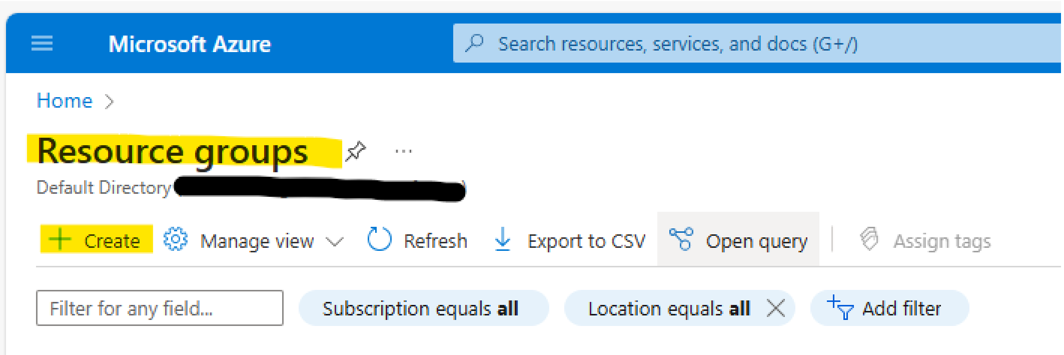
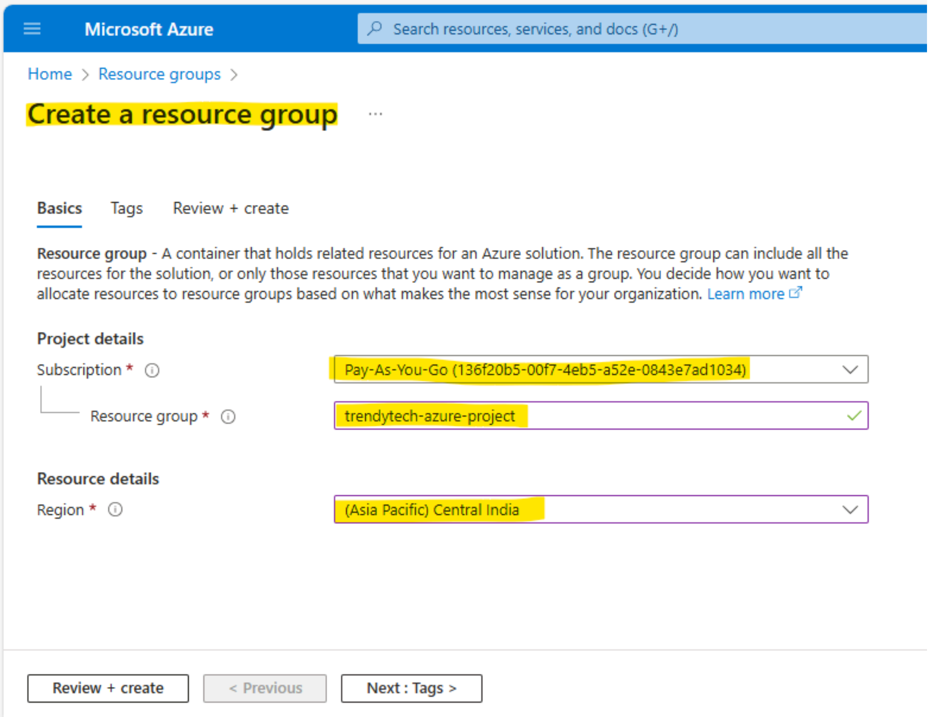
Part 1:

Before creating all the resources we will create the resource group in which we will create all the required resources.

Go to Azure Portal and log in with your Azure account. In the left-hand menu, select "Resource groups". If you don't see it, use the search bar at the top of the page and search for "Resource groups". Click the "+ Create" button or "Add" at the top of the Resource groups page.



Select the subscription under which the resource group will be created. Enter a unique name for your resource group. Choose a location (region) where your resources will reside (e.g., East US, West Europe).



Resource group - A container that holds related resources for an Azure solution. The resource group can include all the resources for the solution, or only those resources that you want to manage as a group. You decide how you want to allocate resources to resource groups based on what makes the most sense for your organization. Learn more

Project details

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

Resource group *: trendytech-azure-project

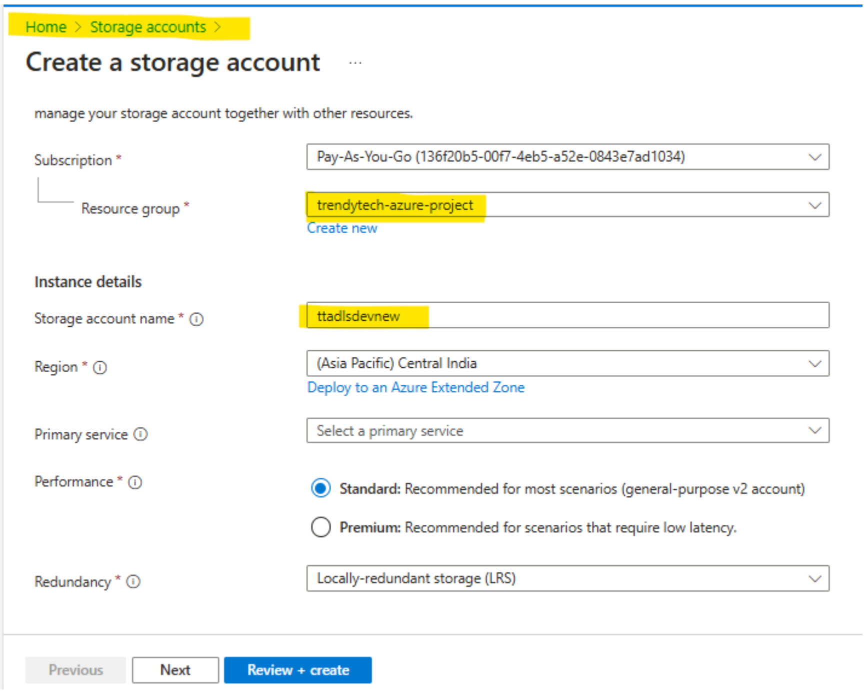
Resource details

Region *: (Asia Pacific) Central India

Click "Review + Create" and then "Create".

Azure Storage Account creation:

Search for "Storage accounts" and click "+ Create". Select subscription, resource group, region, and enter a unique storage account name as shown below.



Home > Storage accounts >

Create a storage account ···

manage your storage account together with other resources.

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

Resource group * trendytech-azure-project

Create new

Instance details

Storage account name * ⓘ ttadlsdevnew

Region * ⓘ (Asia Pacific) Central India

Deploy to an Azure Extended Zone

Primary service ⓘ Select a primary service

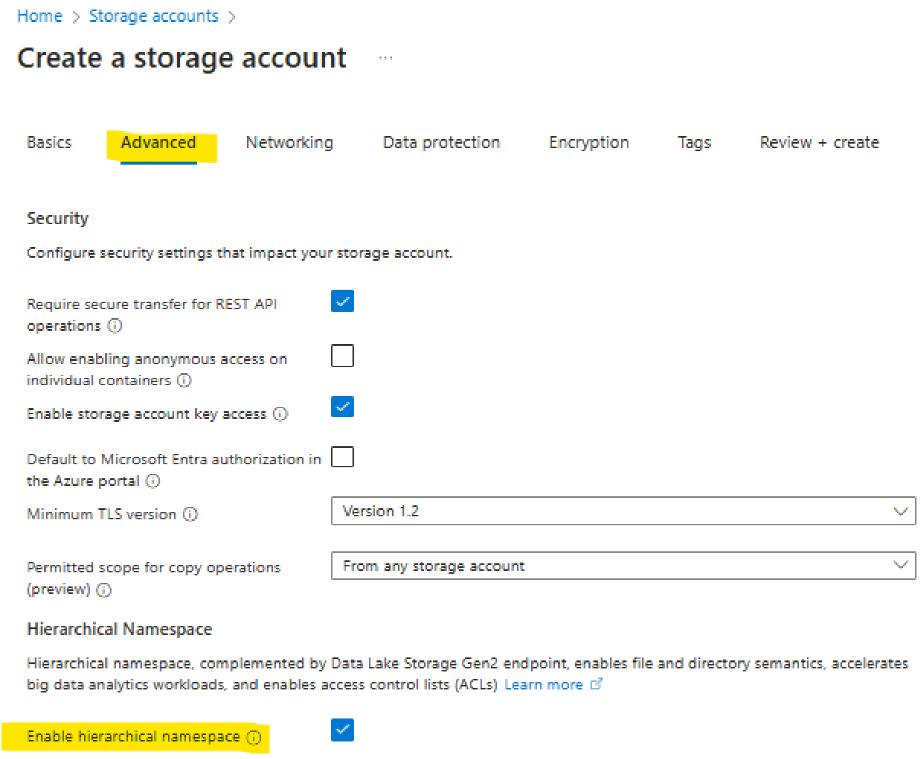
Performance * ⓘ

● **Standard:** Recommended for most scenarios (general-purpose v2 account)

○ **Premium:** Recommended for scenarios that require low latency.

Redundancy * ⓘ Locally-redundant storage (LRS)

Enable ADLS Gen2: Go to the Advanced tab and enable Hierarchical namespace.



Home > Storage accounts >

Create a storage account ···

Security

Configure security settings that impact your storage account.

Require secure transfer for REST API operations ⓘ ☑

Allow enabling anonymous access on individual containers ⓘ □

Enable storage account key access ⓘ ☑

Default to Microsoft Entra authorization in the Azure portal ⓘ □

Minimum TLS version ⓘ Version 1.2

Permitted scope for copy operations (preview) ⓘ From any storage account

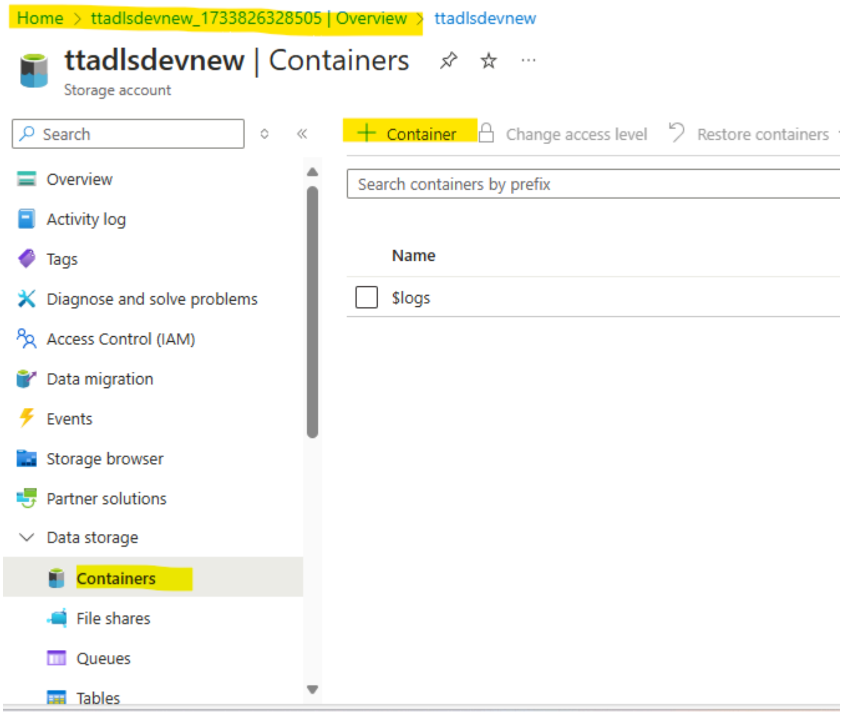
Hierarchical Namespace

Hierarchical namespace, complemented by Data Lake Storage Gen2 endpoint, enables file and directory semantics, accelerates big data analytics workloads, and enables access control lists (ACLs) Learn more

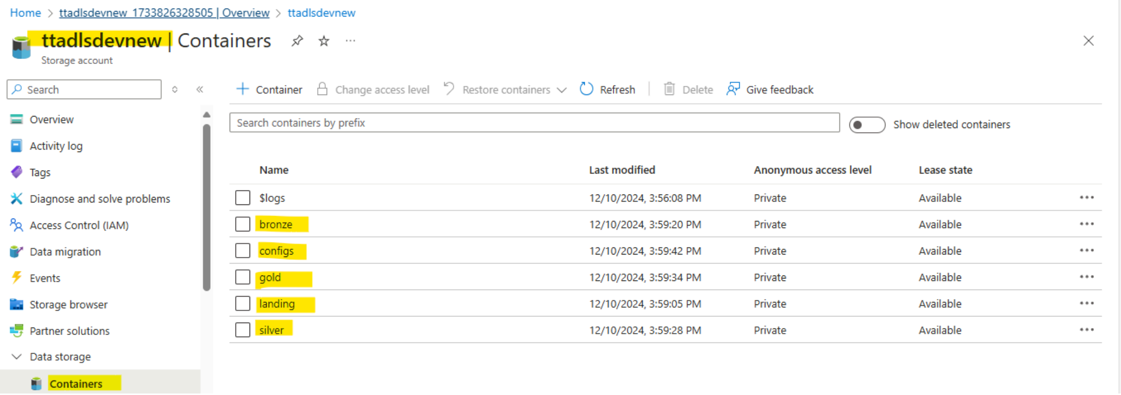
Enable hierarchical namespace ⓘ ☑

Click Review + create, validate settings, and click Create.

Now create the containers “landing”, “bronze”, “silver”, “gold”, “configs” in this storage account as shown below.

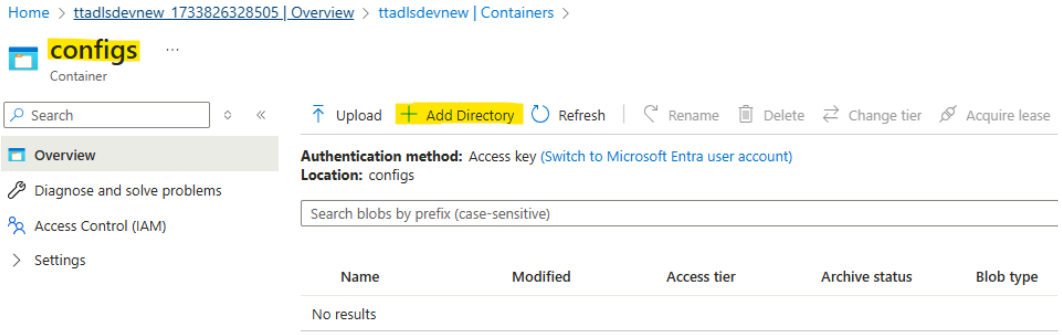


	Name
□	$logs



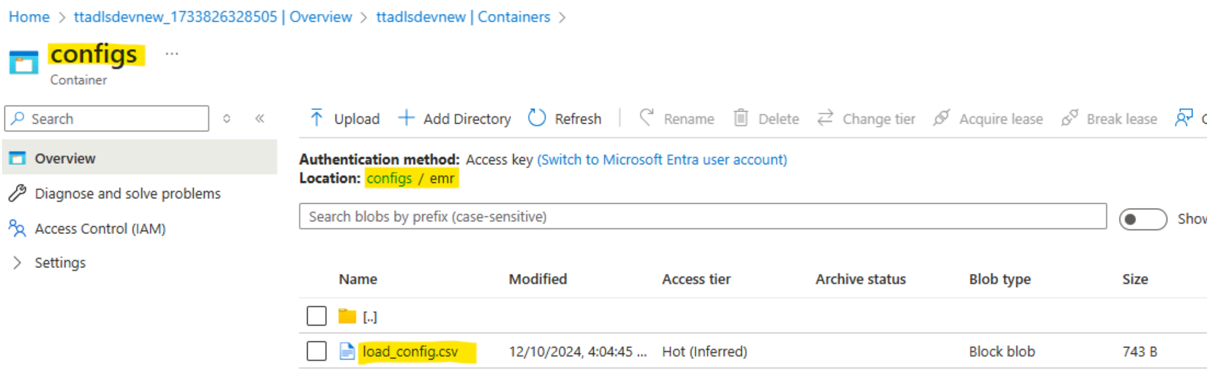
	Name	Last modified	Anonymous access level	Lease state	
□	$logs	12/10/2024, 3:56:08 PM	Private	Available	•••
□	bronze	12/10/2024, 3:59:20 PM	Private	Available	•••
□	configs	12/10/2024, 3:59:42 PM	Private	Available	•••
□	gold	12/10/2024, 3:59:34 PM	Private	Available	•••
□	landing	12/10/2024, 3:59:05 PM	Private	Available	•••
□	silver	12/10/2024, 3:59:28 PM	Private	Available	•••

Then in the configs container create the directory “emr” and then upload the file “load_config.csv” in it.


Home > ttadlsdevnew_1733826328505 | Overview > ttadlsdevnew | Containers >
configs
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Upload
+ Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Search blobs by prefix (case-sensitive)

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: configs

Name	Modified	Access tier	Archive status	Blob type
No results				



Home > ttadlsdevnew_1733826328505 | Overview > ttadlsdevnew | Containers >

configs ...
Container

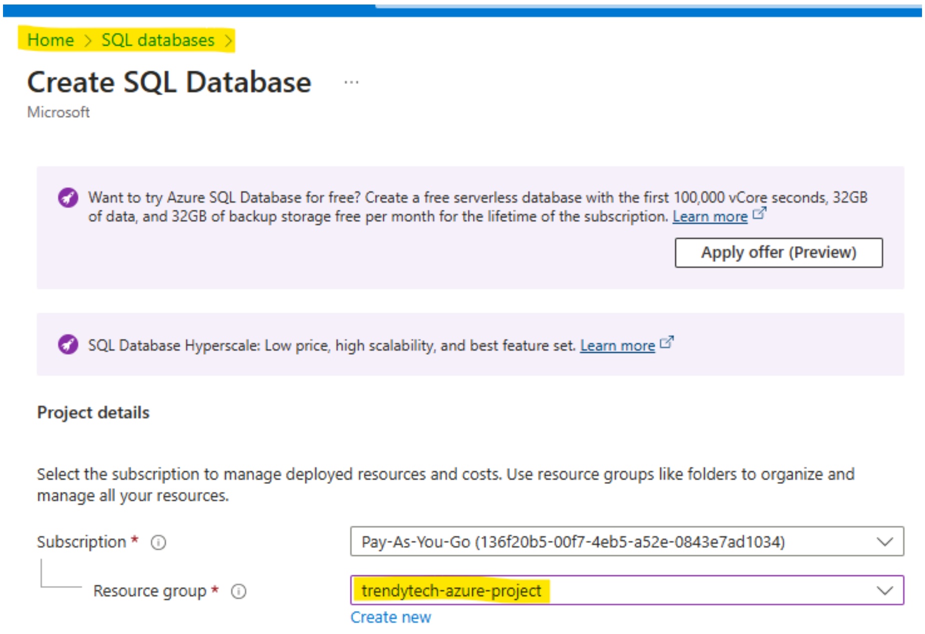
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: configs / emr

	Name	Modified	Access tier	Archive status	Blob type	Size
□	[..]					
□	load_config.csv	12/10/2024, 4:04:45 ...	Hot (Inferred)		Block blob	743 B

Steps to create Azure SQL database:

We will create 2 azure SQL db - trendytech-hospital-a, trendytech-hospital-b

In the search bar, type "SQL Database" and select "SQL Database" from the results. Choose your Subscription and Resource Group. Enter a Database Name. Also create a SQL Server.



Home > SQL databases >

Create SQL Database ...
Microsoft

Want to try Azure SQL Database for free? Create a free serverless database with the first 100,000 vCore seconds, 32GB of data, and 32GB of backup storage free per month for the lifetime of the subscription. Learn more

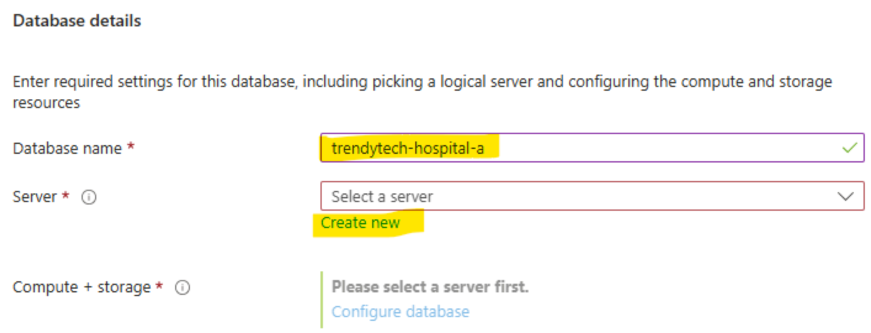
SQL Database Hyperscale: Low price, high scalability, and best feature set. Learn more

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-azure-project

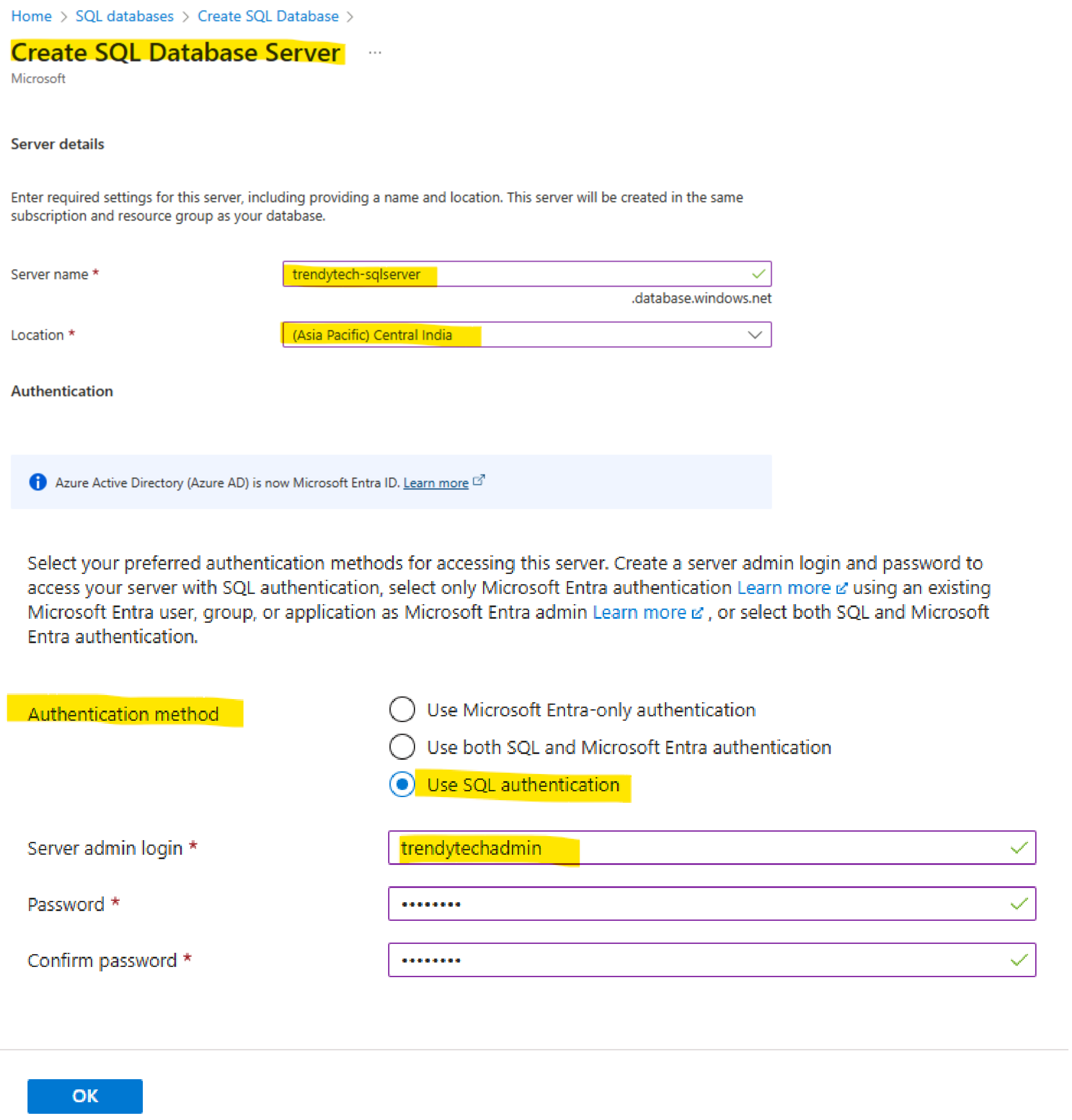


Database details

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

Database name *	trendytech-hospital-a ✓
Server * ⓘ	Select a server Create new
Compute + storage * ⓘ	Please select a server first. Configure database

We will create the server as shown below. Choose a Compute + Storage tier (e.g., Basic, General Purpose).



Create SQL Database Server ...

Microsoft

Server details

Enter required settings for this server, including providing a name and location. This server will be created in the same subscription and resource group as your database.

Server name *	trendytech-sqlserver ✓ .database.windows.net
Location *	(Asia Pacific) Central India

Authentication

ⓘ Azure Active Directory (Azure AD) is now Microsoft Entra ID. Learn more

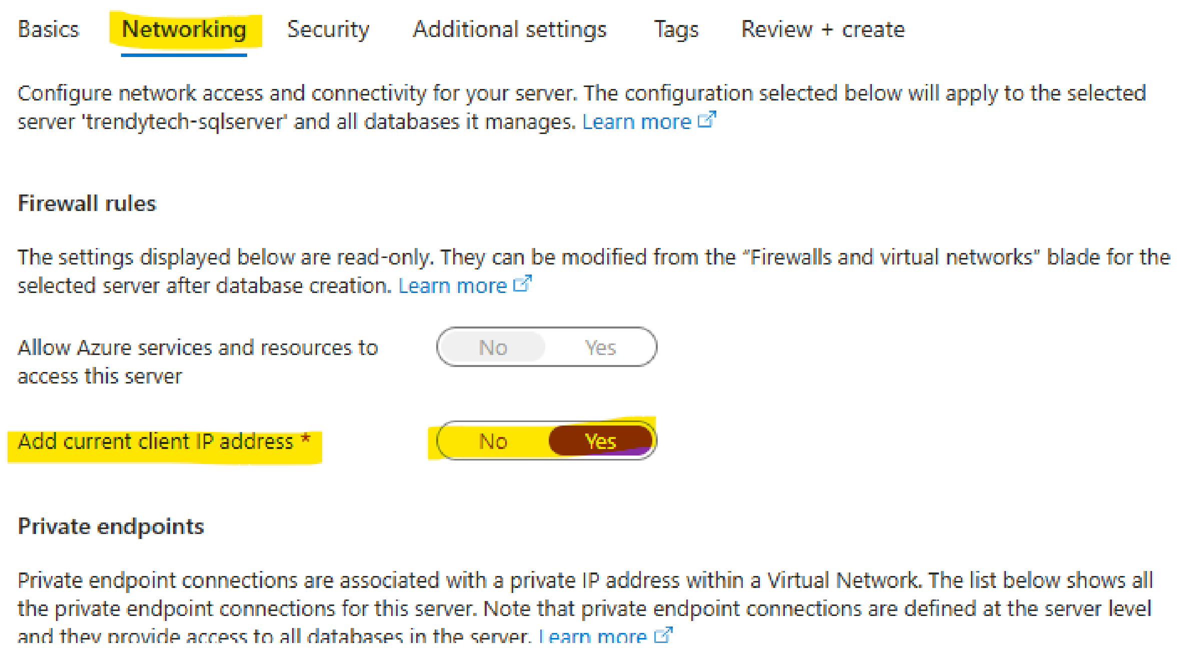
Select your preferred authentication methods for accessing this server. Create a server admin login and password to access your server with SQL authentication, select only Microsoft Entra authentication Learn more using an existing Microsoft Entra user, group, or application as Microsoft Entra admin Learn more , or select both SQL and Microsoft Entra authentication.

Authentication method

○ Use Microsoft Entra-only authentication
○ Use both SQL and Microsoft Entra authentication
● Use SQL authentication

Server admin login *	trendytechadmin ✓
Password *	•••••••• ✓
Confirm password *	•••••••• ✓

After this Go to Networking => In Network connectivity select “Public endpoint” option. Also set yes for the options “Allow Azure services and resources to access this server” and “Add current client IP address”



Configure network access and connectivity for your server. The configuration selected below will apply to the selected server 'trendytech-sqlserver' and all databases it manages. Learn more

Firewall rules

The settings displayed below are read-only. They can be modified from the "Firewalls and virtual networks" blade for the selected server after database creation. Learn more

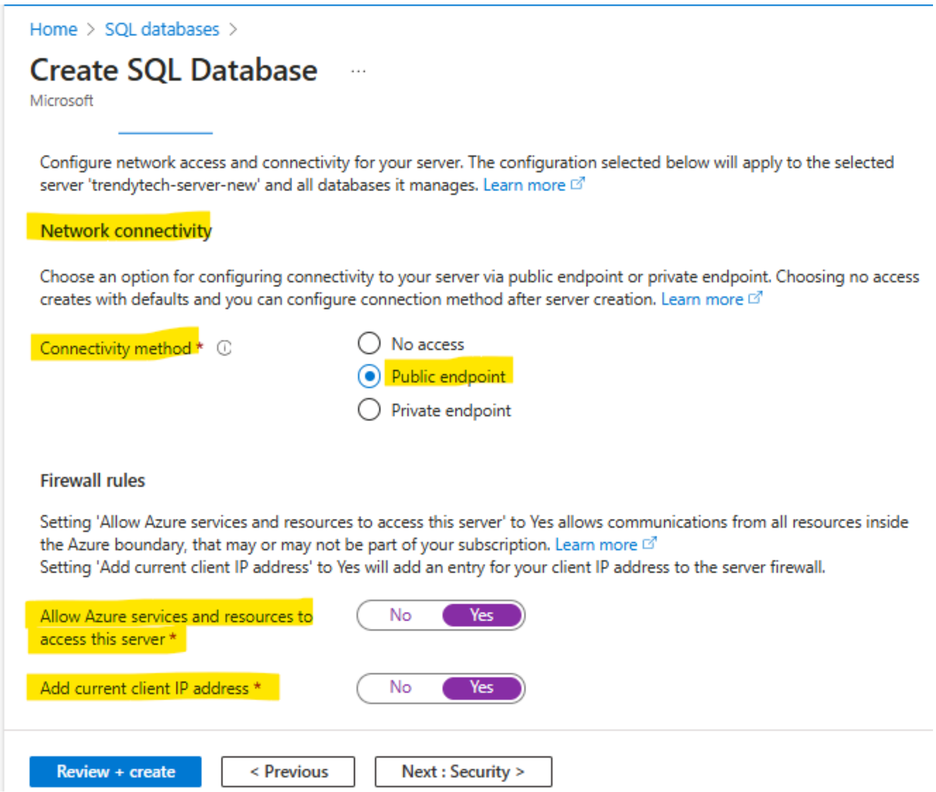
Allow Azure services and resources to access this server: [✓ No] Yes

Add current client IP address *: No [✓ Yes]

Private endpoints

Private endpoint connections are associated with a private IP address within a Virtual Network. The list below shows all the private endpoint connections for this server. Note that private endpoint connections are defined at the server level and they provide access to all databases in the server. Learn more

Note: Please note down this username and password for future reference.



Configure network access and connectivity for your server. The configuration selected below will apply to the selected server 'trendytech-server-new' and all databases it manages. Learn more

Network connectivity

Choose an option for configuring connectivity to your server via public endpoint or private endpoint. Choosing no access creates with defaults and you can configure connection method after server creation. Learn more

Connectivity method *
○ No access
● Public endpoint
○ Private endpoint

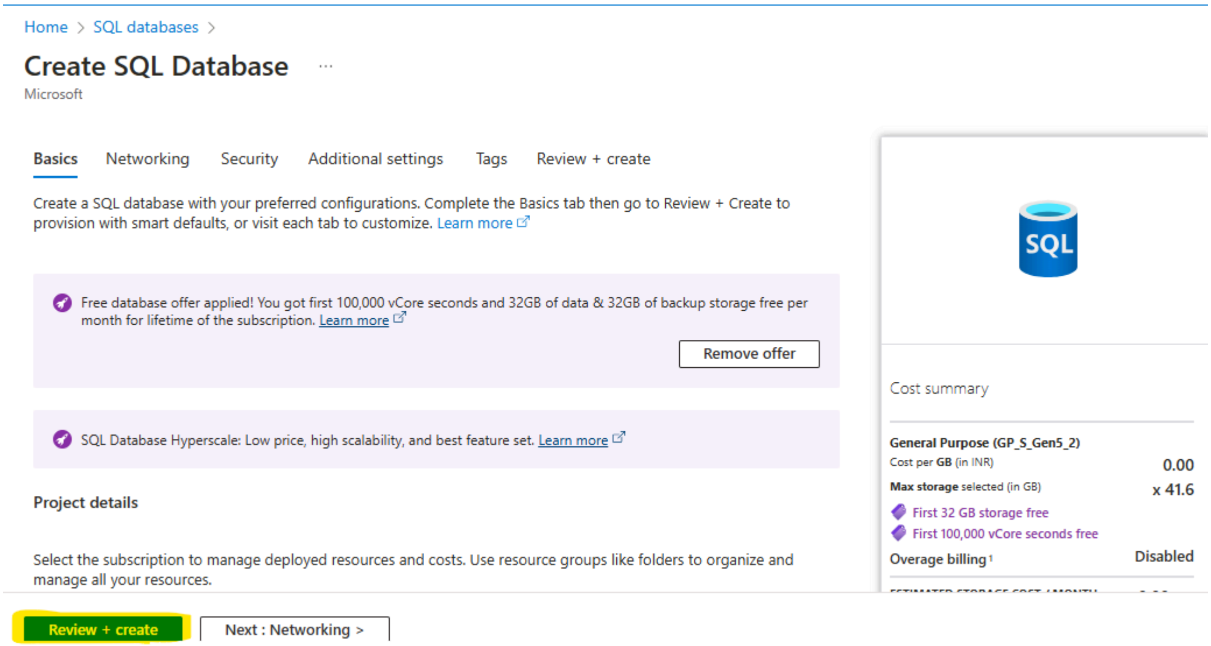
Firewall rules

Setting 'Allow Azure services and resources to access this server' to Yes allows communications from all resources inside the Azure boundary, that may or may not be part of your subscription. Learn more
Setting 'Add current client IP address' to Yes will add an entry for your client IP address to the server firewall.

Allow Azure services and resources to access this server *: No [✓ Yes]

Add current client IP address *: No [✓ Yes]

Click Review + Create, validate the details, and then click Create as shown below.



Home > SQL databases >

Create SQL Database ...

Microsoft

Create a SQL database with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. Learn more

Free database offer applied! You got first 100,000 vCore seconds and 32GB of data & 32GB of backup storage free per month for lifetime of the subscription. Learn more

SQL Database Hyperscale: Low price, high scalability, and best feature set. Learn more

Project details

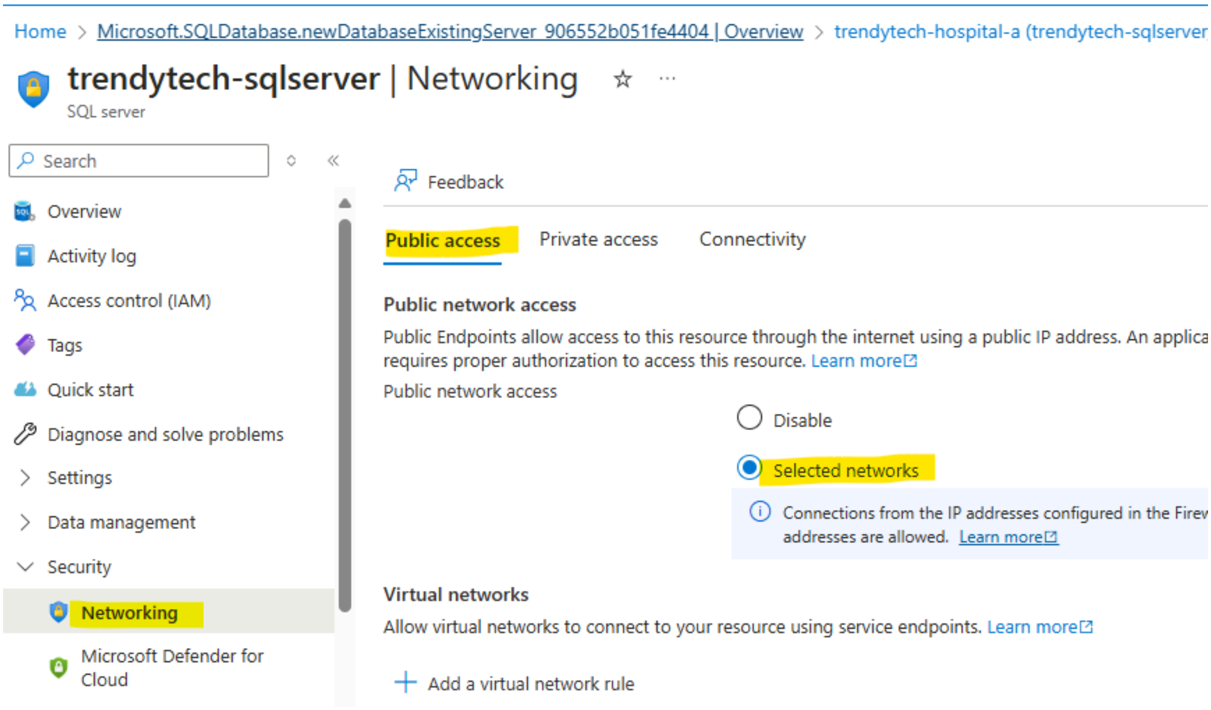
Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Cost summary

General Purpose (GP_S_Gen5_2)	
Cost per GB (in INR)	0.00
Max storage selected (in GB)	x 41.6
First 32 GB storage free	
First 100,000 vCore seconds free	
Overage billing 1	Disabled

Note: While creating database if you are not able to allow public access and add client ip address you can follow below steps:

After creating this database Go to Networking => For Public access (select option Selected networks and save this) =>



Home > Microsoft.SQLDatabase.newDatabaseExistingServer_906552b051fe4404 | Overview > trendytech-hospital-a (trendytech-sqlserver

trendytech-sqlserver | Networking ☆ ...

SQL server

Public network access

Public Endpoints allow access to this resource through the internet using a public IP address. An applica requires proper authorization to access this resource. Learn more

Public network access

○ Disable

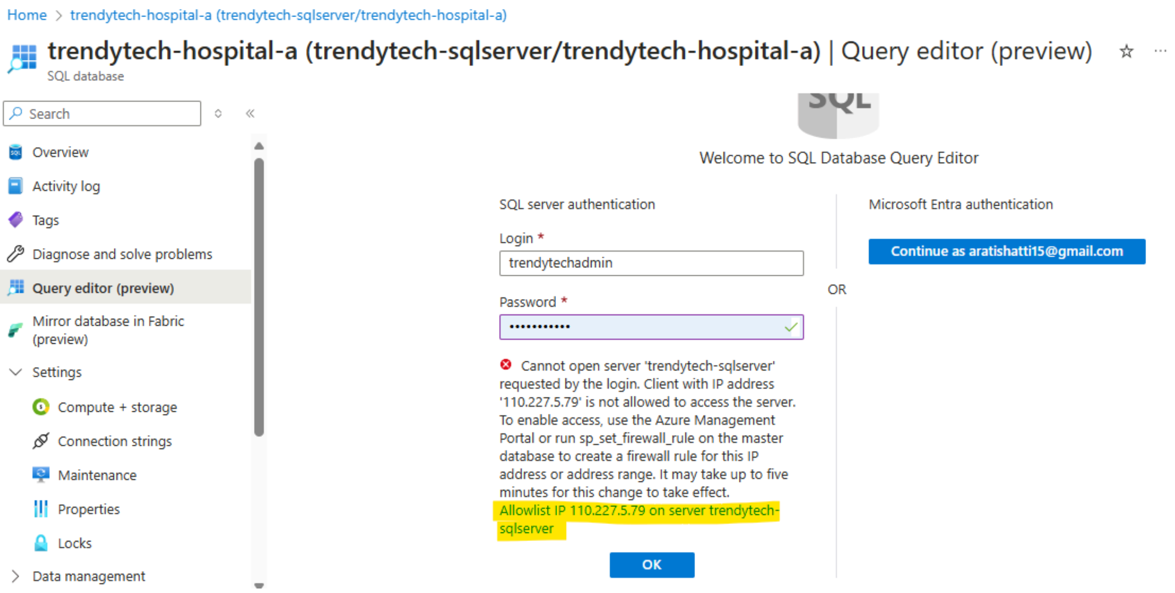
● Selected networks

Connections from the IP addresses configured in the Firev addresses are allowed. Learn more

Virtual networks

Allow virtual networks to connect to your resource using service endpoints. Learn more

Also while using query editor if you face below error click on “Allowing IP for current ip address” as shown below

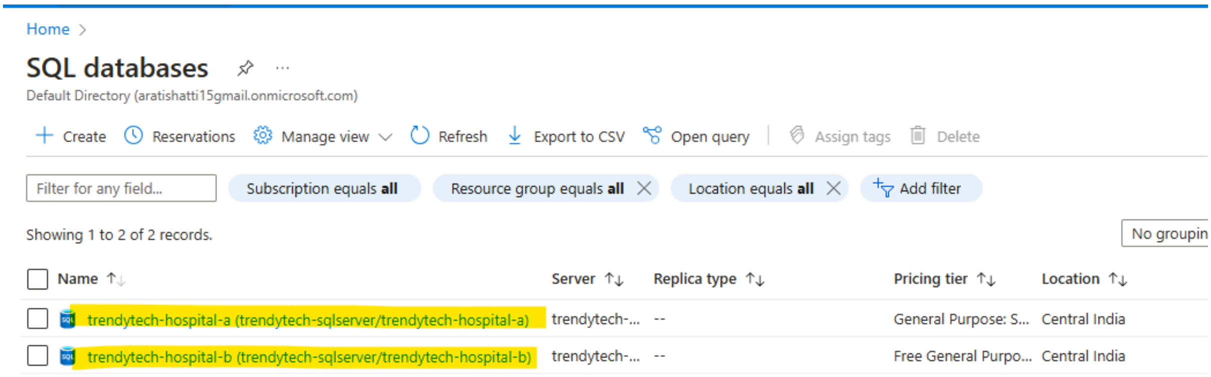

Home > trendytech-hospital-a (trendytech-sqlserver/trendytech-hospital-a)
trendytech-hospital-a (trendytech-sqlserver/trendytech-hospital-a) | Query editor (preview)
SQL database
Search
Overview
Activity log
Tags
Diagnose and solve problems
Query editor (preview)
Mirror database in Fabric (preview)
Settings
Compute + storage
Connection strings
Maintenance
Properties
Locks
Data management
SQL
Welcome to SQL Database Query Editor
SQL server authentication
Microsoft Entra authentication
Continue as aratishatti15@gmail.com
OR
OK

Login *
trendytechadmin

Password *
•••••••••••

Cannot open server 'trendytech-sqlserver' requested by the login. Client with IP address '110.227.5.79' is not allowed to access the server. To enable access, use the Azure Management Portal or run sp_set_firewall_rule on the master database to create a firewall rule for this IP address or address range. It may take up to five minutes for this change to take effect.
Allowlist IP 110.227.5.79 on server trendytech-sqlserver

Similarly we will create another database trendytech-hospital-b(We will use the same server i.e trendytech-sqlserver that we have created while creating trendytech-hospital-a database). Thus we have created 2 databases as shown below.



Showing 1 to 2 of 2 records.

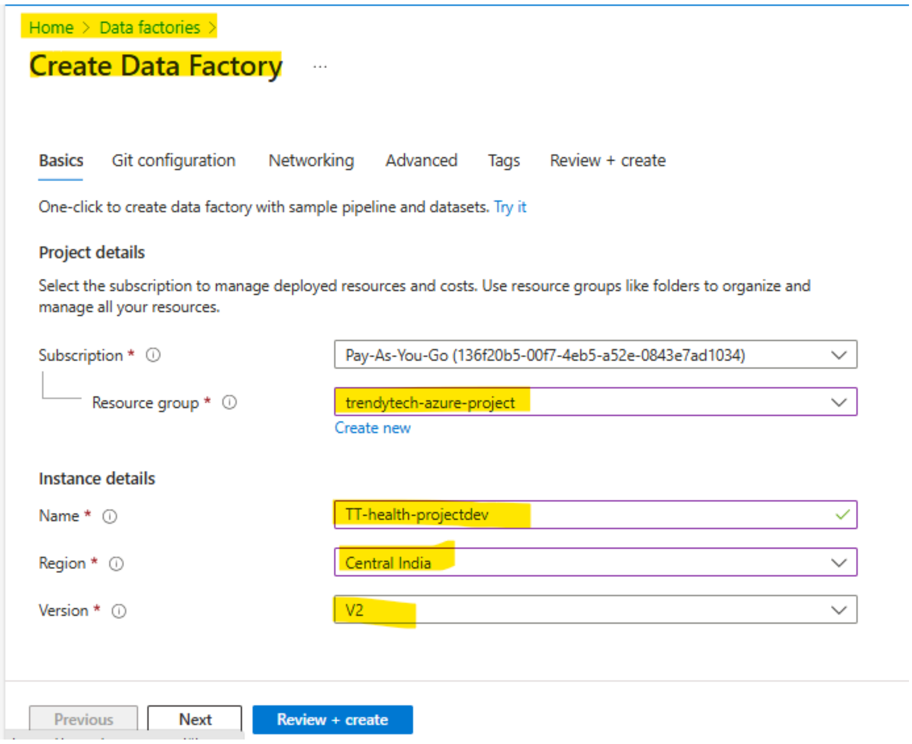
Name	Server	Replica type	Pricing tier	Location
trendytech-hospital-a (trendytech-sqlserver/trendytech-hospital-a)	trendytech-...	--	General Purpose: S...	Central India
trendytech-hospital-b (trendytech-sqlserver/trendytech-hospital-b)	trendytech-...	--	Free General Purpo...	Central India

Then we will create the tables in these databases and for creating tables in the database use below scripts which are present on github account:
For trendytech-hospital-a =>
Trendytech_hospital_A_table_creation_commands

For trendytech-hospital-b =>
Trendytech_hospital_B_table_creation_commands

Steps to create ADF:

In the search bar, type "Data Factory" and select "Data Factory" from the results. Click the "Create" button on the Data Factory page. Provide a globally unique name for your Data Factory instance. Choose V2 (Data Factory Version 2) for the latest features.



Create Data Factory ...

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-azure-project

Create new

Instance details

Name * : TT-health-projectdev

Region * : Central India

Version * : V2

Click "Review + Create" to validate the details. If validation passes, click "Create" to deploy the Data Factory.

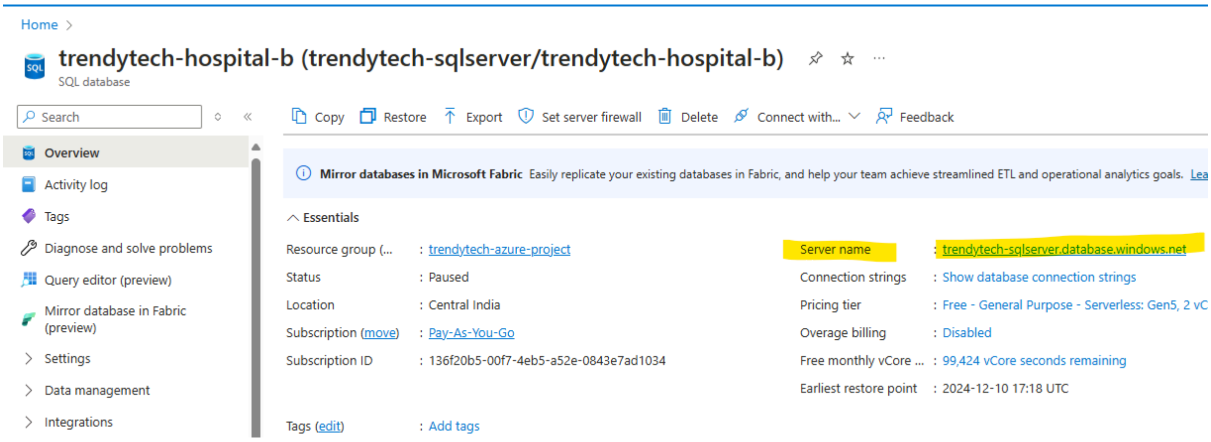
Steps to create ADF pipeline:

Linked Services creation:

In the ADF interface, go to the Manage section on the left-hand panel. Under the Connections section, select Linked Services. Click on New to create a new Linked Service.

1. Azure SQL DB

Note down the server name for that sql database that we have created.



Home >

trendytech-hospital-b (trendytech-sqlserver/trendytech-hospital-b)

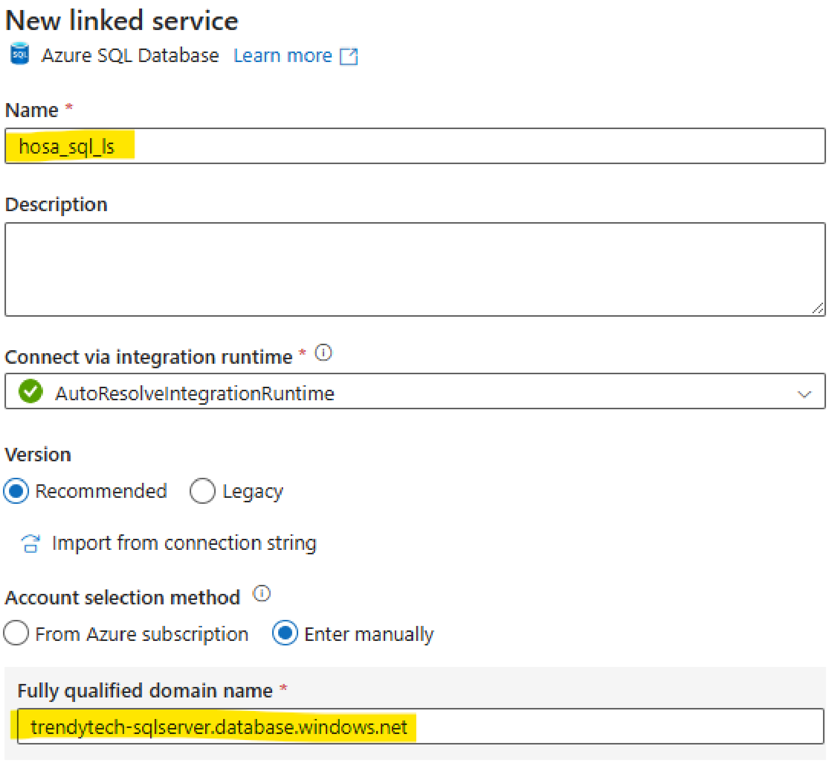
SQL database

Mirror databases in Microsoft Fabric Easily replicate your existing databases in Fabric, and help your team achieve streamlined ETL and operational analytics goals. Lea

Essentials

Resource group (...	: trendytech-azure-project	Server name	: trendytech-sqlserver.database.windows.net
Status	: Paused	Connection strings	: Show database connection strings
Location	: Central India	Pricing tier	: Free - General Purpose - Serverless: Gen5, 2 vC
Subscription (move)	: Pay-As-You-Go	Overage billing	: Disabled
Subscription ID	: 136f20b5-00f7-4eb5-a52e-0843e7ad1034	Free monthly vCore ...	: 99,424 vCore seconds remaining
		Earliest restore point	: 2024-12-10 17:18 UTC
Tags (edit)	: Add tags		

In fully qualified domain names, mention the server name, mention the username and password for sql server and define the parameter db_name and using this parameter we will pass the database name as shown below.



New linked service

Azure SQL Database Learn more

Name *

hosa_sql_ls

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Version

● Recommended ○ Legacy

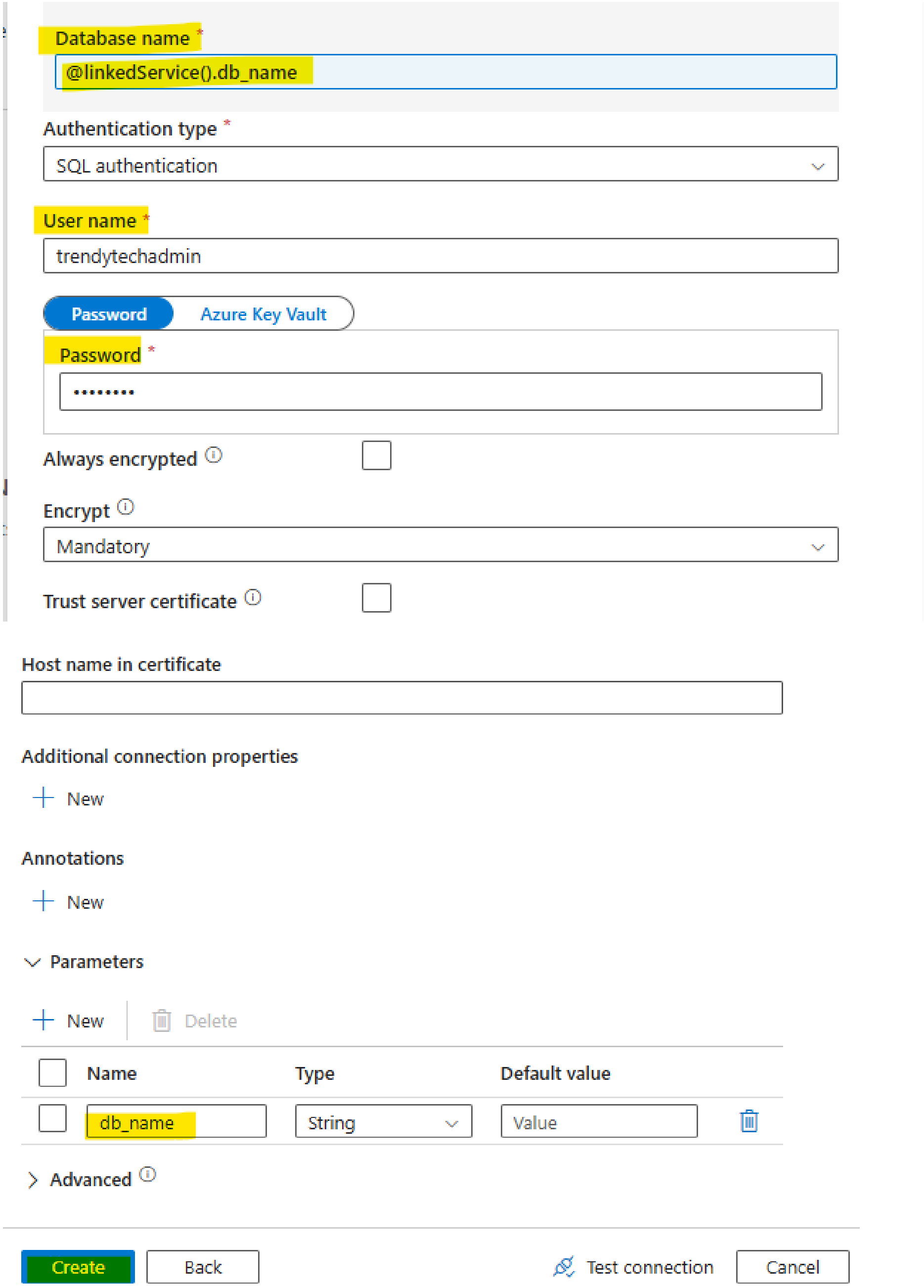
Import from connection string

Account selection method

○ From Azure subscription ● Enter manually

Fully qualified domain name *

trendytech-sqlserver.database.windows.net



Database name *

@linkedService().db_name

Authentication type *

SQL authentication

User name *

trendytechadmin

Password | Azure Key Vault

Password *

••••••••

Always encrypted ⓘ □

Encrypt ⓘ

Mandatory

Trust server certificate ⓘ □

Host name in certificate

Additional connection properties

Annotations

Parameters

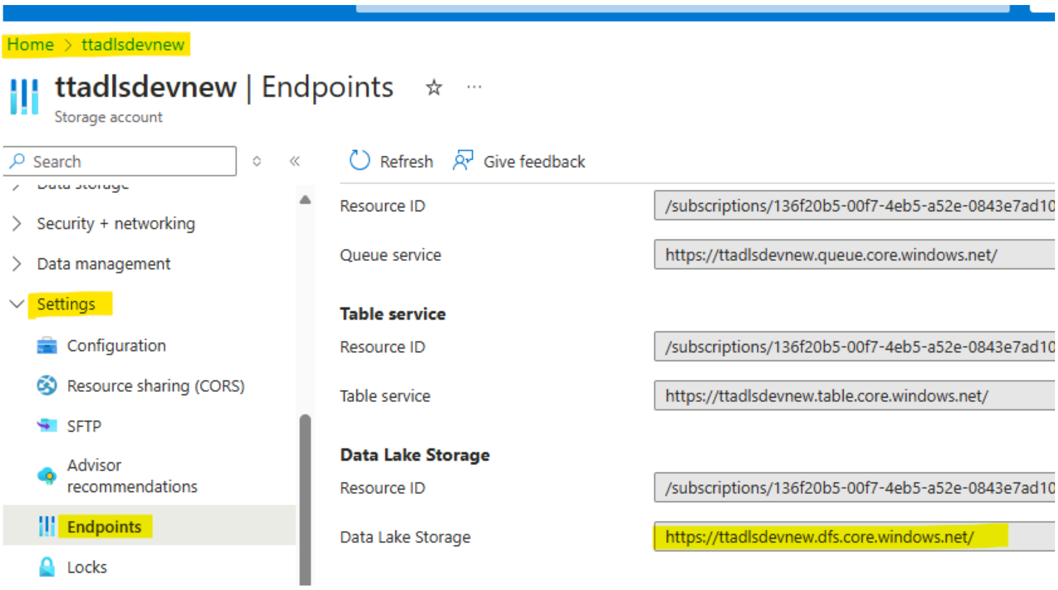
□	Name	Type	Default value
□	db_name	String	Value

And click on create to create the linked service.

2. ADLS GEN2

Select Azure Data Lake Storage1 as the data store. Provide the following details- Name of your Blob Storage account, Authentication. Then click Test Connection to verify and save the Linked Service.

To get the url for Azure Data Lake Storage go to Adls gen2 storage that we have created => Setting => Endpoints => and copy the URL as shown below



Resource ID: /subscriptions/136f20b5-00f7-4eb5-a52e-0843e7ad10

Queue service: https://ttadlsdevnew.queue.core.windows.net/

Table service

Resource ID: /subscriptions/136f20b5-00f7-4eb5-a52e-0843e7ad10

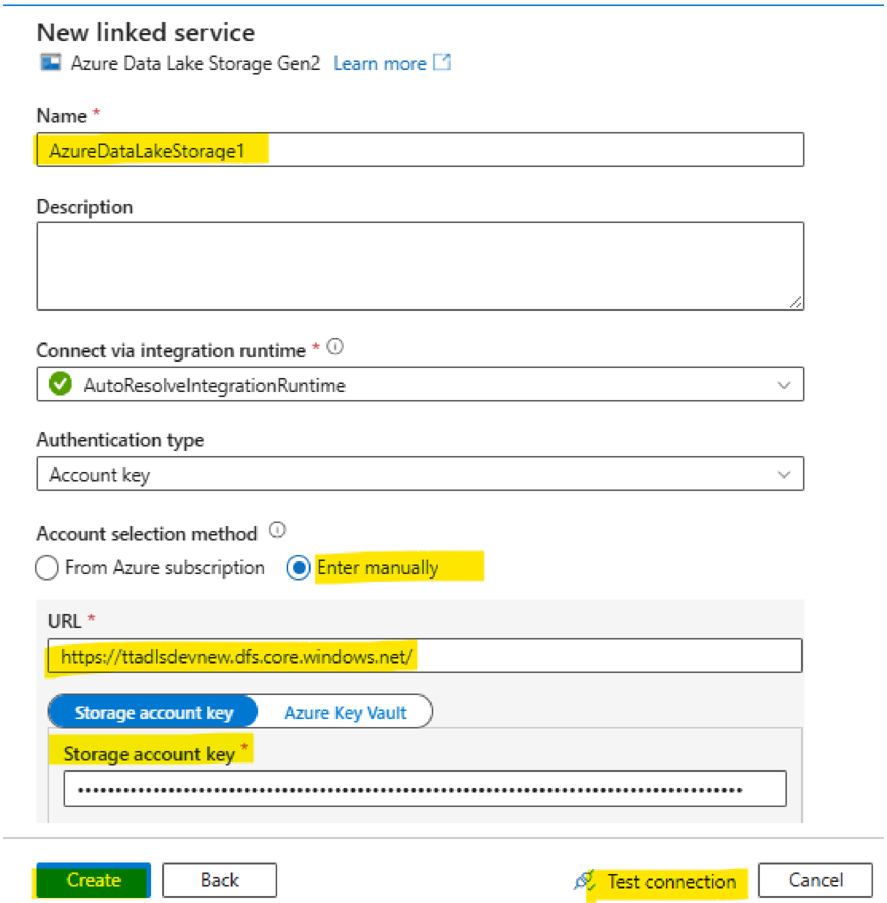
Table service: https://ttadlsdevnew.table.core.windows.net/

Data Lake Storage

Resource ID: /subscriptions/136f20b5-00f7-4eb5-a52e-0843e7ad10

Data Lake Storage: https://ttadlsdevnew.dfs.core.windows.net/

Also copy the access key. Using these details create the linked service as shown below.



Name *: AzureDataLakeStorage1

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Authentication type: Account key

Account selection method: ○ From Azure subscription ● Enter manually

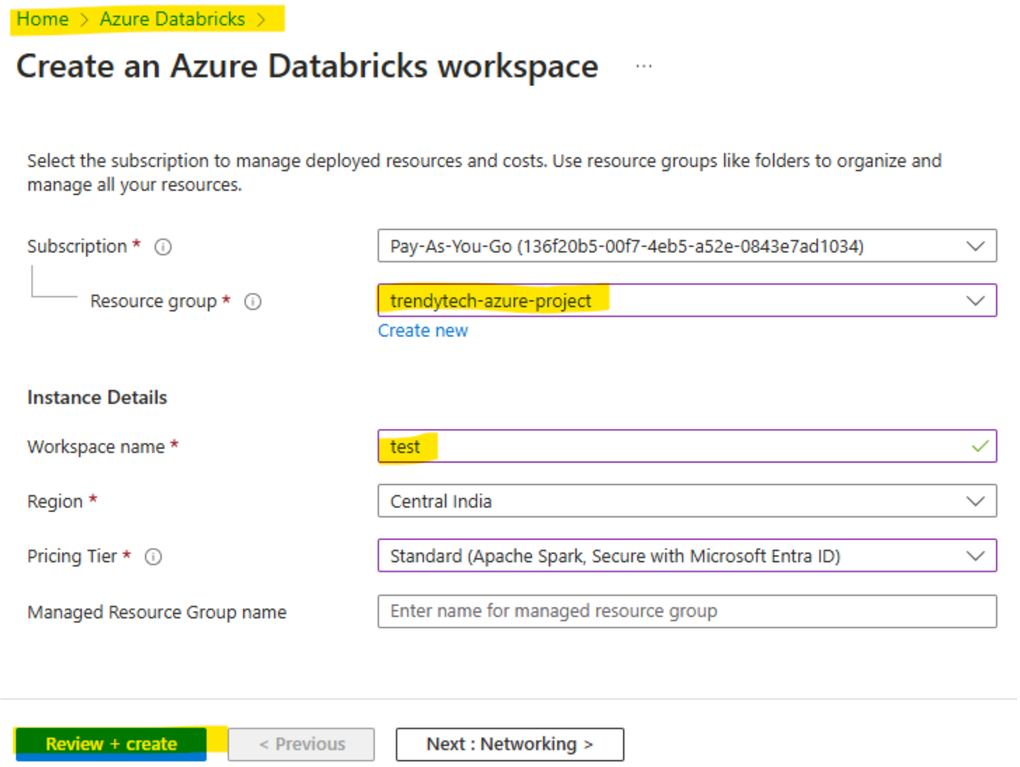
URL *: https://ttadlsdevnew.dfs.core.windows.net/

Storage account key *: ••

3. Delta table - Audit_logs

Create the databricks workspace test and then upload the code notebook "Audit_table_DDL" and start your databricks cluster and create the schema audit table in it using below commands. (Notebook name - audit_table_ddl)

```
create schema if not exists audit;

CREATE TABLE IF NOT EXISTS audit.load_logs (
    id BIGINT GENERATED ALWAYS AS IDENTITY,
    data_source STRING,
    tablename STRING,
    numberofrowscopied INT,
    watermarkcolumnname STRING,
    loaddate TIMESTAMP
);
```



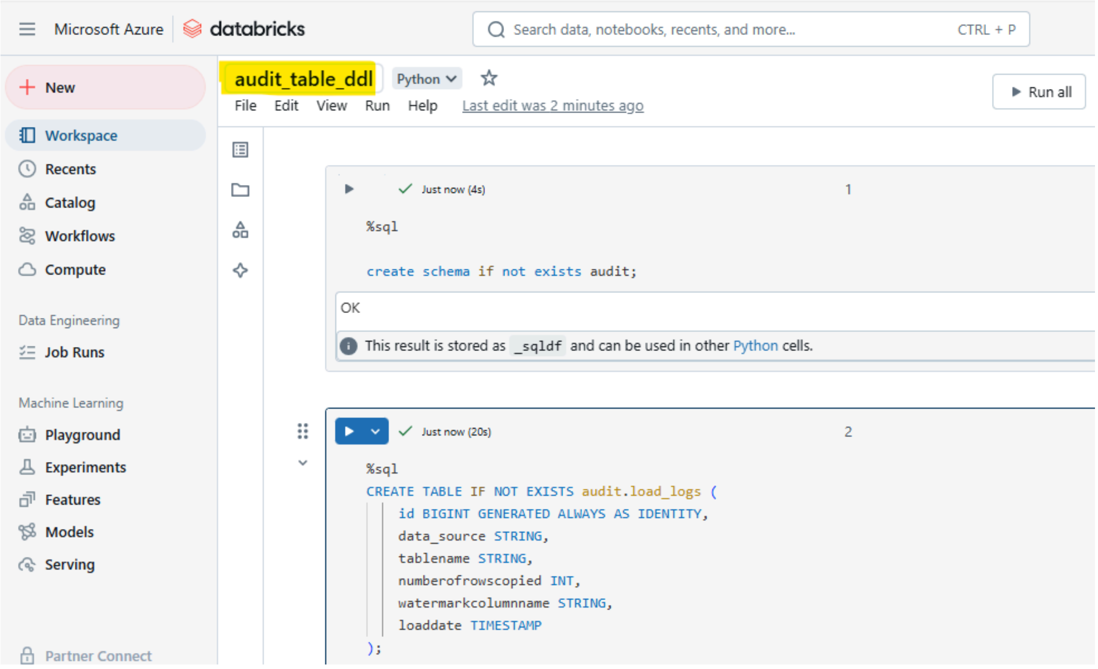
Create an Azure Databricks workspace ...

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-azure-project Create new

Instance Details

Workspace name *	test
Region *	Central India
Pricing Tier *	Standard (Apache Spark, Secure with Microsoft Entra ID)
Managed Resource Group name	Enter name for managed resource group



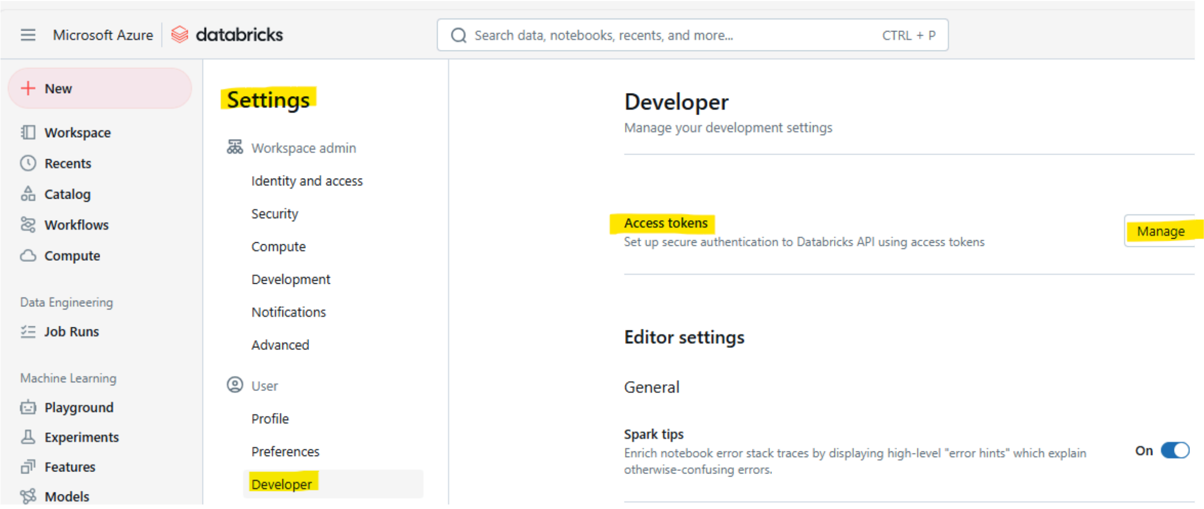
```
%sql

create schema if not exists audit;
```

```
OK
```

This result is stored as `_sqldf` and can be used in other Python cells.

```
%sql
CREATE TABLE IF NOT EXISTS audit.load_logs (
    id BIGINT GENERATED ALWAYS AS IDENTITY,
    data_source STRING,
    tablename STRING,
    numberofrowscopied INT,
    watermarkcolumnname STRING,
    loaddate TIMESTAMP
);
```

Note: To get access token Click your profile icon (top-right corner of the workspace) => Select Settings => Developer (In user setting) => Generate Access Token as shown below



Developer

Manage your development settings

Access tokens

Set up secure authentication to Databricks API using access tokens

Editor settings

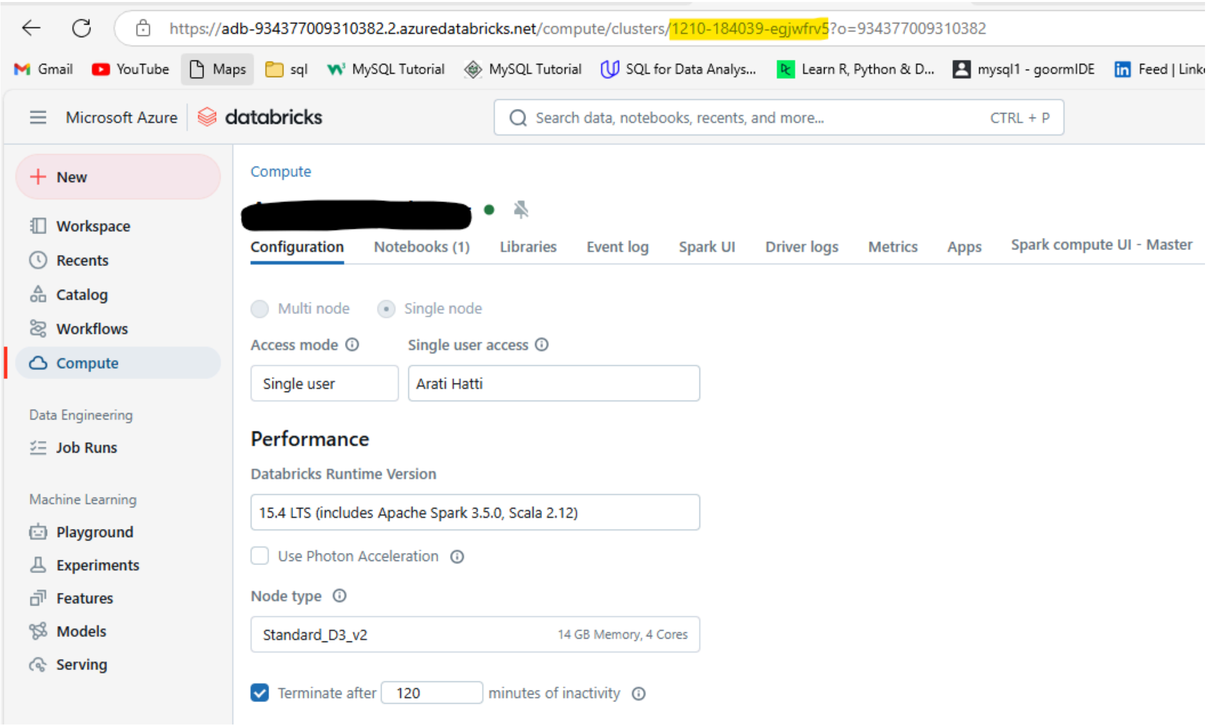
General

Spark tips

Enrich notebook error stack traces by displaying high-level "error hints" which explain otherwise-confusing errors.

Then generate the token and copy for Future use.

While creating linked service in source mention “AzureDatabricksDeltaLake”. Then in the domain mention the URL of databricks workspace. And mention the cluster id for the cluster that we have created. To get Datbricks overview page to get Workspace URL and to get cluster id go to compute => Select the cluster => copy the cluster id as shown below



[redacted]

○ Multi node ● Single node

Access mode: Single user

Single user access: Arati Hatti

Performance

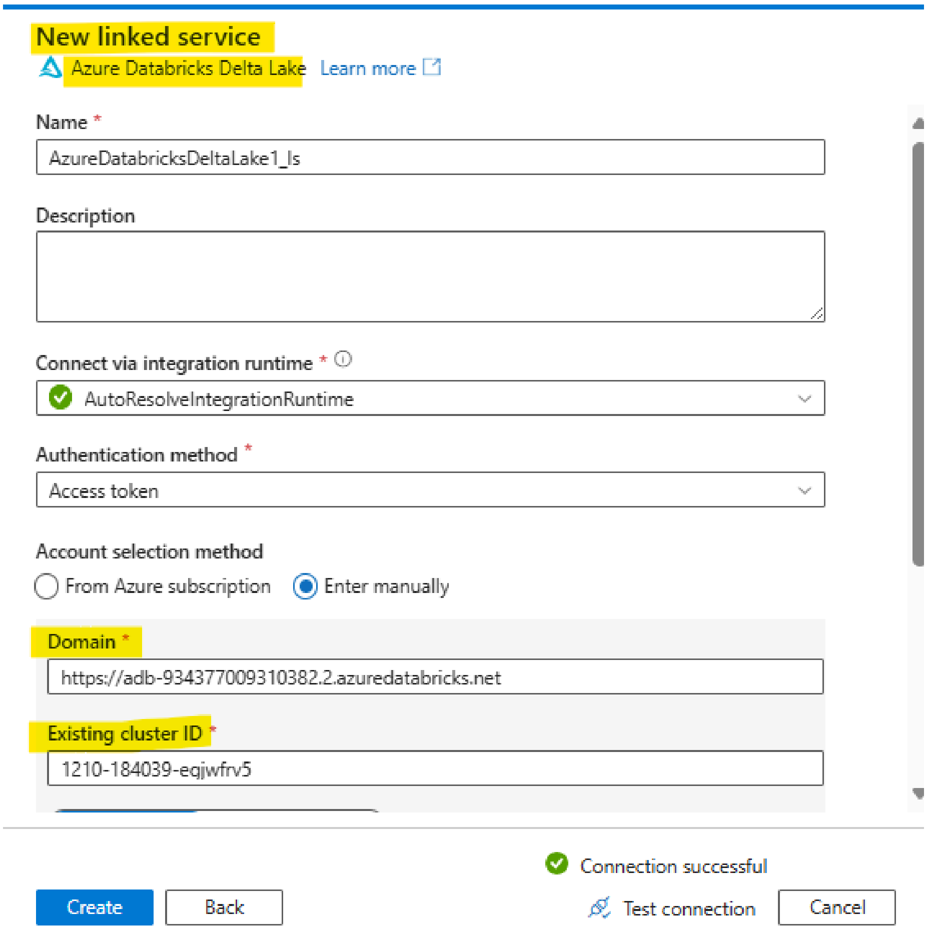
Databricks Runtime Version: 15.4 LTS (includes Apache Spark 3.5.0, Scala 2.12)

□ Use Photon Acceleration

Node type: Standard_D3_v2 14 GB Memory, 4 Cores

☑ Terminate after 120 minutes of inactivity

Refer below screenshot for more details.


New linked service

Azure Databricks Delta Lake

Name *: AzureDatabricksDeltaLake1_ls

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Authentication method *: Access token

Account selection method: ○ From Azure subscription ● Enter manually

Domain *: https://adb-934377009310382.2.azuredatabricks.net

Existing cluster ID *: 1210-184039-eqjwfrv5

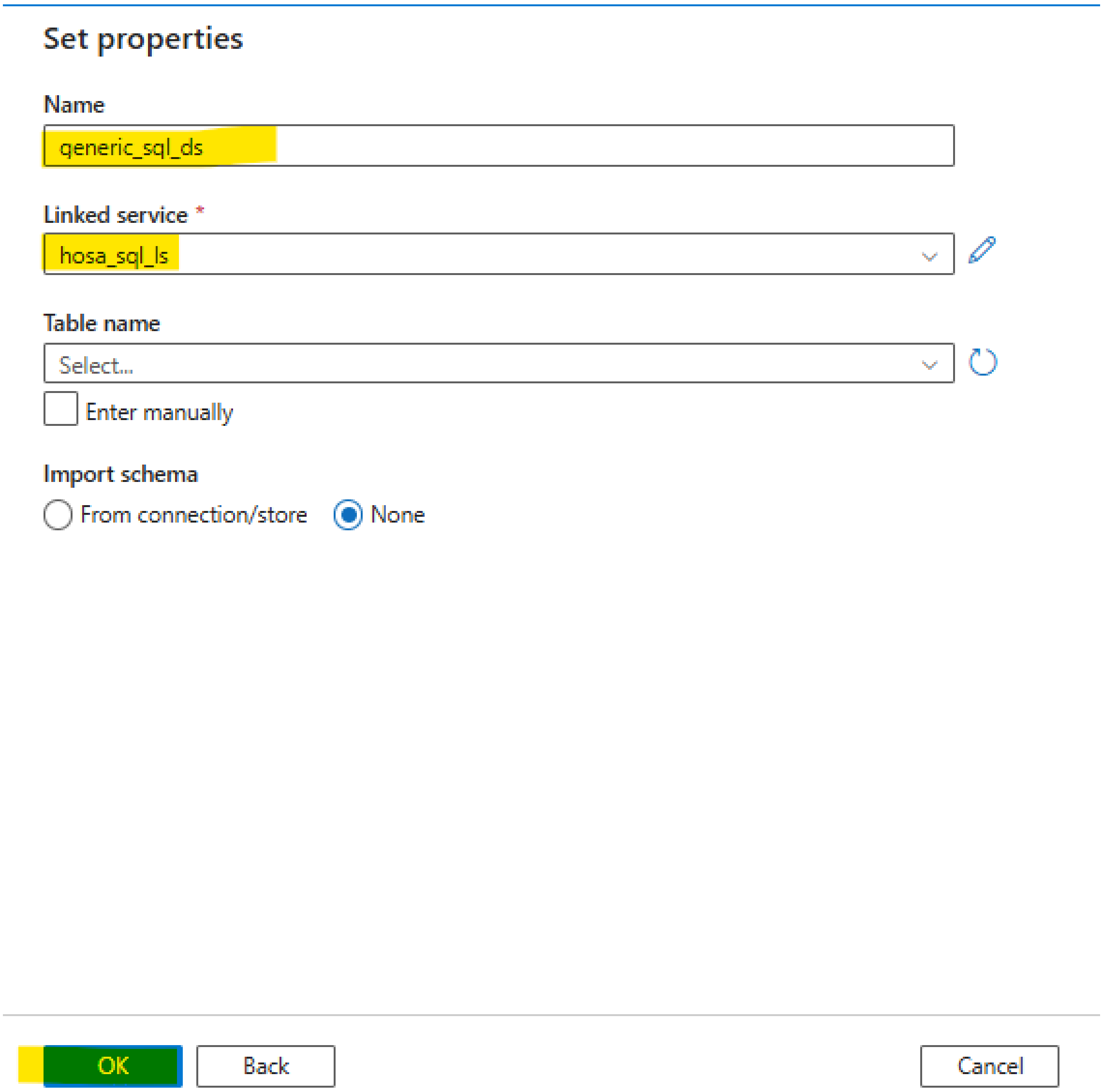
Connection successful

Dataset creation:

In the ADF interface, click on the Author section (left-hand panel). Expand the Datasets option. Click on the "…" next to Datasets in order to create the dataset.

1. Azure SQL DB

We will select the linked service that we have created for the SQL database. To create the datasets for the tables in a parameterized way in the sql database , we will create the parameter db_name, schema_name, table_name.



Set properties

Name: generic_sql_ds

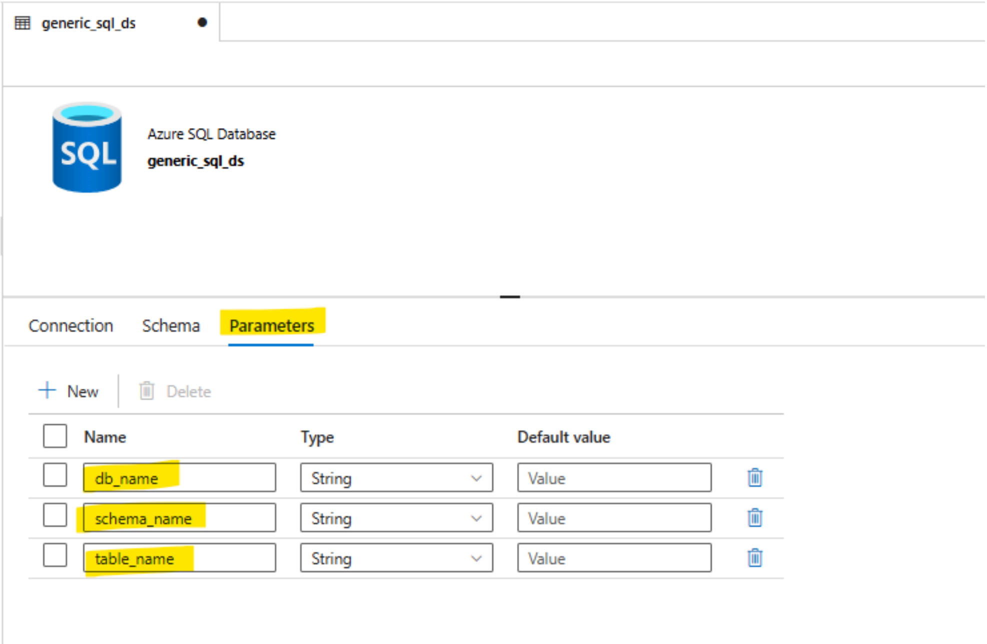
Linked service *: hosa_sql_ls

Table name: Select...

□ Enter manually

Import schema: ○ From connection/store ● None

Then we will create parameters db_name, schema_name, table_name as shown below.

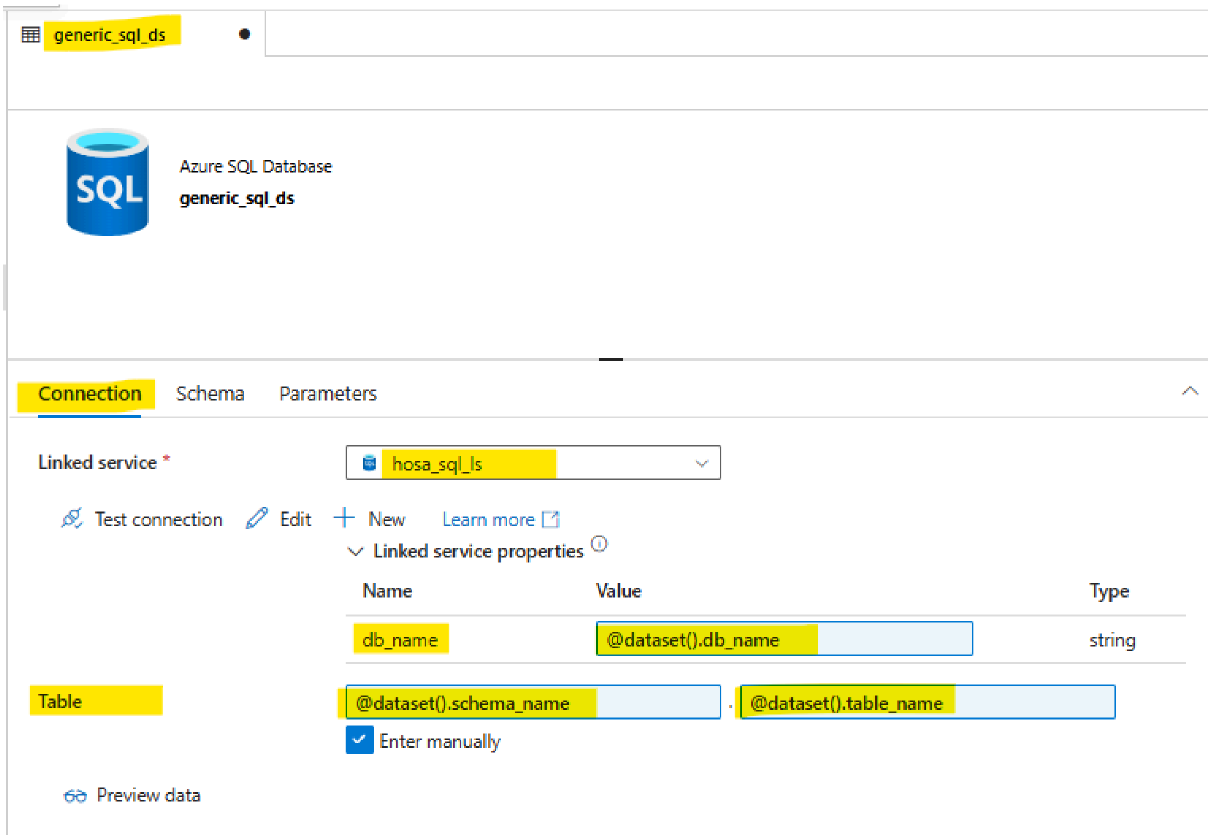


generic_sql_ds ●

Azure SQL Database
generic_sql_ds

□	Name	Type	Default value
□	db_name	String	Value
□	schema_name	String	Value
□	table_name	String	Value

And we will pass dynamic value for table name and schema name as shown below



generic_sql_ds ●

Azure SQL Database
generic_sql_ds

Linked service * : hosa_sql_ls

Linked service properties

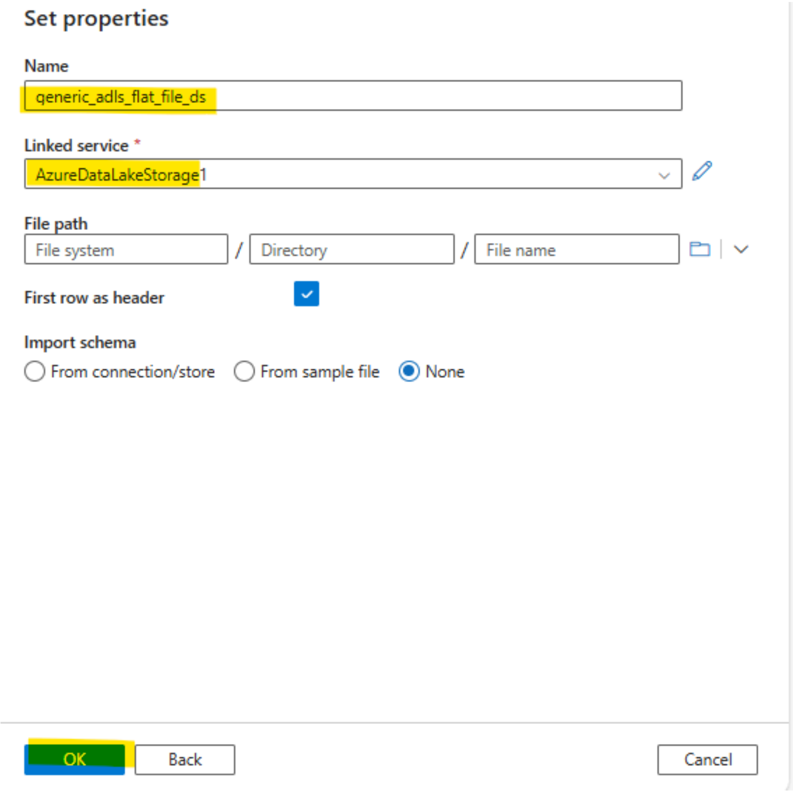
Name	Value	Type
db_name	@dataset().db_name	string

Table: @dataset().schema_name . @dataset().table_name

☑ Enter manually

2. Dataset for Flatfile in ADLS GEN2

Select source as ADLS gen2 and file format as delimited text. Also in order to make it generic we will create the parameter file_name, file_path and container.



Set properties

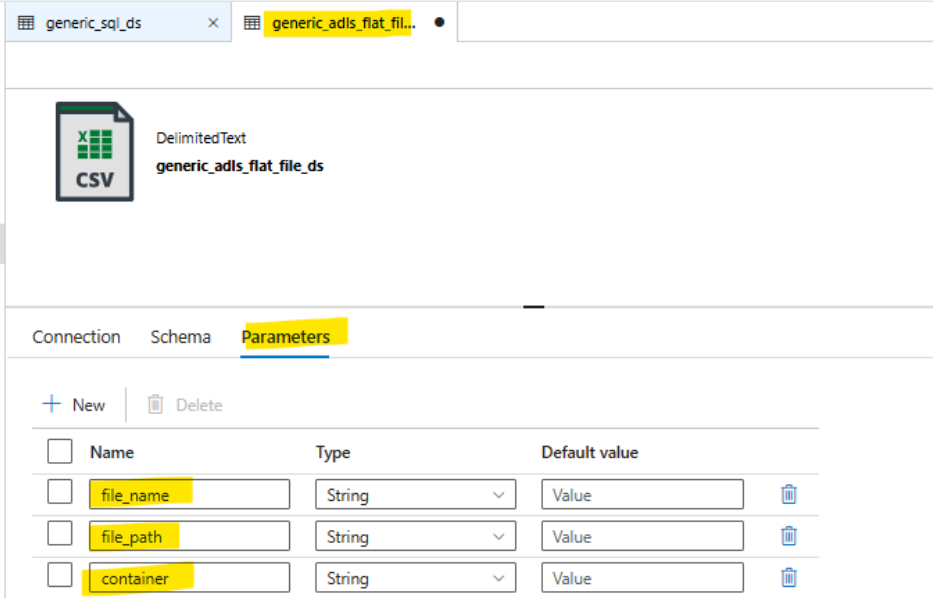
Name: generic_adls_flat_file_ds

Linked service *: AzureDataLakeStorage1

File path: File system / Directory / File name

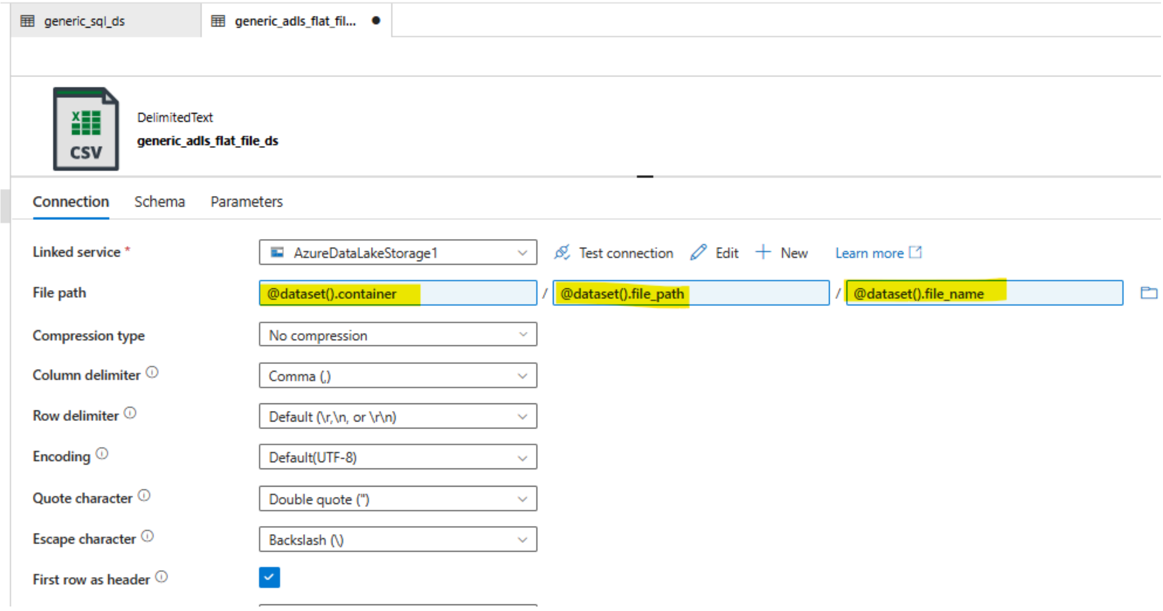
First row as header ☑

Import schema: ○ From connection/store ○ From sample file ● None



DelimitedText
generic_adls_flat_file_ds

	Name	Type	Default value
□	file_name	String	Value
□	file_path	String	Value
□	container	String	Value



DelimitedText
generic_adls_flat_file_ds

Linked service * AzureDataLakeStorage1

File path @dataset().container / @dataset().file_path / @dataset().file_name

Compression type No compression

Column delimiter Comma (,)

Row delimiter Default (\r,\n, or \r\n)

Encoding Default(UTF-8)

Quote character Double quote (")

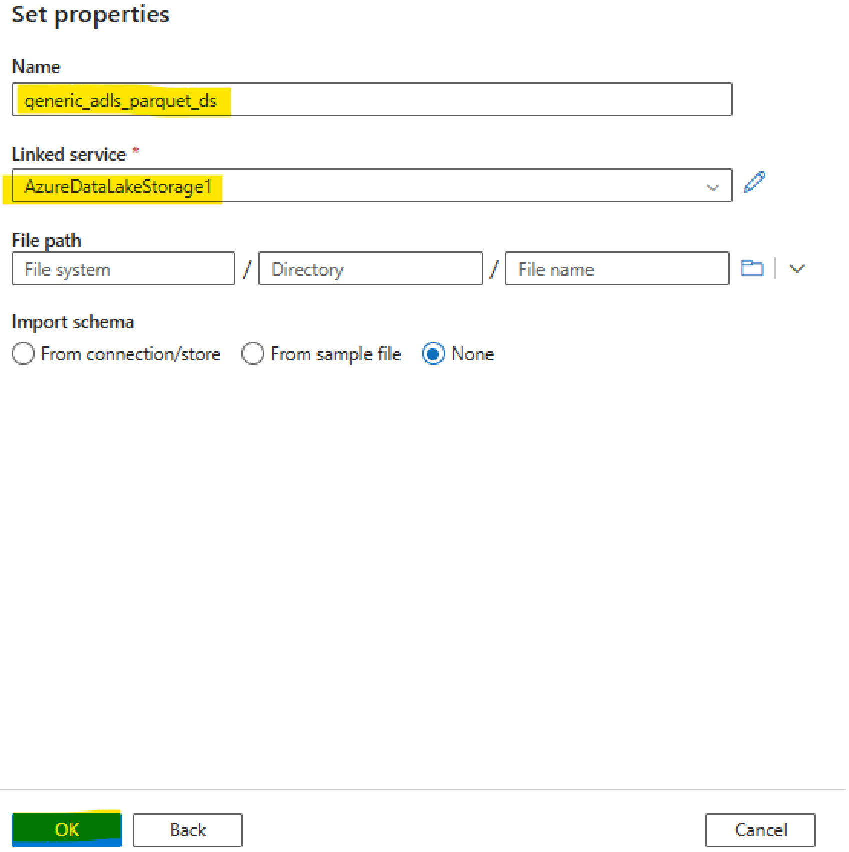
Escape character Backslash (\)

First row as header ☑

Now publish the changes.

3. Dataset for Parquet file in ADLS GEN2

In order to store data in ADLS gen2 in parquet format we will need the dataset.
While creating this dataset we will select source as ADLS gen2, fileformat as parquet and we will create parameters file_name, file_path and container.



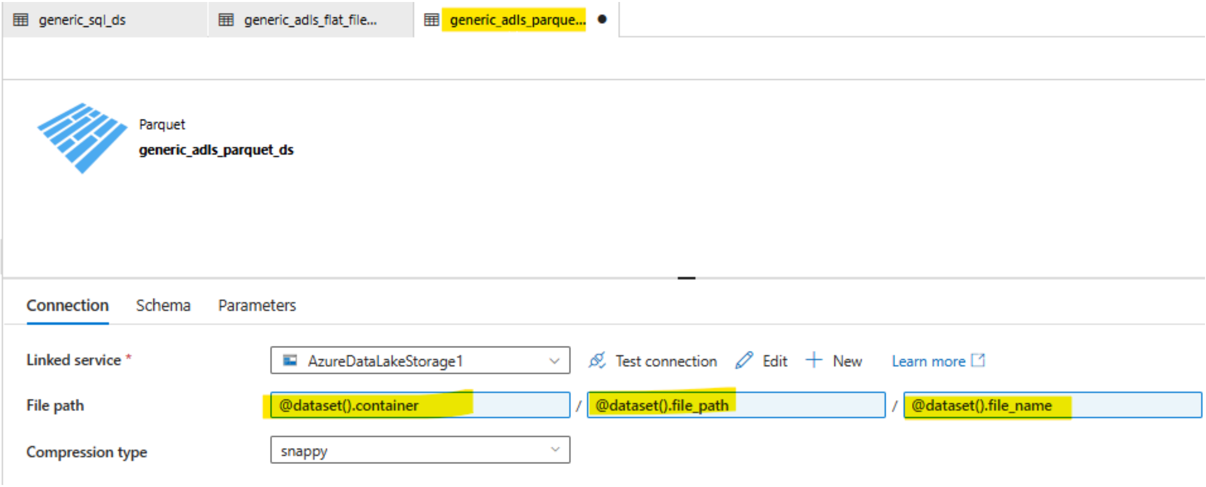
Set properties

Name
generic_adls_parquet_ds

Linked service *
AzureDataLakeStorage1

File path
File system / Directory / File name

Import schema
○ From connection/store ○ From sample file ● None



Parquet
generic_adls_parquet_ds

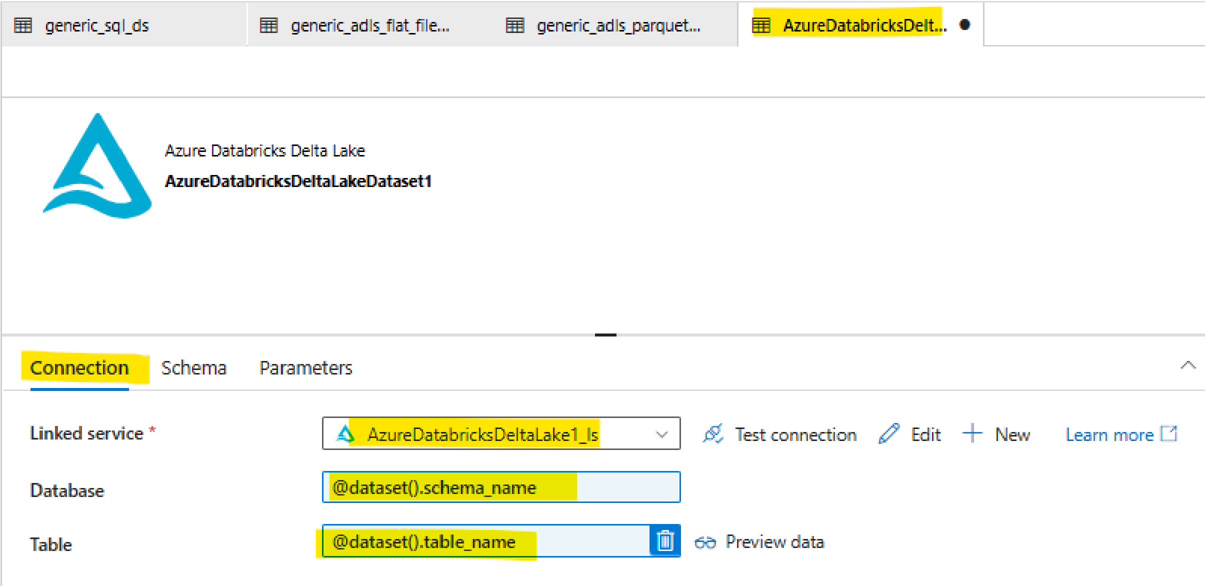
Linked service * AzureDataLakeStorage1

File path @dataset().container / @dataset().file_path / @dataset().file_name

Compression type snappy

4. Databricks Delta Lake for Delta lake

We will select the source as Azure Databricks Delta Lake. For this we will create the parameter schema_name and table_name.



Azure Databricks Delta Lake
AzureDatabricksDeltaLakeDataset1

Linked service * AzureDatabricksDeltaLake1_ls

Database @dataset().schema_name

Table @dataset().table_name

Once all the dataset and linked service are created, publish all in order to save them.

Creation of Pipelines:

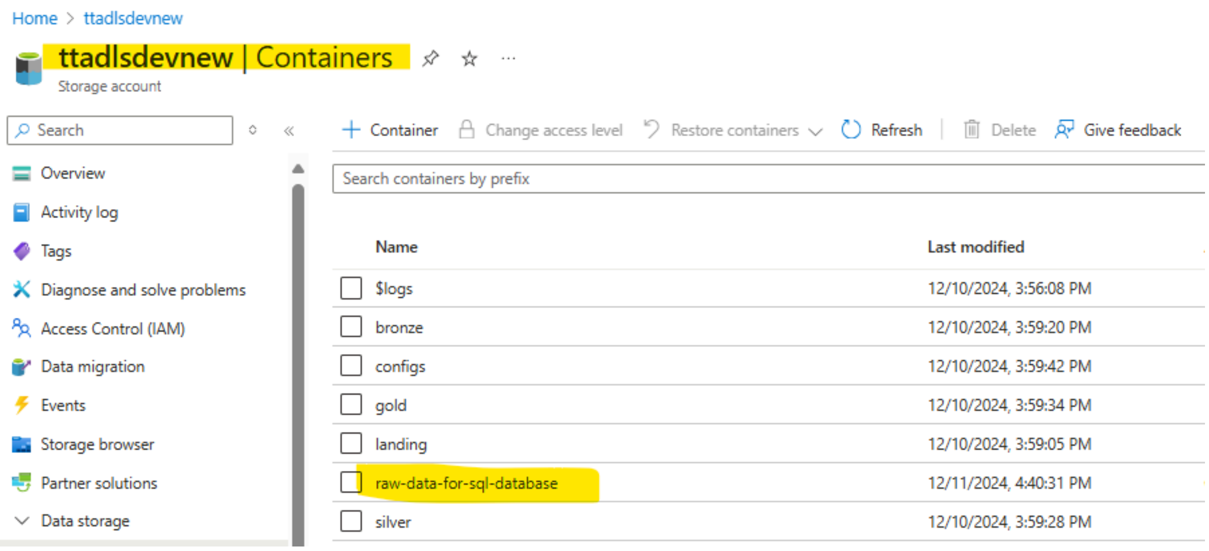
Background activity : Creation of pipeline to copy data into sql tables (pl_to_insert_data_to_sql_table_preprocessing).

Before proceeding with the main pipeline, we will create a simple pipeline in Azure Data Factory (ADF) to copy data from ADLS Gen2 storage into tables in an SQL database. This serves as a prerequisite to ensure that the SQL tables contain the data needed for the main pipeline.

Note: We will create a new container (raw-data-for-sql-database) in the given ADLS Gen2 storage (adlsdevnew) and upload our CSV files, which will serve as the source for the pipeline, along with a lookup file. Additionally, we will create a dataset for the lookup file to use in this pipeline. Using the copy activity in ADF, we will transfer the data into the following tables: Departments, Providers, Encounters, Patients, and Transactions, located in the SQL databases trendytech-hospital-a and trendytech-hospital-b.

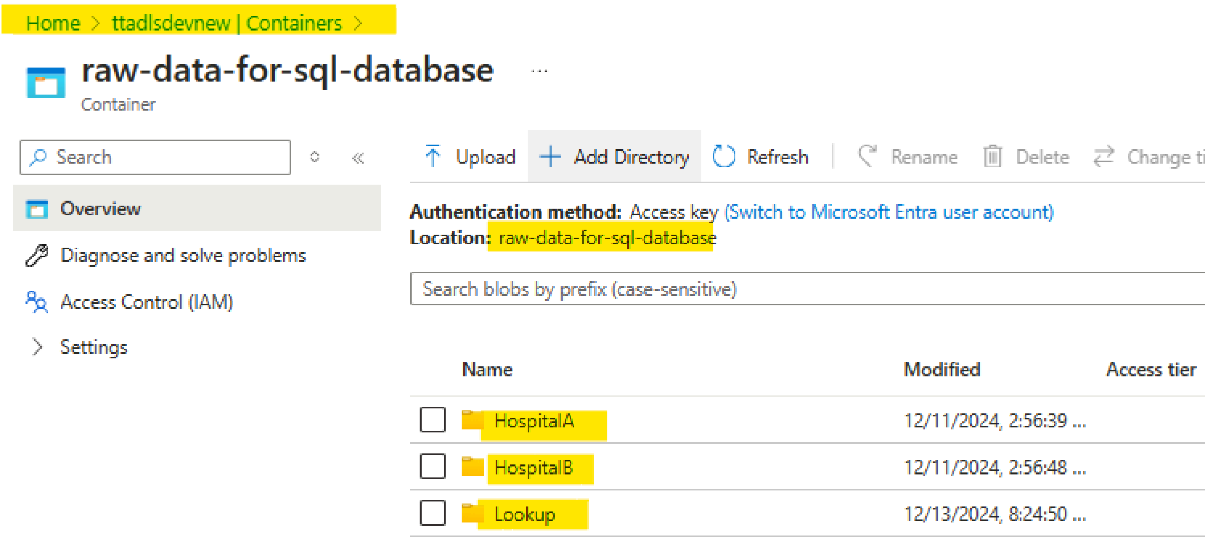
Source: ADLS gen2 -adlsdevnew

We will create a new container (raw-data-for-sql-database) in the given ADLS Gen2 storage (adlsdevnew) and upload our CSV files, along with a lookup file.



Name	Last modified
□ $logs	12/10/2024, 3:56:08 PM
□ bronze	12/10/2024, 3:59:20 PM
□ configs	12/10/2024, 3:59:42 PM
□ gold	12/10/2024, 3:59:34 PM
□ landing	12/10/2024, 3:59:05 PM
□ raw-data-for-sql-database	12/11/2024, 4:40:31 PM
□ silver	12/10/2024, 3:59:28 PM

Folder: HospitalA, HospitalB for datafiles, Lookup for lookup file

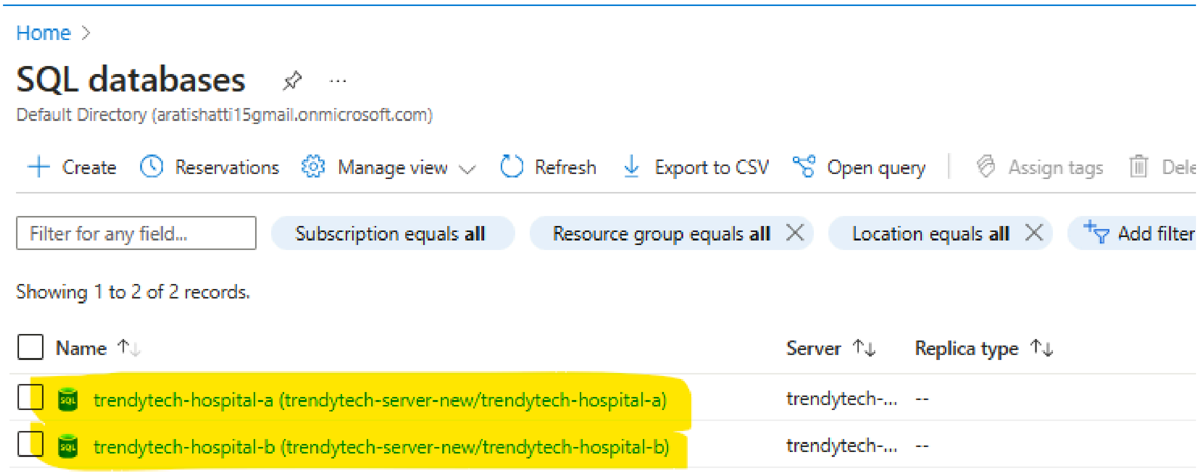

Home > ttadlsdevnew | Containers >
raw-data-for-sql-database ...
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Upload
Add Directory
Refresh
Rename
Delete
Change ti
Search blobs by prefix (case-sensitive)

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: raw-data-for-sql-database

	Name	Modified	Access tier
□	HospitalA	12/11/2024, 2:56:39 ...	
□	HospitalB	12/11/2024, 2:56:48 ...	
□	Lookup	12/13/2024, 8:24:50 ...	

Sink: SQL DB - trendytech-hospital-a, trendytech-hospital-b:

Note: We have already created these databases so no need to create again.



Showing 1 to 2 of 2 records.

□	Name ↑↓	Server ↑↓	Replica type ↑↓
□	trendytech-hospital-a (trendytech-server-new/trendytech-hospital-a)	trendytech-...	--
□	trendytech-hospital-b (trendytech-server-new/trendytech-hospital-b)	trendytech-...	--

Pipeline creation Steps:

1. **Creation of Linked Services:**
   => For ADLS gen2 storage(source):

We will use the same linked service “AzureDataLakeStorage1” that we have created earlier for ADLS Gen2 storage.

=> For SQL DB(Sink):

We will use the same linked service “hosa_sql_ls” that we have created earlier for the database.

2. Creation of Datasets:

=> For source:

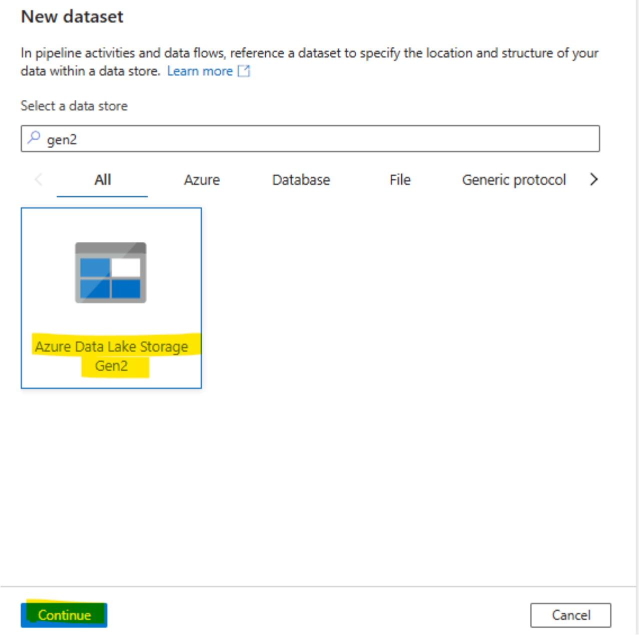
We will use the same generic dataset “generic_adls_flat_file_ds” that we have created earlier.

=> For sink:

We will use the same generic dataset “generic_sql_ds” that we have created earlier.

=>For Lookup we will create a new dataset as shown below.

Select source as ADLS Gen2 storage, then in file format select json as our lookup file is a json file as shown below

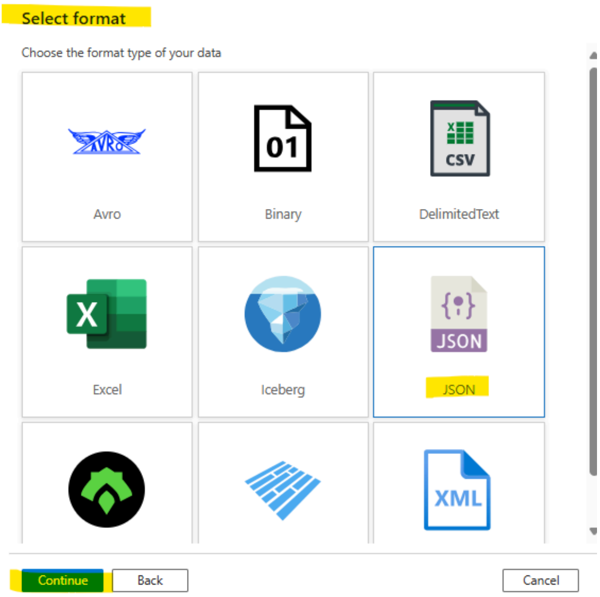


New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more

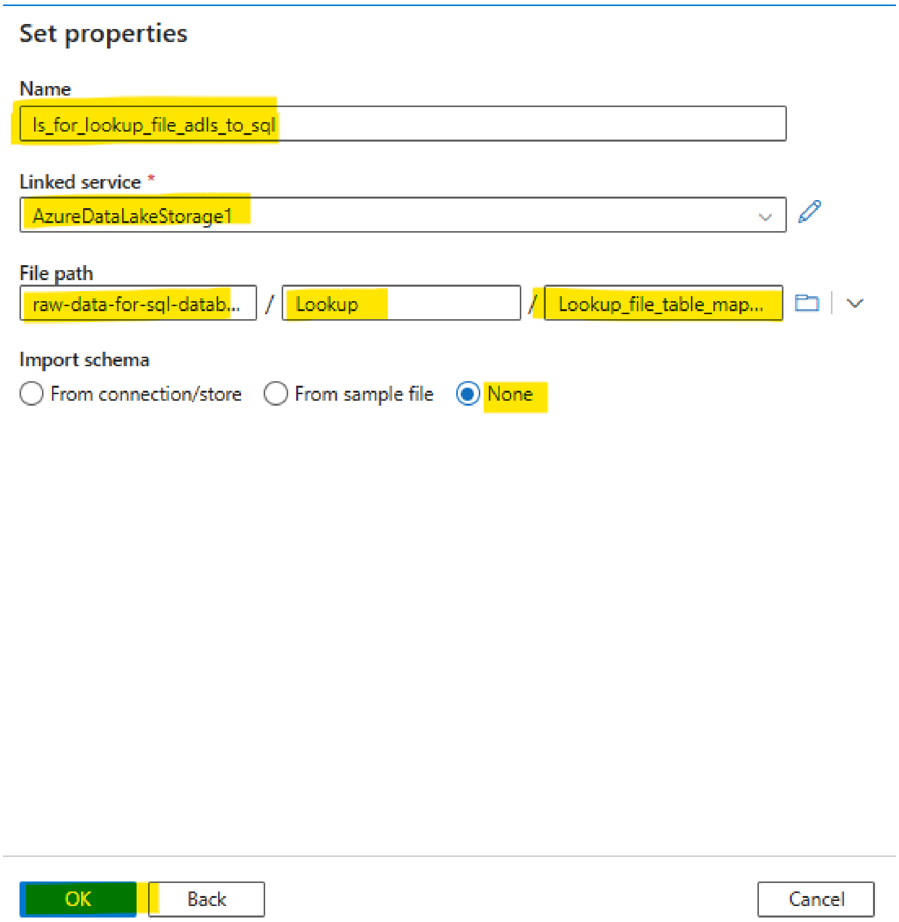
Select a data store

gen2



Select format

Choose the format type of your data



Set properties

Name

ls_for_lookup_file_adls_to_sql

Linked service *

AzureDataLakeStorage1

File path

raw-data-for-sql-datab... / Lookup / Lookup_file_table_map...

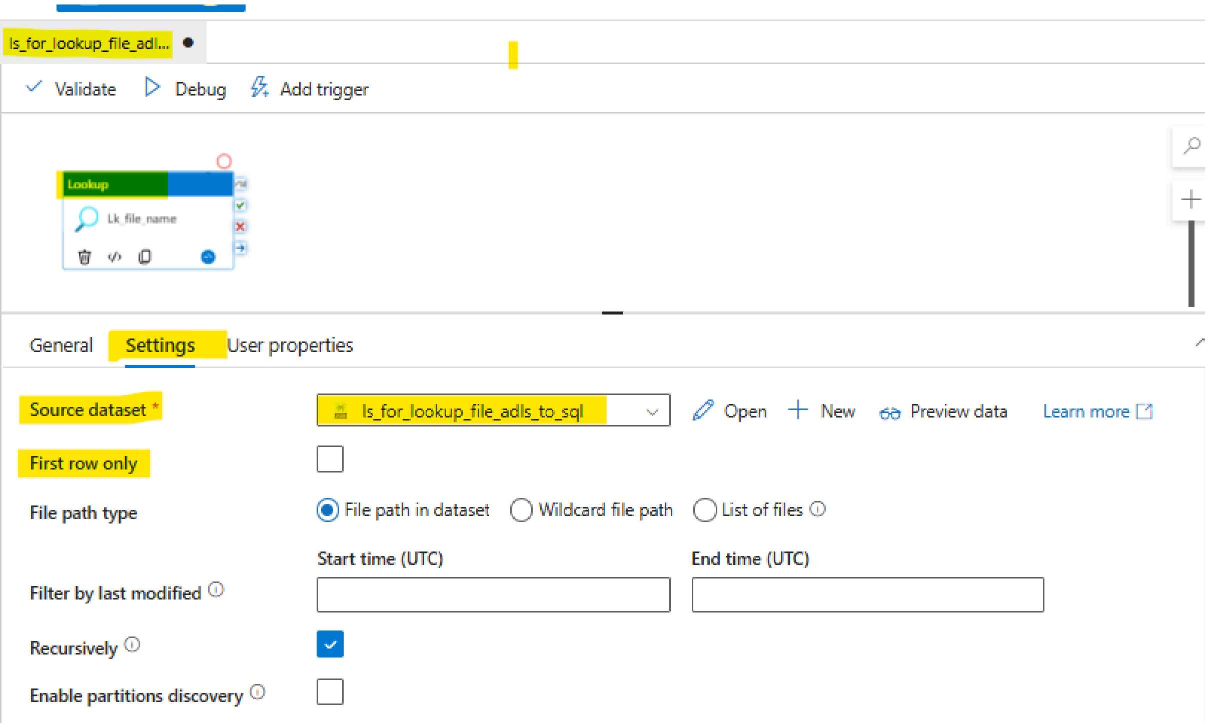
Import schema

○ From connection/store ○ From sample file ● None

Steps to Configure the Pipeline:

Add a Lookup Activity:

Drag a Lookup activity to the canvas. Point it to the mapping CSV dataset.Set First row only to false to read all rows. Refer below screenshot for more clarity.



Source dataset * ls_for_lookup_file_adls_to_sql

First row only □

File path type ● File path in dataset ○ Wildcard file path ○ List of files

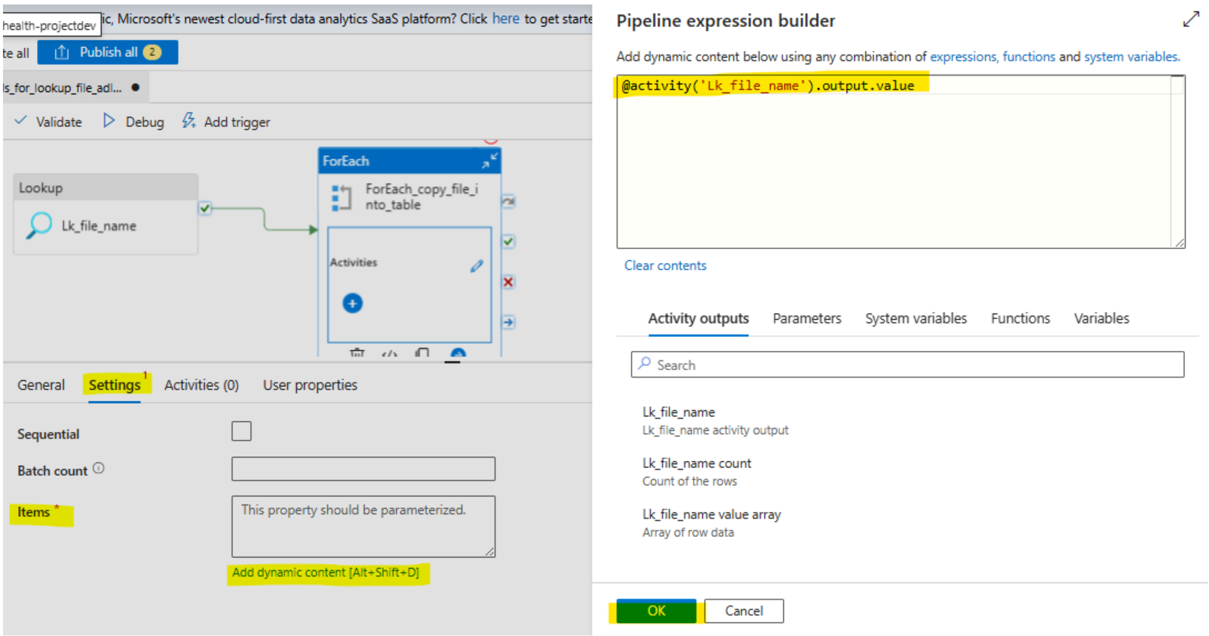
Filter by last modified — Start time (UTC) — End time (UTC)

Recursively ☑

Enable partitions discovery □

Add a ForEach Activity:

Drag a ForEach activity and connect it to the Lookup activity.
Set its Items property to @activity('Lk_file_name').output.value
Refer below screenshot for more clarity.



Sequential □

Batch count

Items * This property should be parameterized.

Add dynamic content [Alt+Shift+D]

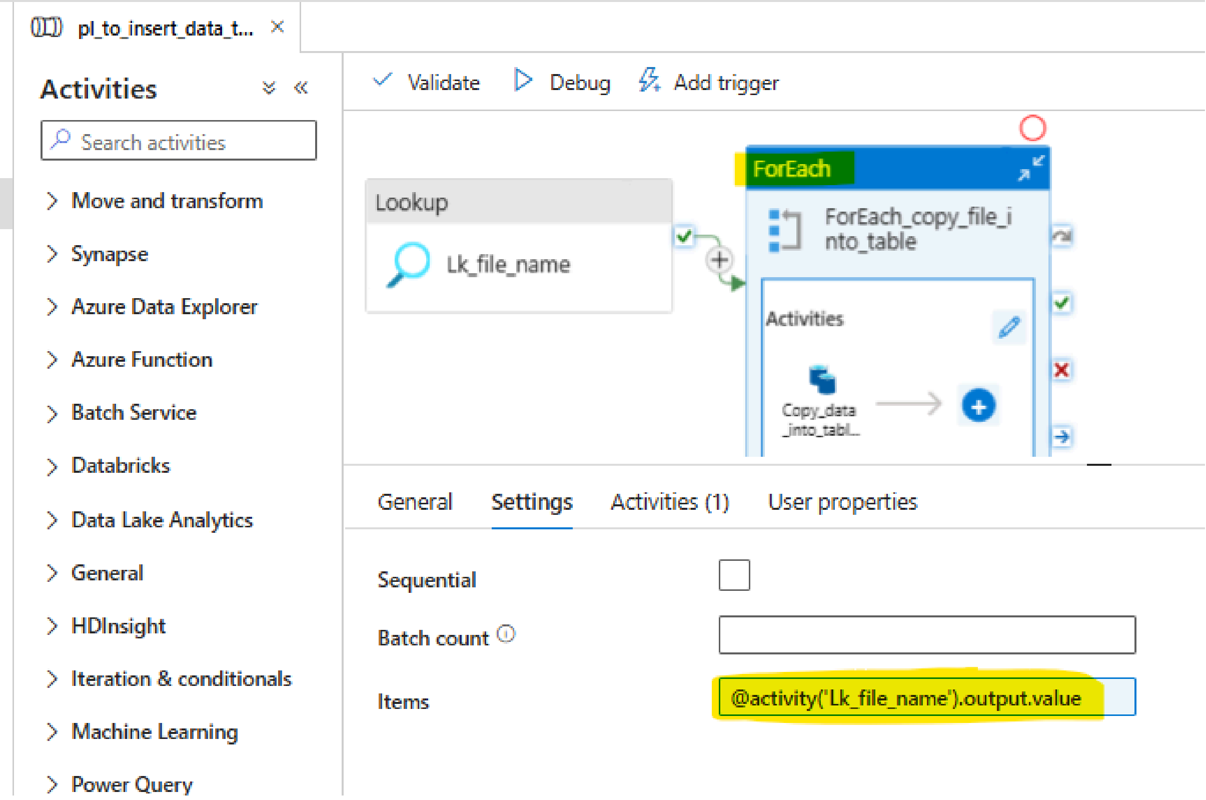
Add dynamic content below using any combination of expressions, functions and system variables.

```
@activity('Lk_file_name').output.value
```

Lk_file_name
Lk_file_name activity output

Lk_file_name count
Count of the rows

Lk_file_name value array
Array of row data



Sequential □

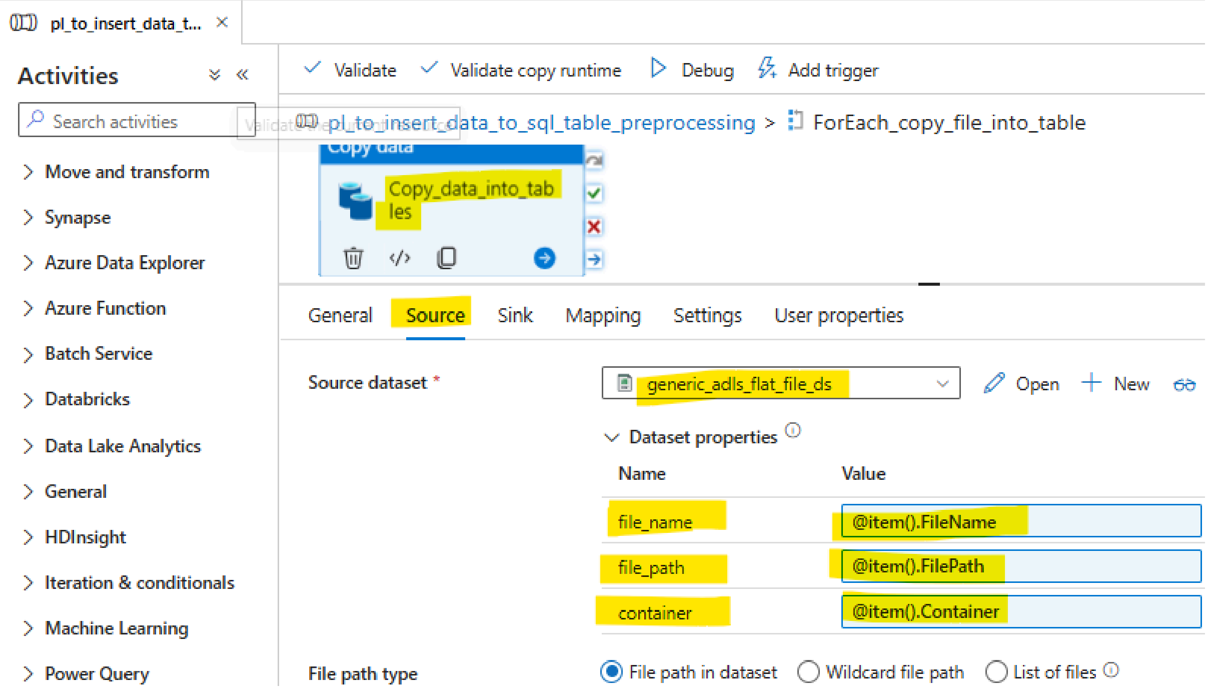
Batch count

Items @activity('Lk_file_name').output.value

Configure the ForEach Activity:

Inside the ForEach activity, add a Copy Data activity.

Source: Use the source dataset.



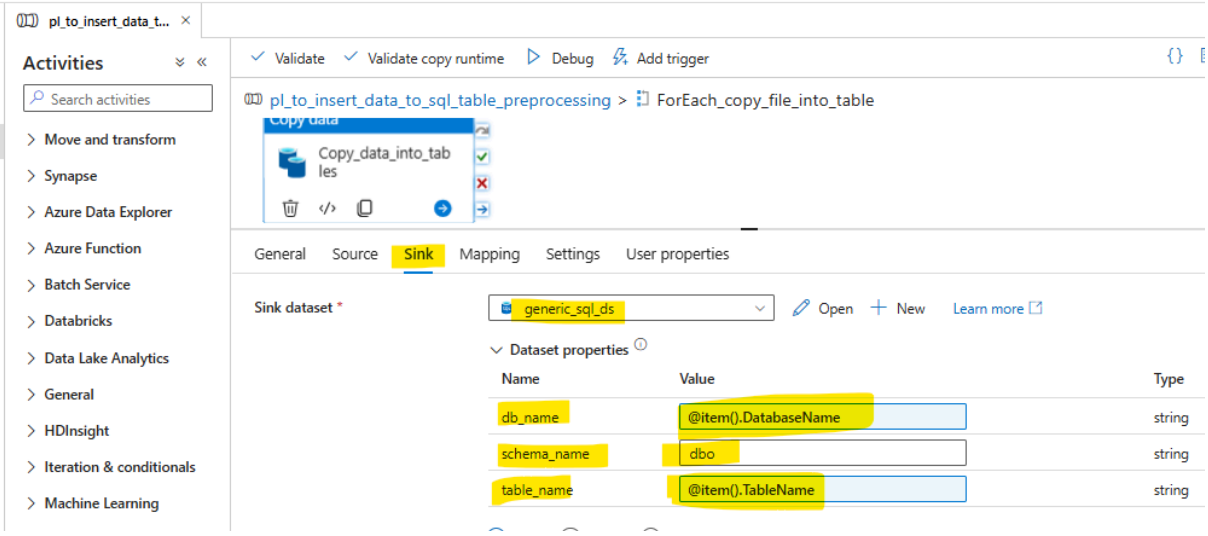
Source dataset * generic_adls_flat_file_ds

Dataset properties

Name	Value
file_name	@item().FileName
file_path	@item().FilePath
container	@item().Container

File path type ● File path in dataset ○ Wildcard file path ○ List of files

Sink: Use the destination dataset.

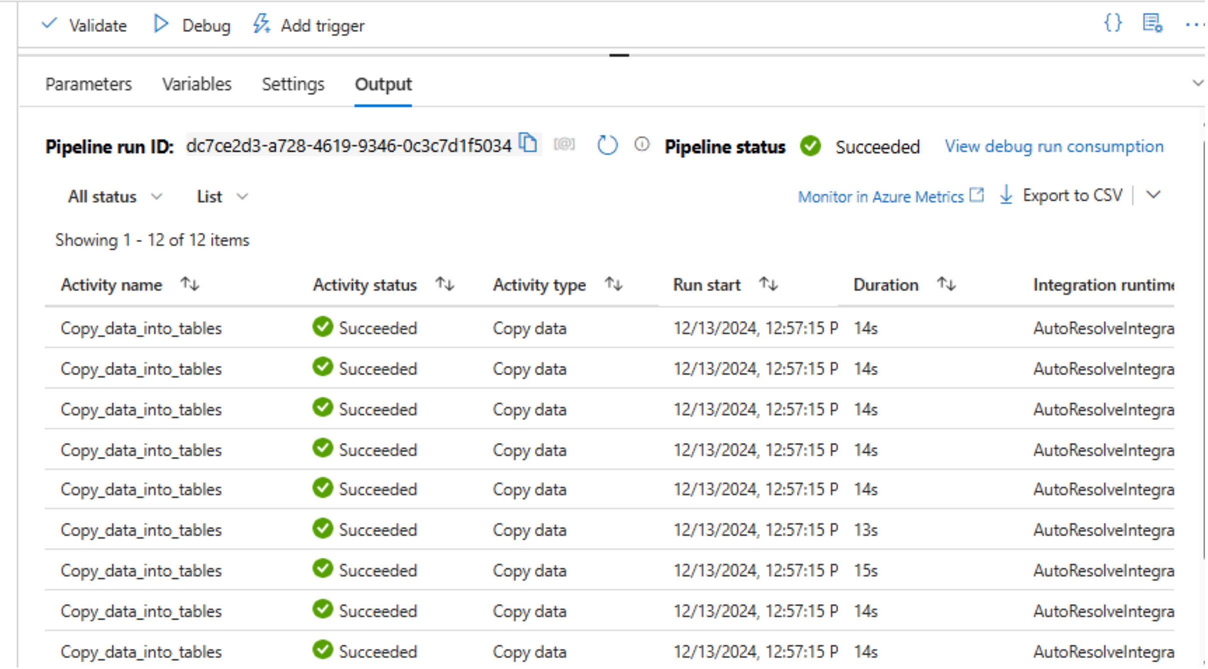


Sink dataset * generic_sql_ds

Dataset properties

Name	Value	Type
db_name	@item().DatabaseName	string
schema_name	dbo	string
table_name	@item().TableName	string

This pipeline will copy the data from the file into the tables in the sql database. On successfully running the pipeline we will get below output.



Pipeline run ID: dc7ce2d3-a728-4619-9346-0c3c7d1f5034 Pipeline status Succeeded

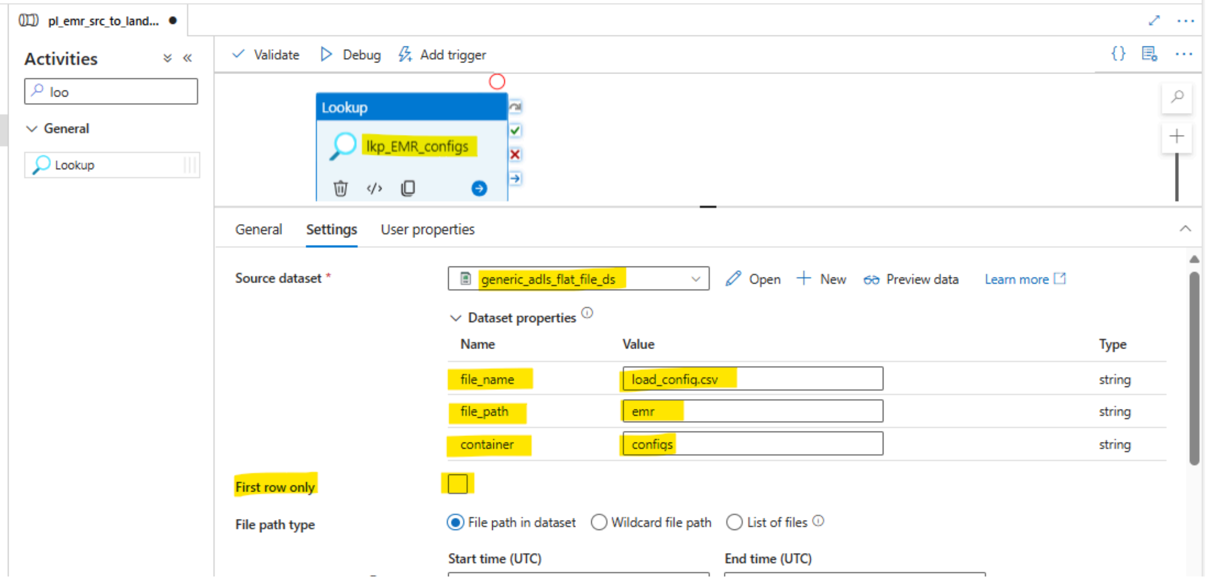
Showing 1 - 12 of 12 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtim
Copy_data_into_tables	Succeeded	Copy data	12/13/2024, 12:57:15 P	14s	AutoResolveIntegra
Copy_data_into_tables	Succeeded	Copy data	12/13/2024, 12:57:15 P	14s	AutoResolveIntegra
Copy_data_into_tables	Succeeded	Copy data	12/13/2024, 12:57:15 P	14s	AutoResolveIntegra
Copy_data_into_tables	Succeeded	Copy data	12/13/2024, 12:57:15 P	14s	AutoResolveIntegra
Copy_data_into_tables	Succeeded	Copy data	12/13/2024, 12:57:15 P	14s	AutoResolveIntegra
Copy_data_into_tables	Succeeded	Copy data	12/13/2024, 12:57:15 P	13s	AutoResolveIntegra
Copy_data_into_tables	Succeeded	Copy data	12/13/2024, 12:57:15 P	15s	AutoResolveIntegra
Copy_data_into_tables	Succeeded	Copy data	12/13/2024, 12:57:15 P	14s	AutoResolveIntegra
Copy_data_into_tables	Succeeded	Copy data	12/13/2024, 12:57:15 P	14s	AutoResolveIntegra

Pipeline to copy data from Azure Sql db to Landing Folder in ADLS Gen2

1. To read the config file we will use Lookup activity.

In this for source dataset will be for configs file and we will pass the parameter values as shown below.



Source dataset * generic_adls_flat_file_ds

Open New Preview data Learn more

Dataset properties

Name	Value	Type
file_name	load_config.csv	string
file_path	emr	string
container	configs	string

First row only □

File path type ● File path in dataset ○ Wildcard file path ○ List of files

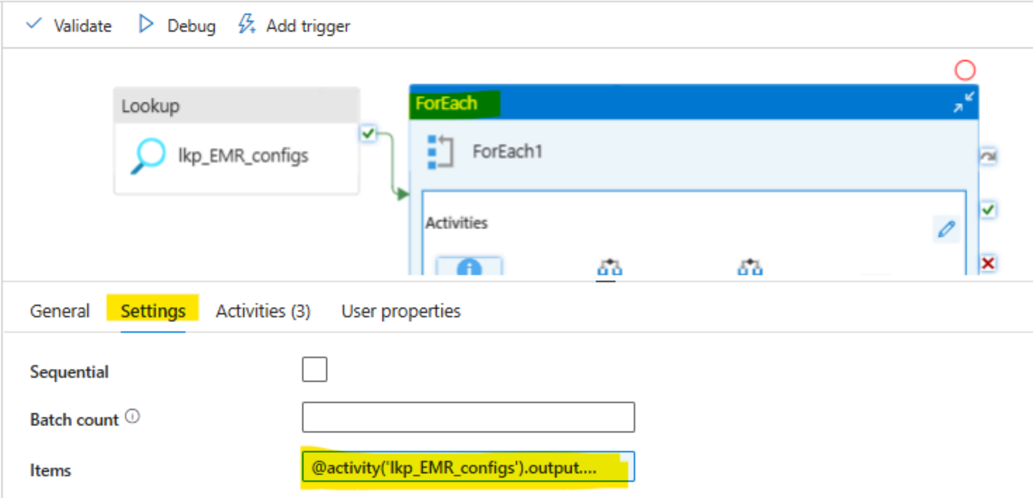
Start time (UTC) End time (UTC)

Then additionally we can preview the data.

2. In order to iterate through each row of configuration data we will use ForEach Activity.

Processing Logic Within ForEach Activity:

@activity('lkp_EMR_configs').output.value



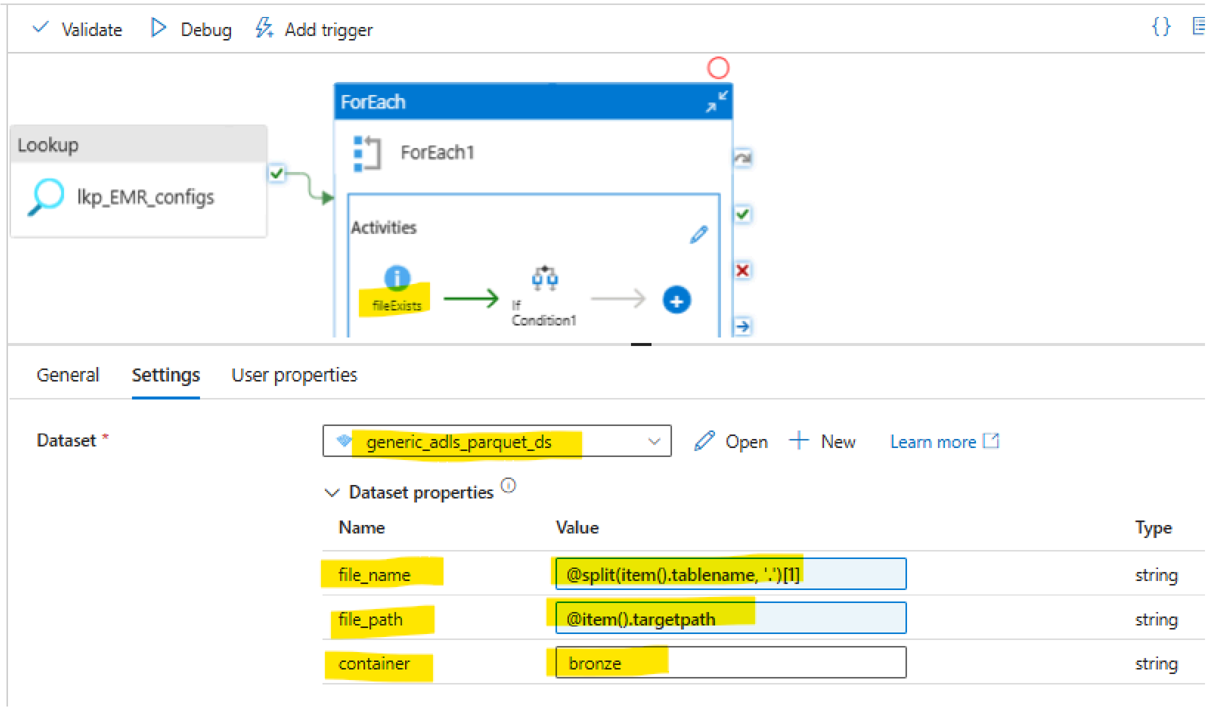
Sequential □

Batch count

Items @activity('lkp_EMR_configs').output....

a. We will use get metadata activity in order to check whether file exists in Bronze container:

To file name we will use below logic - @split(item().tablename, '.')[1]
file_path is present in lookup file as targetpath
And container name we will explicitly mention as bronze as shown below



Dataset * generic_adls_parquet_ds

Dataset properties

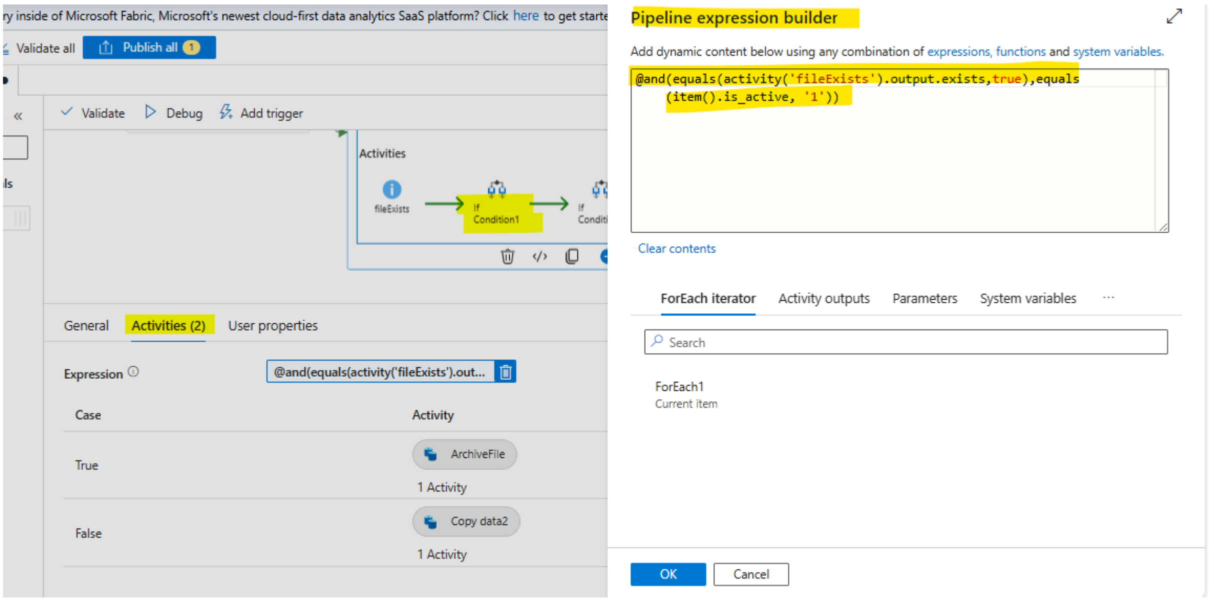
Name	Value	Type
file_name	@split(item().tablename, '.')[1]	string
file_path	@item().targetpath	string
container	bronze	string

This will check if the file exists in the Bronze container. Based on the file's presence or absence, we will use an If Condition activity to determine the subsequent processing steps.

b. Use an If Condition activity based on the file's existence.

Condition 1: File Exists (True) => Move the file to the Archive folder.

condition:
@and(equals(activity('fileExists').output.exists,true),equals(item().is_active, '1'))



ry inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get starte

Validate all | Publish all 1

Validate | Debug | Add trigger

Activities: fileExists → If Condition1 → If Conditi

General | Activities (2) | User properties

Expression: @and(equals(activity('fileExists').out...

Case	Activity
True	ArchiveFile 1 Activity
False	Copy data2 1 Activity

Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@and(equals(activity('fileExists').output.exists,true),equals
    (item().is_active, '1'))
```

Clear contents

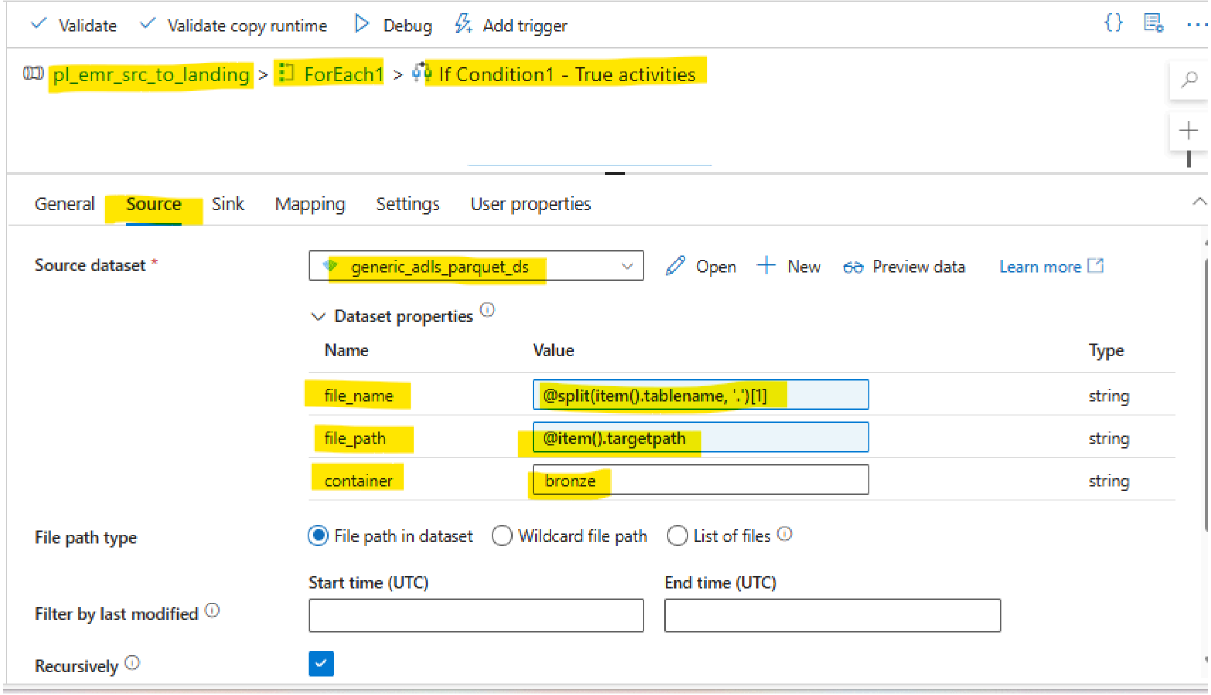
ForEach iterator | Activity outputs | Parameters | System variables | ···

Search

ForEach1
Current item

OK | Cancel

Source: Container: Bronze, Path: hosa, File: encounters



Validate | Validate copy runtime | Debug | Add trigger

pl_emr_src_to_landing > ForEach1 > If Condition1 - True activities

General | Source | Sink | Mapping | Settings | User properties

Source dataset *: generic_adls_parquet_ds | Open | New | Preview data | Learn more

Dataset properties

Name	Value	Type
file_name	@split(item().tablename, '.')[1]	string
file_path	@item().targetpath	string
container	bronze	string

File path type: ● File path in dataset ○ Wildcard file path ○ List of files

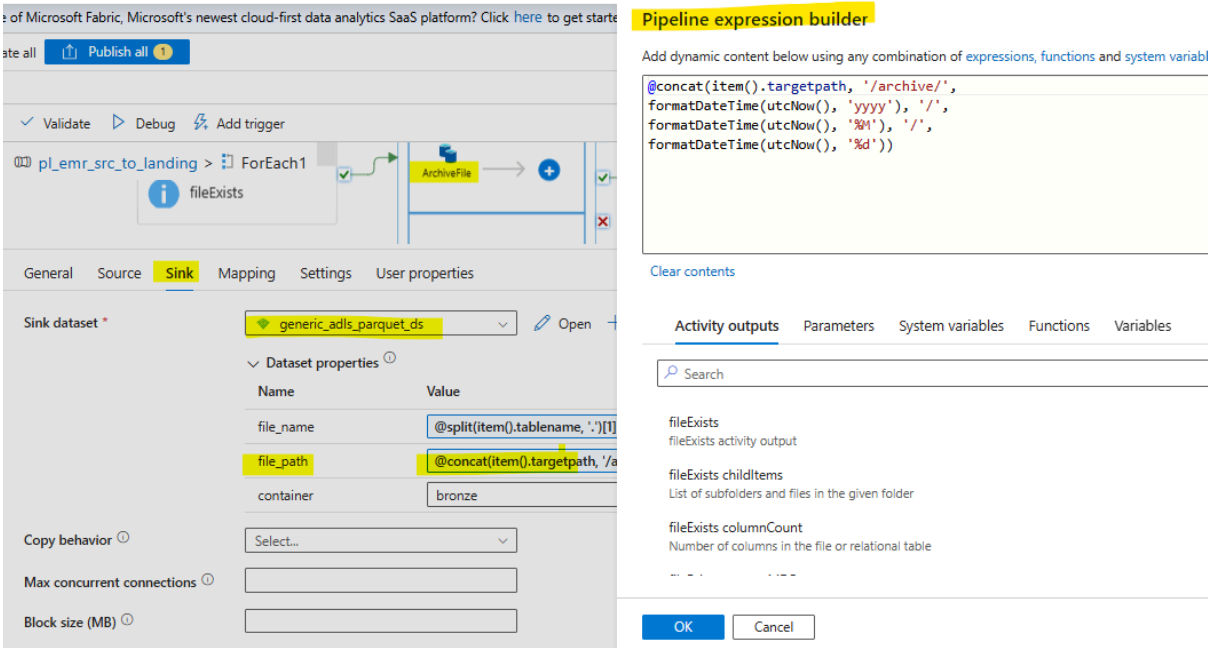
Filter by last modified: Start time (UTC) | End time (UTC)

Recursively: ☑

Target: Container: Bronze,

File_path -Path: hosa/archive/<year>/<month>/<day> =>
@concat(item().targetpath, '/archive/',
formatDateTime(utcNow(), 'yyyy'), '/',
formatDateTime(utcNow(), '%M'), '/',
formatDateTime(utcNow(), '%d'))

File_name - @split(item().tablename, '.')[1]



Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variabl

```
@concat(item().targetpath, '/archive/',
formatDateTime(utcNow(), 'yyyy'), '/',
formatDateTime(utcNow(), '%M'), '/',
formatDateTime(utcNow(), '%d'))
```

Sink dataset * generic_adls_parquet_ds

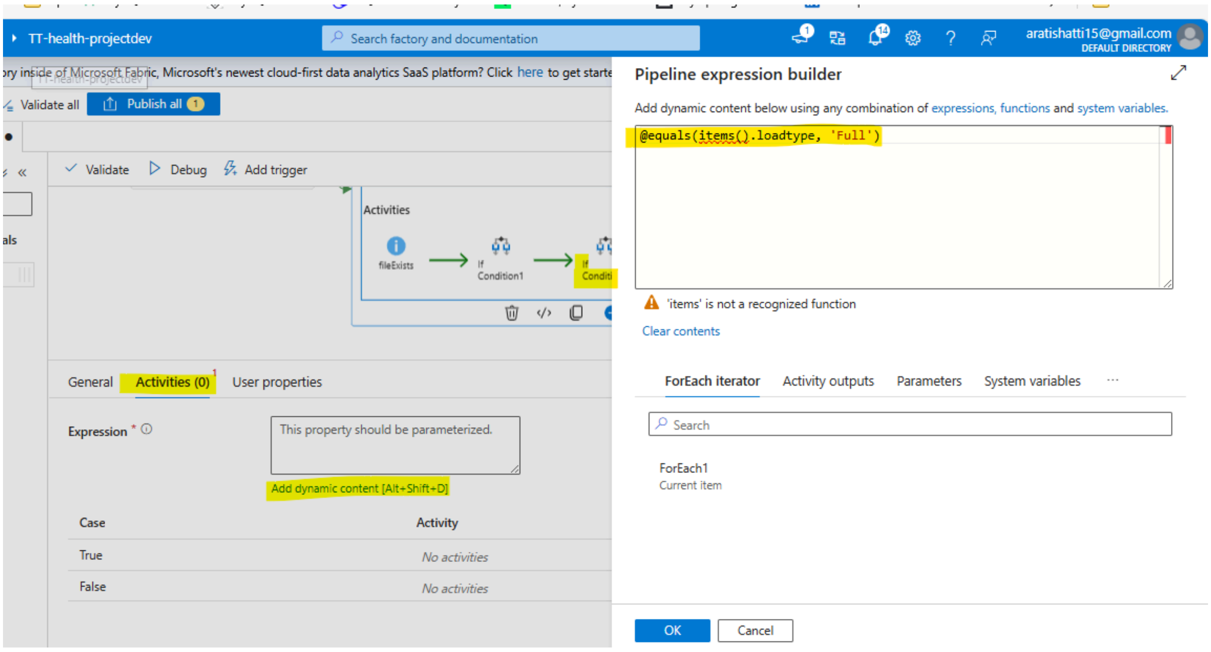
Dataset properties

Name	Value
file_name	@split(item().tablename, '.')[1]
file_path	@concat(item().targetpath, '/a
container	bronze

Copy behavior: Select...

c. Determine if it’s a full load or incremental load using If condition.

@equals(items().loadtype, 'Full')



Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@equals(items().loadtype, 'Full')
```

⚠ 'items' is not a recognized function

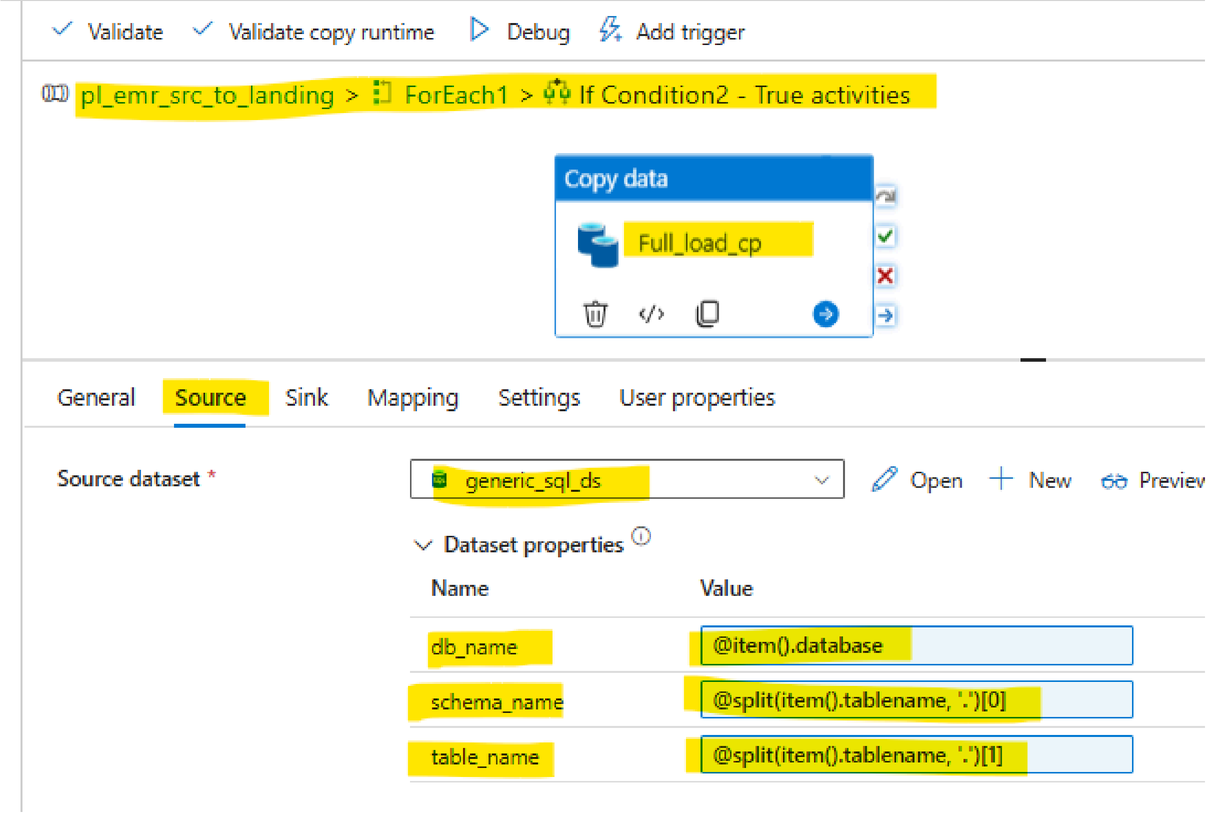
Expression * This property should be parameterized.

Add dynamic content [Alt+Shift+D]

Case	Activity
True	No activities
False	No activities

If “If condition” holds true => Full Load => Copy all data from the database table. => Enter Log details in the audit table:

Folder and File Structure
Bronze Container:
Source Path: bronze/hosa
Target Path for Data Loads: bronze/<target-path>



✓ Validate ✓ Validate copy runtime ▷ Debug Add trigger

pl_emr_src_to_landing > ForEach1 > If Condition2 - True activities

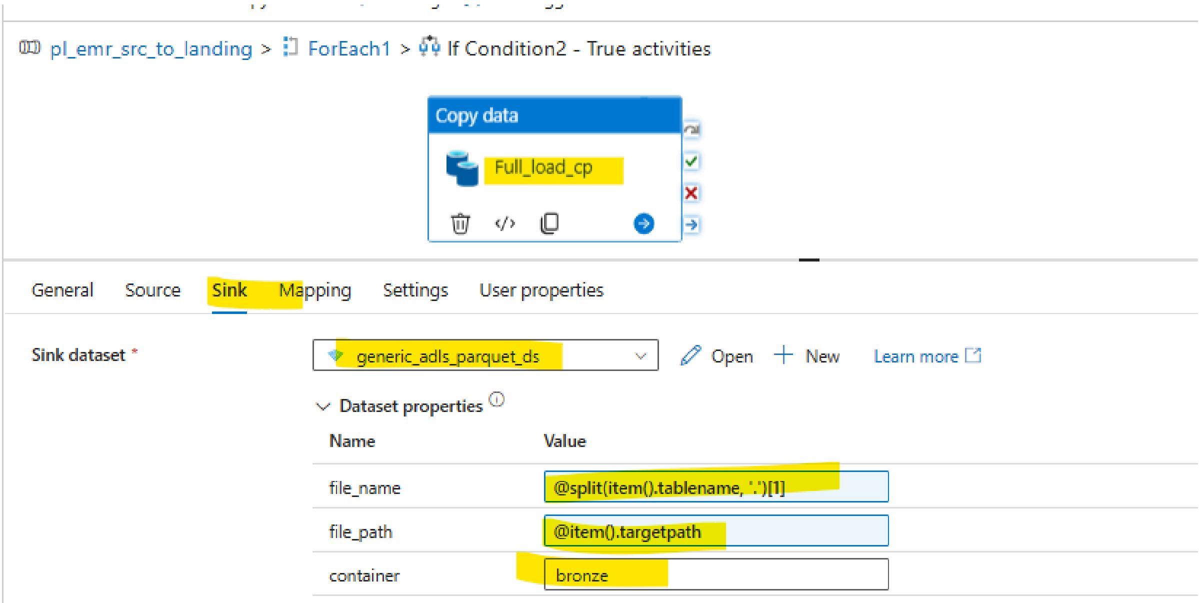
Copy data
Full_load_cp

General | Source | Sink | Mapping | Settings | User properties

Source dataset * generic_sql_ds | Open | + New | Preview

Dataset properties

Name	Value
db_name	@item().database
schema_name	@split(item().tablename, '.')[0]
table_name	@split(item().tablename, '.')[1]



pl_emr_src_to_landing > ForEach1 > If Condition2 - True activities

Copy data
Full_load_cp

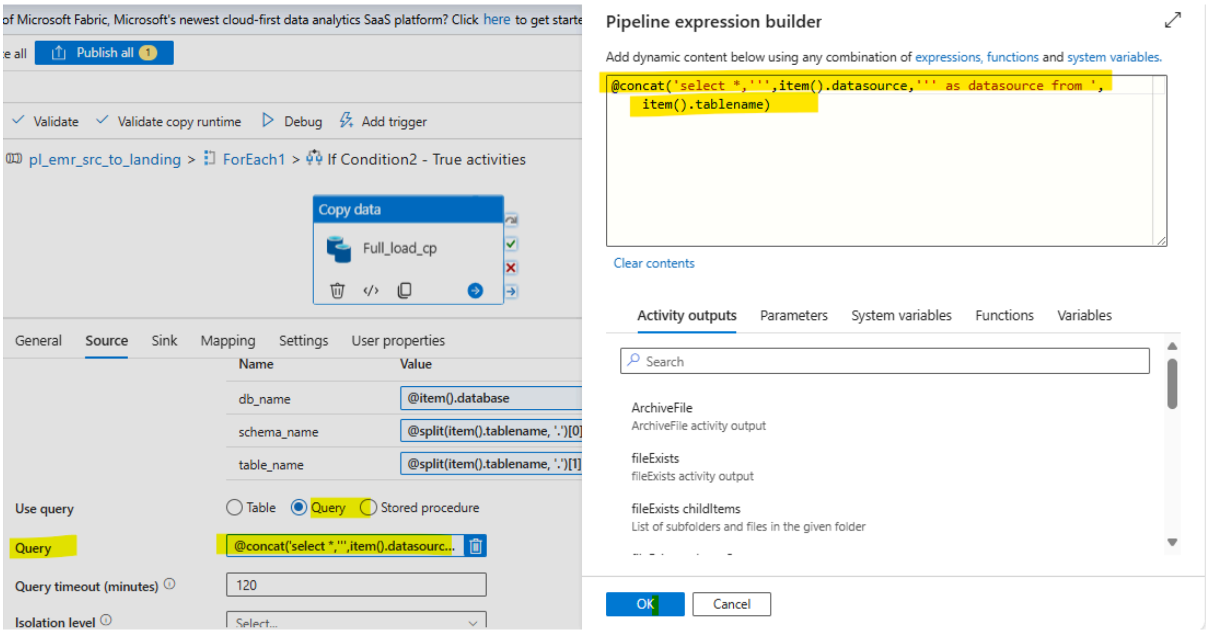
General | Source | Sink | Mapping | Settings | User properties

Sink dataset * generic_adls_parquet_ds | Open | + New | Learn more

Dataset properties

Name	Value
file_name	@split(item().tablename, '.')[1]
file_path	@item().targetpath
container	bronze

Query: @concat('select *,''',item().datasource,''' as datasource from ',item().tablename)



Name	Value
db_name	@item().database
schema_name	@split(item().tablename, '.')[0]
table_name	@split(item().tablename, '.')[1]

Use query ○ Table ● Query ○ Stored procedure

Query @concat('select *,''',item().datasourc...

Query timeout (minutes) 120

Isolation level Select...

Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@concat('select *,''',item().datasource,''' as datasource from ',
    item().tablename)
```

Clear contents

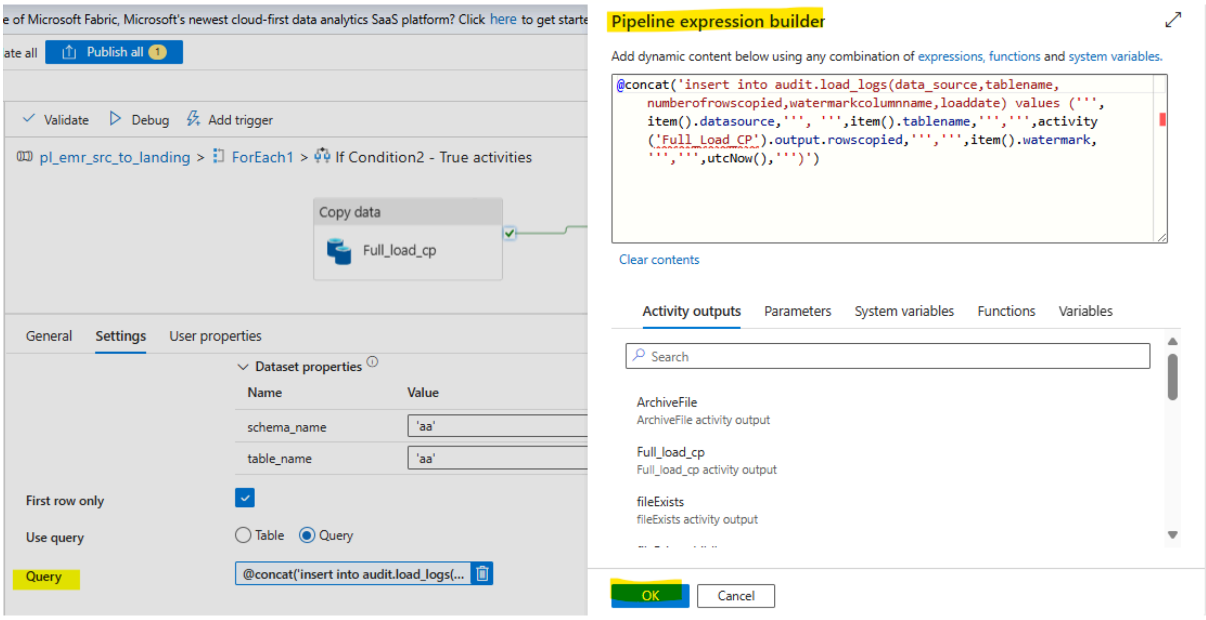
ArchiveFile
ArchiveFile activity output

fileExists
fileExists activity output

fileExists childItems
List of subfolders and files in the given folder

Enter Log details in the audit table:

Query: @concat('insert into audit.load_logs(data_source,tablename,numberofrowscopied,watermarkcolumnname,loaddate) values (''',item().datasource,''', ''',item().tablename,''',''',activity('Full_Load_CP').output.rowscopied,''',''',item().watermark,''',''',utcNow(),''')')


e of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get starte
ate all
Publish all 1
Validate
Debug
Add trigger
pl_emr_src_to_landing > ForEach1 > If Condition2 - True activities
Copy data
Full_load_cp
General
Settings
User properties
Activity outputs
Parameters
System variables
Functions
Variables
Search
OK
Cancel

Dataset properties

Name	Value
schema_name	'aa'
table_name	'aa'

First row only ☑

Use query ○ Table ● Query

Query @concat('insert into audit.load_logs(...

Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@concat('insert into audit.load_logs(data_source,tablename,
    numberofrowscopied,watermarkcolumnname,loaddate) values (''',
    item().datasource,''', ''',item().tablename,''',''',activity
    ('Full_Load_CP').output.rowscopied,''',''',item().watermark,
    ''',''',utcNow(),''')')
```

Clear contents

ArchiveFile
ArchiveFile activity output

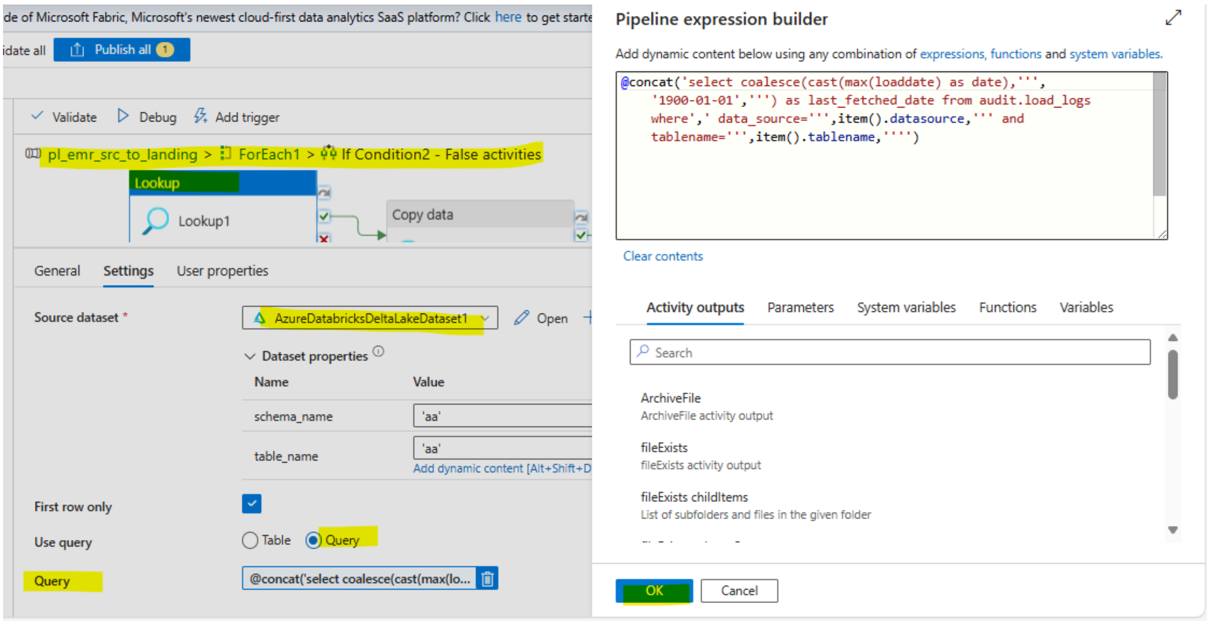
Full_load_cp
Full_load_cp activity output

fileExists
fileExists activity output

If condition is false => Incremental Load

(Fetch incremental data using the last fetched date) using Lookup=> Incremental load using copy activity =>Enter log details in the audit table:

Lookup:



de of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get starte
idate all
Publish all 1
Validate
Debug
Add trigger
pl_emr_src_to_landing > ForEach1 > If Condition2 - False activities
Lookup
Lookup1
Copy data
General
Settings
User properties

Source dataset * AzureDatabricksDeltaLakeDataset1

Dataset properties

Name	Value
schema_name	'aa'
table_name	'aa'

Open
Add dynamic content [Alt+Shift+D

First row only ☑

Use query ○ Table ● Query

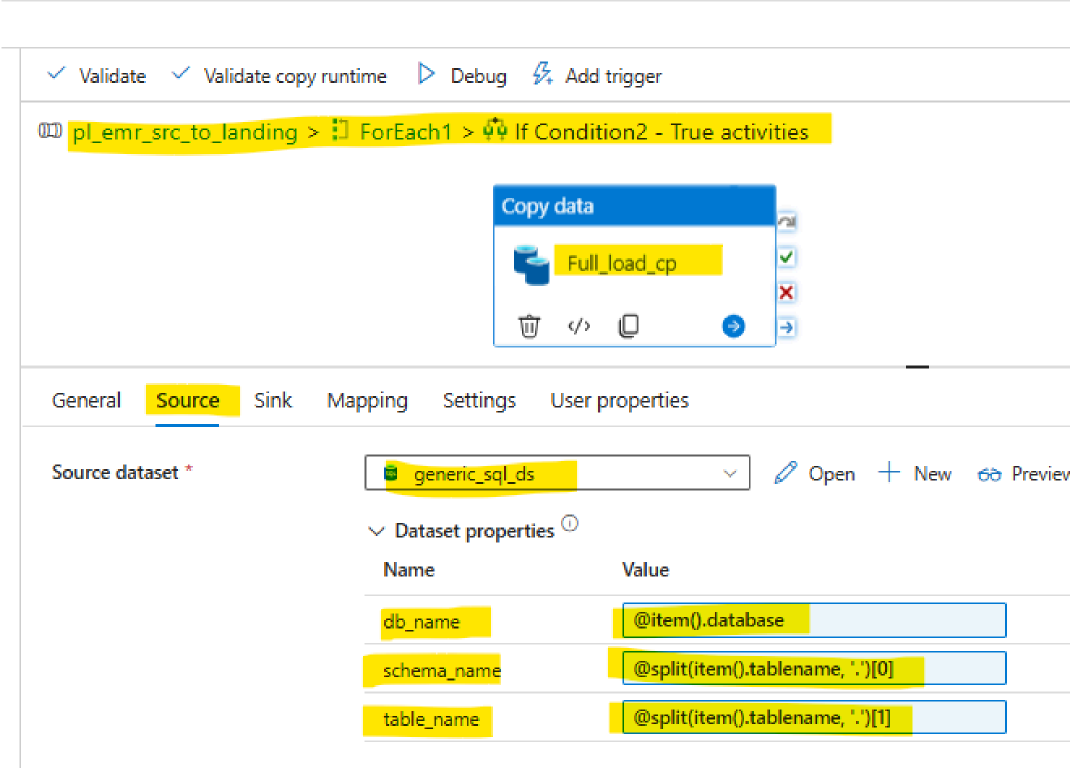
Query @concat('select coalesce(cast(max(lo...

Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@concat('select coalesce(cast(max(loaddate) as date),''',
    '1900-01-01',''') as last_fetched_date from audit.load_logs
    where',' data_source=''',item().datasource,''' and
    tablename=''',item().tablename,'''')
```

Incremental load:
Source Path: bronze/hosa

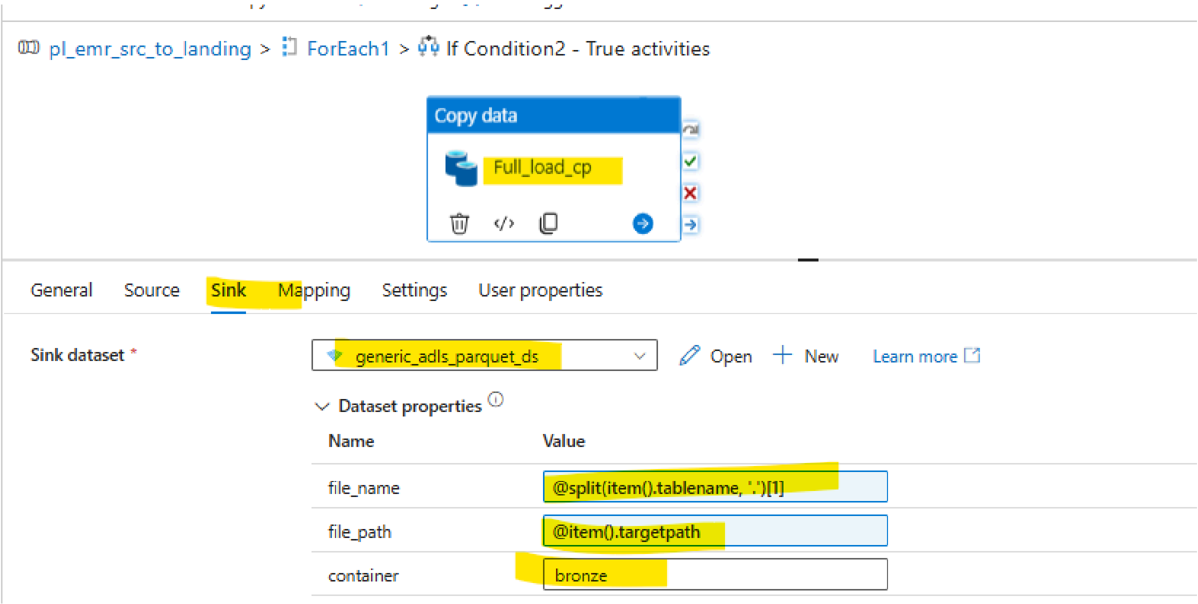


Source dataset * generic_sql_ds

Dataset properties

Name	Value
db_name	@item().database
schema_name	@split(item().tablename, '.')[0]
table_name	@split(item().tablename, '.')[1]

Target Path for Data Loads: bronze/<target-path>



pl_emr_src_to_landing > ForEach1 > If Condition2 - True activities

Copy data: Full_load_cp

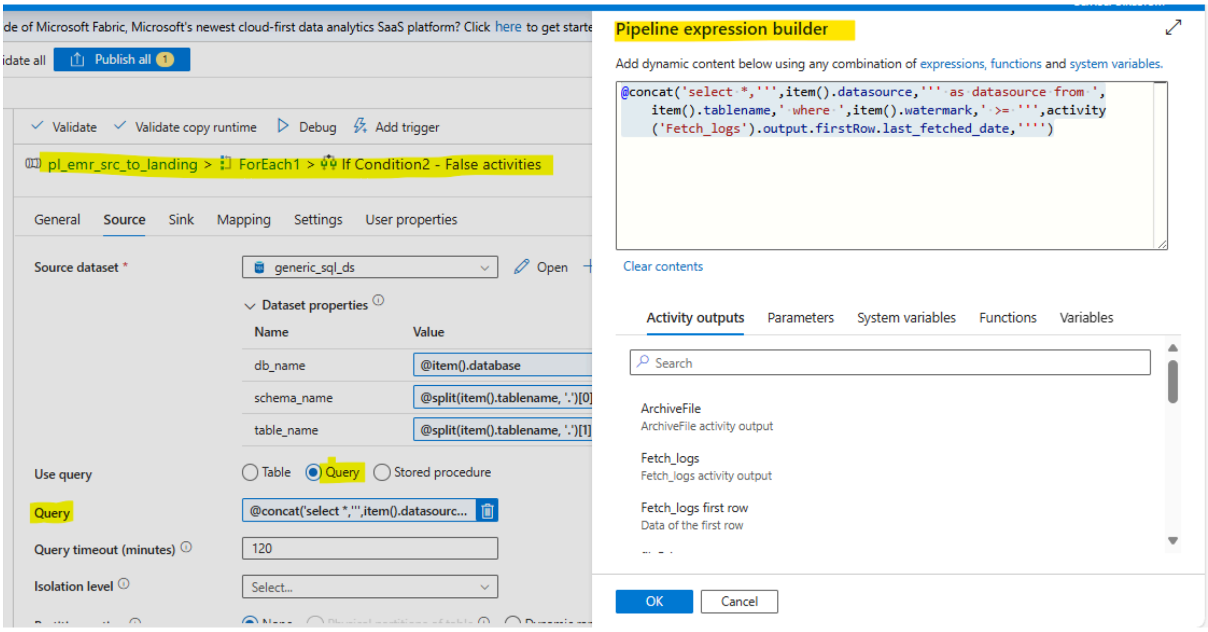
General | Source | Sink | Mapping | Settings | User properties

Sink dataset * generic_adls_parquet_ds

Dataset properties

Name	Value
file_name	@split(item().tablename, '.')[1]
file_path	@item().targetpath
container	bronze

Query: @concat('select *,''',item().datasource,''' as datasource from ',item().tablename,' where ',item().watermark,' >= ''',activity('Fetch_logs').output.firstRow.last_fetched_date,'''')



de of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get starte

pl_emr_src_to_landing > ForEach1 > If Condition2 - False activities

General | Source | Sink | Mapping | Settings | User properties

Source dataset * generic_sql_ds

Dataset properties

Name	Value
db_name	@item().database
schema_name	@split(item().tablename, '.')[0]
table_name	@split(item().tablename, '.')[1]

Use query: ○ Table ● Query ○ Stored procedure

Query: @concat('select *,''',item().datasourc...

Query timeout (minutes): 120

Isolation level: Select...

Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

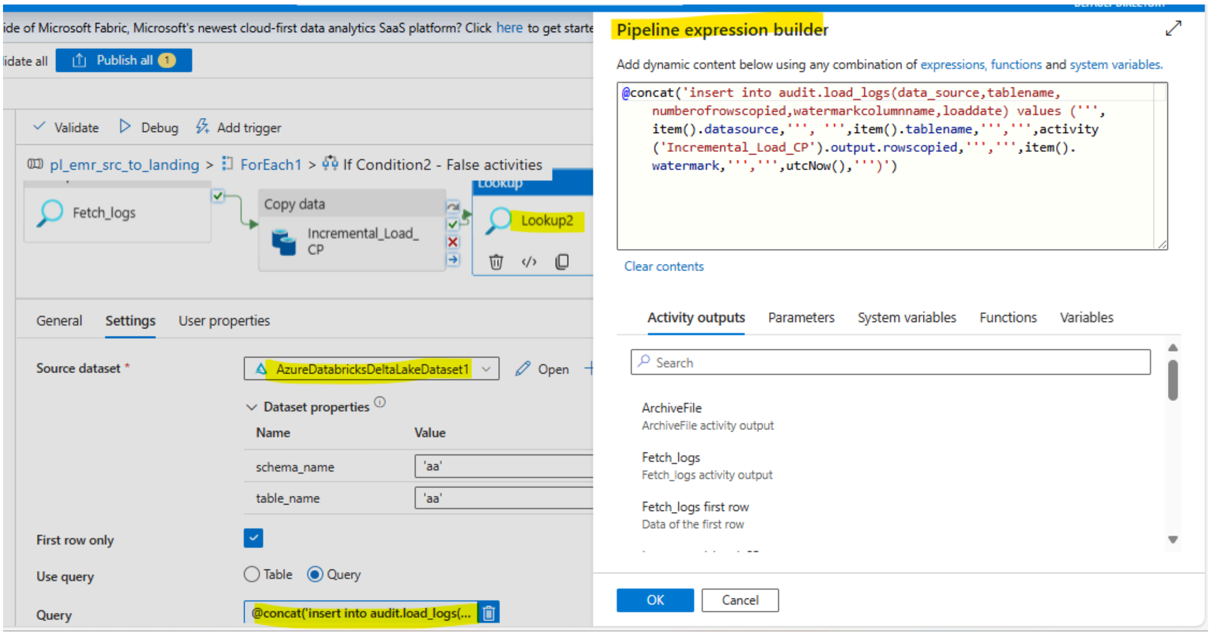
```
@concat('select *,''',item().datasource,''' as datasource from ',
    item().tablename,' where ',item().watermark,' >= ''',activity
    ('Fetch_logs').output.firstRow.last_fetched_date,'''')
```

ArchiveFile
ArchiveFile activity output

Fetch_logs
Fetch_logs activity output

Fetch_logs first row
Data of the first row

Lookup:



Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@concat('insert into audit.load_logs(data_source,tablename,
    numberofrowscopied,watermarkcolumnname,loaddate) values ('''',
    item().datasource,''', ''',item().tablename,''',''',activity
    ('Incremental_Load_CP').output.rowscopied,''',''',item().
    watermark,''',''',utcNow(),''')')
```

Source dataset * AzureDatabricksDeltaLakeDataset1

Dataset properties

Name	Value
schema_name	'aa'
table_name	'aa'

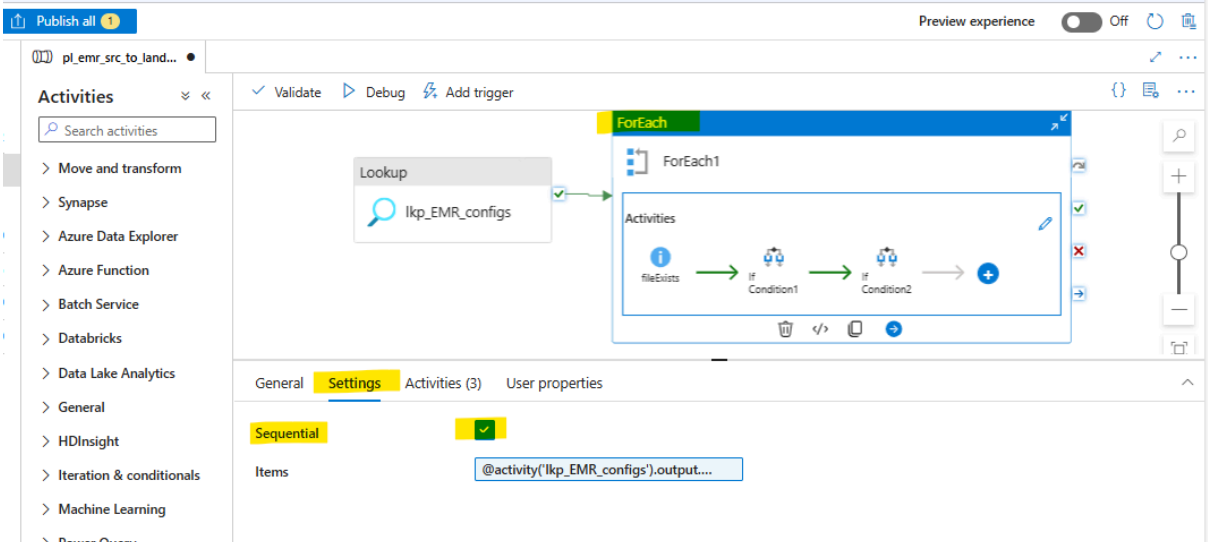
First row only ☑

Use query ○ Table ● Query

Query @concat('insert into audit.load_logs(...

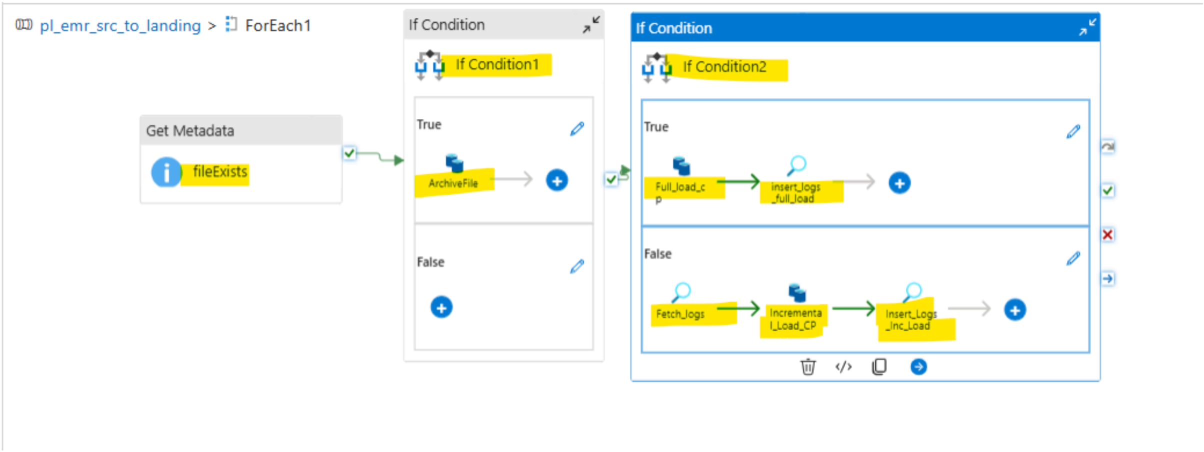
This is our complete pipeline:

Before running the pipeline for each activity select the “sequential” option as shown below.



Sequential ☑

Items @activity('lkp_EMR_configs').output....



But limitation with this pipeline is it is sequential which will we resolve in part 2

Part 2:

In this section, we will focus on improving our data pipeline and governance by implementing the following:

Clean and organize raw data into a structured format.
Apply Common Data Model (CDM) standards and implement Slowly Changing Dimensions (SCD2) for historical tracking.

Use Delta tables for efficient data storage and updates.

Build Fact and Dimension tables for better data analysis and reporting.

Secure sensitive data by integrating Azure Key Vault for managing secrets and credentials.

Standardize names across datasets, pipelines, and tables for better organization and understanding.

Optimize Azure Data Factory pipelines to run multiple processes at the same time, reducing execution time.

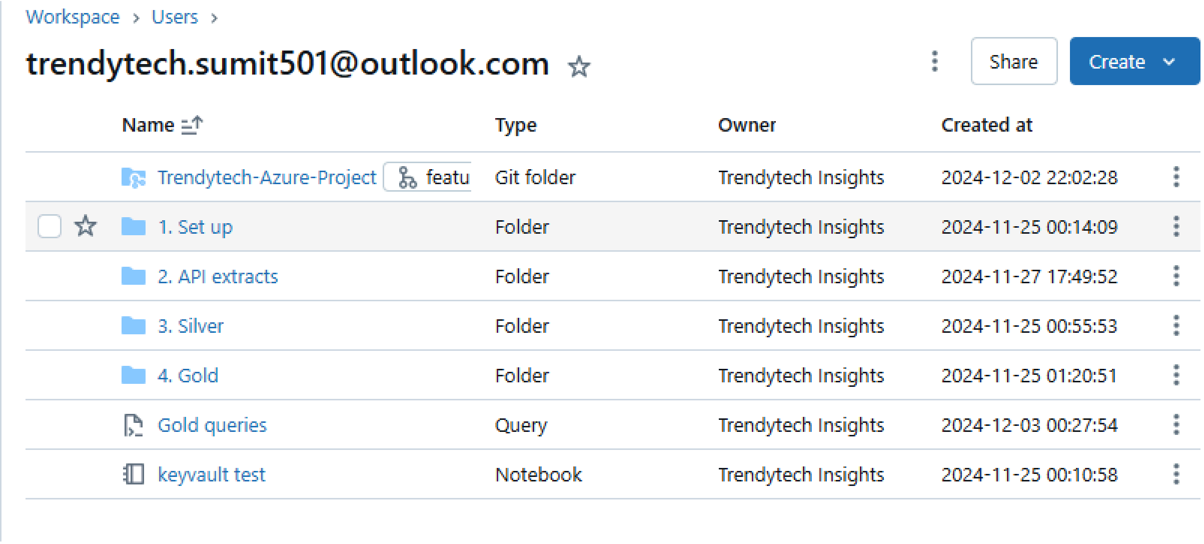
Integrate external APIs for dynamic data processing.
Handle Claims Data effectively, including implementing an is_active flag for tracking record statuses.

Transition from a local Hive Metastore to Databricks Unity Catalog for centralized metadata management and improved data governance.

We will first create new databricks workspace “tt-hc-adb-ws”, select the “Premium (+ Role-based access controls)” while creating workspace

Note: Also you can name the resource group as “TT-HeathProjectDev”

Also to organize the notebook we will create the folder as shown below.

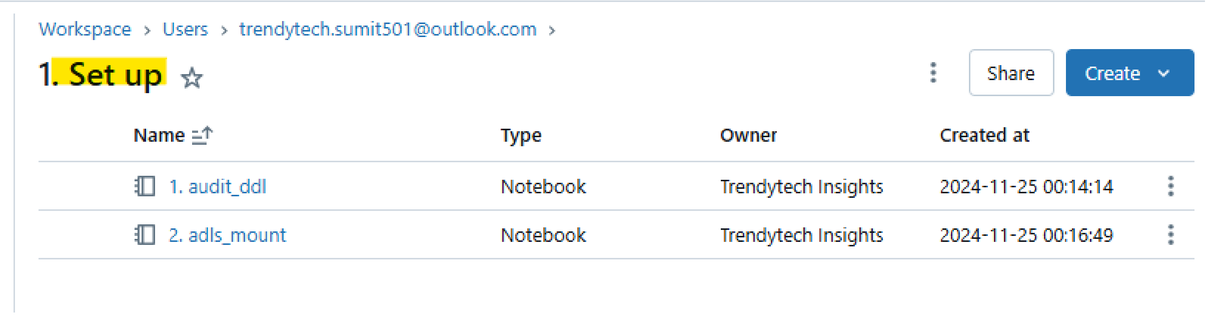


Workspace › Users ›

trendytech.sumit501@outlook.com ☆

Name	Type	Owner	Created at
Trendytech-Azure-Project featu	Git folder	Trendytech Insights	2024-12-02 22:02:28
1. Set up	Folder	Trendytech Insights	2024-11-25 00:14:09
2. API extracts	Folder	Trendytech Insights	2024-11-27 17:49:52
3. Silver	Folder	Trendytech Insights	2024-11-25 00:55:53
4. Gold	Folder	Trendytech Insights	2024-11-25 01:20:51
Gold queries	Query	Trendytech Insights	2024-12-03 00:27:54
keyvault test	Notebook	Trendytech Insights	2024-11-25 00:10:58

1. Set up:

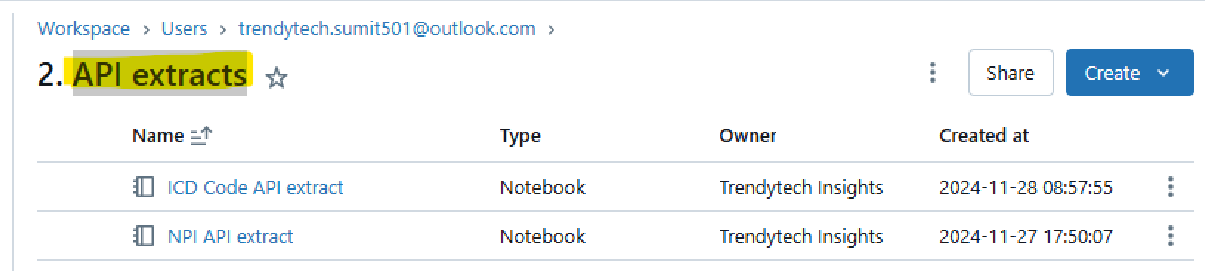


Workspace › Users › trendytech.sumit501@outlook.com ›

1. Set up ☆

Name	Type	Owner	Created at
1. audit_ddl	Notebook	Trendytech Insights	2024-11-25 00:14:14
2. adls_mount	Notebook	Trendytech Insights	2024-11-25 00:16:49

2. API extracts

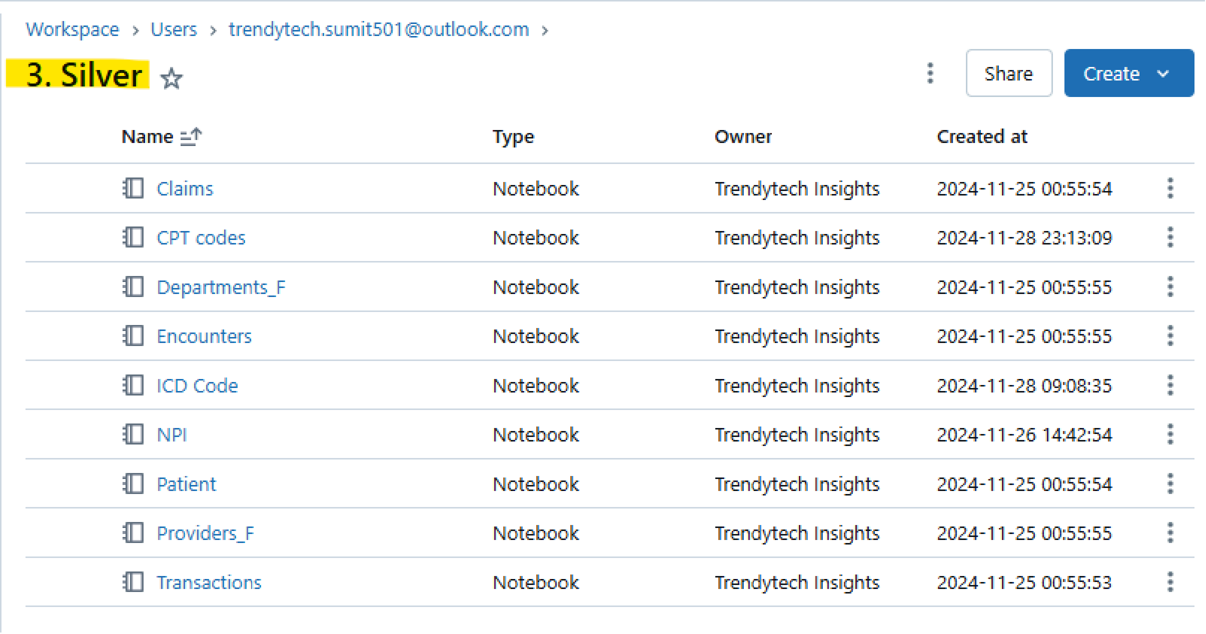


Workspace › Users › trendytech.sumit501@outlook.com ›

2. API extracts ☆

Name	Type	Owner	Created at
ICD Code API extract	Notebook	Trendytech Insights	2024-11-28 08:57:55
NPI API extract	Notebook	Trendytech Insights	2024-11-27 17:50:07

3. Silver

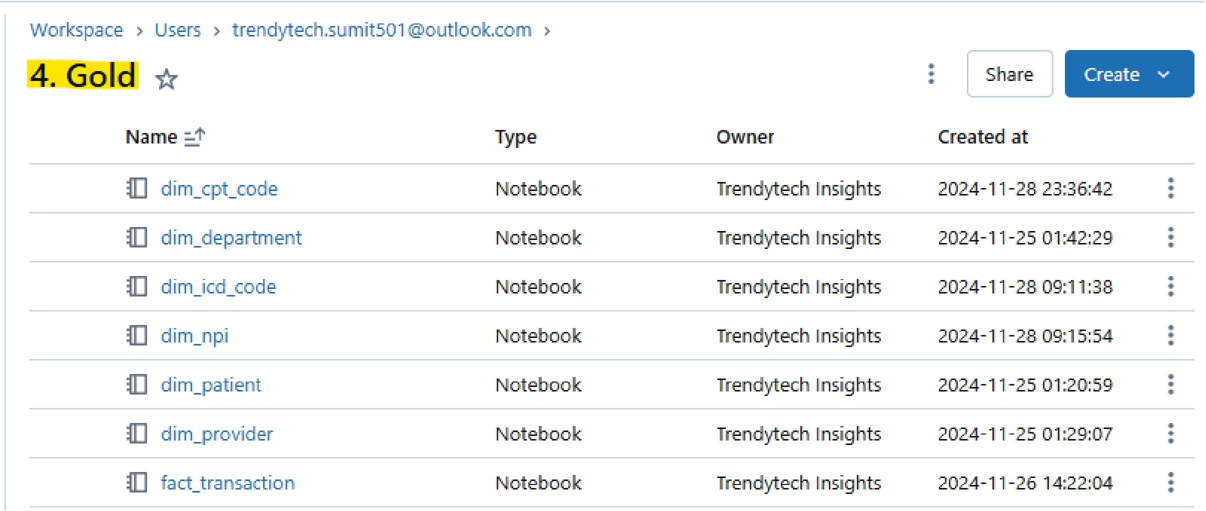


Workspace › Users › trendytech.sumit501@outlook.com ›

3. Silver ☆

Name	Type	Owner	Created at
Claims	Notebook	Trendytech Insights	2024-11-25 00:55:54
CPT codes	Notebook	Trendytech Insights	2024-11-28 23:13:09
Departments_F	Notebook	Trendytech Insights	2024-11-25 00:55:55
Encounters	Notebook	Trendytech Insights	2024-11-25 00:55:55
ICD Code	Notebook	Trendytech Insights	2024-11-28 09:08:35
NPI	Notebook	Trendytech Insights	2024-11-26 14:42:54
Patient	Notebook	Trendytech Insights	2024-11-25 00:55:54
Providers_F	Notebook	Trendytech Insights	2024-11-25 00:55:55
Transactions	Notebook	Trendytech Insights	2024-11-25 00:55:53

4. Gold



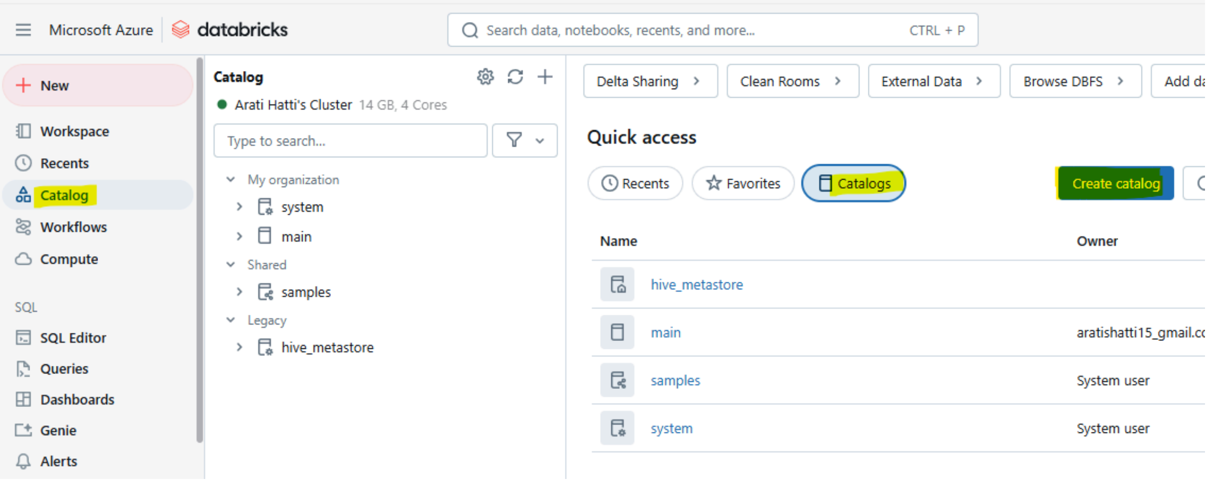
Workspace › Users › trendytech.sumit501@outlook.com ›

4. Gold ☆

Name	Type	Owner	Created at
dim_cpt_code	Notebook	Trendytech Insights	2024-11-28 23:36:42
dim_department	Notebook	Trendytech Insights	2024-11-25 01:42:29
dim_icd_code	Notebook	Trendytech Insights	2024-11-28 09:11:38
dim_npi	Notebook	Trendytech Insights	2024-11-28 09:15:54
dim_patient	Notebook	Trendytech Insights	2024-11-25 01:20:59
dim_provider	Notebook	Trendytech Insights	2024-11-25 01:29:07
fact_transaction	Notebook	Trendytech Insights	2024-11-26 14:22:04

Note: After creating the databricks workspace enable the DBFS.

Create the catalog “tt-hc-adb-ws” as shown below.



Name	Owner
hive_metastore	
main	aratishatti15_gmail.c
samples	System user
system	System user

Now we will create the audit database using as shown below