

Azure Data Engineering Project Report

Technologies Used

- **Azure Data Factory:** For data orchestration and pipeline management
- **Azure Databricks:** For advanced data processing and analytics
- **Azure Synapse Analytics:** For data warehousing and serverless SQL
- **Azure Storage Account:** For data lake storage
- **Power BI:** For data visualization and reporting

Step 01

Create resource group to keep all the necessary resources

Step 02

Create storage account. Choose Data Lake Gen2 to create the file system in hierarchical structure.

Blob Storage for:

storing files, images, logs, backups, etc.

Don't need analytics or a hierarchical file structure.

Data Lake Gen2 for:

building a data lake for analytics or ML.

need Hadoop/Spark/Hive integration.

have a file system-like structure with directories.

Step 3

Create Azure Data Factory. Create 3 containers for bronze, silver, gold in Storage Account

Implementing Data Pipeline in Azure Data Factory

1.Static Connection Implementation

The project begins with establishing static connections to data sources. Static connections represent the traditional approach where data files are loaded individually using manual copy operations.

Key Characteristics of Static Approach:

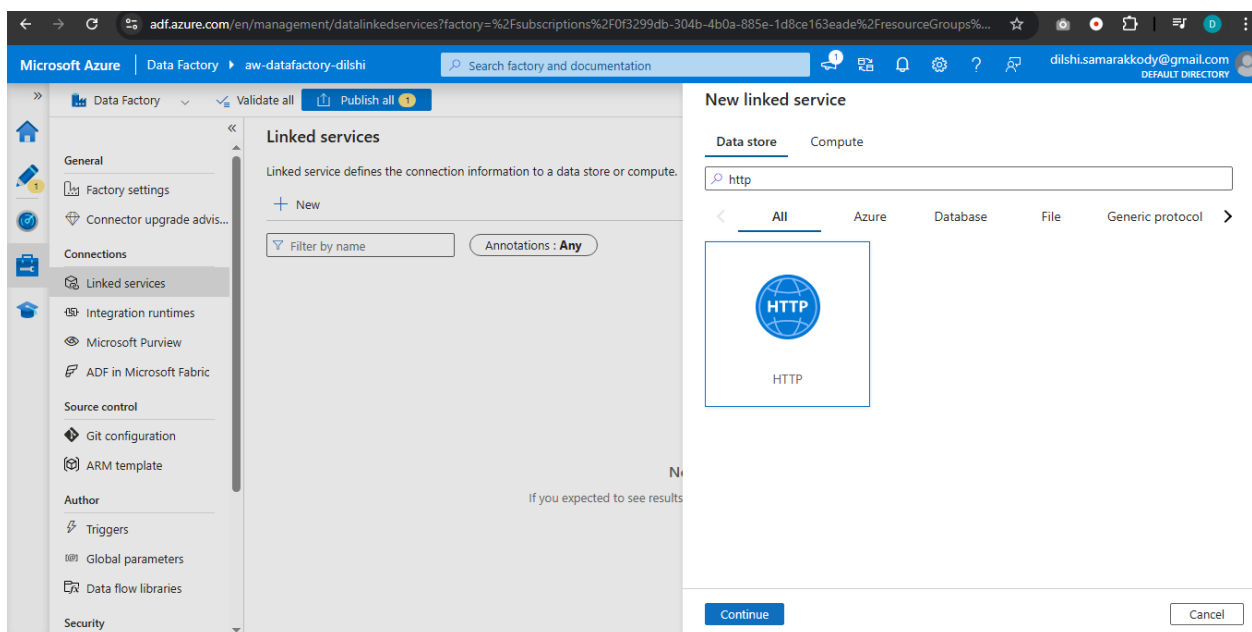
- Manual intervention required for each data file
- Individual file processing
- Limited scalability
- Suitable for small, infrequent data loads

Copy Data Activity Configuration

The static implementation uses Azure Data Factory's Copy Data activity with the following configurations:

- Source dataset definition
- Destination dataset configuration

To create a connection with source (Github) create http linked service



Create Destination linked service for Azure Data Lake Storage Gen2

Microsoft Azure | Data Factory | aw-datafactory-dilshi

Search factory and documentation

Validate all Publish all

Preview experience Off

General

- Factory settings
- Connector upgrade advis...

Connections

- Linked services
- Integration runtimes
- Microsoft Purview
- ADF in Microsoft Fabric

Source control

- Git configuration
- ARM template

Author

- Triggers
- Global parameters
- Data flow libraries

Security

Linked services

Linked service defines the connection information to a data store or compute. [Learn more](#)

+ New

Filter by name Annotations: Any

Showing 1 - 2 of 2 items

Name	Type	Related	Annotations
azuredatalakestorage	Azure Data Lake Storage Gen2	0	
httplinkedservice	HTTP	0	

Set properties for source and sink

Microsoft Azure | Data Factory | aw-datafactory-dilshi

Search factory and documentation

Validate all Publish all

raw_data_load

Activities

- Search activities
- Move and transform
 - Copy data
 - Data flow
- Synapse
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning

Copy data

copy_raw_data

General Source Sink Mapping Settings User properties

Source dataset *

Select...

Set properties

Name: ds_http

Linked service *: httplinkedservice

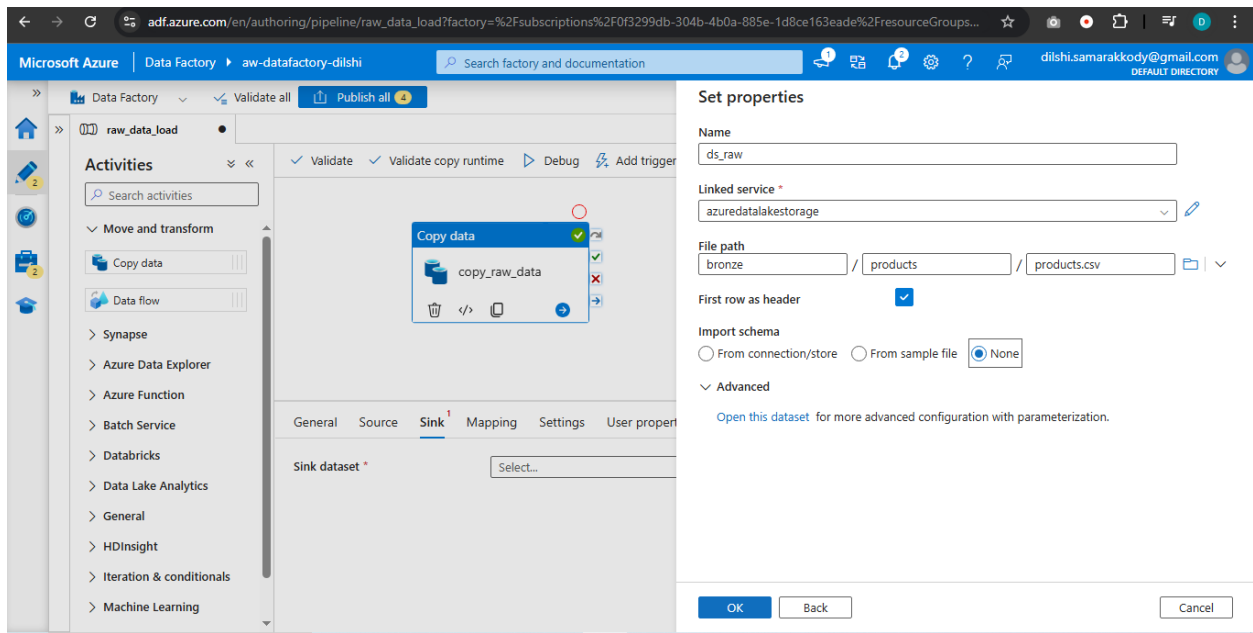
Relative URL: /Azure-Data-Engineering-Project/refs/heads/pro-01/data/AdventureWorks_Products.csv

First row as header: ☒

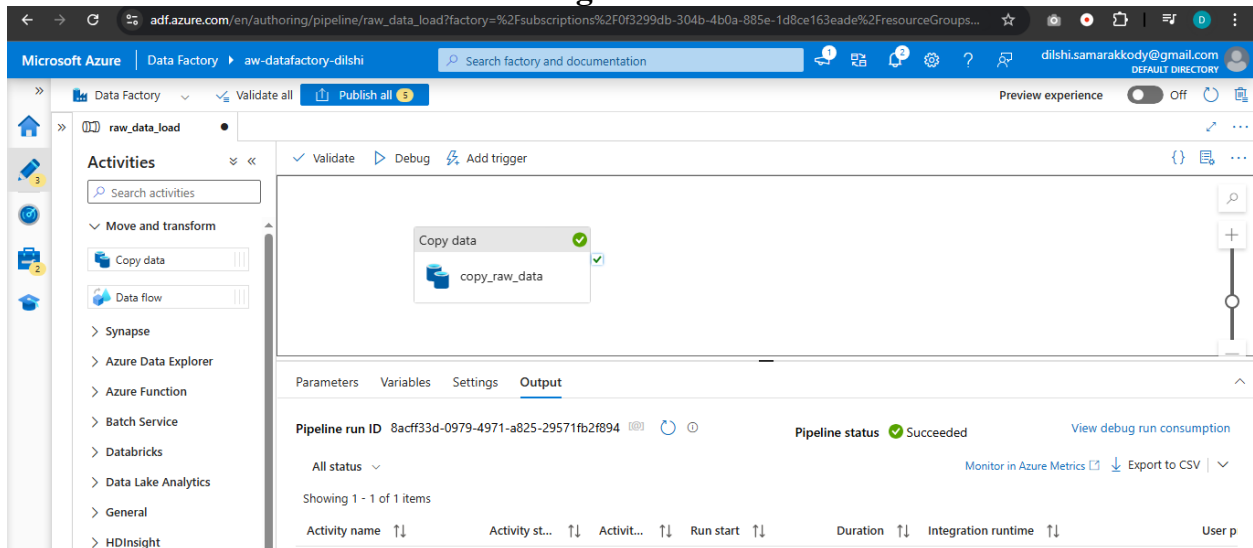
Import schema: ☐ From connection/store ☐ From sample file ☒ None

> Advanced

OK Back Cancel



Load raw data from GitHub to ADLS gen 2



2. Dynamic Pipeline Architecture

The dynamic pipelines represent a significant improvement in automation and scalability. Dynamic pipelines can automatically process multiple files and adapt to changing data requirements.

Key Components:

Lookup Activity: Retrieves metadata and configuration information

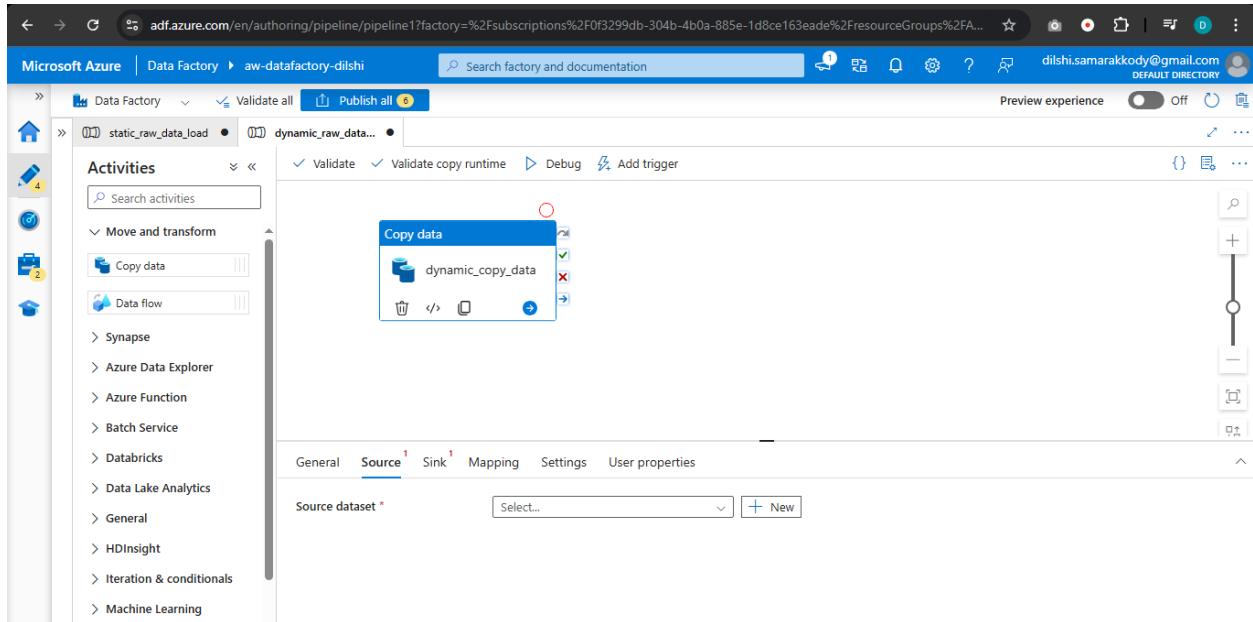
ForEach Activity: Iterates through data sources dynamically

Copy Data Activity: Performs data transfer operations

Data Flow: Handles complex transformations

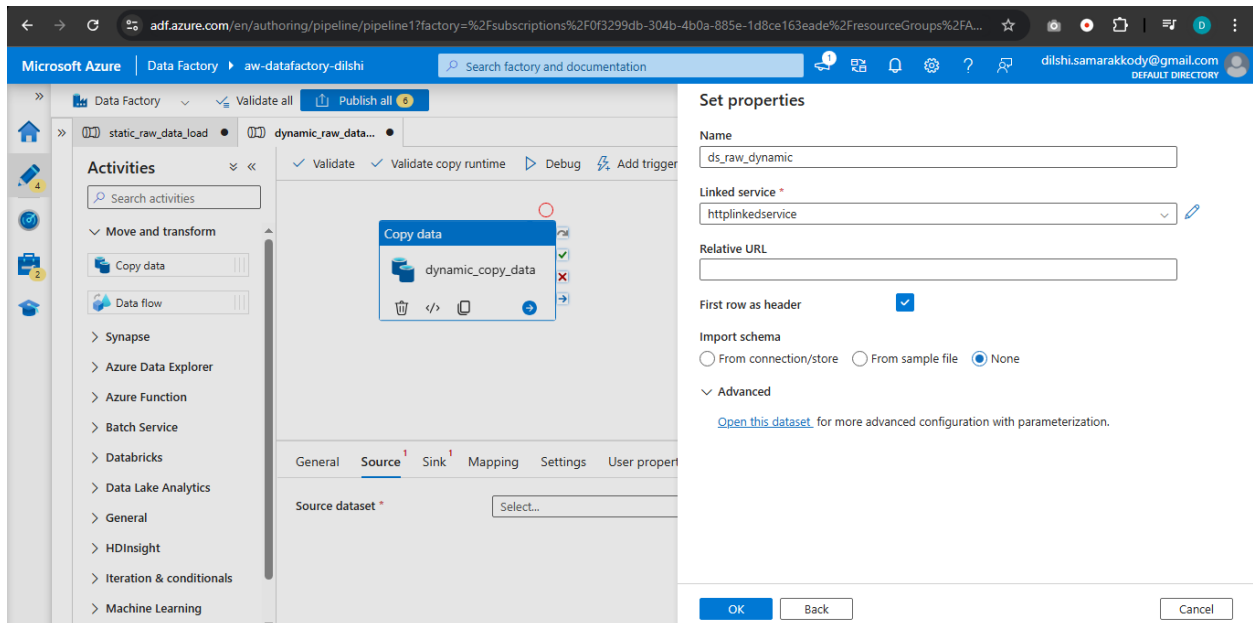
Step 5

Create parameterized copy data tool



Step 06

And set properties



Step 7

Create dynamic parameter for relative URL

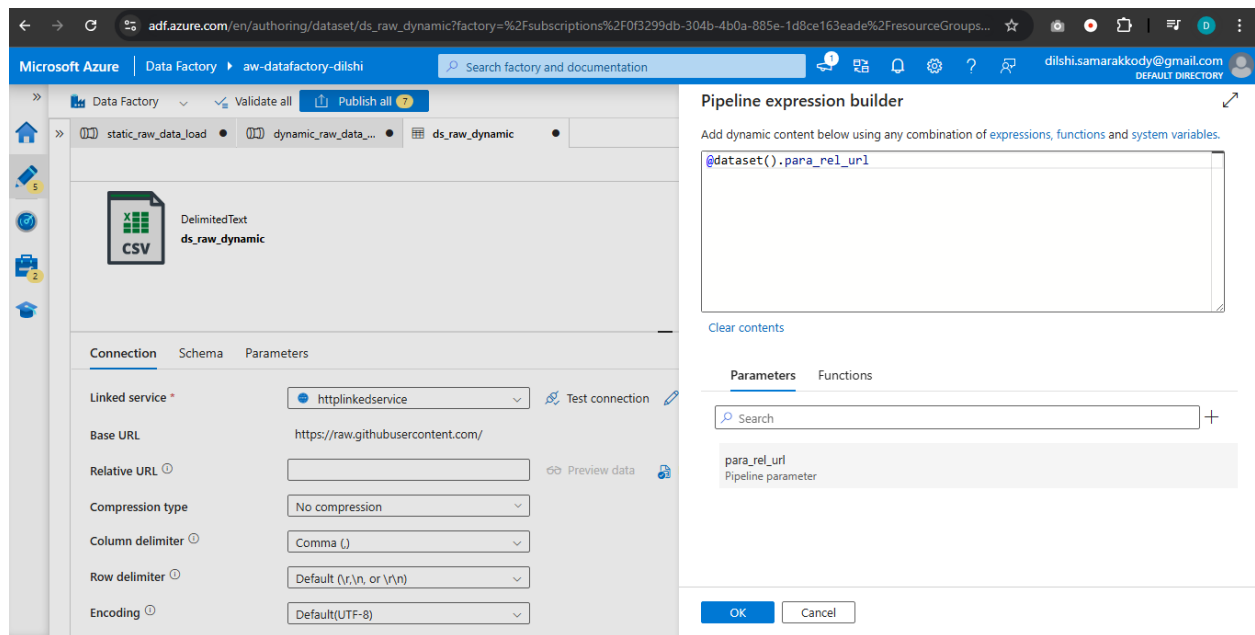
The screenshot displays the Microsoft Azure Data Factory portal interface. The top navigation bar shows the user is logged in as 'dilshi.samarakkody@gmail.com' and is viewing the 'aw-datafactory-dilshi' workspace. The main content area shows the configuration for a dataset named 'ds_raw_dynamic', which is a 'DelimitedText' type. The configuration is divided into three tabs: 'Connection', 'Schema', and 'Parameters'. The 'Connection' tab is active, showing the following settings:

- Linked service: `httplinkedservice` (with a dropdown arrow, 'Test connection' button, 'Edit' button, '+ New' button, and 'Learn more' link).
- Base URL: `https://raw.githubusercontent.com/`
- Relative URL: (empty text box) with a 'Preview data' button and a 'Detect format' button. Below the text box is a link: 'Add dynamic content [Alt+Shift+D]'.
- Compression type: `No compression` (with a dropdown arrow).
- Column delimiter: `Comma (,)` (with a dropdown arrow).
- Row delimiter: `Default (\r\n, or \n\n)` (with a dropdown arrow).
- Encoding: `Default(UTF-8)` (with a dropdown arrow).

Below the configuration panel, a 'New parameter' dialog box is open. It contains the following fields:

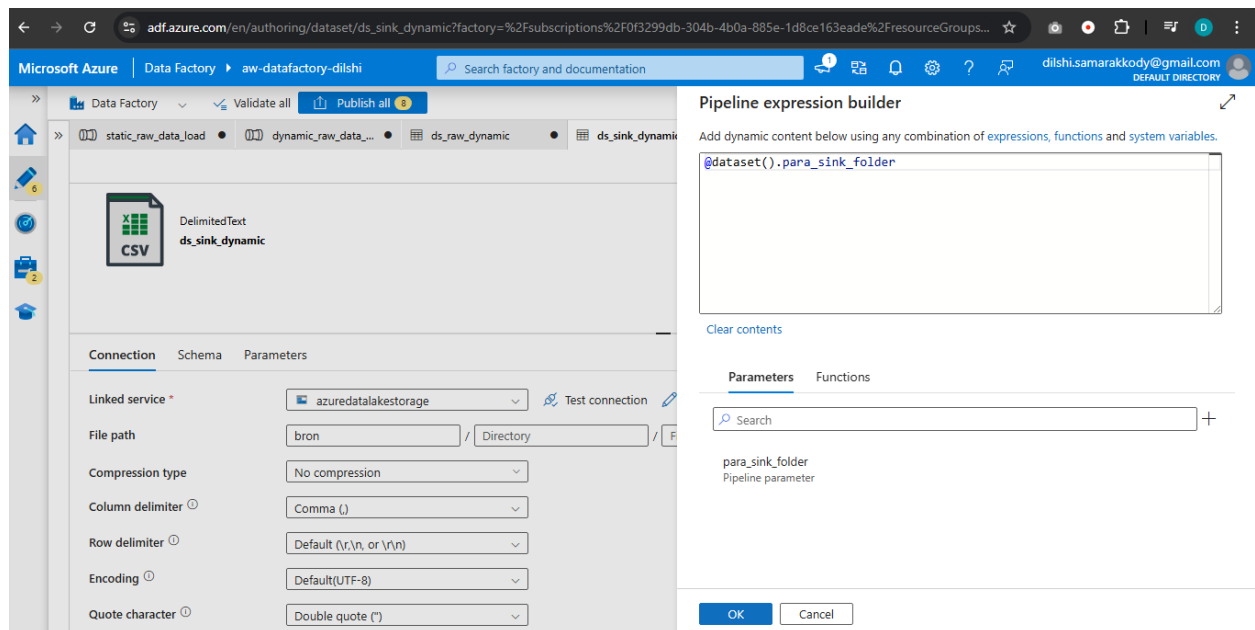
- Name: `para_rel_url`
- Type: `String` (with a dropdown arrow).
- Default value: (empty text box).

At the bottom of the dialog box are 'Save' and 'Cancel' buttons.



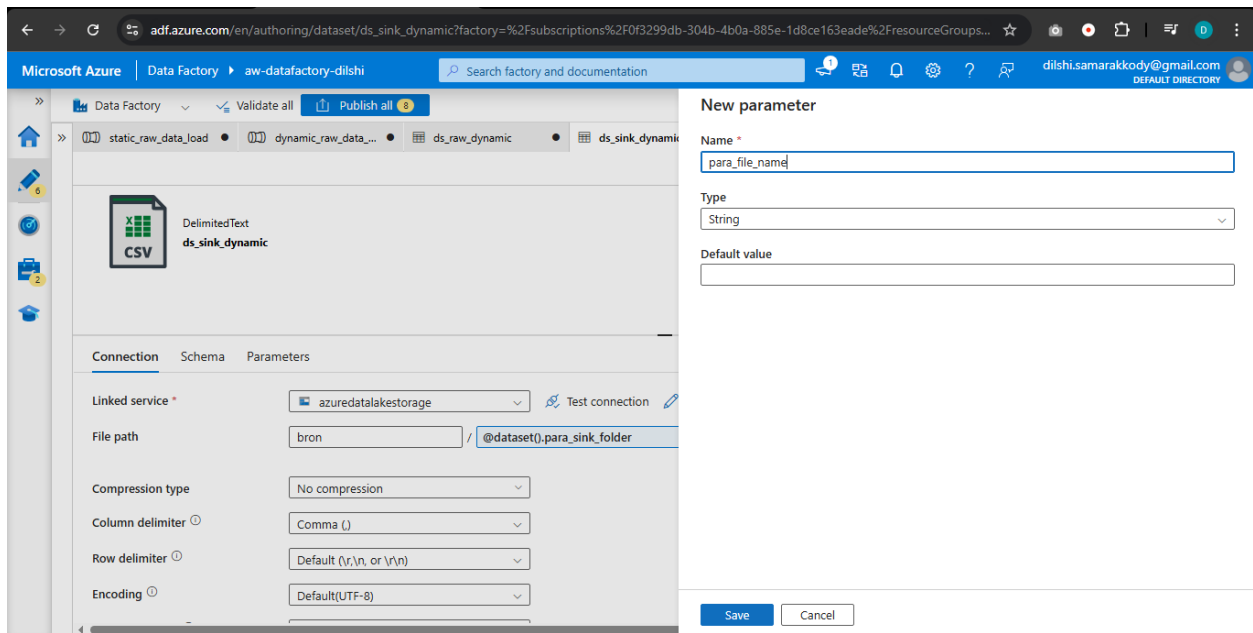
Step 8

Create dynamic parameter for sink folder



Step 9

Create dynamic parameter for sink filename



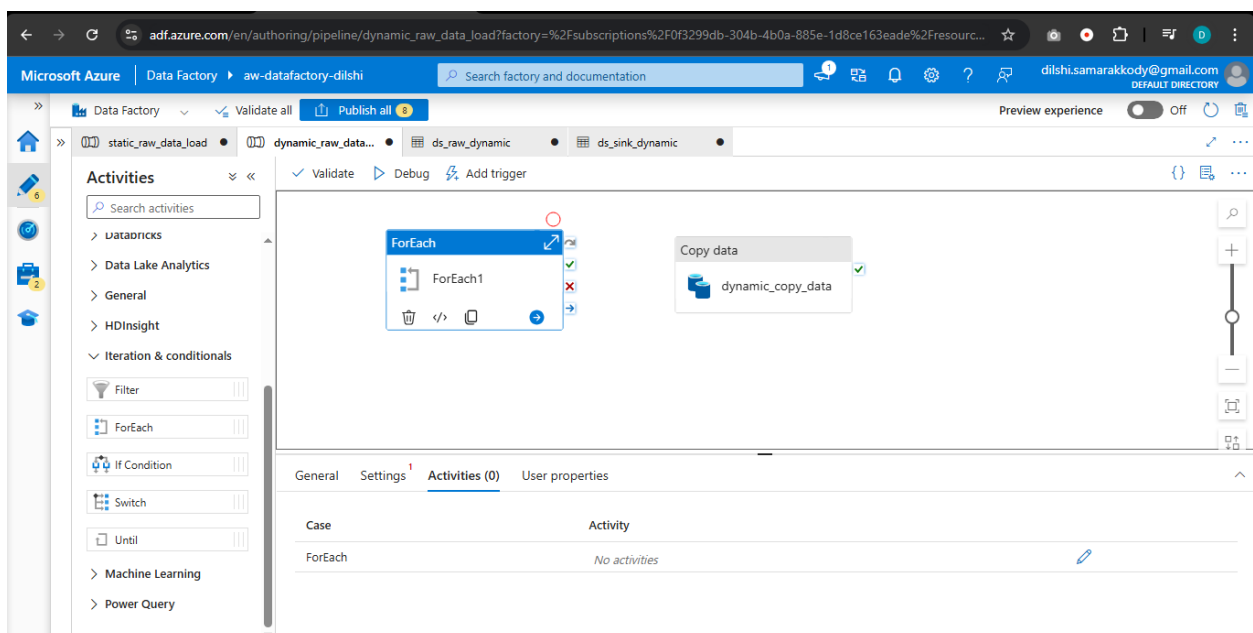
ForEach Loop Implementation and provide values for the dynamic parameters

The ForEach activity enables parallel processing of multiple data sources:

- Iterates through lookup results
- Executes child activities for each item
- Supports sequential and parallel execution modes
- Provides item-level error handling

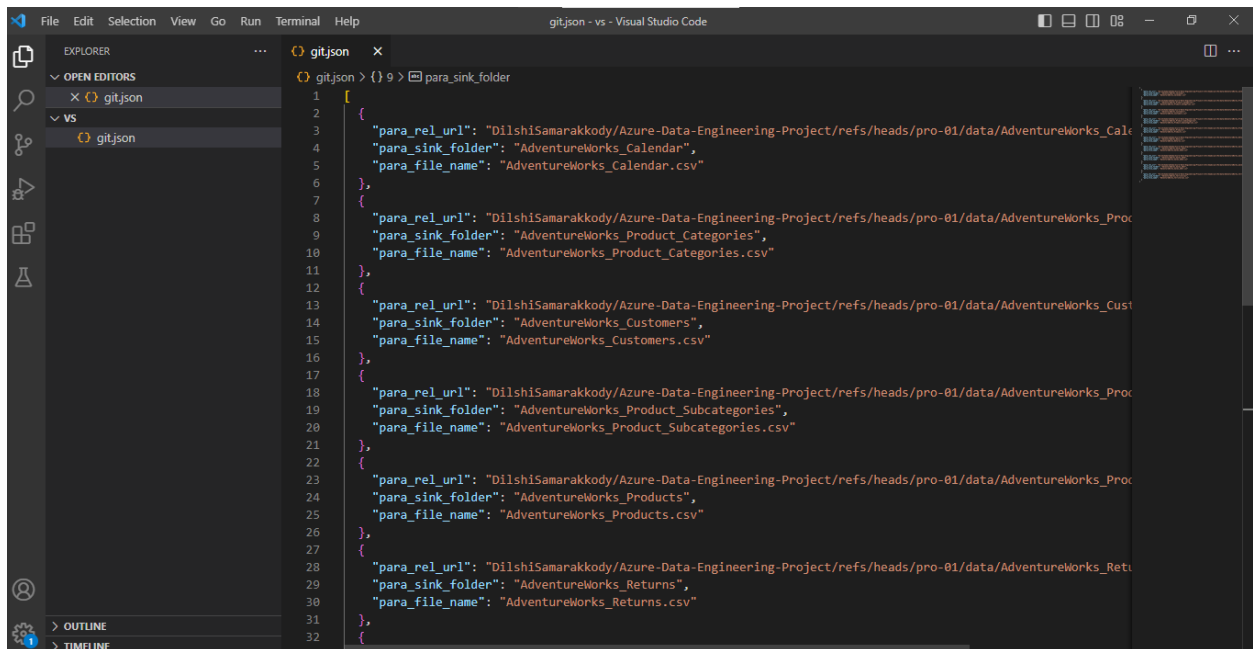
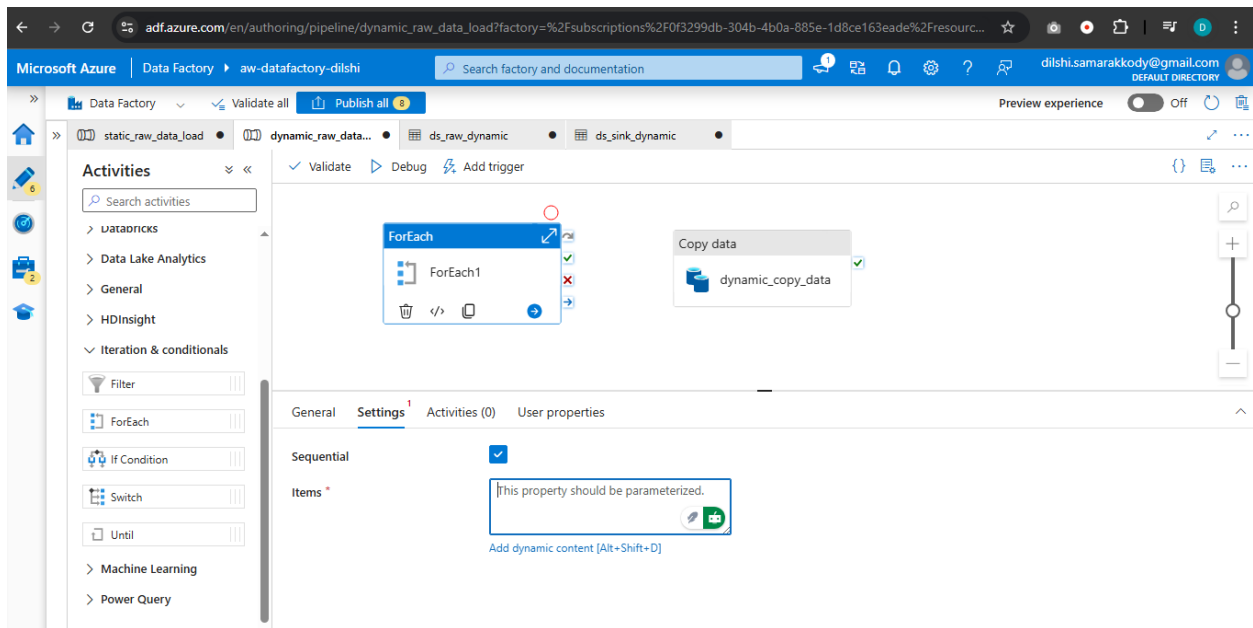
Step 10

Create ForEach activity



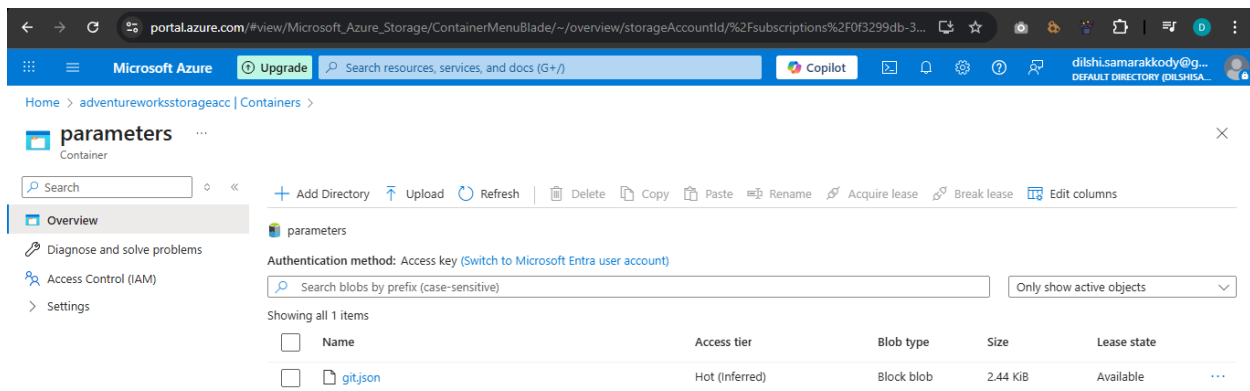
Step 11

Pass a sequential list of values as an array in json format



Step 12

Create a folder and upload the json file



Lookup Activity Configuration

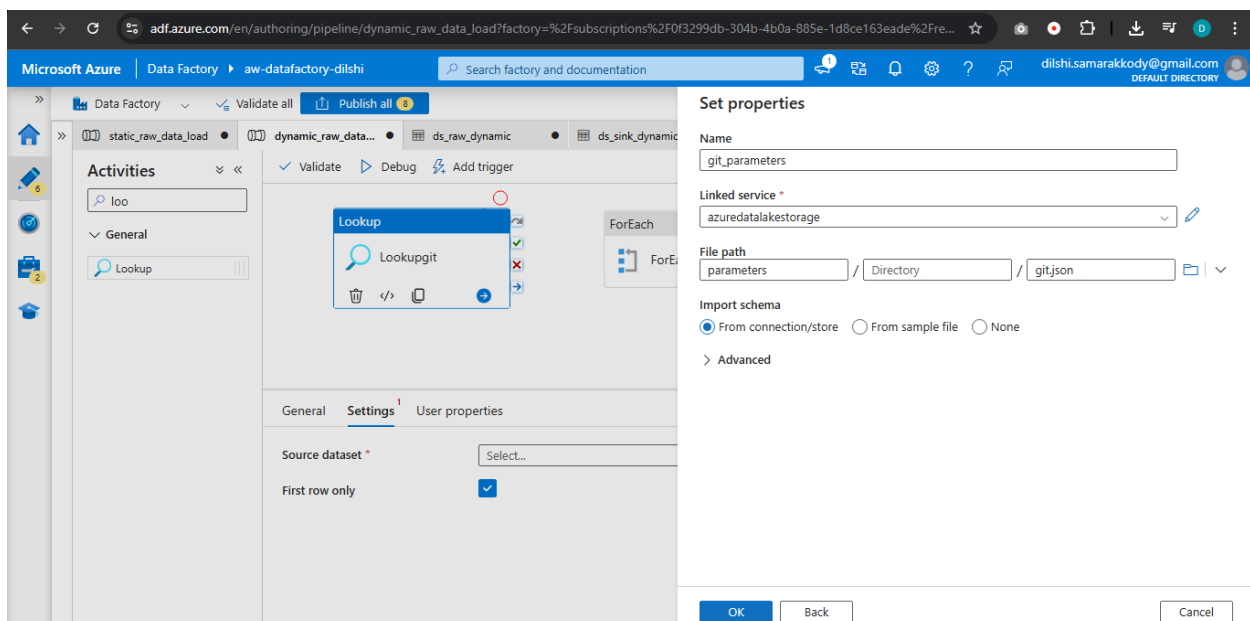
The Lookup activity serves as the foundation for dynamic processing:

- Queries configuration tables or files
- Retrieves source and destination information
- Provides parameters for subsequent activities
- Enables conditional processing logic

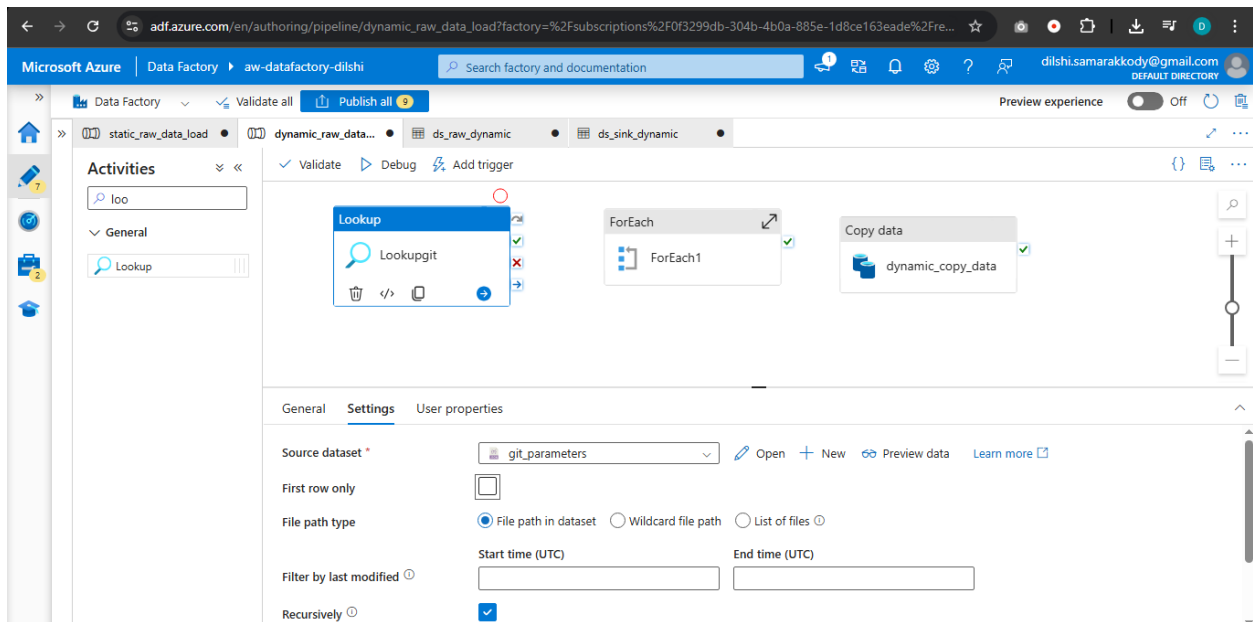
Step 13

Create a Lookup activity to check the output

Select ADLS and Json format for data source

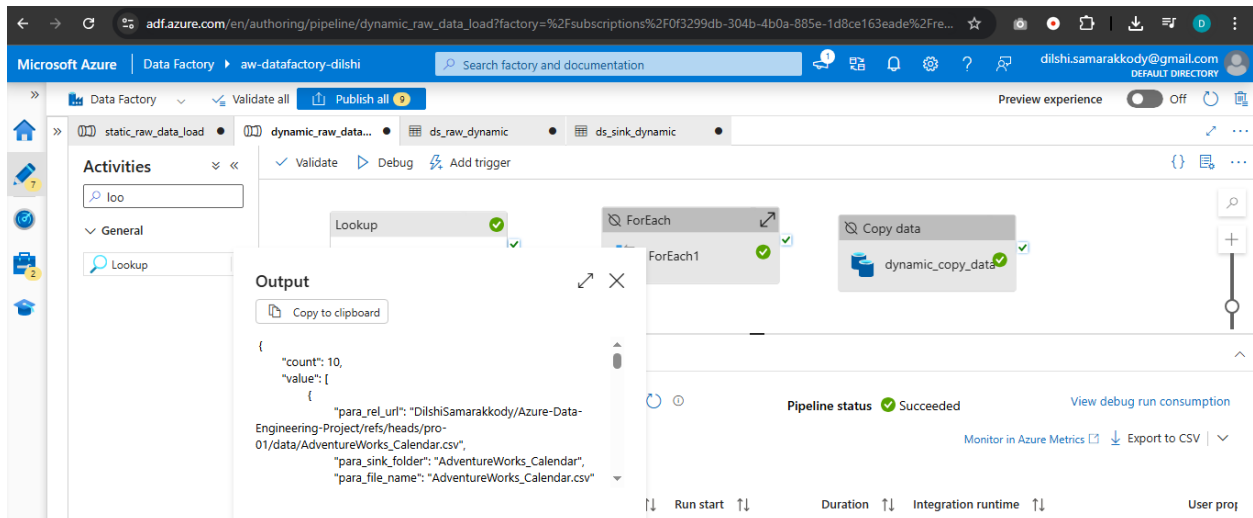


Uncheck first row only



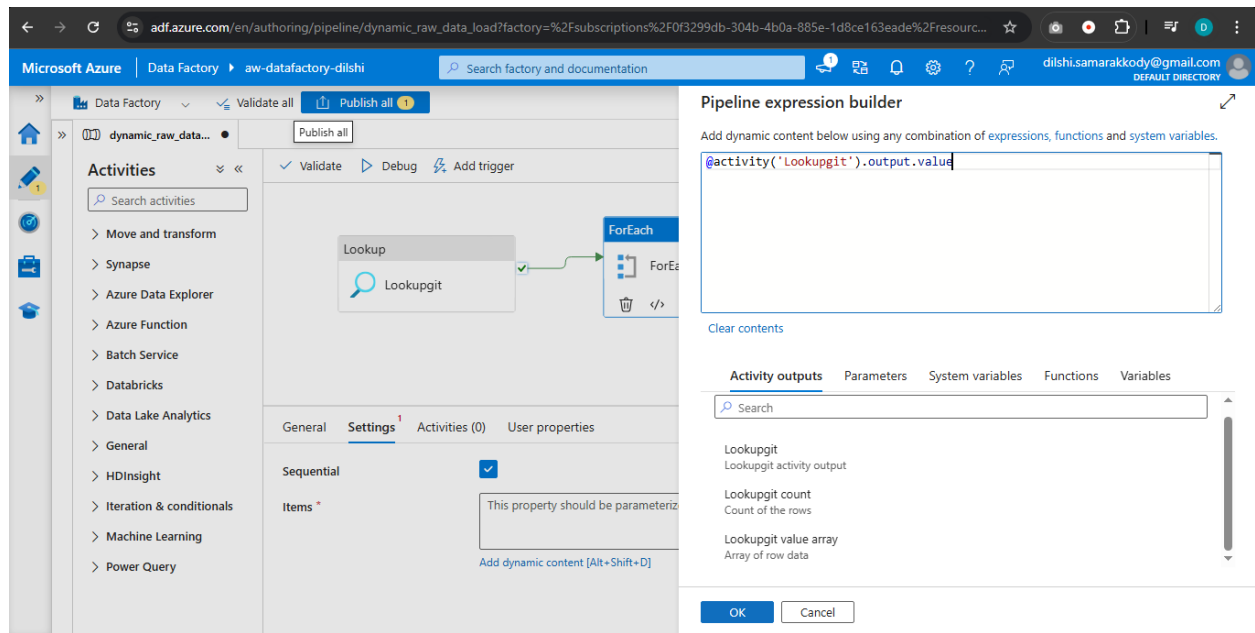
Step 14

Deactivate other activities and Debug only Lookup activity. Check the output.



Step 15

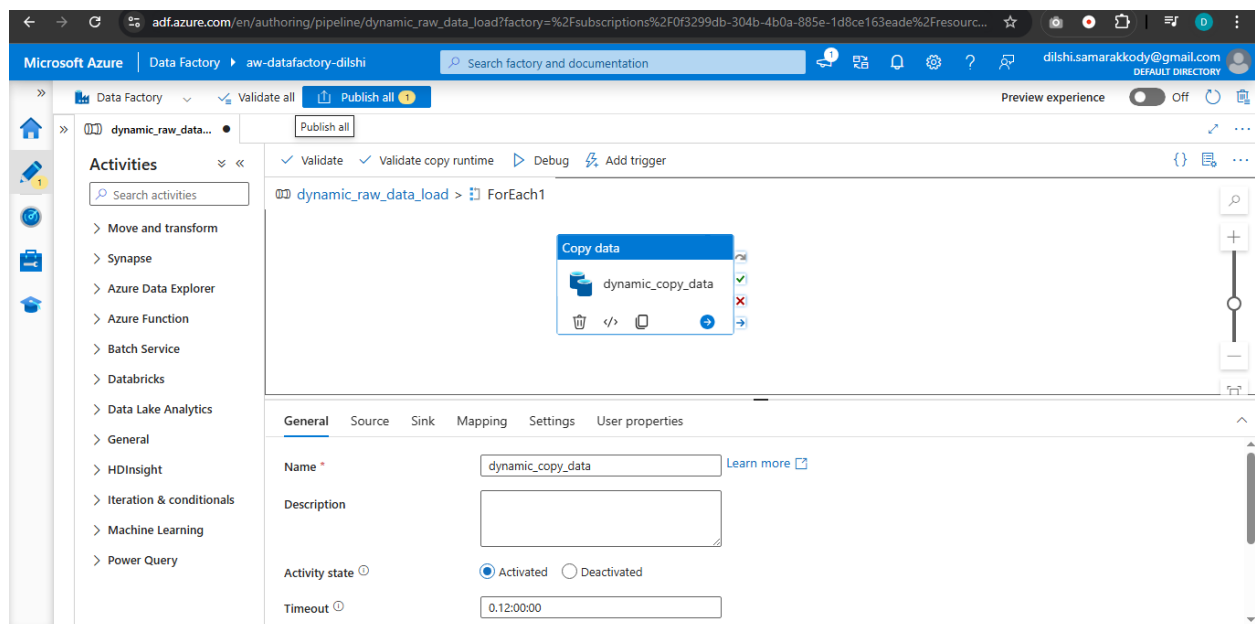
There are 3 outputs choose only value array.



Step 16

Activate other activities. Connect the Lookup activity to ForEach.

Select ForEach activity and go to activities and copy and paste Copy data activity.



Step 17

Pass the values for the source and sink dynamic parameters

The image shows two screenshots from the Microsoft Azure Data Factory interface. The top screenshot displays the 'Pipeline expression builder' dialog box, which is used to add dynamic content to a pipeline. The 'ForEach iterator' tab is selected, and the expression '@item().para_rel_url' is entered in the text area. The bottom screenshot shows the 'Sink dataset' configuration for 'ds_sink_dynamic'. The 'Dataset properties' section is expanded, showing a table with the following data:

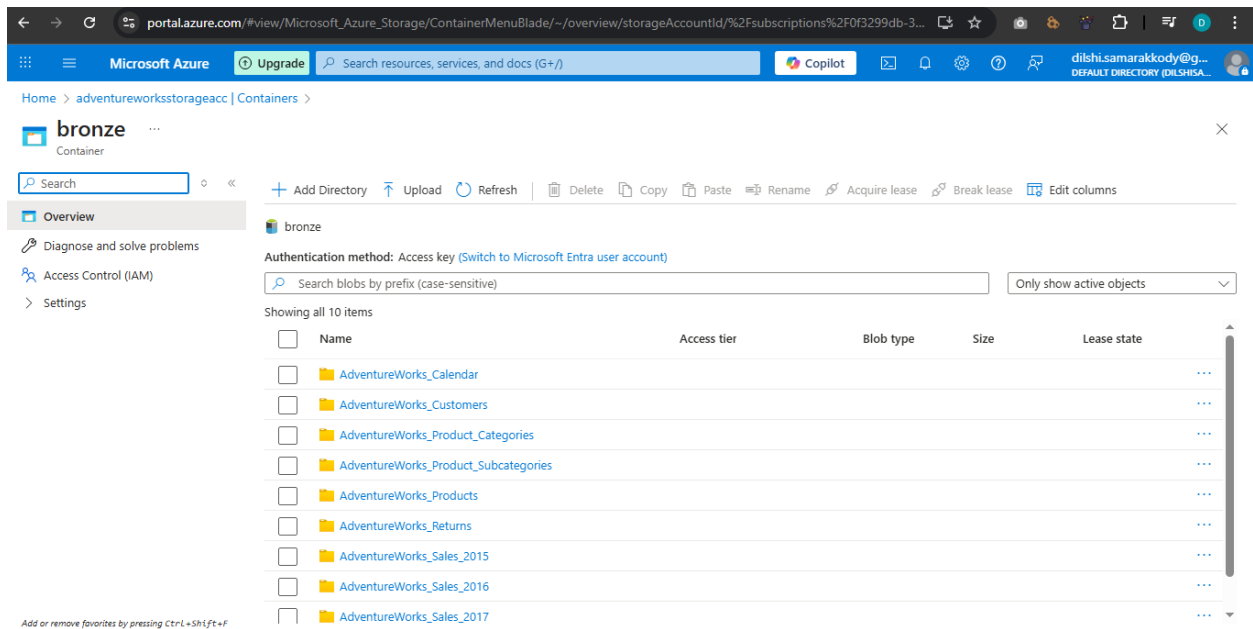
Name	Value	Type
para_sink_folder	@item().para_sink_folder	String
para_file_name	@item().para_file_name	String

Steps 18

Debug the dynamic pipeline

The image shows the 'Pipeline run results' page in the Microsoft Azure Data Factory interface. The pipeline run ID is 'a38ed721-f963-4b97-8dfa-66a9f21ee222'. The pipeline status is 'Succeeded'. The 'Output' tab is selected, showing a table with the following data:

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime	User p
Lookup	Lookupgit					
ForEach	ForEach1					
dynamic_c	dynamic_c					

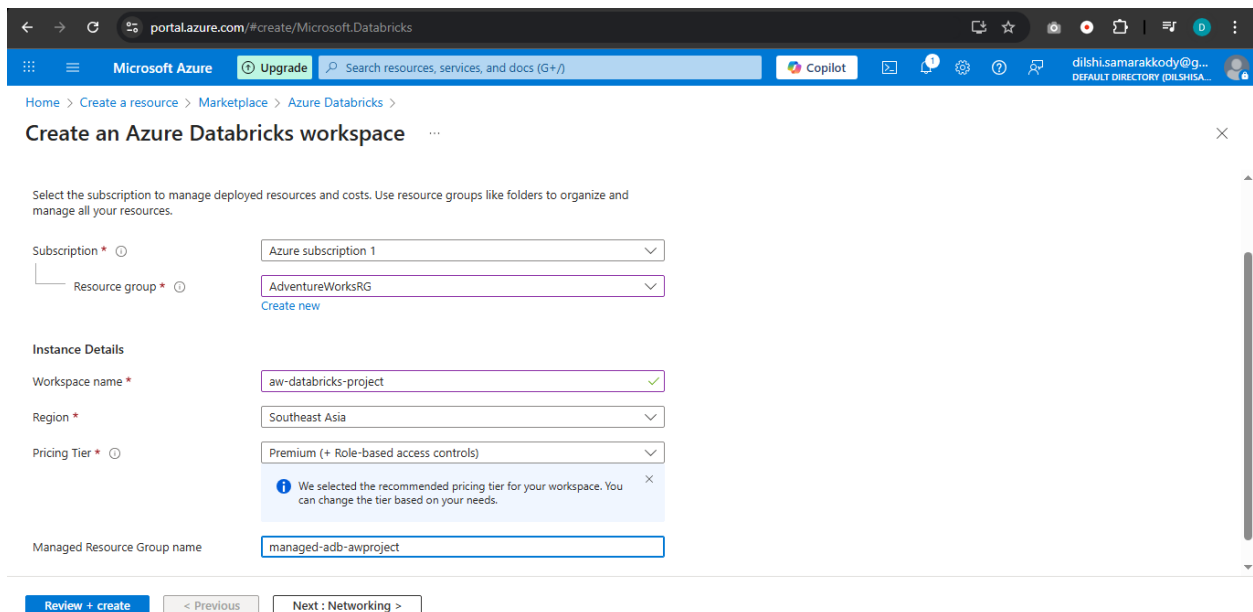


3. Azure Databricks Integration

Azure Databricks provides advanced analytics and machine learning capabilities.

Step 19

Create Databricks workspace.



Step 20

Create a new cluster for workspace

The screenshot shows the Databricks web interface. The left sidebar contains navigation options: Workspace, Recents, Catalog, Jobs & Pipelines, Compute (selected), Marketplace, SQL, SQL Editor, Queries, Dashboards, Genie, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, and Data Ingestion. The main area displays the configuration for a cluster named 'aw-project-cluster-dilshi'. The 'Configuration' tab is active, showing settings for Policy (Unrestricted), Access mode (No isolation shared), Performance (Databricks Runtime Version 14.3 LTS), and Node type (Standard_D4s_v3, 16 GB Memory, 4 Cores). A 'Summary' panel on the right shows details: 1 Driver, 16 GB Memory, 4 Cores, Runtime 14.3.x-scala2.12, and Standard_D4s_v3 0.75 DBU/h. The 'Terminate after' checkbox is checked, set to 120 minutes of inactivity.

4. Authentication and Authorization

Create an application that accesses data in ADLS and provide credential to Azure Databricks.

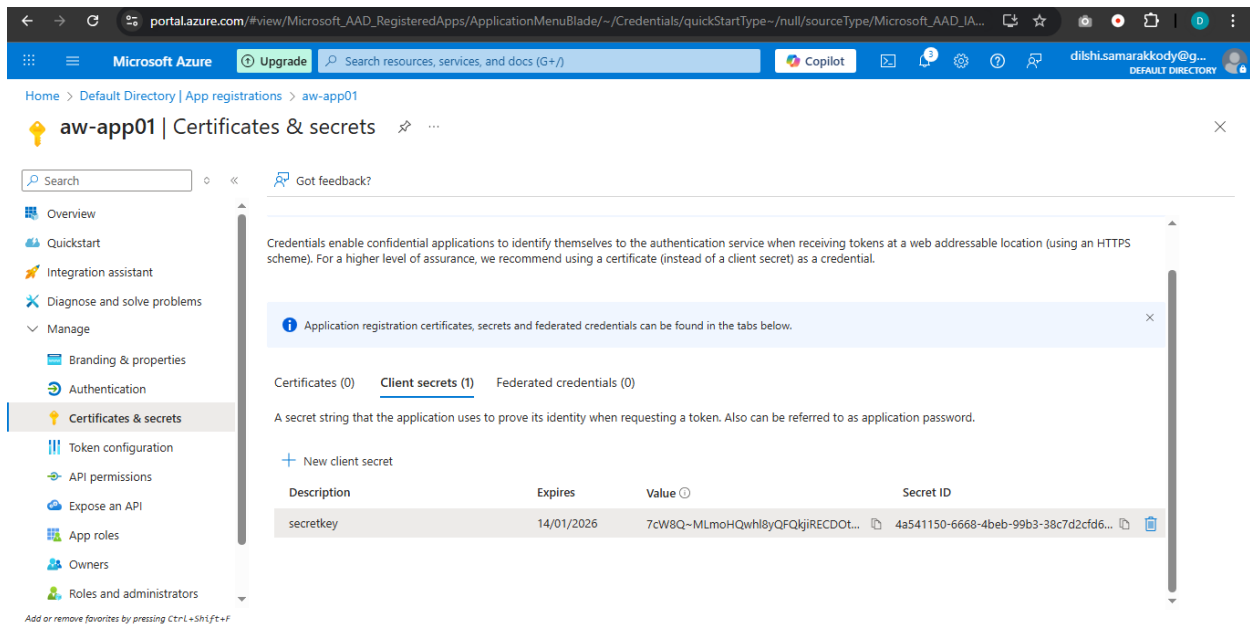
Step 21

Go to App Registration and create a new app

The screenshot shows the Microsoft Azure portal's 'Register an application' page. The breadcrumb trail is 'Home > Default Directory | App registrations >'. The page title is 'Register an application'. The 'Name' field is labeled '* Name' and contains the text 'aw-app01'. Below this, the 'Supported account types' section asks 'Who can use this application or access this API?'. Four radio button options are listed: 'Accounts in this organizational directory only (Default Directory only - Single tenant)' (selected), 'Accounts in any organizational directory (Any Microsoft Entra ID tenant - Multitenant)', 'Accounts in any organizational directory (Any Microsoft Entra ID tenant - Multitenant) and personal Microsoft accounts (e.g. Skype, Xbox)', and 'Personal Microsoft accounts only'. A 'Help me choose...' link is provided. At the bottom, there is a link to 'Microsoft Platform Policies' and a 'Register' button.

Step 22

Create a Secret key



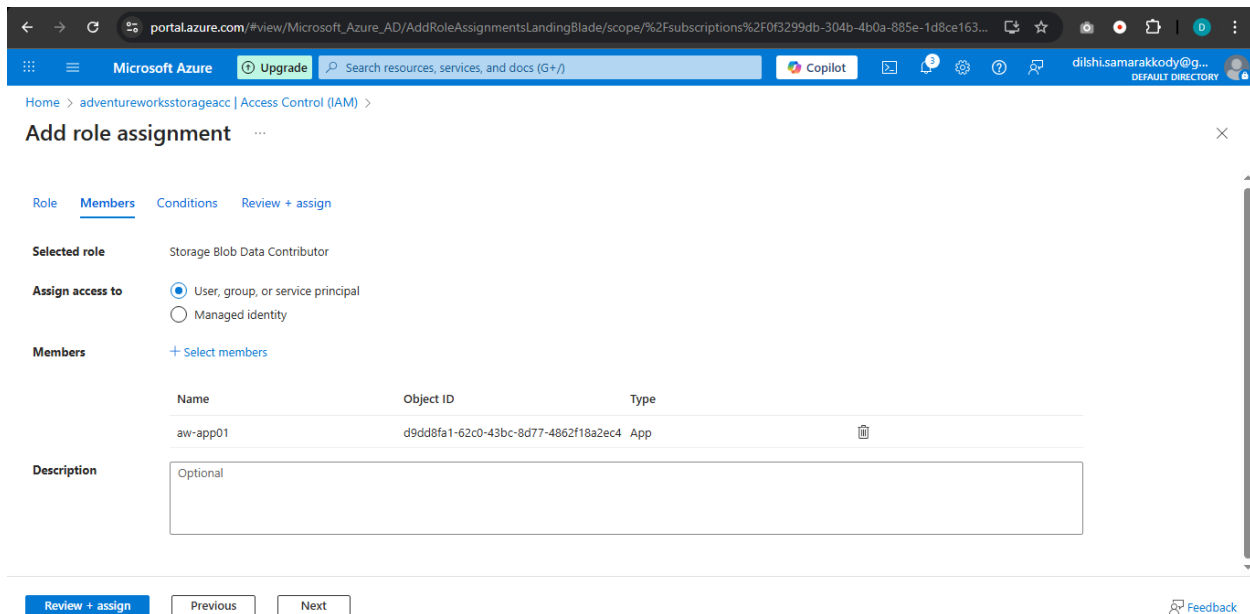
The screenshot shows the Microsoft Azure portal interface. The breadcrumb navigation is: Home > Default Directory | App registrations > aw-app01. The page title is "aw-app01 | Certificates & secrets". The left sidebar shows the "Certificates & secrets" option selected under the "Manage" section. The main content area has a heading "Certificates (0) Client secrets (1) Federated credentials (0)". Below this, it says "A secret string that the application uses to prove its identity when requesting a token. Also can be referred to as application password." There is a "+ New client secret" button. A table lists the existing secret:

Description	Expires	Value	Secret ID
secretkey	14/01/2026	7cW8Q~MLmoHQwhl8yQFQkjRECDOT...	4a541150-6668-4beb-99b3-38c7d2cfd6...

Assign a role to the application that can access the ADLS.

Step 23

Go to storage account and select Access control (IAM). Add a role assignment and select Storage Blob Contributor which gives both read and write permission to ADLS

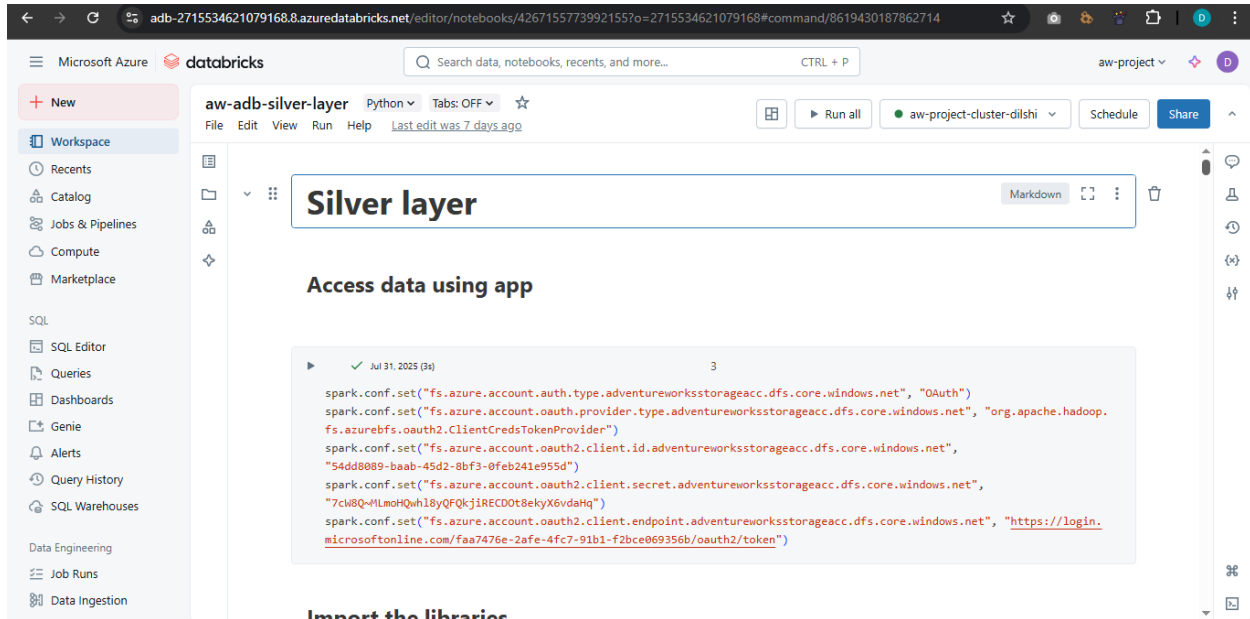


The screenshot shows the "Add role assignment" page in the Microsoft Azure portal. The breadcrumb navigation is: Home > adventureworksstorageacc | Access Control (IAM) >. The page title is "Add role assignment". The "Role" tab is selected, showing "Storage Blob Data Contributor" as the "Selected role". Under "Assign access to", the "User, group, or service principal" option is selected. The "Members" section shows a table with one member:

Name	Object ID	Type
aw-app01	d9dd8fa1-62c0-43bc-8d77-4862f18a2ec4	App

Below the table is a "Description" field with the text "Optional". At the bottom, there are buttons for "Review + assign", "Previous", and "Next".

5. Data access using app



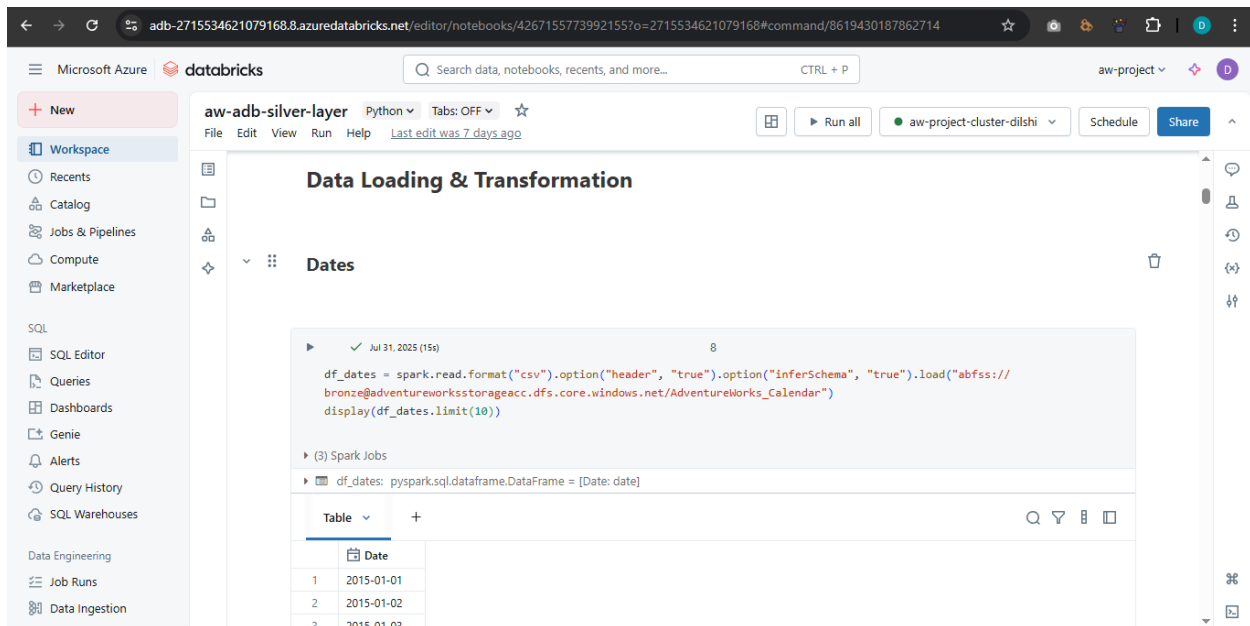
The screenshot shows the Databricks interface for a workspace named 'aw-adb-silver-layer'. The notebook is titled 'Silver layer' and contains a section 'Access data using app'. The code in the notebook sets up Spark configuration for Azure authentication and connects to the 'AdventureWorks' database. The code is as follows:

```
spark.conf.set("fs.azure.account.auth.type.adventureworksstorageacc.dfs.core.windows.net", "OAuth")
spark.conf.set("fs.azure.account.oauth.provider.type.adventureworksstorageacc.dfs.core.windows.net", "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider")
spark.conf.set("fs.azure.account.oauth2.client.id.adventureworksstorageacc.dfs.core.windows.net", "54dd8089-baab-45d2-8bf3-0feb241e955d")
spark.conf.set("fs.azure.account.oauth2.client.secret.adventureworksstorageacc.dfs.core.windows.net", "7cW8Q~tLmoHQwh18yQFQkjIRECD0t8ekyX6vdaHq")
spark.conf.set("fs.azure.account.oauth2.client.endpoint.adventureworksstorageacc.dfs.core.windows.net", "https://login.microsoftonline.com/faa7476e-2afe-4fc7-91b1-f2bce069356b/oauth2/token")
```

Below the code, there is a section titled 'Import the libraries'.

Step 24

Implement the Silver Layer



The screenshot shows the Databricks interface for a workspace named 'aw-adb-silver-layer'. The notebook is titled 'Data Loading & Transformation' and contains a section 'Dates'. The code in the notebook reads data from a CSV file and displays the first 10 rows. The code is as follows:

```
df_dates = spark.read.format("csv").option("header", "true").option("inferSchema", "true").load("abfss://bronze@adventureworksstorageacc.dfs.core.windows.net/AdventureWorks_Calendar")
display(df_dates.limit(10))
```

Below the code, there is a table showing the first 3 rows of the data:

	Date
1	2015-01-01
2	2015-01-02
3	2015-01-03

6. Azure Synapse Analytics Implementation

Step 25

Create synapse workspace

portal.azure.com/#create/Microsoft.Synapse

Microsoft Azure Upgrade Search resources, services, and docs (G+/) Copilot dilshi.samarakkody@g... DEFAULT DIRECTORY (DILSHISA...)

Home > Create a resource > Marketplace > Azure Synapse Analytics >

Create Synapse workspace

Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name * aw-synapse-project-dilshi ✓

Region * (Asia Pacific) Southeast Asia ✓

Select Data Lake Storage Gen2 * ☒ From subscription ☐ Manually via URL

Account name * (New) defaultsynsesgdilshi ✓
[Create new](#)

File system name * (New) defaultfs ✓
[Create new](#)

☒ Assign myself the Storage Blob Data Contributor role on the Data Lake Storage Gen2 account to interactively query it in the workspace.

i We will automatically grant the workspace identity data access to the specified Data Lake Storage Gen2 account, using the [Storage Blob Data Contributor](#) role. To enable other users to use this storage account after you create it, you must assign them the role.

[Review + create](#) < Previous Next: Security >

portal.azure.com/#create/Microsoft.Synapse

Microsoft Azure Upgrade Search resources, services, and docs (G+/) Copilot dilshi.samarakkody@g... DEFAULT DIRECTORY (DILSHISA...)

Home > Create a resource > Marketplace > Azure Synapse Analytics >

Create Synapse workspace

Authentication

Choose the authentication method for access to workspace resources such as SQL pools. The authentication method can be changed later on. [Learn more](#)

Authentication method * ☒ Use both local and Microsoft Entra ID authentication ☐ Use only Microsoft Entra ID authentication

SQL Server admin login * sqladminuser

SQL Password * ✓

Confirm password ✓

System assigned managed identity permission

Select to grant the workspace network access to the Data Lake Storage Gen2 account using the workspace system identity. [Learn more](#)

☐ Allow network access to Data Lake Storage Gen2 account. **i** The selected Data Lake Storage Gen2 account does not restrict network access using any network access rules, or you selected a storage account manually via URL under Basics tab. [Learn more](#)

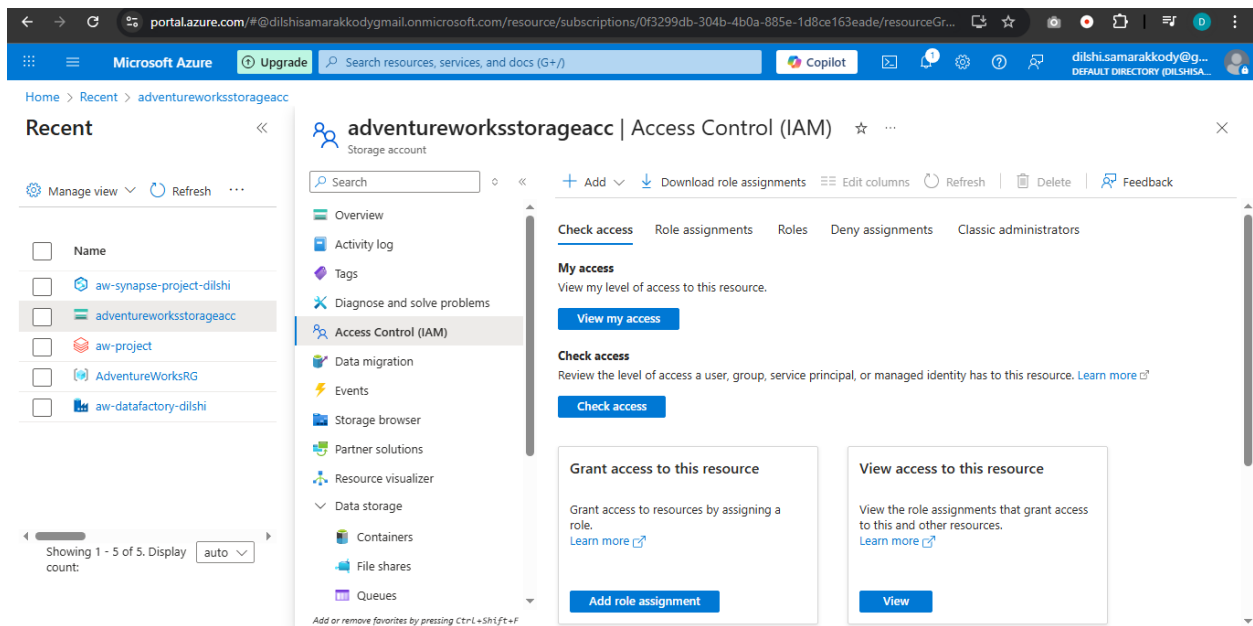
Workspace authentication

[Review + create](#) < Previous Next: Networking >

Allow Azure Synapse Analytics to access data stored in ADLS using managed identity

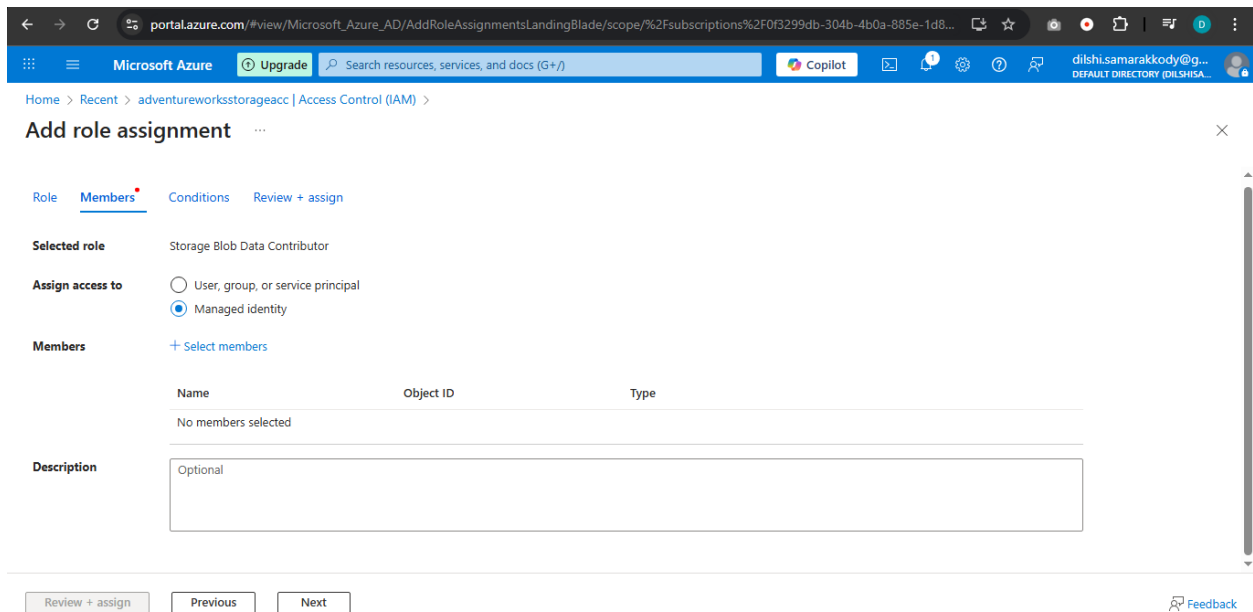
Step 26

Create Managed identity



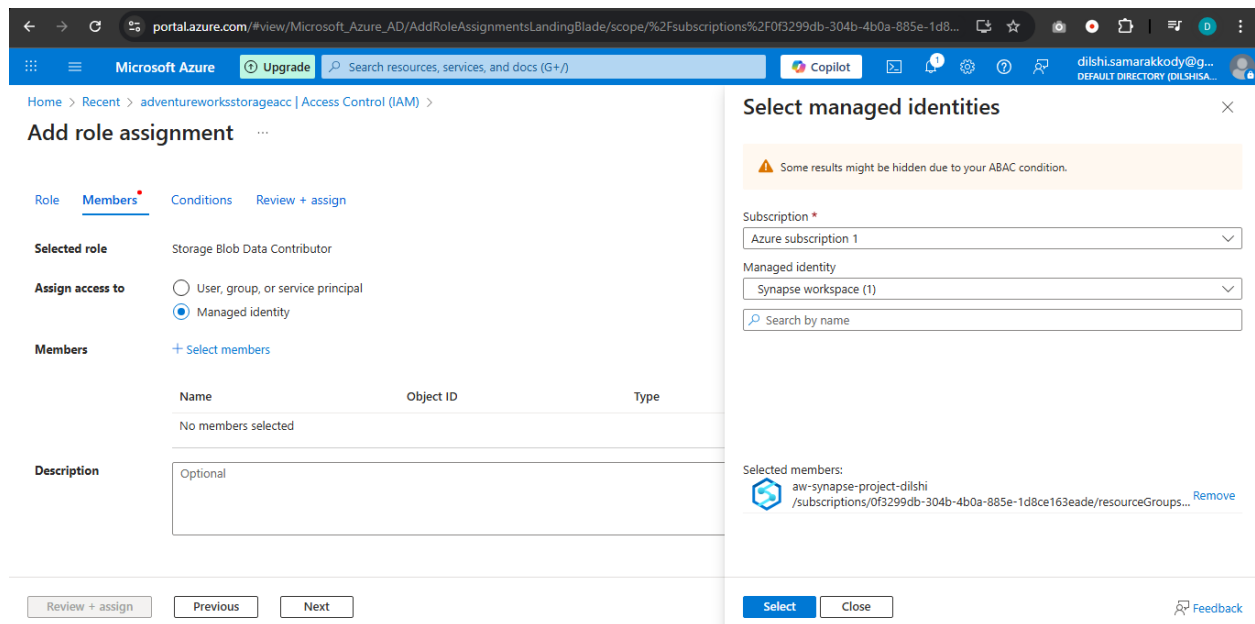
Step 27

Select managed identity



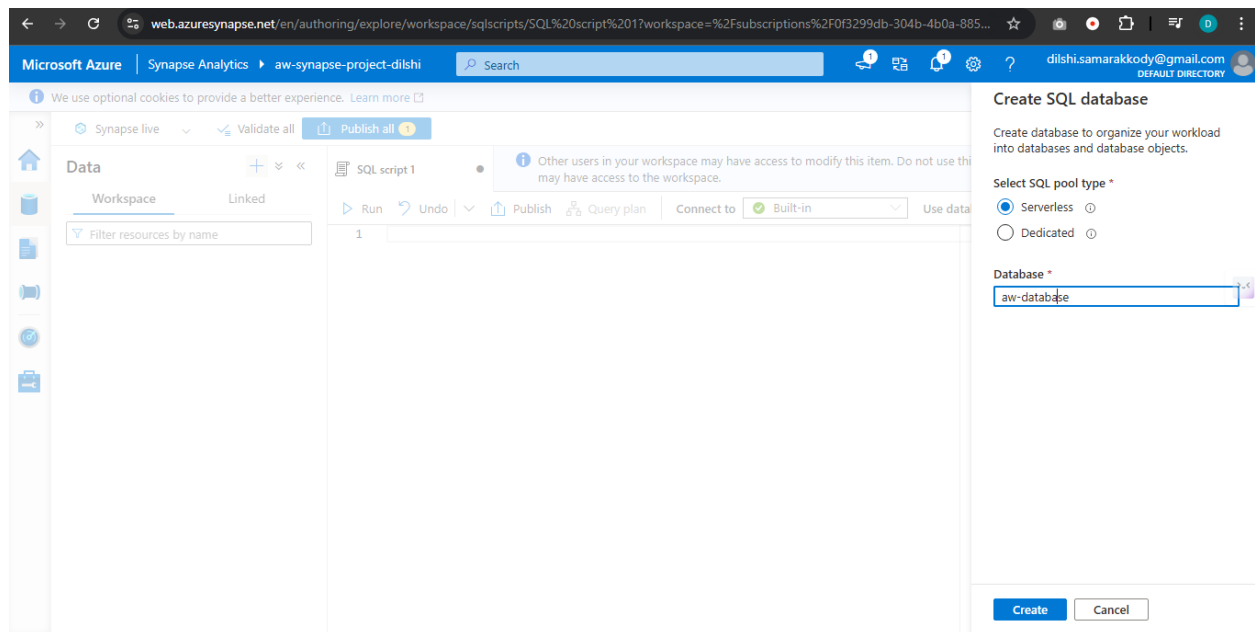
Step 28

Assign a role on Managed Identity



Step 29

Create Serverless database



Step 30

Go to storage account and create role assign for me to access data in ADLS

portal.azure.com/#view/Microsoft_Azure_AD/AddRoleAssignmentsLandingBlade/scope/%2Fsubscriptions%2F03299db-304b-4b0a-885e-1d8...

Microsoft Azure Upgrade Search resources, services, and docs (G+)

Home > adventureworksstorageacc | Access Control (IAM) >

Add role assignment

Role **Members** Conditions Review + assign

Selected role Storage Blob Data Contributor

Assign access to ☒ User, group, or service principal ☐ Managed identity

Members + Select members

Name	Object ID	Type
No members selected		

Description Optional

Review + assign Previous Next

Select members

Search: di

Selected members:

Dilshi Samarakkody(Guest)
dilshi.samarakkody_gmail.com#EXT#@dilshisamarakkodygmail.onmicrosoft.c...

Select Close

Check data preview

web.azure.synapse.net/en/authoring/analyze/sqlscripts/SQL%20script%201?workspace=%2Fsubscriptions%2F03299db-304b-4b0a-885e-1d8ce163...

Microsoft Azure Synapse Analytics aw-synapse-project-dilshi Search

We use optional cookies to provide a better experience. Learn more

Accept Reject More options

Synapse live Validate all Publish all

Develop

Filter resources by name

- SQL scripts 1
 - SQL script 1

```
1 SELECT
2 * FROM
3 OPENROWSET (
4     BULK'https://adventureworksstorageacc.blob.core.windows.net/silver/AdventureWorks_Calender/',
5     FORMAT = 'PARQUET'
6 ) as a query01
```

Run Undo Publish Query plan Connect to Built-in Use database aw-database

Results

Messages

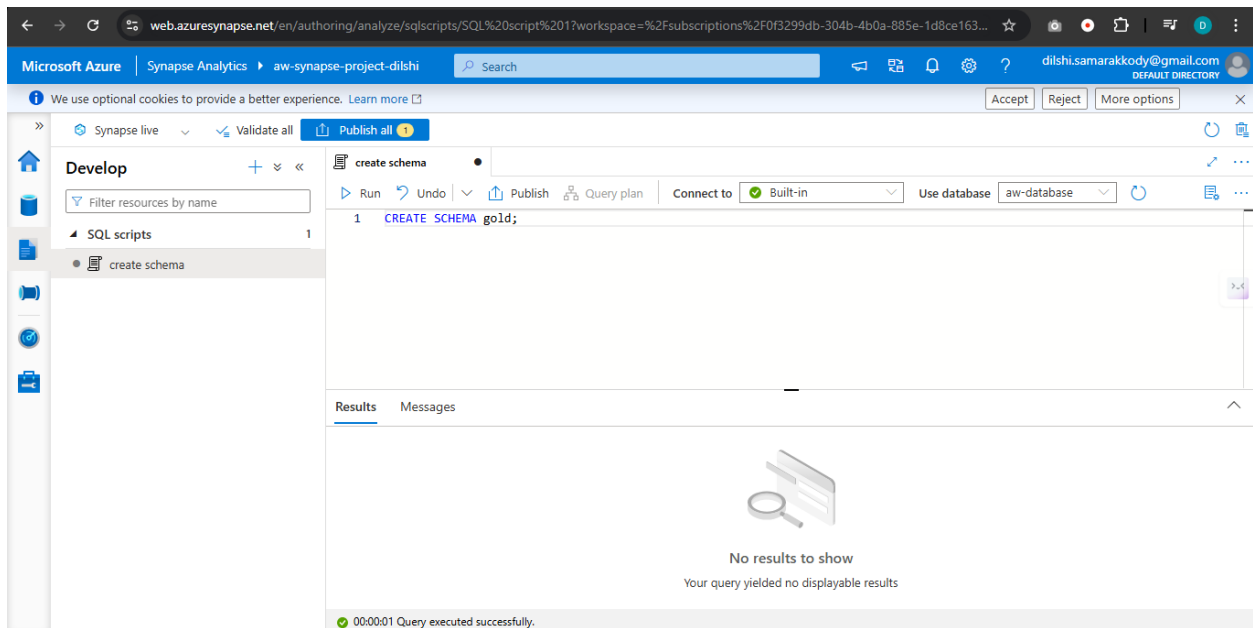
View Table Chart Export results

Date	Month	Year
2015-01-01	1	2015
2015-01-02	1	2015

00:00:06 Query executed successfully.

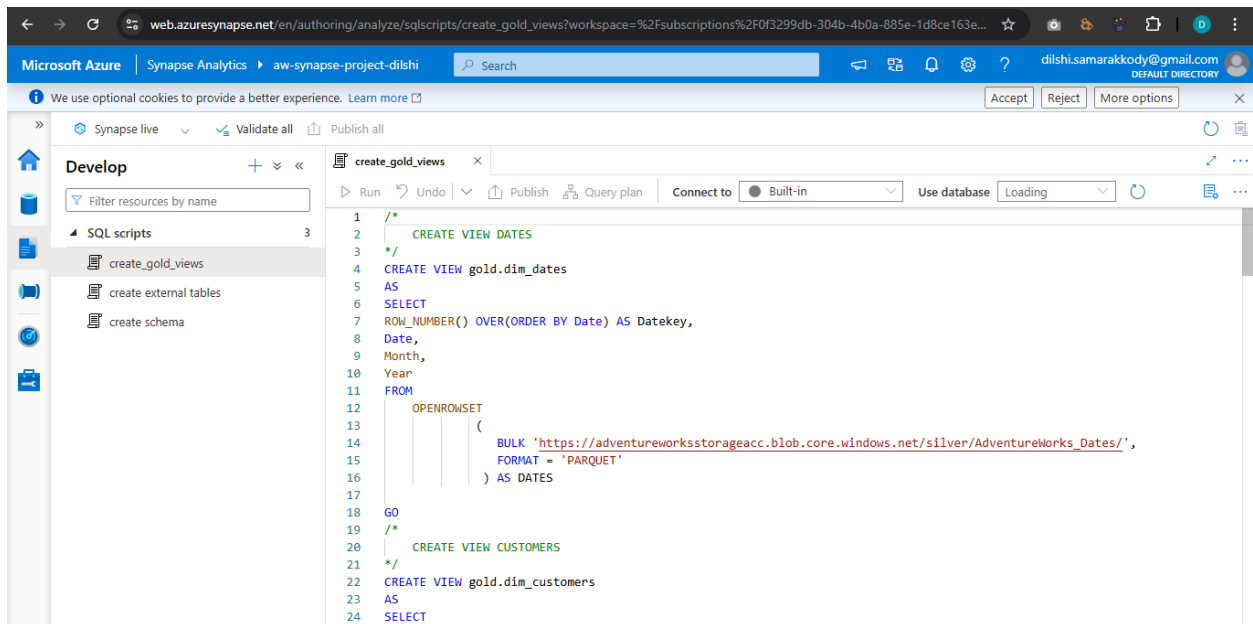
Step 31

Create a schema



Step 32

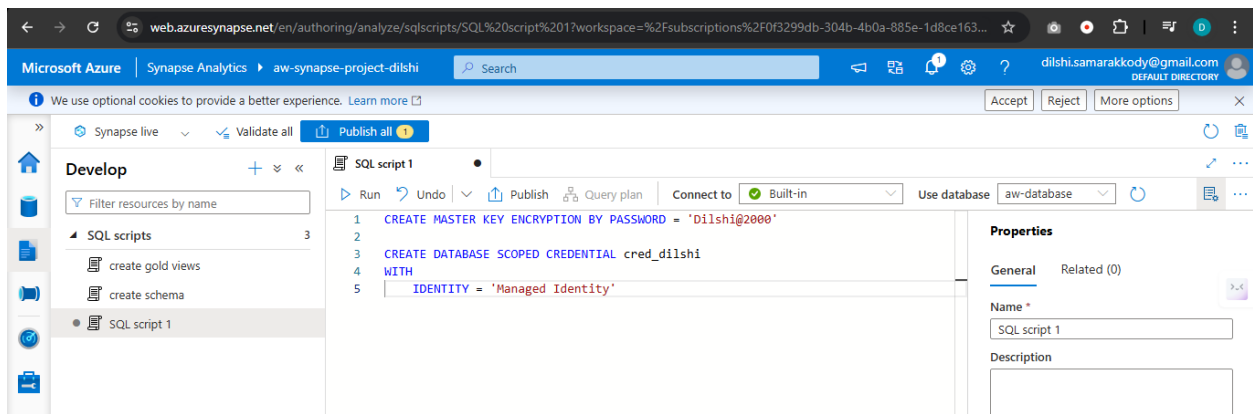
Create views for dimension and fact tables



Create master key if not available

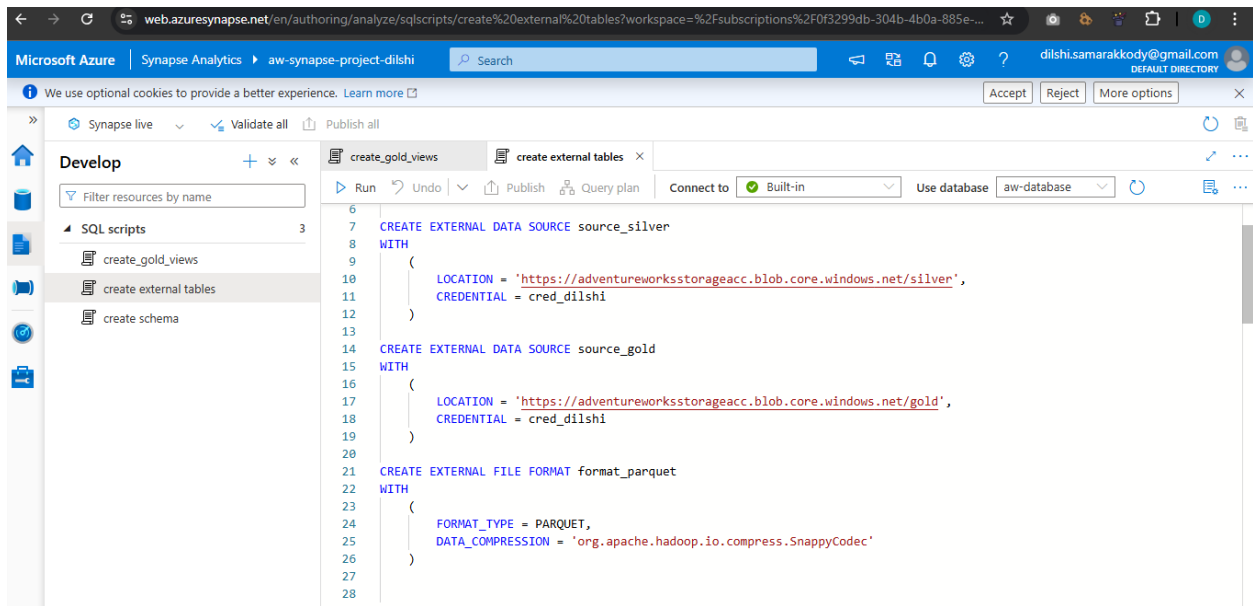
Step 33

Create database scoped credential

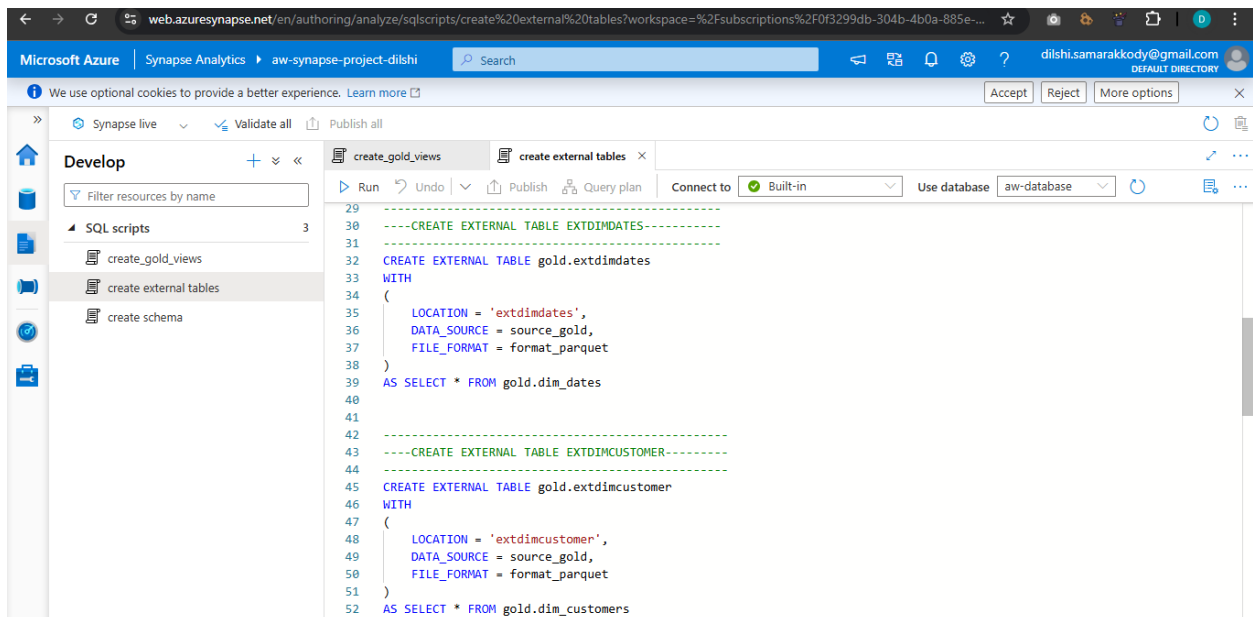


Step 34

Creating external data source and create external file format



Create external table



Check data in gold layer

The screenshot shows the Microsoft Azure Synapse Analytics web interface. The left sidebar displays the 'Develop' section with a list of resources: 'create_gold_views', 'create external tables', 'create schema', and 'SQL script 1'. The main area shows the 'SQL script 1' editor with a query: `SELECT * FROM gold.fact_sales`. Below the editor, the 'Results' tab is active, displaying a table of data. The table has columns: OrderNumber, OrderDate, OrderLineItem, OrderQuantity, StockDate, ProductKey, CustomerKey, and TerritoryKey. The data is filtered to show 5 rows. A status bar at the bottom indicates '00:00:17 Query executed successfully.'

OrderNumber	OrderDate	OrderLineItem	OrderQuantity	StockDate	ProductKey	CustomerKey	TerritoryKey
SO45086	2015-01-02	1	1	2001-12-18	47	7598	9
SO45085	2015-01-02	1	1	2001-10-09	45	7597	9
SO45085	2015-01-02	1	1	2001-10-09	45	7597	9
SO45093	2015-01-03	1	1	2001-10-03	45	7753	9
SO45093	2015-01-03	1	1	2001-10-03	45	7753	9

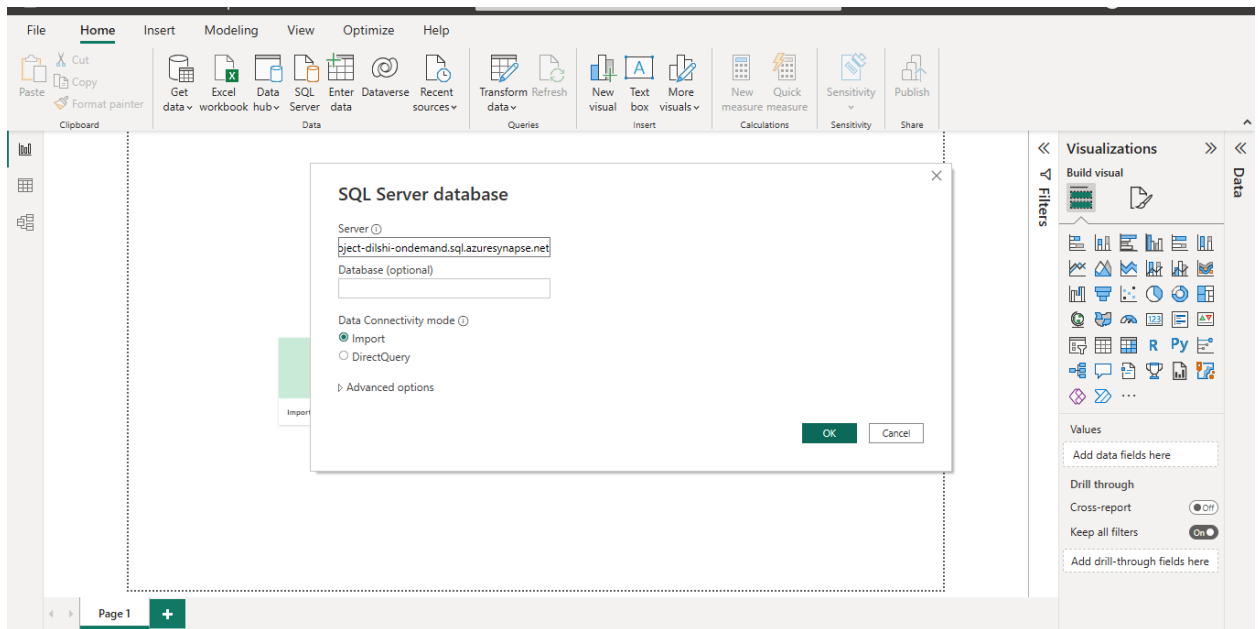
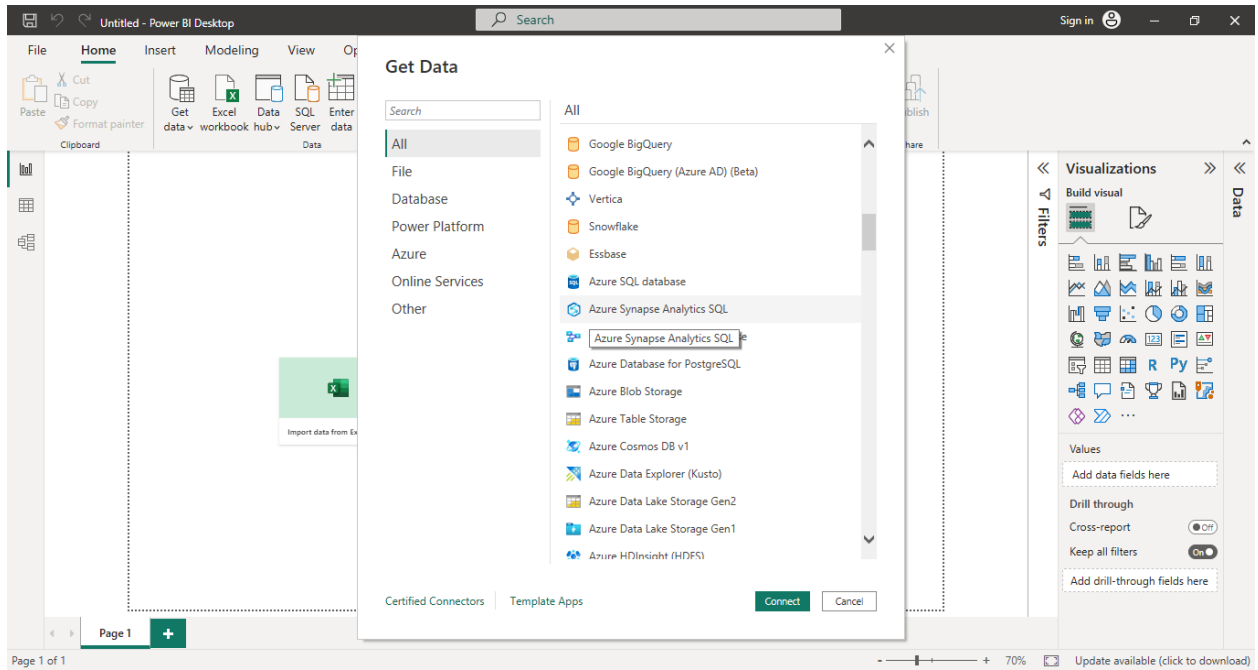
Copy the serverless sql end point and create connection with power bi

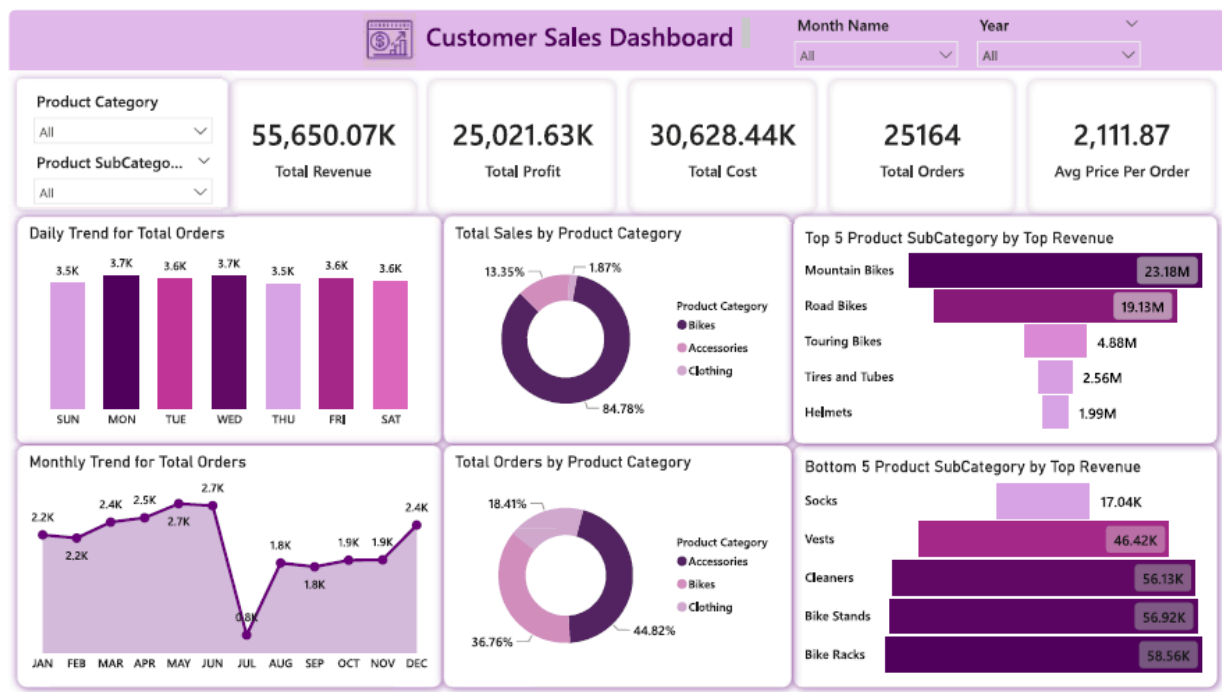
The screenshot shows the Microsoft Azure portal interface for the 'aw-synapse-project-dilshi' Synapse workspace. The left sidebar displays the 'Overview' section with a list of resources: 'Activity log', 'Access control (IAM)', 'Tags', 'Diagnose and solve problems', 'Resource visualizer', 'Settings', 'Analytics pools', 'Security', 'Monitoring', 'Automation', and 'Help'. The main area shows the 'Essentials' section, which displays various configuration details for the workspace. A 'Copied' tooltip is visible over the 'Serverless SQL endpoint' value.

Property	Value
Resource group	AdventureWorksRG
Status	Succeeded
Location	Southeast Asia
Subscription	Azure subscription 1
Subscription ID	0f3299db-304b-4b0a-885e-1d8ce163eade
Managed virtual network	No
Managed identity object	3b626d3e-2ed3-4fe2-aff3-4f55b70565b5
Workspace web URL	https://web.azuresynapse.net/workspace=%2fsubs...
Tags	Add tags
Networking	Show firewall settings
Primary ADLS Gen2 acco...	https://defaultsynapsesgdilshi.dfs.core.windows.net
Primary ADLS Gen2 file s...	defaultfs
SQL admin username	sqladminuser
SQL Microsoft Entra admin	live.com#dilshi.samarakody@gmail.com
Dedicated SQL endpoint	aw-synapse-project-dilshi.sqlazuresynapse.net
Serverless SQL endpoint	aw-synapse-project-dilshi-ondemand.sqlazuresyn...
Development endpoint	https://aw-synapse-project-dilshi.dev.azuresynapse...

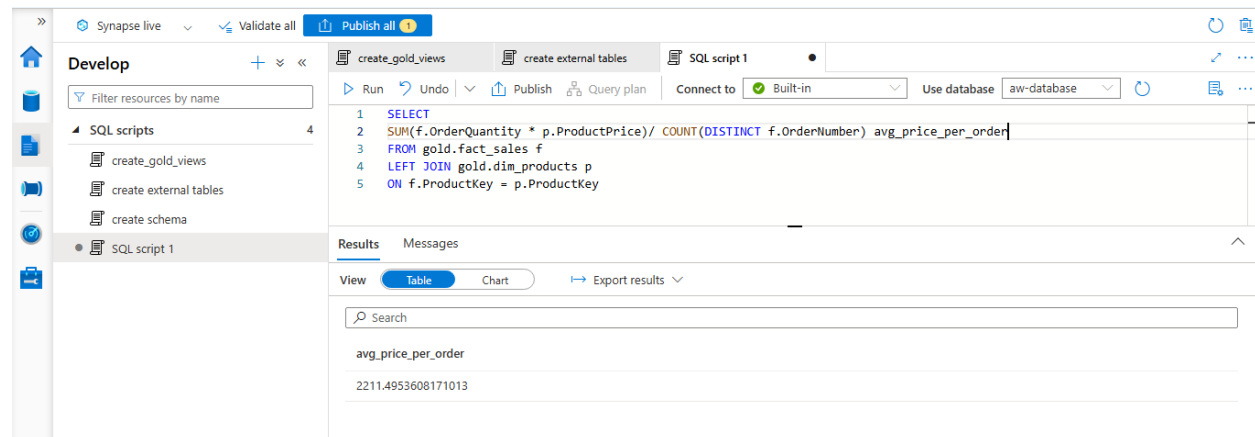
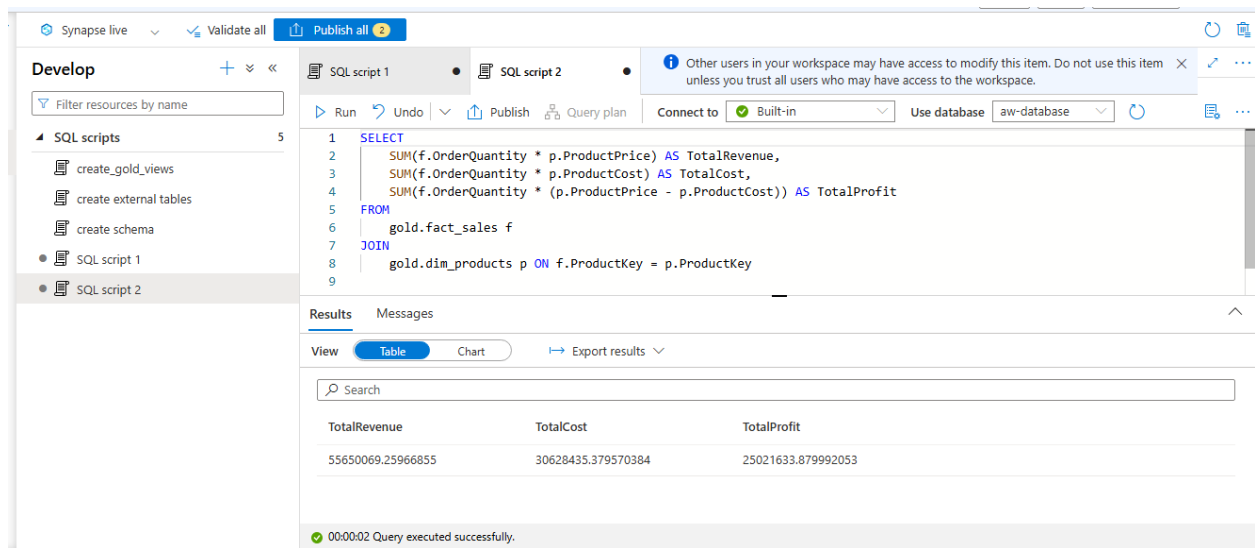
7. Power BI dashboard implementation

Load data from Azure Synapse into Power BI





KPI validation



We use optional cookies to provide a better experience. [Learn more](#) Accept Reject More options ×

Synapse live Validate all Publish all 2

Develop + ≡ «

Filter resources by name

SQL scripts 5

- create_gold_views
- create external tables
- create schema
- SQL script 1
- SQL script 2

SQL script 1

```
1 select
2 p.ProductCategory,
3 SUM(f.OrderQuantity * p.ProductPrice) as tot_price
4 from gold.fact_sales f
5 LEFT JOIN gold.dim_products p
6 ON f.ProductKey = p.ProductKey
7 group by p.ProductCategory
```

Results Messages

View Table Chart Export results

Search

ProductCategory	tot_price
Accessories	7427842.503841094
Bikes	47180481.20360897
Clothing	1041745.5521997635

00:00:02 Query executed successfully.

web.azure.synapse.net/en/authoring/analyze/sqlscripts/SQL%20script%201?workspace=%2Fsubscriptions%2F0f3299db-304b-4b0a-885e-1d8ce163... Search dilshi.samarakody@gmail.com DEFAULT DIRECTORY

Microsoft Azure | Synapse Analytics | aw-synapse-project-dilshi Search Accept Reject More options ×

Synapse live Validate all Publish all 2

Develop + ≡ «

Filter resources by name

SQL scripts 5

- create_gold_views
- create external tables
- create schema
- SQL script 1
- SQL script 2

SQL script 1

```
1 select
2 p.ProductSubCategory,
3 SUM(f.OrderQuantity * p.ProductPrice) as tot_price
4 from gold.fact_sales f
5 LEFT JOIN gold.dim_products p
6 ON f.ProductKey = p.ProductKey
7 group by p.ProductSubCategory
```

Results Messages

View Table Chart Export results

Search

ProductSubCategory	tot_price
Bike Racks	58560
Bike Stands	56922
Bottles and Cages	1701606.5899979584

00:00:02 Query executed successfully.