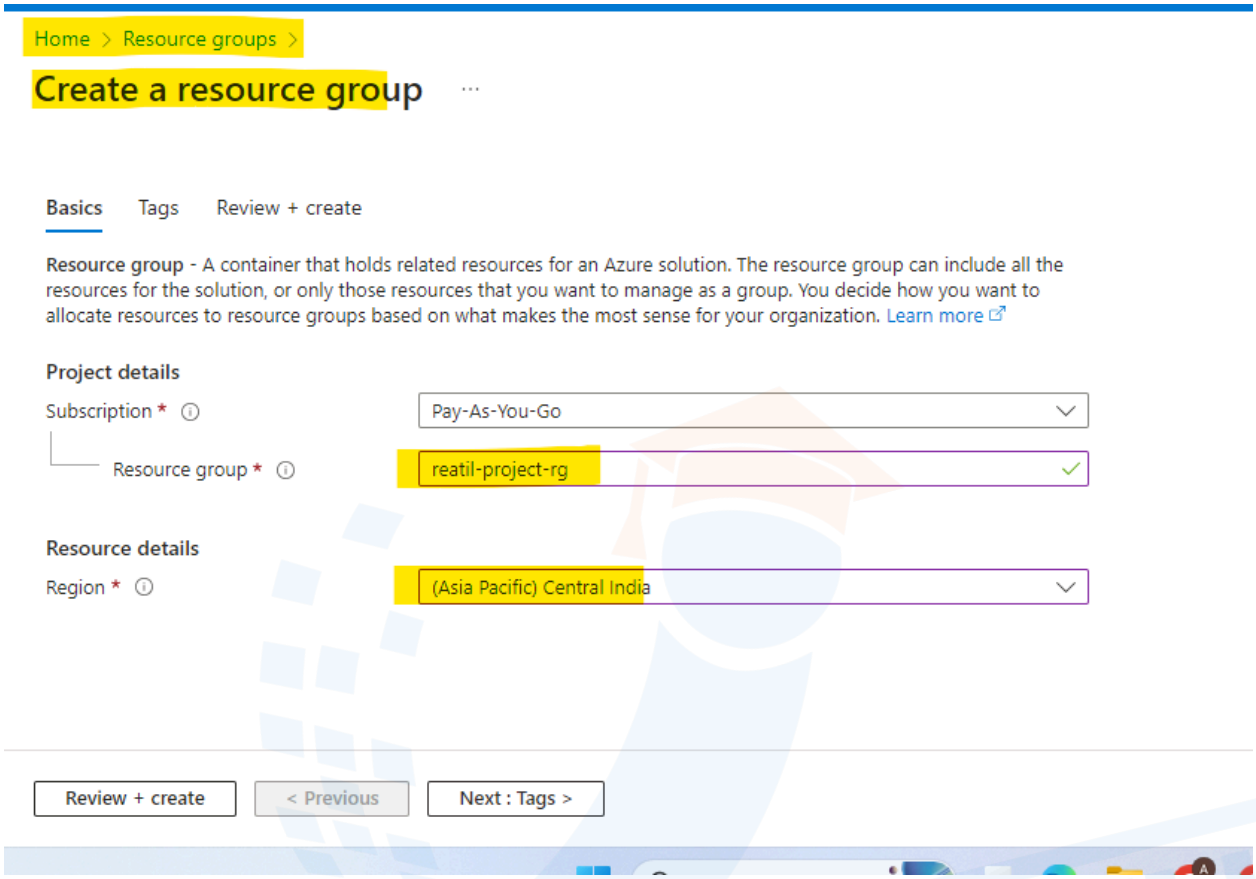
Requirement:
Products.csv will be updated by third party vendors regularly. Currently in the blob storage and we need to bring the products dataset to Datalake.

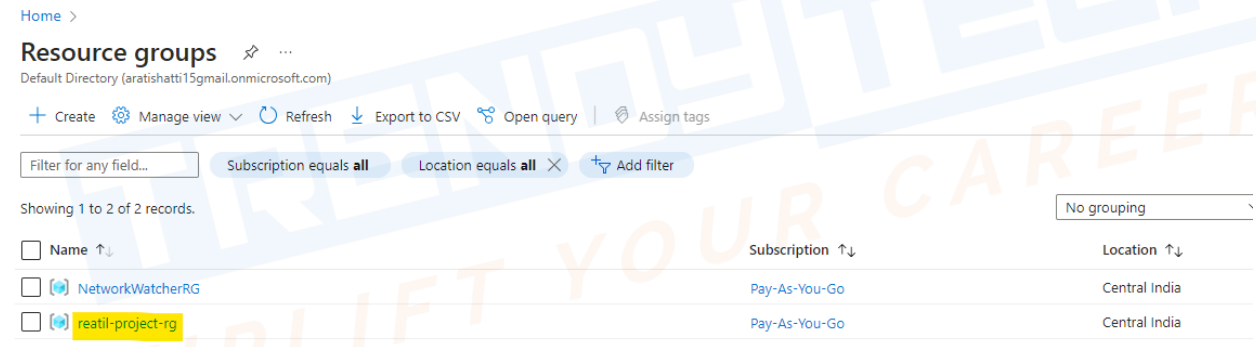
Note: You can create a Dashboard for the project to organize the resources related to the project at one place.

Click on Dashboard => Create => Custom =>



Create Resource Group :

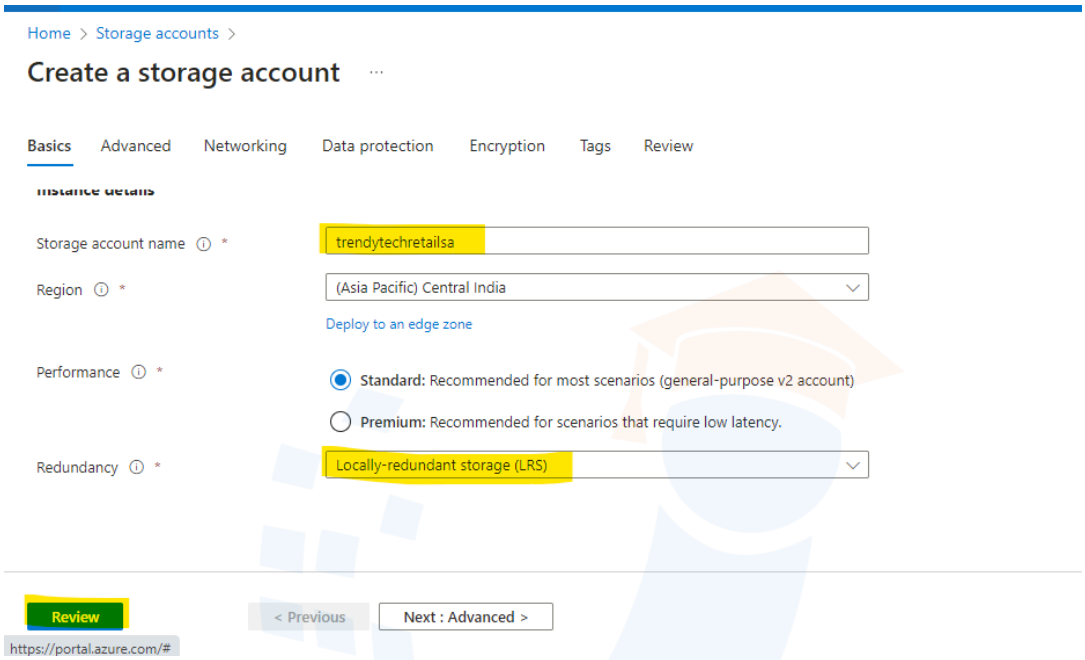




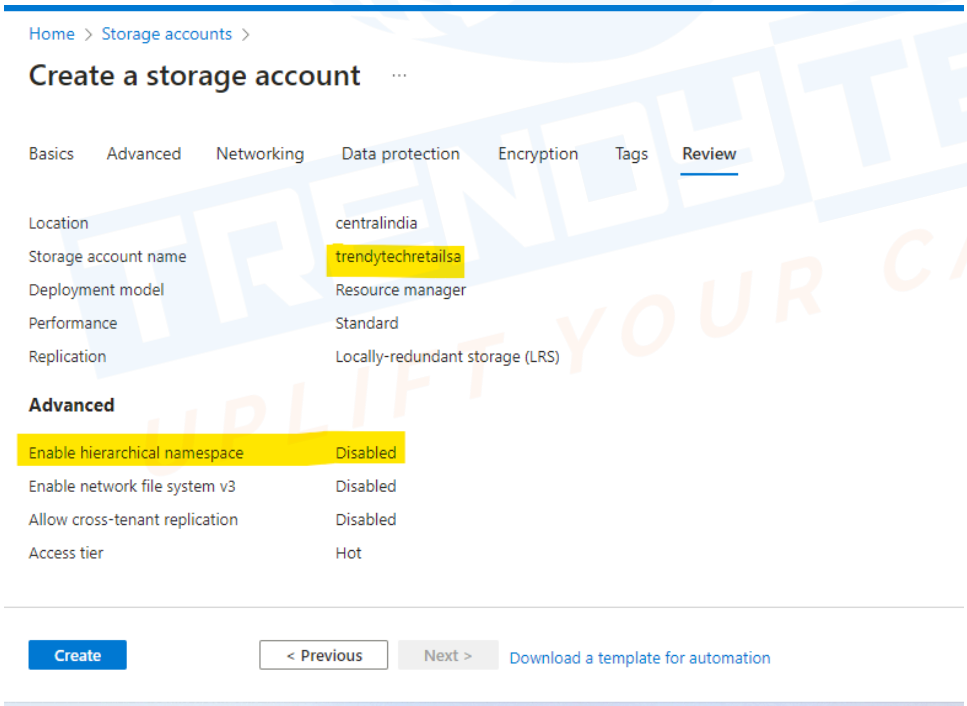
Note: Once a resource has been created, you have the option to add it to the dashboard for quick access.

Create Storage Account Resource - **“**trendytechretailsa”

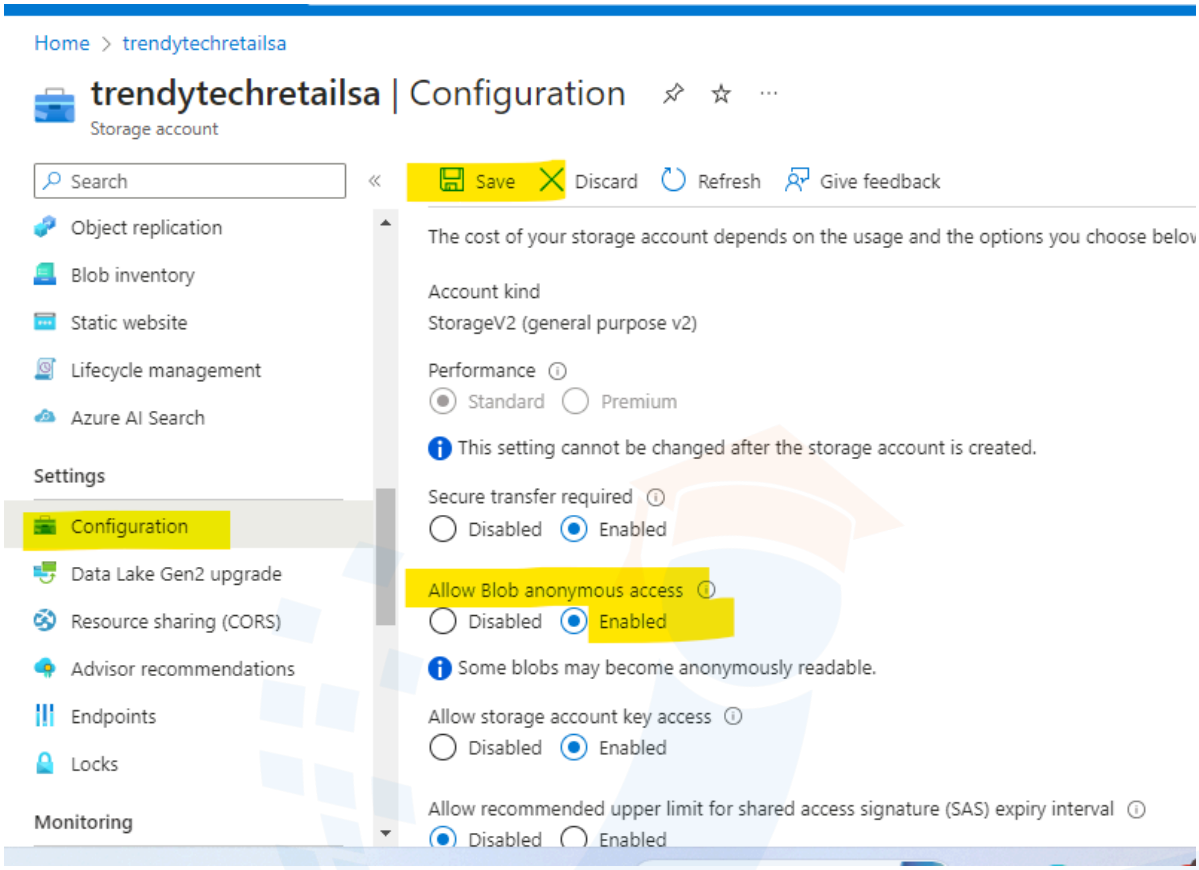
To create a Storage Account blob storage follow below steps:



Note: In case of blob storage Hierarchical namespace will be disabled.

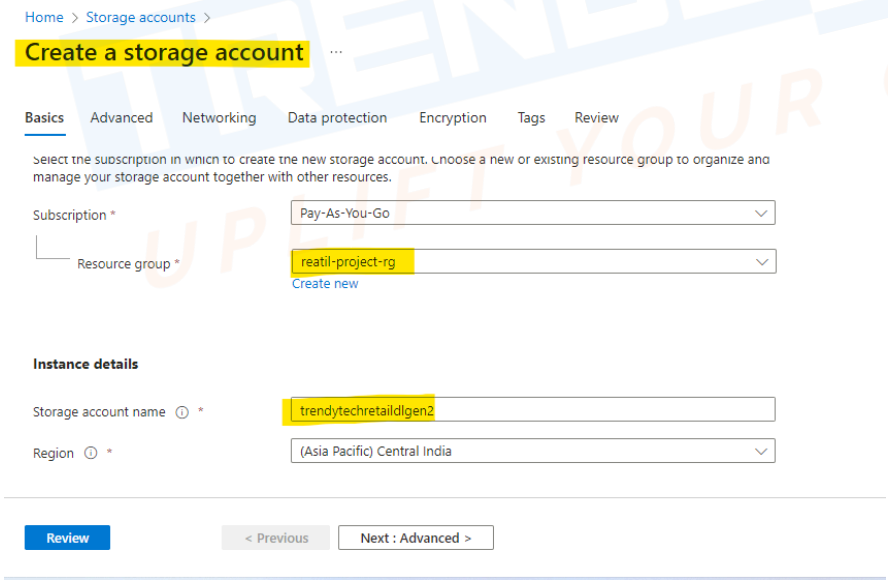


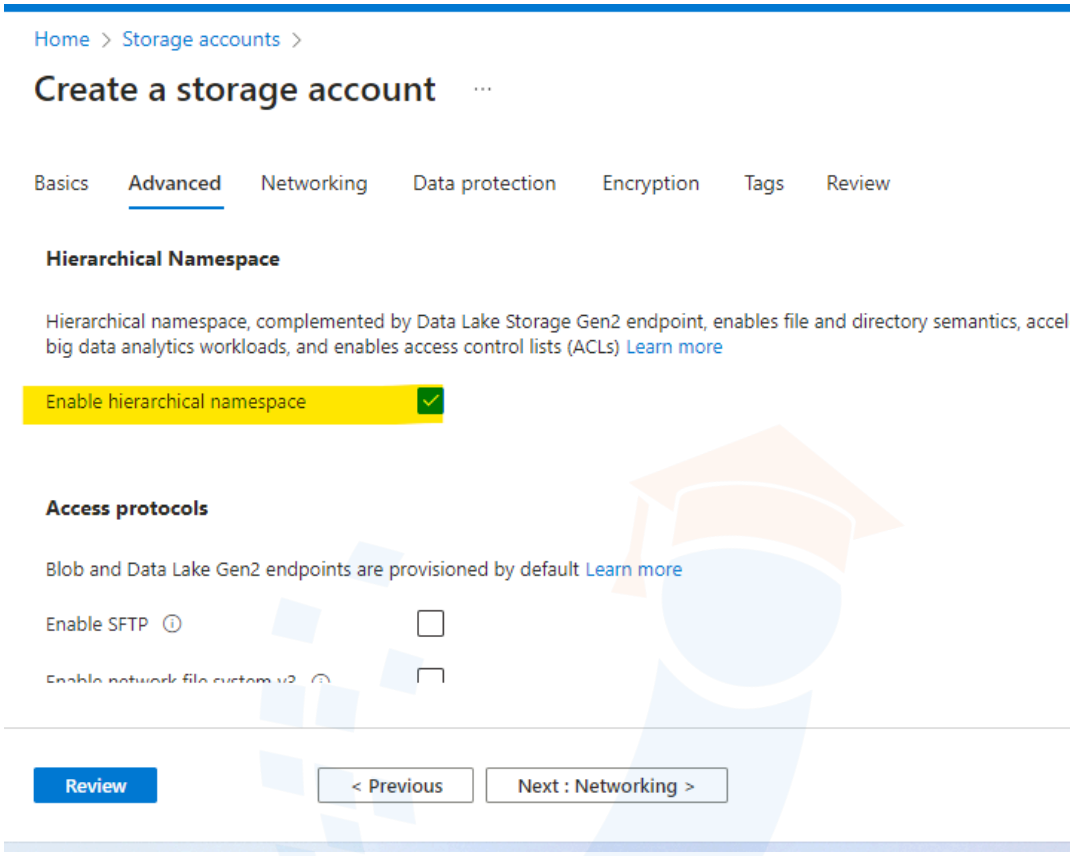
Note : After creating resources go to Setting => Configuration => Enable (Allow Blob anonymous access) => click on “Save”



Note: Once a resource has been created, you have the option to add it to the dashboard for quick access.

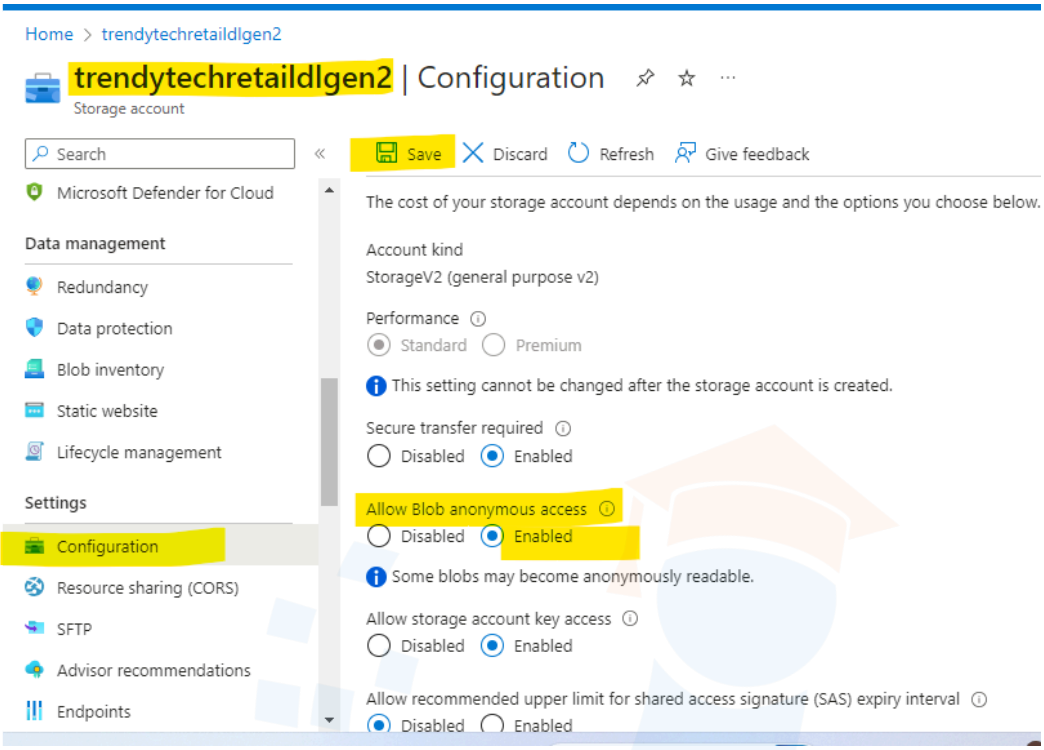
Create Storage Account Resource with hierarchical namespace enabled under the resource group -



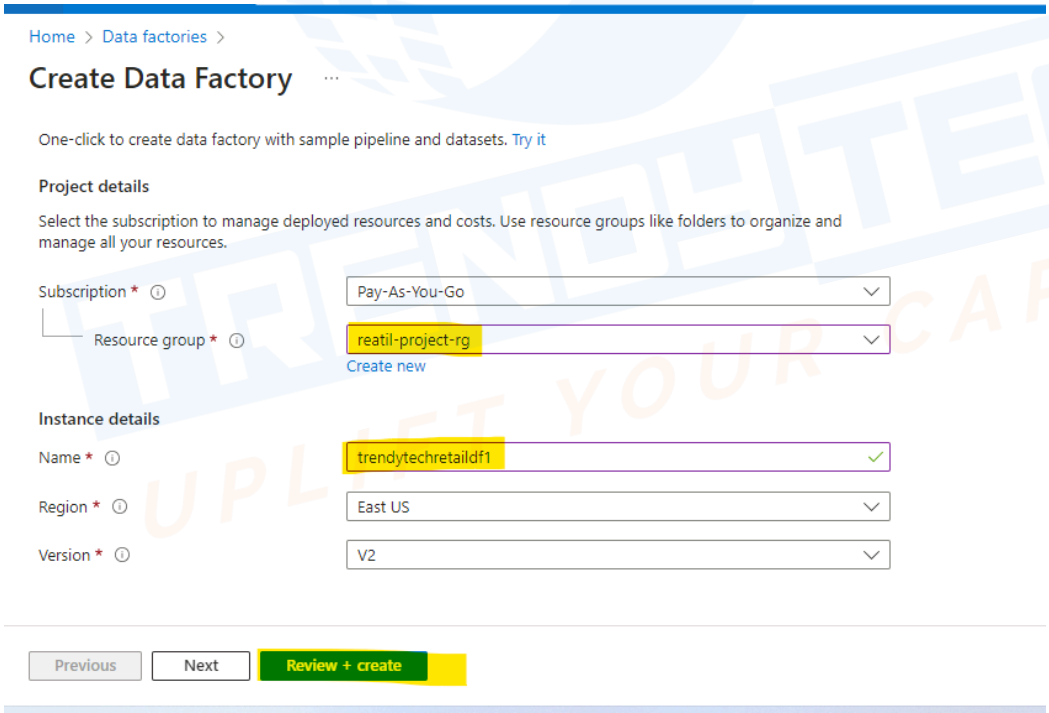


Note : After creating resources go to Setting => Configuration => Enable (Allow Blob anonymous access) => click on “Save”

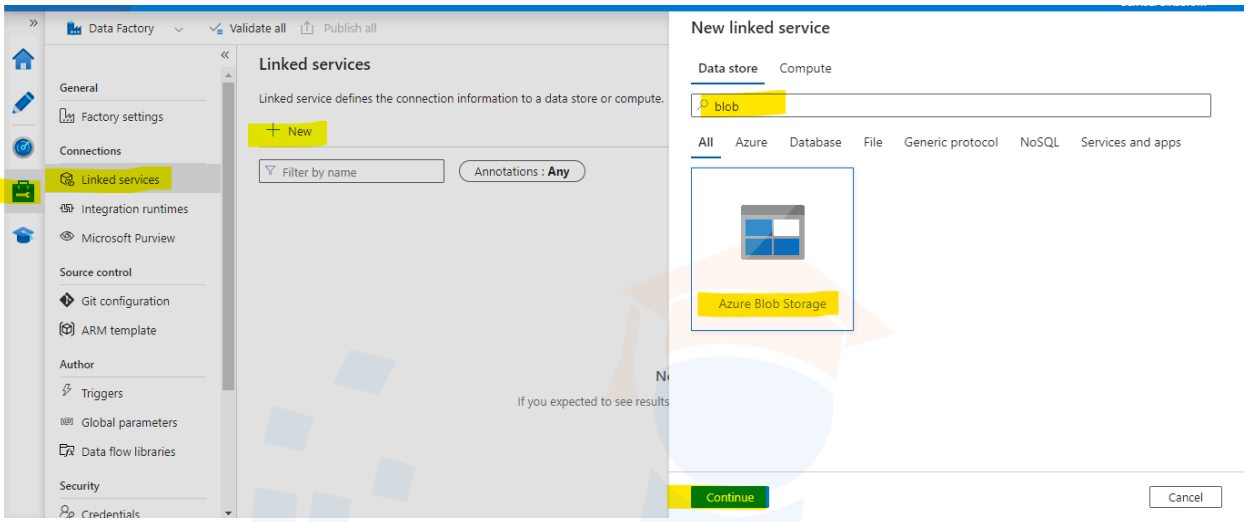
TRENDYTECH
UPLIFT YOUR CAREER!



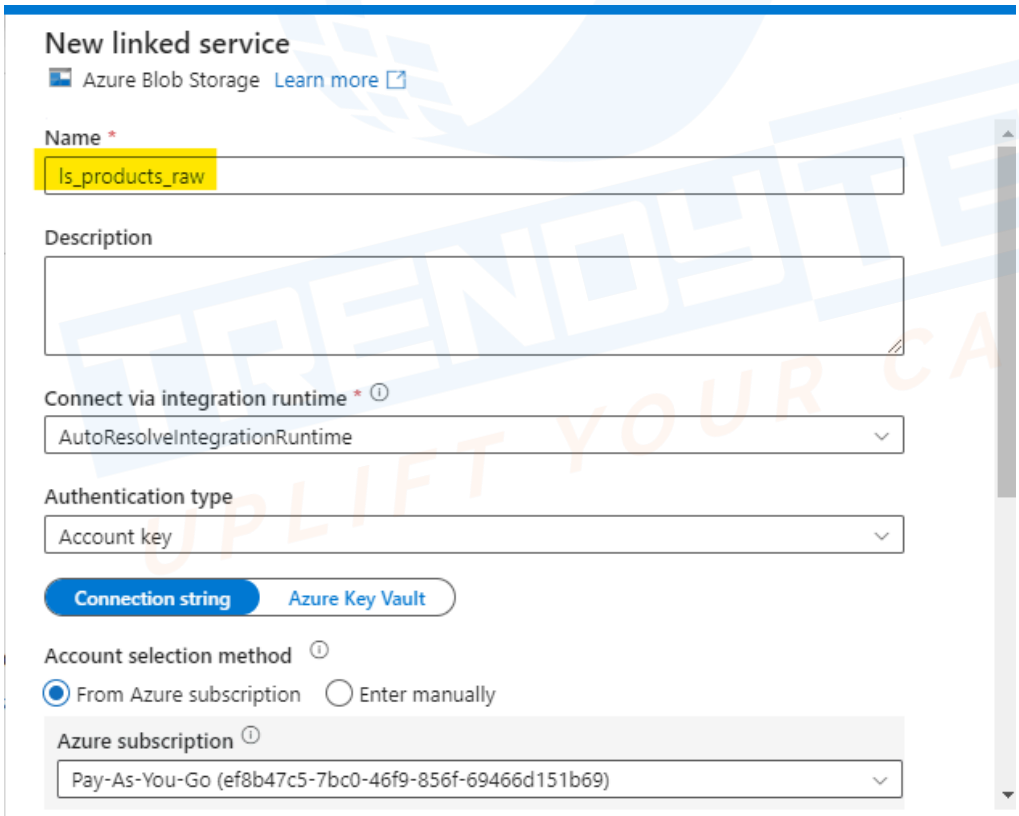
Create Data Factory Resource

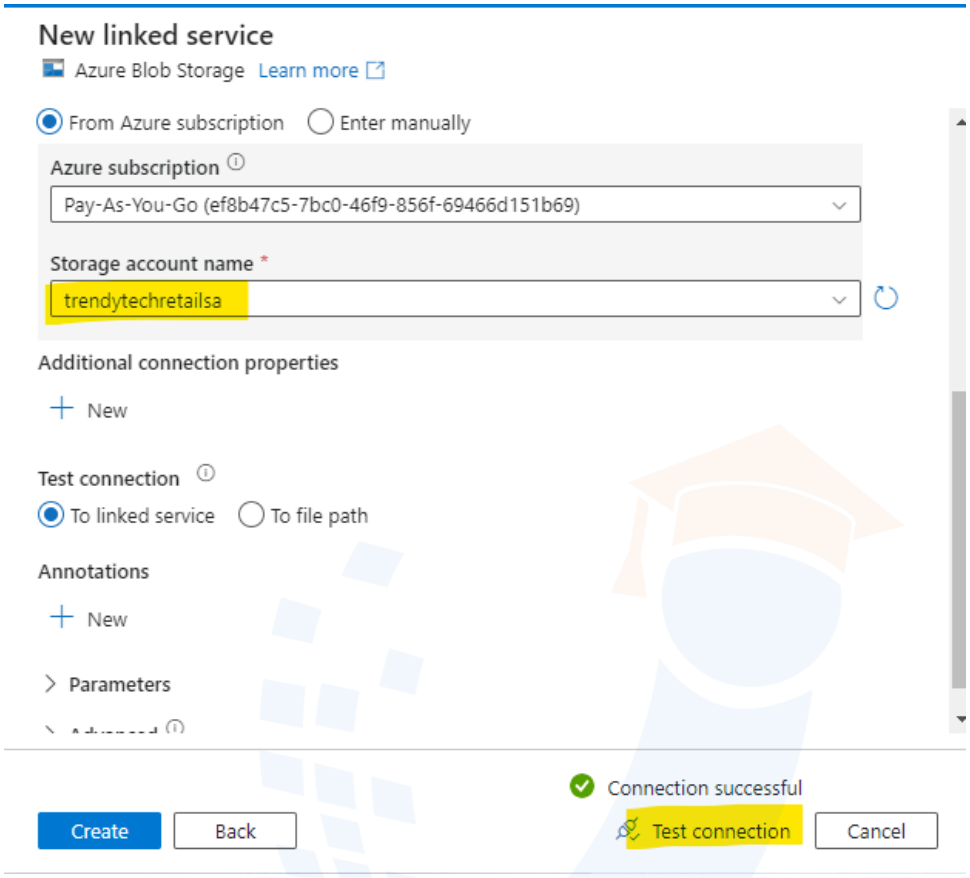


Create two Linked Services in ADF, one pointing the Source (BlobStorage) and other pointing the Sink(ADLSGen2)

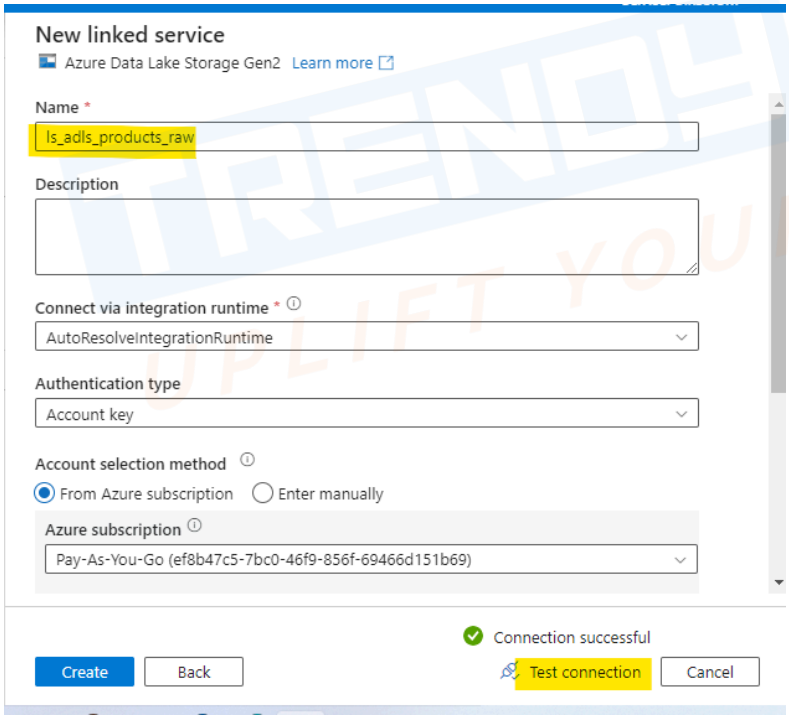


Creation of linked service pointing the Source (BlobStorage)





Creation of linked service pointing the Source (BlobStorage)

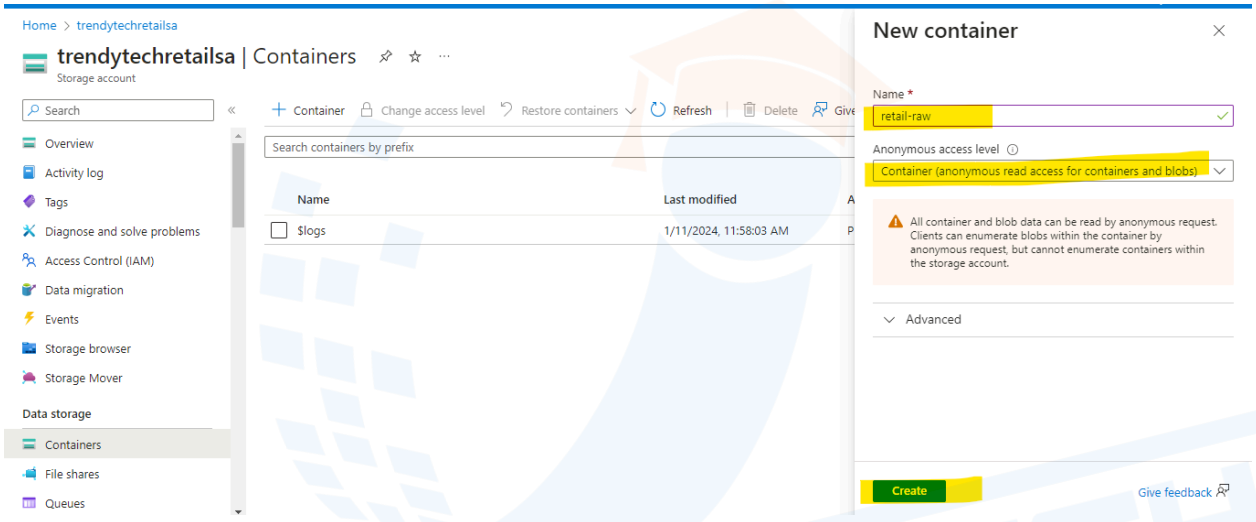


Note: After creating linked services in Azure Data Factory, be sure to publish these changes to make them active and available for use in your data workflows.

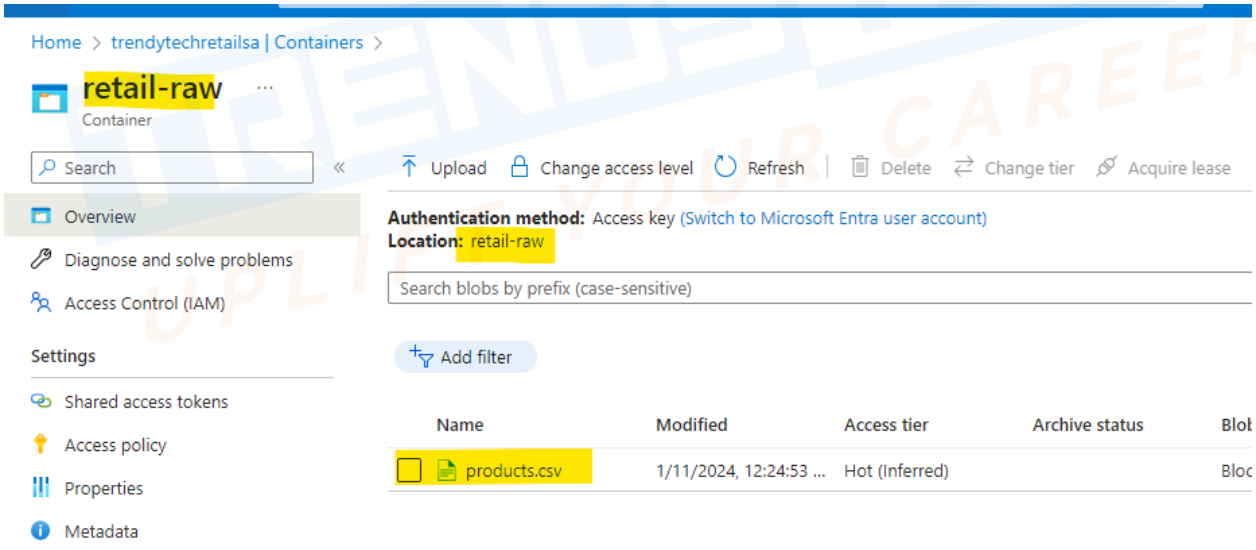
Upload the products.csv dataset to a container “retail-raw”in blob storage “trendytechretailsa”

First create the container “retail-raw”in blob storage:

Click on Container option in Data Storage => new container **(retail-raw)**=> set Anonymous access level to **“Container (anonymous access for container and blob)”**

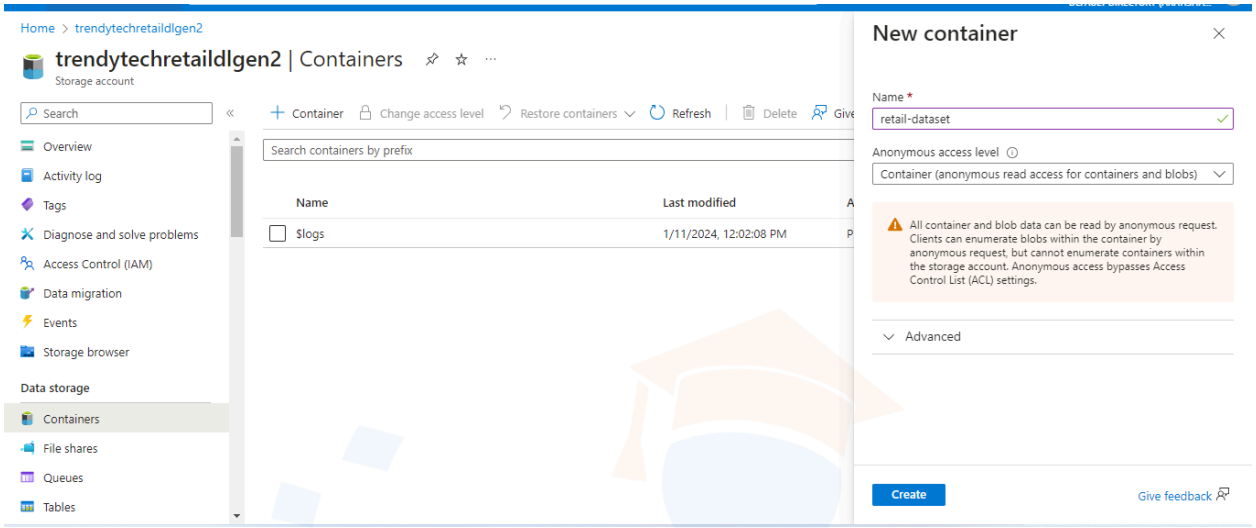


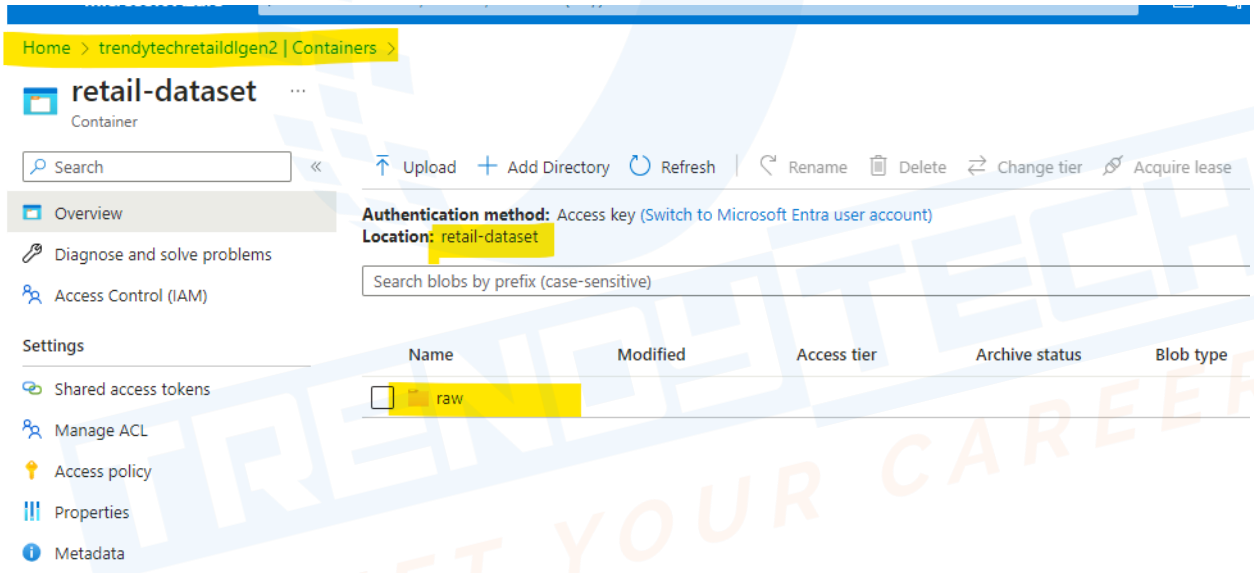
Then upload the products.csv dataset in it.



Also in Adls gen2 storage create container “retail-dataset” and ad directory “raw

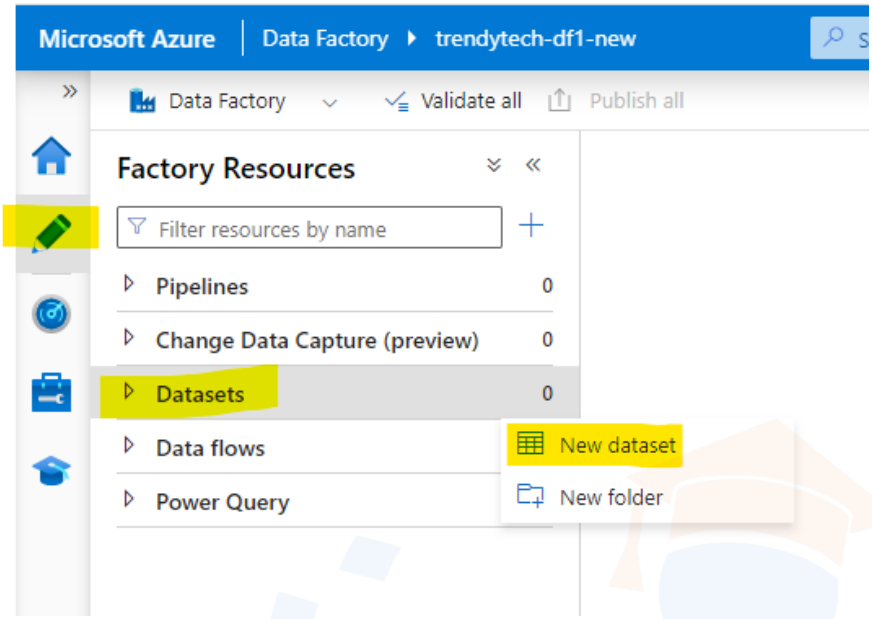
”



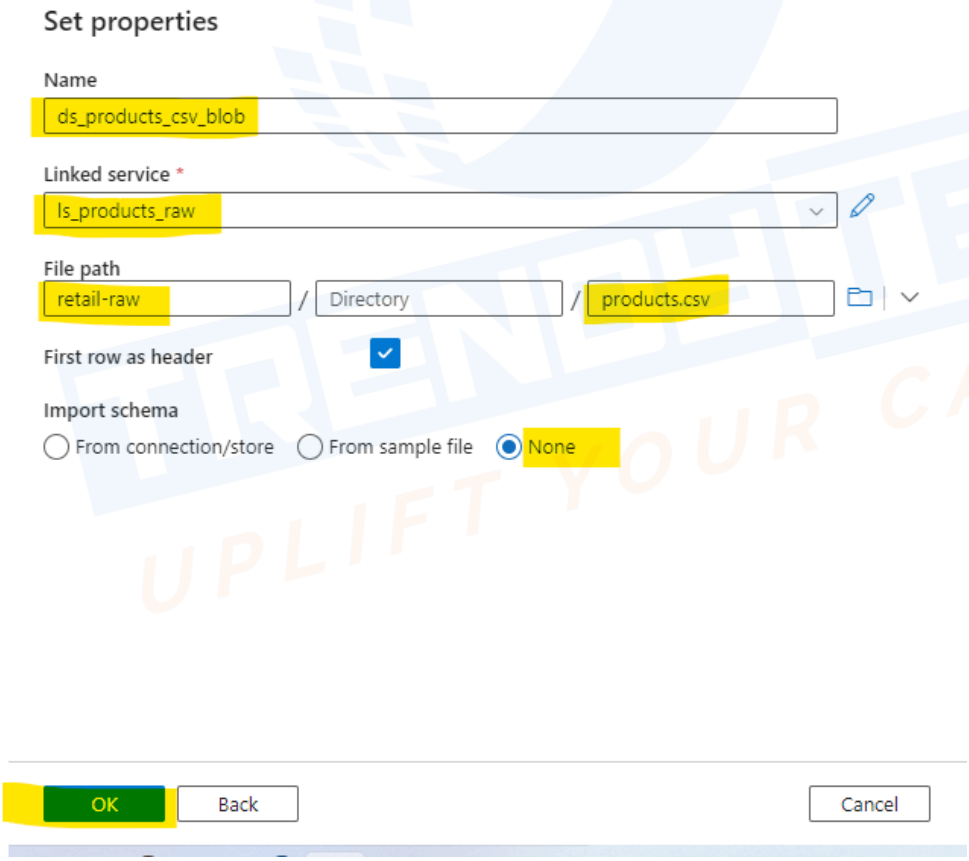


Create two Datasets, one for the products.csv source file in blob storage and the target dataset in csv format in ADLSGen2.

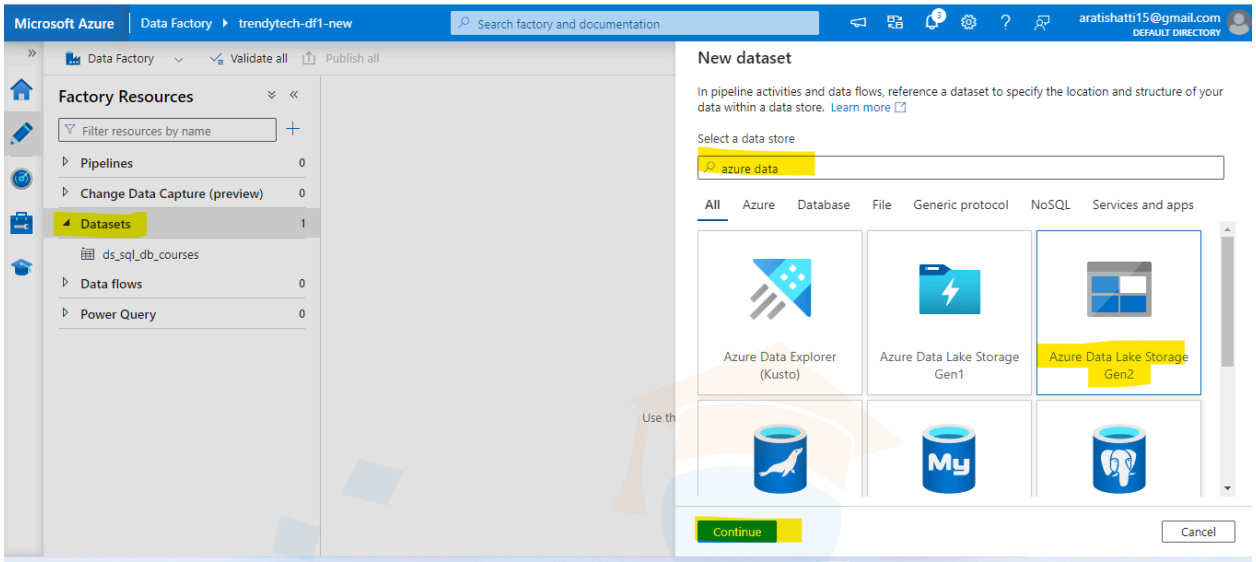
To create dataset click on Author => Datasets => New datasets



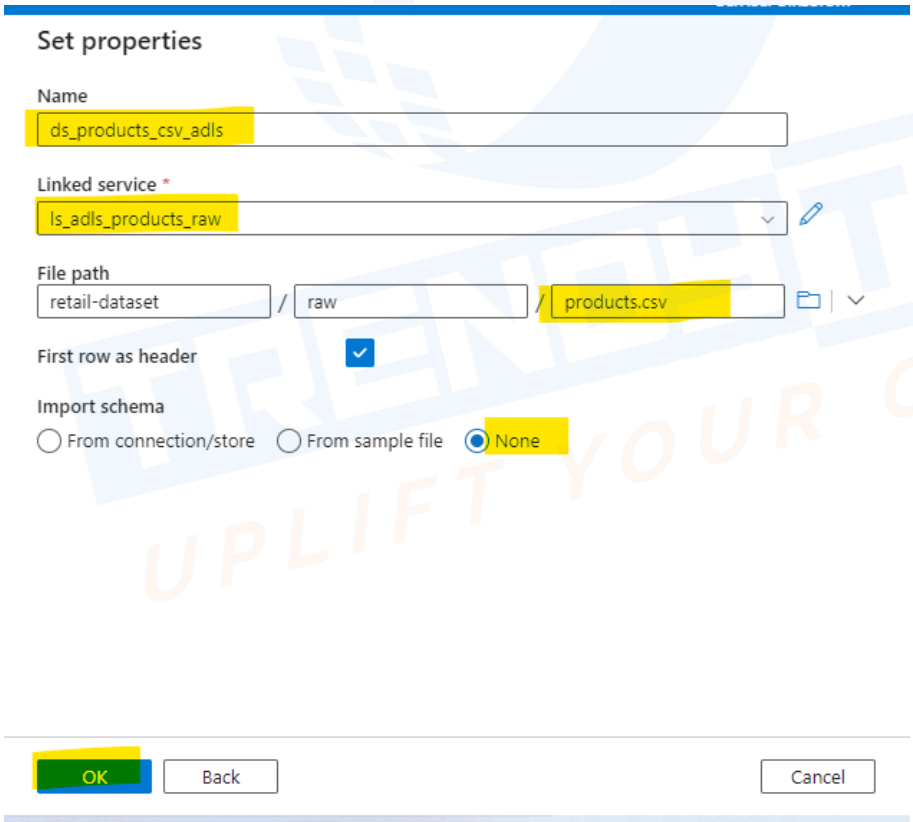
Creating a Dataset, one for the products.csv source file in blob storage



Dataset creation for sink (ADLS gen2) please check below steps:



While creating a dataset for sink after browsing the directory mention file name "products.csv" explicitly.

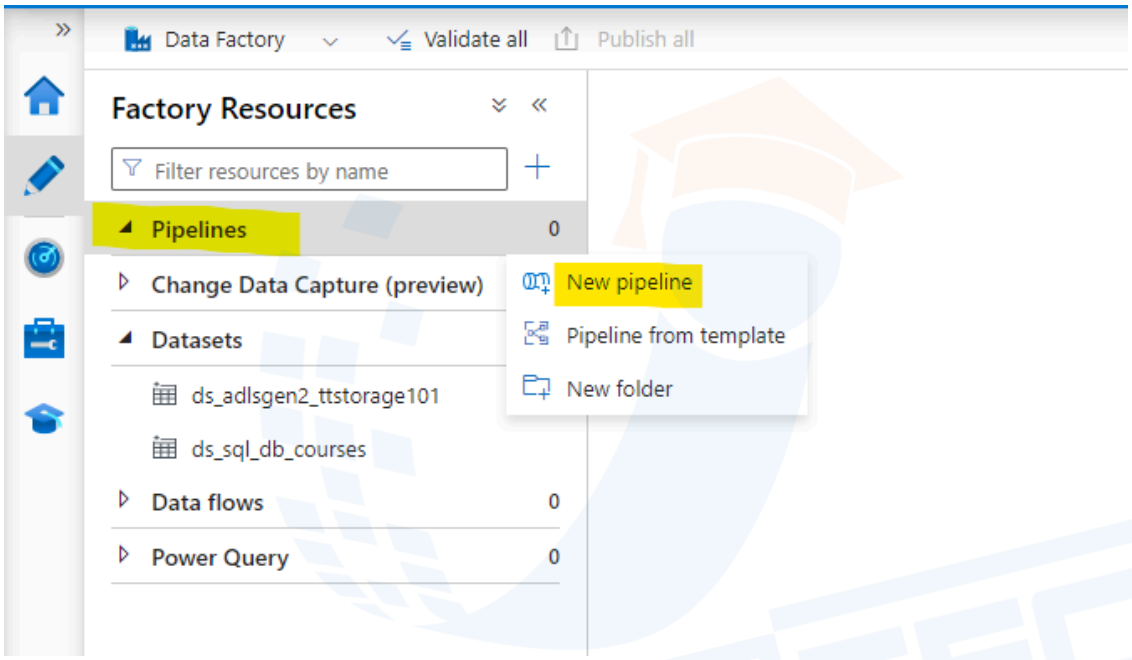


Note: After creating datasets in Azure Data Factory, be sure to publish these changes to make them active and available for use in your data workflows.

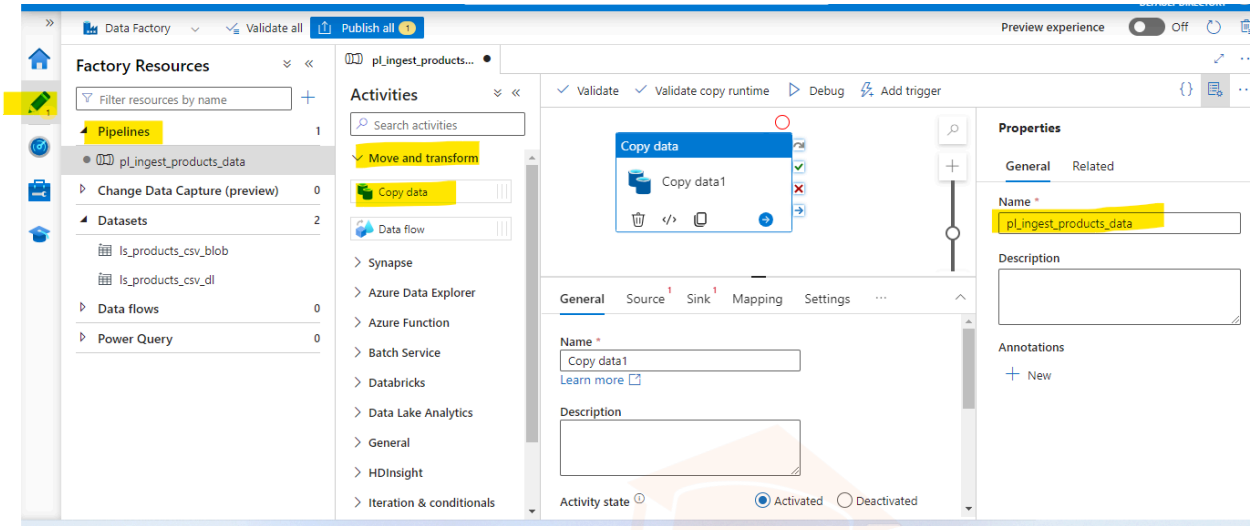
Create Pipeline and choose the source and sink to execute the required copyactivity.

Now we will create the pipeline and we will keep the copy activity in it –pl_ingest_products_data

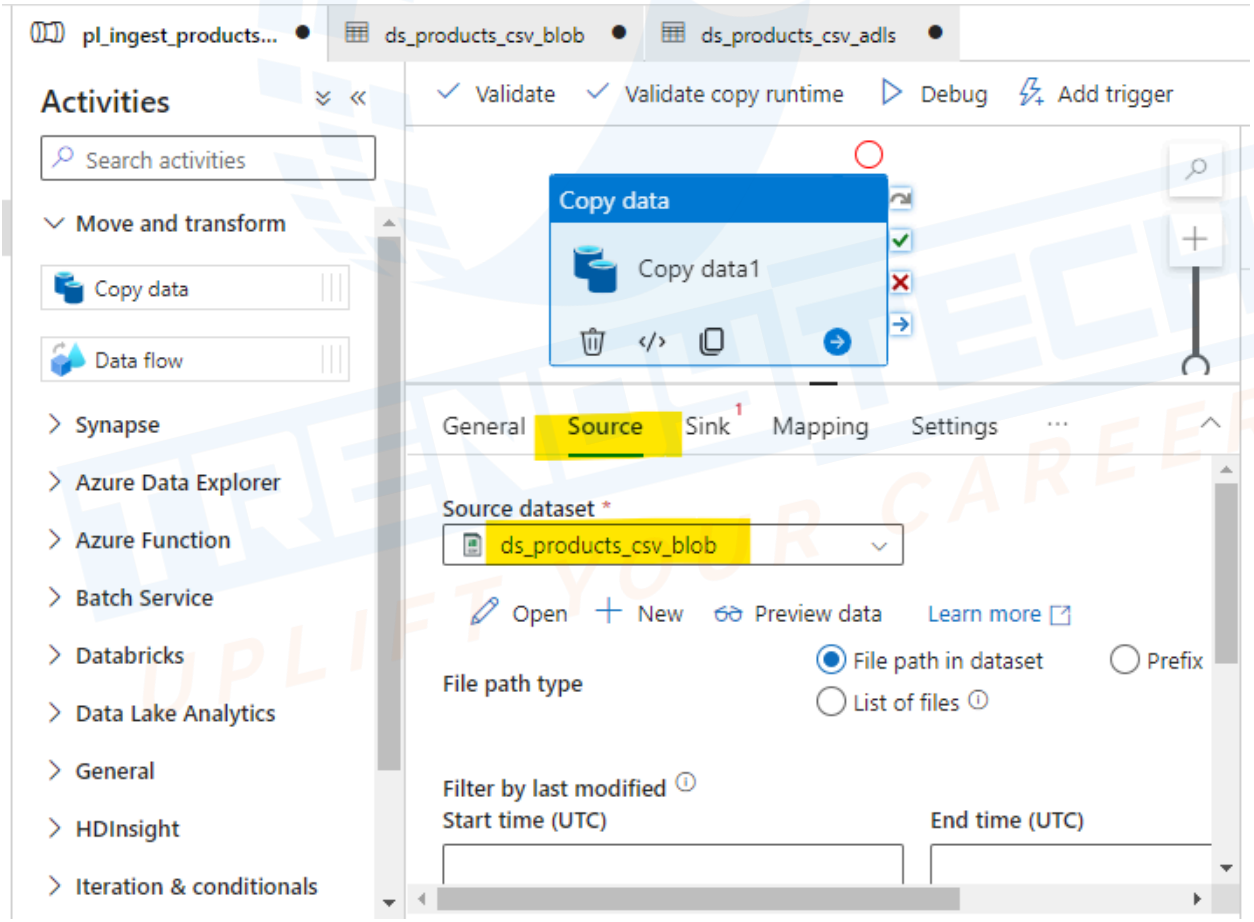
Click on three dots for pipelines option you will get below screen



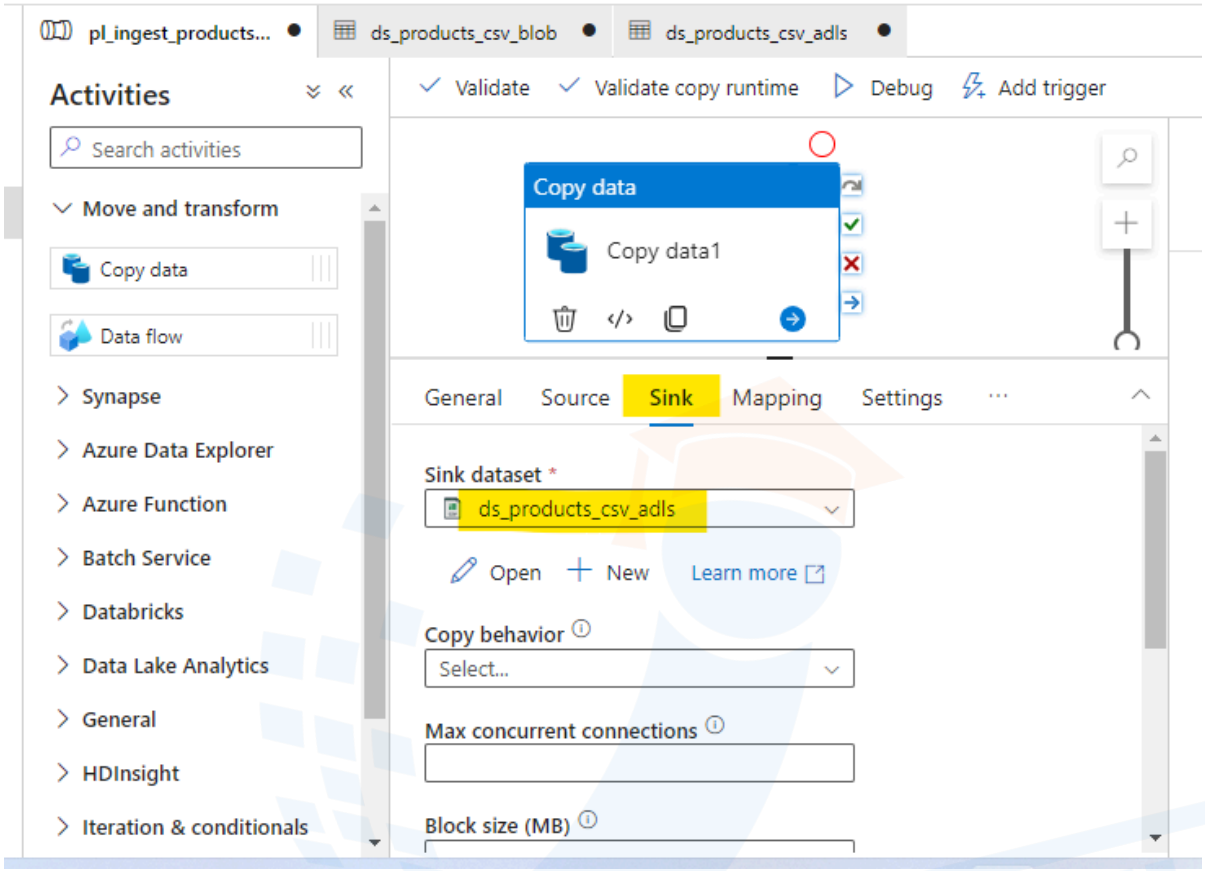
TRENDYTECH
UPLIFT YOUR CAREER!



Now in source select dataset for products.csv in blob storage **“ds_products_csv_blob”**

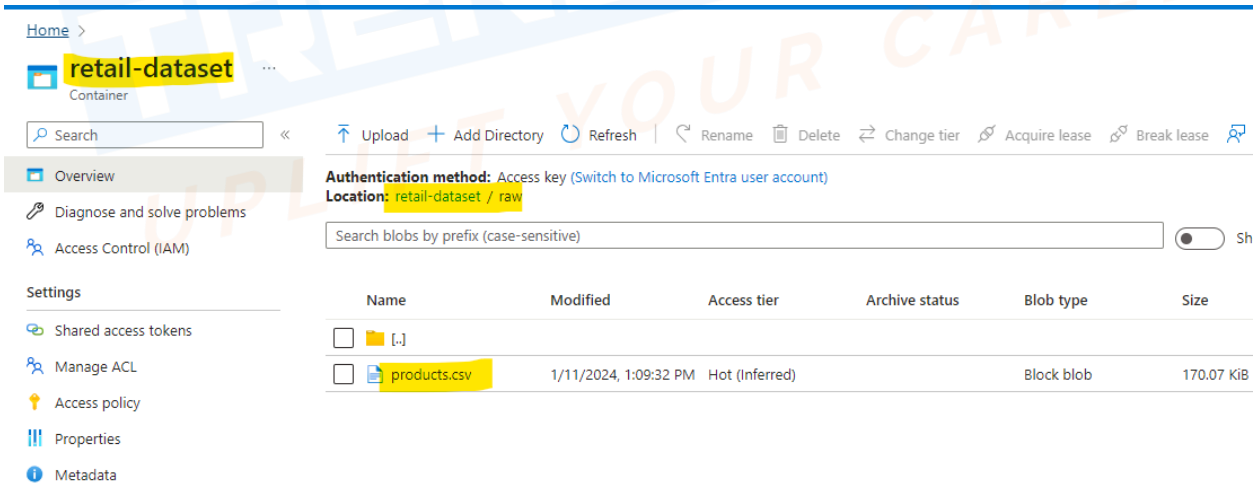


Now in sink select dataset for sink here **“ds_products_csv_adls”**



- Debug the pipeline and validate it. And if this is successful then publish the pipeline.

Also in your ADLS gen2 storage you will see the data copied after successfully running the pipeline. Refer below screenshot.

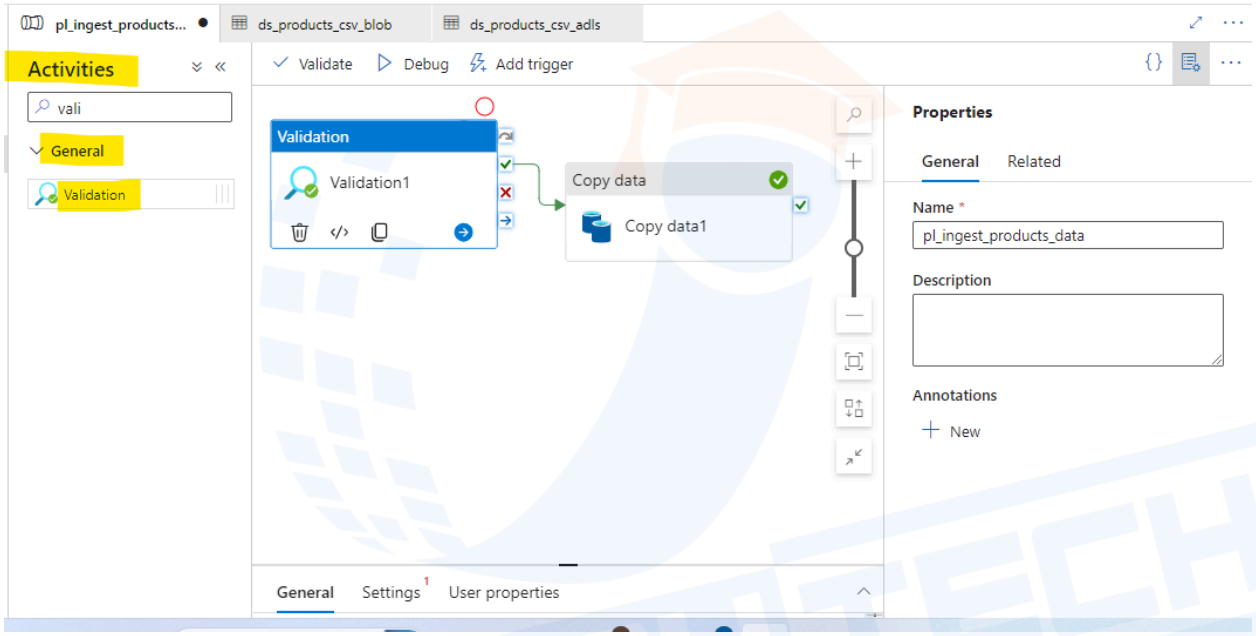


Note: After creating new pipelines in Azure Data Factory, be sure to publish these changes to make them active and available for use in your data workflows.

To Enhance The Above Process To Make It Production Ready

1.Pick the file as soon as it is available in blob storage without manual intervention.

Bring in the **"Validation activity"** present under the **"general category "**.

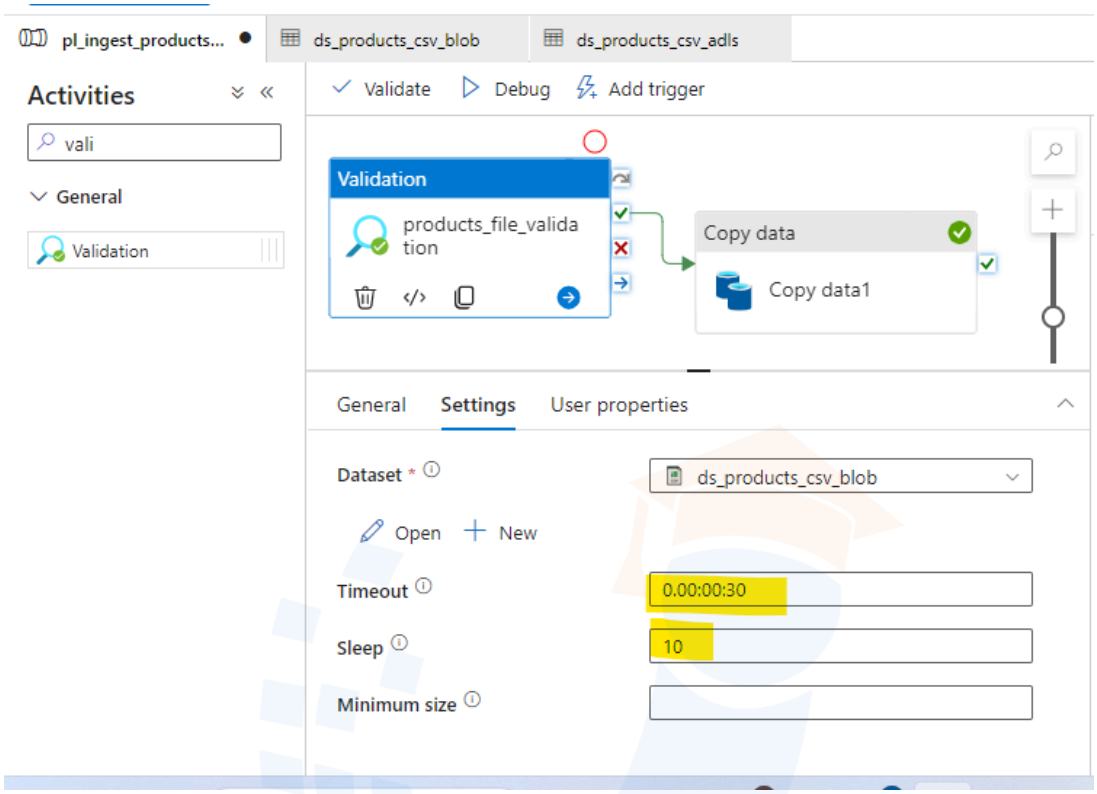


Set the values for timeout and sleep parameters to automatically pick any new files that are added to the source.

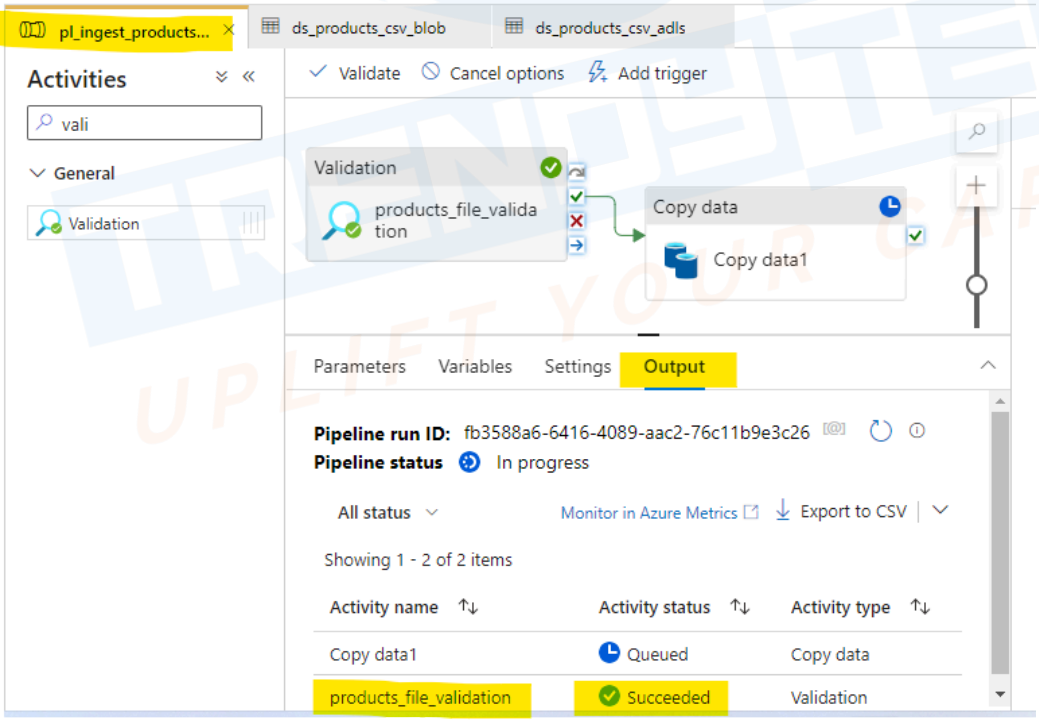
Note: Before this please remove the file in Source(Blob) and sink(ADLS gen) .

Click on Debug => Then within the mentioned time upload the file (products.csv) in blob storage => Check whether validation activity has been passed or not.

Ex: Mention 10 sec as time and can check validation activity mentioned above.

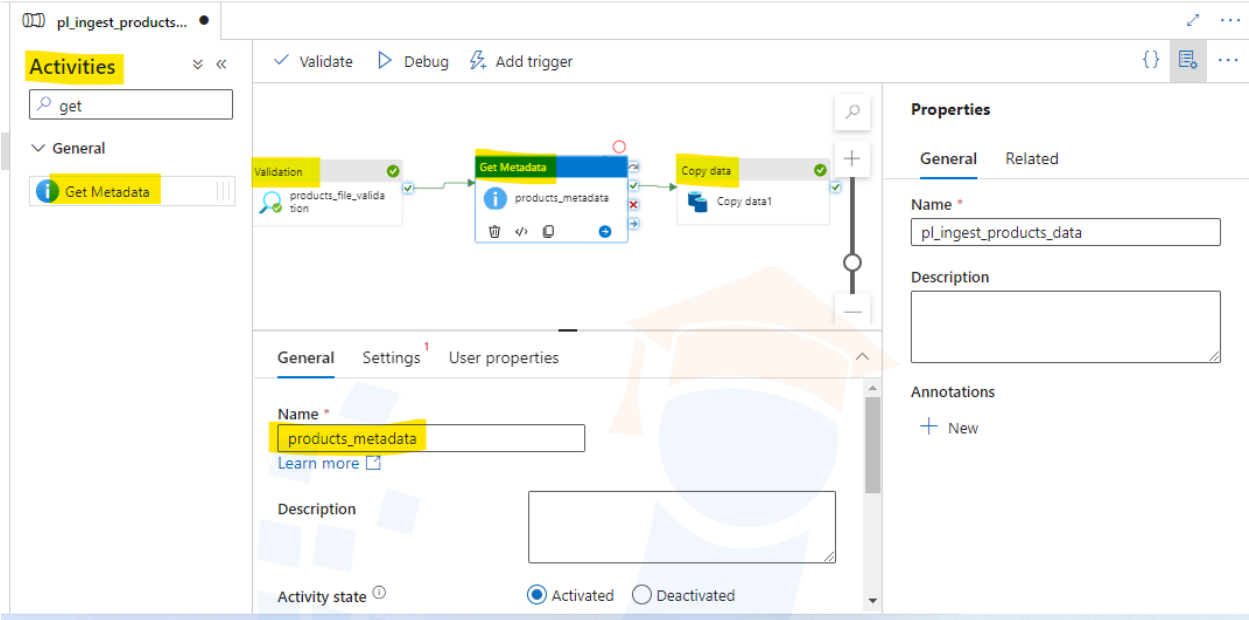


If validation has been successful it will invoke copy data activity, you can see the output as below

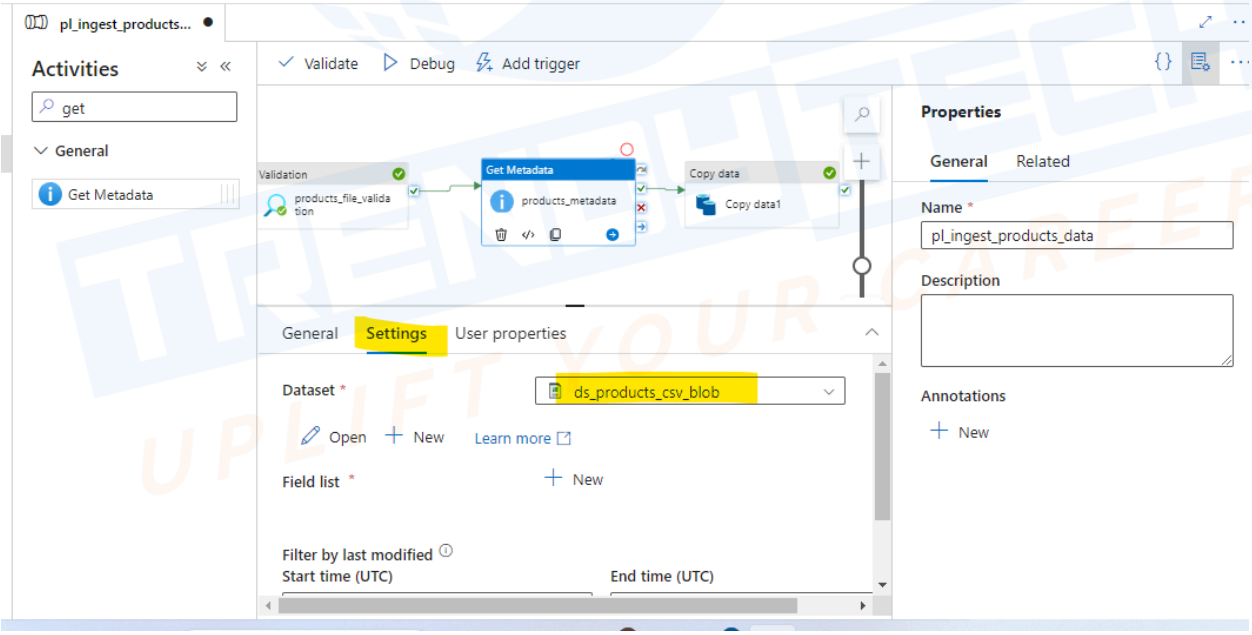


2.**Perform Basic Validation/sanity check before ingesting the Datalake.**

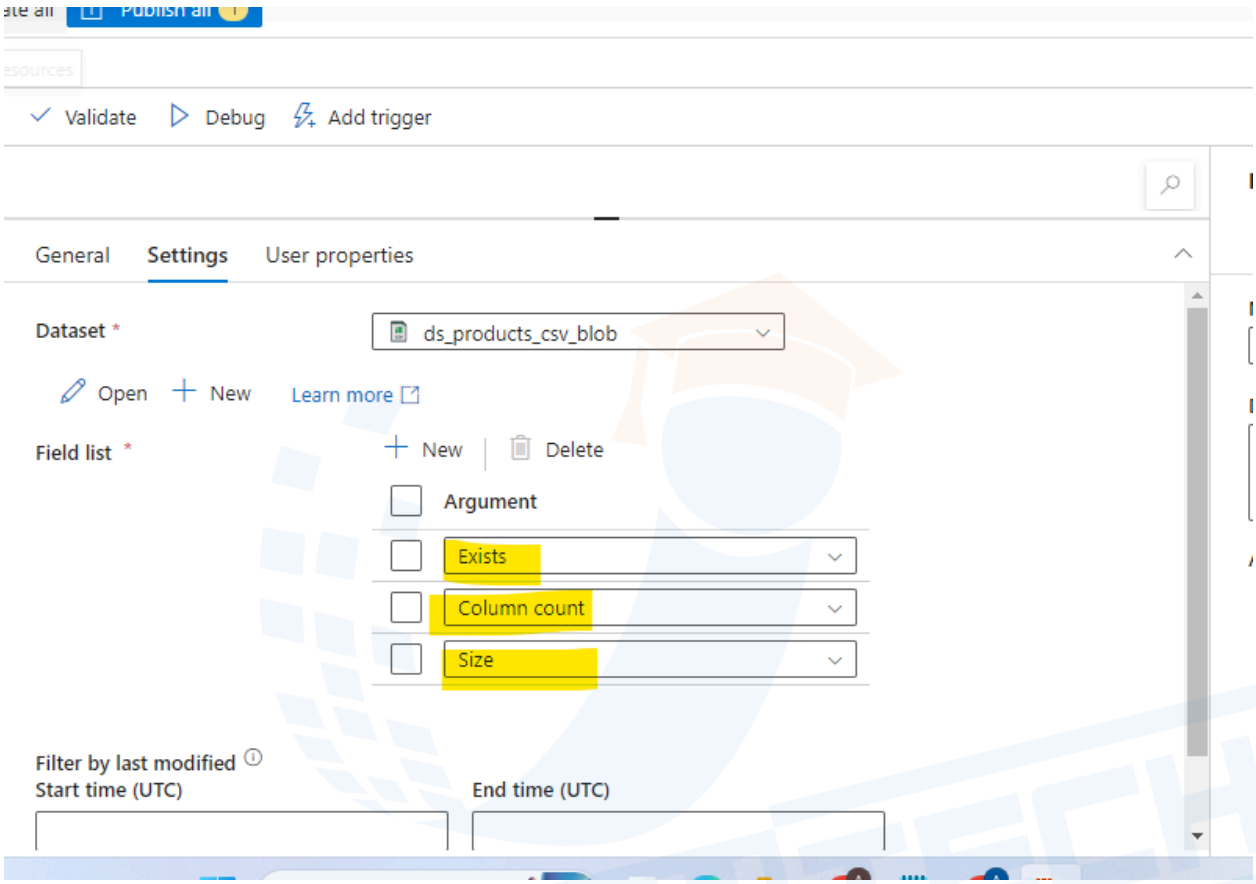
Use the **“Get Metadata activity”** to perform **basic validations like column count, file size.**



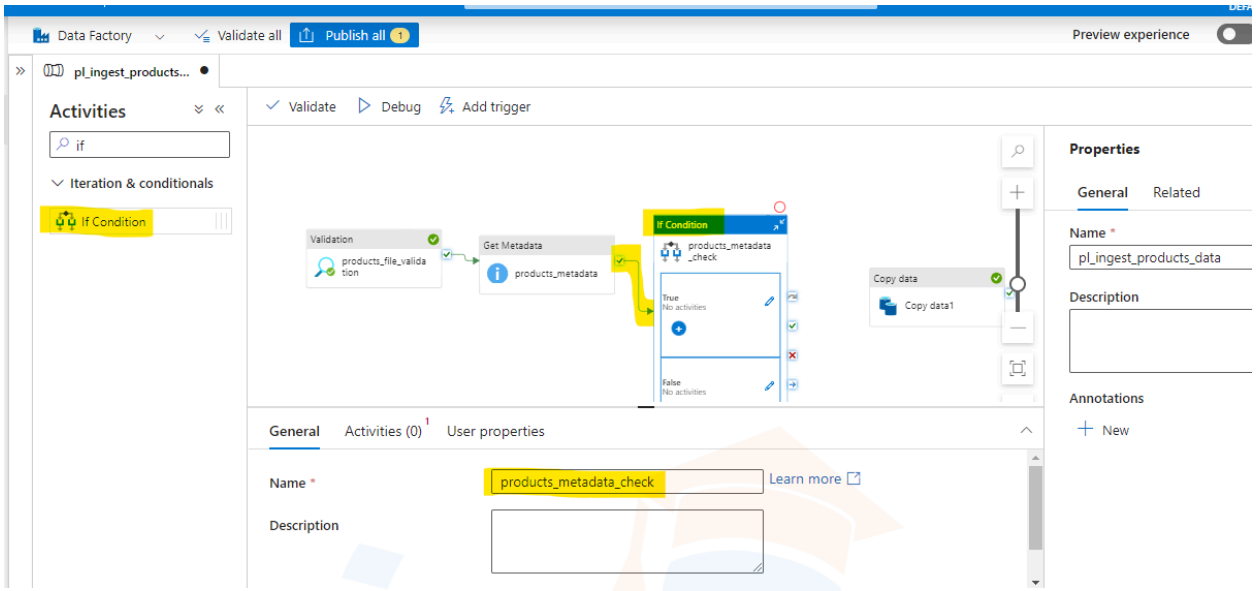
Mentioning the source dataset refer the attached screenshot, you can keep other setting as it is.



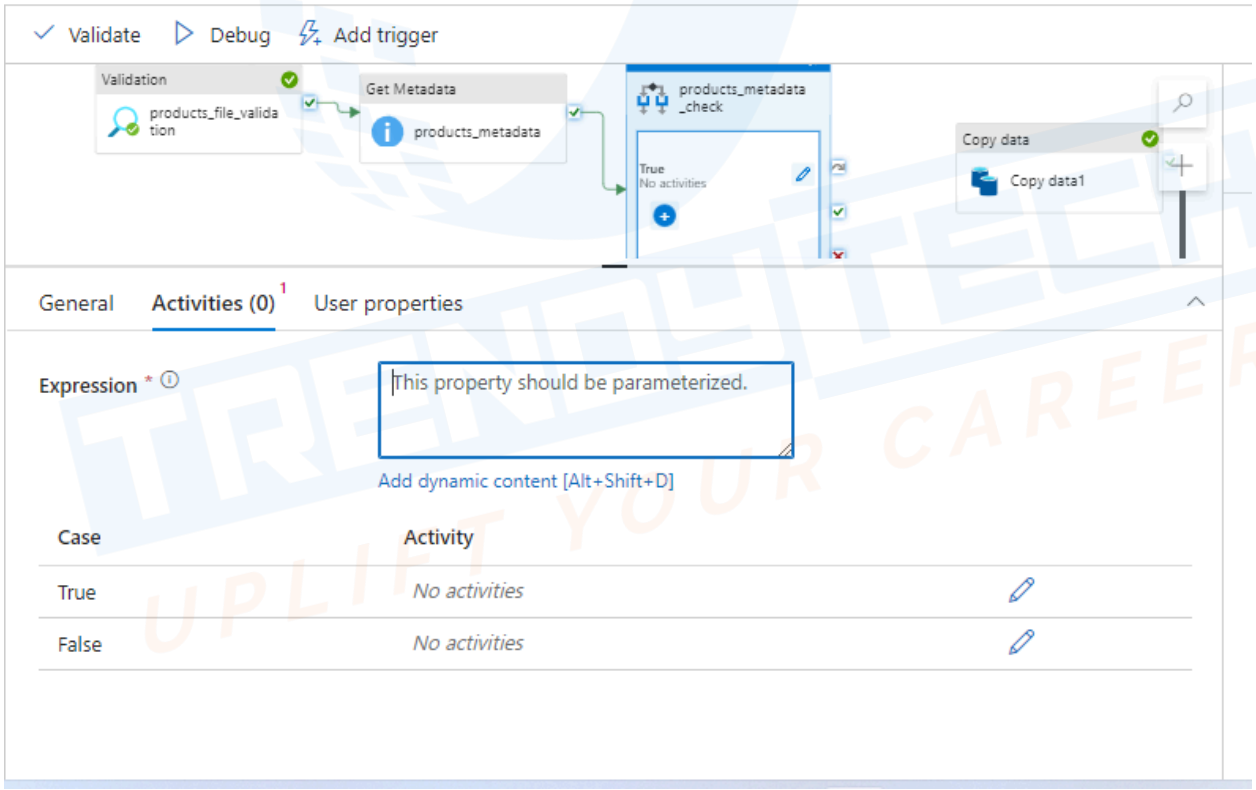
Note: In the Get Metadata activity, by utilizing the "field list" option, we can specify the file properties we desire, such as existence, column count, size, and other relevant attributes.. Refer attached screenshot



To validate the column count, an additional activity, specifically an "If Condition," will be necessary.



To mention the activity click on “Add dynamic content”

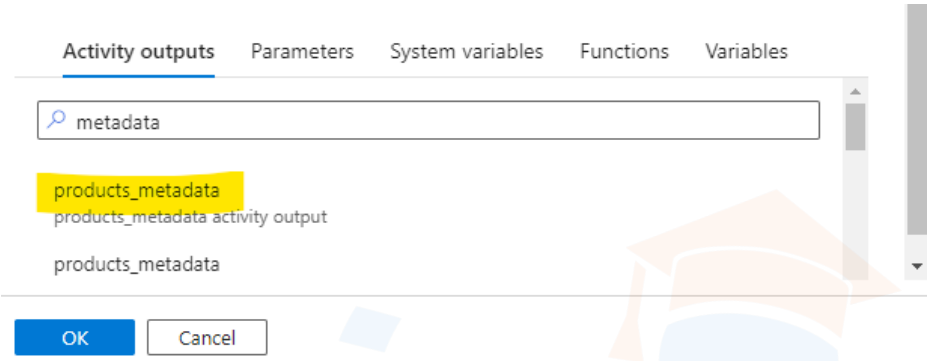


Now to check the column count, basically it should be 6 as per this products.csv file we will follow below process

a. First mention **“@equals()”**

=> `@equals()`

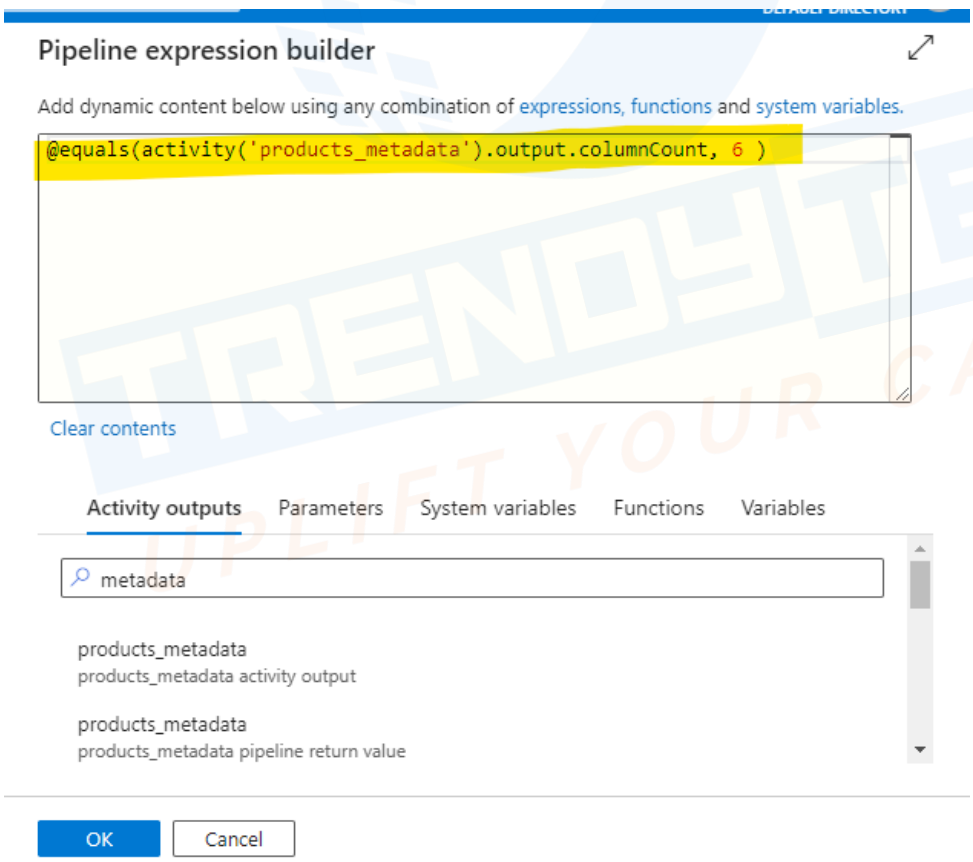
b. Then select **“products metadata output”**



=> `@equals(activity('products_metadata').output)`

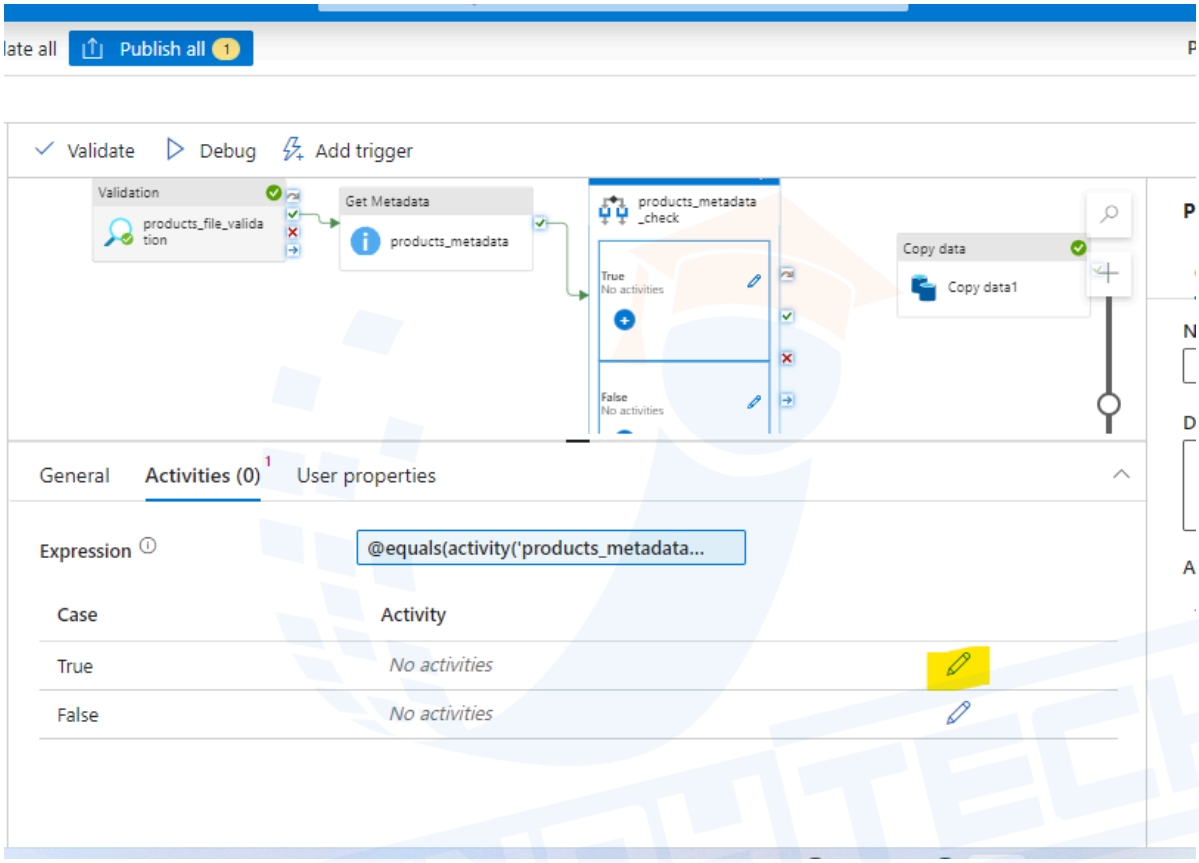
C. To calculate column count add **“.columnCount”**

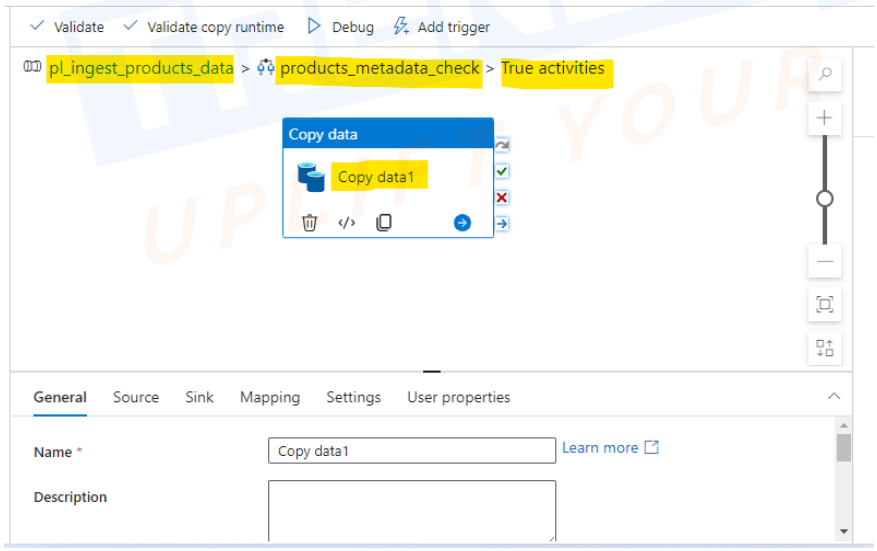
=> `@equals(activity('products_metadata').output.columnCount, 6 )`



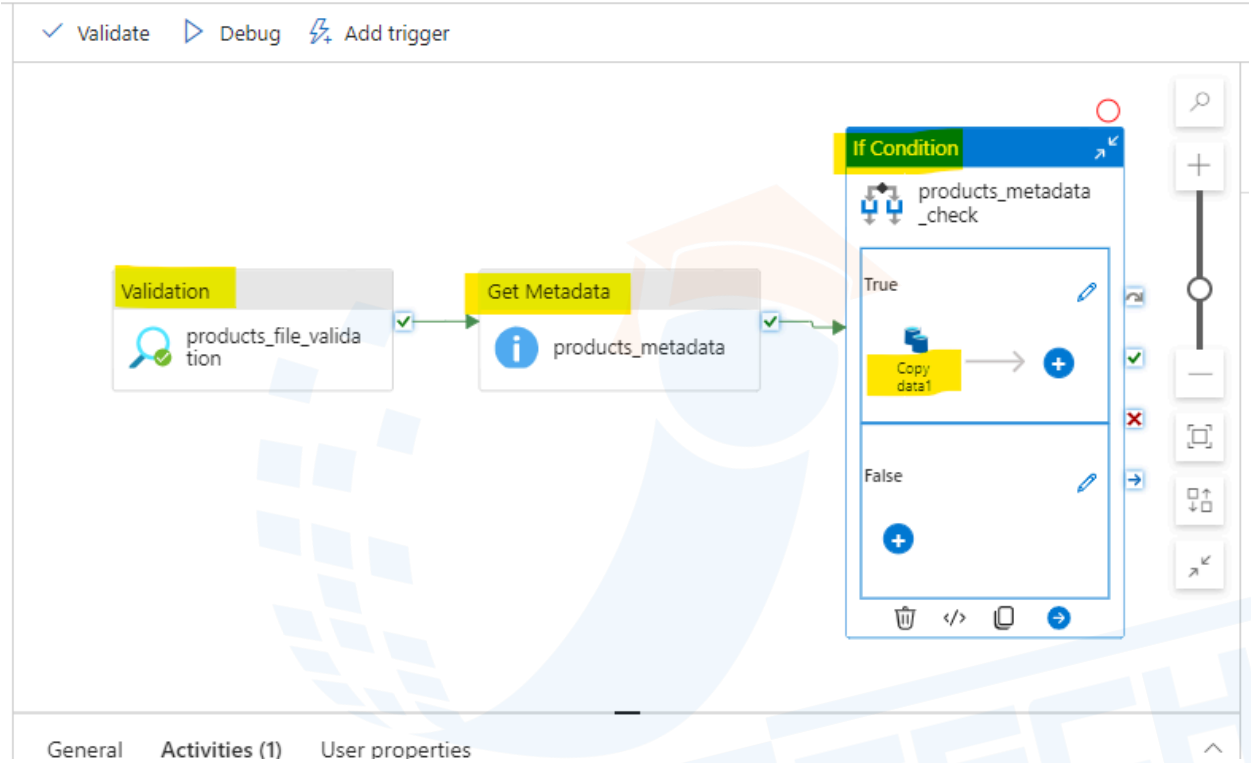
When this "If Condition" evaluates to "True," it is essential to execute the Copy Data activity.

To accomplish this, click on the **"Edit"** option for the **"True"** branch, and then **paste** the previously created **Copy Data activity (“copy data 1”)**.





The resulting pipeline configuration is illustrated in the image below. After the validation process is completed, the Get Metadata activity is executed. If the column count validation is successful, the file will be copied from the source(Blob storage) to the sink (ADLS gen2).

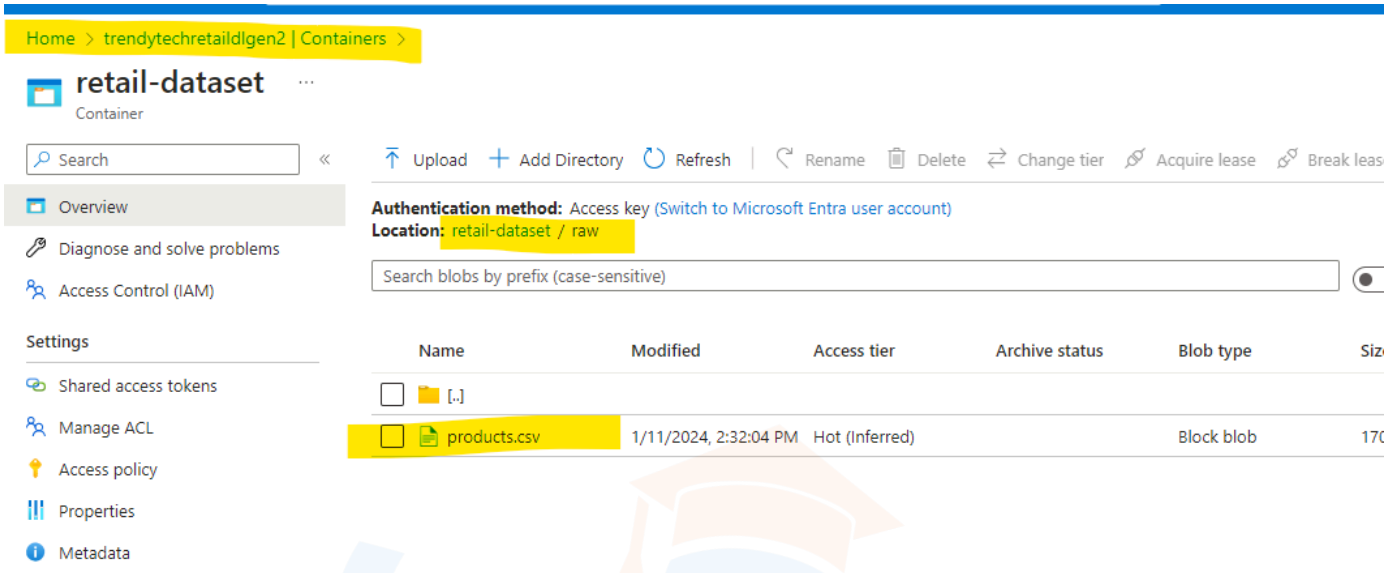


Debug the pipeline and validate it.

Note: Before this please remove the file in Source(Blob) and sink(ADLS gen) .

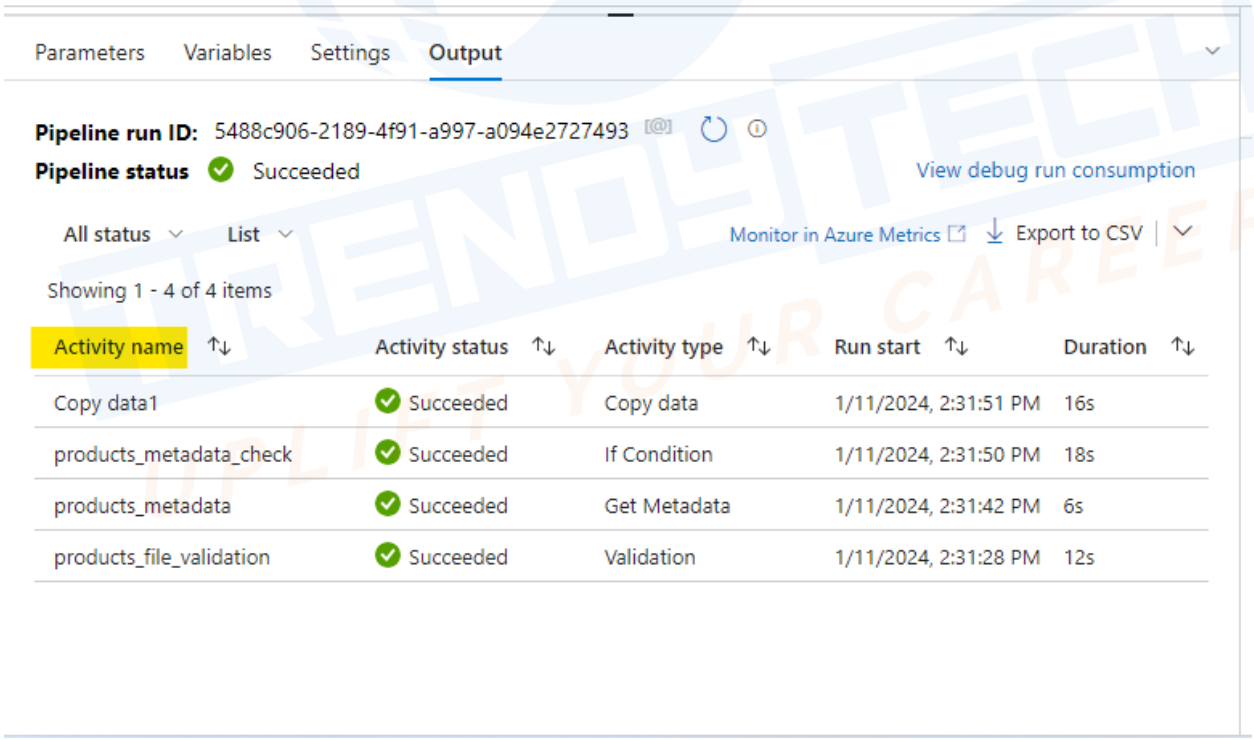
Click on Debug => Then within the mentioned time upload the file (products.csv) in blob storage => Check pipeline runs

Upon the successful execution of the pipeline, the "products.csv" file will be transferred to the specified sink, namely "trendytechretaildlgen2," within the "raw" folder of the "retail-dataset" container.



Note: After creating new pipelines in Azure Data Factory, be sure to publish these changes to make them active and available for use in your data workflows.

You can check the output of each activity.



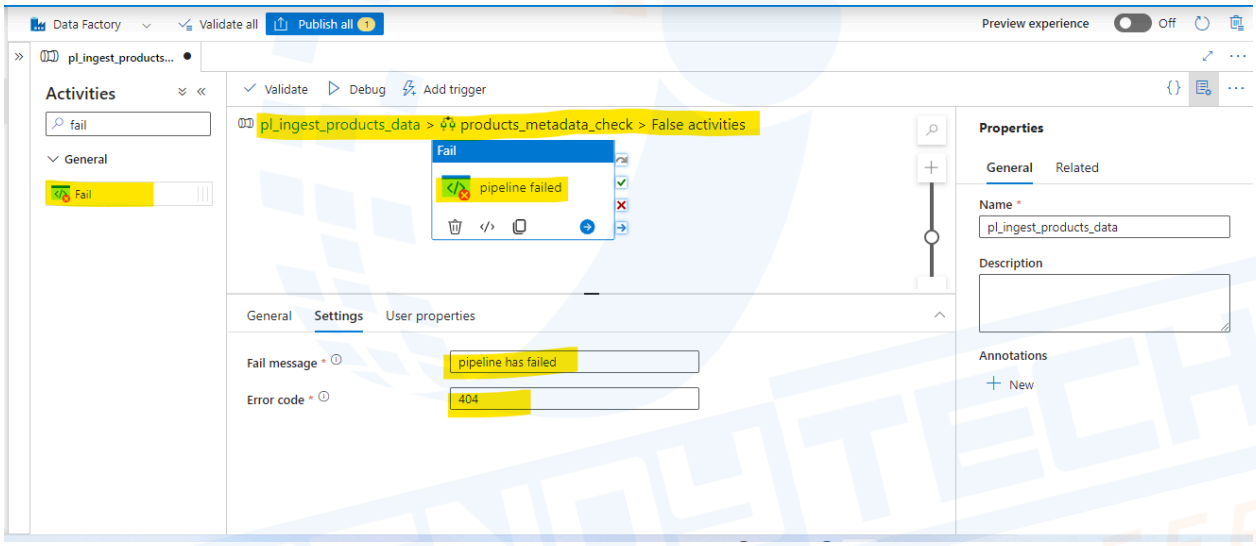
3.Notify When The Pipeline Execution Fails.

Note: Before this please remove the file in Source(Blob) and sink(ADLS gen) .

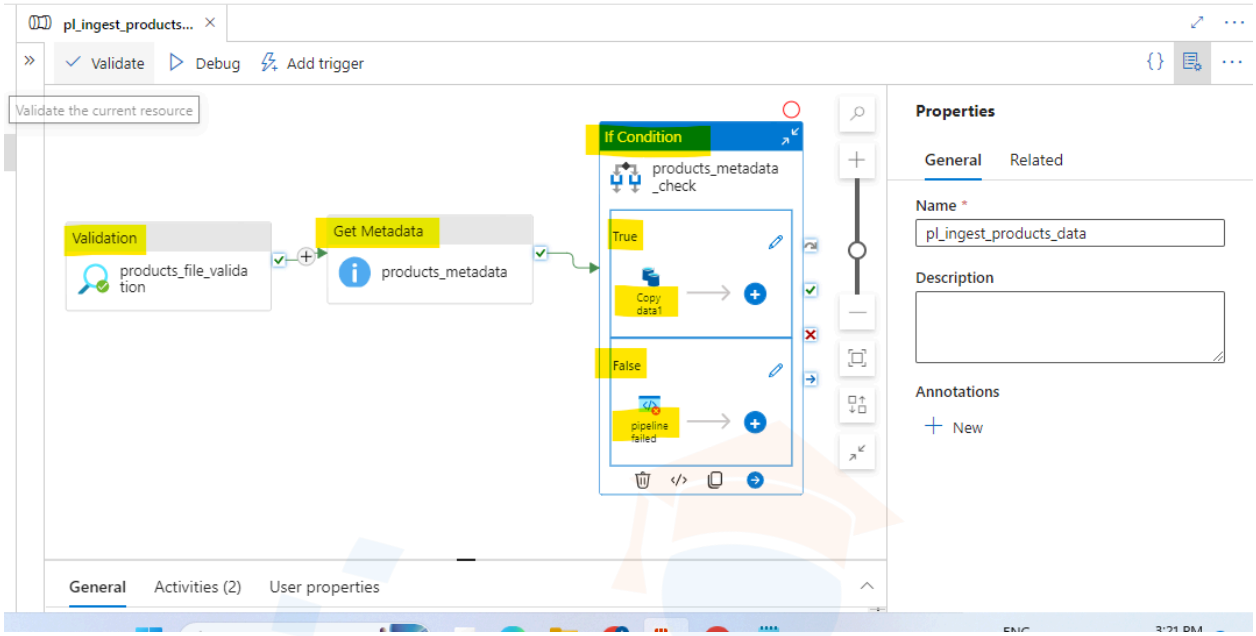
Use Fail activity to indicate in case the above If condition data validation process fails and use Alert Rule to notify the pipeline failure.

When this "If Condition" evaluates to "False," the pipeline will fail and it should show us the message regarding its failure.

To accomplish this, click on the **"Edit"** option for the **"False"** branch, and then **add the "Fail activity" and add message pipeline failed with error code suppose 404.**



The depicted pipeline configuration is presented in the image below. Following the validation process, the Get Metadata activity is triggered. Upon successful column count validation, the file is transferred from the source (Blob storage) to the sink (ADLS Gen2). In the event of a validation failure, the process proceeds to the "False" branch, displaying a failure message in the pipeline logs.



Debug the pipeline and validate it.

Note: Before this please remove the file in Source(Blob) and sink(ADLS gen) .

Furthermore, make a copy of the "pipeline.csv" file to a different location, such as the desktop. Subsequently, edit the copied file by removing the value of the last column in the first row. Utilize this modified file for debugging purposes.

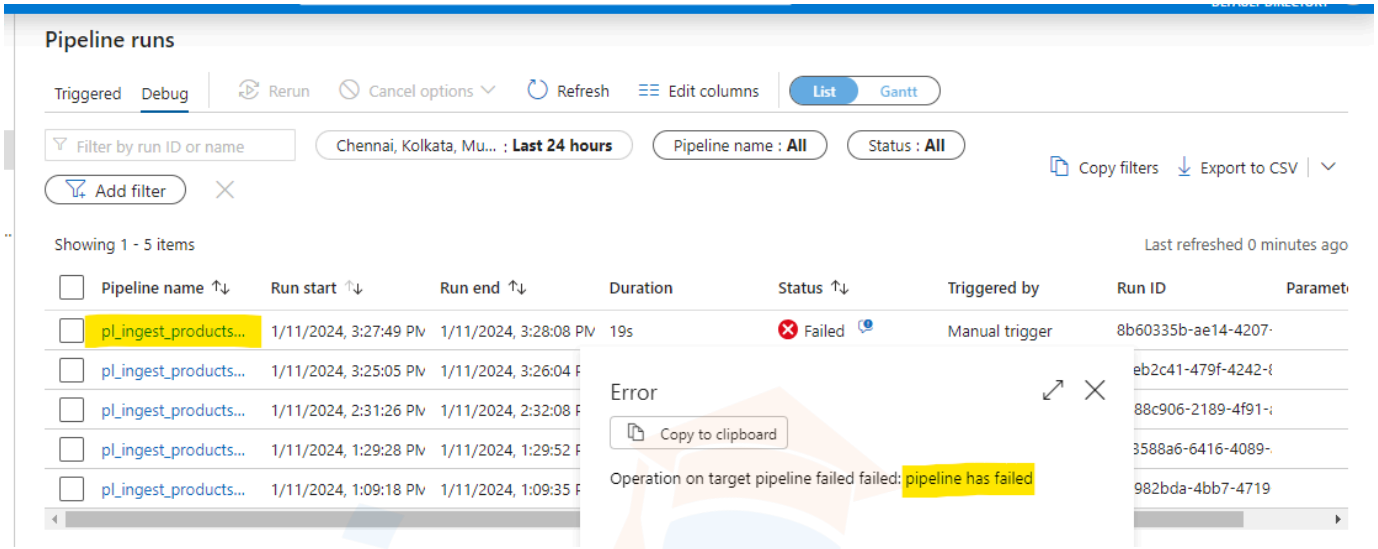
In this modified file, the last column has been removed, resulting in the following structure.

```
1,2,Quest Q64 10 FT. x 10 FT. Slant Leg Instant U,,59.98
2,2,Under Armour Men's Highlight MC Football Clea,,129.99,http://images.acmesports.sports/Under+Armour+Men%27s+Highlight+MC+Football+Cleat
3,2,Under Armour Men's Renegade D Mid Football Cl,,89.99,http://images.acmesports.sports/Under+Armour+Men%27s+Renegade+D+Mid+Football+Cleat
4,2,Under Armour Men's Renegade D Mid Football Cl,,89.99,http://images.acmesports.sports/Under+Armour+Men%27s+Renegade+D+Mid+Football+Cleat
5,2,Riddell Youth Revolution Speed Custom Footbal,,199.99,http://images.acmesports.sports/Riddell+Youth+Revolution+Speed+Custom+Football+Helmet
6,2,Jordan Men's VI Retro TD Football Cleat,,134.99,http://images.acmesports.sports/Jordan+Men%27s+VI+Retro+TD+Football+Cleat
7,2,Schutt Youth Recruit Hybrid Custom Football H,,99.99,http://images.acmesports.sports/Schutt+Youth+Recruit+Hybrid+Custom+Football+Helmet+2014
8,2,Nike Men's Vapor Carbon Elite TD Football Cle,,129.99,http://images.acmesports.sports/Nike+Men%27s+Vapor+Carbon+Elite+TD+Football+Cleat
9,2,Nike Adult Vapor Jet 3.0 Receiver Gloves,,50.0,http://images.acmesports.sports/Nike+Adult+Vapor+Jet+3.0+Receiver+Gloves
10,2,Under Armour Men's Highlight MC Football Clea,,129.99,http://images.acmesports.sports/Under+Armour+Men%27s+Highlight+MC+Football+Cleat
11,2,Fitness Gear 300 lb Olympic Weight Set,,209.99,http://images.acmesports.sports/Fitness+Gear+300+lb+Olympic+Weight+Set
12,2,Under Armour Men's Highlight MC Alter Ego Fla,,139.99,http://images.acmesports.sports/Under+Armour+Men%27s+Highlight+MC+Alter+Ego+Flash+Football...
13,2,Under Armour Men's Renegade D Mid Football Cl,,89.99,http://images.acmesports.sports/Under+Armour+Men%27s+Renegade+D+Mid+Football+Cleat
```

Publish the pipeline, then click on "Debug." Within the specified timeframe (10 seconds in this case), upload the file ("products.csv") to the Blob storage. Finally, review the pipeline run for the desired outcomes.

As pipeline has failed it is showing message highlighted in below screenshot **“pipeline has failed”**

.

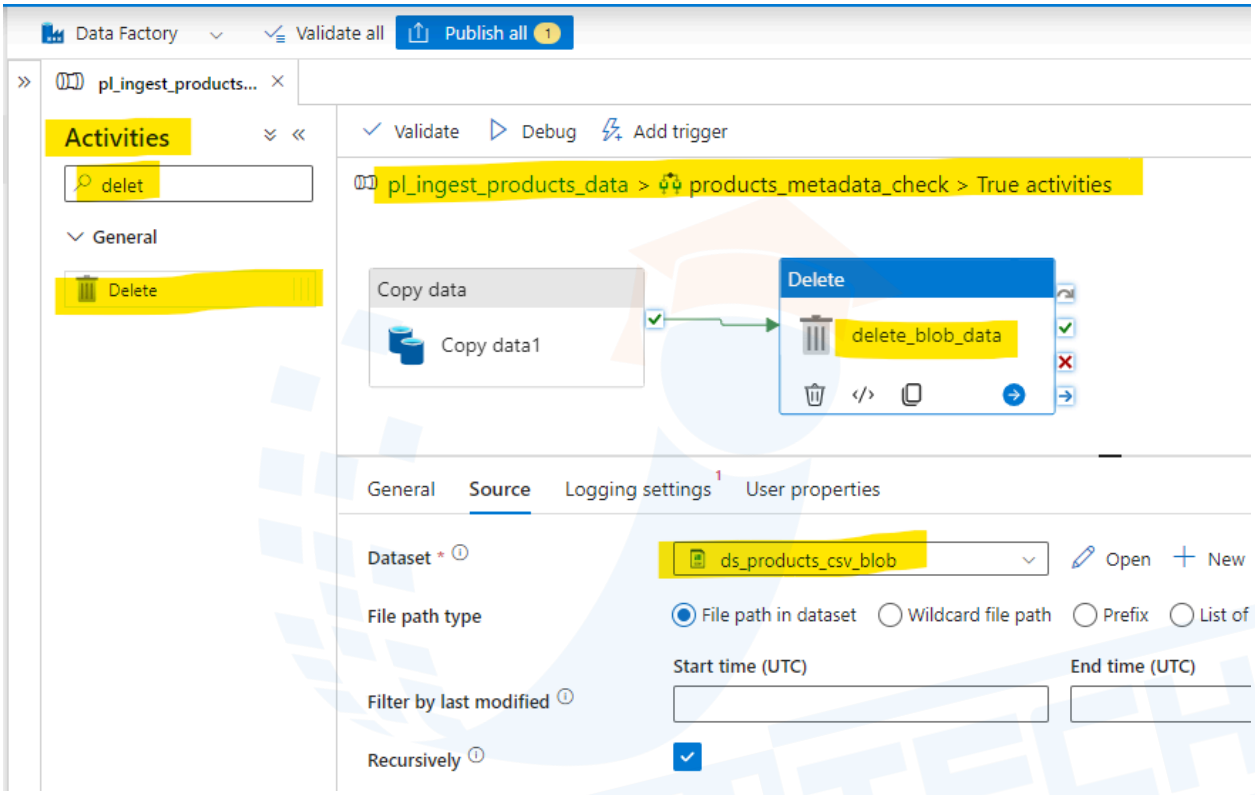


Also the column count is 5 refer below screenshot.

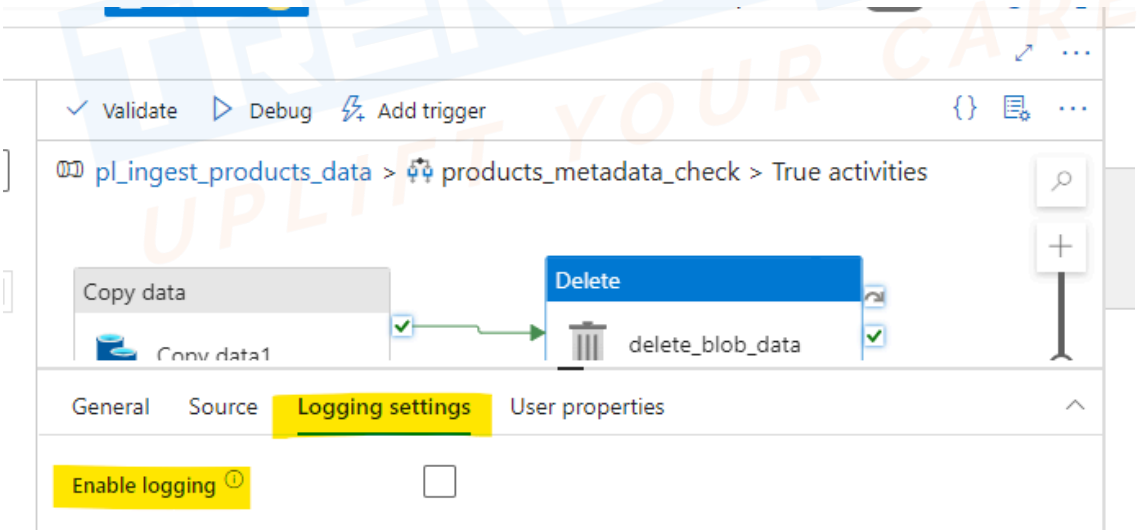
To delete the source data after copy activity has been done.

Follow below steps:

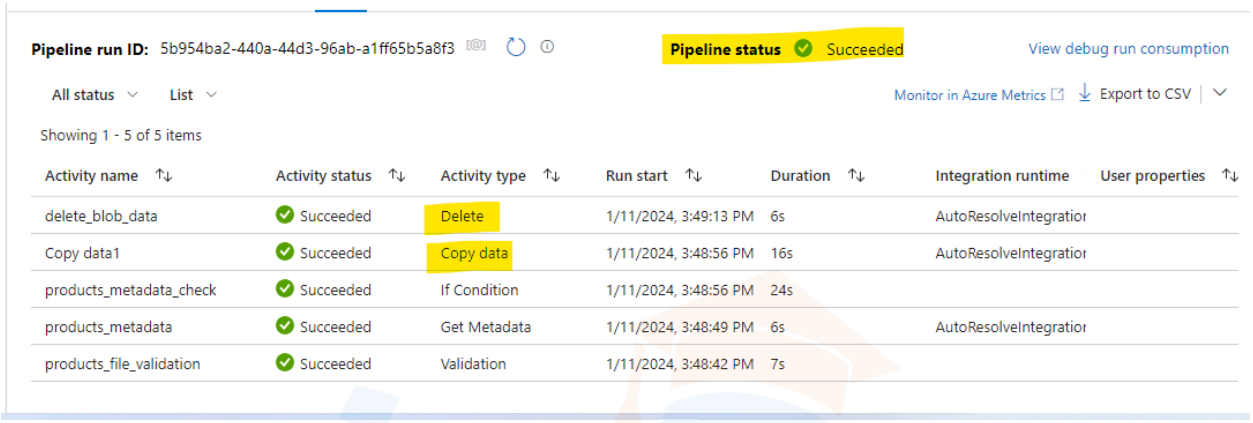
Drag delete activity in **“True”** blogger of if condition=>Renamed it **(“delete_blob_data”)** => In Source add the blob storage dataset of pipeline.csv **“ds_products_csv_blob”**



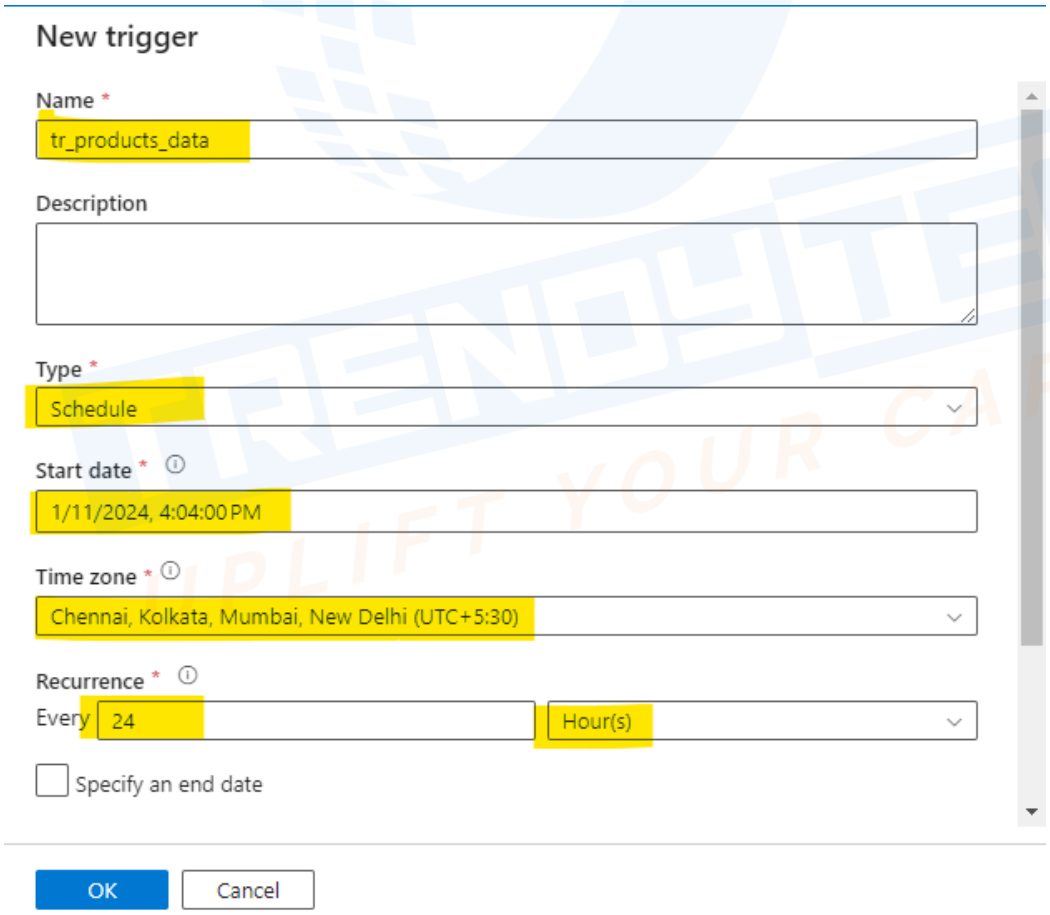
Also disable logging

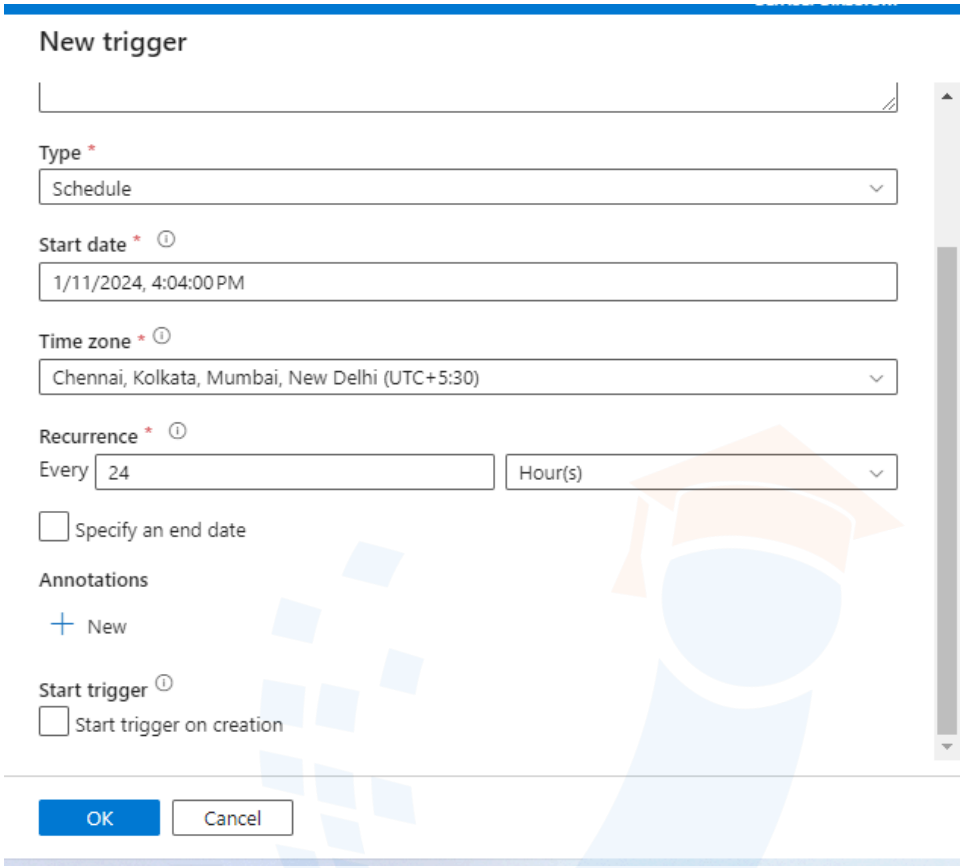


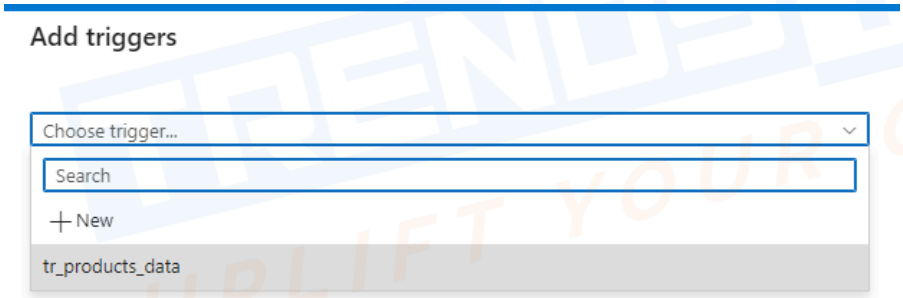
Publish the pipeline, then click on "Debug." Within the specified timeframe (10 seconds in this case), upload the file ("products.csv") to the Blob storage. Finally, review the pipeline run for the desired outcomes.

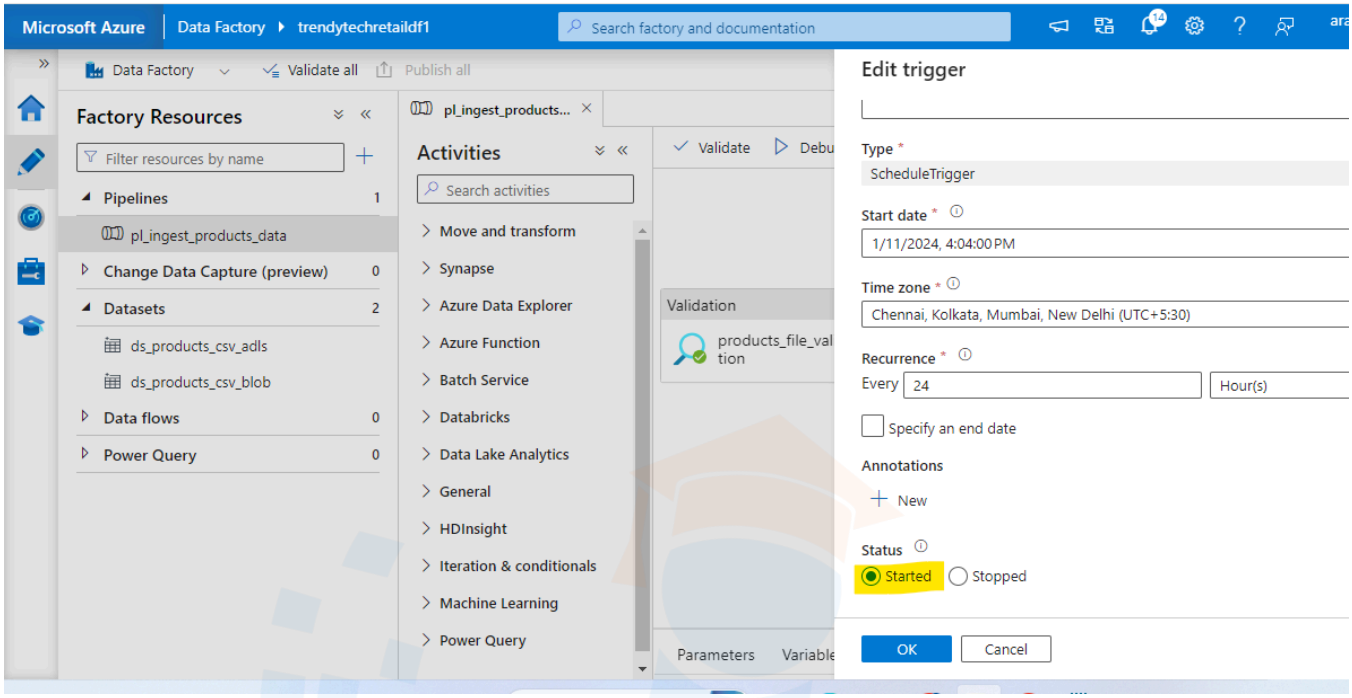


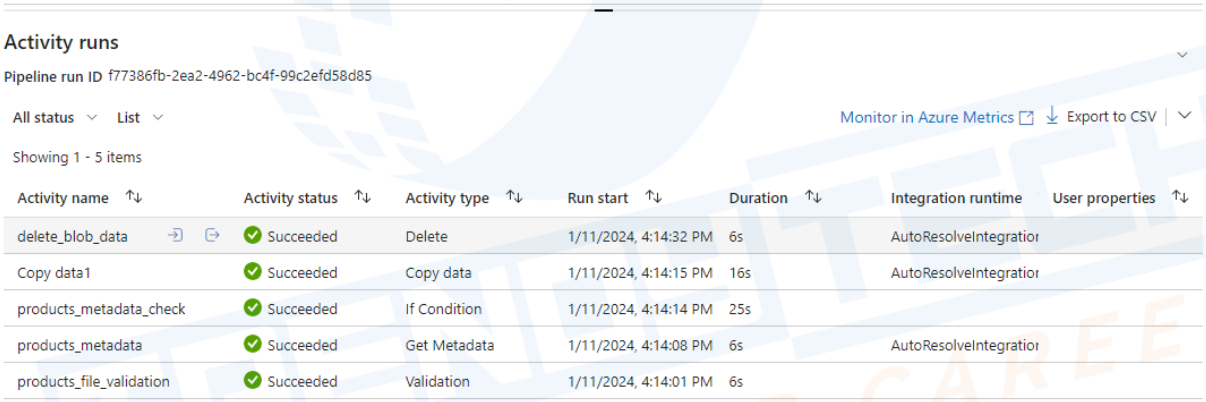
4.Create a Trigger and associate it to the pipeline to schedule the pipeline execution.



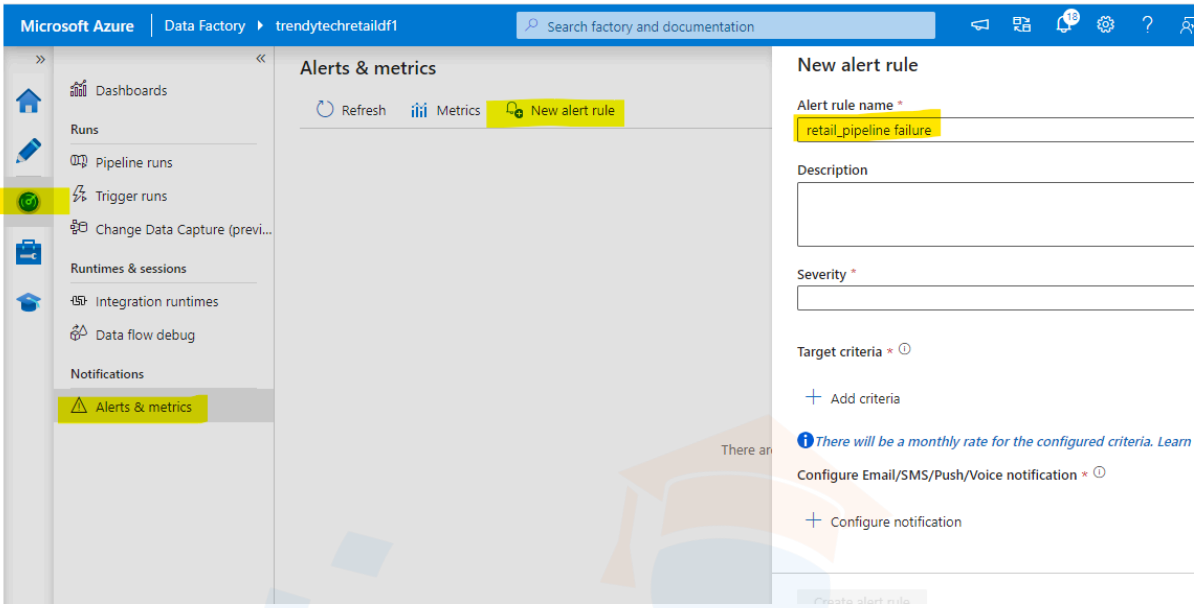

New trigger
Type *
Schedule
Start date *
1/11/2024, 4:04:00 PM
Time zone *
Chennai, Kolkata, Mumbai, New Delhi (UTC+5:30)
Recurrence *
Every
24
Hour(s)
Specify an end date
Annotations
New
Start trigger
Start trigger on creation
OK
Cancel

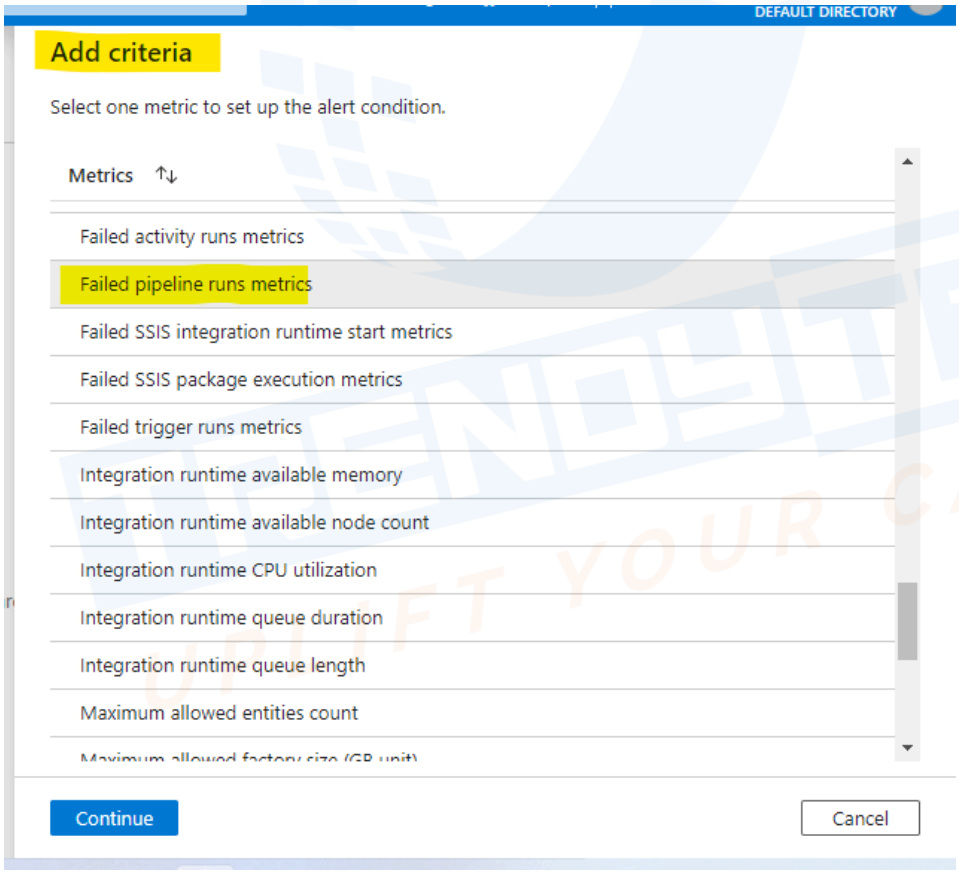

Add triggers
Choose trigger...
Search
New
tr_products_data



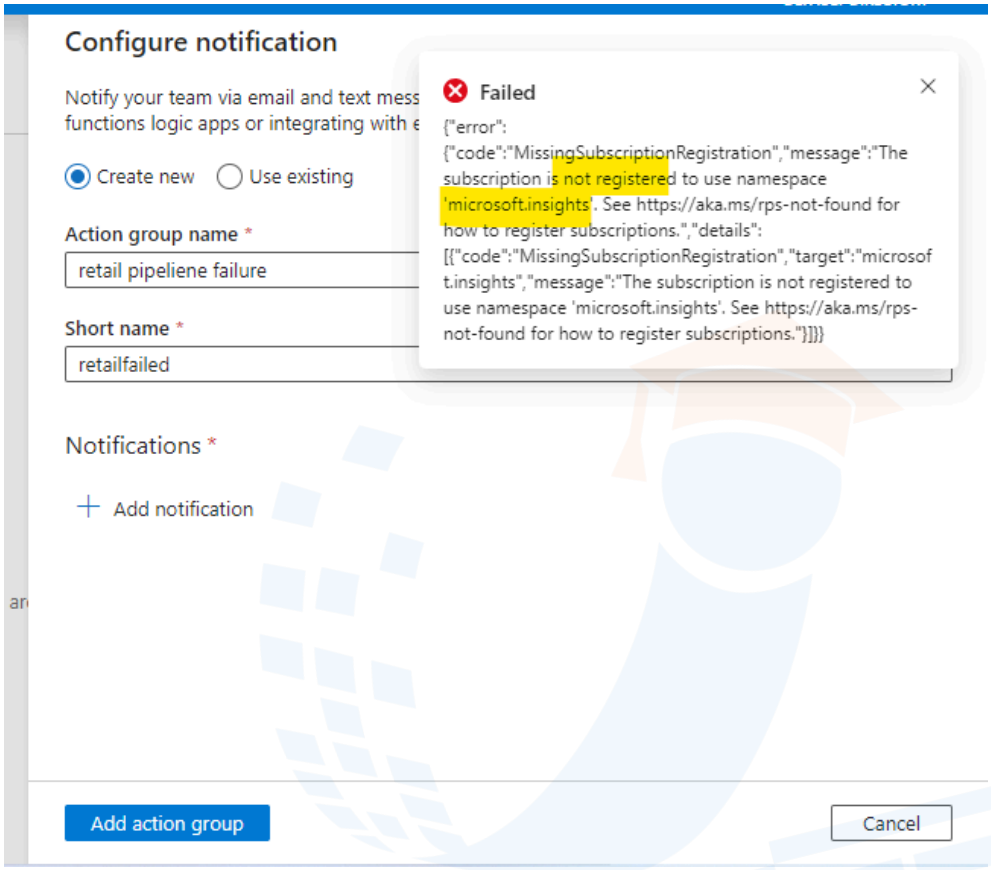


Steps to create the alert click on Monitor => Alert & Metrics => New Alert Rule :

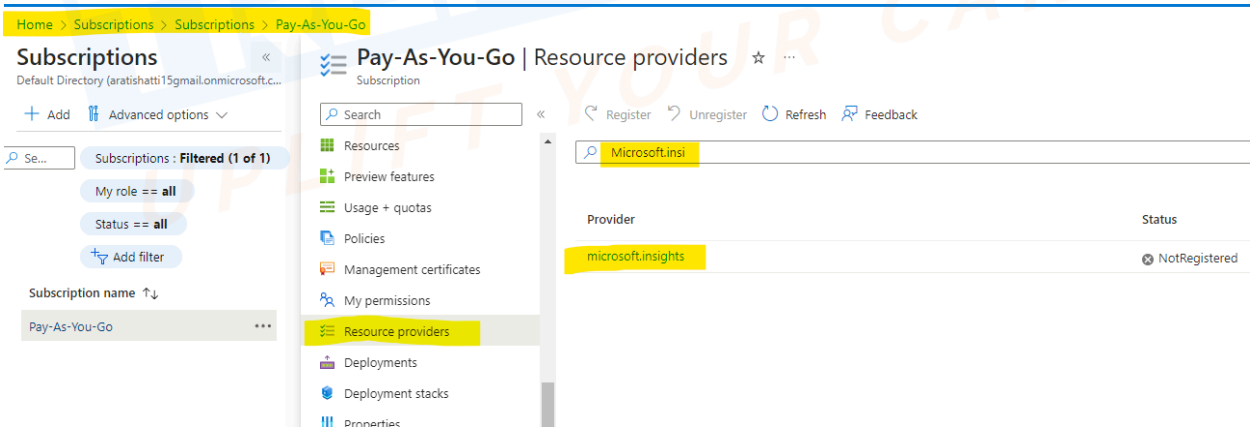

Microsoft Azure
Data Factory
trendytechretaildf1
Search factory and documentation
Dashboards
Runs
Pipeline runs
Trigger runs
Change Data Capture (previ...
Runtimes & sessions
Integration runtimes
Data flow debug
Notifications
Alerts & metrics
Alerts & metrics
Refresh
Metrics
New alert rule
New alert rule
Alert rule name *
retail_pipeline failure
Description
Severity *
Target criteria *
Add criteria
There will be a monthly rate for the configured criteria. Learn
Configure Email/SMS/Push/Voice notification *
Configure notification

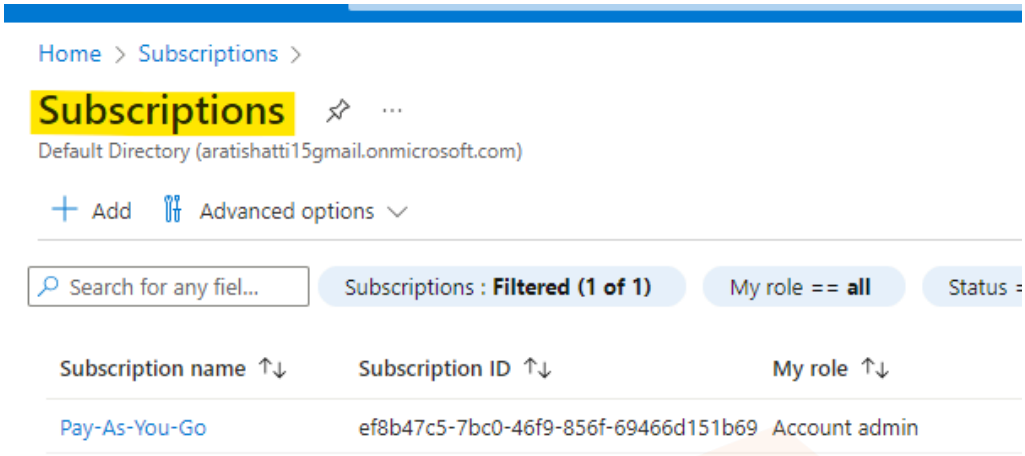

DEFAULT DIRECTORY
Add criteria
Select one metric to set up the alert condition.
Metrics
Failed activity runs metrics
Failed pipeline runs metrics
Failed SSIS integration runtime start metrics
Failed SSIS package execution metrics
Failed trigger runs metrics
Integration runtime available memory
Integration runtime available node count
Integration runtime CPU utilization
Integration runtime queue duration
Integration runtime queue length
Maximum allowed entities count
Continue
Cancel

Note: In case you encounter an error during the configuration of notifications, please refer to the following steps.

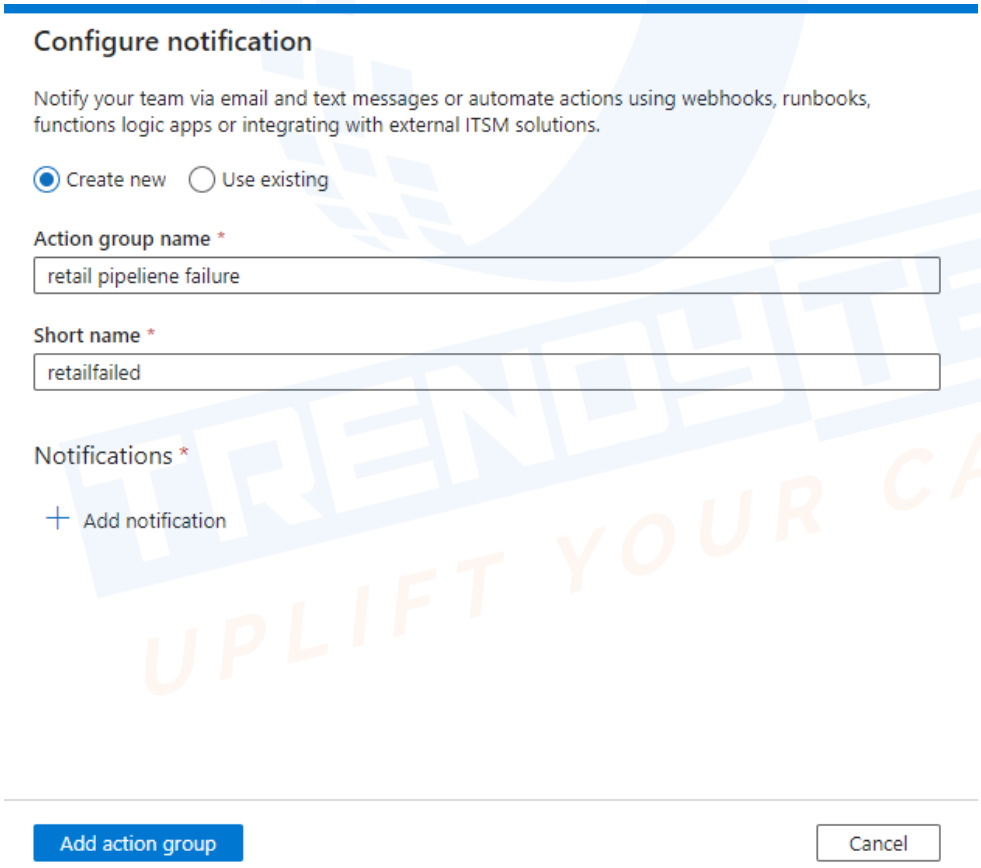


Go to Subscriptions => select your subscription => Resource Providers => microsoft.insights (register this service)

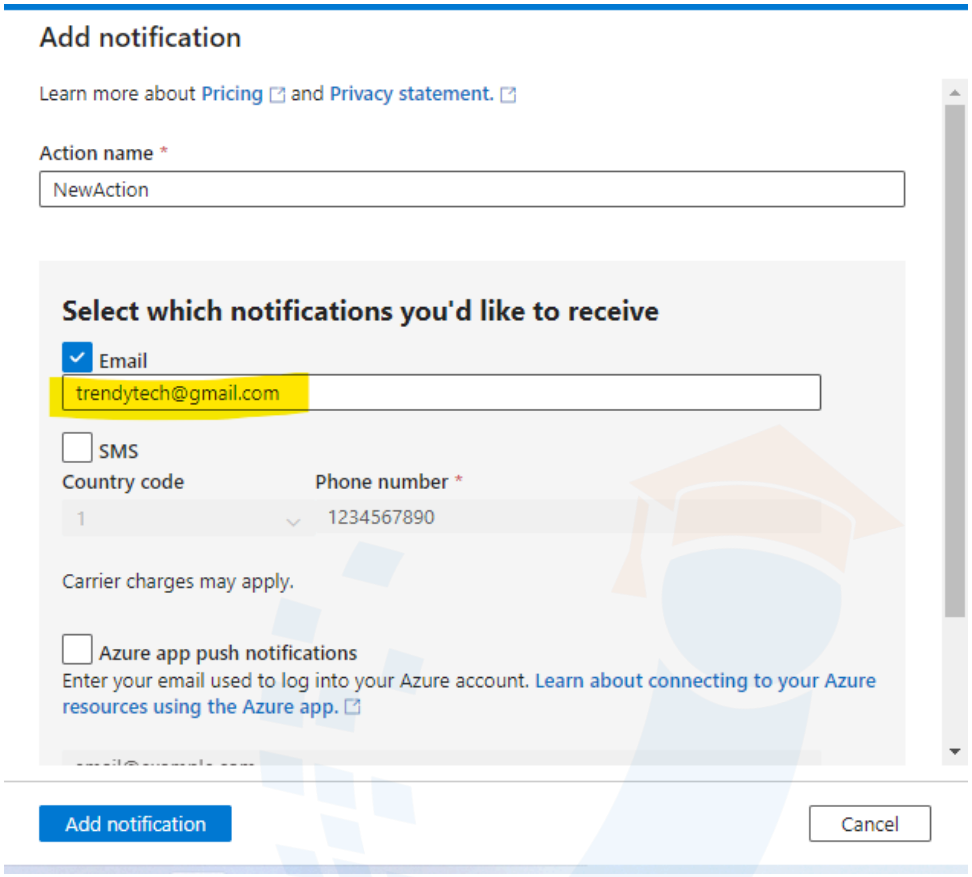




Continuation of Alert Creation :

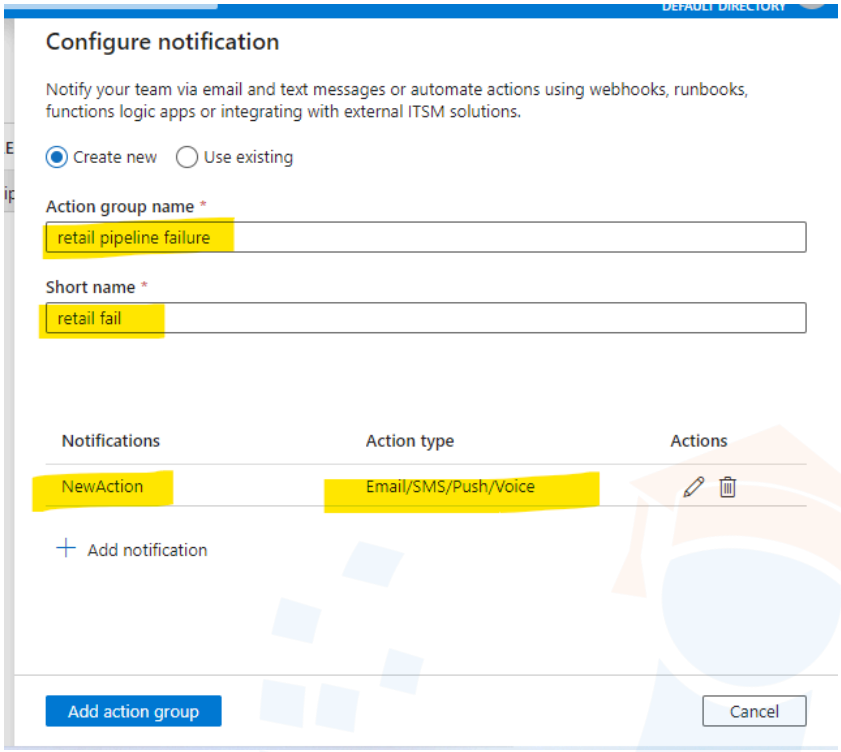


Add the **email id** on which you want to receive the emails and click on **“Add notification”.**



Mention the name of the Action group and its short name.

TRENDYTECH
UPLIFT YOUR CAREER!

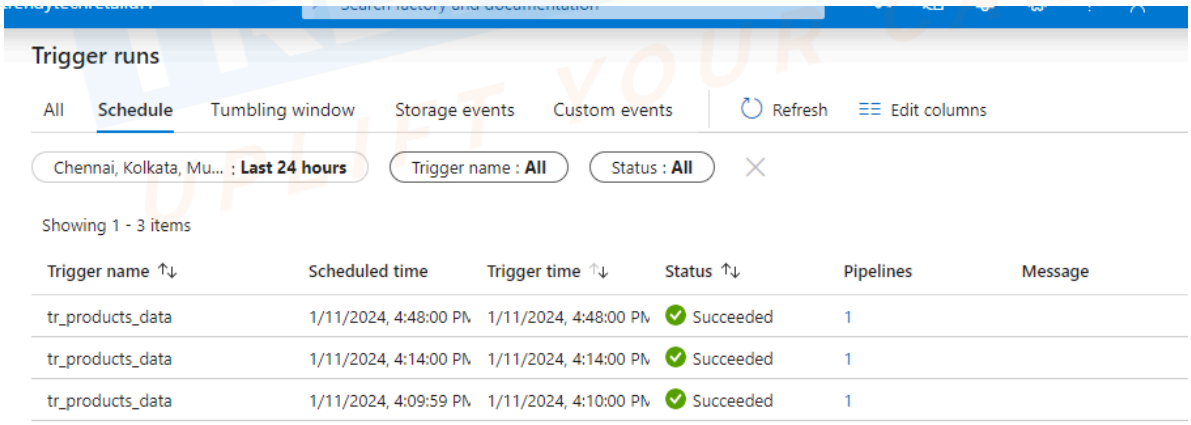


And Create Alert.

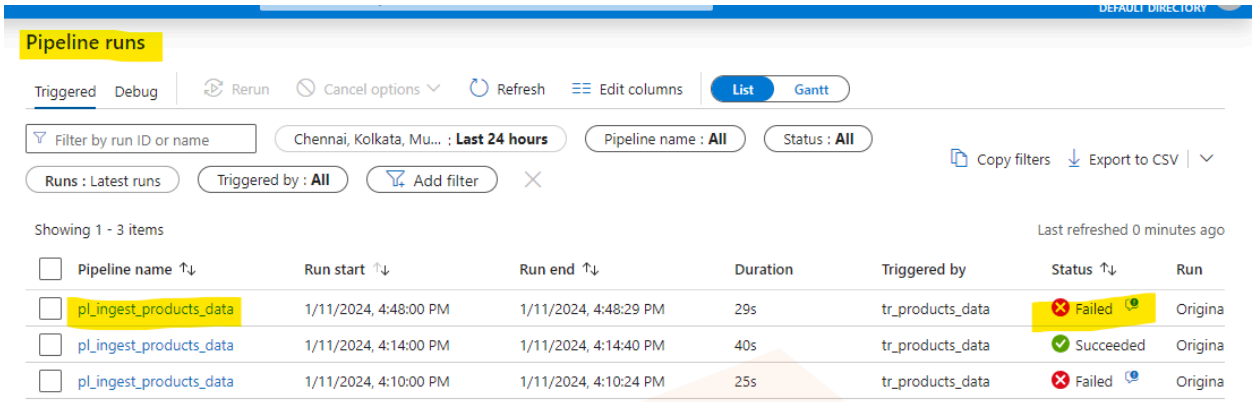
Now again set the trigger for future time for the pipeline with the file which has 5 columns.

(Use a copy of the "pipeline.csv" file that was created to a different location, such as the desktop by removing the value of the last column in the first row. Utilize this modified file for debugging purpose.)

Note : In this case Trigger won't fail.

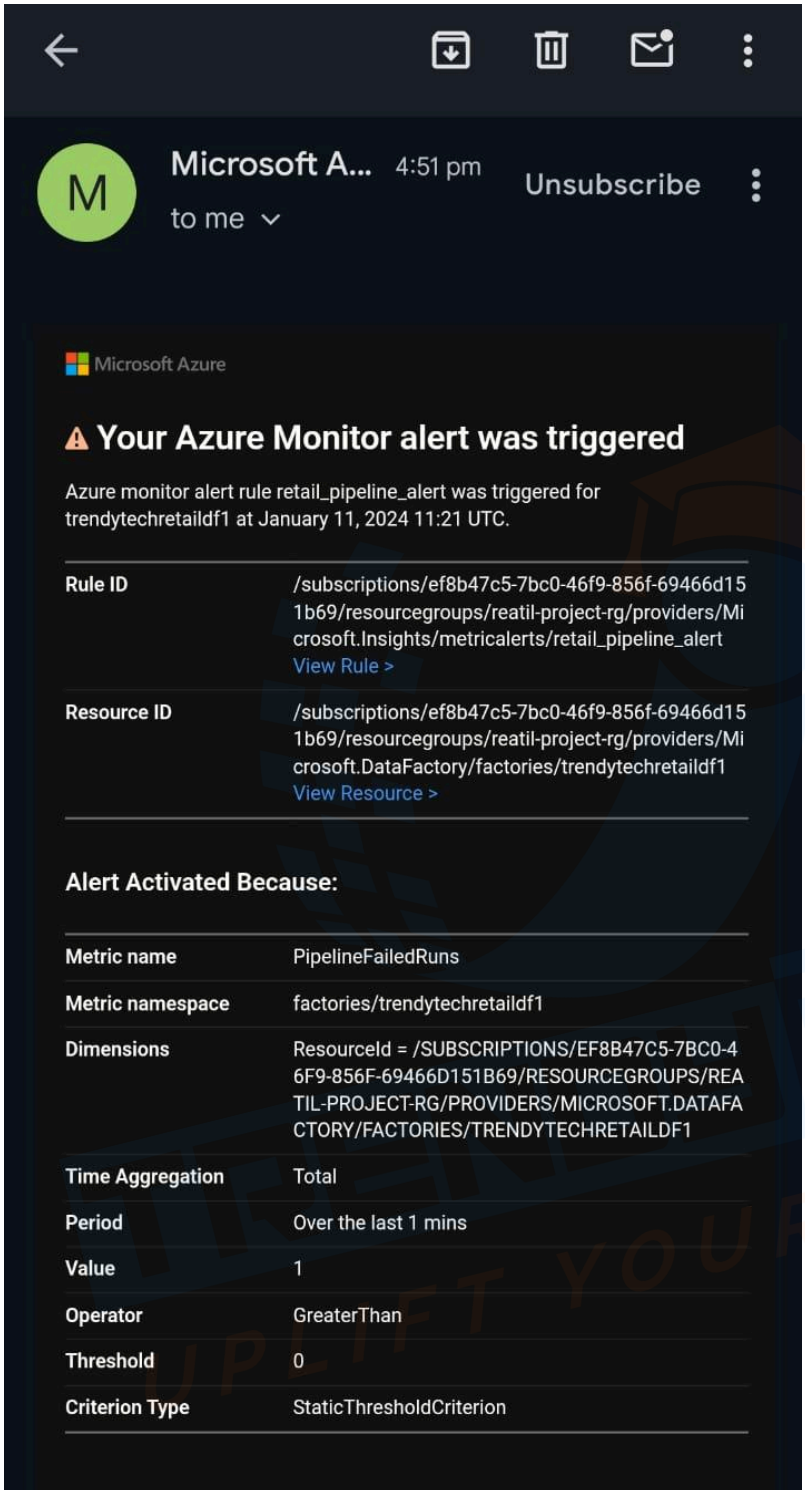


But pipeline will be failed refer the attached



Also in case of alert you will receive email over your registered mail id as follows

TRENDYTECH
UPLIFT YOUR CAREER!



Note: At the end delete all the resources that you have created.