

1. Pulling data from API
2. Powerful orchestration tool : Azure data factory
3. Building dynamic parameter
4. Pulling data from source and land data in bronze layer
5. bronze(raw) layer: keep data as it is as a replica. Do not apply any transformation.
6. Push data to silver layer with some transformations
7. Serving layer: serving data to stake holders
8. We r making azure warehouse which is Azure synapse analytics
9. Analysis on power bi

Medallion architecture

Medallion Architecture is a layered approach to organizing data processing pipelines, where data flows through multiple stages (or “medallions”) — typically Bronze, Silver, and Gold layers. Each layer refines the data progressively, improving quality and usability.

Layers in Medallion Architecture

1. Bronze Layer (Raw Data)

- Contains raw, unprocessed, or minimally processed data.
- Data is ingested as-is from various sources (logs, IoT devices, databases, streaming sources).
- Usually stores data in its original format.
- May contain duplicates, errors, and inconsistencies.
- Acts as the single source of truth for original data.

2. Silver Layer (Cleaned and Enriched Data)

- Contains cleaned, filtered, and possibly joined data.
- Data quality issues are fixed here (null values handled, invalid records removed).
- Data is structured and normalized for easier querying.
- Can join multiple sources, enrich data with reference datasets.
- Serves as the basis for detailed analytics and business intelligence.

3. Gold Layer (Aggregated and Business-Level Data)

- Contains aggregated, highly curated, and business-friendly datasets.
- Optimized for end-user consumption — dashboards, reports, machine learning models.
- Data here is often denormalized and aggregated by key business dimensions.
- Typically used for KPI reporting, machine learning feature sets, or exporting to other systems.

Why use Medallion Architecture?

- Data Quality Improvement: Step-by-step cleansing and validation.
- Separation of Concerns: Different teams can work independently on each layer.
- Reusability: Silver data can feed multiple gold datasets.
- Scalability: Efficient to process and store data incrementally.
- Traceability: Ability to trace back to raw source data (Bronze) when needed.

Create the resource group

Add tags for doing categorization

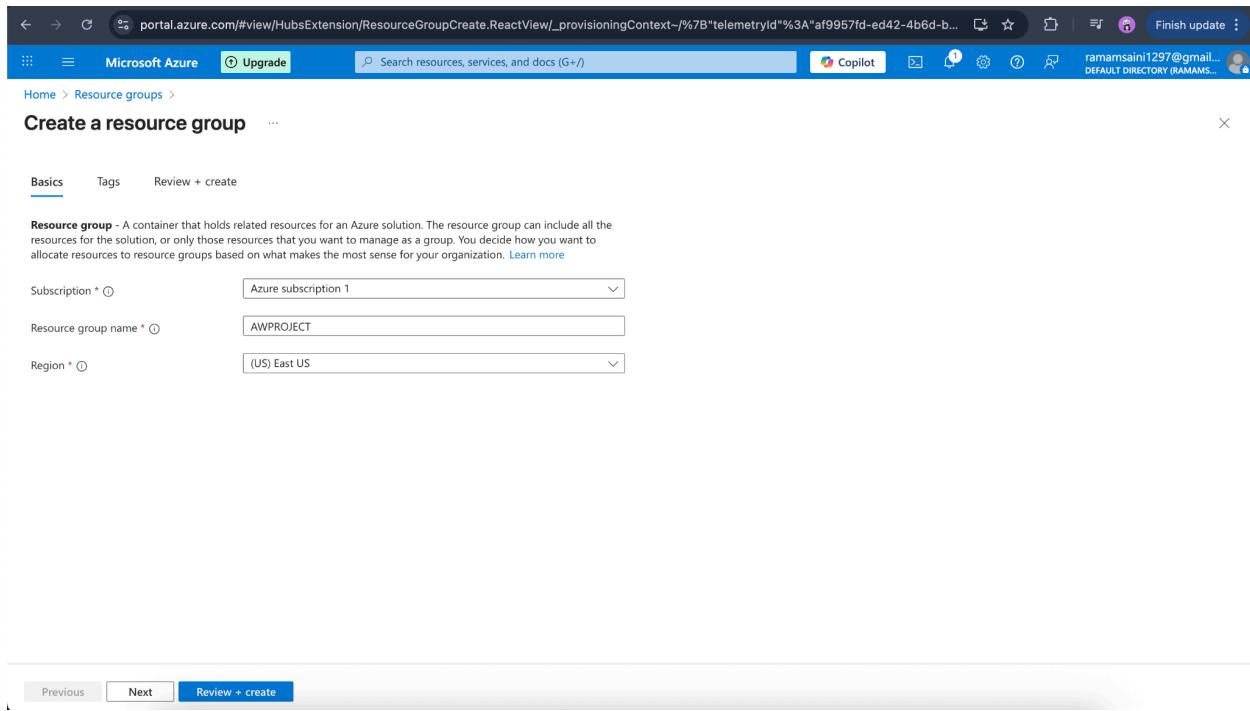
LET'S BEGIN

1. Azure Resource Setup

1.1 Create a Resource Group

A resource group is a container for related Azure resources.

1. Navigate to the Azure portal and create a new Resource Group.
2. Provide a Subscription, a Resource group name (e.g., AWPROJECT), and a Region.
3. Optionally, add tags for categorization.



1.2 Create an Azure Data Lake Storage Account

A data lake is a storage account designed for big data analytics. To create one:

1. Create a Storage Account.
2. Provide a Subscription, the Resource group created earlier, and a Storage account name (e.g., **adventureworkstoraged1**).
3. Select the Region and Performance tier. For Redundancy, Locally-redundant storage (LRS) stores replicas in the same data center, while Geo-redundant storage (GRS) replicates data to different regions.
4. In the Advanced tab, enable the Hierarchical namespace to create an Azure Data Lake Gen2 account instead of a standard blob storage account. This allows for a hierarchical file structure with folders.
5. After creation, create containers inside this storage account to represent the layers of the Medallion Architecture:
bronze, **silver**, and **gold**

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

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

Finish update

ramamsaini1297@gmail.com

DEFAULT DIRECTORY (RAMAMS...)

Home > Create a storage account ...

Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more about Azure storage accounts](#)

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * Azure subscription 1

Resource group * AWPROJECT

Create new

Instance details

Storage account name * adventureworkstoragedl

Region * (US) East US Deploy to an Azure Extended Zone

Primary service Azure Blob Storage or Azure Data Lake Storage Gen 2

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

Premium: Recommended for scenarios that require low latency.

Redundancy * Locally-redundant storage (LRS)

Previous Next Review + create Give feedback

Chrome File Edit View History Bookmarks Profiles Tab Window Help Fri 8 Aug 3:55 PM

Inbox (11,516) – ramamsaini1297@gmail.com | What is TensorFlow? | (10) Azure End-To-End Data ... | Create a storage account – Microsoft Azure

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

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

Copilot

ramamsaini1297@gmail.com

DEFAULT DIRECTORY (RAMAMS...)

Home > Create a 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

Access protocols

Blob and Data Lake Gen2 endpoints are provisioned by default [Learn more](#)

Enable SFTP

Enable network file system v3

Blob storage

Allow cross-tenant replication

ⓘ Cross-tenant replication and hierarchical namespace cannot be enabled simultaneously.

Access tier Hot: Optimized for frequently accessed data and everyday usage scenarios

Cool: Optimized for infrequently accessed data and backup scenarios

Cold: Optimized for rarely accessed data and backup scenarios

Azure Files

Enable large file shares

Previous Next Review + create Give feedback

This screenshot shows the 'Create a storage account' wizard on the Azure portal. It's the second step, 'Instance details'. The 'Project details' section is above, showing a subscription and a resource group named 'AWPROJECT'. The 'Instance details' section includes fields for the storage account name ('adventureworkstoragedl'), region ('(US) East US'), primary service ('Azure Blob Storage or Azure Data Lake Storage Gen 2'), performance level ('Standard'), and redundancy ('Locally-redundant storage (LRS)'). Below this is a 'Hierarchical Namespace' section with a checked checkbox for enabling it. The 'Access protocols' section shows checkboxes for 'Enable SFTP' and 'Enable network file system v3'. Under 'Blob storage', there's a note about 'Cross-tenant replication' and a radio button for 'Hot' access tier. The 'Azure Files' section has a checkbox for 'Enable large file shares'. At the bottom are 'Previous', 'Next', and 'Review + create' buttons, along with a 'Give feedback' link.

Use data lake when need a hierarchical file structure (folders).

Azure Data Factory

Azure Data Factory is a cloud-based data integration service provided by Microsoft Azure. It allows you to create, schedule, and orchestrate data pipelines to move and transform data from various sources to your desired destination — typically for analytics or storage.

Azure Data Factory = ETL / ELT tool on Azure

- Extract → from data sources (e.g., SQL, APIs, blob storage, on-prem)
- Transform → using data flows or external compute like Azure Databricks
- Load → into destinations like Azure Synapse, Data Lake, SQL Database, etc.

Create containers that indicates for zones (bronze, silver and gold)

Create the zones inside azure data factory (adventureworkstorage)

2.2 Creating a Static Pipeline

A static pipeline is a simple, non-dynamic data copy process.

1. Create an Azure Data Factory instance in the `AWPROJECT` resource group.
2. In the ADF Studio, navigate to the Manage section to create Linked Services.
3. Create an HTTP linked service to connect to the source API (e.g., GitHub).
4. Create a second linked service to connect to the Azure Data Lake Storage Gen2 account (`storageDataLake`).
5. In the Author section, create a new pipeline.
6. Add a Copy data activity to the pipeline.
7. Configure the Source dataset, using the HTTP linked service and a relative URL to the CSV file.
8. Configure the Sink dataset, using the data lake linked service. Set the File path to the `bronze` container with a subfolder (e.g., `/bronze/products/products.csv`).
9. Debug the pipeline to ensure the data is successfully copied to the data lake.
10. Publish the changes to save your work. The data should now be in the `bronze` layer of your data lake. This is the raw data, kept as a replica without any transformations.

Microsoft Azure Upgrade Search resources, services, and docs (G+) Copilot Home > Create Data Factory ...

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

Subscription * Azure subscription 1
Resource group * AWPROJECT Create new

Instance details
Name * awlinstance1
Region * East US
Version * V2

Previous Next Review + create Give feedback

Microsoft Azure Microsoft.DataFactory-20250808155826 | Overview

awlinstance1 Data factory (V2)

Search Delete

Overview

- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems
- Resource visualizer
- Settings
- Getting started
- Monitoring
- Automation
- Help

Essentials

Resource group (move) : AWPROJECT	Type : Data factory (V2)
Status : Succeeded	Getting started : Quick start
Location : East US	
Subscription (move) : Azure subscription 1	
Subscription ID : 8095fb85-ef2a-4467-9263-7324217c1456	

Azure Data Factory Studio

Launch studio

Quick Starts Tutorials Template Gallery Training Modules

Monitoring

Add or remove favorites by pressing Cmd+Shift+F

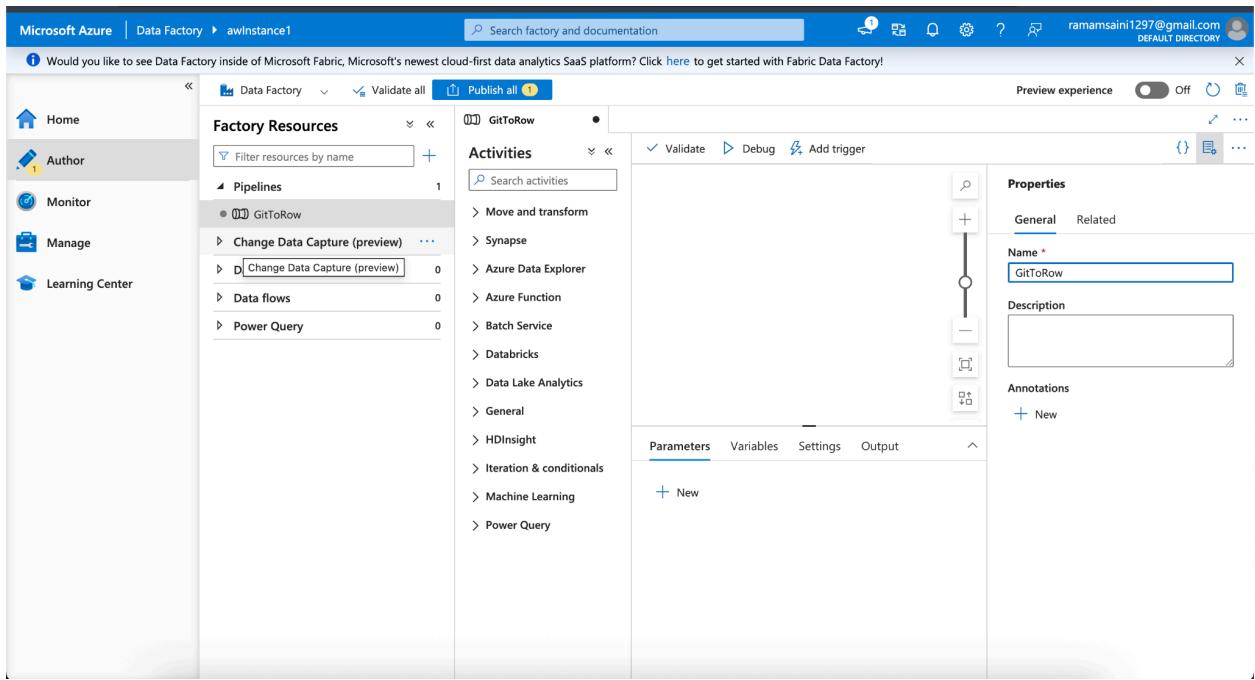
Author : creating all pipelines

Monitor: monitor our pipelines, failed pipelines , reasons of failed pipelines

Manage:manage things like repo github connections, devops connections and link services

Learning center: read resources, documentation pick avail. Data sets

Create the containers for all zones in data lake (adventureworkstoragedl)



ETL pipeline with Azure Data Factory

Link services: connection that raw data has build with source and destination

Put Data from github and add in data lake

Create 2 linked services

Like pick data from http and push to data lake

Http linked service as taking data from git hub

Microsoft Azure | Data Factory > awinstance1

Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click [here](#) to get started.

Home Author Monitor Manage Learning Center

Data Factory Validate all Publish all

General Factory settings Connector upgrade advisor

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 Credentials Customer managed key Outbound rules

Linked services

Linked service defines the connection information to a data store

+ New Filter by name Annotations : Any

If you expected to see

New linked service

HTTP Learn more

Name * HttpLinkedService

Description

Connect via integration runtime * AutoResolveIntegrationRuntime

Base URL * <https://raw.githubusercontent.com>

Information will be sent to the URL specified. Please ensure you trust the URL entered.

Server certificate validation Enable Disable

Authentication type * Anonymous

Auth headers New

Annotations New

Annotations

Connection successful

Create Back Test connection Cancel

This screenshot shows the 'New linked service' dialog for an 'HTTP' service. The 'Name' field is set to 'HttpLinkedService'. Under 'Connect via integration runtime', the 'AutoResolveIntegrationRuntime' checkbox is checked. The 'Base URL' field contains the URL 'https://raw.githubusercontent.com'. The 'Authentication type' dropdown is set to 'Anonymous'. At the bottom right, there are 'Create', 'Back', 'Test connection', and 'Cancel' buttons. A success message 'Connection successful' is displayed above the 'Test connection' button.

Microsoft Azure | Data Factory > awinstance1

Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click [here](#) to get started.

Home Author Monitor Manage Learning Center

Data Factory Validate all Publish all

General Factory settings Connector upgrade advisor

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 Credentials Customer managed key Outbound rules

Linked services

Linked service defines the connection information to a data store

+ New Filter by name Annotations : Any

Showing 1 - 1 of 1 items

Name ↑

● HttpLinkedService

New linked service

Azure Data Lake Storage Gen2 Learn more

Name * storageDataLake

Description

Connect via integration runtime * AutoResolveIntegrationRuntime

Authentication type Account key

Account selection method From Azure subscription Enter manually

Azure subscription Select all

Storage account name * Select all adventureworkstoragedl

Test connection To linked service To file path

Annotations New

Create Back Test connection Cancel

This screenshot shows the 'New linked service' dialog for an 'Azure Data Lake Storage Gen2' service. The 'Name' field is set to 'storageDataLake'. Under 'Connect via integration runtime', the 'AutoResolveIntegrationRuntime' checkbox is checked. The 'Authentication type' dropdown is set to 'Account key'. The 'Account selection method' dropdown is set to 'From Azure subscription'. The 'Azure subscription' dropdown is set to 'Select all'. The 'Storage account name' dropdown is set to 'adventureworkstoragedl'. At the bottom right, there are 'Create', 'Back', 'Test connection', and 'Cancel' buttons.

Microsoft Azure | Data Factory > awinstance1

Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click [here](#) to get started with Fabric Data Factory!

Search factory and documentation

Data Factory Validate all Publish all

Preview experience Off

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
HttpLinkedService	HTTP	0	
storageDataLake	Azure Data Lake Storage Gen2	0	

Create the data set out of it
 Created the static pipeline
 Create the source data set as
 author—> source—> New—> http—> csv

Microsoft Azure | Data Factory > awinstance1

Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click [here](#) to get started with Fabric Data Factory!

Search factory and documentation

Data Factory Validate all Publish all

Set properties

Factory Resources

Pipelines 1

GitToRow

Activities

Move and transform

Synapse

Azure Data Explorer

Azure Function

Batch Service

Databricks

Data Lake Analytics

General

HDInsight

Iteration & conditionals

Machine Learning

Power Query

Name: ds_https

Linked service: HttpLinkedService

Relative URL: ramandeep-12/azureDataProject/ref/heads/main/AdventureWorks_Product_Subcategory

First row as header:

Import schema:

- From connection/store
- From sample file
- None

Source database: Open

Request method: GET

Additional headers: Request body

OK Back Cancel

The screenshot shows the Microsoft Azure Data Factory Author interface. On the left, the navigation bar includes Home, Author (selected), Monitor, Manage, and Learning Center. The main area displays 'Factory Resources' under Pipelines, with one pipeline named 'GitToRow'. The pipeline details pane shows the 'Activities' section with a 'Copy data' activity selected. The 'Source' tab is active, showing 'Source dataset' as 'ds_http'. Below it, 'Request method' is set to 'GET', and there are fields for 'Additional headers' and 'Request body'. The top right corner has a 'Preview experience' toggle set to 'Off'.

For sink

New → data lake gen 2 → csv

The screenshot shows the 'Set properties' dialog for a sink dataset named 'ds_raw'. The dialog includes fields for 'Name' (set to 'ds_raw'), 'Linked service' (set to 'storageDataLake'), 'File path' (set to 'bronze/products/products.csv'), 'First row as header' (checkbox checked), 'Import schema' (radio button set to 'None'), and an 'Advanced' section. The 'Sink dataset' tab is selected at the bottom. At the bottom right are 'OK', 'Back', and 'Cancel' buttons.

Debug

The screenshot shows the Microsoft Azure Data Factory interface. On the left, there's a navigation bar with options like Home, Author, Monitor, Manage, and Learning Center. The main area is titled 'Factory Resources' and shows a list of Pipelines, Datasets, Data flows, and Power Query. A specific pipeline named 'ds_http' is selected. Within this pipeline, a 'Copy data' activity named 'CopyRawData' is highlighted. The pipeline status is shown as 'Succeeded'. Below the pipeline details, there's a table showing the run history.

Activity name	Activity st...	Activit...	Run start	Duration
CopyRawData	Succeeded	Copy data	8/8/2025, 5:16:15 PM	12s

The screenshot shows the Microsoft Azure Storage Explorer interface. It displays a container named 'bronze' which contains a directory 'products' and a file 'products.csv'. The file 'products.csv' has a size of 56.76 KiB and was last modified on 08/08/2025, 17:16:25. The blob type is 'Block blob' and the lease state is 'Available'.

Name	Last modified	Access tier	Blob type	Size	Lease state
products.csv	08/08/2025, 17:16:25	Hot (Inferred)	Block blob	56.76 KiB	Available

2.3 Creating a Dynamic Pipeline

A dynamic pipeline handles multiple files or changing parameters automatically.

1. Create a new container in your data lake, named **parameters**.

2. Upload a JSON file (e.g., `dummy.json`) to this container. This file will contain the relative URL, folder name, and file name for each source file.
3. In the ADF pipeline, add a Lookup activity to read the contents of the `dummy.json` file.
4. Add a For Each activity and set its Items property to the output of the Lookup activity using the expression `@activity('LookupGit').output.value`.
5. Inside the For Each activity, add the Copy data activity.
6. For the Source dataset, create a dynamic dataset with a parameter for the relative URL. Set the value to `@item().parameter_relative_url`.
7. For the Sink dataset, create a dynamic dataset with parameters for the sink folder and file name. Set their values to `@item().p_sink_folder` and `@item().p_sink_fileName` respectively.

Publish this as it saves our work in data factory and whenever we come then able to see all the progress otherwise we need to rebuild everything

Dynamic Pipeline

Using iterations and conditions

1. Relative url
 2. Folder where we store data
 3. File name
- Above 3 got changes everytime so instead of changing manually we create the parameters

For each Activity(for loop)

Put copy activity inside for each activity these things keep on moving till our iterations got completed

Need one data set that contains data of all the files... one data set to pick all the data set instead of creating multiple data sets called parameterization

Microsoft Azure | Data Factory > awinstance1

Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click [here](#) to get started.

Home Author Monitor Manage Learning Center

Data Factory > Validate all Publish all 1

Factory Resources < >

Pipelines 2

- GitToRow
- DynamicGitToRaw

Change Data Capture (preview) 0

Datasets 2

- ds_http
- ds_raw

Data flows 0

Power Query 0

DynamicGitToRaw ●

Validate Validate copy runtime

Copy data

Copy data1

Relative URL

First row as header

Import schema

- From connection/store
- From sample file
- None

Advanced

Open this dataset for more advanced configuration with parameterization.

General Source 1 Sink 1 Mapping

Source dataset * Select...

OK Back Cancel

Set properties

Name: ds_dynamic

Linked service: HttpLinkedService

Relative URL:

First row as header: checked

Import schema: None

Source dataset: ds_dynamic

Microsoft Azure | Data Factory > awinstance1

Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click [here](#) to get started.

Home Author Monitor Manage Learning Center

Data Factory > Validate all Publish all 2

Factory Resources < >

Pipelines 2

- GitToRow
- DynamicGitToRaw

Change Data Capture (preview) 0

Datasets 3

- ds_dynamic
- ds_http
- ds_raw

Data flows 0

Power Query 0

DynamicGitToRaw ● ds_dynamic

DelimitedText ds_dynamic

CSV

Connection Schema Parameters

Linked service: HttpLink

Base URL: https://raw.

Relative URL:

Compression type: No compress

Column delimiter: Comma (,)

Row delimiter: Default (\r\n)

Encoding: Default(UTF)

Quote character: Double quote ("")

Escape character: Backslash (\)

Parameters

parameter_relative_url

Pipeline parameter

OK Cancel

Pipeline expression builder

Add dynamic content below using any combination of [expressions](#), [functions](#) and [system variables](#).

@dataset().parameter_relative_url

Clear contents

Parameters Functions

Search

parameter_relative_url

Pipeline parameter

parameter_relative_url

The screenshot shows the Microsoft Azure Data Factory Author interface. On the left, the navigation bar includes Home, Author (selected), Monitor, Manage, and Learning Center. The main area displays 'Factory Resources' with sections for Pipelines, Datasets, Data flows, and Power Query. Under Datasets, 'ds_dynamic' is selected. The right pane shows the dataset configuration with tabs for Connection, Schema, and Parameters. The Connection tab is active, showing a linked service named 'HttpLinkedService' connected to a base URL of 'https://raw.githubusercontent.com'. Other parameters include relative URL (@dataset().parameter_relative_url), compression type (No compression), column delimiter (Comma (,), row delimiter (Default (\r\n, or \n)), encoding (Default(UTF-8)), quote character (Double quote (")), and escape character (Backslash (\)).

Created sink dynamic parameter

The screenshot shows the Microsoft Azure Data Factory Author interface. The left sidebar shows Home, Author (selected), Monitor, Manage, and Learning Center. The main area shows 'Factory Resources' with Pipelines, Datasets, Data flows, and Power Query sections. A pipeline named 'DynamicGitToRaw' is selected. The right pane shows the pipeline configuration with activities like 'Copy data', 'Move and transform', 'Azure Data Explorer', and 'Iteration & conditionals'. A 'Set properties' dialog is open for the 'Copy data' activity, which is a 'sink dataset'. The dialog shows the name 'ds_sink_dynamic', linked service 'storageDataLake', file path (File system / Directory / File name), and import schema options (From connection/store, From sample file, None). Advanced options are also visible.

The screenshot shows the Microsoft Azure Data Factory Author interface. On the left sidebar, under 'Factory Resources', 'Datasets' is selected, showing four datasets: ds_dynamic, ds_http, ds_raw, and ds_sink_dynamic. The ds_sink_dynamic dataset is currently selected. The main panel displays the 'Connection' tab for this dataset. The 'Linked service' dropdown is set to 'storageDataLake'. The 'File path' field contains the expression '@dataset().p_sink_folder' followed by '/@dataset().p_sink_fileName'. Other connection settings include 'No compression' for compression type, 'Comma (,) for column delimiter, and 'Default (\r\n, or \n\r)' for row delimiter. Encoding is set to 'Default(UTF-8)', quote character to 'Double quote ("')', and escape character to 'Backslash (\')'. The 'First row as header' checkbox is checked.

For each activity

In items need the entities which should be run to properly

Make one parameter container where upload one json file which contains all the relative_url , foldername and file name

The screenshot shows the Microsoft Azure Data Factory Author interface. On the left sidebar, under 'Activities', 'Iteration & conditionals' is expanded, showing 'ForEach' selected. In the main workspace, a 'ForEach' activity is connected to a 'Copy data' activity. The 'ForEach' activity has a child activity named 'ForEachGit'. The 'Copy data' activity is named 'Copy data1'. Below the activities, the 'General' tab of the 'ForEach' activity is visible, showing the name 'ForEachGit', a description field, and an 'Activity state' section with 'Activated' selected.

New container

Name: parameters

Anonymous access level: Private (no anonymous access)

The access level is set to private because anonymous access is disabled on this storage account.

Create **Give feedback**

Upload blob

1 file(s) selected: dummy.json

Drag and drop files here or [Browse for files](#)

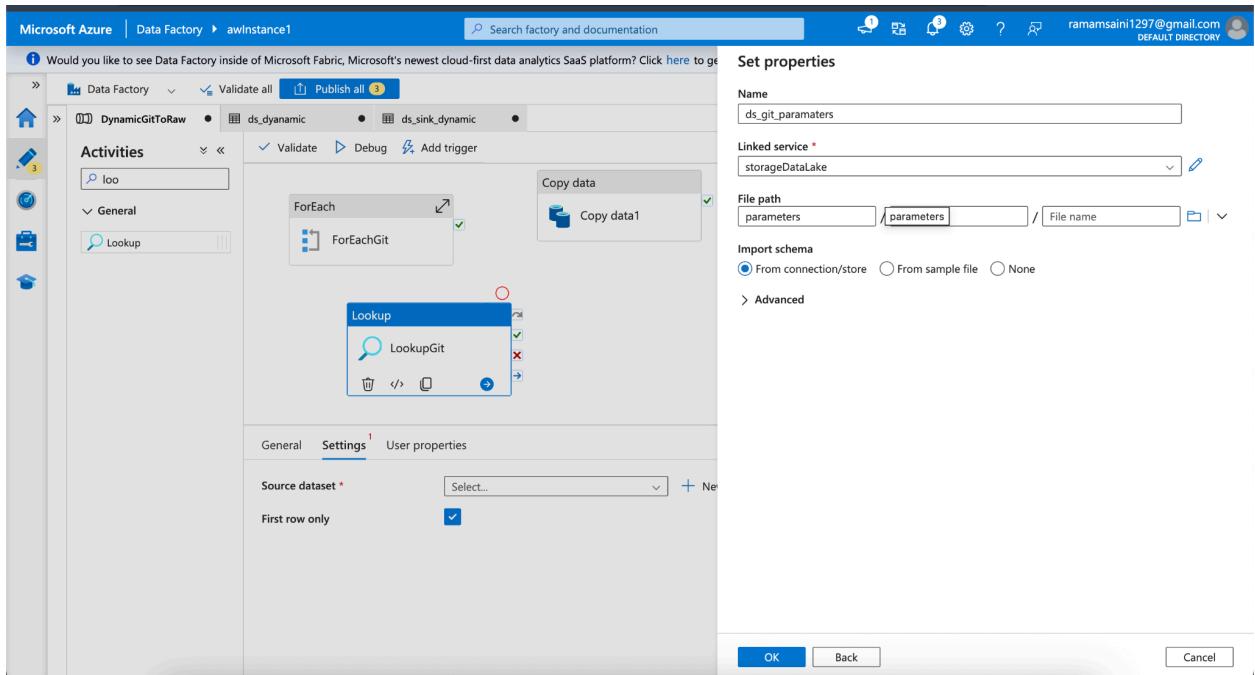
Overwrite if files already exist

Upload **Give feedback**

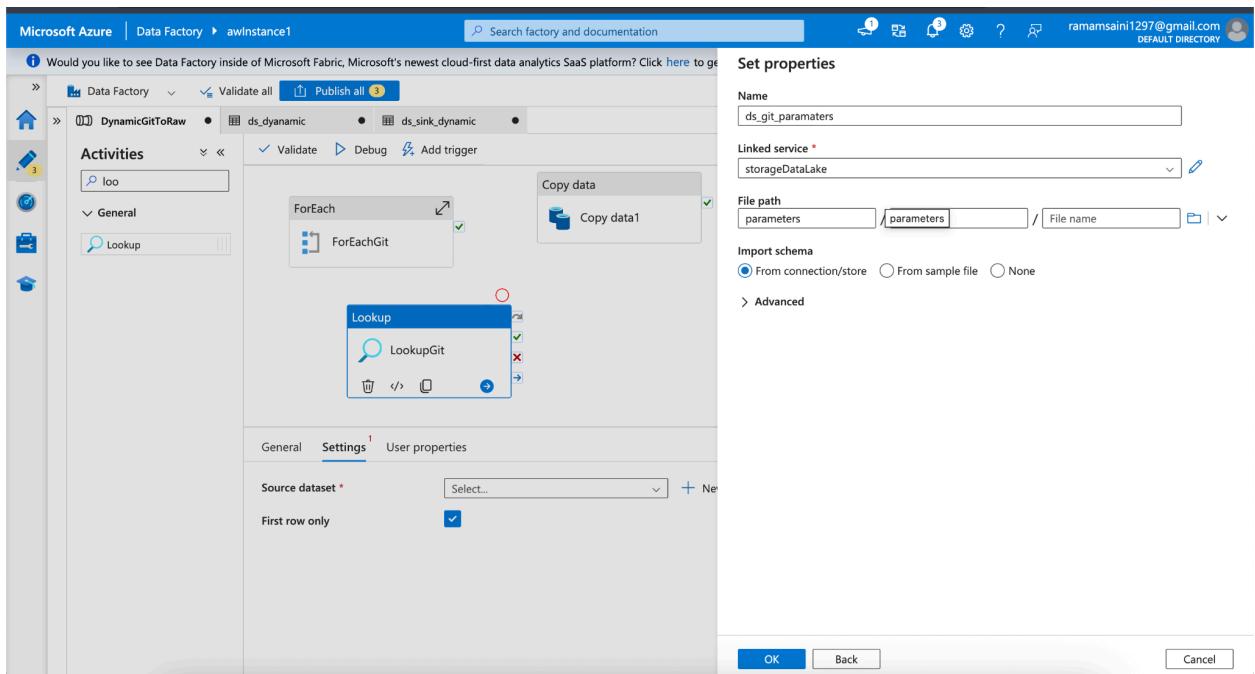
Lookup activity (gives the output of the data)

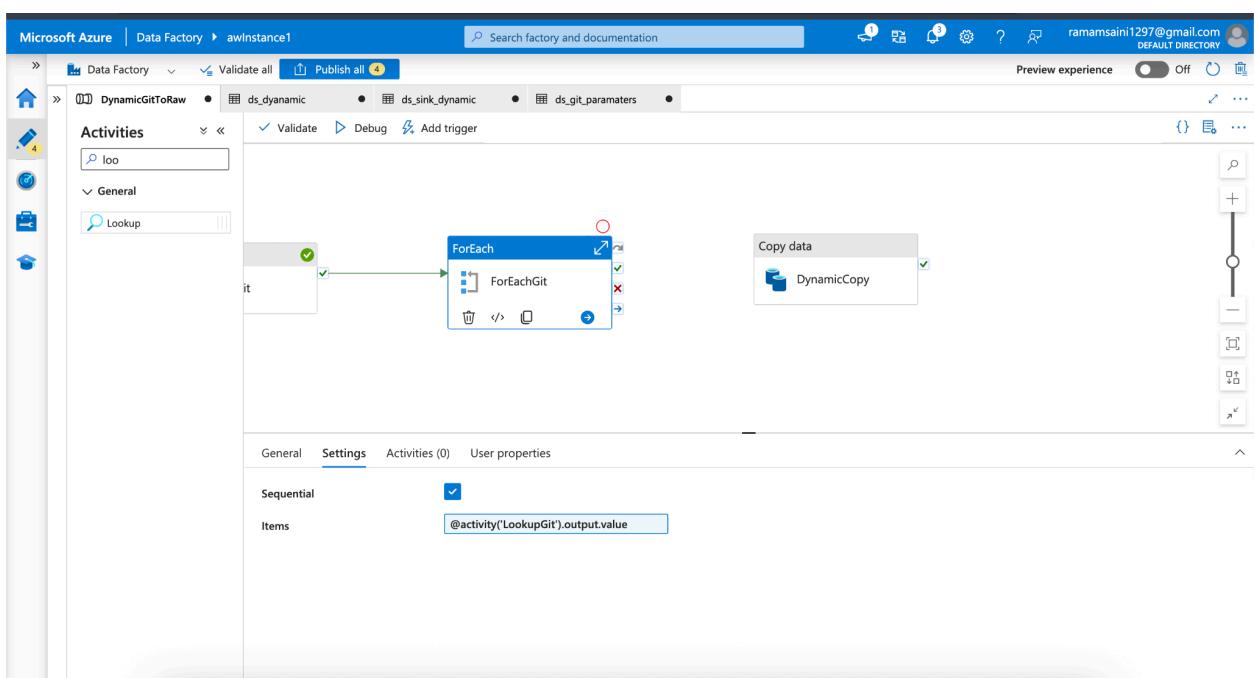
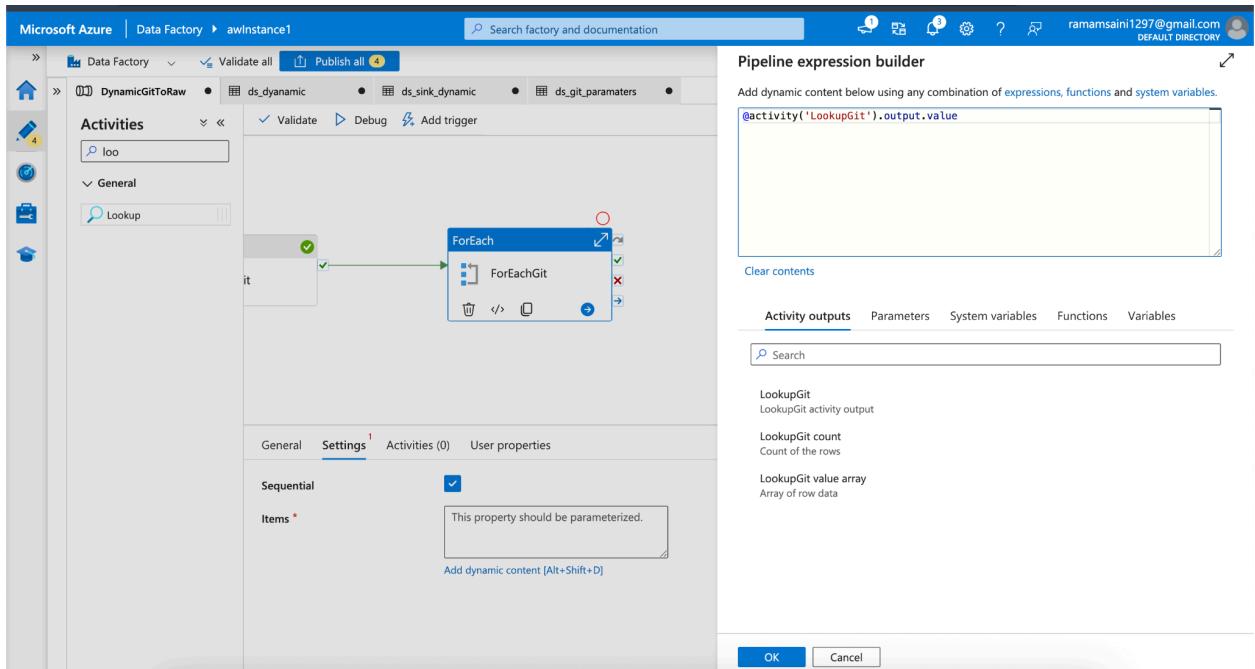
In Azure Data Factory (ADF), the Lookup activity is used to read data from a source (like a database, CSV, JSON, etc.) and make it available to other activities in the pipeline.

In forEach



Go to settings and of foreach and paste the dynamic copy there
.value should be written .. we ve the data





Microsoft Azure | Data Factory > awinstance1

Search factory and documentation

Preview experience Off

Data Factory > Validate all Publish all 4

Activities > General > DynamicGitToRaw > ForEachGit

Copy data

DynamicCopy

Source dataset: ds_dynamic

Dataset properties:

Name	Type
parameter_relative_url	String

Request method: GET

Additional headers:

General Source Sink Mapping Settings User properties

Source dataset: ds_dynamic

Dataset properties:

Name	Type
parameter_relative_url	String

Request method: GET

Additional headers:

Microsoft Azure | Data Factory > awinstance1

Search factory and documentation

Pipeline expression builder

Add dynamic content below using any combination of [expressions](#), [functions](#) and [system variables](#).

@item().parameter_relative_url

Clear contents

ForEach iterator Activity outputs Parameters System variables ...

Search

ForEachGit Current item

General Source Sink Mapping Settings User properties

Source dataset: ds_dynamic

Dataset properties:

Name	Type
parameter_relative_url	String

Request method: GET

Additional headers:

Microsoft Azure | Data Factory > awinstance1

Validate all Publish all 4

Search factory and documentation

ramamsaini1297@gmail.com DEFAULT DIRECTORY

Activities

DynamicGitToRaw > ForEachGit

Copy data

DynamicCopy

General Sink Mapping Settings User properties

Sink dataset * ds_sink_dynamic

Dataset properties

Name	Value
p_sink_folder	@item().p_sink_folder
p_sink_fileName	[Value Add dynamic content [Alt+T]]

Copy behavior Select...

OK Cancel

Pipeline expression builder

Add dynamic content below using any combination of [expressions](#), [functions](#) and [system variables](#).

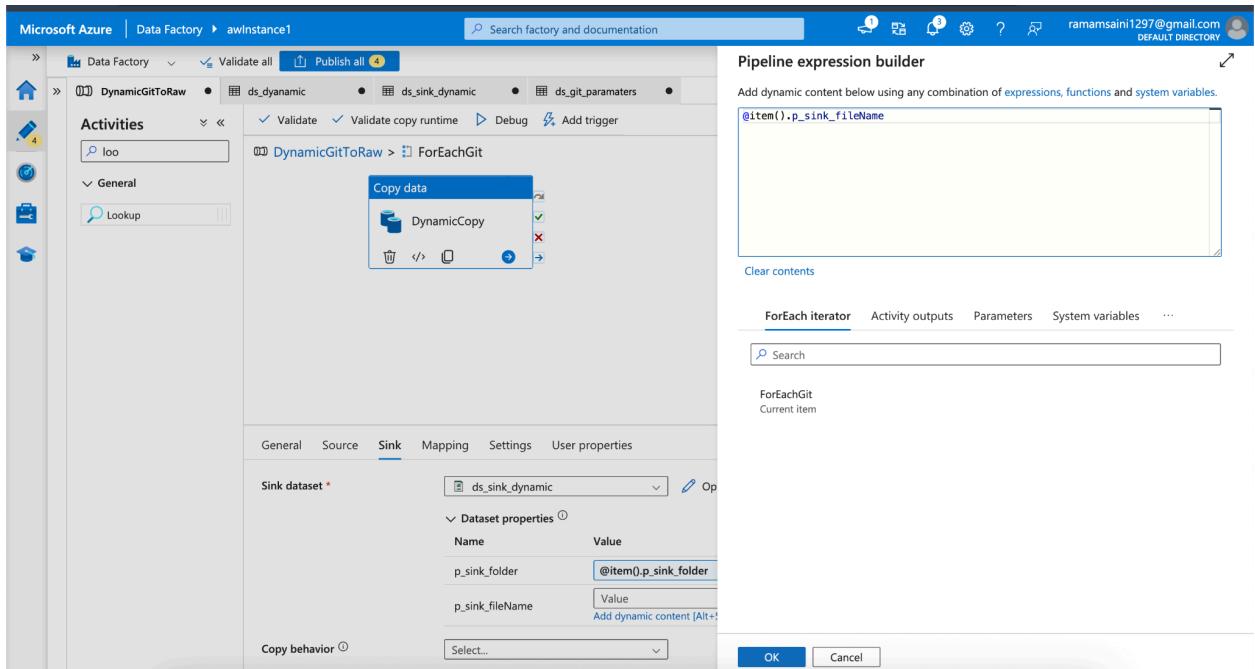
```
@item().p_sink_fileName
```

Clear contents

ForEach iterator Activity outputs Parameters System variables ...

Search

ForEachGit Current item



Microsoft Azure | Data Factory > awinstance1

Validate all Publish all 4

Search factory and documentation

ramamsaini1297@gmail.com DEFAULT DIRECTORY

Preview experience Off

Activities

DynamicGitToRaw > ForEachGit

Copy data

DynamicCopy

General Source Sink Mapping Settings User properties

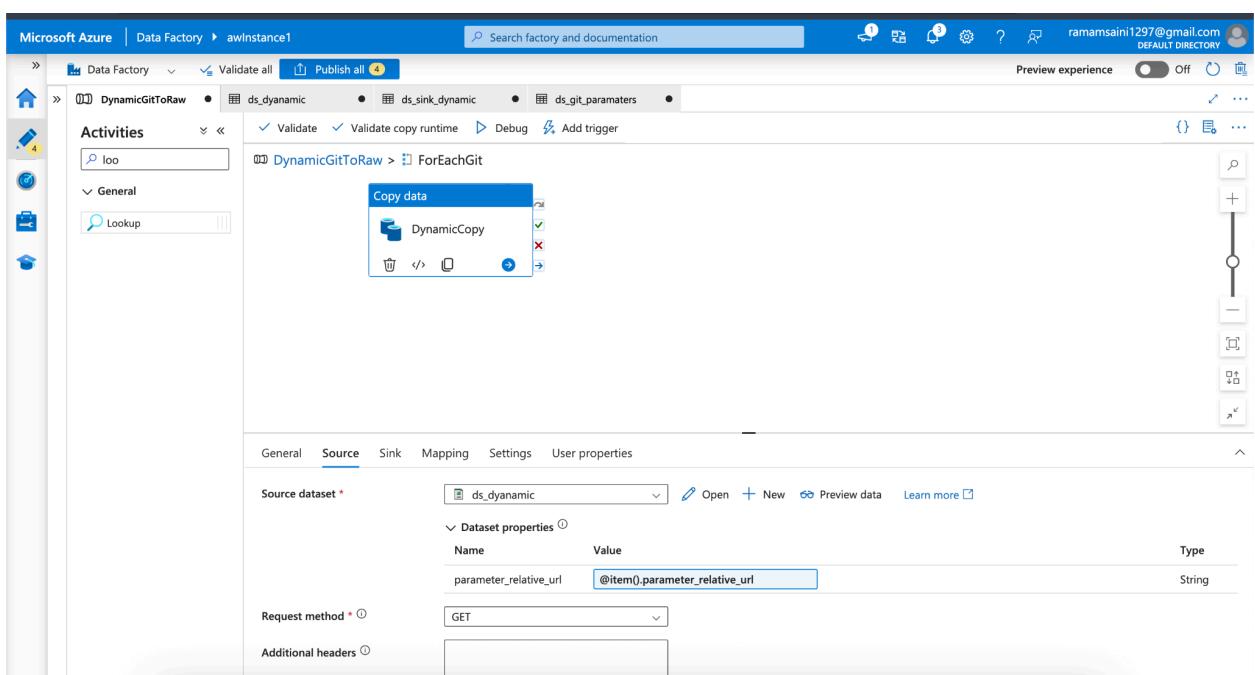
Source dataset * ds_dyanamic

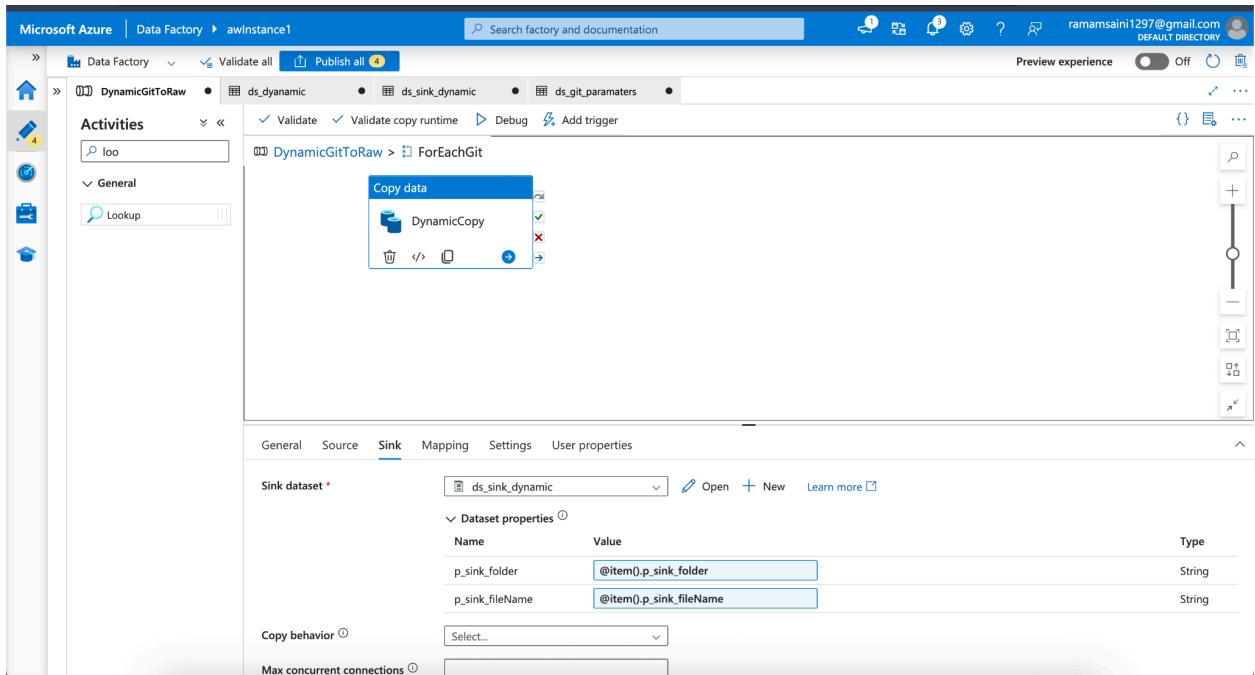
Dataset properties

Name	Type
parameter_relative_url	String

Request method * GET

Additional headers





AZURE DATABRICKS

3. Data Transformation with Azure Databricks

3.1 Setting up Databricks

Azure Databricks is a managed Apache Spark service for big data analytics.

1. Create an Azure Databricks workspace in the Azure portal. A managed resource group will be created automatically to house the virtual machines and virtual network for the cluster.
2. Launch the workspace and create a Cluster under the Compute section.
3. To allow Databricks to access your data lake, you need to register an application in Microsoft Entra ID (formerly Azure Active Directory). This application will act as a service principal.
4. Register a new application (e.g., `awproject_app`), create a Client Secret for it, and save the Application (client) ID, Directory (tenant) ID, and Client Secret.
5. Assign the `awproject_app` the Storage Blob Data Contributor role on your data lake storage account to give it read/write permissions.

3.2 Performing Transformations

1. In your Databricks workspace, create a new Notebook.
2. Attach the notebook to your cluster.
3. Write code (e.g., using PySpark or Scala) to perform transformations. This involves reading the raw data from the **bronze** container, cleaning and enriching it, and then writing the refined data to the **silver** container.

Azure databricks create the managed resource groups where it creates all the VM / VNet

After that just create it

Go to resource —> launch workspace

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

ramamsaini1297@gmail.com DEFAULT DIRECTORY (RAMAMS...)

Home > Create an Azure Databricks workspace ...

Basics Networking Encryption Security & compliance Tags Review + create

Project Details

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

Subscription * Azure subscription 1

Resource group * AWPROJECT [Create new](#)

Instance Details

Workspace name * azuredatabricks

Region * East US

Pricing Tier * Premium (+ Role-based access controls)

Managed Resource Group name managed-azuredatabricks-rg

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

In Azure Databricks, the term "compute" usually refers to the clusters or compute resources that run your code — whether that's Spark jobs, SQL queries, ML models, or data transformations.

The screenshot shows the 'Compute' section of the Databricks UI, specifically the 'Simple form: OFF' configuration for an 'azure Cluster'. The 'Configuration' tab is selected. Key settings include:

- Policy:** Unrestricted
- Access mode:** Single user or group access (Dedicated (formerly: Single user) - Ramandeep Saini)
- Performance:** Databricks Runtime Version: 16.4 LTS (includes Apache Spark 3.5.2, Scala 2.12), Use Photon Acceleration (unchecked), Node type: Standard_D4ds_v5 (16 GB Memory, 4 Cores), Terminate after 20 minutes of inactivity (checked).
- Tags:** No custom tags, Automatically added tags.

A summary panel on the right provides cluster details: 1 Driver, 16 GB Memory, 4 Cores, Runtime: 16.4.x-scala2.12, Unity Catalog: Standard_D4ds_v5, 1 DBU/h.

The screenshot shows the 'App registrations' section of the Azure Active Directory (Azure AD) portal under the 'Default Directory' tenant. The 'Owned applications' tab is selected. The interface includes:

- Sidebar navigation: Overview, Preview features, Diagnose and solve problems, Manage (Users, Groups, External identities, Roles and administrators, Administrative units, Delegated admin partners, Enterprise applications, Devices), App registrations (Identity Governance, Application proxy, Custom security attributes, Licenses, Cross-tenant synchronization).
- Header: Microsoft Azure, Upgrade, Search resources, services, and docs, Copilot, Home > Default Directory.
- Content area: A message about the deprecation of ADAL and Graph starting June 30th, 2020. It also states that the account isn't listed as an owner of any applications and provides links to view all applications in the directory or from personal accounts.

Transformation

The screenshot shows the Microsoft Azure 'Register an application' interface. The user has entered 'awproject_app' as the name. The 'Supported account types' section includes options for organizational accounts, Microsoft accounts, and personal Microsoft accounts. A 'Redirect URI (optional)' field is present, and a note states it's optional but required for most scenarios. A 'Register' button is at the bottom.

Your access to a Databricks workspace is controlled through Microsoft Entra ID.

The screenshot shows the Microsoft Azure 'Overview' page for the app 'awproject_app'. It displays basic information such as the display name, application (client) ID, object ID, and directory (tenant) ID. It also shows supported account types ('My organization only'). Two informational banners are present: one about the new improved registration and another about the end of support for ADAL and AAD Graph.

The workspace admin grants you permissions by adding your Entra ID account to the workspace or to an Azure resource role (like *Contributor*, *Reader*, etc.).

- If you want to access data in Azure Storage without using an access key, Databricks can authenticate through Entra ID using your logged-in identity or a service principal.

Microsoft Azure Upgrade Search resources, services, and docs (G+) Copilot ... ramamsaini1297@gmail... DEFAULT DIRECTORY (RAMAMS...)

Home > Default Directory | App registrations > awproject_app

awproject_app | Certificates & secrets

Search Got feedback?

- Overview
- Quickstart
- Integration assistant
- Diagnose and solve problems
- Manage
 - Branding & properties
 - Authentication (Preview)
 - Certificates & secrets**
 - Token configuration
 - API permissions
 - Expose an API
 - App roles
 - Owners
 - Roles and administrators
 - Manifest
- Support + Troubleshooting

Credentials enable confidential applications to identify themselves to the authentication service when receiving tokens at a web addressable location (using an HTTPS scheme). For a higher level of assurance, we recommend using a certificate (instead of a client secret) as a credential.

Application registration certificates, secrets and federated credentials can be found in the tabs below.

Certificates (0)	Client secrets (0)	Federated credentials (0)								
A secret string that the application uses to prove its identity when requesting a token. Also can be referred to as application password.										
New client secret <table border="1"> <thead> <tr> <th>Description</th> <th>Expires</th> <th>Value</th> <th>Secret ID</th> </tr> </thead> <tbody> <tr> <td>No client secrets have been created for this application.</td> <td></td> <td></td> <td></td> </tr> </tbody> </table>			Description	Expires	Value	Secret ID	No client secrets have been created for this application.			
Description	Expires	Value	Secret ID							
No client secrets have been created for this application.										

Add or remove favorites by pressing Cmd + Shift + F

Microsoft Azure Upgrade Search resources, services, and docs (G+) Copilot ... ramamsaini1297@gmail... DEFAULT DIRECTORY (RAMAMS...)

Home > Default Directory | App registrations > awproject_app

awproject_app | Certificates & secrets

Search Got feedback?

Got a second to give us some feedback? →

Credentials enable confidential applications to identify themselves to the authentication service when receiving tokens at a web addressable location (using an HTTPS scheme). For a higher level of assurance, we recommend using a certificate (instead of a client secret) as a credential.

Application registration certificates, secrets and federated credentials can be found in the tabs below.

Certificates (0)	Client secrets (1)	Federated credentials (0)								
A secret string that the application uses to prove its identity when requesting a token. Also can be referred to as application password.										
New client secret <table border="1"> <thead> <tr> <th>Description</th> <th>Expires</th> <th>Value</th> <th>Secret ID</th> </tr> </thead> <tbody> <tr> <td>awprojectsecret</td> <td>09/02/2026</td> <td>1ji8Q~pn7tNCdu4HOr0u0C~Nij2024PP2...</td> <td>738f24fd-f201-4f6a-93c5-c294b03498bd</td> </tr> </tbody> </table>			Description	Expires	Value	Secret ID	awprojectsecret	09/02/2026	1ji8Q~pn7tNCdu4HOr0u0C~Nij2024PP2...	738f24fd-f201-4f6a-93c5-c294b03498bd
Description	Expires	Value	Secret ID							
awprojectsecret	09/02/2026	1ji8Q~pn7tNCdu4HOr0u0C~Nij2024PP2...	738f24fd-f201-4f6a-93c5-c294b03498bd							

Add or remove favorites by pressing Cmd + Shift + F

1. Creating the application

Save this info after registering application

We create the certificate and secret

2. Assigning role to this application so it access the datalake

Home → storage tab → Access control (IAM) → Add → storage blob contributor(read_write permission)

A role definition is a collection of permissions. You can use the built-in roles or you can create your own custom roles. [Learn more](#)

Copilot can help pick a role

Name	Description	Type	Category	Details
Defender CSPM Storage Data Scanner	Grants access to read blobs and files. This role is used by the data scanner of Defender CSPM.	BuiltinRole	None	View
Defender for Storage Data Scanner	Grants access to read blobs and update index tags. This role is used by the data scanner of Defender for Storage.	BuiltinRole	None	View
Storage Blob Data Contributor	Allows for read, write and delete access to Azure Storage blob containers and data	BuiltinRole	Storage	View
Storage Blob Data Owner	Allows for full access to Azure Storage blob containers and data, including assigning POSIX access control.	BuiltinRole	Storage	View
Storage Blob Data Reader	Allows for read access to Azure Storage blob containers and data	BuiltinRole	Storage	View
Storage Blob Delegator	Allows for generation of a user delegation key which can be used to sign SAS tokens	BuiltinRole	Storage	View

3.

Review + assign Previous Next

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

Select members:

Name	Type
awproject_app	Application

Select Close

First notebook in Azure databricks

Databricks→workspace→create folder

Inside that folder→create the notebook

Transformation

Turn on our cluster →connect→select ur cluster that created in compute

Refer to notebook created

<https://adb-1557609439492944.4.azuredatabricks.net/editor/notebooks/4280593522547844?o=1557609439492944#command/5655344618087748>

Synapse analytics

4. Serving Data with Azure Synapse Analytics

4.1 Introduction to Synapse Analytics

Azure Synapse Analytics is a unified platform for enterprise analytics, combining data warehousing, big data analytics, and data integration. It integrates with ADF and Spark, providing a single environment for data professionals.

4.2 Setting up Synapse Analytics

1. Create a Synapse workspace. You will also need to create a default storage account for it.
2. Assign the Synapse workspace's managed identity the Storage Blob Data Contributor role on your primary data lake to allow it to access the data

Whenever create synapse analytics workspace need to create the default storage account as well.

Then just create

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

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

Create Synapse workspace

***Basics** ***Security** Networking Tags Review + create

Create a Synapse workspace to develop an enterprise analytics solution in just a few clicks.

Project details

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

Subscription * Azure subscription 1 The Synapse and SQL resource providers are now registered with this subscription.

Resource group * AWPROJECT Create new

Managed resource group synapse-rg

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 * synapse-aw-ws

Region * (US) East US

Select Data Lake Storage Gen2 * From subscription

Account name * (New) synapselfstoragedefault123 Create new

Review + create < Previous Next: Security >

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

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

Create Synapse workspace

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 * synapse-aw-ws

Region * (US) East US

Select Data Lake Storage Gen2 * From subscription

Account name * (New) synapselfstoragedefault123 Create new

File system name * (New) defaultfilesystem Create new

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

Info 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 your workspace, perform these tasks:

- Assign other users to the **Contributor** role on workspace
- Assign other users the appropriate [Synapse RBAC roles](#) using Synapse Studio
- Assign yourself and other users to the **Storage Blob Data Contributor** role on the storage account

[Learn more](#)

Review + create < Previous Next: Security >

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

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

Create Synapse workspace

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 *

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.

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 encryption

⚠ Double encryption configuration cannot be changed after opting into using a customer-managed key at the time of workspace creation.

Choose to encrypt all data at rest in the workspace with a key managed by you (customer-managed key). This will provide double encryption with encryption at the infrastructure layer that uses platform-managed keys. [Learn more](#)

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

Unified platform bec we can combine

Adf + spak+warehousing

Can create pipeline inside integrate section of azure

Microsoft Azure | Synapse Analytics > synapseanalyticsaw

We use optional cookies to provide a better experience. [Learn more](#)

Accept Reject More options

Integrate Pipeline 1

Filter resources by name

Pipelines Pipeline 1

Activities

Synapse Move and transform Azure Data Explorer Azure Function Batch Service Databricks Data Lake Analytics General HDInsight Iteration & conditionals Machine Learning

Properties

Name * Pipeline 1

Description

Annotations

Parameters Variables Settings Output

New

These are spark pool having same ui like databricks but it is not databricks we call it spark pool . as both uses synapse cluster

The screenshot shows the Microsoft Azure Synapse Analytics Develop workspace. On the left, there's a navigation sidebar with options: Home, Data, Develop, Integrate, Monitor, and Manage. The main area is titled 'Develop' and shows a 'Notebooks' section with 'Notebook 1'. A warning message at the top right says: 'Please select a Spark pool to attach before running cell!'. To the right, there's a 'Properties' panel for 'Notebook 1' with tabs for General and Related (0). The General tab shows the Name as 'Notebook 1' and a Description field. It also includes sections for Type (.ipynb notebook), Size (618 bytes), Notebook settings (checkboxes for 'Include cell output when saving' and 'Enable unpublished notebook reference'), and Session (Configure session).

Data warehouse

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace. The left sidebar has the same navigation options as the Develop workspace. The main area is titled 'Data' and shows a 'Workspace' section with 'Notebook 1'. A warning message at the top right says: 'Please select a Spark pool to attach before running cell!'. To the right, there's a 'Properties' panel for 'Notebook 1' with tabs for General and Related (0). The General tab shows the Name as 'Notebook 1' and a Description field. It also includes sections for Type (.ipynb notebook), Size (618 bytes), Notebook settings (checkboxes for 'Include cell output when saving' and 'Enable unpublished notebook reference'), and Session (Configure session).

4.3 The Lakehouse Concept

Synapse Analytics supports a

Lakehouse concept, which combines a data lake with a data warehouse.

- Serverless SQL Pool: This option allows you to query data directly from your data lake without storing it in a traditional database. An abstraction layer (metadata) is created on top of the data in the data lake, which is then used by the SQL pool to process queries.
- Dedicated SQL Pool: This is a traditional data warehouse where data is physically stored in the database.

Lake db

It gives ability to use sparks. In spark when create db then it is called lake house or lake db but in databricks we called as lake house.

If creating tables using spark pool then called lake db in synapse and in databricks we called as lake house

Create the data warehouse

Synapse db synapsedatalake

Both are azure products so no need to include any 3rd party app., not need to register any kind of app.

Directly access data lake using synapse-analytics by allowing workspace the permission. By default synapse analytics has credentials and we just need to assign the access , assign the role to that cred

Allow synapse workspace to access data using the identity, using the cred that synapse workspace or any other azure resource has by default

IQ

Name of that identity: managed Identity or system managed identity

Add Role

Storage data lake→IAM→ select storage blob data contributor→ next

Creating sql scripts

Develop→new→ sql script

First need to create db

Data→ new→ sql db→2options serverless sql pool and dedicated sql pool

Dedicated sql pool: traditional way of storing data where data actually resides in db.

Traditional db on cloud and optimized for query reads for big data for data warehousing

Serverless sql pool: like lakehouse where we do not store the data in db.

Lakehouse concept

Data resides in data lake why? Bec its cheap

Having data in csv format in data lake and want and need to create the serveless db

We want to apply select statement on this file but do not want to store the data in db traditionally

Data resides in data lake and it creates the abstraction layer (metadata layer) stores the metadata like columns, headers, all info regarding data

Whenever user queries the data we write select statement from my table

It will apply this metadata to the columns on the data stored in data lake

And serveless pool does all the work behind the scene

It pulls the data and applies the metadata layer and returns the result

That's the lake house concept where we use our data lake but at the same time we want our data to perform as a data warehouse

Data warehouse + data lake = lake house

4.4 Creating the Gold Layer

1. In Synapse Analytics, go to the Develop section and create a new SQL Script.

2. Create a schema (e.g., `gold`) to organize your objects.
3. Use the `OPENROWSET()` function to query data from the `silver` layer in your data lake.
4. Create Views on top of the silver layer data. A view is a saved query that doesn't store data itself.
5. Use CETAS (Create External Table As Select) to create external tables in the `gold` layer. This command reads data from the silver layer and writes it as a highly curated dataset in the `gold` folder of your data lake.
6. Before creating external tables, you need to create a master key, database-scoped credentials, external data sources for the silver and gold locations, and external file formats (e.g., `PARQUET`).

Pick the data from silver layer

Using `openrowset()`:- helps to apply abstraction layer on the data residing in data lake

As we assign role of storage blob contributor to our synapse workspace , need to assign one more to role but to urself

Whenever query data residing in data lake , we also should have the permission to access the data

datalake→ IAM→ add→ storage blob contributor→user, grp or service contributor→select ur mail id → assign

(gold layer synapse analytics code)

External tables

External tables: that keep the data

Managed table: that do not store the data databricks or data lake etc env. store the data

Steps to create the external table in synapse analytics

1. Create Credentials

2. External data source(whenever need to pick data then we need to mention the url again and again so to avoid this we create external data source. Keep url at container level and rest of the url at location)
 3. External file format (csv, json, parquet)
1. Need to create master key for this db

We created the view on the top of silver layer , we read data from silver layer and wrote in gold layer using CETAS, and it creates external table over it.

CETAS(create external table as select)

Views: not store the data, it just the query when create views then it creates the views then it stores the query not the data

Ext table: we store data in data lake → gold layer→folder

To establish the connection between power BI and synapse workspace, we use SQL endpoints

Connect to power BI

synapseAnalytics→copy serverless SQL endpoint

Power bi→get data→Azure→synapse analytics sql→ add sql endpoint in server

userName: adminRaman

Password: Raman@!23

Microsoft Azure | Synapse Analytics > synapseanalyticsaw

We use optional cookies to provide a better experience. [Learn more](#)

Accept | Reject | More options | X

Synapse live | Validate all | Publish all

Develop + <

Filter resources by name

SQL scripts 5

- create schema
- create views for gold layer
- external table
- fetch data from views
- script1

create schema x external table fetch data from views create views for gold...
Run Undo Publish Query plan Connect to Built-in Use database awdatabase ...

```
1 CREATE SCHEMA gold
```

Microsoft Azure | Synapse Analytics > synapseanalyticsaw

We use optional cookies to provide a better experience. [Learn more](#)

Accept | Reject | More options | X

Synapse live | Validate all | Publish all

Develop + <

Filter resources by name

SQL scripts 5

- create schema
- create views for gold layer
- external table
- fetch data from views
- script1

create schema x external table fetch data from views create views for gold...
Run Undo Publish Query plan Connect to Built-in Use database awdatabase ...

```
1 --- create view calender
2
3 CREATE VIEW gold.calender
4 AS
5 SELECT *
6 FROM OPENROWSET (
7 | BULK 'https://adventureworkstoraged1.dfs.core.windows.net/silver/Calender/' , FORMAT = 'PARQUET'
8 ) as Query1
9
10 -----
11 --- create view customer
12
13 CREATE VIEW gold.customers AS SELECT * FROM OPENROWSET (
14 | BULK 'https://adventureworkstoraged1.dfs.core.windows.net/silver/Customers/' , FORMAT = 'PARQUET'
15 ) as Query1
16
17 -----
18 --- create view products
19
20 CREATE VIEW gold.products AS SELECT * FROM OPENROWSET (
21 | BULK 'https://adventureworkstoraged1.dfs.core.windows.net/silver/Products/' , FORMAT = 'PARQUET'
22 ) as Query1
23
24 -----
25 --- create view returns
26
27 CREATE VIEW gold.returnProduct AS SELECT * FROM OPENROWSET (
28 | BULK 'https://adventureworkstoraged1.dfs.core.windows.net/silver>Returns/' , FORMAT = 'PARQUET'
29 ) as Query1
30
31 -----
32 --- create view sales_2015
33
34 -----
```

Microsoft Azure | Synapse Analytics > synapseanalyticsaw 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 5

- create schema
- create views for gold layer
- external table
- fetch data from views
- script1

```
25 --- create view returns
26
27 CREATE VIEW gold.returnProduct AS SELECT * FROM OPENROWSET (
28   BULK 'https://adventureworkstoraged1.dfs.core.windows.net/silver>Returns/' , FORMAT = 'PARQUET'
29 ) as Query1
30
31 -----
32 --- create view sales 2015
33
34 CREATE VIEW gold.sales2015 AS SELECT * FROM OPENROWSET (
35   BULK 'https://adventureworkstoraged1.dfs.core.windows.net/silver\Sales2015/' , FORMAT = 'PARQUET'
36 ) as Query1
37
38 -----
39 --- create view sales
40
41 CREATE VIEW gold.sales AS SELECT * FROM OPENROWSET (
42   BULK 'https://adventureworkstoraged1.dfs.core.windows.net/silver/Sales/' , FORMAT = 'PARQUET'
43 ) as Query1
44
45 -----
46 --- create view Sub_Categories
47
48 CREATE VIEW gold.Sub_Categories AS SELECT * FROM OPENROWSET (
49   BULK 'https://adventureworkstoraged1.dfs.core.windows.net/silver\Sub_Categories/' , FORMAT = 'PARQUET'
50 ) as Query1
51
52
53 -----
54 --- create view Territories
55
56 CREATE VIEW gold.Territories AS SELECT * FROM OPENROWSET (
57   BULK 'https://adventureworkstoraged1.dfs.core.windows.net/silver\Territories/' , FORMAT = 'PARQUET'
58 ) as Query1
```

Microsoft Azure | Synapse Analytics > synapseanalyticsaw 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 5

- create schema
- create views for gold layer
- external table
- fetch data from views
- script1

```
1 | SELECT * FROM gold.customers
| Publish (⌘+S)
```

```

1 CREATE MASTER KEY ENCRYPTION BY PASSWORD = 'Password@123'
2
3 -- create db credentials
4
5 CREATE DATABASE SCOPED CREDENTIAL cred_raman
6 WITH IDENTITY='Managed Identity'
7
8 CREATE DATABASE SCOPED CREDENTIAL silver_sales_cred
9 WITH IDENTITY='Managed Identity'
10
11 -- create external data source (silver data source)
12 -- create 2 silver(read data) and gold(push data)
13
14 CREATE EXTERNAL DATA SOURCE silver_source WITH (
15     LOCATION = 'https://adventureworkstoraged1.blob.core.windows.net/silver',
16     CREDENTIAL = cred_raman
17 )
18
19 CREATE EXTERNAL DATA SOURCE silver_sales_source WITH (
20     LOCATION = 'https://adventureworkstoraged1.blob.core.windows.net/silver',
21     CREDENTIAL = silver_sales_cred
22 )
23
24 CREATE EXTERNAL DATA SOURCE gold_source WITH (
25     LOCATION = 'https://adventureworkstoraged1.blob.core.windows.net/gold',
26     CREDENTIAL = cred_raman
27 )
28
29 -- External file format
30
31 CREATE EXTERNAL FILE FORMAT parquet_format WITH(
32     FORMAT_TYPE= PARQUET,
33     DATA_COMPRESSION= 'org.apache.hadoop.io.compress.SnappyCodec'
34 )

```

```

28
29 -- External file format
30
31 CREATE EXTERNAL FILE FORMAT parquet_format WITH(
32     FORMAT_TYPE= PARQUET,
33     DATA_COMPRESSION= 'org.apache.hadoop.io.compress.SnappyCodec'
34 )
35
36
37
38 -- craete external table as external sales
39 CREATE EXTERNAL TABLE gold.extsales WITH(
40     LOCATION = 'extsales',
41     DATA_SOURCE = gold_source,
42     FILE_FORMAT = parquet_format
43 ) AS SELECT * FROM gold.sales2015
44
45 CREATE EXTERNAL TABLE gold.extCompletesales WITH(
46     LOCATION = 'extTotalSales',
47     DATA_SOURCE = gold_source,
48     FILE_FORMAT = parquet_format
49 ) AS SELECT * FROM gold.sales
50
51 CREATE EXTERNAL TABLE gold.combinedsales WITH(
52     LOCATION = 'extSales',
53     DATA_SOURCE = silver_sales_source,
54     FILE_FORMAT = parquet_format
55 ) AS SELECT * FROM gold.sales
56
57 SELECT * FROM gold.extCompletesales
58
59 SELECT * FROM gold.combinedsales
60
61 SELECT * FROM gold.extsales

```

5. Data Visualization with Power BI

The final step is to serve the curated data to stakeholders. This can be done using Power BI for analysis and visualization.

1. In Synapse Analytics, copy the serverless SQL endpoint.
2. In Power BI Desktop, select Get Data > Azure > Azure Synapse Analytics SQL.
3. Paste the SQL endpoint into the Server field.
4. Use the SQL server admin username and password to establish the connection.
5. You can now access the views and external tables in your gold layer to create reports and dashboards.