# Comprehensive Approach to the Azure Data Pipeline Lifecycle

## DOCUMENTATION

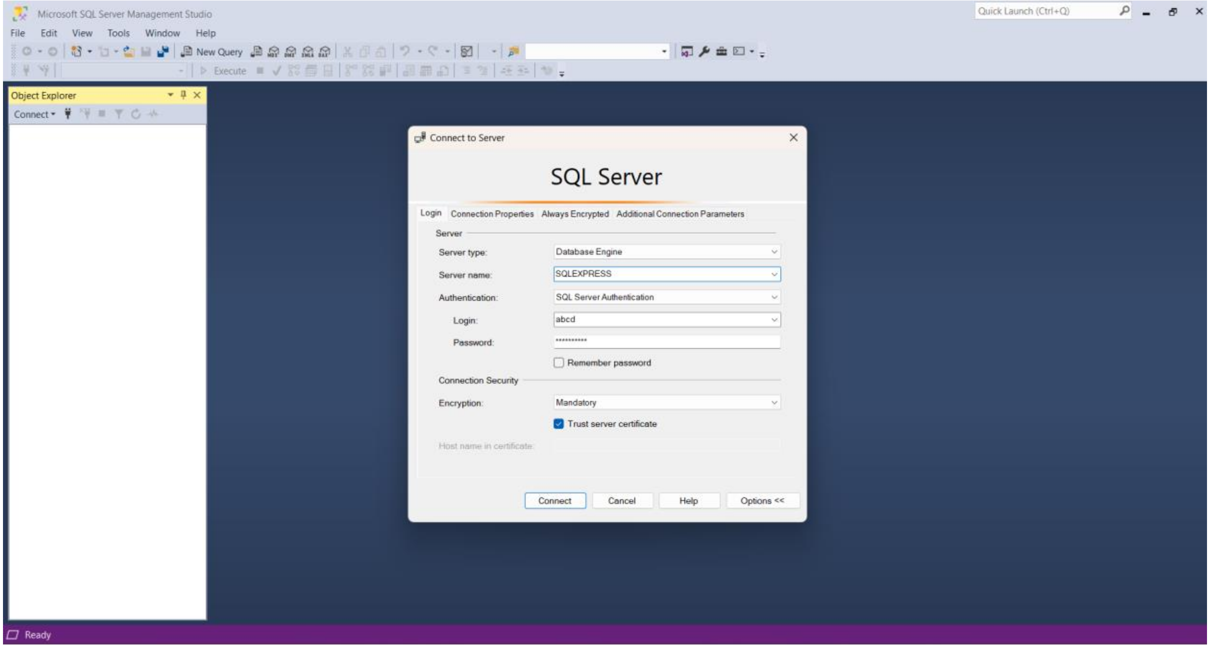
The pipeline is designed to ingest, transform, and load data from SQL Server into Azure Synapse Analytics and Power BI.

It contains three levels:

- Microsoft SQL Server and SSMS
- Azure Portal
- Power BI

## Level 1: Microsoft SQL Server and SSMS

- Open “SQL Server Management Studio
- Connect to Database engine using appropriate credentials.
- Upload data into your database.
- Query the tables for specific requirements (if necessary)



Server

Server type: Database Engine

Server name: SQLEXPRESS

Authentication: SQL Server Authentication

Login: abcd

Password: **********

☐ Remember password

Connection Security

Encryption: Mandatory

☑ Trust server certificate

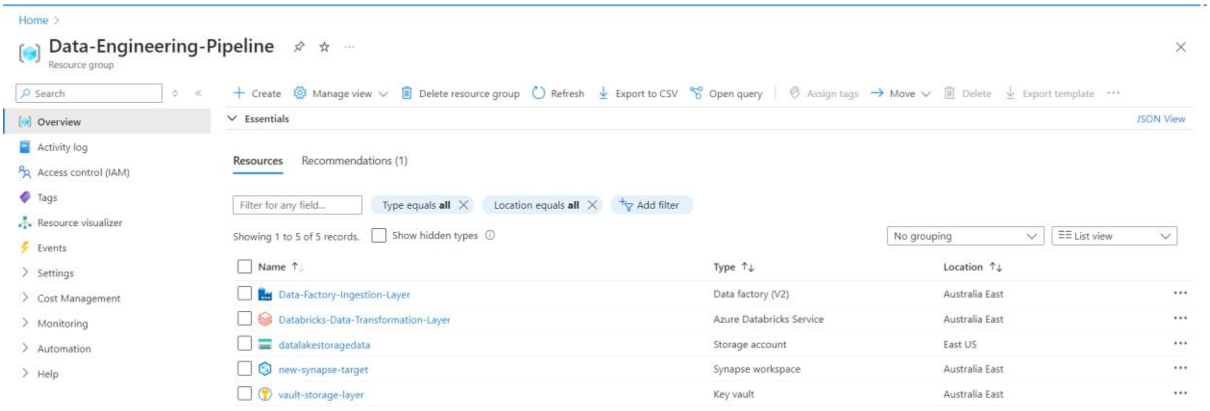
Host name in certificate:

## Level 2: Azure portal

- Sign up/Login using credentials
- Create a Resource group

  List of resources to be created under this resource group are:

  - Azure Data Factory (V2)
  - Azure Databricks service
  - Azure Key vault
  - Storage Account (Azure Data Lake Storage Gen 2 (ADLS Gen 2))
  - Azure Synapse Analytics Workspace



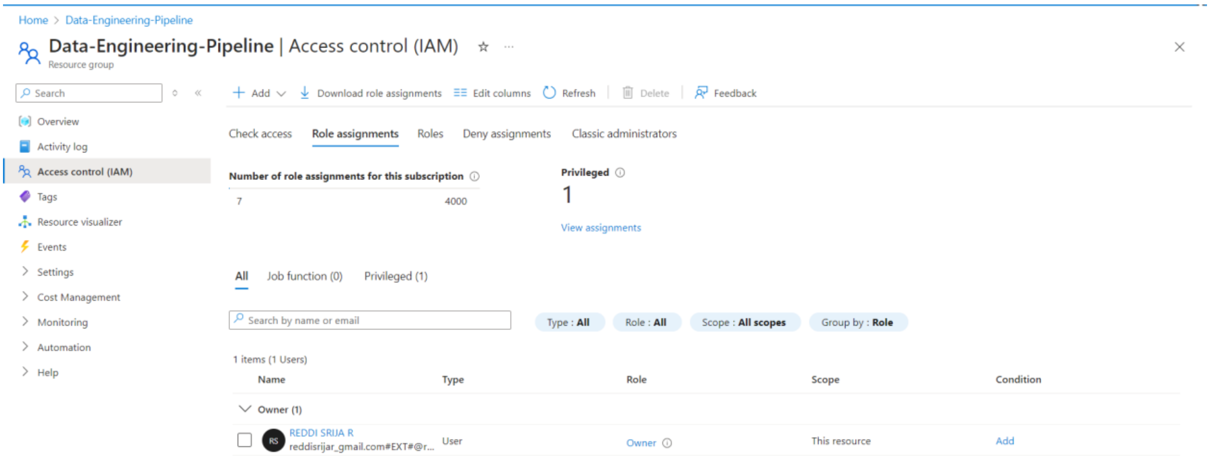
Data-Engineering-Pipeline

Resource group

Showing 1 to 5 of 5 records.

| Name | Type | Location |
|---|---|---|
| Data-Factory-Ingestion-Layer | Data factory (V2) | Australia East |
| Databricks-Data-Transformation-Layer | Azure Databricks Service | Australia East |
| datalakestoragedata | Storage account | East US |
| new-synapse-target | Synapse workspace | Australia East |
| vault-storage-layer | Key vault | Australia East |

- Under “Access Control (IAM)” section of your resource group – Add “Owner” role assignment. (Not Managed Identity) (Select user or group option)



Data-Engineering-Pipeline | Access control (IAM)

Resource group

Number of role assignments for this subscription: 7 / 4000

Privileged: 1

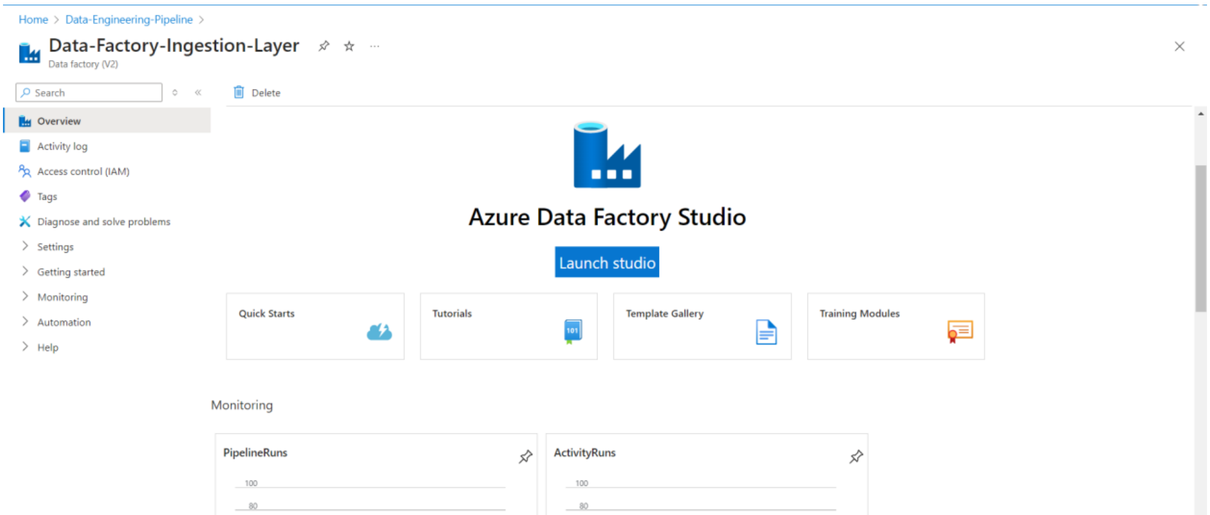
1 items (1 Users)

| Name | Type | Role | Scope | Condition |
|---|---|---|---|---|
| Owner (1) | | | | |
| REDDI SRIJA R<br>reddisrijar_gmail.com#EXT#@r... | User | Owner | This resource | Add |

- Your project pipeline will be built in four layers:
  - Data Ingestion Layer
  - Data Transformation Layer
  - Data Loading Layer
  - Data Reporting Layer

## 1. Data Ingestion Layer

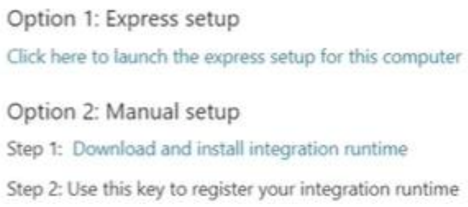
- Launch “Azure Data Factory Studio”



- Navigate to “Manage” section in left pane

- In “Integration service”
  - Default service named “Auto Resolve Integration Runtime” will be available – used to connect any cloud-based resources.

  - Create a new “Integration Runtime” – used to connect any on-premise resources. (As we are ingesting data from on-prem SQL Server)
  - Select either option and follow the instructions.
    - If Option 1: Click the link to download -> run the .exe file -> you can choose to either start or stop the created integration runtime service.
    - If Option 2: Click the link to download -> use provided either of the key to complete the setup.



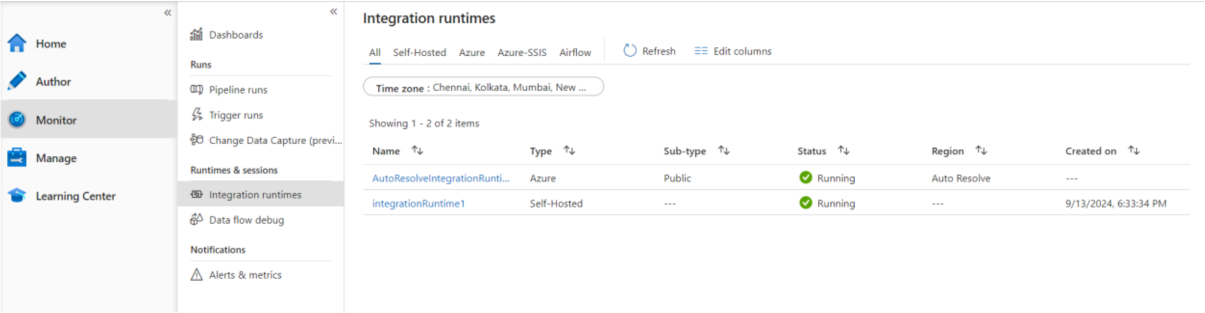
Option 1: Express setup

Click here to launch the express setup for this computer

Option 2: Manual setup

Step 1: Download and install integration runtime

Step 2: Use this key to register your integration runtime



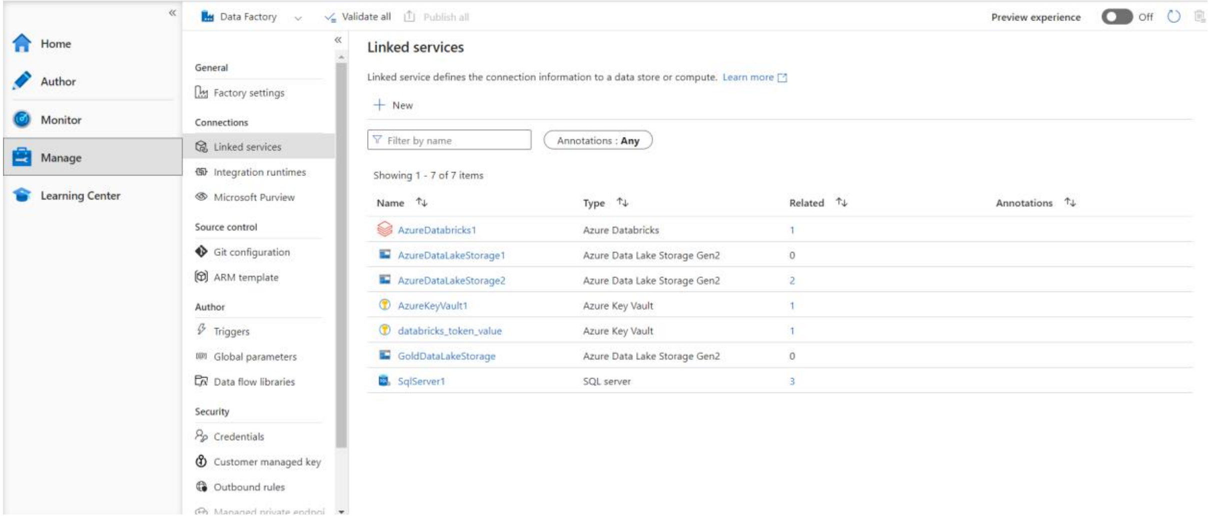
Integration runtimes

Time zone : Chennai, Kolkata, Mumbai, New ...

Showing 1 - 2 of 2 items

| Name | Type | Sub-type | Status | Region | Created on |
|---|---|---|---|---|---|
| AutoResolveIntegrationRunti... | Azure | Public | ✓ Running | Auto Resolve | --- |
| integrationRuntime1 | Self-Hosted | --- | ✓ Running | --- | 9/13/2024, 6:33:34 PM |

- Create a connection under “Linked Services” section for:
    - SQL server (use integration runtime)
    - Key vault (to automatically detect credentials for connection between azure services)
    - ADLS Gen 2 (use auto resolve runtime)



Linked services

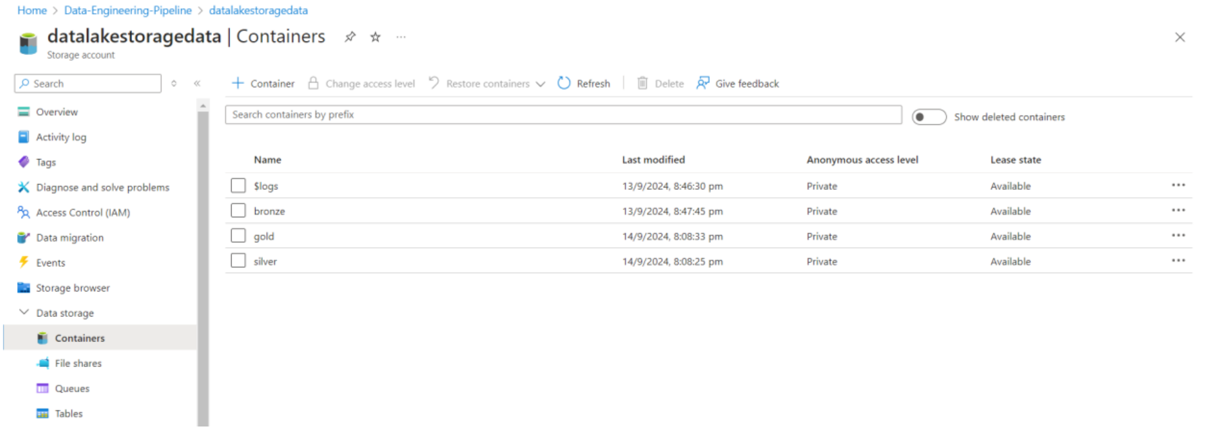
Linked service defines the connection information to a data store or compute. Learn more

Showing 1 - 7 of 7 items

| Name | Type | Related | Annotations |
|---|---|---|---|
| AzureDatabricks1 | Azure Databricks | 1 | |
| AzureDataLakeStorage1 | Azure Data Lake Storage Gen2 | 0 | |
| AzureDataLakeStorage2 | Azure Data Lake Storage Gen2 | 2 | |
| AzureKeyVault1 | Azure Key Vault | 1 | |
| databricks_token_value | Azure Key Vault | 1 | |
| GoldDataLakeStorage | Azure Data Lake Storage Gen2 | 0 | |
| SqlServer1 | SQL server | 3 | |

- Navigate to “Author” section in left pane.
- Create a new pipeline to load data automatically
    - Select “Lookup Tables” – to copy tables from our on-prem SQL Server in ‘parquet’ data format; from “Activities” section.

      Click “Publish” or “Publish All” to save your work in ADF (Azure Data Factory).
    - Create a “Bronze”, “Silver” and “Gold” containers in ‘storage account’ resource (Data storage section) in your resource group.
    - Provide that Bronze container’s location to the pipeline for raw data to get ingested.



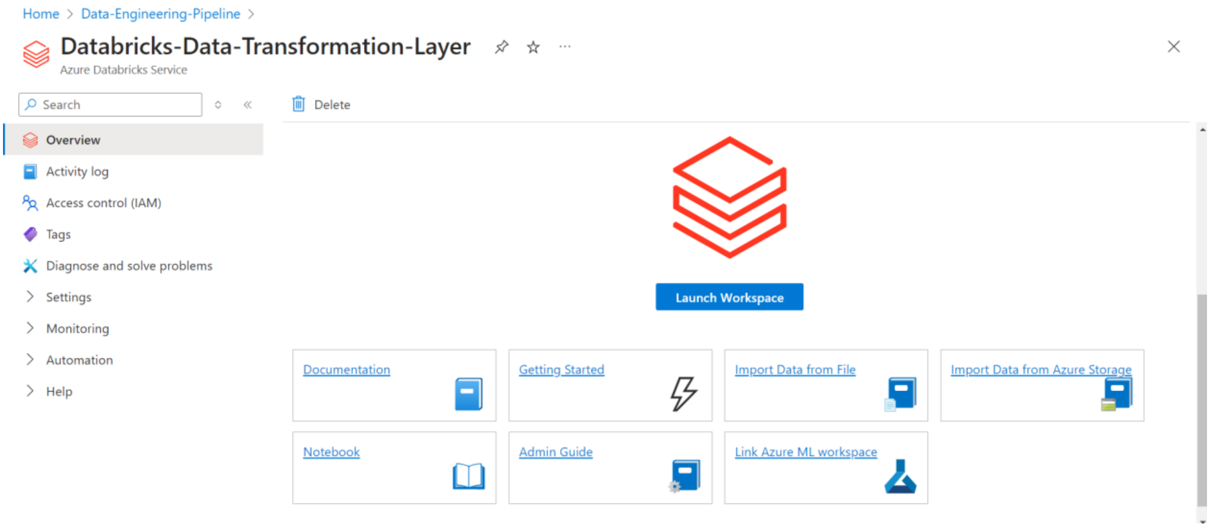
datalakestoragedata | Containers

| Name | Last modified | Anonymous access level | Lease state |
|---|---|---|---|
| □ $logs | 13/9/2024, 8:46:30 pm | Private | Available |
| □ bronze | 13/9/2024, 8:47:45 pm | Private | Available |
| □ gold | 14/9/2024, 8:08:33 pm | Private | Available |
| □ silver | 14/9/2024, 8:08:25 pm | Private | Available |

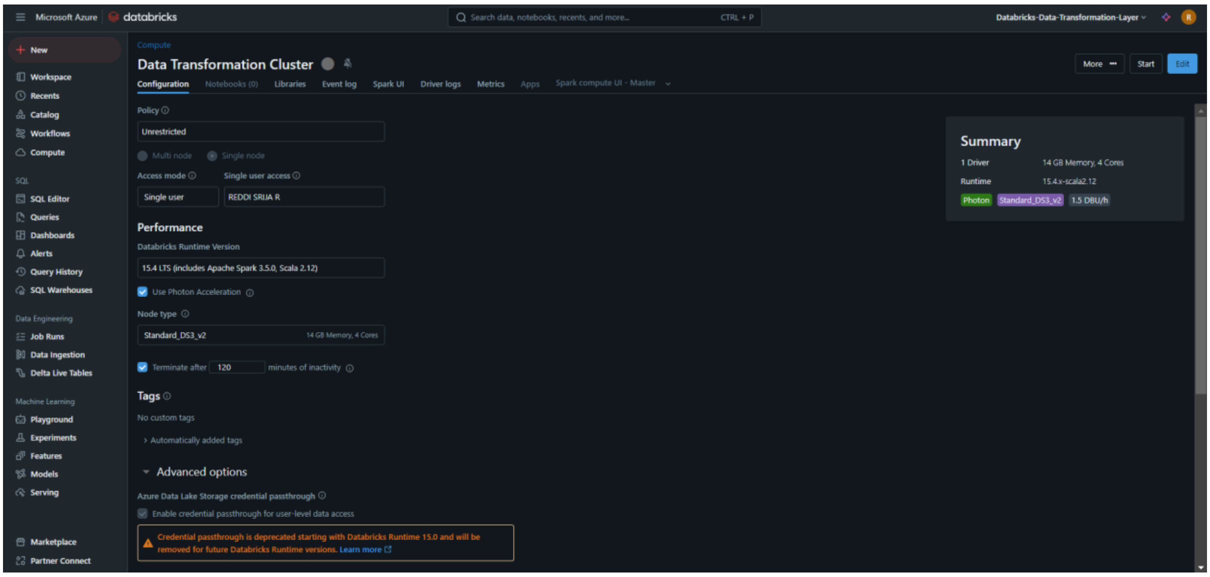
- Under Activities, select “Copy Each” and integrate it with “For Each” Activity – to run the activity for each table inside your database.

## 2. Data Transformation Layer

- Launch Databricks workspace



- Under “Compute” section – create new cluster
- Under “Workspace” section – inside ‘shared’ folder – create three notebooks
    - Storage Mount – to mount our cloud resources
    - Bronze to silver notebook – to store transformed data
    - Silver to gold notebook – to stored fully transformed data



Data Transformation Cluster

Policy: Unrestricted

○ Multi node ● Single node

Access mode: Single user

Single user access: REDDI SRIJA R

Performance

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

☑ Use Photon Acceleration

Node type: Standard_DS3_v2 14 GB Memory, 4 Cores

☑ Terminate after 120 minutes of inactivity

Tags

No custom tags

› Automatically added tags

Advanced options

Azure Data Lake Storage credential passthrough

☑ Enable credential passthrough for user-level data access

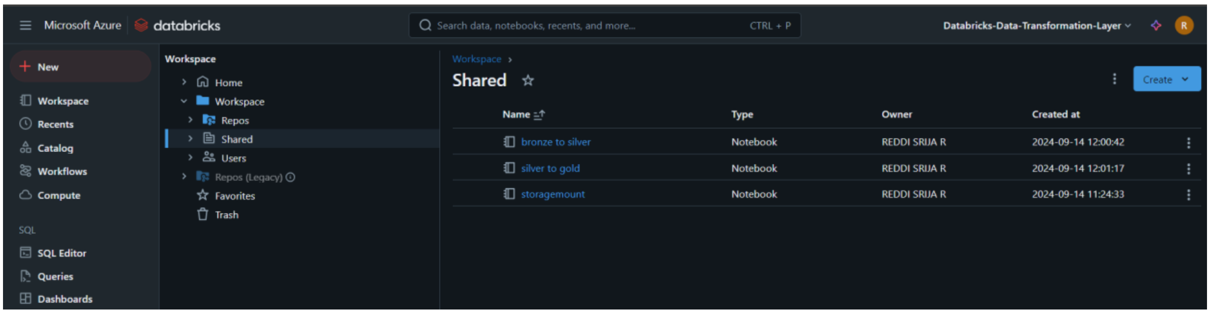
Credential passthrough is deprecated starting with Databricks Runtime 15.0 and will be removed for future Databricks Runtime versions. Learn more

Summary

1 Driver 14 GB Memory, 4 Cores

Runtime 15.4.x-scala2.12

Photon Standard_DS3_v2 1.5 DBU/h



Workspace >

Shared

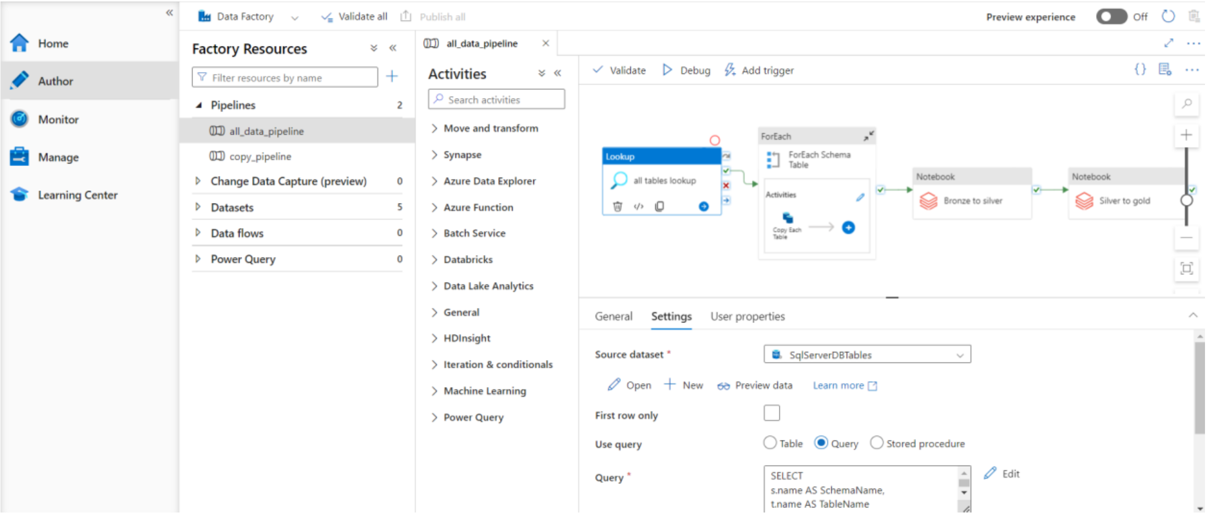
| Name | Type | Owner | Created at |
|---|---|---|---|
| bronze to silver | Notebook | REDDI SRIJA R | 2024-09-14 12:00:42 |
| silver to gold | Notebook | REDDI SRIJA R | 2024-09-14 12:01:17 |
| storagemount | Notebook | REDDI SRIJA R | 2024-09-14 11:24:33 |

- Navigate to Azure data factory:
    - In Manage -> linked services -> create a connection for “Databricks” with an access token.
    - Access token can be created in “Databricks” resource in your resource group.

    - Go to your previously created pipeline and select ‘Databricks’ from Activities section.
    - Connect your “Bronze to silver” and “Silver to gold” notebooks.

    - Notebook’s can contain code that performs any type of data transformation necessary for your use case.

Code for the notebooks is uploaded in the github with same filename.



Source dataset * SqlServerDBTables

First row only □

Use query ○ Table ● Query ○ Stored procedure

Query *

```
SELECT
s.name AS SchemaName,
t.name AS TableName
```

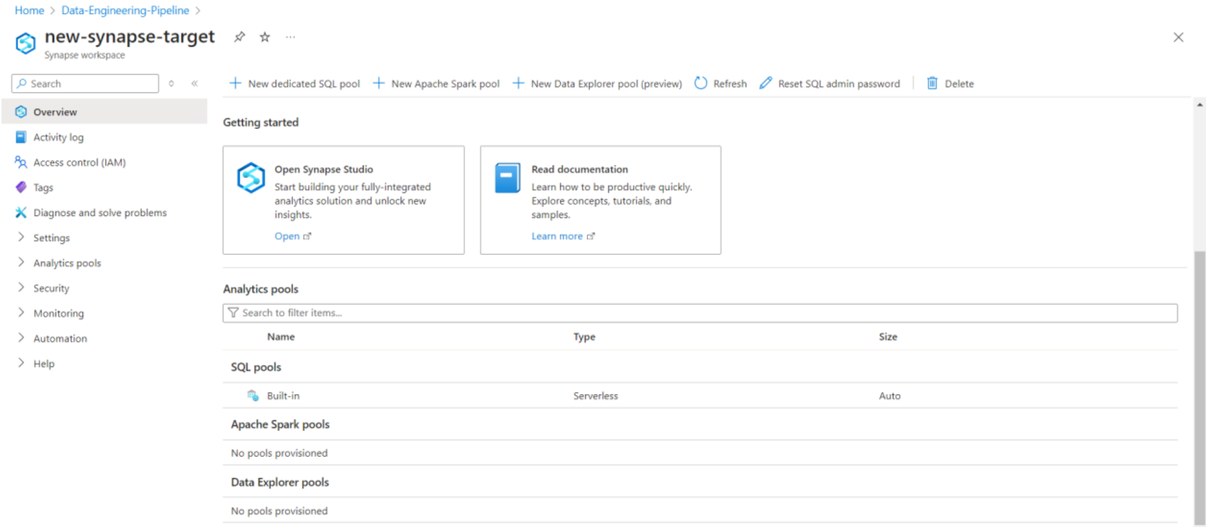
- You can “Debug” and check the working of pipeline

- Click on “Add Trigger” to run the entire pipeline and monitor the progress of your pipeline under “Monitor” section in left pane.

- Trigger has two options: If “Trigger now” is opted, then you have to manually run the process; whereas if opt for “New/Edit”, then it will create a trigger which will automatically run our pipeline upon any modifications of your data in your database.

## 3. Data Loading Layer

- Download OpenJDK for synapse workspace.
- Launch “Open Synapse Studio”

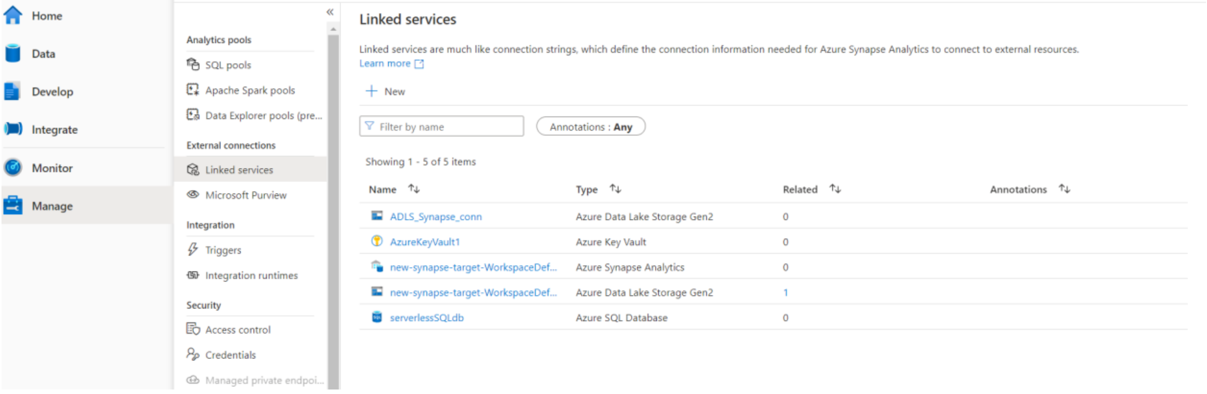


Start building your fully-integrated analytics solution and unlock new insights.

Learn how to be productive quickly. Explore concepts, tutorials, and samples.

| Name | Type | Size |
|---|---|---|
| SQL pools | | |
| Built-in | Serverless | Auto |
| Apache Spark pools | | |
| No pools provisioned | | |
| Data Explorer pools | | |
| No pools provisioned | | |

- Create a connection from “Linked services” from Manage section in left pane for “Azure SQL Database”.



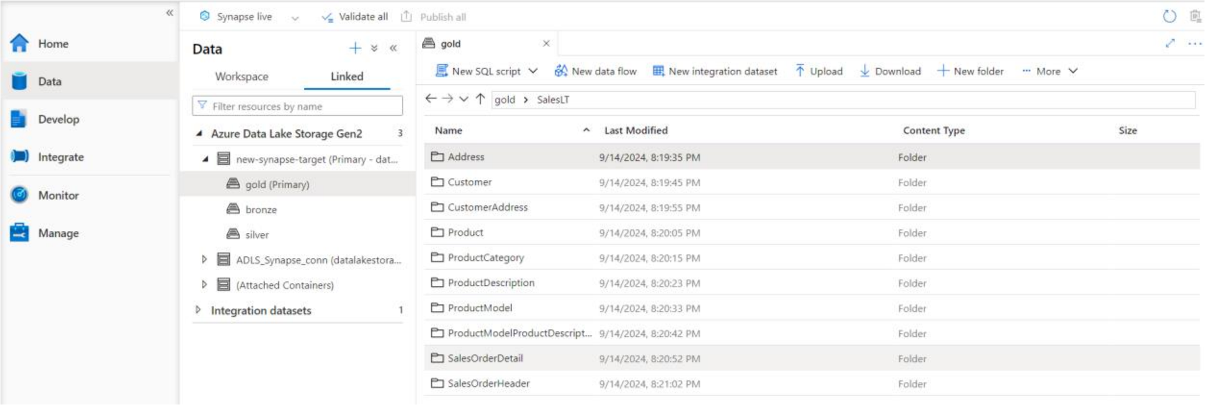
Linked services are much like connection strings, which define the connection information needed for Azure Synapse Analytics to connect to external resources.

Showing 1 - 5 of 5 items

| Name | Type | Related | Annotations |
|---|---|---|---|
| ADLS_Synapse_conn | Azure Data Lake Storage Gen2 | 0 | |
| AzureKeyVault1 | Azure Key Vault | 0 | |
| new-synapse-target-WorkspaceDef... | Azure Synapse Analytics | 0 | |
| new-synapse-target-WorkspaceDef... | Azure Data Lake Storage Gen2 | 1 | |
| serverlessSQLdb | Azure SQL Database | 0 | |

- Create a “Serverless SQL Database” from “Data” section in left pane. It will be reflected in ‘workspace’ block.

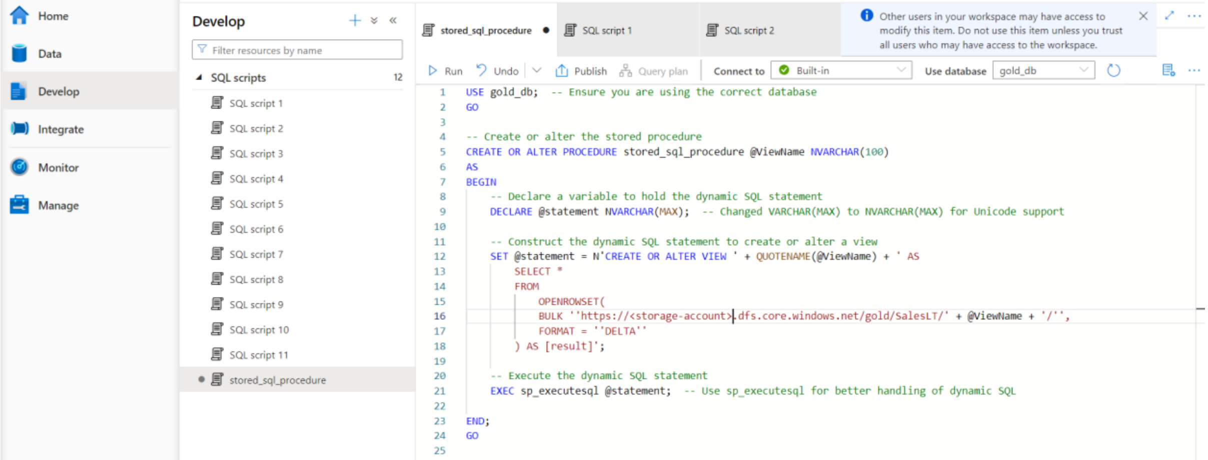
- Under ‘Linked’ block, you can expand the sections to find you data containers. Right click and select “Top 1000 rows” to have a preview of your data.



| Name | Last Modified | Content Type | Size |
|---|---|---|---|
| Address | 9/14/2024, 8:19:35 PM | Folder | |
| Customer | 9/14/2024, 8:19:45 PM | Folder | |
| CustomerAddress | 9/14/2024, 8:19:55 PM | Folder | |
| Product | 9/14/2024, 8:20:05 PM | Folder | |
| ProductCategory | 9/14/2024, 8:20:15 PM | Folder | |
| ProductDescription | 9/14/2024, 8:20:23 PM | Folder | |
| ProductModel | 9/14/2024, 8:20:33 PM | Folder | |
| ProductModelProductDescript... | 9/14/2024, 8:20:42 PM | Folder | |
| SalesOrderDetail | 9/14/2024, 8:20:52 PM | Folder | |
| SalesOrderHeader | 9/14/2024, 8:21:02 PM | Folder | |

- From “Develop” section in left pane, create a .sql script file or import existing .sql file which should contain SQL code to that creates views of your data into the database you created in synapse workspace. Run and select your previously created database.

- A ‘stored procedure’ sql file will be developed to automatically detect data from your containers present in ADLS Gen 2 storage account and load into your database as Views.



Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

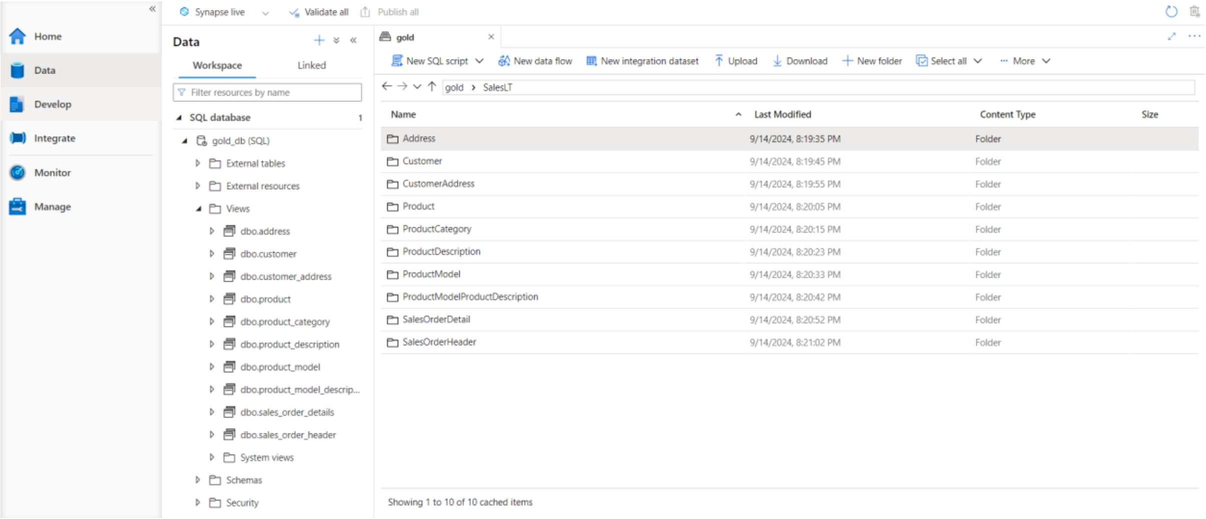
Connect to: Built-in

Use database: gold_db

```sql
USE gold_db;  -- Ensure you are using the correct database
GO

-- Create or alter the stored procedure
CREATE OR ALTER PROCEDURE stored_sql_procedure @ViewName NVARCHAR(100)
AS
BEGIN
    -- Declare a variable to hold the dynamic SQL statement
    DECLARE @statement NVARCHAR(MAX);  -- Changed VARCHAR(MAX) to NVARCHAR(MAX) for Unicode support

    -- Construct the dynamic SQL statement to create or alter a view
    SET @statement = N'CREATE OR ALTER VIEW ' + QUOTENAME(@ViewName) + ' AS
        SELECT *
        FROM
            OPENROWSET(
            BULK ''https://<storage-account>.dfs.core.windows.net/gold/SalesLT/' + @ViewName + '/'',
            FORMAT = ''DELTA''
        ) AS [result]';

    -- Execute the dynamic SQL statement
    EXEC sp_executesql @statement;  -- Use sp_executesql for better handling of dynamic SQL

END;
GO
```

- From synapse workspace – properties – collect “Serverless SQL Endpoint” to use it whenever necessary as “Server name” value.

- From “Integrate” section in left pane, create a pipeline for views creation.
- From activities section – select “Get metadata” option – opt for binary data format while connection and provide “Gold” container’s location.
- Select “For Each” from activities section.
- As for the activity of this “For Each” block, select “Stored procedure” from activities section and connect your stored procedure .sql file developed.
- Use Debug or Trigger accordingly.



| Name | Last Modified | Content Type | Size |
|---|---|---|---|
| Address | 9/14/2024, 8:19:35 PM | Folder | |
| Customer | 9/14/2024, 8:19:45 PM | Folder | |
| CustomerAddress | 9/14/2024, 8:19:55 PM | Folder | |
| Product | 9/14/2024, 8:20:05 PM | Folder | |
| ProductCategory | 9/14/2024, 8:20:15 PM | Folder | |
| ProductDescription | 9/14/2024, 8:20:23 PM | Folder | |
| ProductModel | 9/14/2024, 8:20:33 PM | Folder | |
| ProductModelProductDescription | 9/14/2024, 8:20:42 PM | Folder | |
| SalesOrderDetail | 9/14/2024, 8:20:52 PM | Folder | |
| SalesOrderHeader | 9/14/2024, 8:21:02 PM | Folder | |

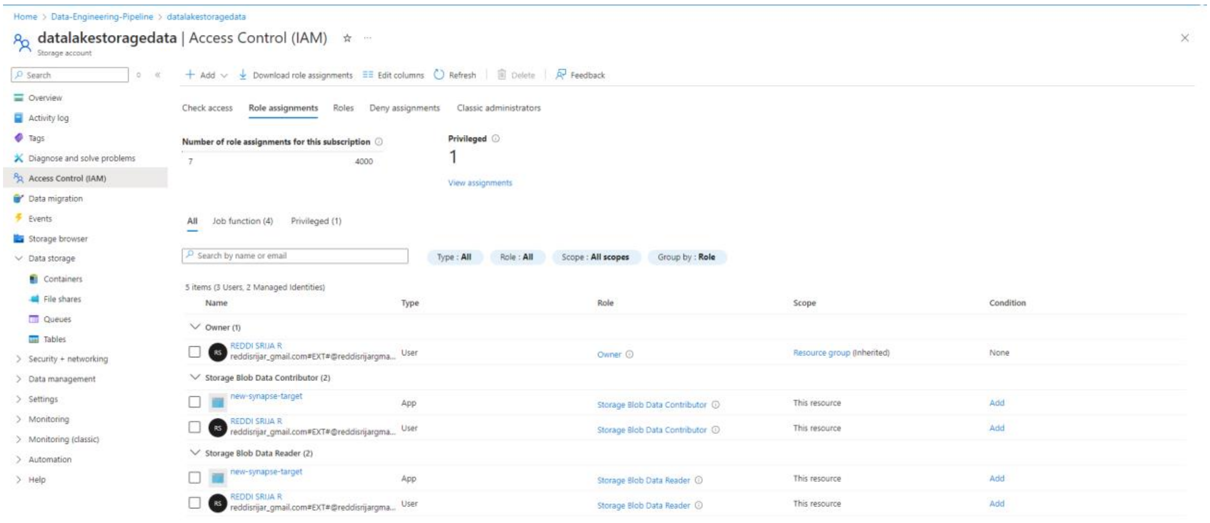
Showing 1 to 10 of 10 cached items

## 4. Data Reporting

- Power BI
    - Get Data - Azure - Synapse Analytics (SQL DW)
    - Provide Credentials and Sign in your Microsoft account
    - Select and Load Data Tables
    - Create visuals and Develop relevant Dashboard

## Access Control (IAM)

- Add role assignments
    - Owner
    - Storage Blob Data Contributor (User and Managed Identity)
    - Storage Blob Data Reader (User and Manager Identity)
    - While creation of Managed Identity – select synapse workspace.



Number of role assignments for this subscription

7 4000

Privileged

1

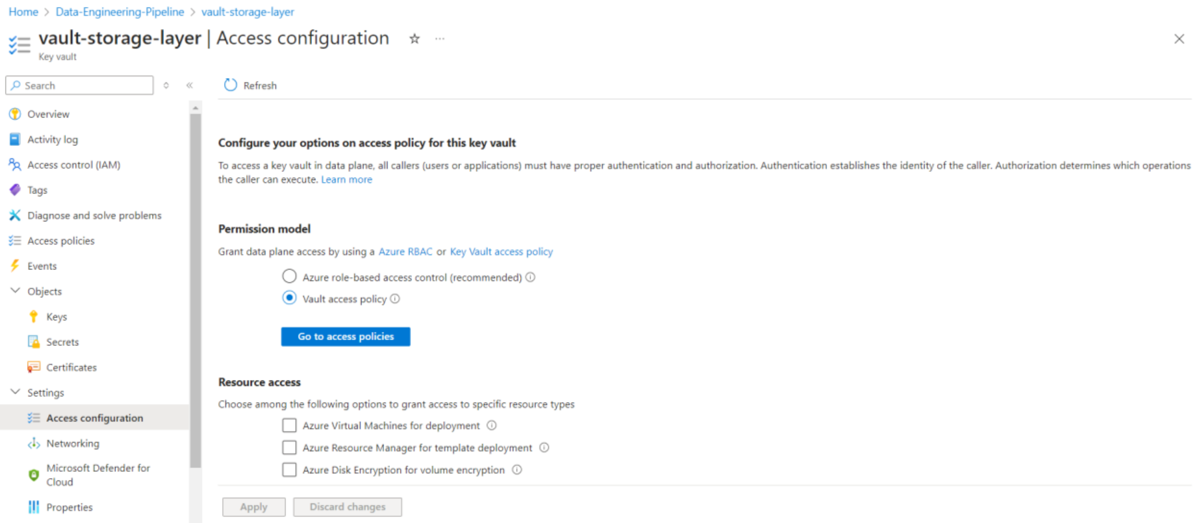
View assignments

5 items (3 Users, 2 Managed Identities)

| Name | Type | Role | Scope | Condition |
|---|---|---|---|---|
| Owner (1) | | | | |
| ☐ REDDI SRIJA R<br>reddisrijar_gmail.com#EXT#@reddisrijargma... | User | Owner | Resource group (Inherited) | None |
| Storage Blob Data Contributor (2) | | | | |
| ☐ new-synapse-target | App | Storage Blob Data Contributor | This resource | Add |
| ☐ REDDI SRIJA R<br>reddisrijar_gmail.com#EXT#@reddisrijargma... | User | Storage Blob Data Contributor | This resource | Add |
| Storage Blob Data Reader (2) | | | | |
| ☐ new-synapse-target | App | Storage Blob Data Reader | This resource | Add |
| ☐ REDDI SRIJA R<br>reddisrijar_gmail.com#EXT#@reddisrijargma... | User | Storage Blob Data Reader | This resource | Add |

## Key Vault

- Generate “Username”, “Password” and “Databricks token” under “SECRETS” from “Objects” section.
- To give access of these secrets to other services/resources while connection creation, change “Access Configuration” to “Value access policy”.
- Except for “Azure Key Vault” resource, other resource’s “Access Configuration” should be “Azure role-based access control” (ADBC control)



Home > Data-Engineering-Pipeline > vault-storage-layer

vault-storage-layer | Access configuration

Key vault

**Configure your options on access policy for this key vault**

To access a key vault in data plane, all callers (users or applications) must have proper authentication and authorization. Authentication establishes the identity of the caller. Authorization determines which operations the caller can execute. Learn more

**Permission model**

Grant data plane access by using a Azure RBAC or Key Vault access policy

○ Azure role-based access control (recommended)

● Vault access policy

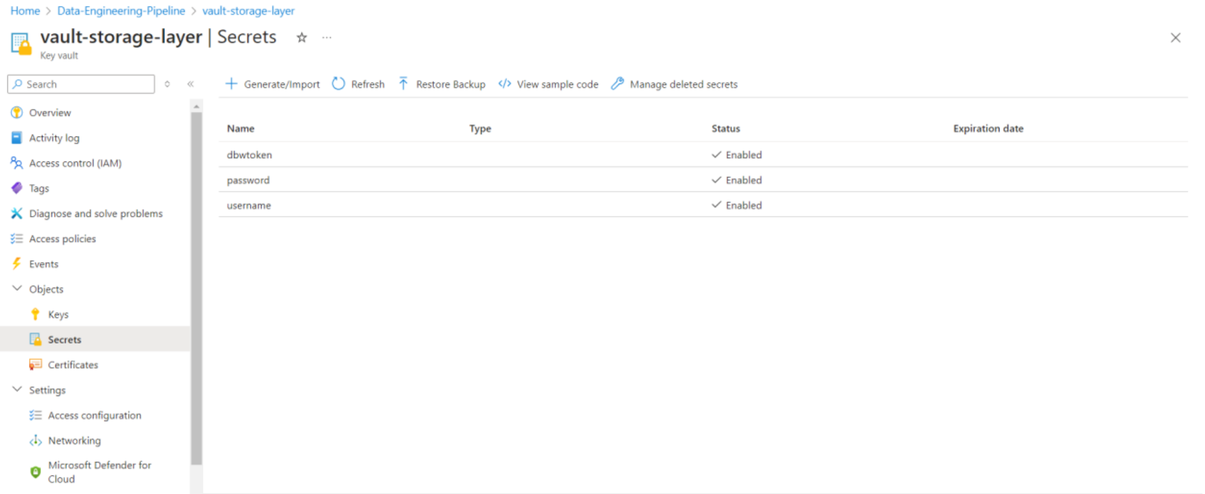
**Resource access**

Choose among the following options to grant access to specific resource types

□ Azure Virtual Machines for deployment

□ Azure Resource Manager for template deployment

□ Azure Disk Encryption for volume encryption



Home > Data-Engineering-Pipeline > vault-storage-layer

vault-storage-layer | Secrets

Key vault

| Name | Type | Status | Expiration date |
|---|---|---|---|
| dbwtoken | | ✓ Enabled | |
| password | | ✓ Enabled | |
| username | | ✓ Enabled | |

## Additional Resources:

- https://docs.microsoft.com/en-us/azure/data-factory/

- https://docs.microsoft.com/en-us/azure/databricks/

- https://docs.microsoft.com/en-us/azure/synapse-analytics/

- Mount ADLS Gen 2 storage Documentation: https://learn.microsoft.com/en-us/azure/databricks/archive/credential-passthrough/adls-passthrough