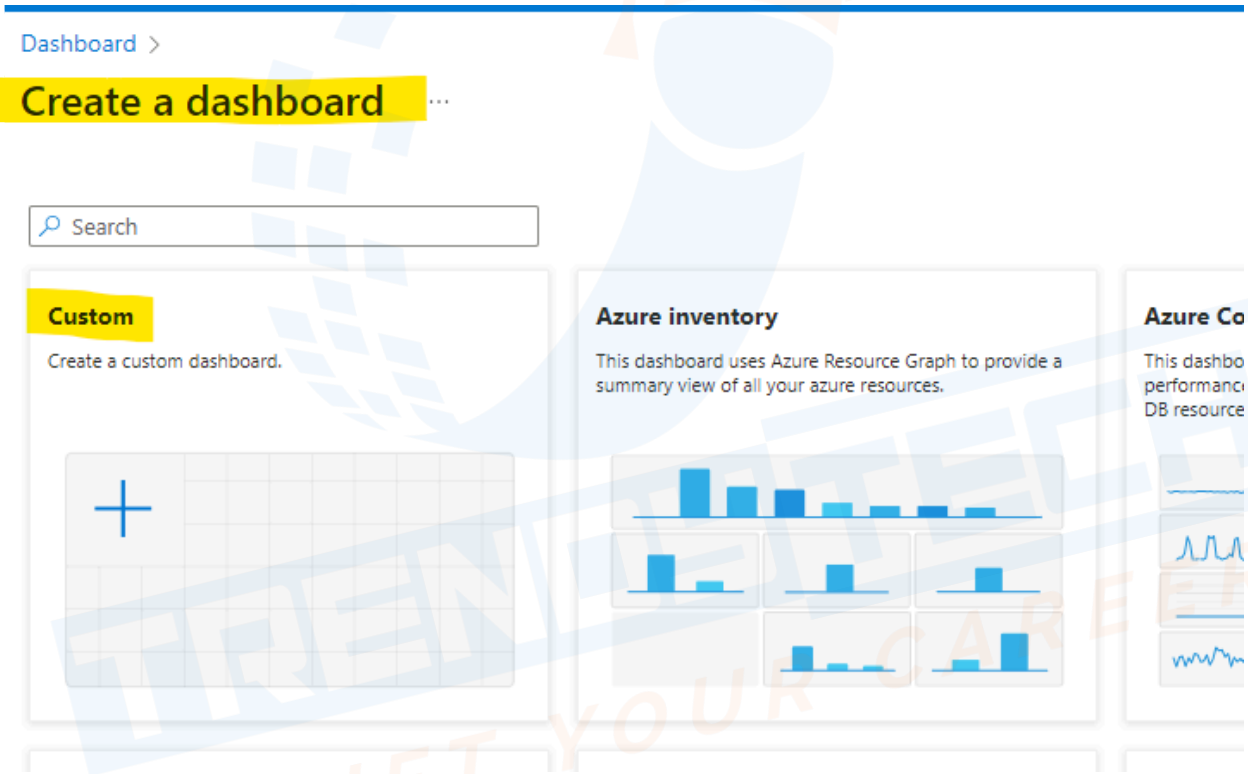
Use case 2 - Ingesting Data from Amazon S3 to Azure ADLS Gen2

Ingestion:

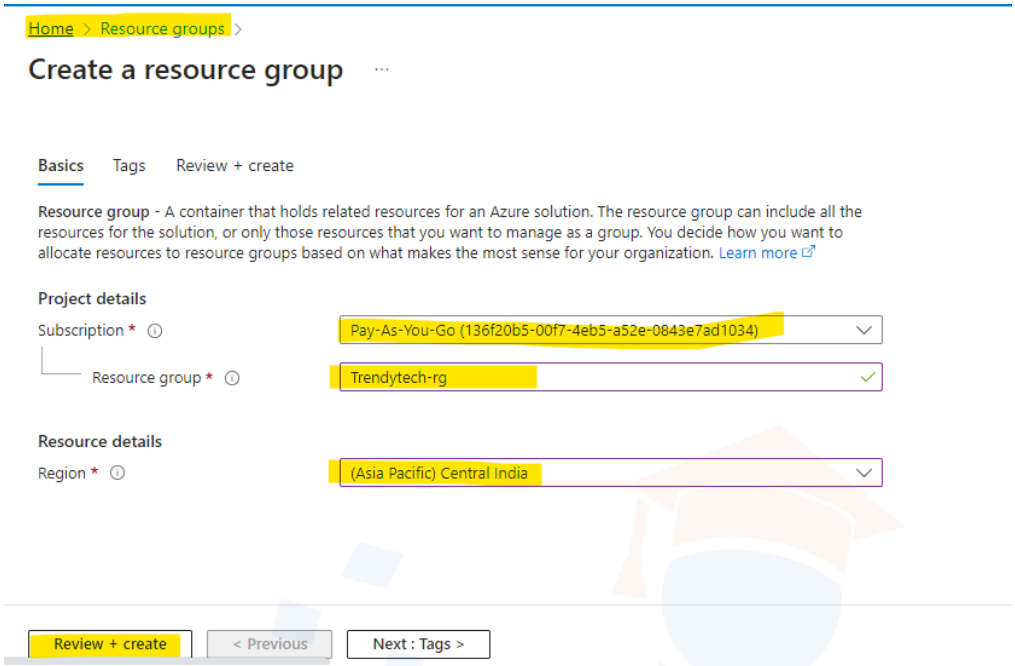
1. Create a Resource Group and pin it to the dashboard.

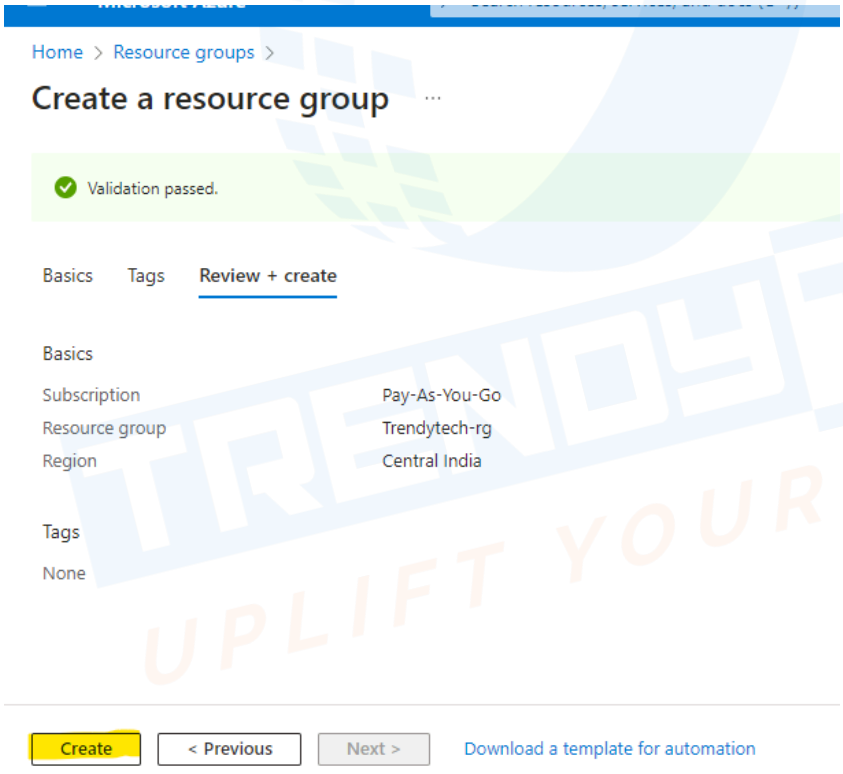
Create a Dashboard and a Resource Group for the project to organize the resources related to the project at one place.

To create dashboard: Click on Dashboard => Create => Custom =>

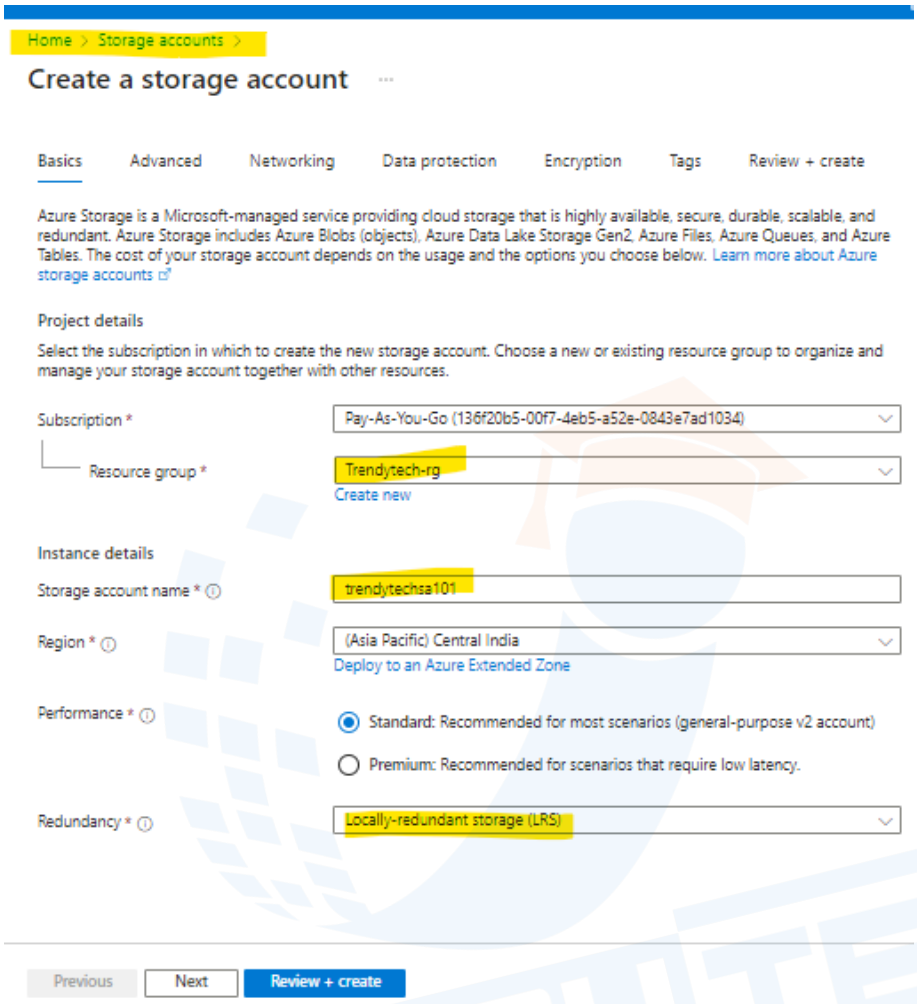


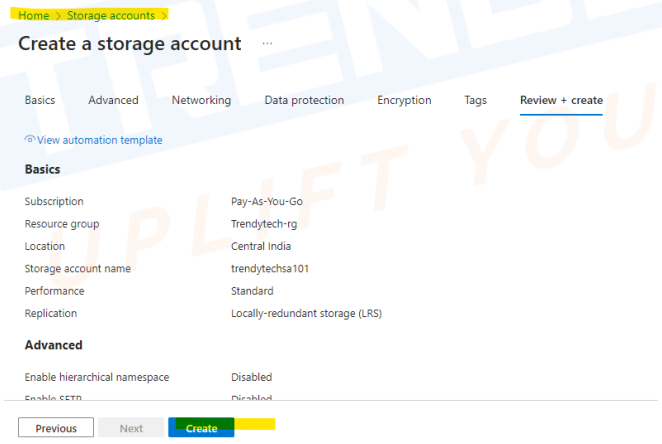
To create Resource group (trendytech-rg):





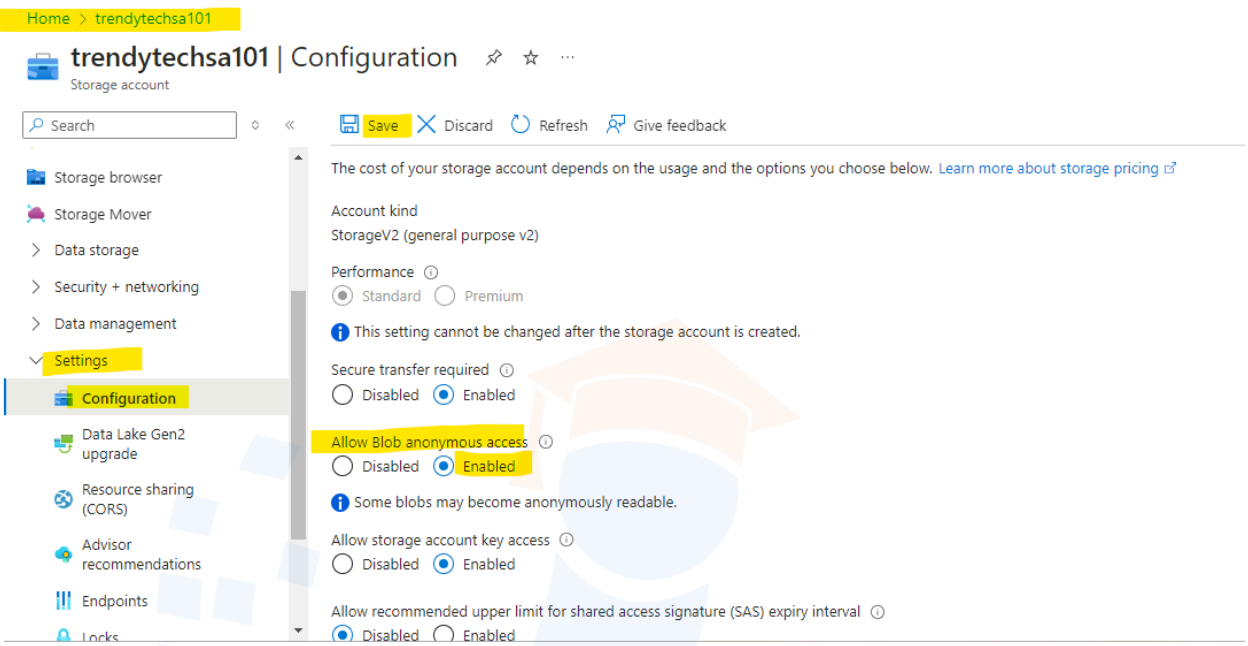
2. Create Storage Accounts a Normal Blob Storage Account (trendytechsa101)



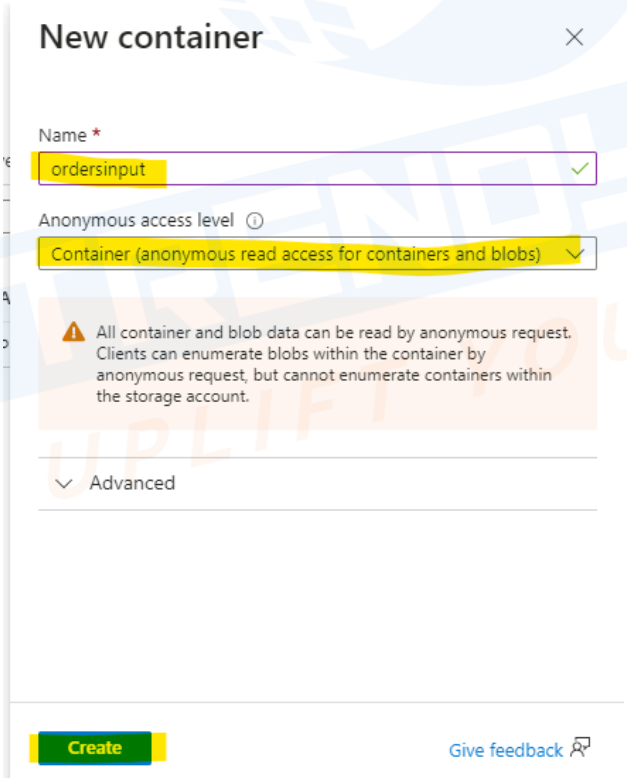


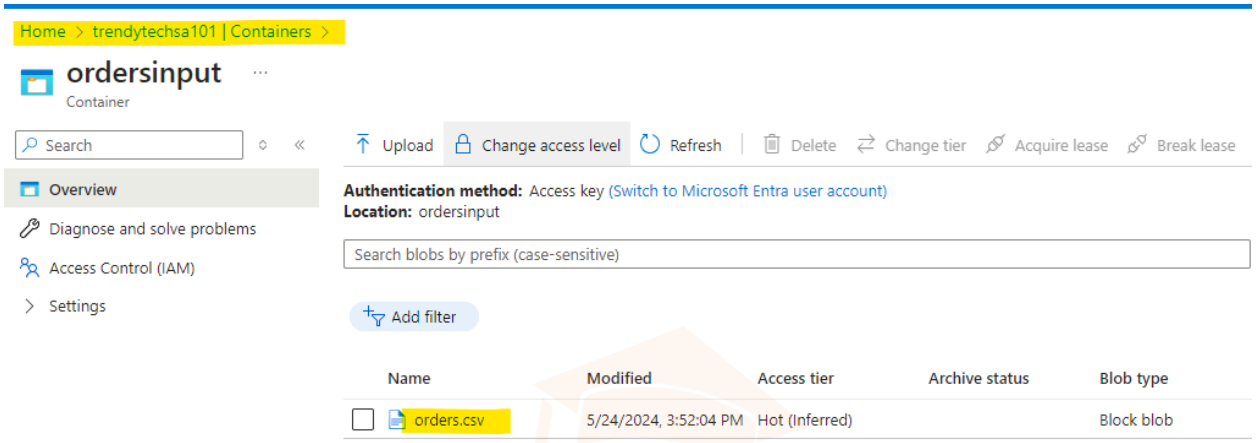
Click on review and create and then create the storage account.

Note: Allow the “blob anonymous access” for the storage account

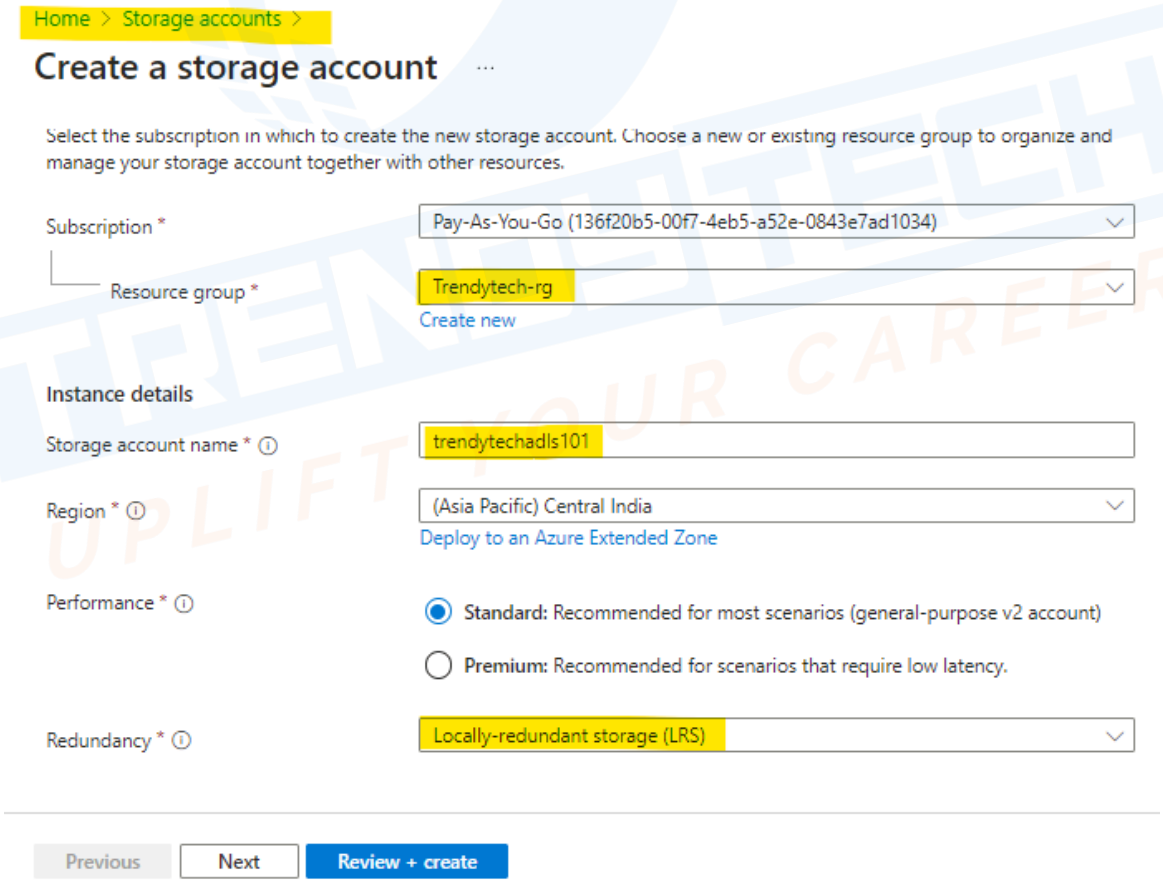


Now create the folder order_input and upload the file orders.csv in it.

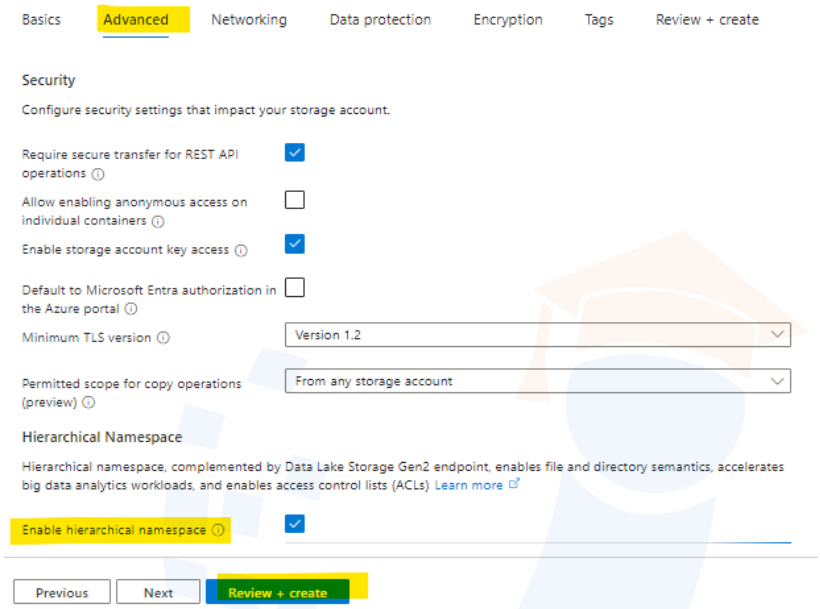




3. ADLS Gen2 Storage Account (with hierarchical namespace enabled)(trendytechadls101)

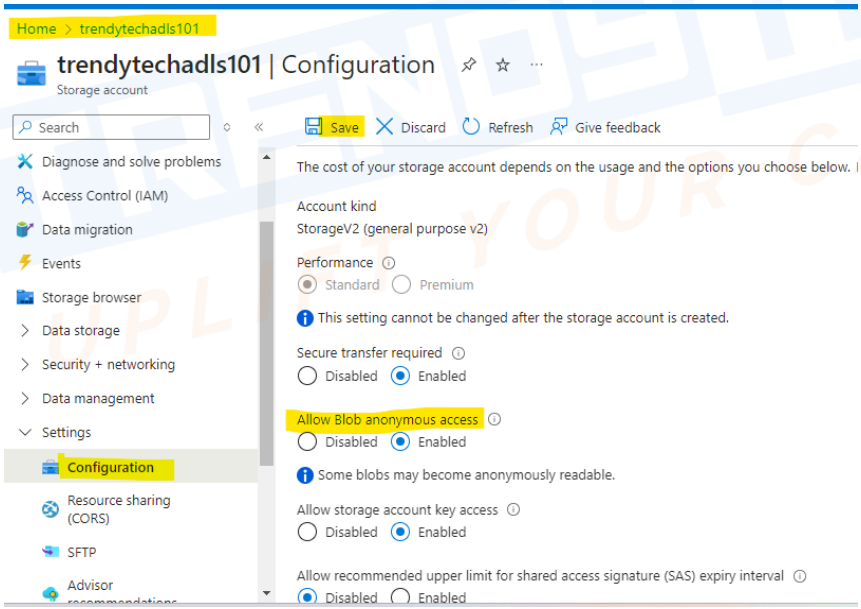


In advance tab, select option “**Enable hierarchical namespace**”, refer attached screenshot

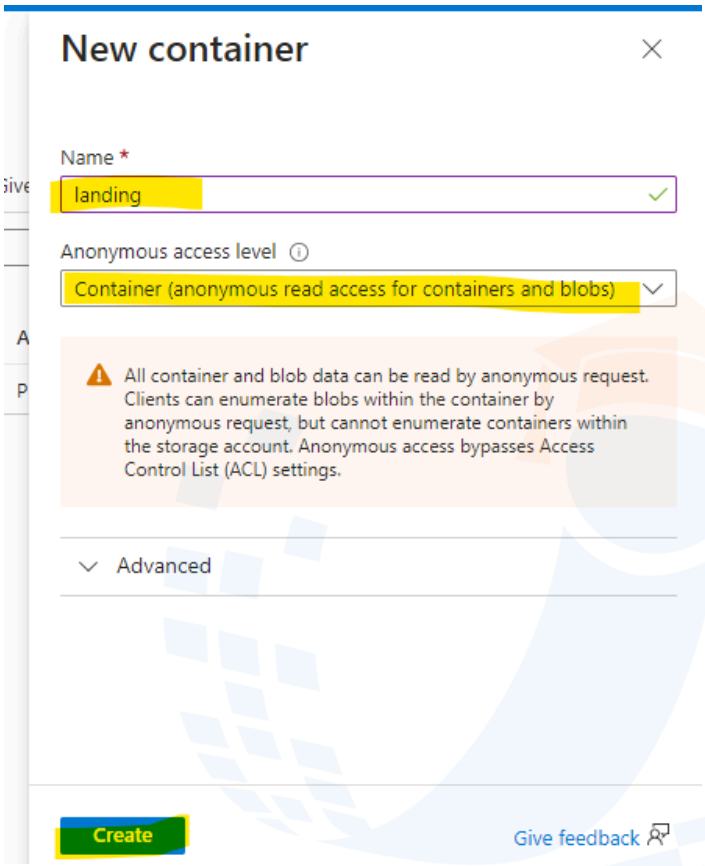


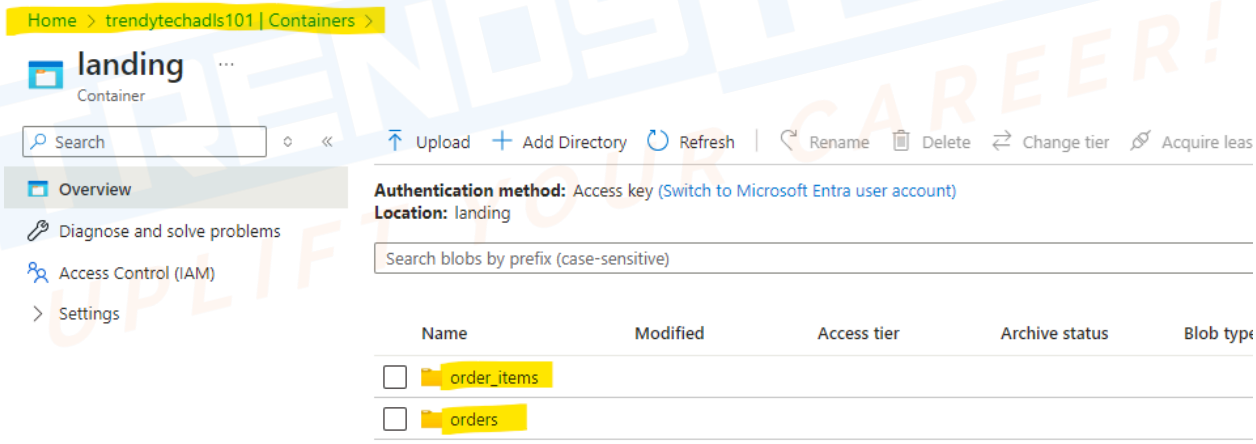
Click on review and create and then create the storage account.

Note: Allow the “blob anonymous access” for the storage account



Create the container “landing” in the adls gen2 storage and in landing create the folders orders and order_items





4. Create an Azure Data Factory within a Resource Group (trendytech-df-1001)

Home > Data factories >

Create Data Factory

Basics | Git configuration | Networking | Advanced | Tags | Review + create

One-click to create data factory with sample pipeline and datasets. Try it

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * : Pay-As-You-Go (136f20b5-00f7-4eb5-a52e-0843e7ad1034)

Resource group * : Trendytech-rg

Create new

Instance details

Name * : trendytech-df-1001

Region * : East US

Version * : V2

Previous | Next | Review + create

Home > Data factories >

Create Data Factory

Basics | Git configuration | Networking | Advanced | Tags | Review + create

View automation template

TERMS

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. See the Azure Marketplace Terms for additional details.

Basics

Subscription	Pay-As-You-Go
Resource group	Trendytech-rg
Name	trendytech-df-1001
Region	East US
Version	V2

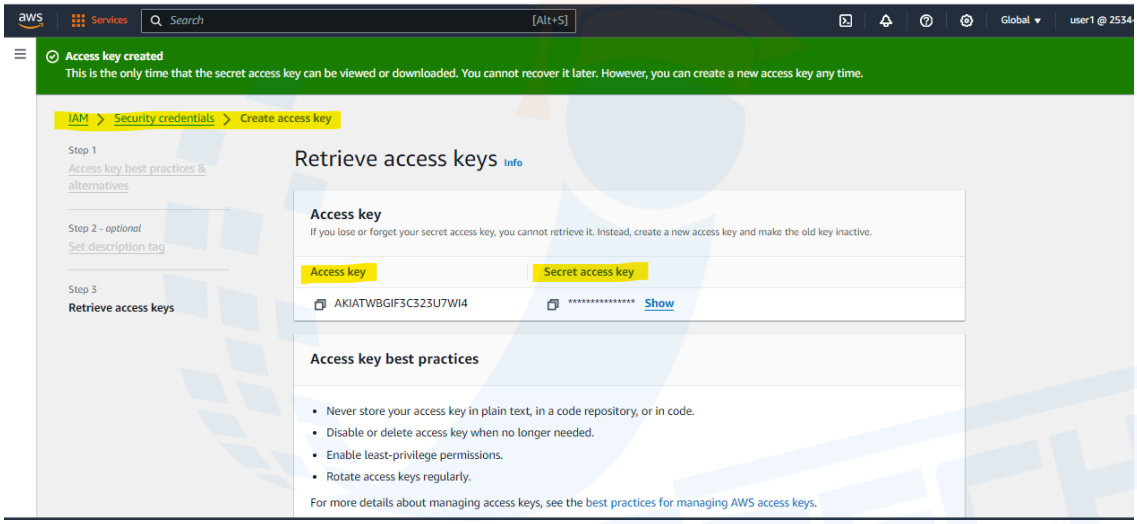
Previous | Next | Create

5. Create an Amazon Web Services Account with S3 Storage

First create the IAM user and then login to AWS console using that user and create access key.

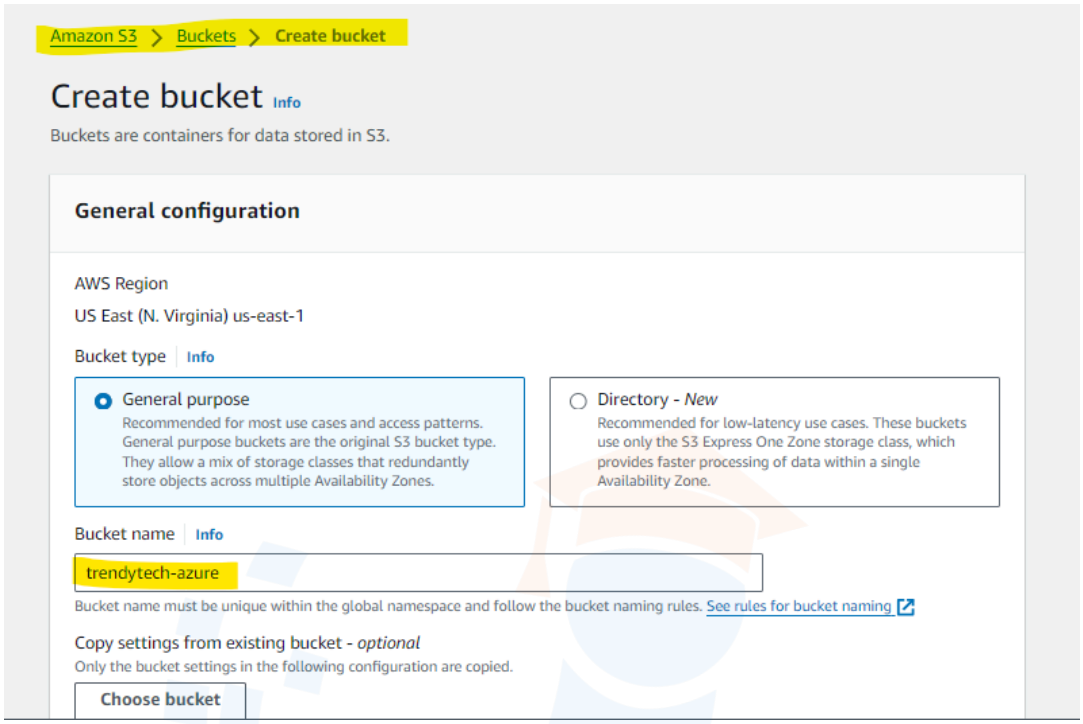
To create an access key you can refer to this document.

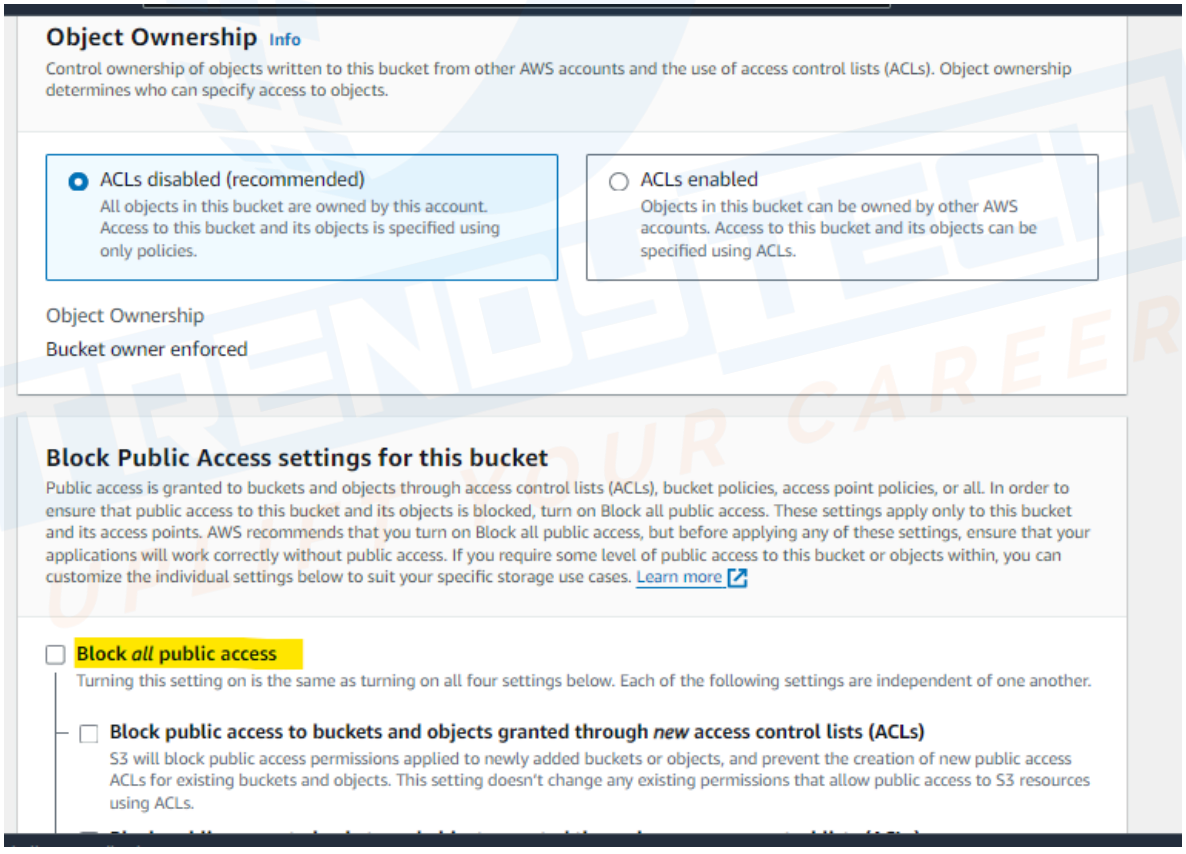
https://docs.aws.amazon.com/IAM/latest/UserGuide/id_credentials_access-keys.html



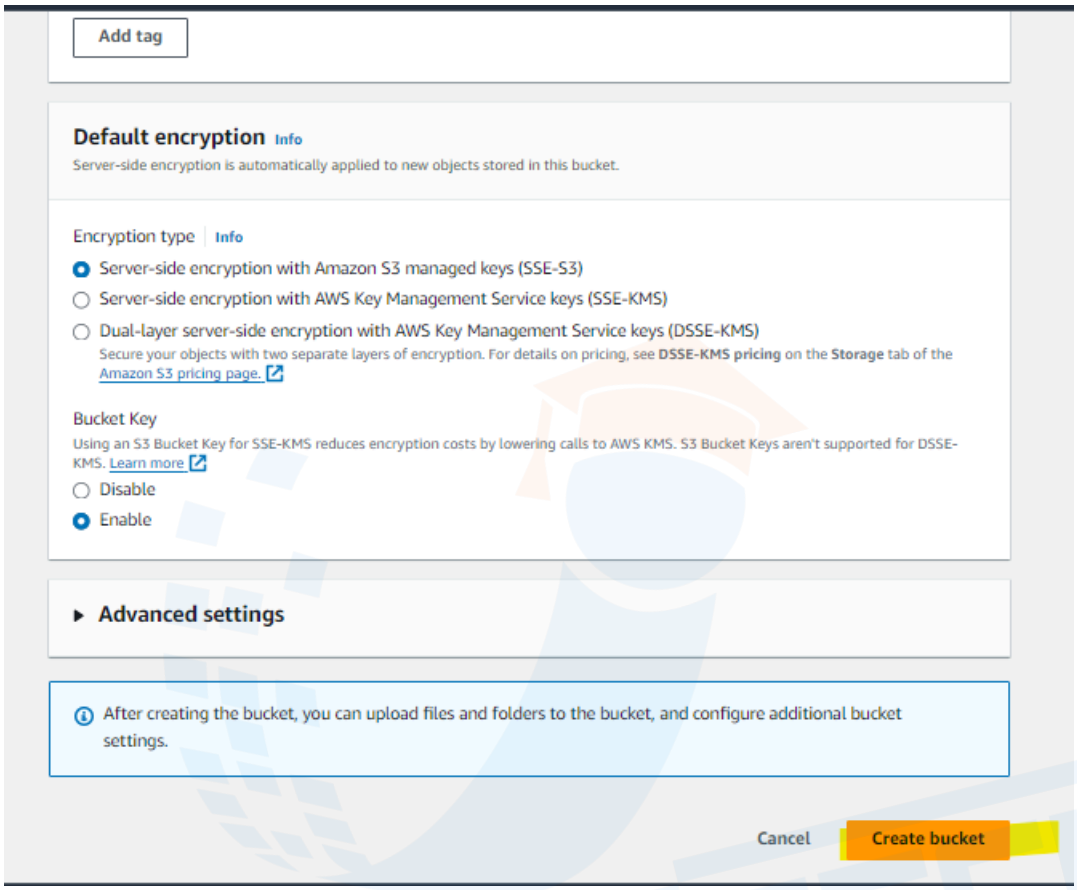
Create a bucket (trendytech-azure) within the S3 Storage and add the order_items.csv file to the bucket.

While creating the s3 bucket, deselect the option” Block public access” and keep other settings as it is and create the bucket.

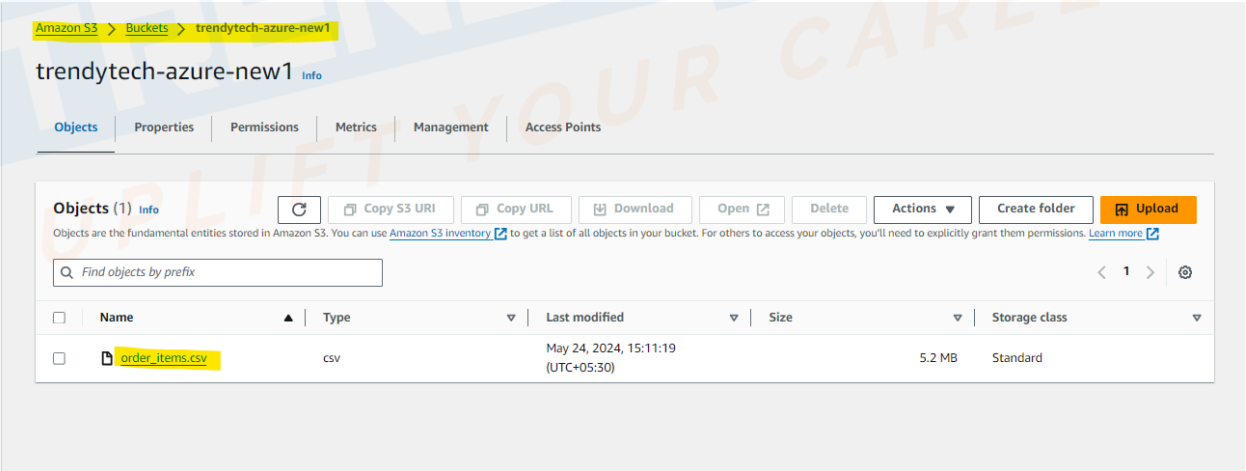

Amazon S3 > Buckets > Create bucket
Create bucket Info
Buckets are containers for data stored in S3.
General configuration
AWS Region
US East (N. Virginia) us-east-1
Bucket type Info
General purpose
Recommended for most use cases and access patterns. General purpose buckets are the original S3 bucket type. They allow a mix of storage classes that redundantly store objects across multiple Availability Zones.
Directory - New
Recommended for low-latency use cases. These buckets use only the S3 Express One Zone storage class, which provides faster processing of data within a single Availability Zone.
Bucket name Info
trendytech-azure
Bucket name must be unique within the global namespace and follow the bucket naming rules. See rules for bucket naming
Copy settings from existing bucket - optional
Only the bucket settings in the following configuration are copied.
Choose bucket


Object Ownership Info
Control ownership of objects written to this bucket from other AWS accounts and the use of access control lists (ACLs). Object ownership determines who can specify access to objects.
ACLs disabled (recommended)
All objects in this bucket are owned by this account. Access to this bucket and its objects is specified using only policies.
ACLs enabled
Objects in this bucket can be owned by other AWS accounts. Access to this bucket and its objects can be specified using ACLs.
Object Ownership
Bucket owner enforced
Block Public Access settings for this bucket
Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to this bucket and its objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to this bucket or objects within, you can customize the individual settings below to suit your specific storage use cases. Learn more
Block all public access
Turning this setting on is the same as turning on all four settings below. Each of the following settings are independent of one another.
Block public access to buckets and objects granted through new access control lists (ACLs)
S3 will block public access permissions applied to newly added buckets or objects, and prevent the creation of new public access ACLs for existing buckets and objects. This setting doesn't change any existing permissions that allow public access to S3 resources using ACLs.

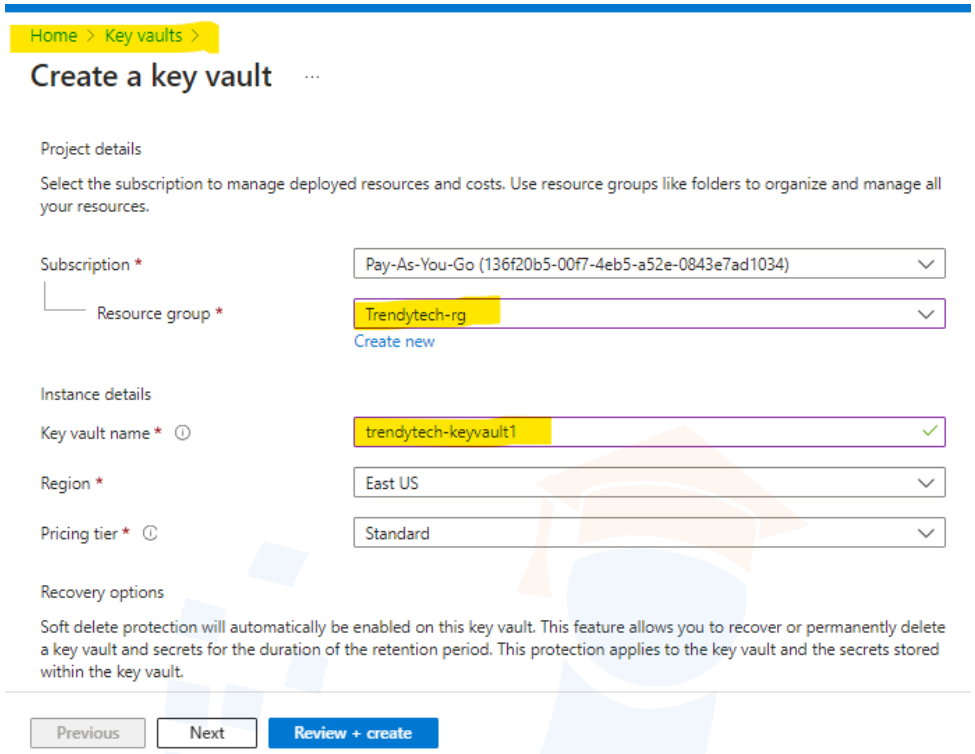
Then upload the file order_items.csv.



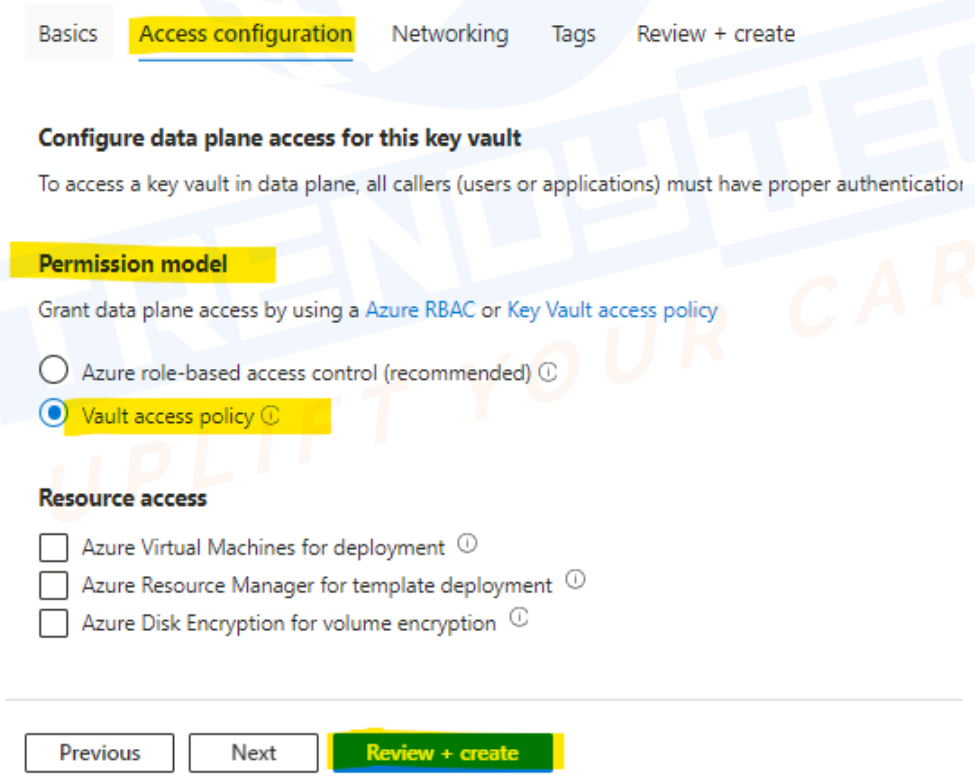
And once it is created upload file order_items.csv in it.

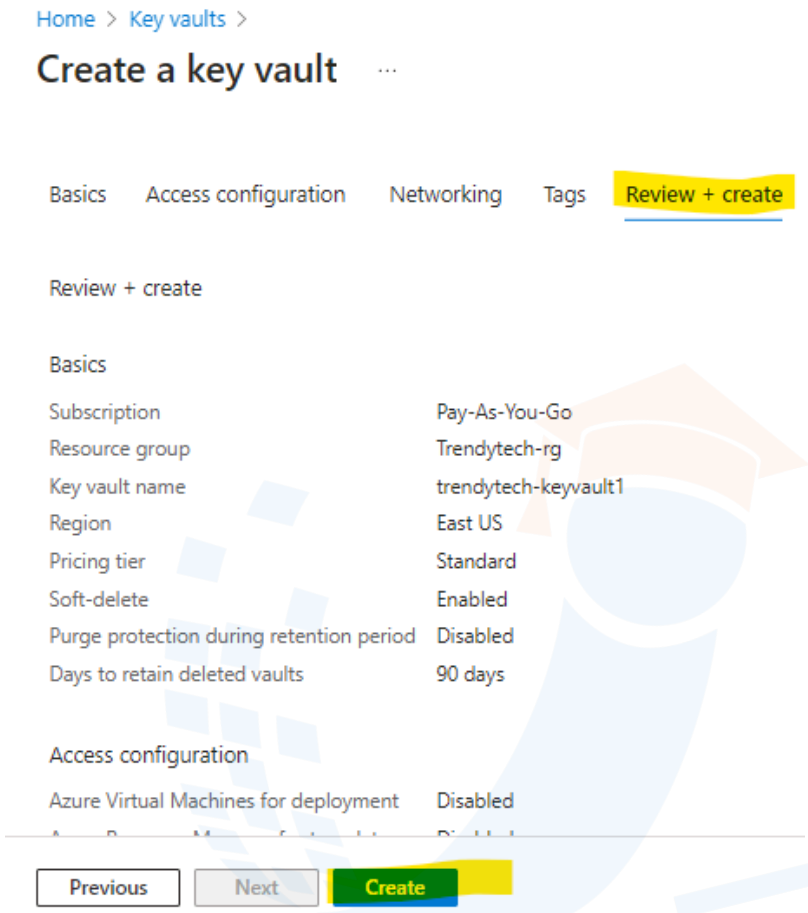


6. Create a Key vault (trendytech-keyvault1)



Note: In Access configuration select option “vault access policy”

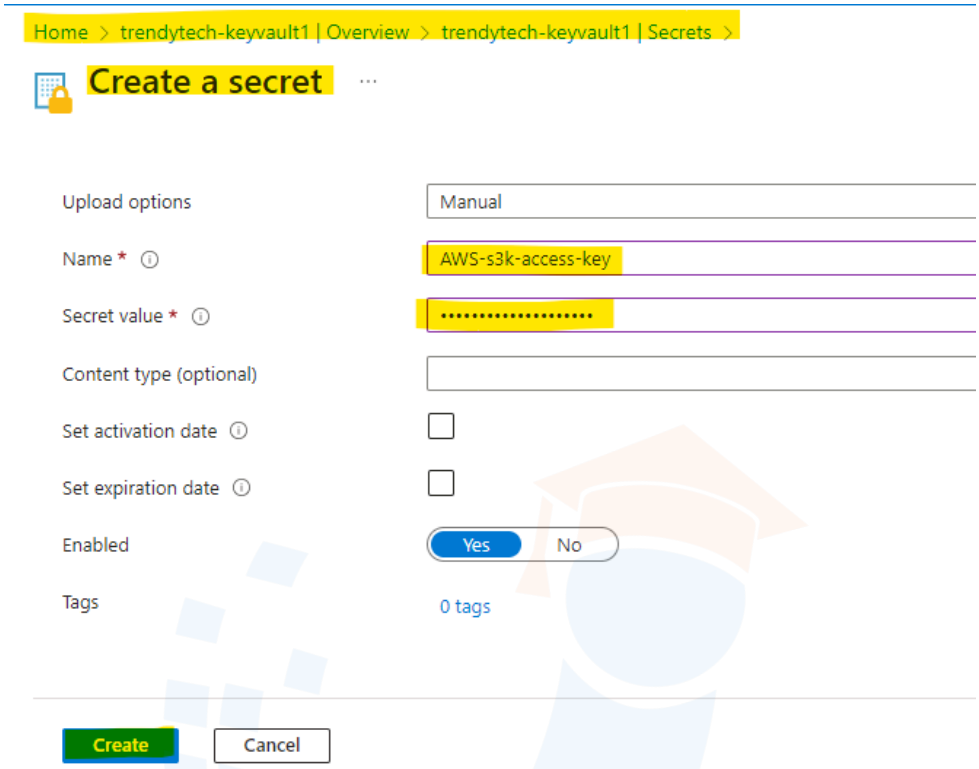


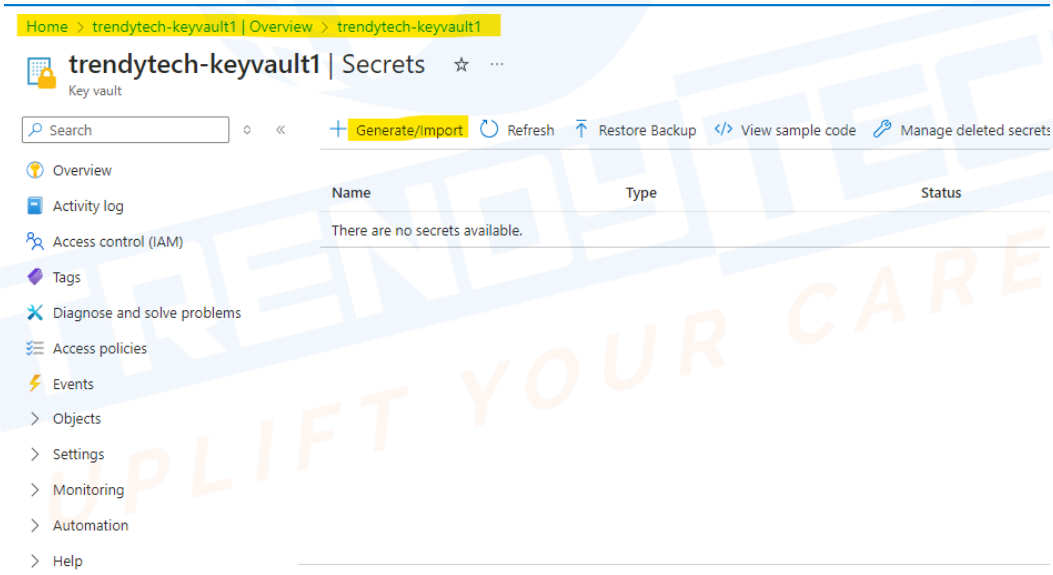


And click on the option “create”.

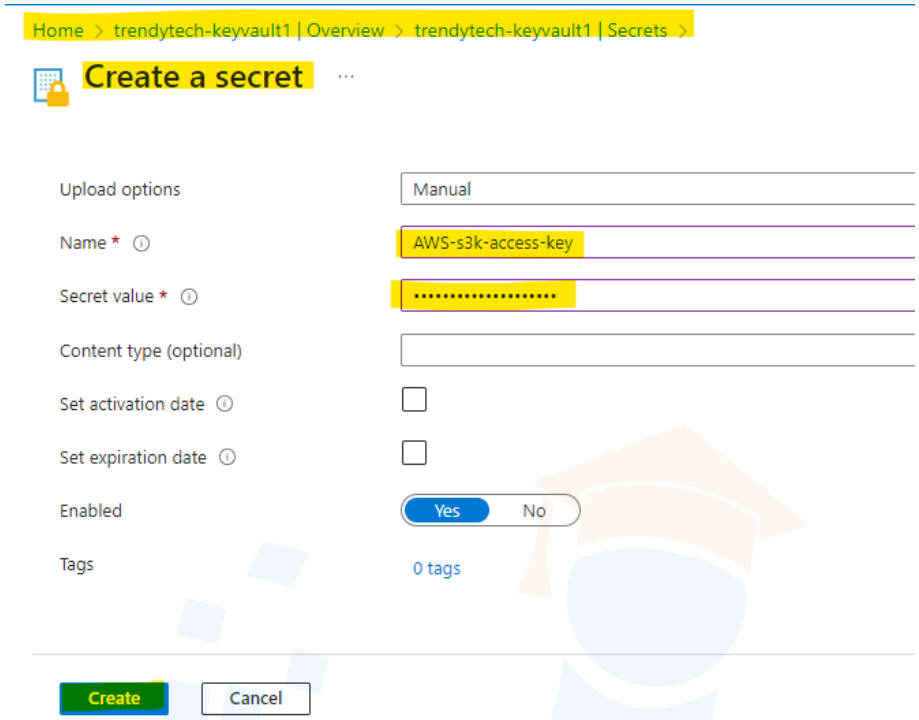
a. Store access key and access key secret in keyvault

In the key vault select the option “Generate/import”.

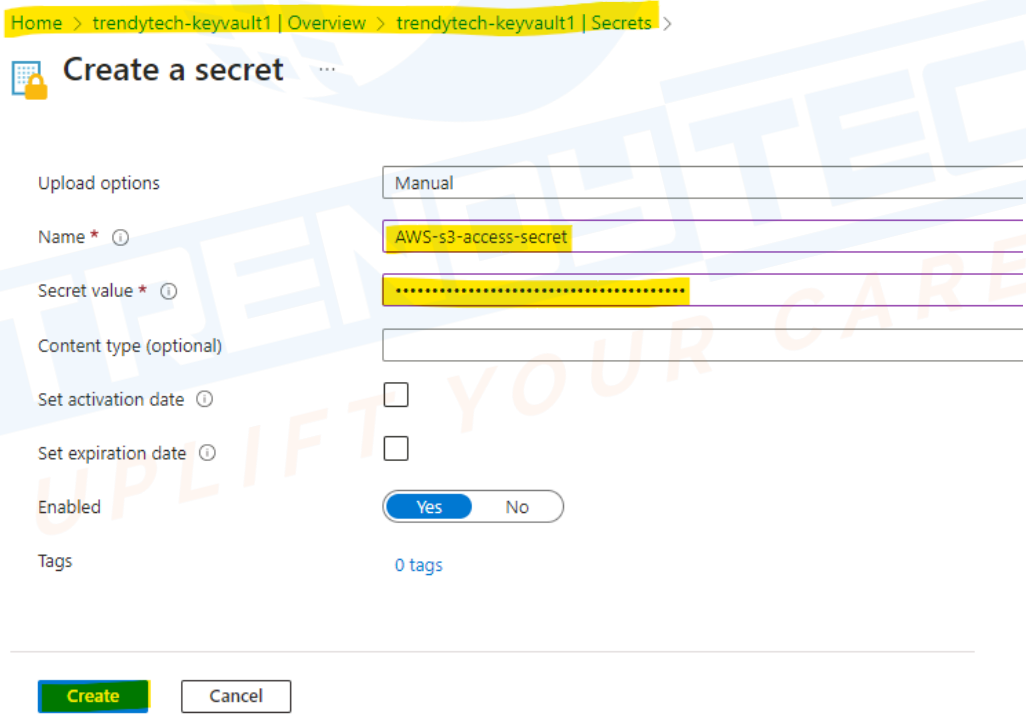


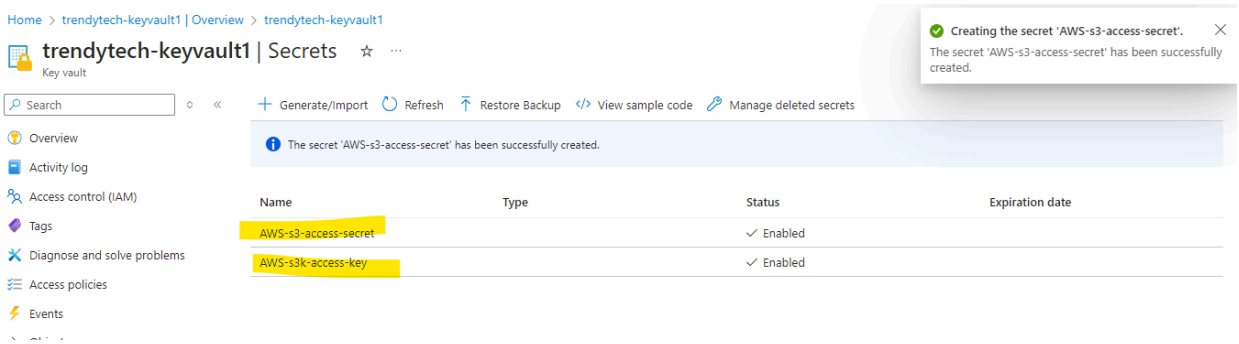


First we will store the access key

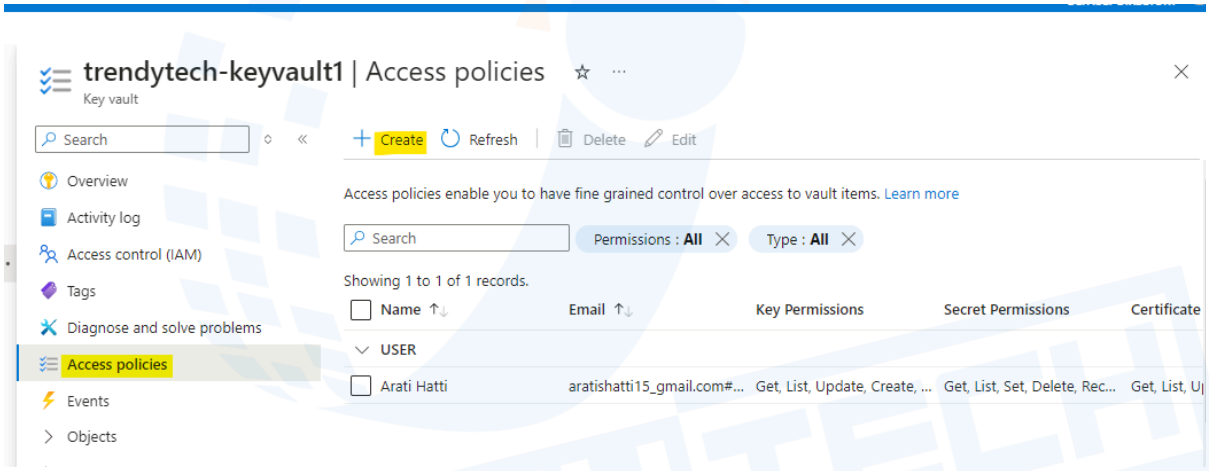


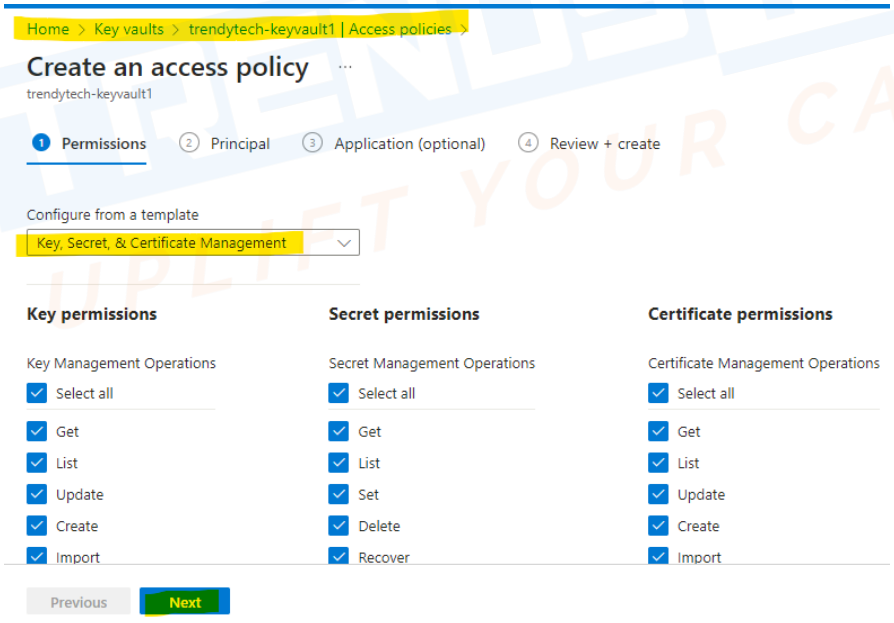
And click on option create to create an access key.

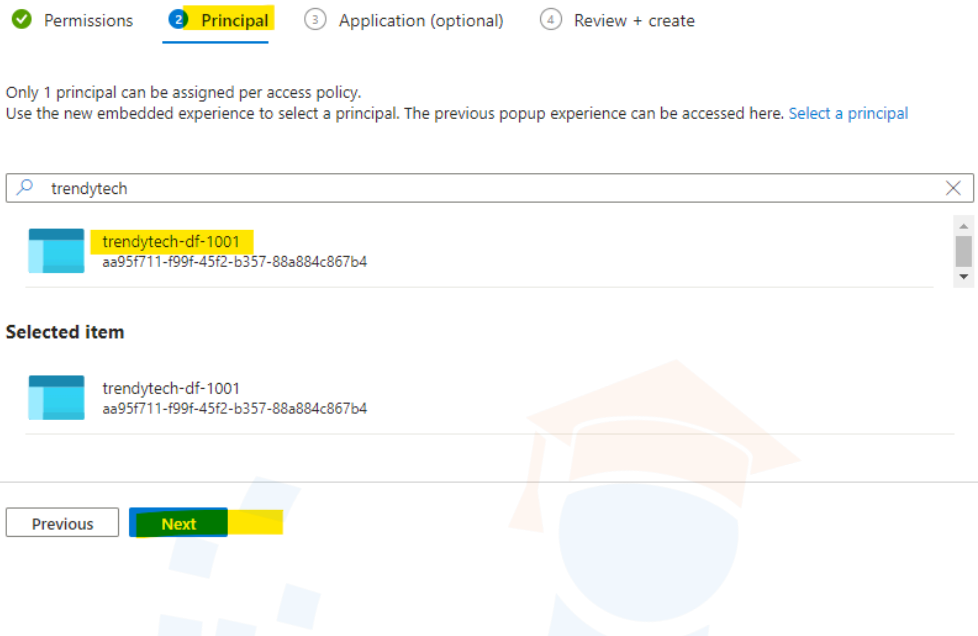


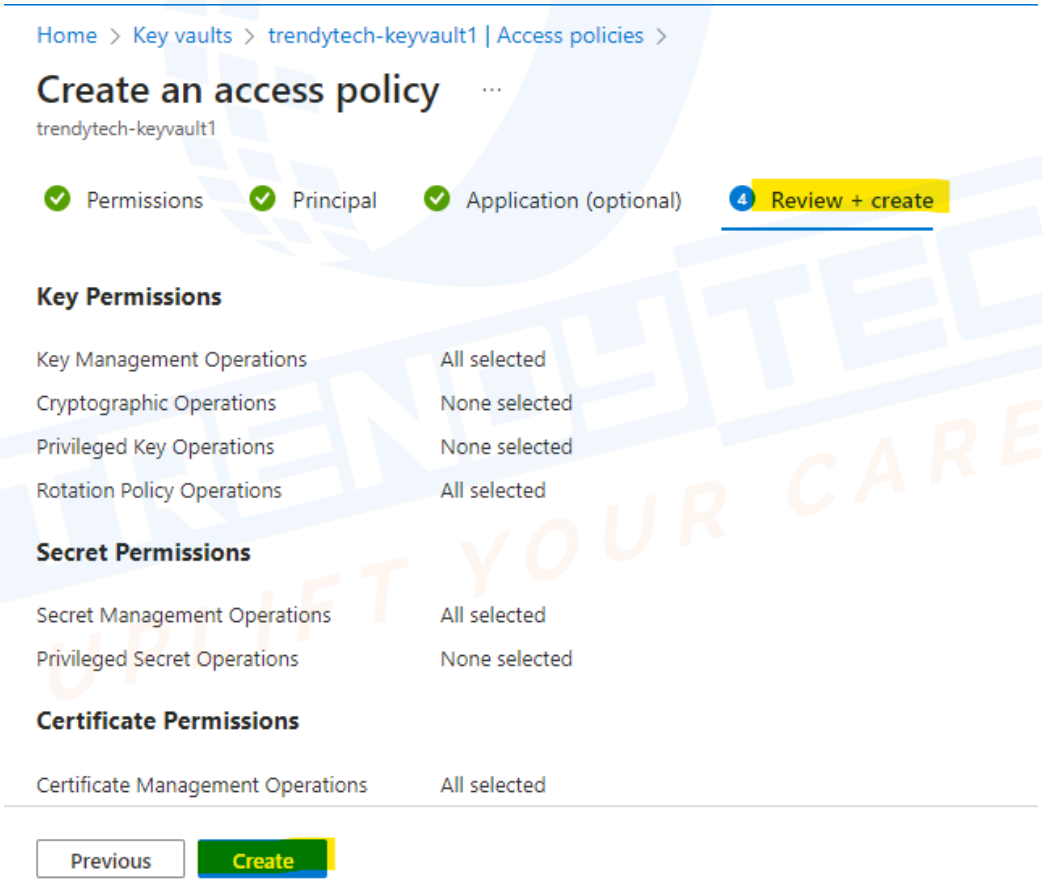


In Key vault grant access permission to the Data Factory service principal under the access policies.





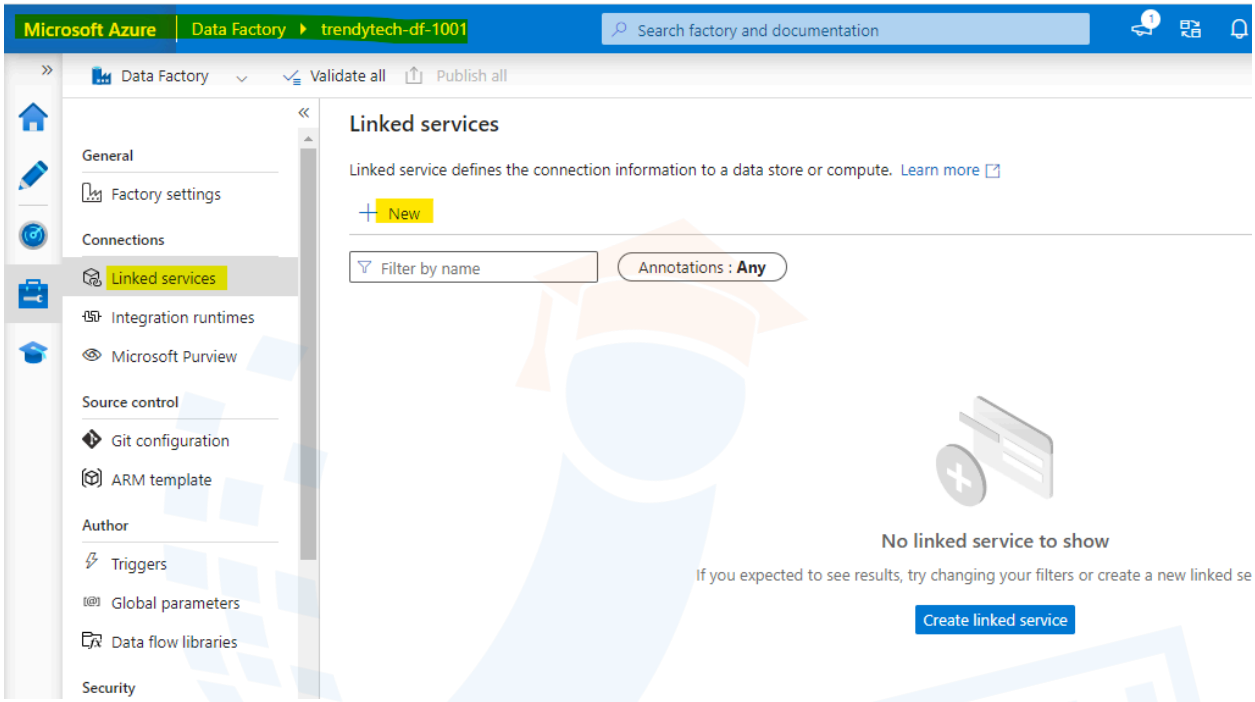

Permissions
2 Principal
3 Application (optional)
4 Review + create
Only 1 principal can be assigned per access policy.
Use the new embedded experience to select a principal. The previous popup experience can be accessed here. Select a principal
trendytech
trendytech-df-1001
aa95f711-f99f-45f2-b357-88a884c867b4
Selected item
trendytech-df-1001
aa95f711-f99f-45f2-b357-88a884c867b4
Previous
Next


Home > Key vaults > trendytech-keyvault1 | Access policies >
Create an access policy
trendytech-keyvault1
Permissions
Principal
Application (optional)
4 Review + create
Key Permissions
Key Management Operations All selected
Cryptographic Operations None selected
Privileged Key Operations None selected
Rotation Policy Operations All selected
Secret Permissions
Secret Management Operations All selected
Privileged Secret Operations None selected
Certificate Permissions
Certificate Management Operations All selected
Previous
Create

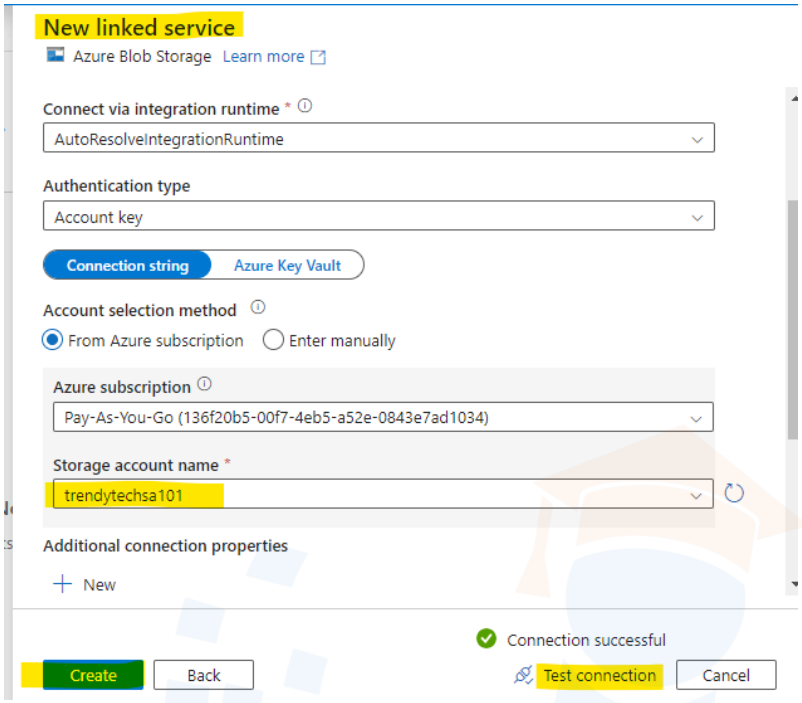
Launch Azure Studio and Create Linked Services:

Linked service - For key vault , AWS s3, Datalake, blob

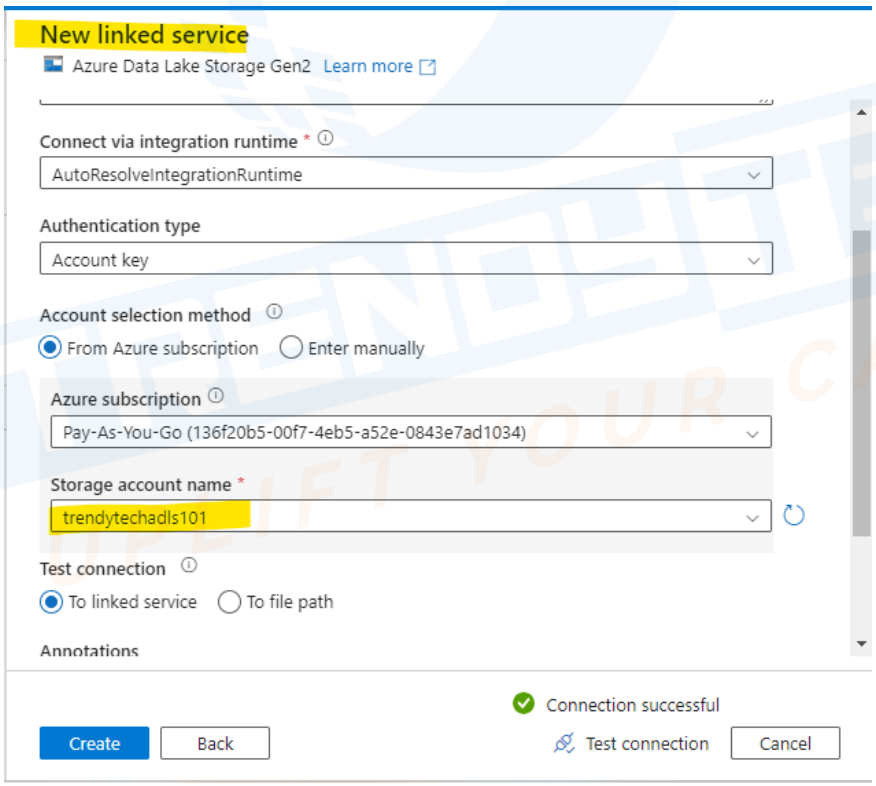
Go to Monitor => Linked Service => New



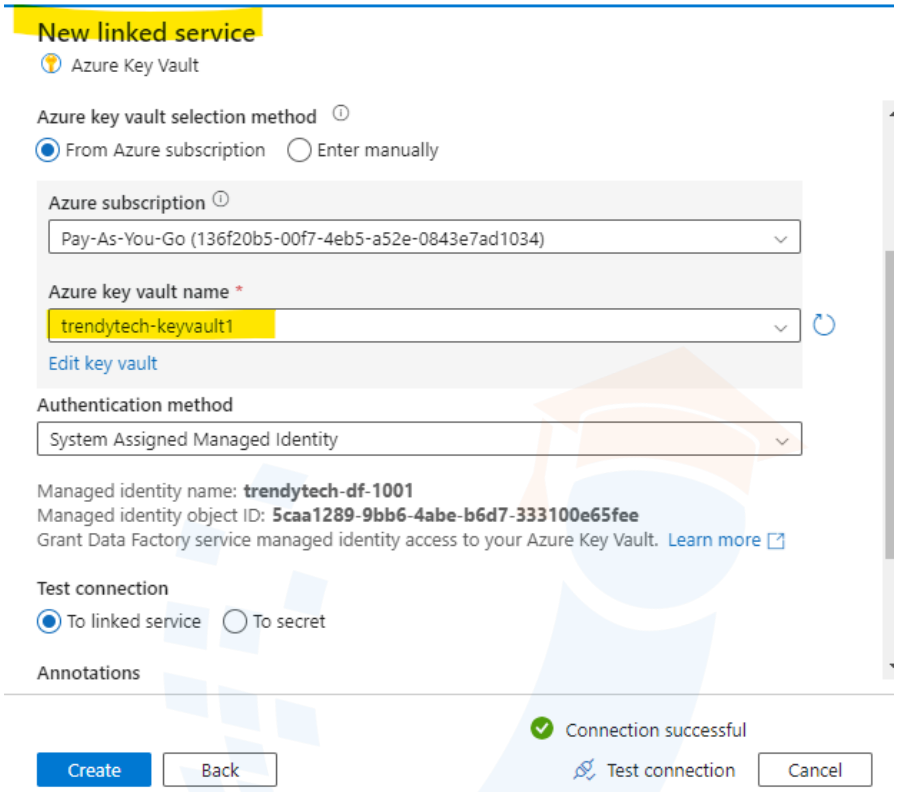
We will create linked service for Blob storage:



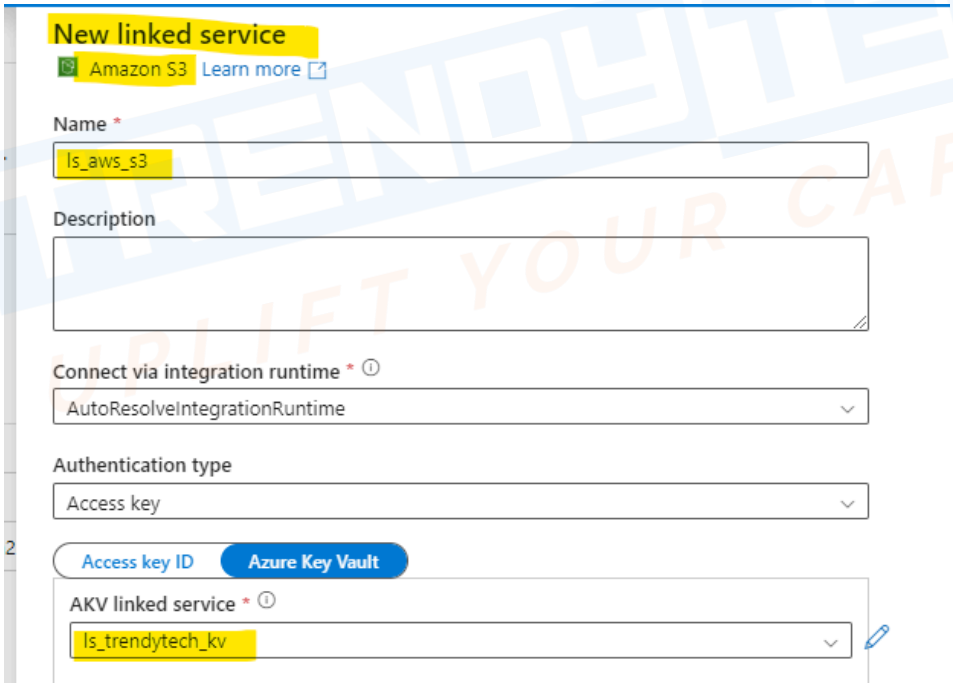
We will create linked service for Azure Data Lake Storage gen2:

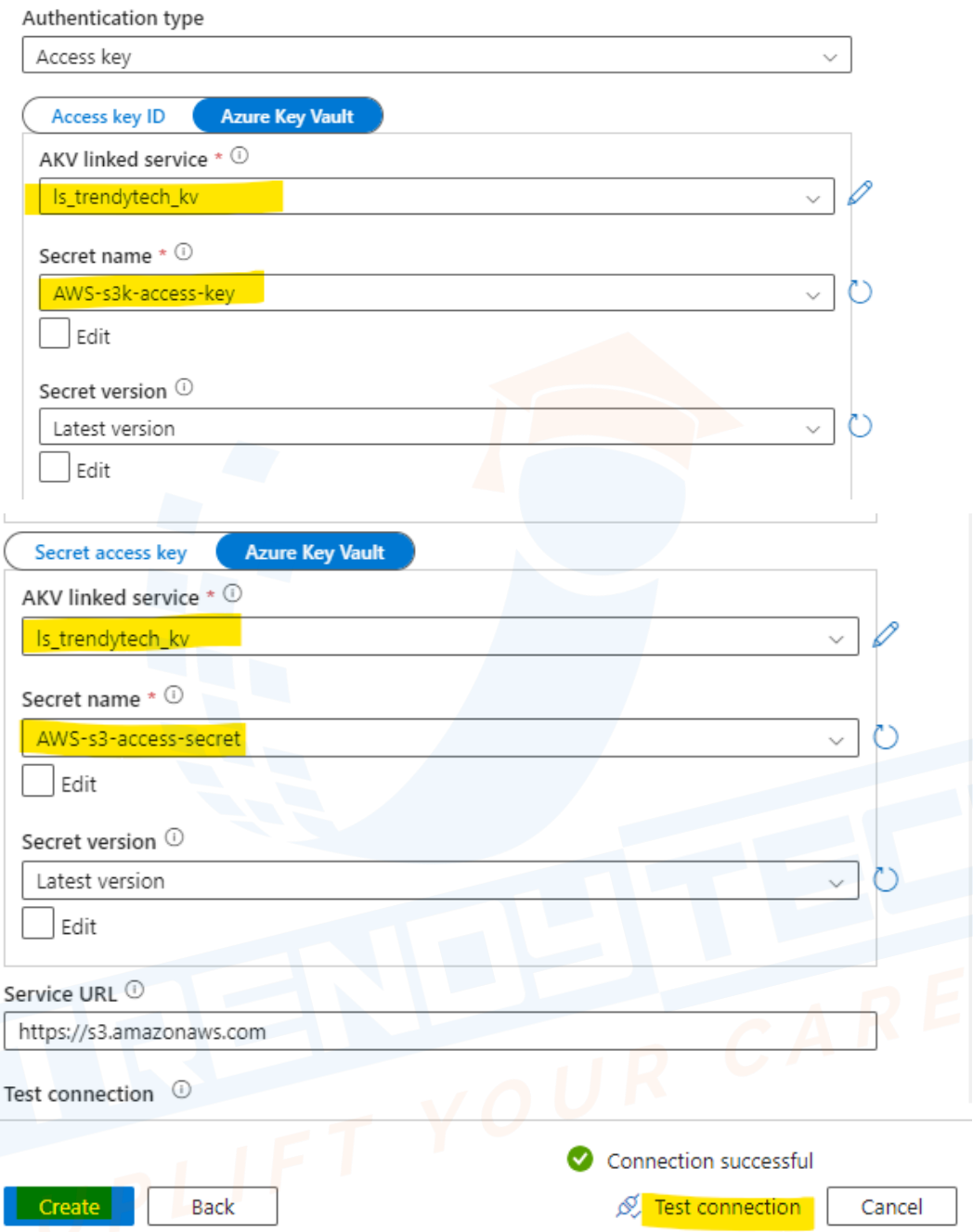


We will create linked service for Key Vault:

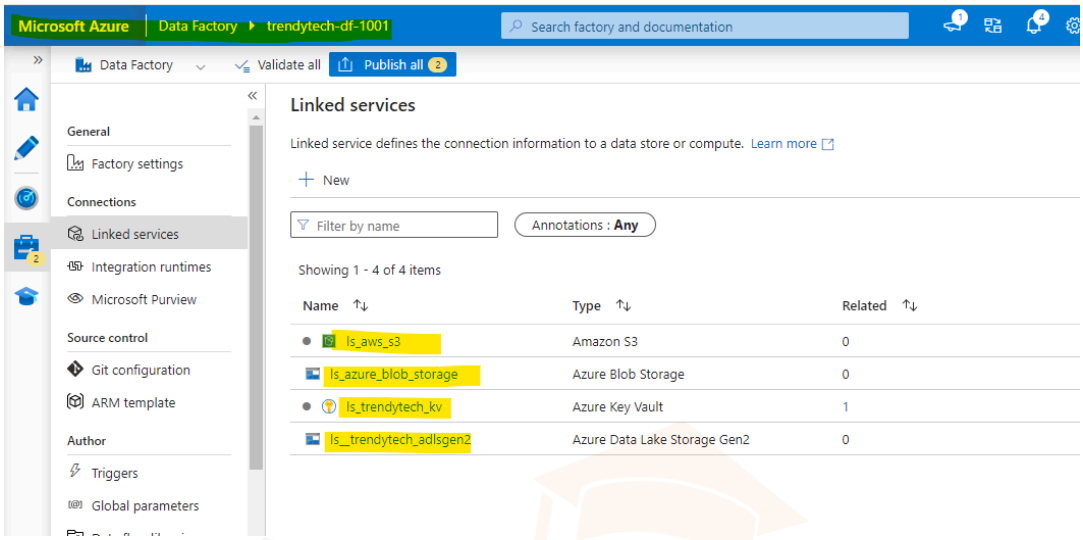


We will create linked service for Amazon S3:



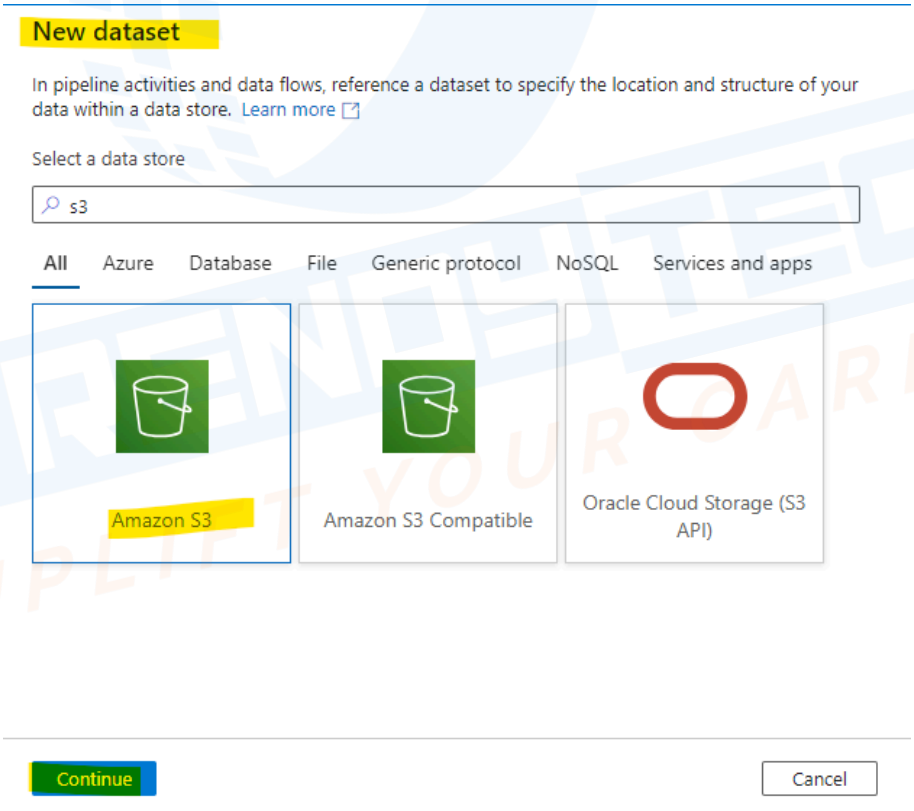


These are all the link services for all the resources that we have have created

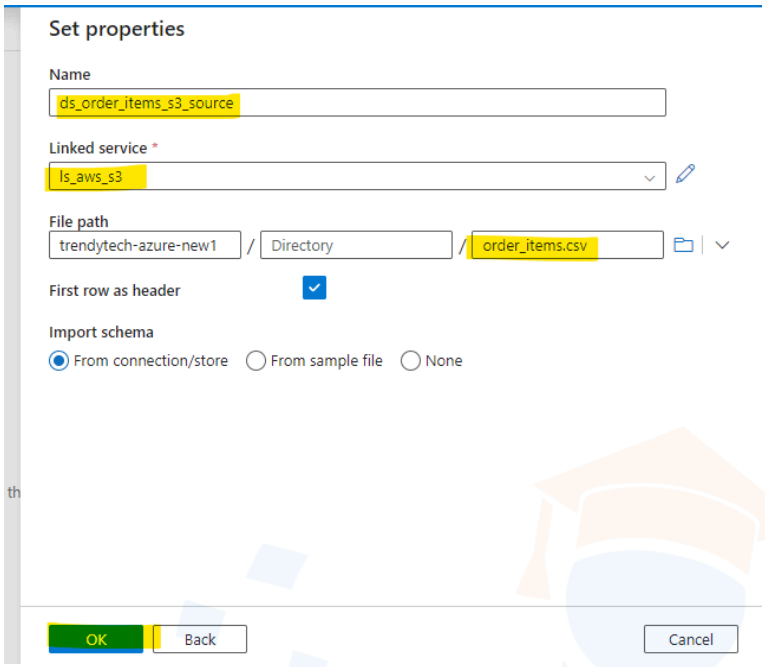


Create Datasets:
Source: Amazon S3, blob

1. Creating dataset for orders_items.csv in s3:

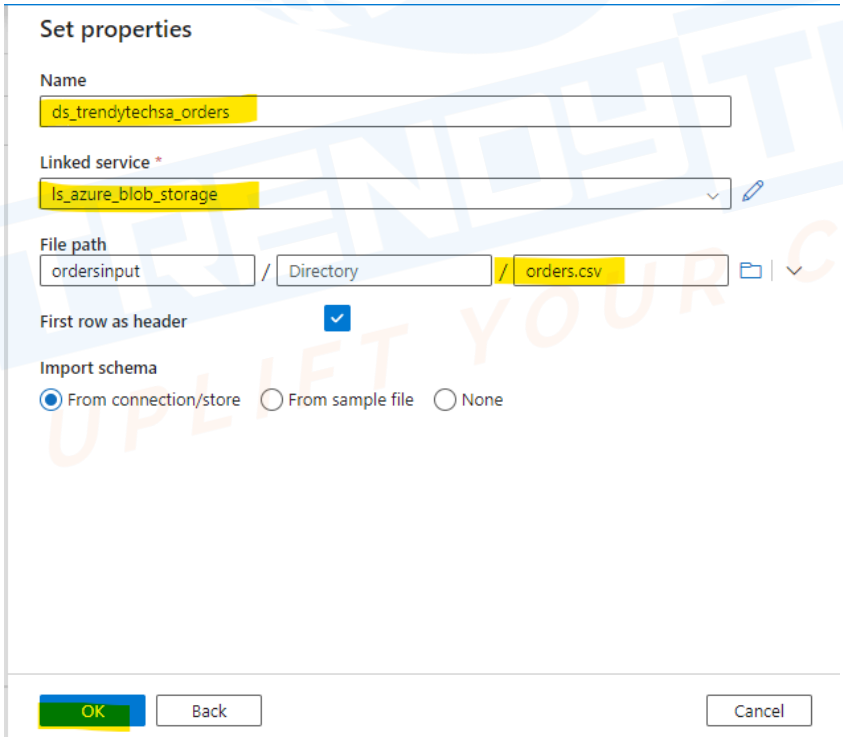


Select option “delimited text”



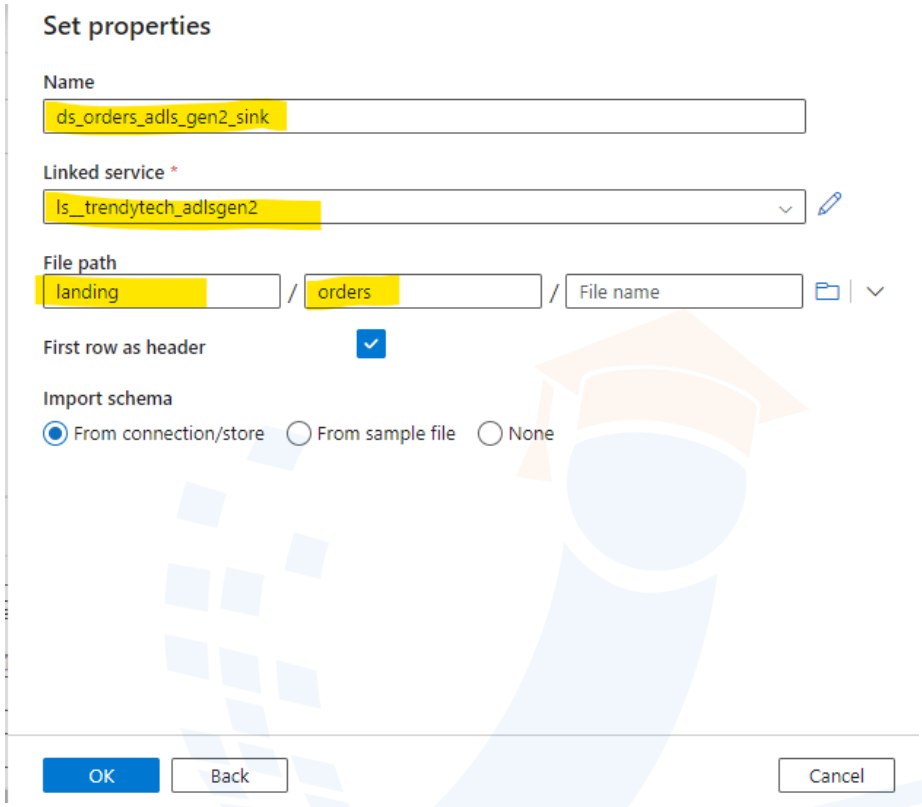
2. For orders.csv file in blob storage

Click on New dataset => select “Azure Blob Storage” => delimited text file =>

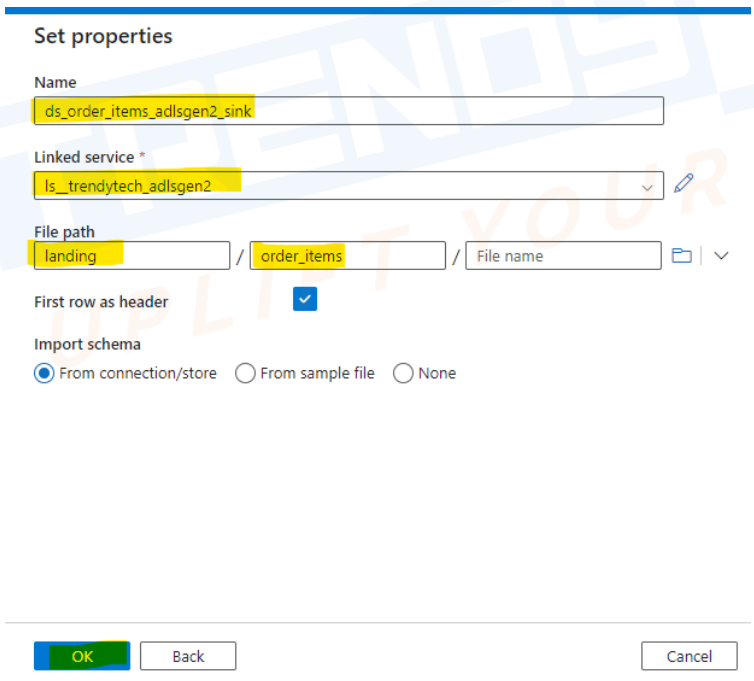


Sink: ADLS Gen2 (landing folder - orders, order_items)

For orders:



For order_items:

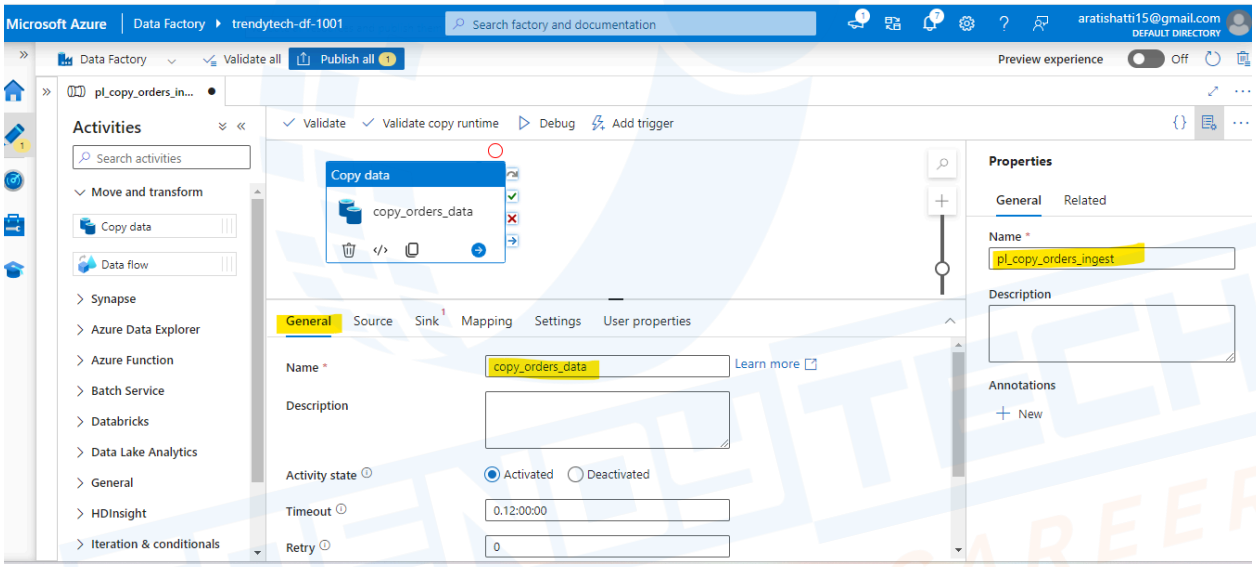


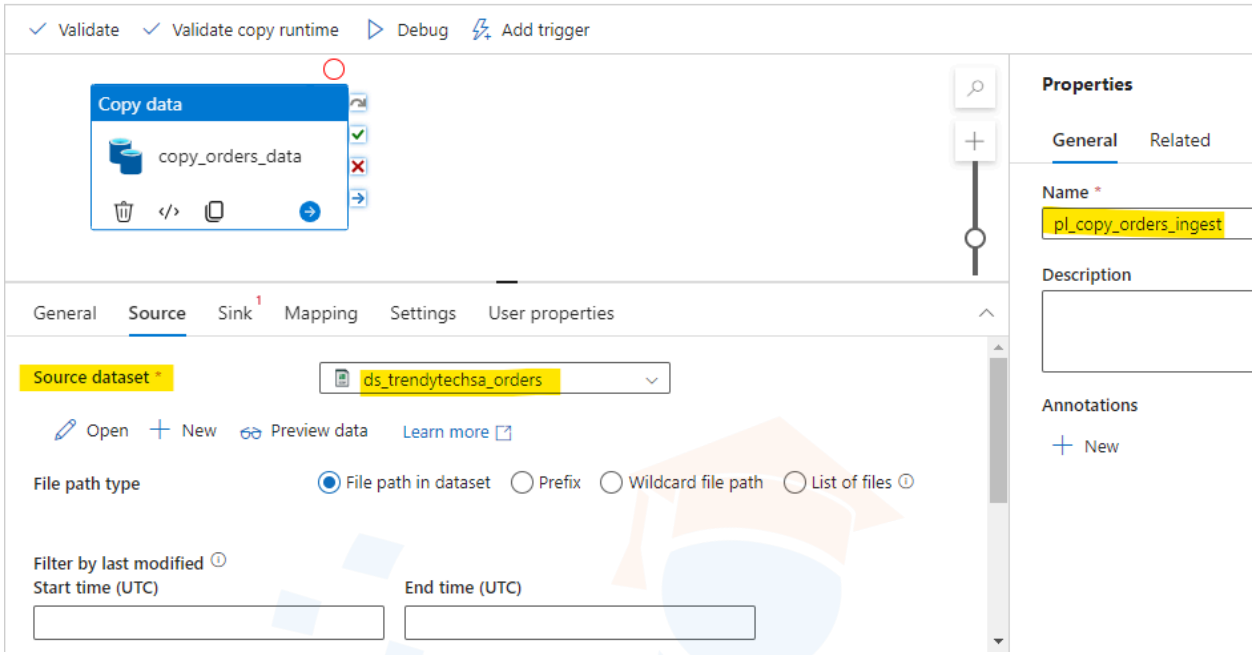
Note: Publish all the resources that you have created.

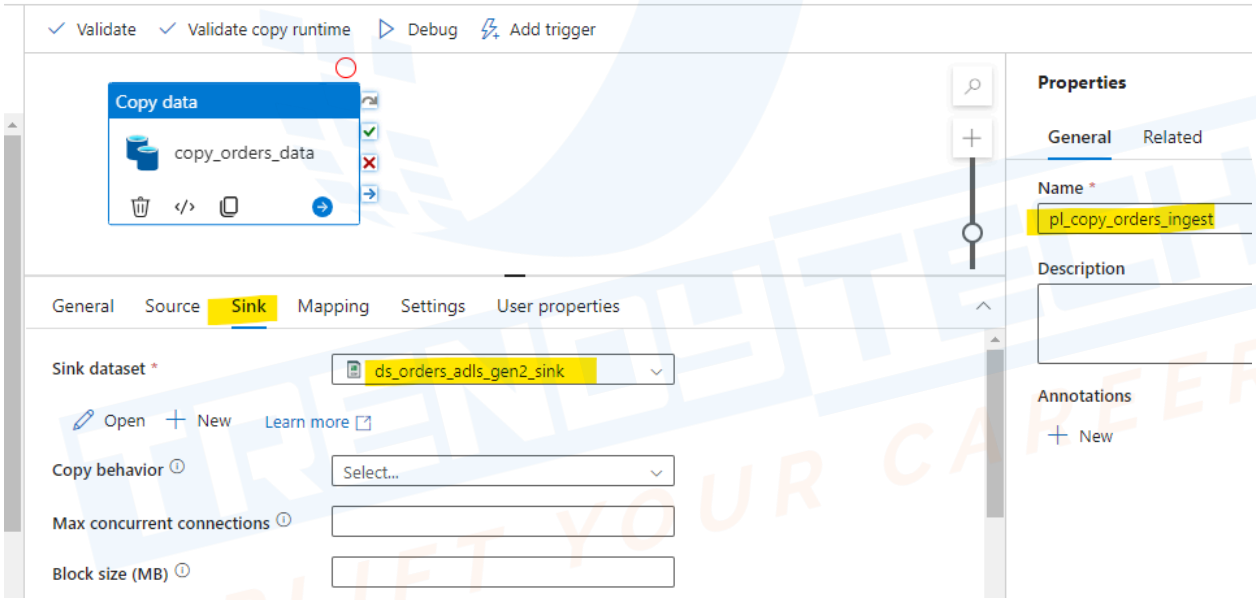
Now we will create pipelines for ingestion, processing and execution.

1. For copying the order.csv from s3 => adlsgen2 (Debug and verify it)

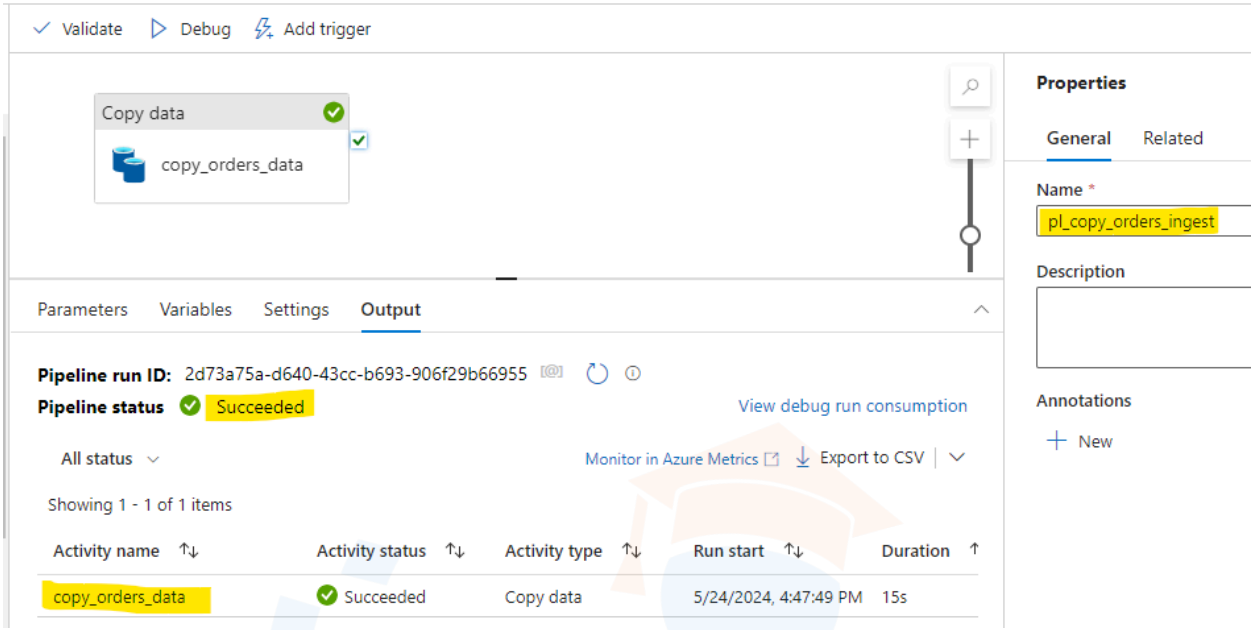
   Add the copy data activity and set source and sink dataset as shown below.





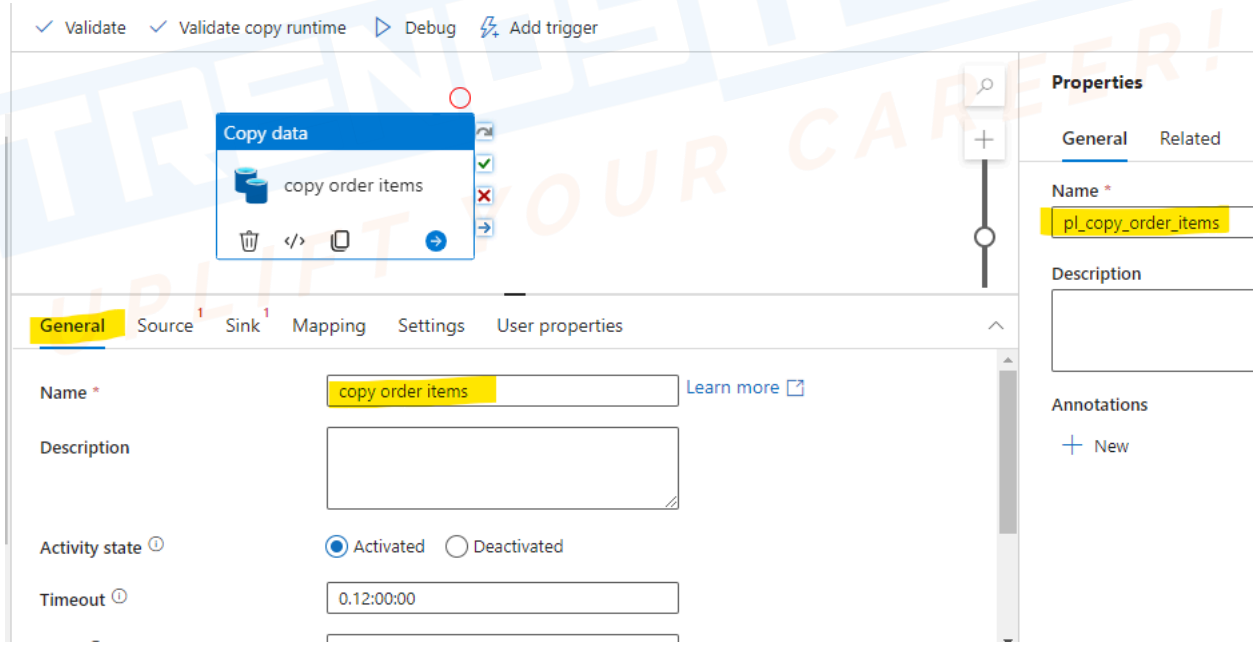


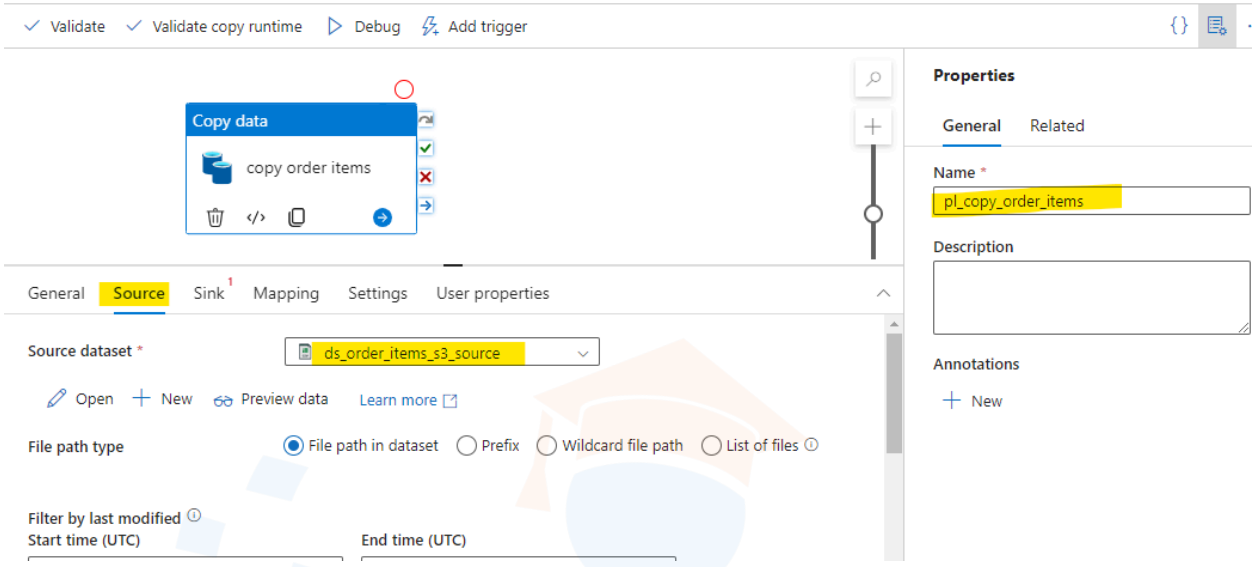
Now click on Debug and check the pipeline



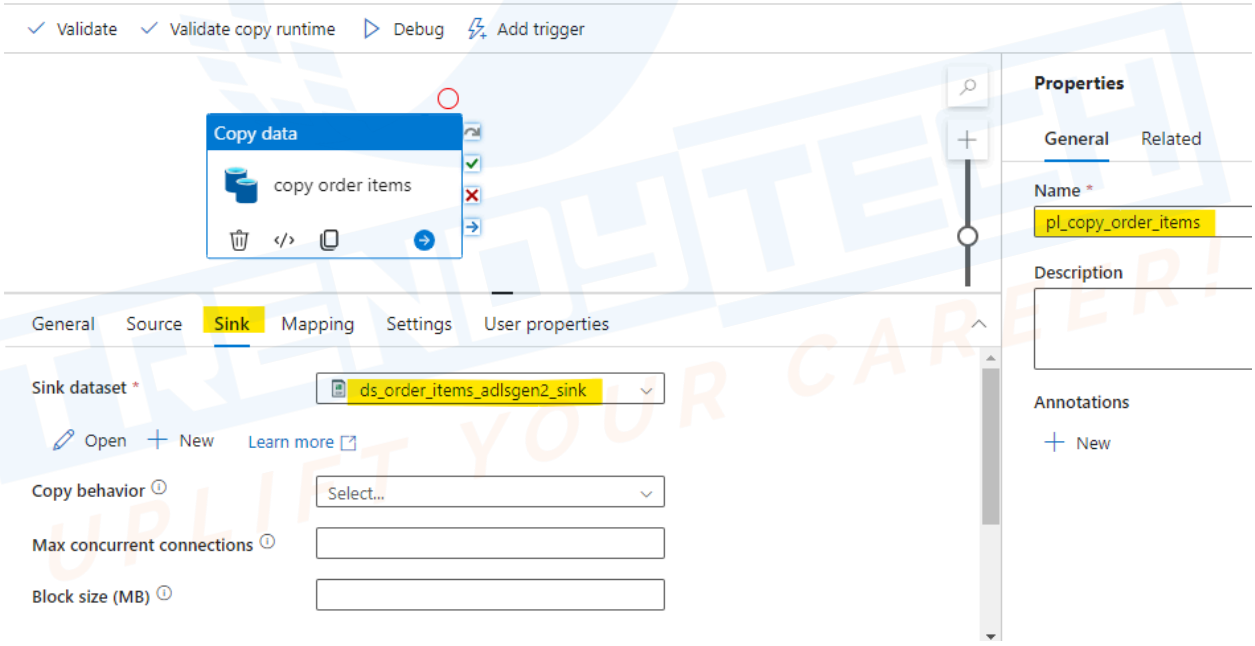
2. For copying the order_items.csv from blob => adlsgen2 (Debug and verify it)

Add the copy data activity and set source and sink dataset as shown below.

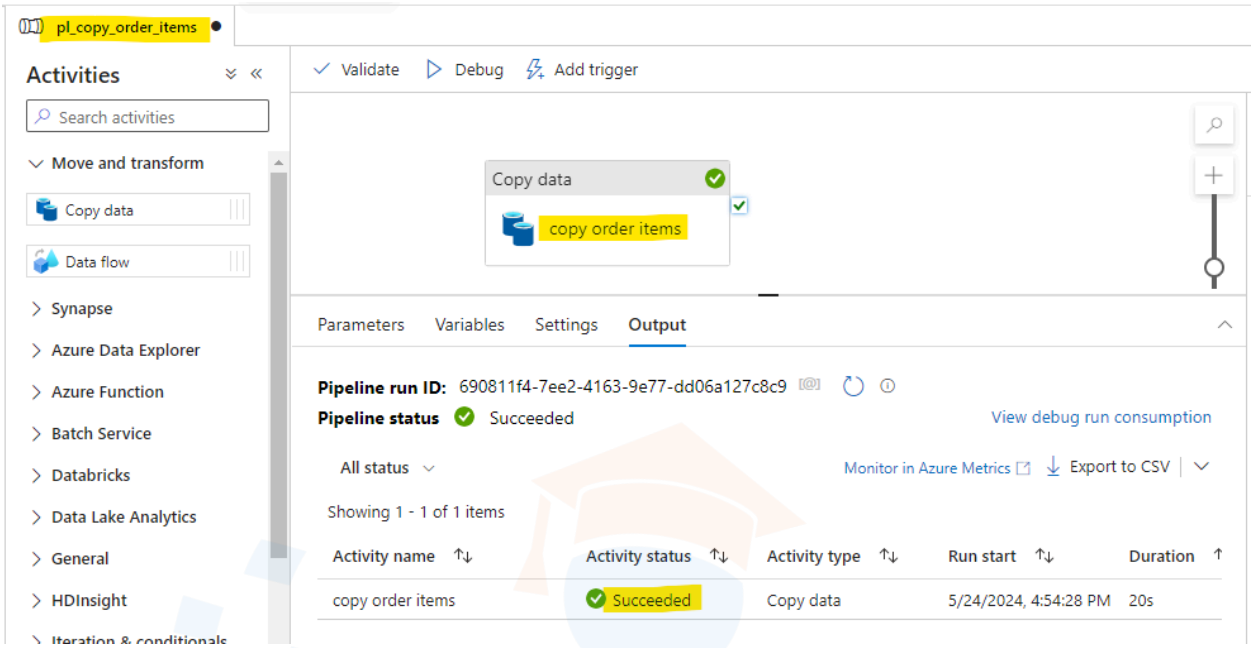




Now click on Debug and check the pipeline



Now click on Debug and check the pipeline



3. Create new pipeline for processing the data

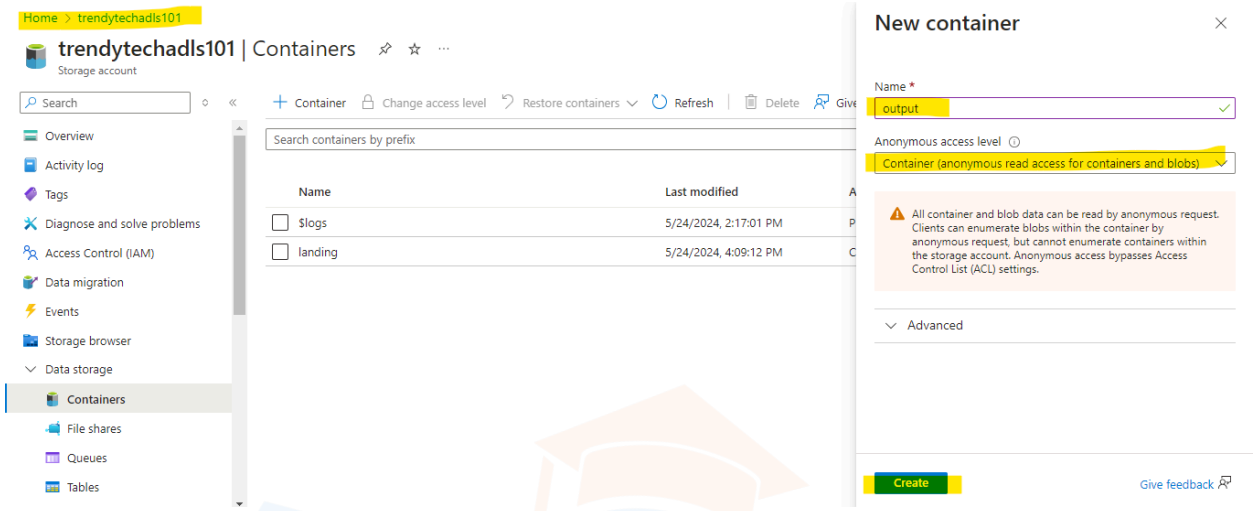
But before creating a pipeline we will create a Data Flow.

For processing data using Data Flow

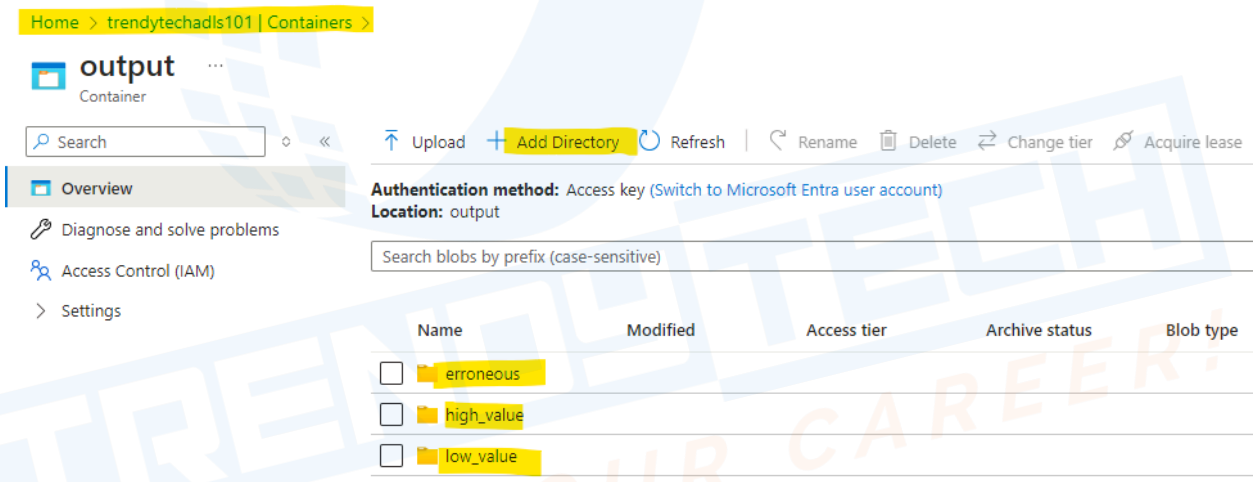
In Adls gen 2 create container **“output”** and segregate data into 3 categories:
High value orders (>500)
Low value orders (<=500)
Erroneous orders (no order amount).

Create 3 folders for the above categories in the output folder of the storage container as **“high_value_orders”, “low_value_orders”, “erroneous”**.

.



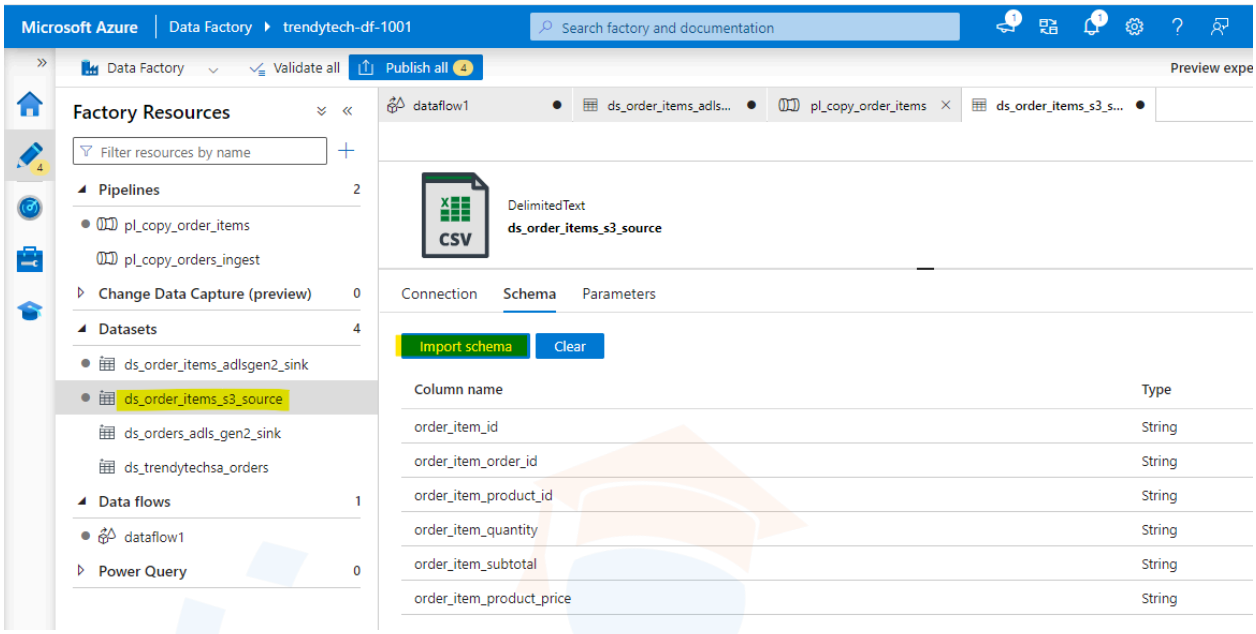
Click on the Add directory and add the folder high_value, low value and erroneous as shown.



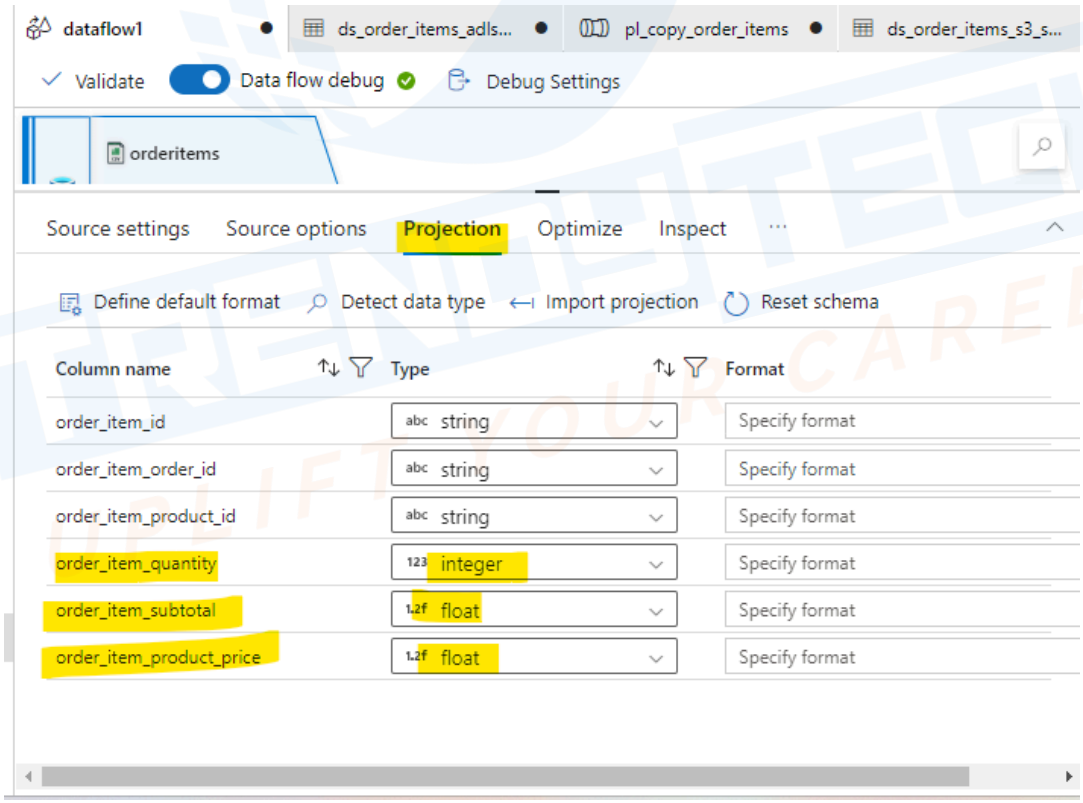
Now create a New Data Flow:

-Add Source for order_item dataset (in adls gen2) and import the schema (*.csv).

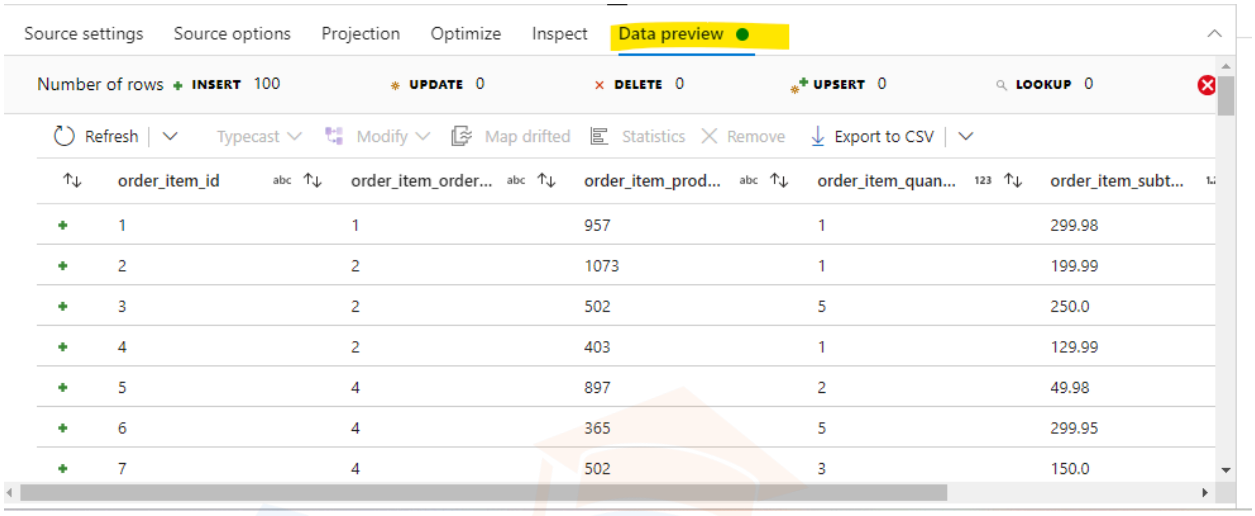
Note: for order_item dataset also import the schema as shown below.



Preview the data to check if it is in the desired form. Change data types if required (e.g., order_item_quantity from string to integer, order_item_subtotal and order_item_product_price to float). Refer the below screenshot



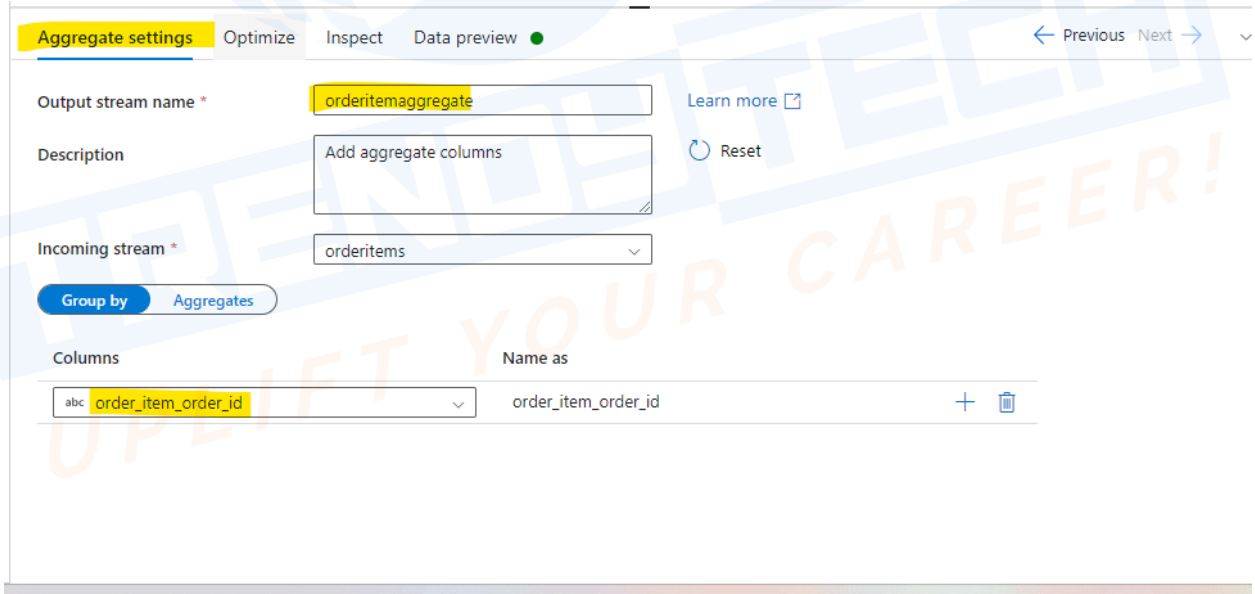
And preview the data

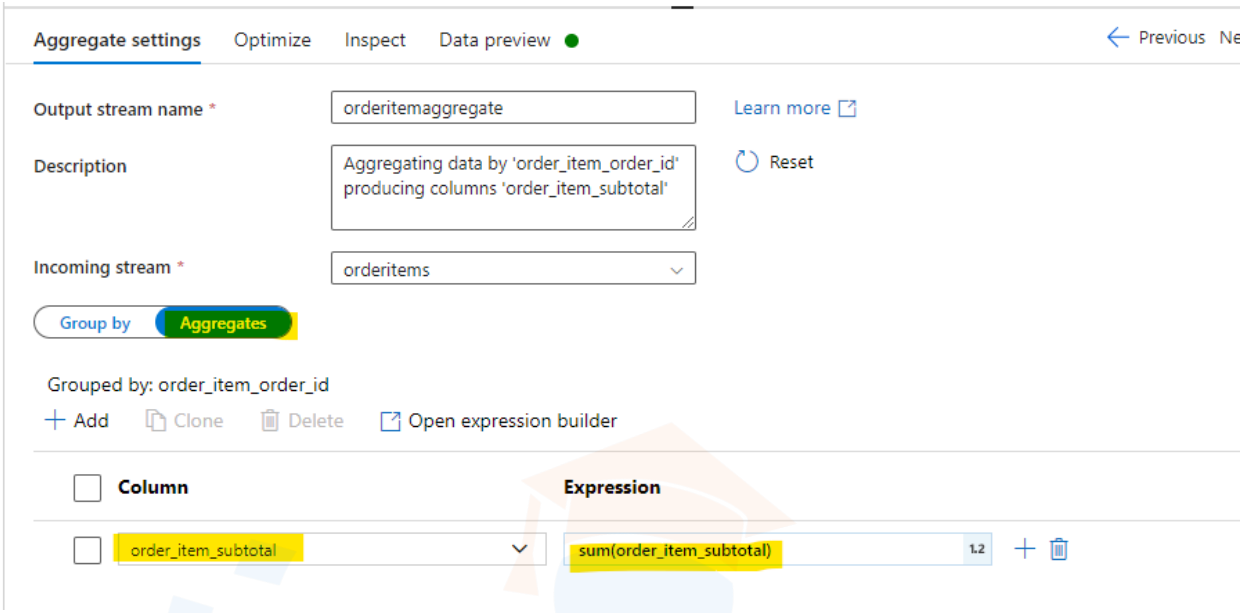


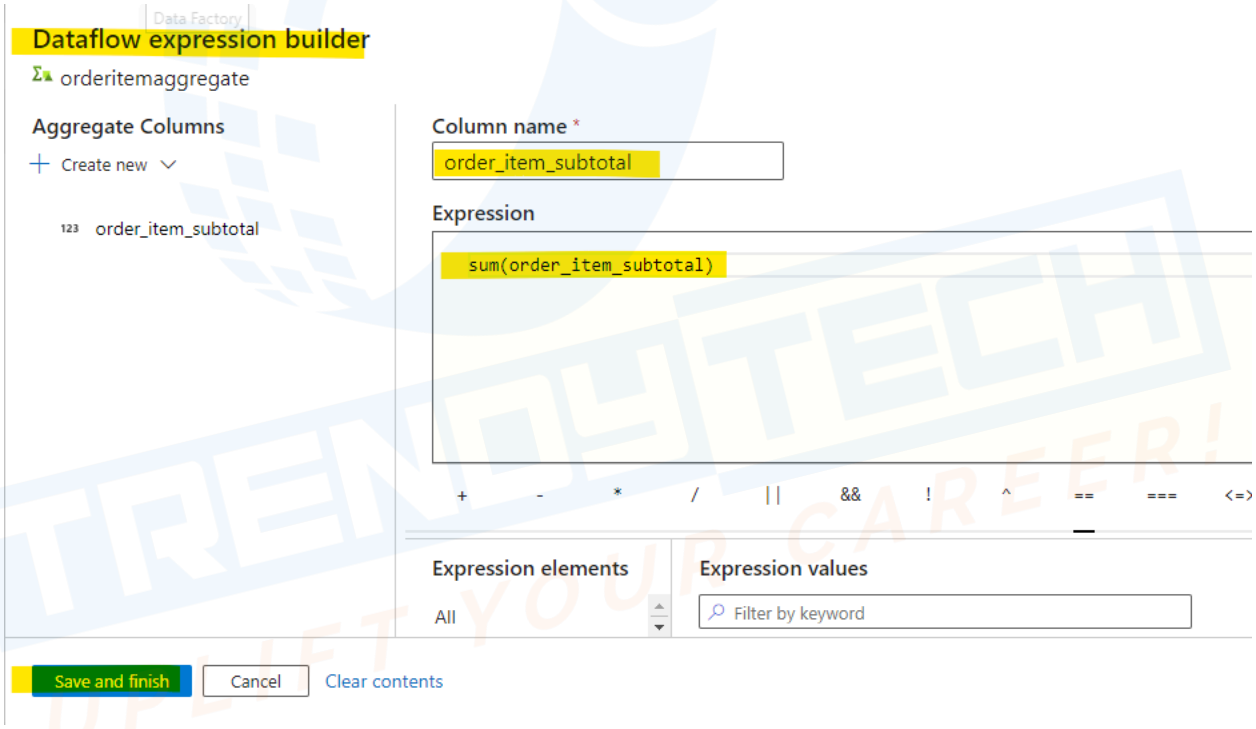
Turn on the Debug mode to check if the processing is as per the requirement.

- Use aggregate transformation activity to calculate the subtotal of respective order items.

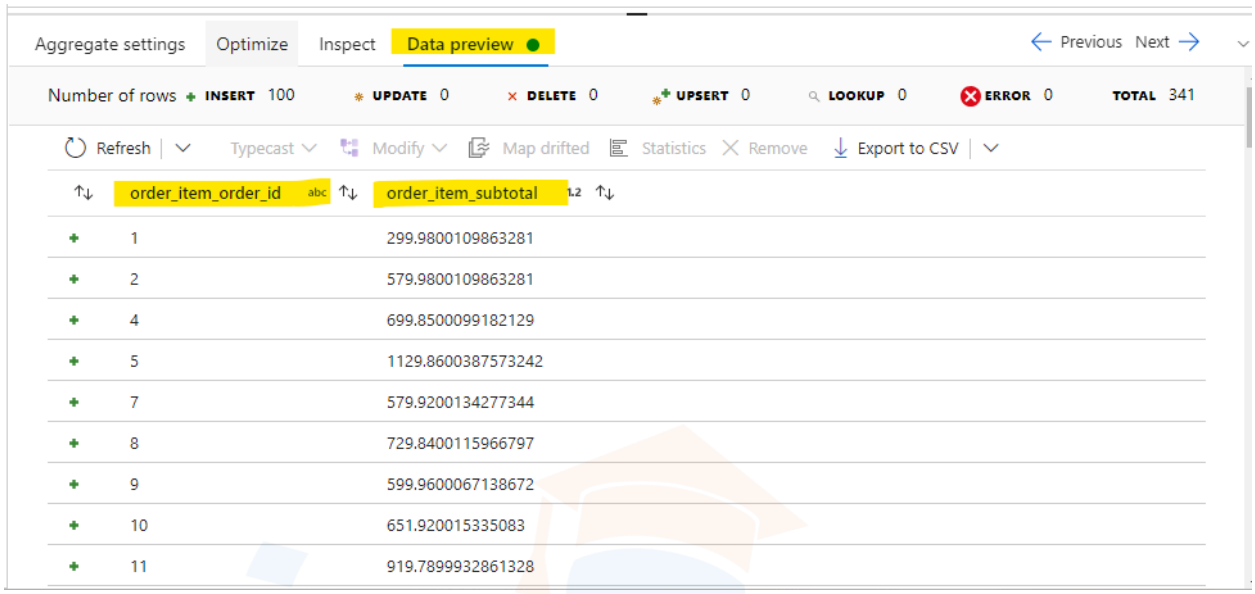
Group by column “order_item_order_id”





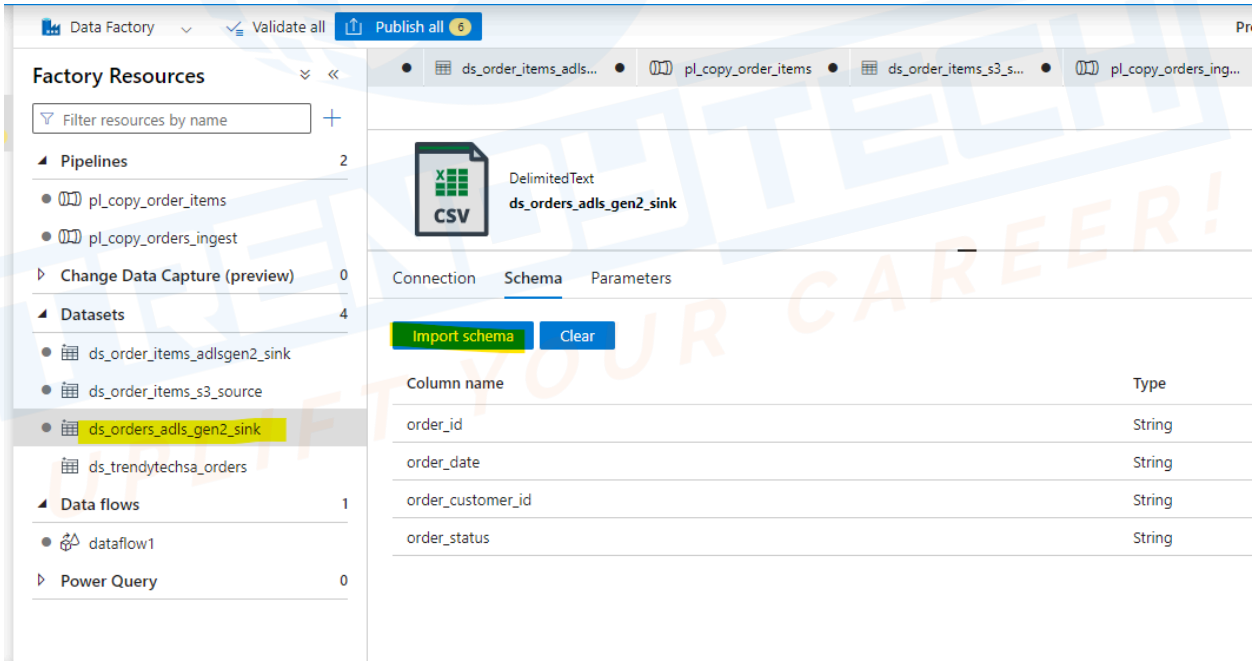


Click on save and finish and Click on Data preview as shown below.

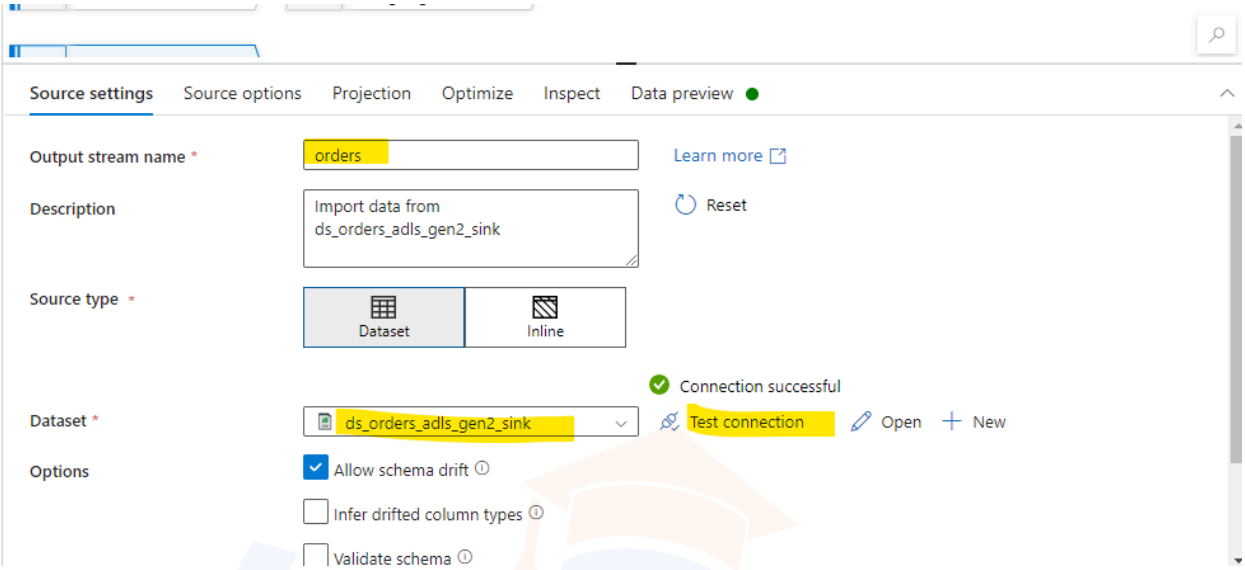


- Add Source for orders dataset and import the schema (*.csv). Preview the data to check if it is in the desired form.

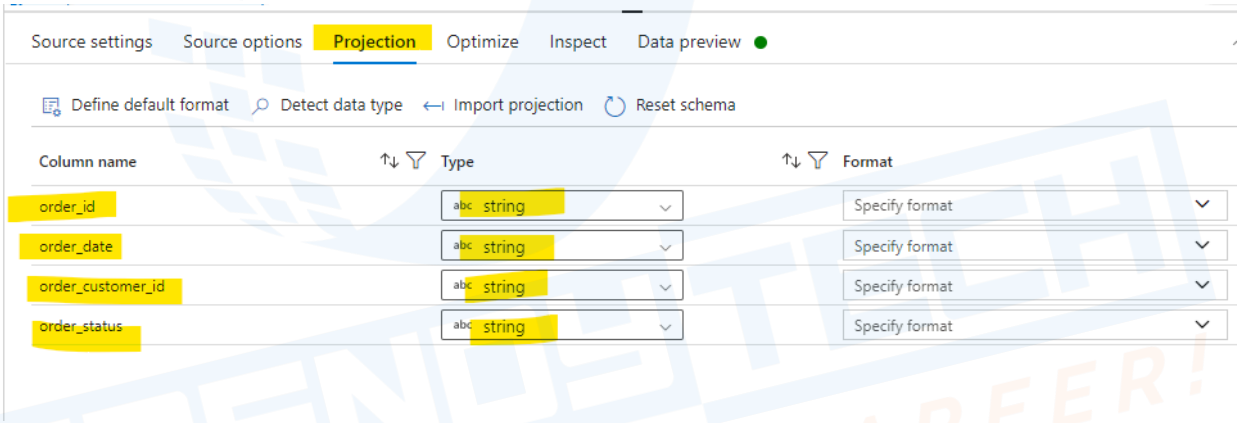
Again import the schema of orders dataset as shown below.



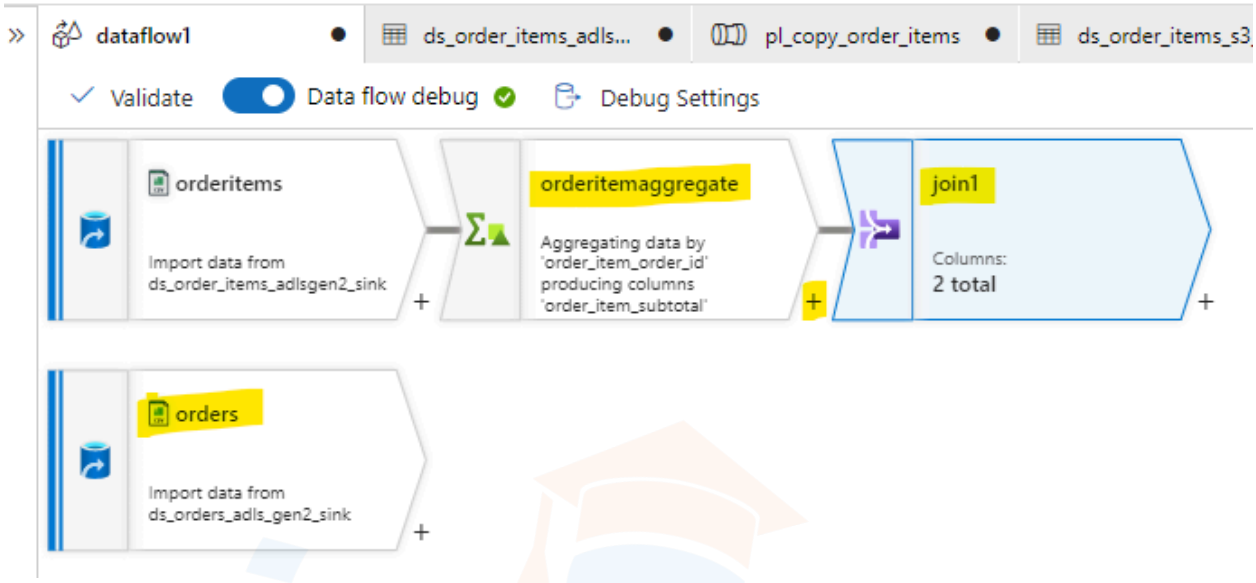
Now add the source in Data flow.

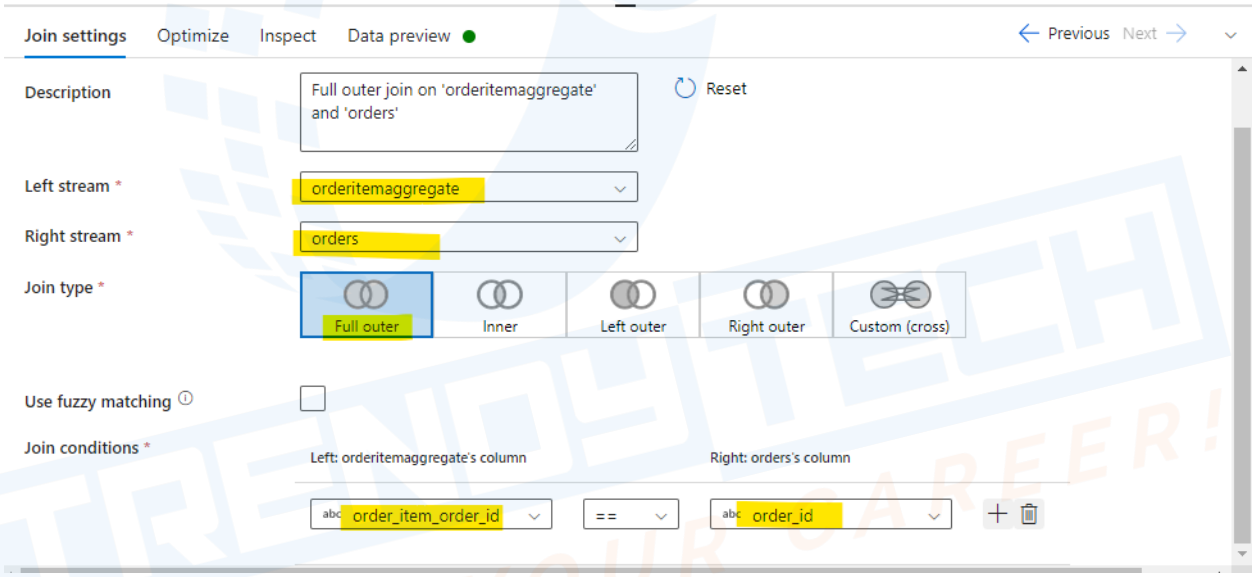


Import the schema and preview the data

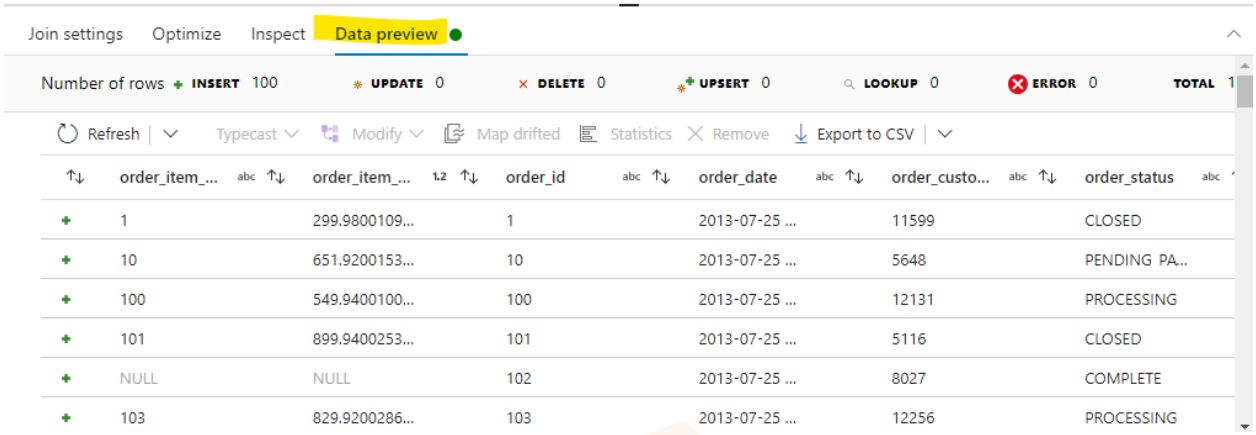


-Use the transformation Join to perform a full outer join for orders and order_item datasets on the condition order_item_id == order_id.

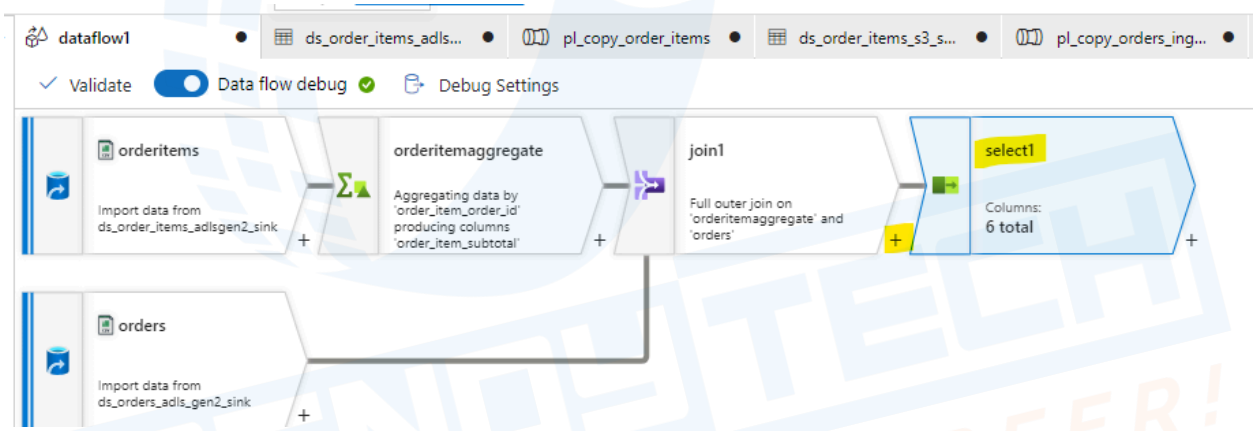




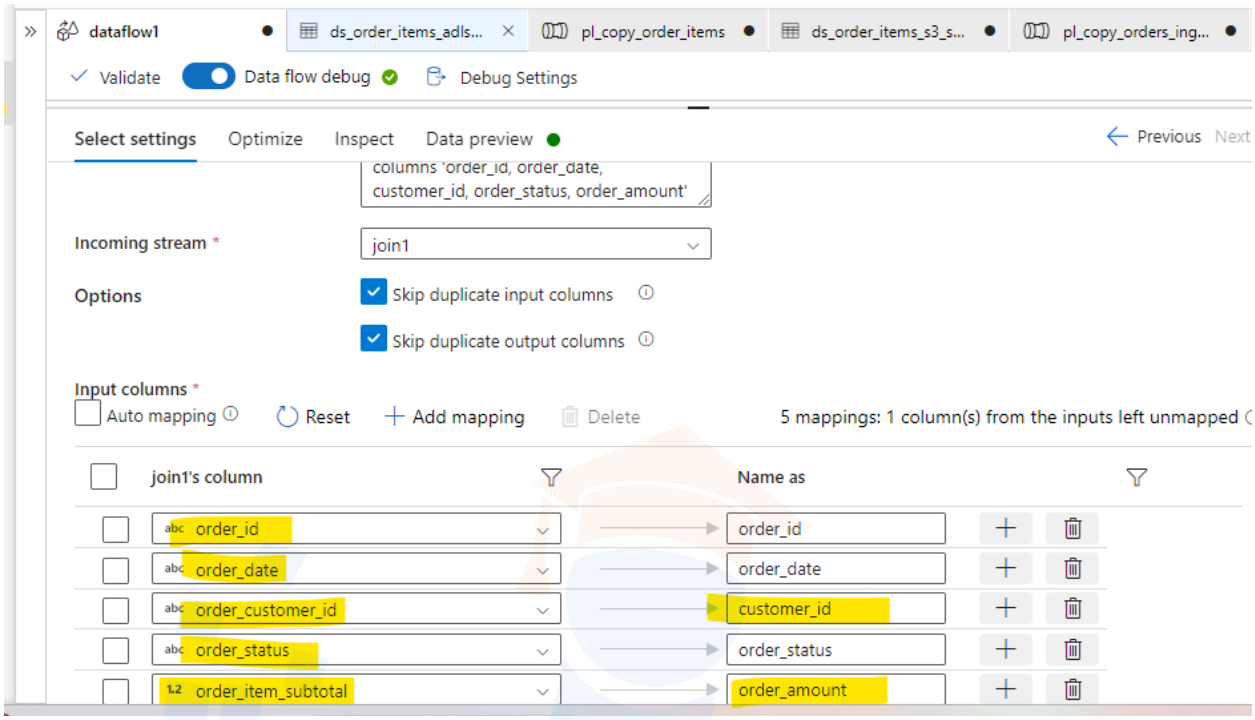
Click on the option “data preview” and preview the data



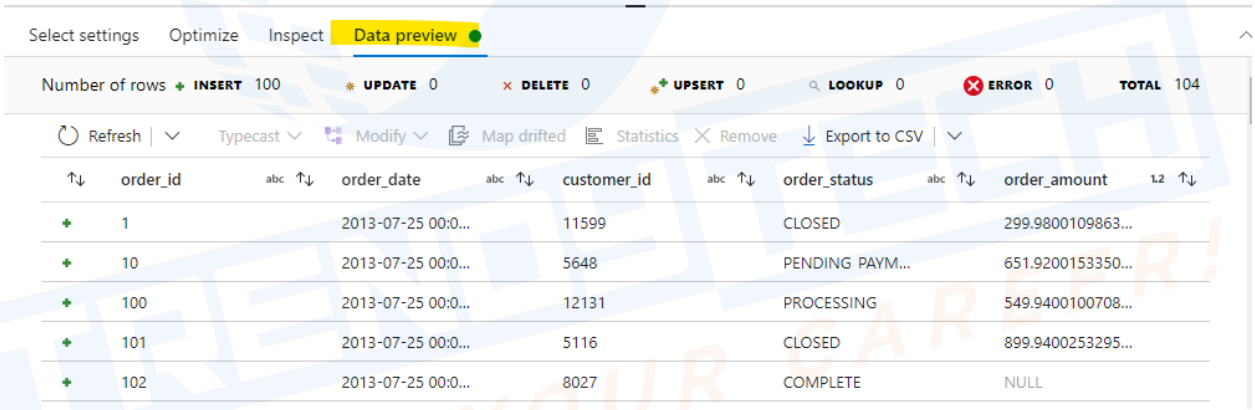
-Use the transformation Select to remove duplicate columns like order_item_order_id and rename the columns if required.



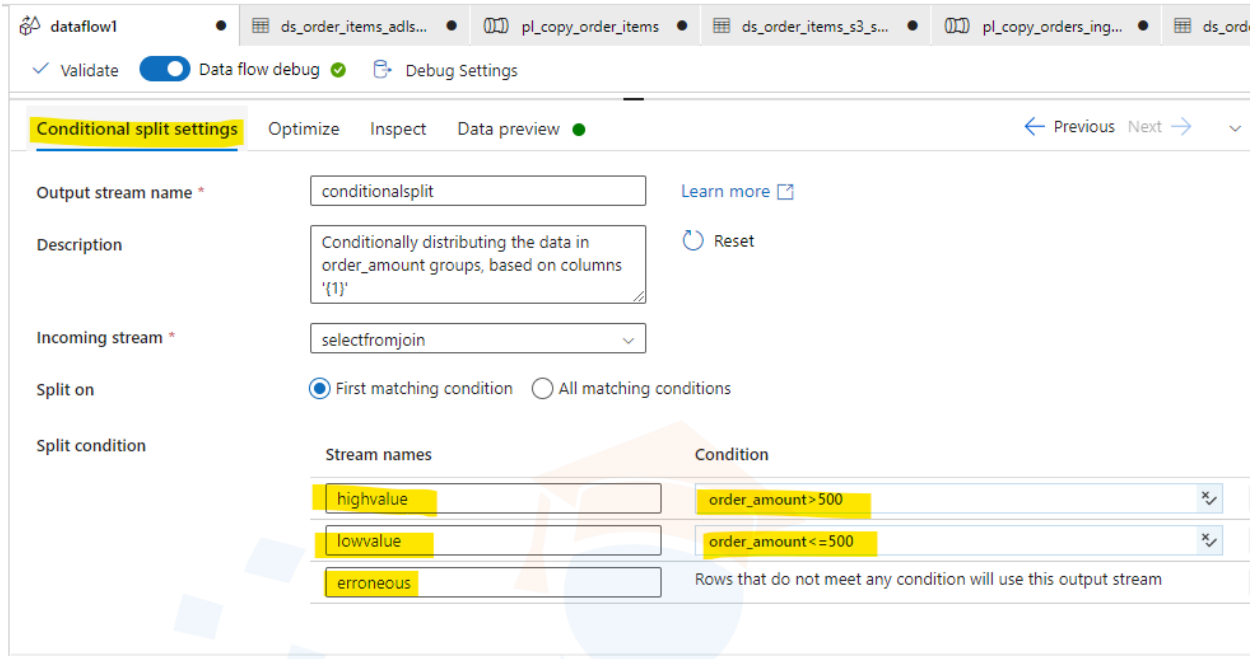
We will change the column name order_customer_id to customer_id and will resume the column order_item_order_id as shown below.

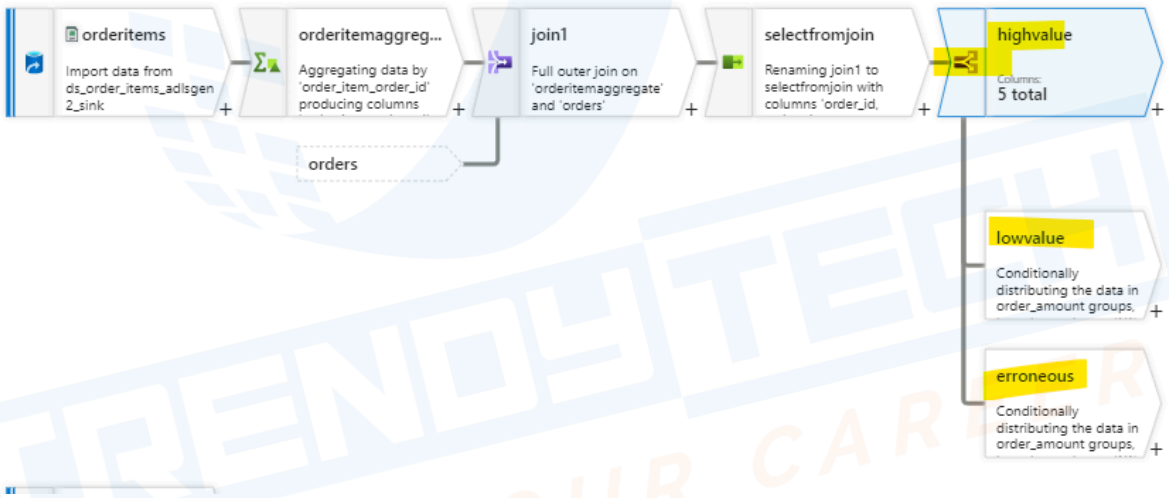


Click on the option “Data Preview”



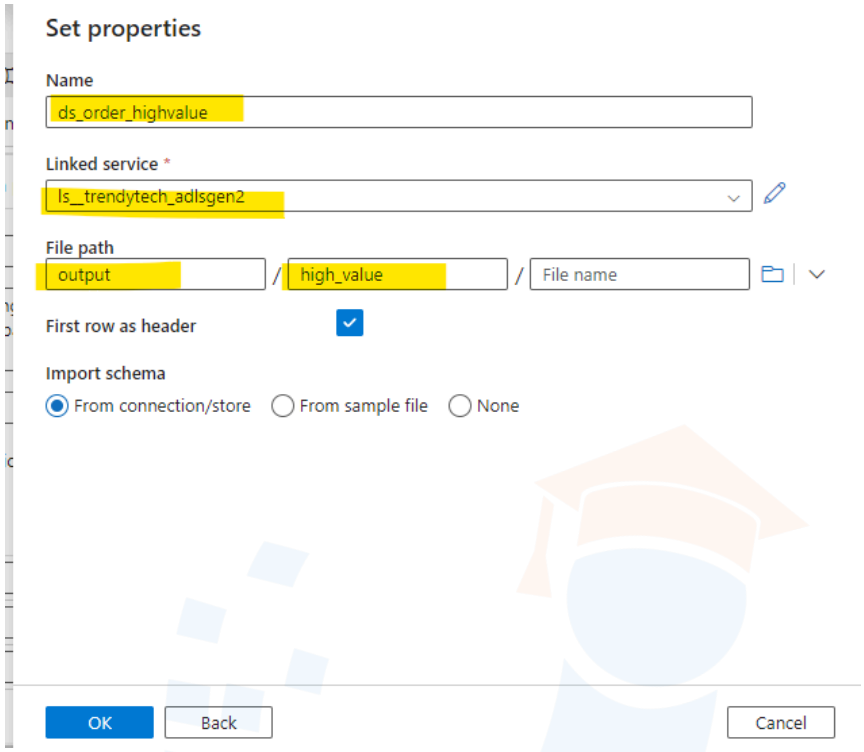
-Use the transformation Conditional Split to segregate the data into high value, low value, and erroneous.

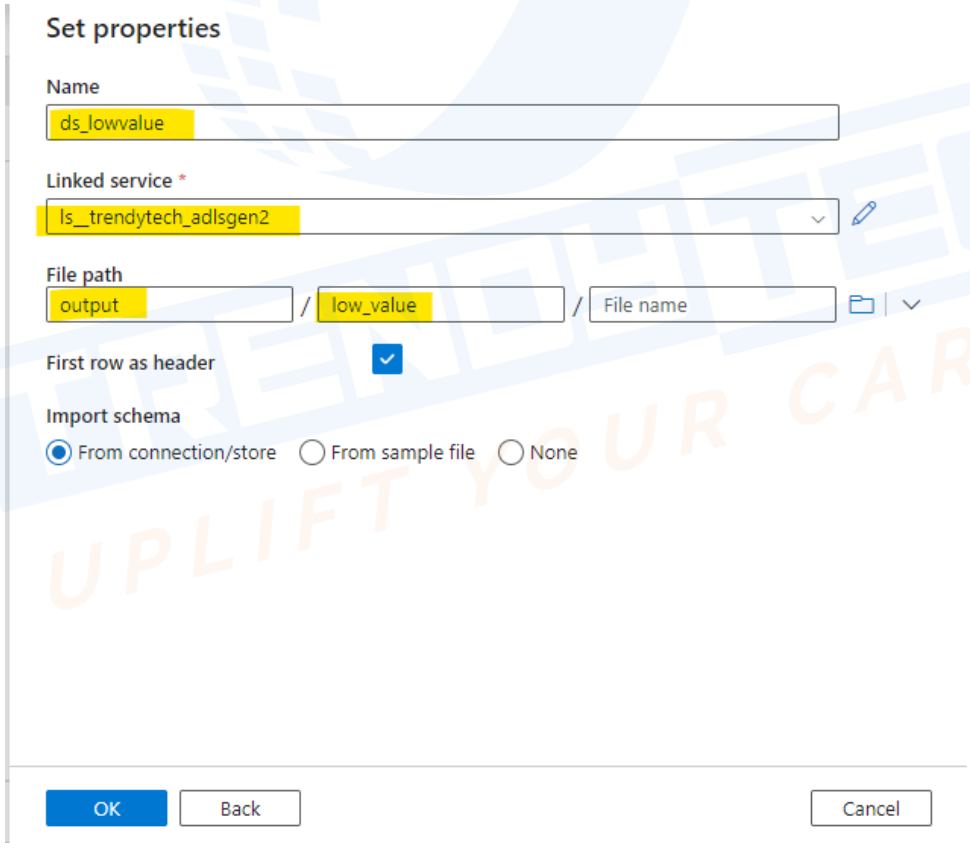


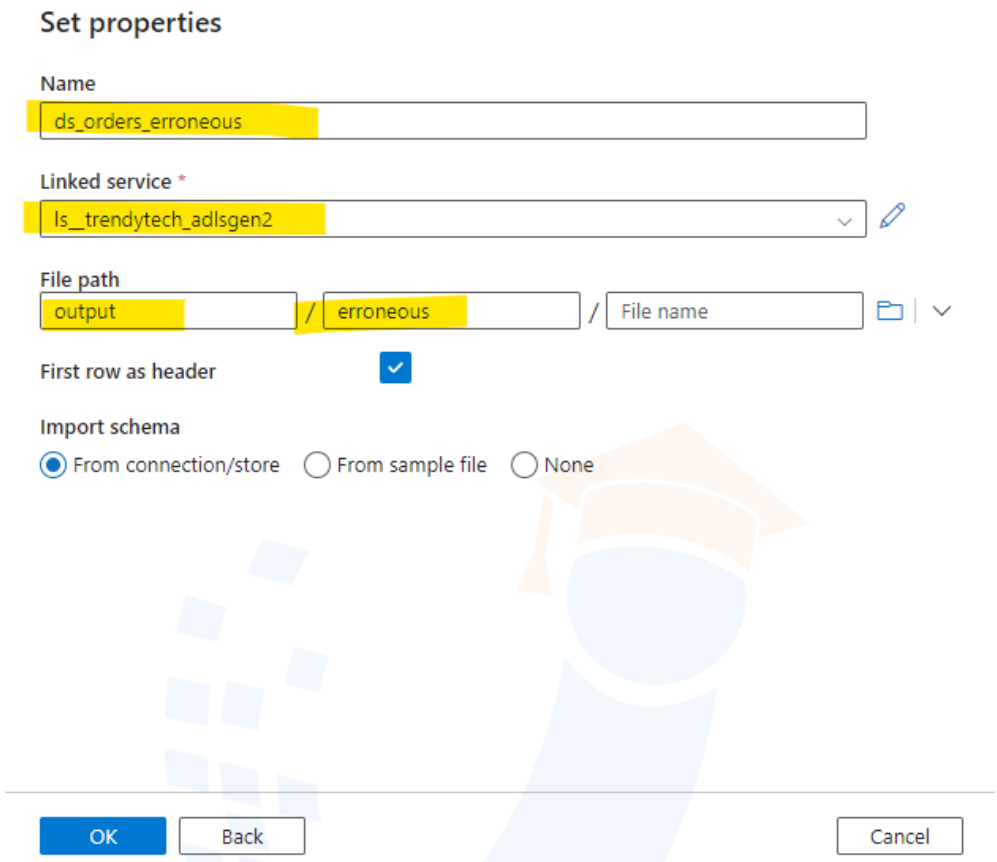


-Write the outputs (high value, low value, erroneous) to their respective sinks in the output folder of ADLS GEN2.

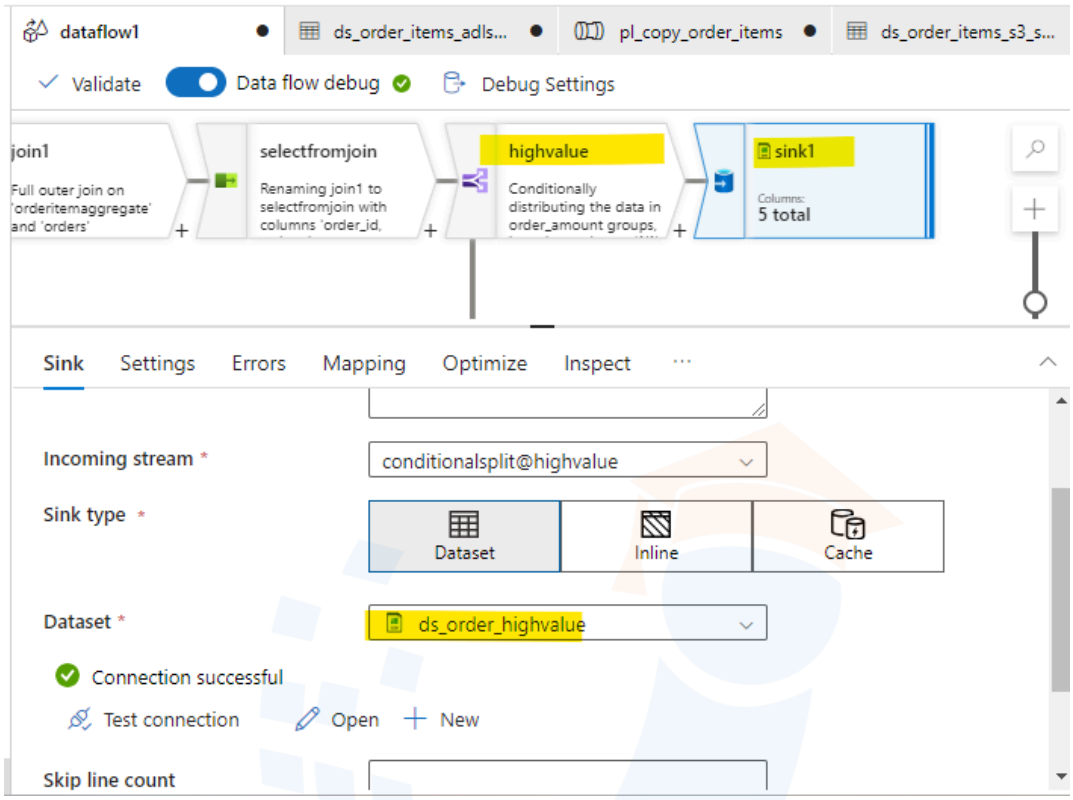
Create dataset for storing “**high_value_orders”, “low_value_orders”, “erroneous”** data in adls gen2, refer the below screenshot.

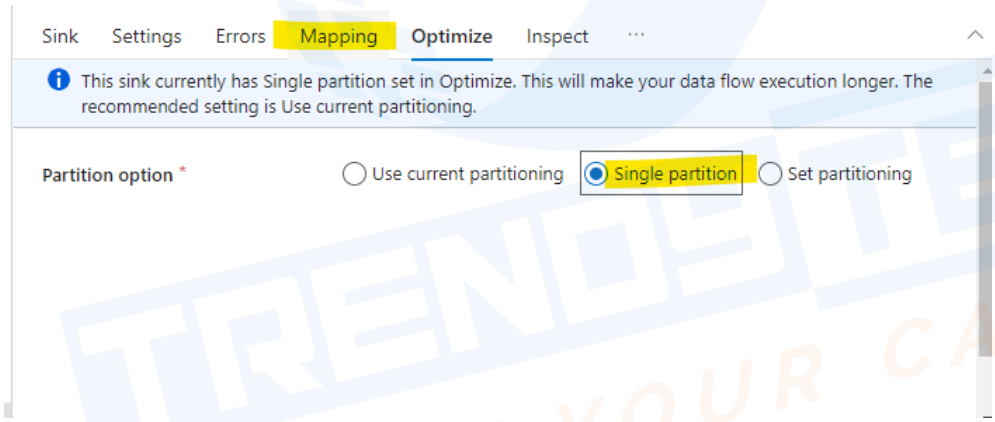




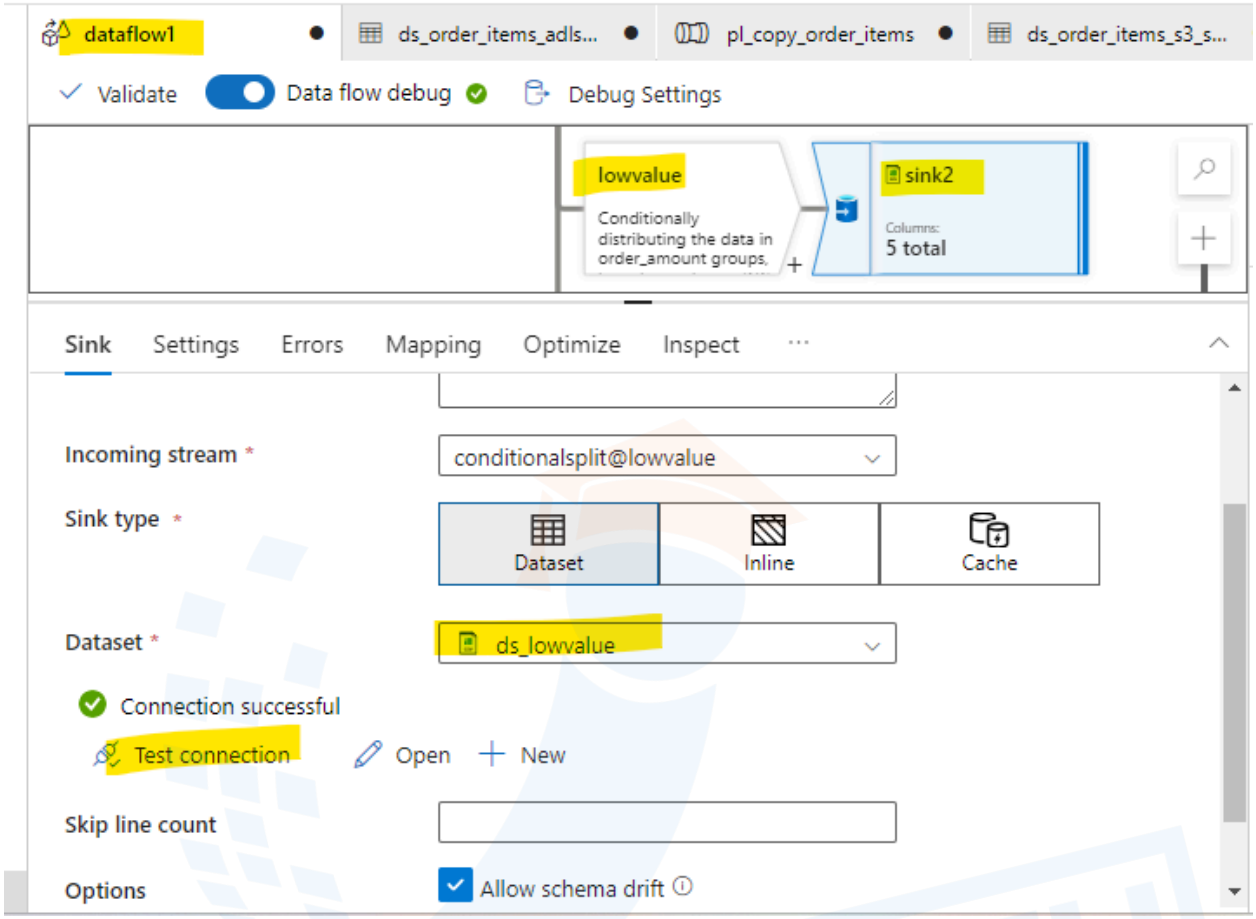


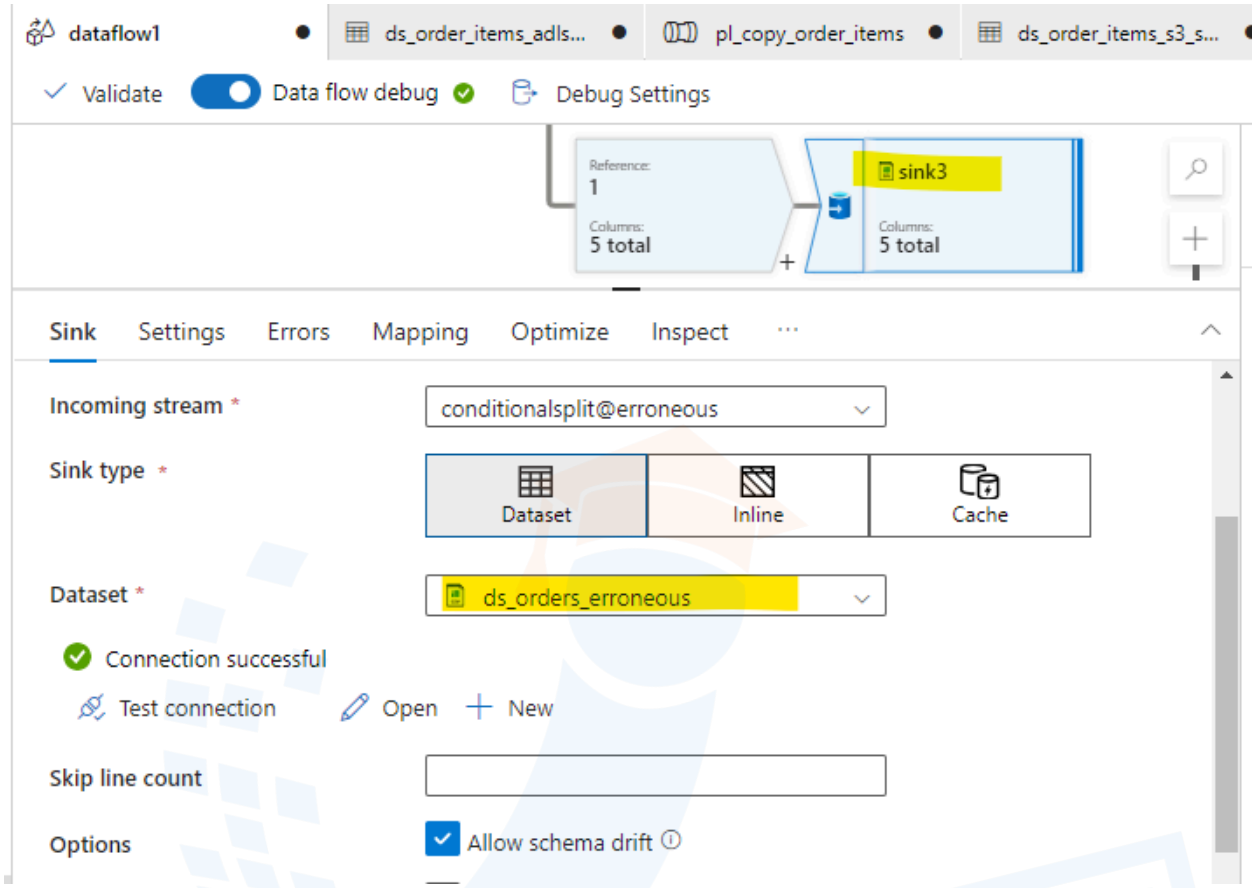
Note: Use single partition option while storing the data





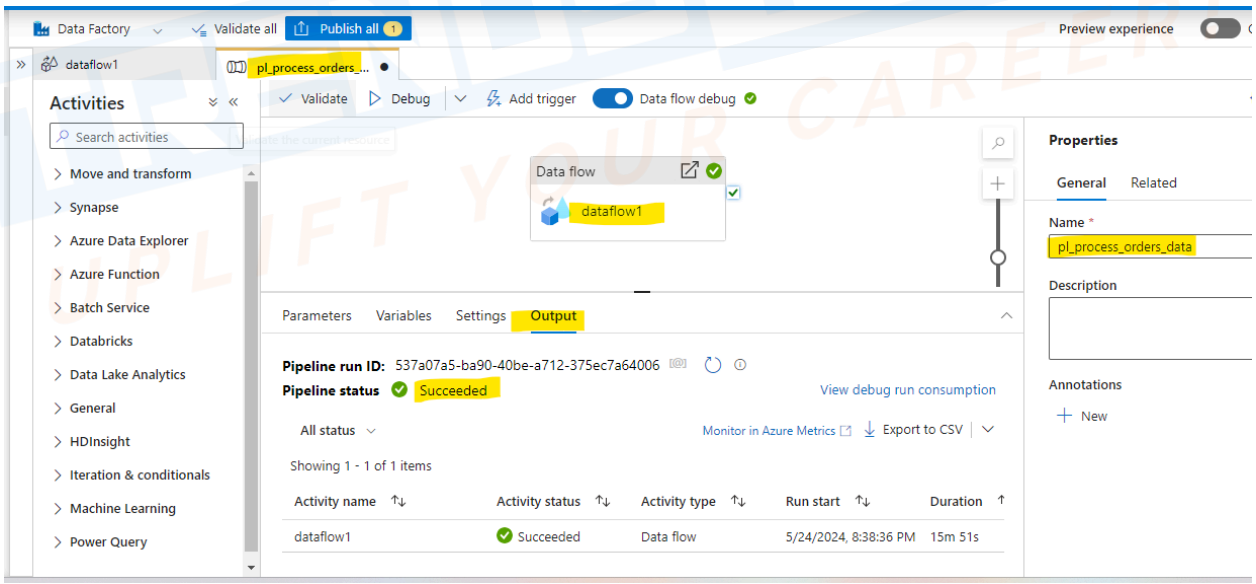
Similarly we will set sink for low value and erroneous values also.





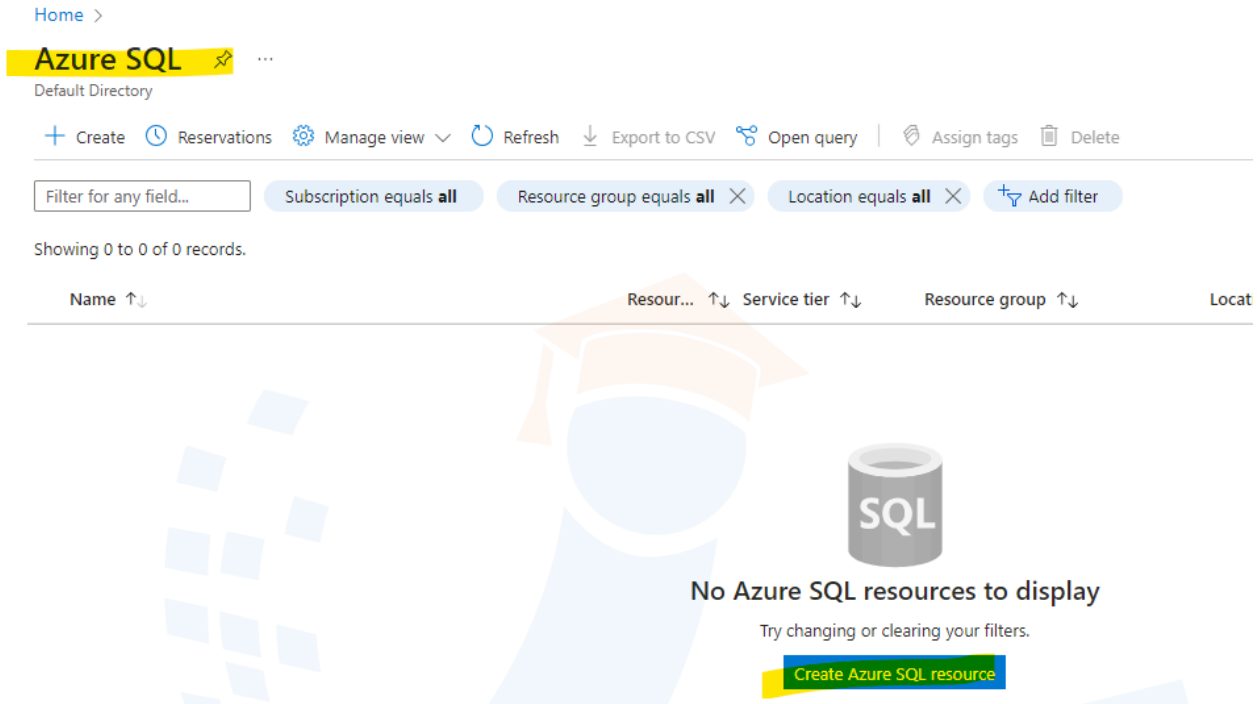
Publish all the changes.

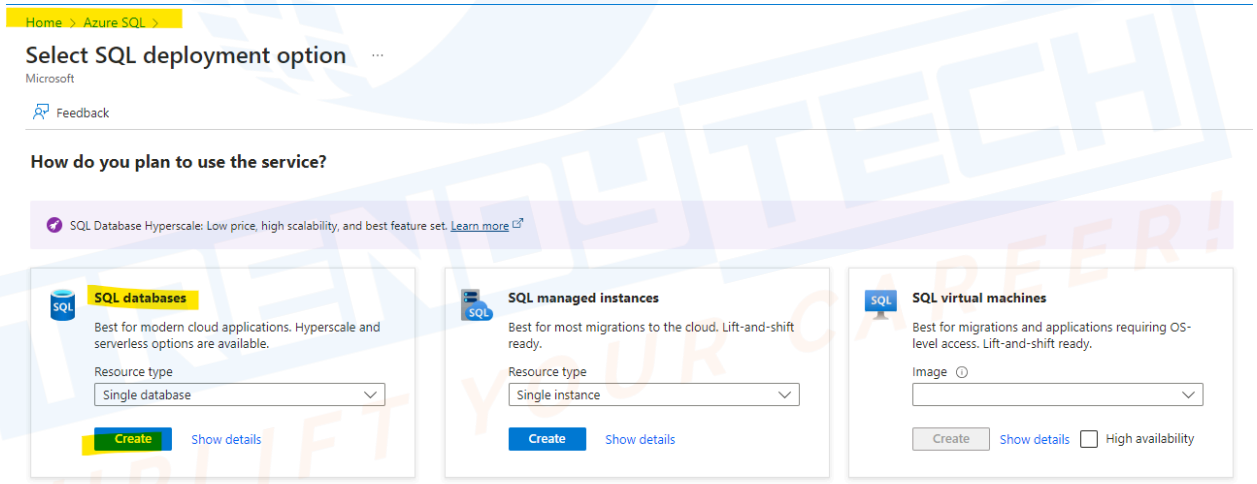
Create pipeline and add this dataflow in it and debug that pipeline



4. For Populating the SQL Database for Reporting Team:

Create and Deploy an Azure SQL Database





Home > Azure SQL > Select SQL deployment option >

Create SQL Database

...

Microsoft

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ: Pay-As-You-Go (136f20b5-00f7-4eb5-a52e-0843e7ad1034)

Resource group * ⓘ: Trendytech-rg

Create new

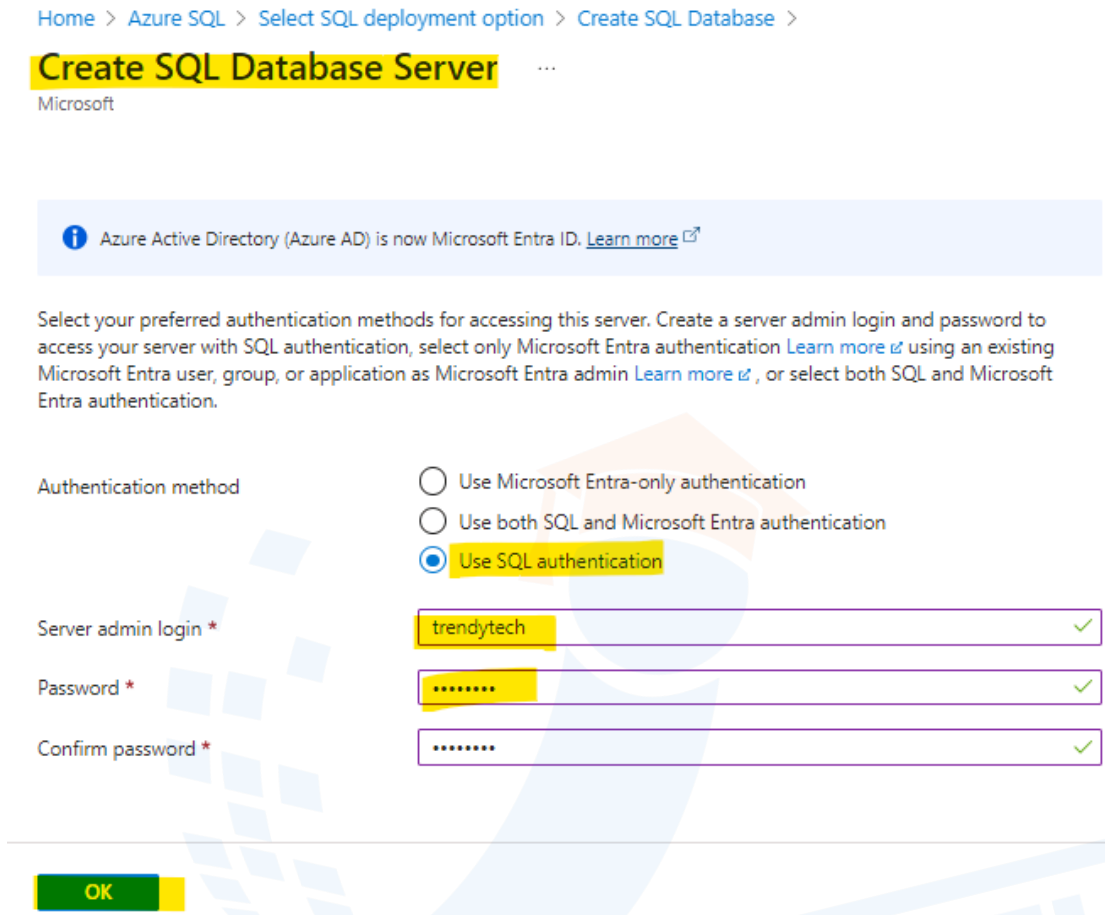
Database details

Enter required settings for this database, including picking a logical server and configuring the compute and storage resources

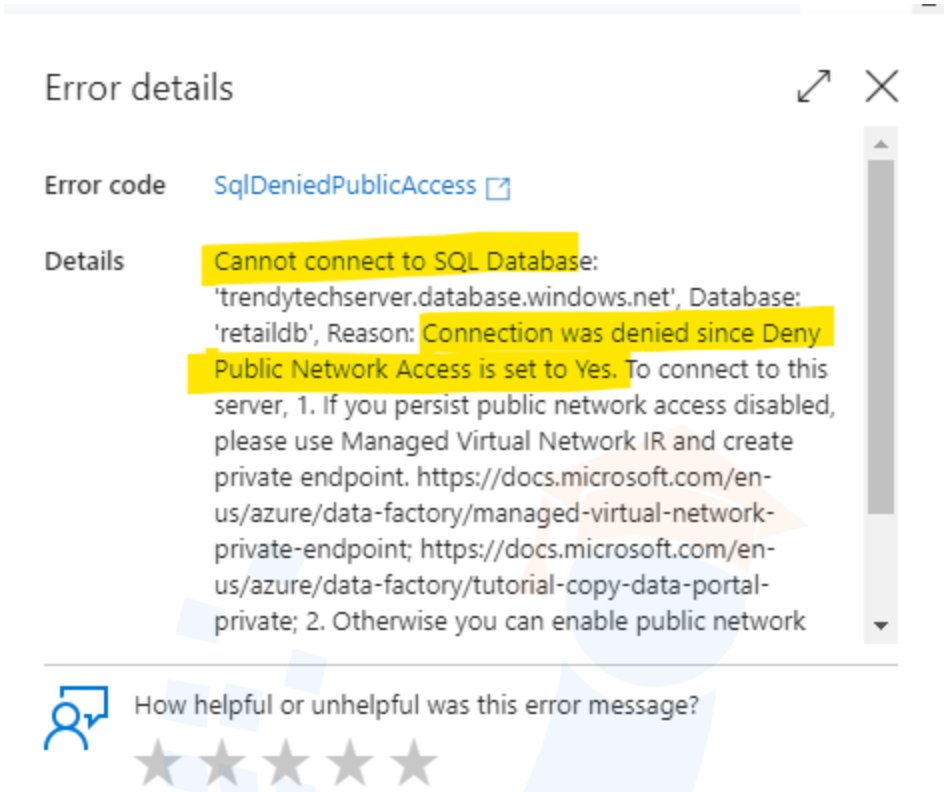
Database name *: retaildb

Server * ⓘ: Select a server

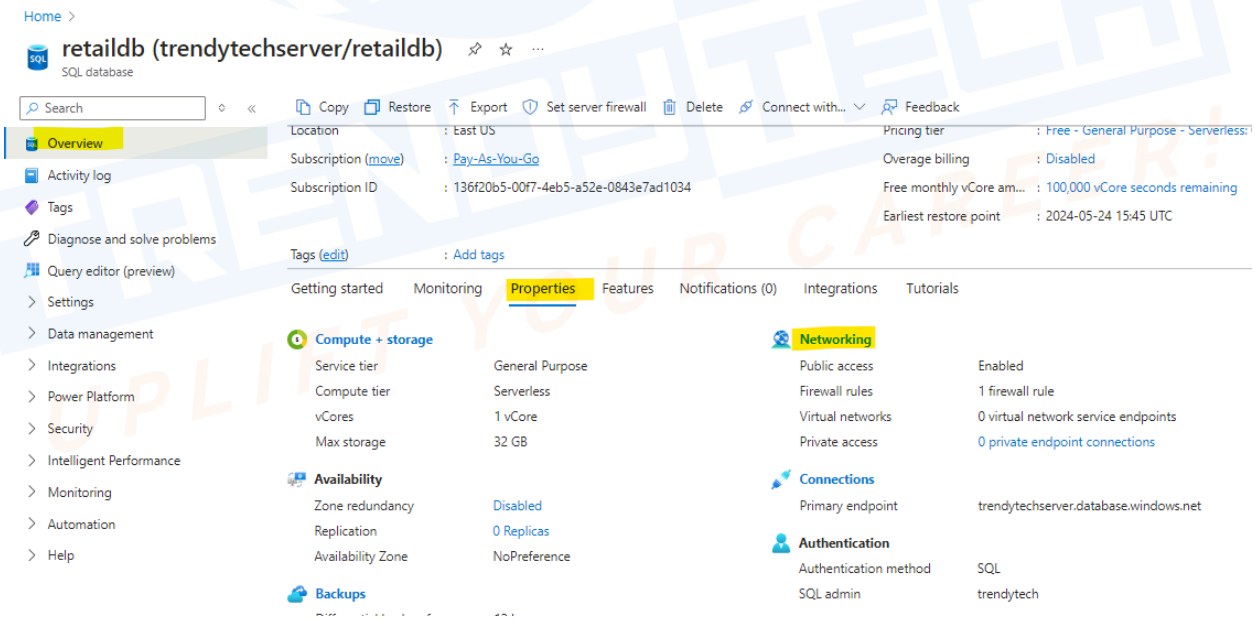
Create new

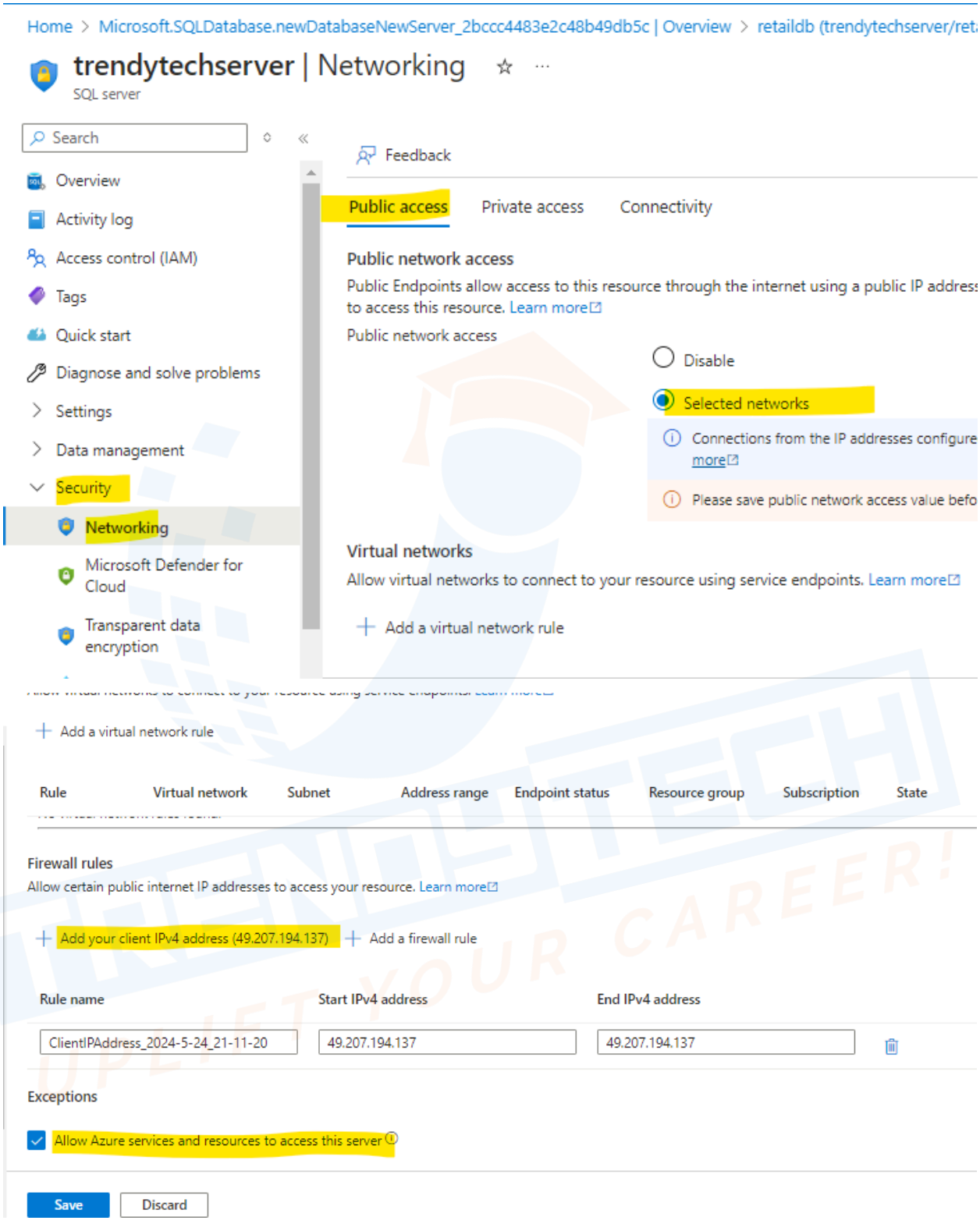


And create database.
If you get below error while connecting to database:



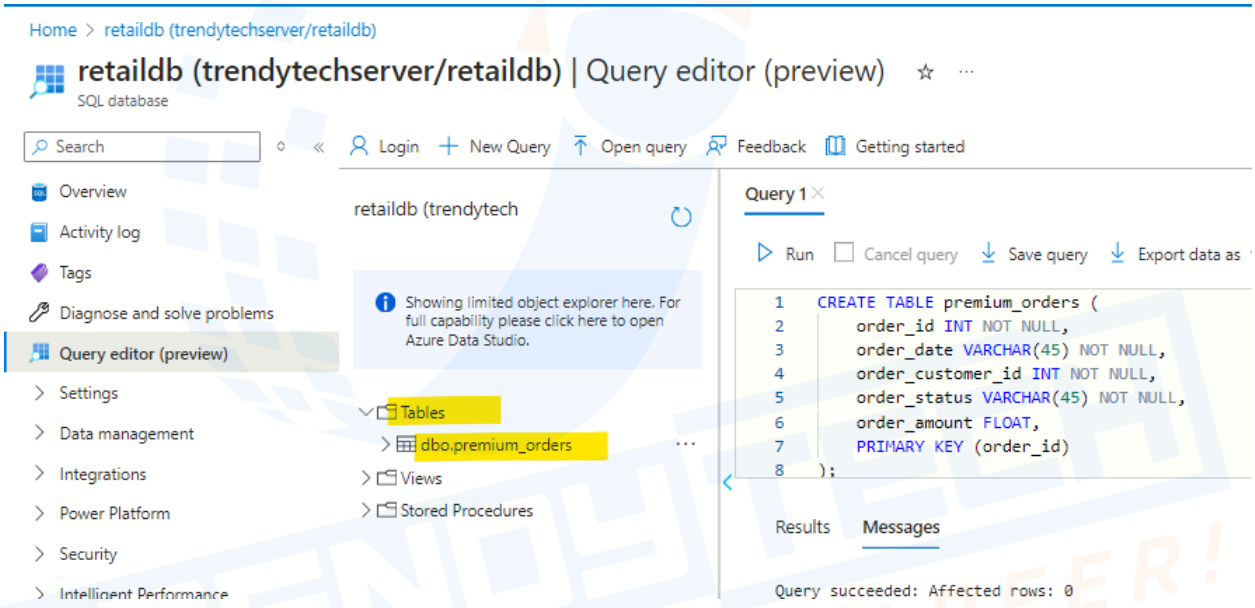
Follow below steps:
Go to “Overview” => Properties => Networking




Home > Microsoft.SQLDatabase.newDatabaseNewServer_2bccc4483e2c48b49db5c | Overview > retaildb (trendytechserver/ret
trendytechserver | Networking
SQL server
Search
Feedback
Overview
Activity log
Access control (IAM)
Tags
Quick start
Diagnose and solve problems
Settings
Data management
Security
Networking
Microsoft Defender for Cloud
Transparent data encryption
Public access
Private access
Connectivity
Public network access
Public Endpoints allow access to this resource through the internet using a public IP address
to access this resource. Learn more
Public network access
Disable
Selected networks
Connections from the IP addresses configure
more
Please save public network access value befo
Virtual networks
Allow virtual networks to connect to your resource using service endpoints. Learn more
Add a virtual network rule
Add a virtual network rule
Rule
Virtual network
Subnet
Address range
Endpoint status
Resource group
Subscription
State
Firewall rules
Allow certain public internet IP addresses to access your resource. Learn more
Add your client IPv4 address (49.207.194.137)
Add a firewall rule
Rule name
Start IPv4 address
End IPv4 address
ClientIPAddress_2024-5-24_21-11-20
49.207.194.137
49.207.194.137
Exceptions
Allow Azure services and resources to access this server
Save
Discard

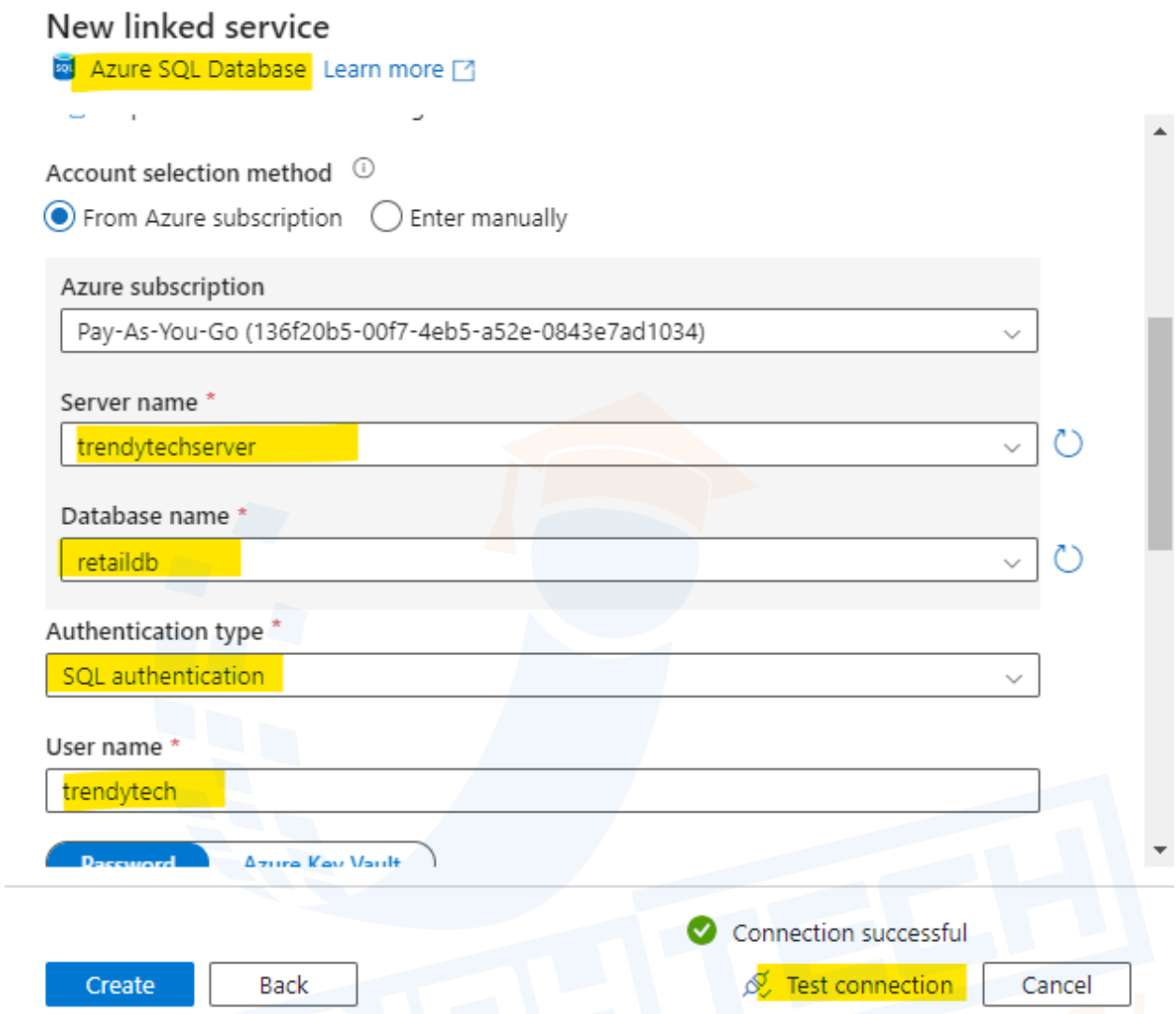
In database create table using below schema

```
CREATE TABLE premium_orders (
    order_id INT NOT NULL,
    order_date VARCHAR(45) NOT NULL,
    order_customer_id INT NOT NULL,
    order_status VARCHAR(45) NOT NULL,
    order_amount FLOAT,
    PRIMARY KEY (order_id)
);
```



Create Linked Services:

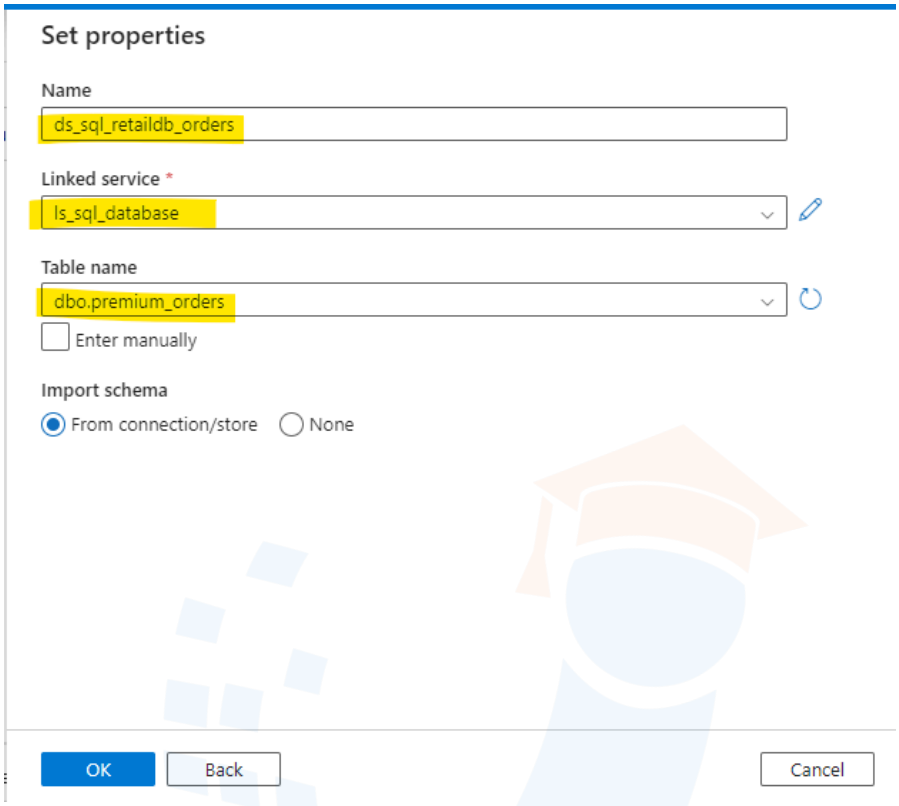
Pointing to SQL DB



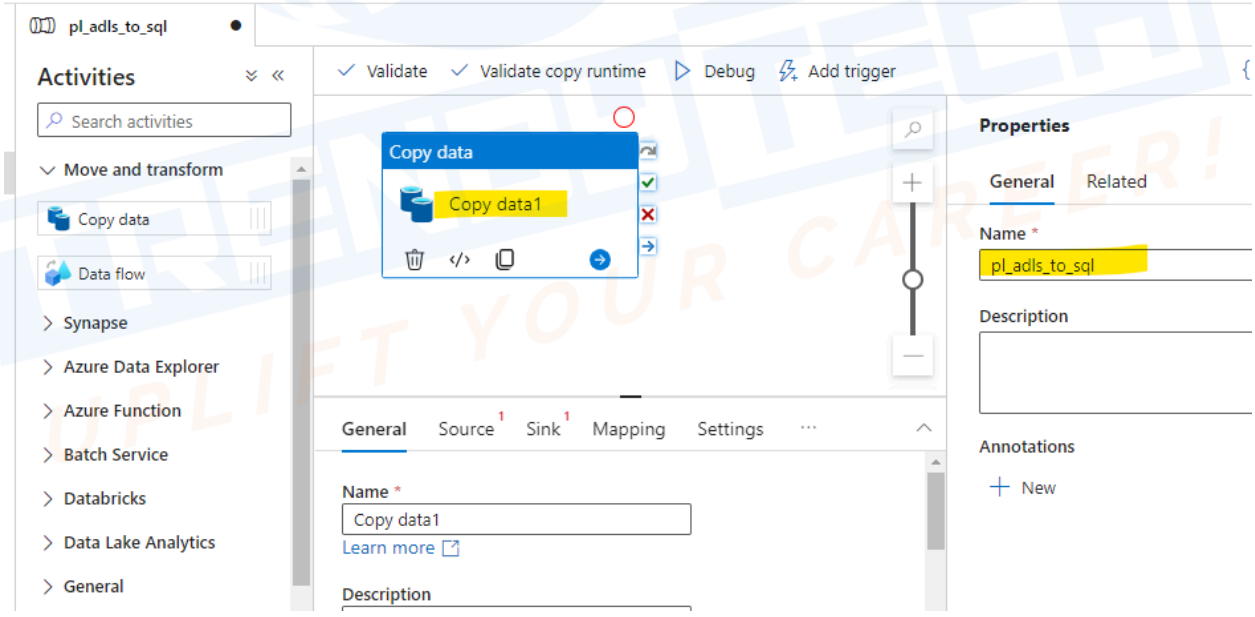
Pointing to ADLS GEN2 where the high value order output is stored.(already created)

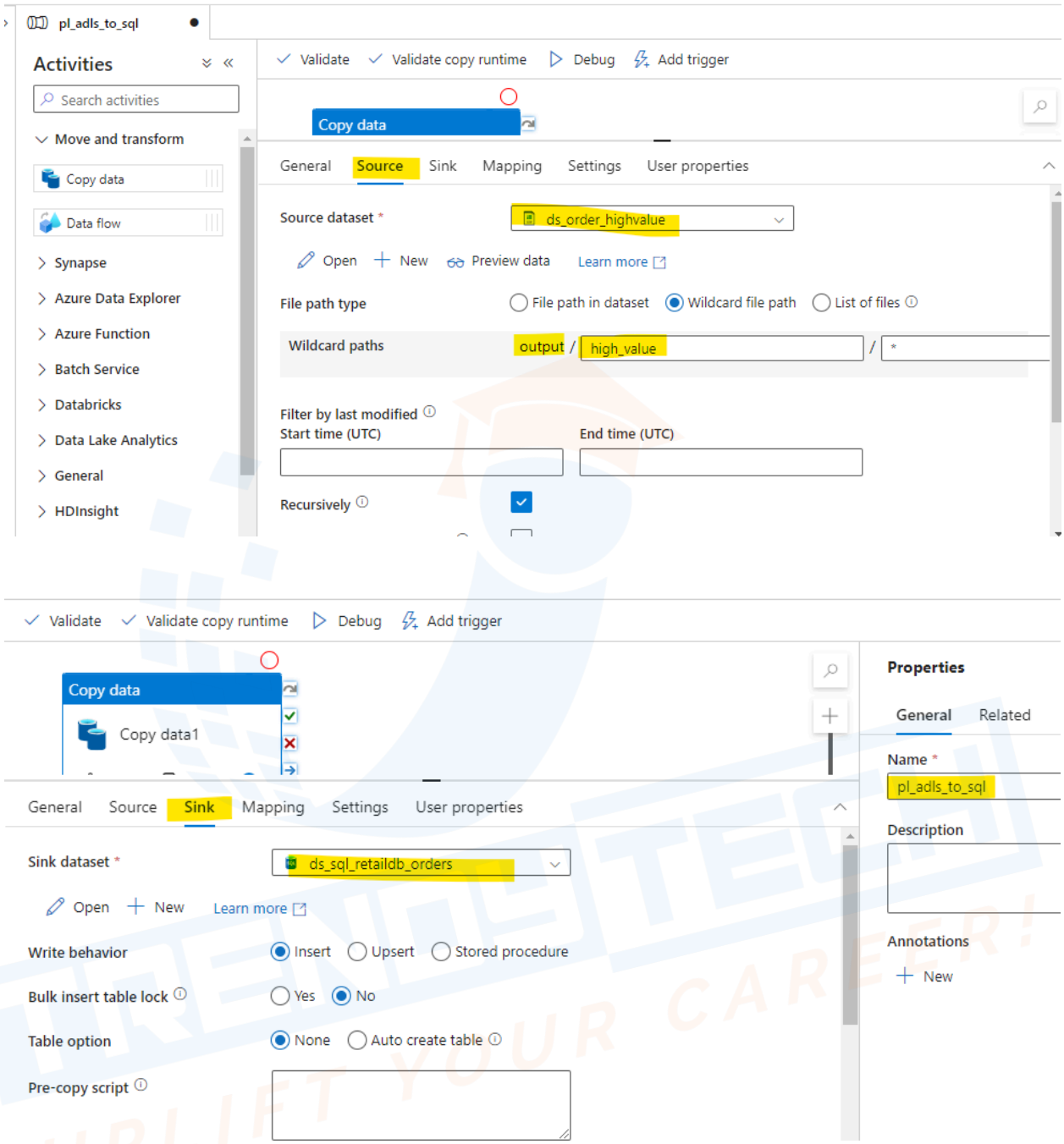
Create Datasets:

Pointing to the high value order data. (already created)
Pointing to the table in SQL DB.

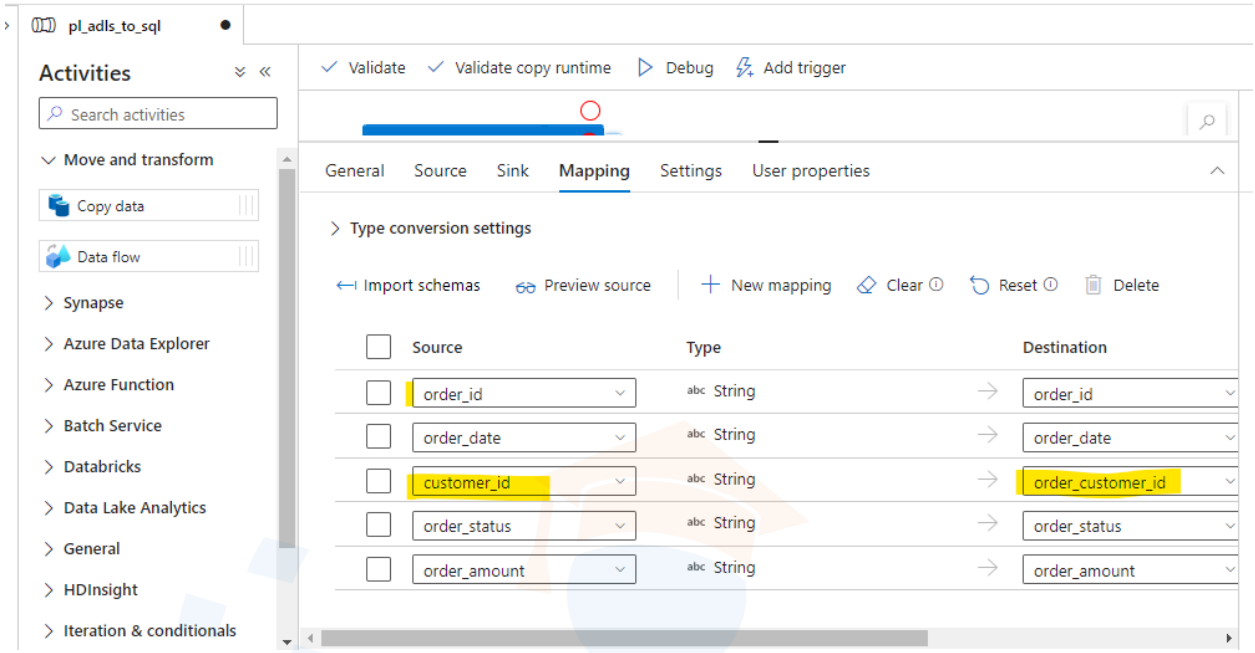


Create a Pipeline (copy pipeline adls to sql):





And define the mapping as shown below.



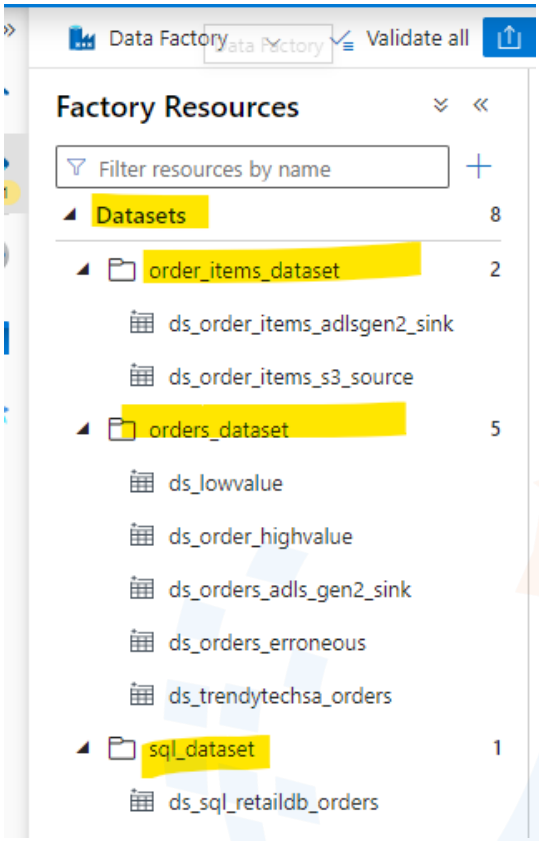
Push data from ADLS GEN2 (Source) to Azure SQL DB (Sink).

Debug the pipeline.

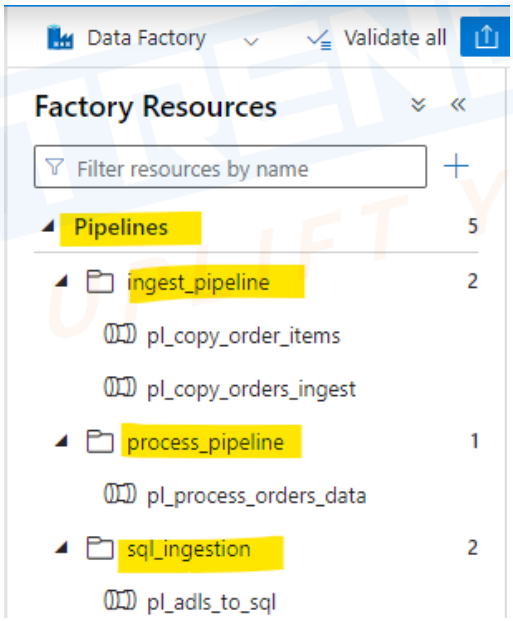
Note: remove the data from all the folders and truncate table also before executing the final pipeline.

Organize Datasets and Pipelines into Folders:

1. Create folders for datasets and add the datasets to their respective folders (orders dataset, order_items dataset, sql).



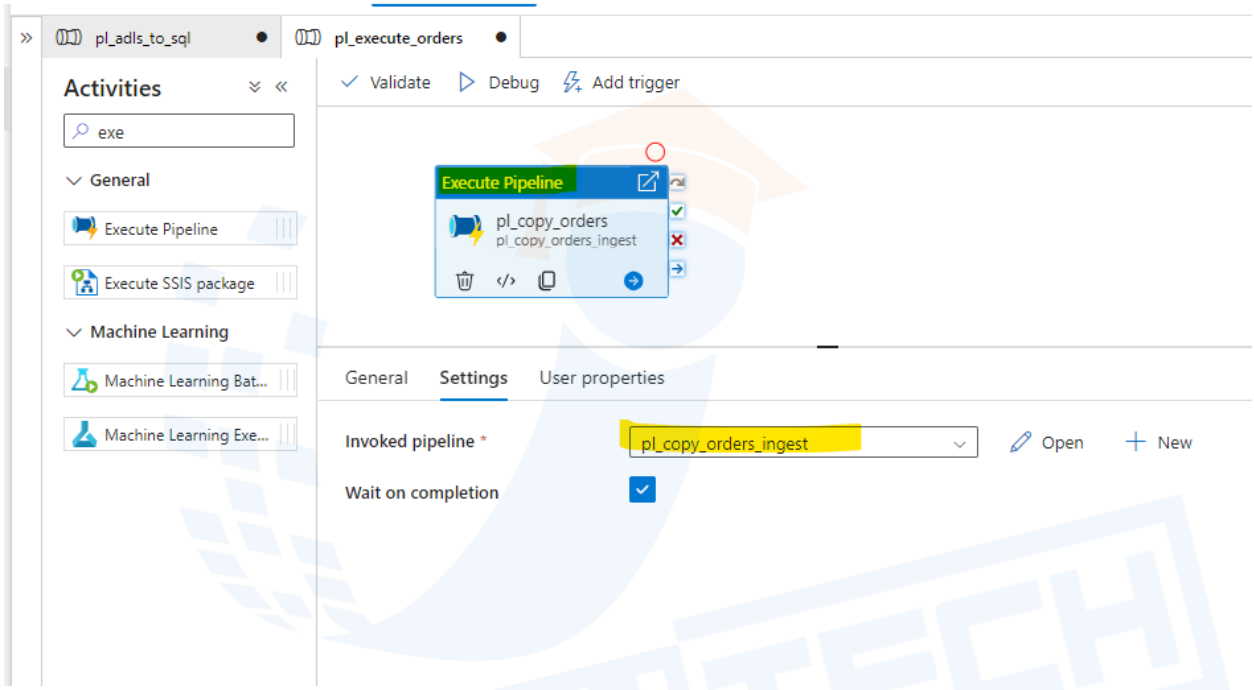
2. Create folders for pipelines and add the pipelines associated to specific activities to their respective folders (ingest, process, sql ingestion).



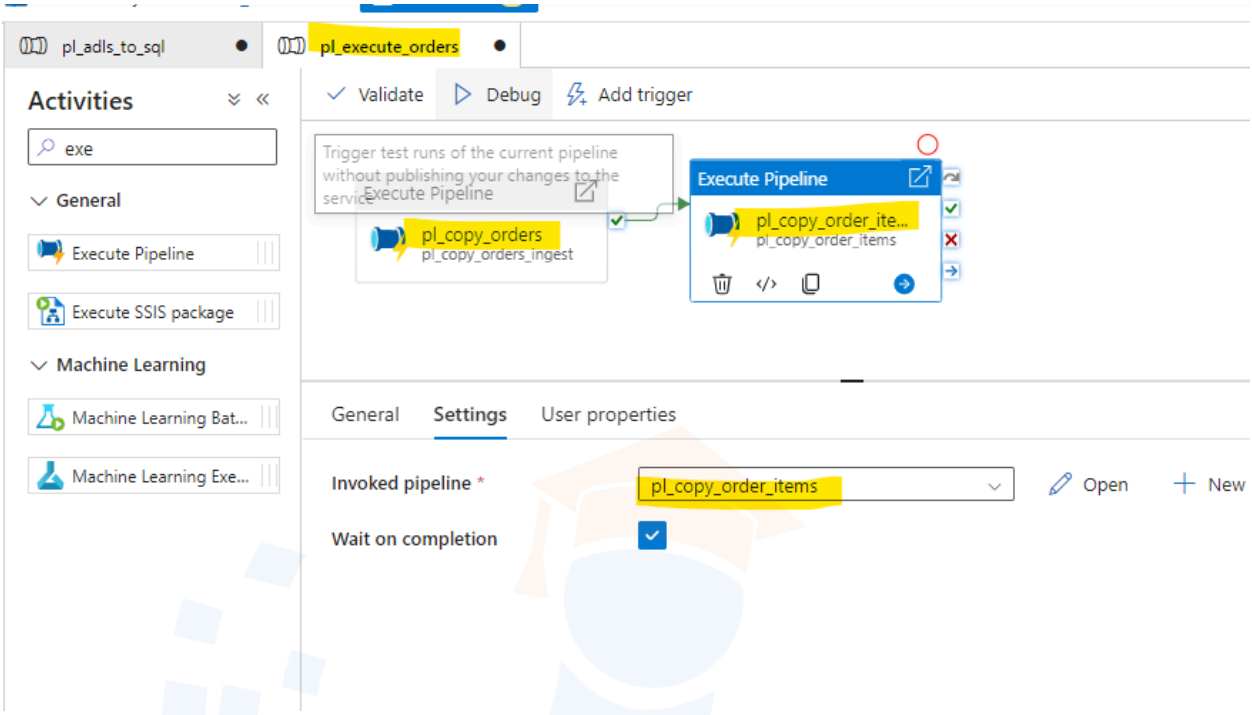
3. Create an Execute Pipeline Activity:

Chain all the pipelines to get executed in a specific order.

We will first add the pipeline pl_copy_orders using the option execution pipeline.

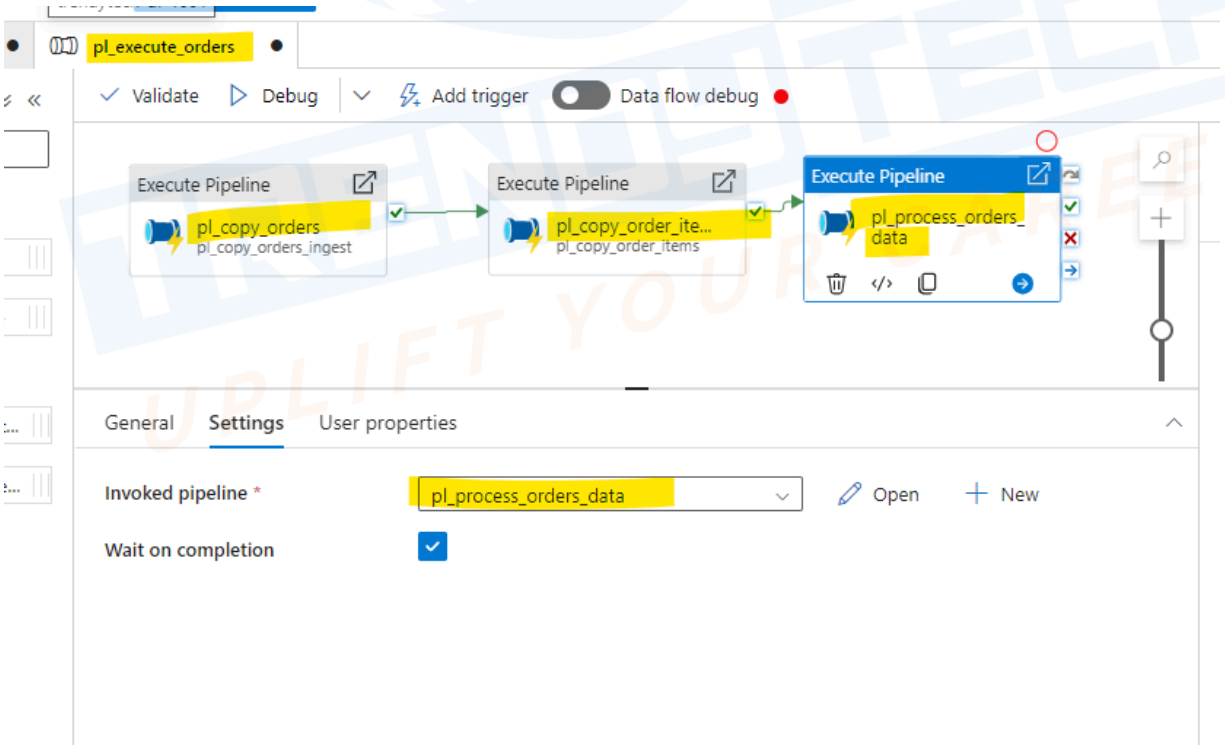


Now we will add the pipeline “pl_copy_order_items” as shown below using the option execution pipeline.

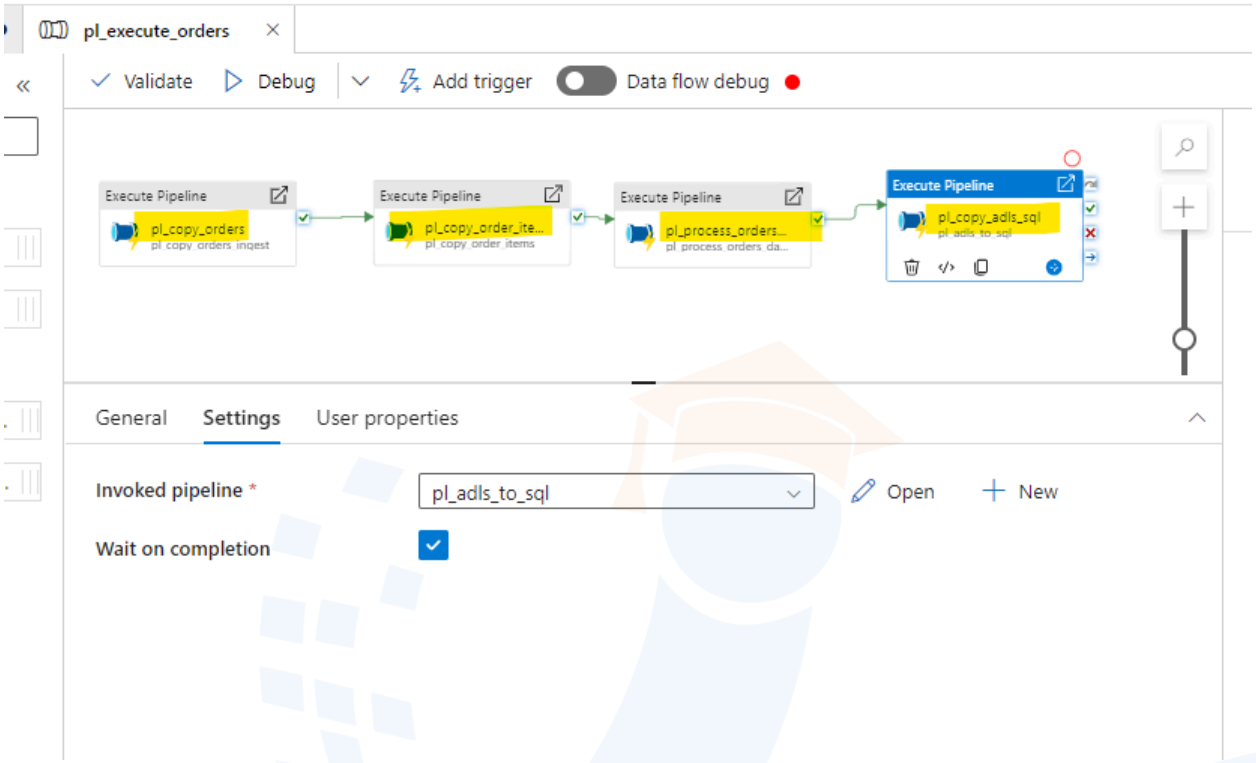


We will add the pipeline pl_process_orders_data as shown below using the option execution pipeline.

.

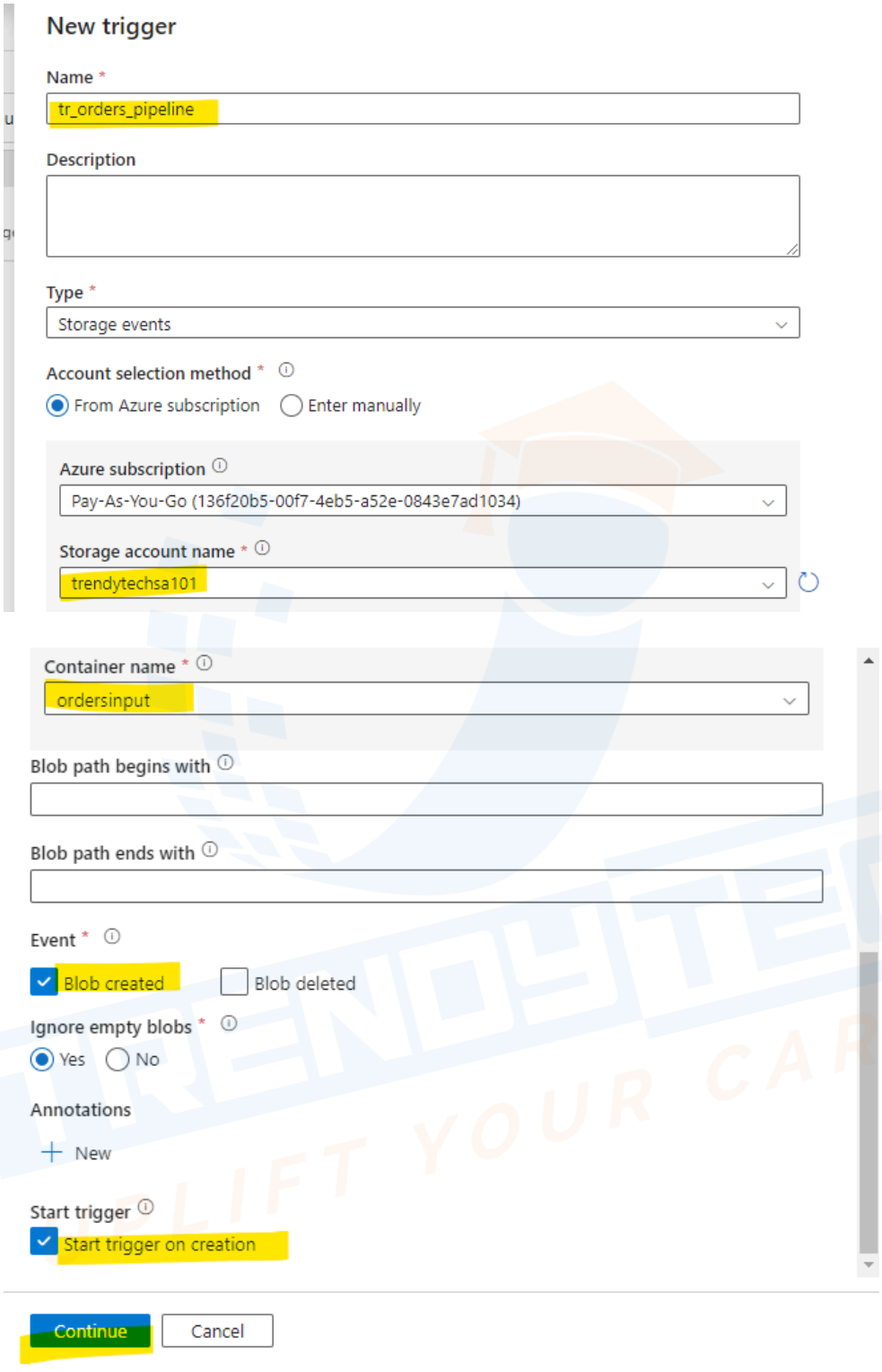


Our final pipeline will be as shown below.



4. Create a Trigger:

   Ingest and process data with Storage Event Type (the trigger will get initiated whenever a new data file arrives at the blob storage as mentioned in the storage account and container name fields while creating the trigger).



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.

5. Attach the Created Trigger to the Execute Pipeline:

Trigger the pipeline execution without any manual intervention whenever a new file gets added to the blob storage.

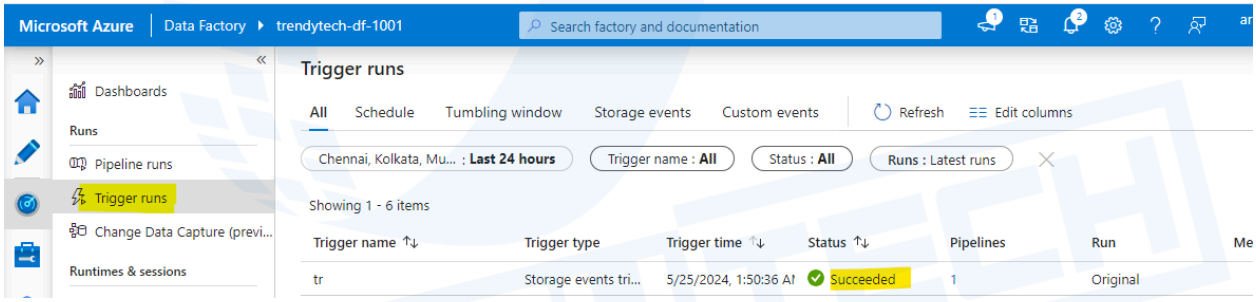
If you get below error

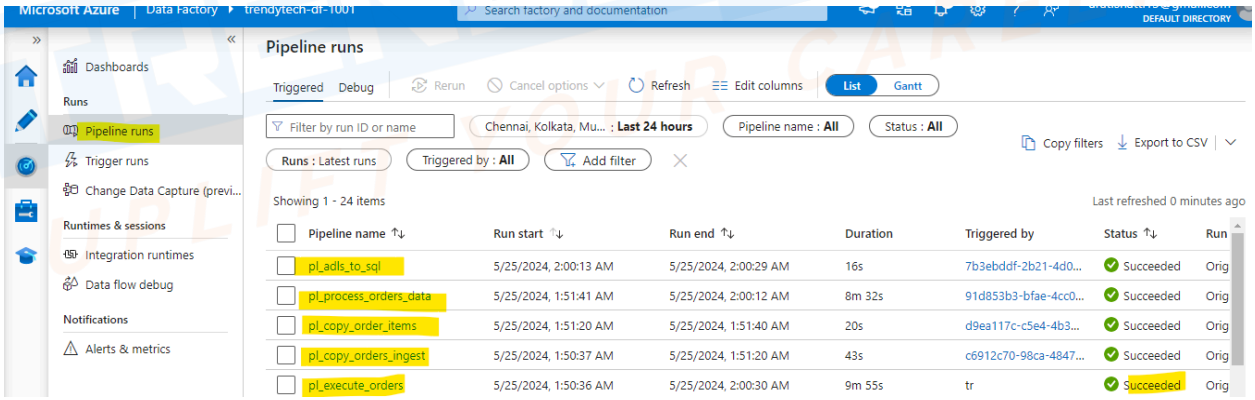
Register Azure Event Grid resource provider to your subscription before creating an event trigger. Learn more here

Follow this document:

https://learn.microsoft.com/en-us/azure/azure-resource-manager/management/resource-providers-and-types#azure-portal

Monitor the trigger and pipeline under the monitor tab.





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