Assessment Title
Azure Practical Scenario Evaluation

CCC602| Assessment-3

The Student Name | Geni Bahar
Student ID: 27085022

Tutor/Accessor | Ovesh Vohra

Programme
Diploma in Cloud Engineering and Cyber Security (120 Credits)

Course
CCC602: Data Engineering in the Public Cloud
(Level 6, 30 Credits)

Date: 17th January 2026

Table of Contents

Task 1: Cloud Vendor Evaluation (Theoretical Evaluation)

Question 1: Comparative Evaluation: Microsoft Azure vs. Amazon Web Services (AWS)

Big Data and Analytics Capabilities

- Both Microsoft Azure and AWS offer mature big data and analytics services, but they differ in how seamlessly these services work together. From my learning and hands-on experience during this project, Azure provides a well-connected analytics ecosystem that includes Azure Data Factory, Azure Data Lake Storage Gen2, Azure Databricks, and Azure Synapse Analytics. I found that these services were easier to connect and manage together when building an end-to-end data pipeline. These services are designed to work together, which reduces setup complexity and learning overhead.
- AWS offers equivalent services such as Amazon S3, AWS Glue, Amazon EMR, and Amazon Redshift. While these services are powerful and widely used, they often require additional configuration and integration effort to achieve the same end-to-end workflow that Azure provides more naturally. From a learning and implementation perspective, Azure felt more structured and easier to manage for a complete data pipeline.

Integration with Business Intelligence (BI) Tools

- Integration with BI tools is a key requirement for analysing and presenting insights from the Tokyo Olympic dataset. Azure integrates natively with Power BI, which I found particularly useful as it allows simple connectivity and real-time data visualisation without requiring complex additional configuration. Since Power BI is part of the Microsoft ecosystem, it works smoothly with Azure services without additional connectors or complex configurations.
- AWS supports BI tools such as Amazon QuickSight and Tableau. While these tools are effective, integration with Power BI is less direct and often requires extra setup. For organisations already using Microsoft products, Azure offers a more seamless and user-friendly BI experience.

Cost Structures and Pricing Comparison

- Azure provides flexible pricing models such as pay-as-you-go, reserved capacity, and serverless options. Azure Data Lake Storage Gen2 offers low-cost storage for large datasets, and Azure Synapse serverless SQL allows querying data without provisioning dedicated resources, which helps control costs. Azure Databricks also allows scaling compute resources only when required.
- AWS offers similar pricing models for Amazon S3 and EMR; however, EMR clusters can become expensive if not carefully managed. In comparison, Azure's serverless and consumption-based services provide better cost transparency for smaller projects and student-level environments.

Scalability and Automation

- Both platforms support scalability and automation through Infrastructure as Code. Azure supports automation through ARM templates, which helped me understand how infrastructure can be deployed in a repeatable and consistent way without manually creating each resource through the Azure Portal. AWS uses CloudFormation, which provides similar functionality but has a steeper learning curve.
- Azure's automation tools felt easier to understand and implement, especially when deploying multiple resources such as storage, data factory, and analytics services together.

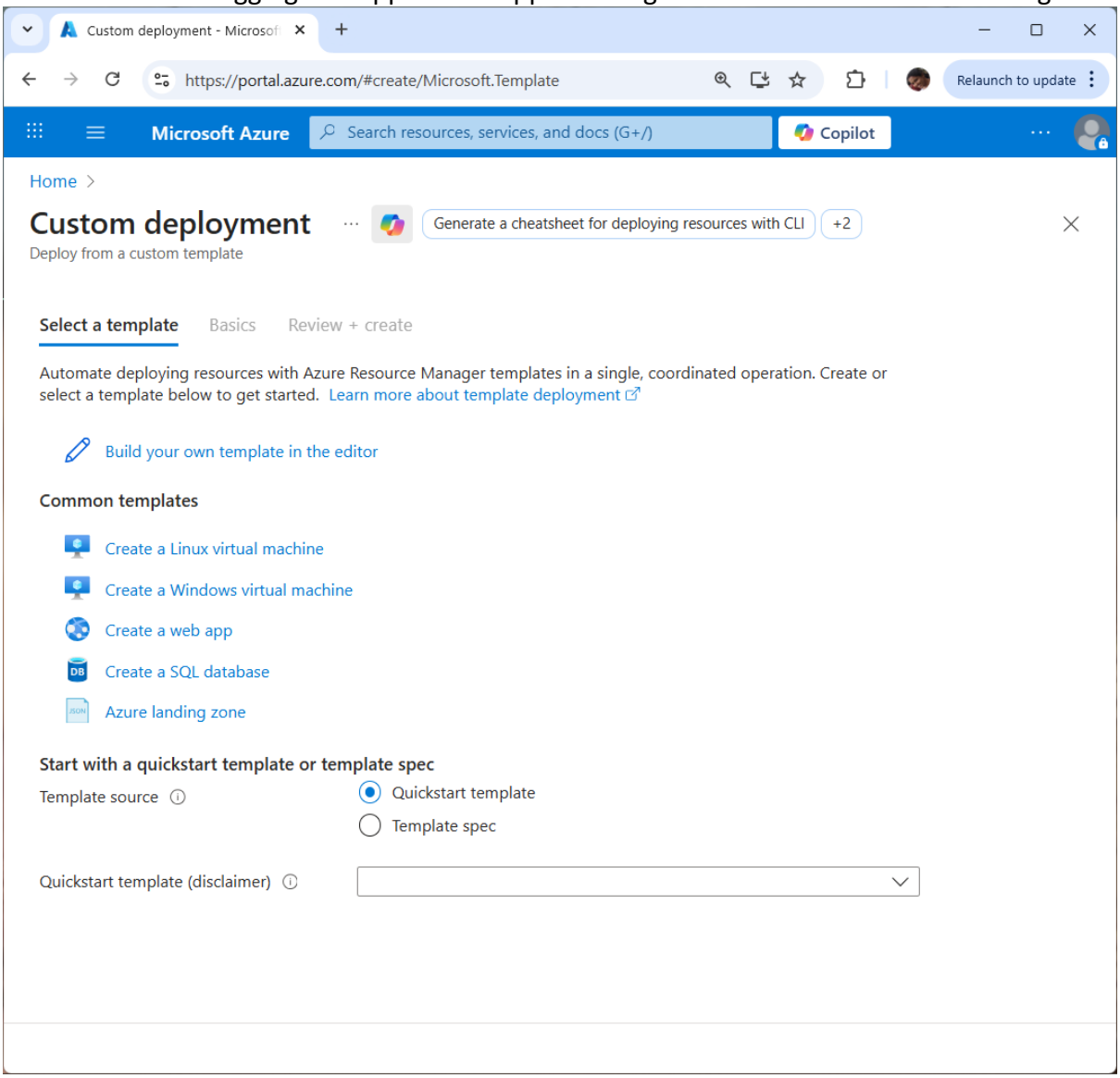
Vendor Selection Report

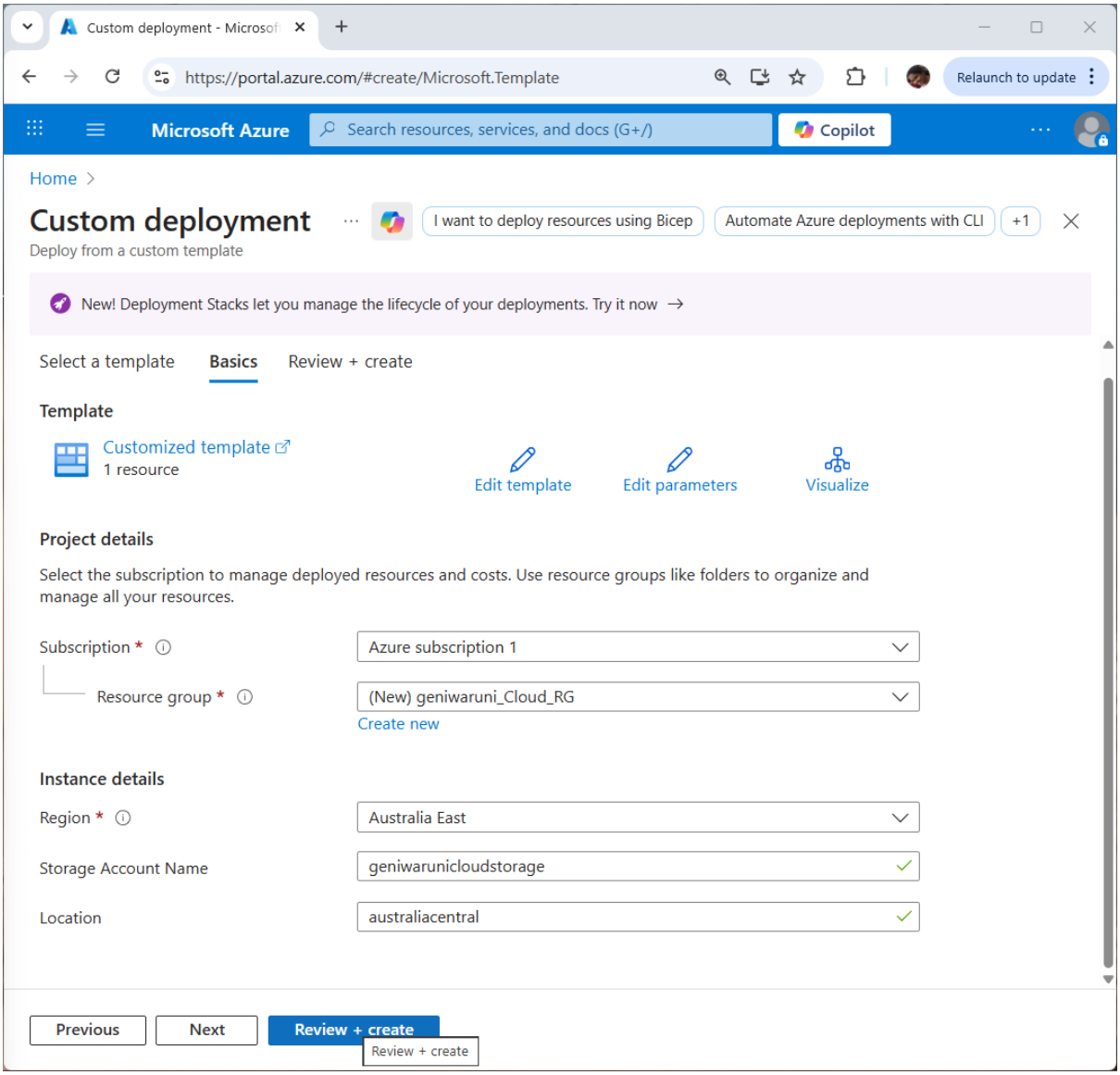
- Microsoft Azure was selected as the preferred cloud platform due to its strong integration across analytics services, native compatibility with Power BI, flexible cost structure, and robust automation capabilities using ARM templates. Azure aligns well with data analytics workloads and the Microsoft ecosystem, making it a practical and efficient choice for building a scalable and cost-effective data engineering solution.

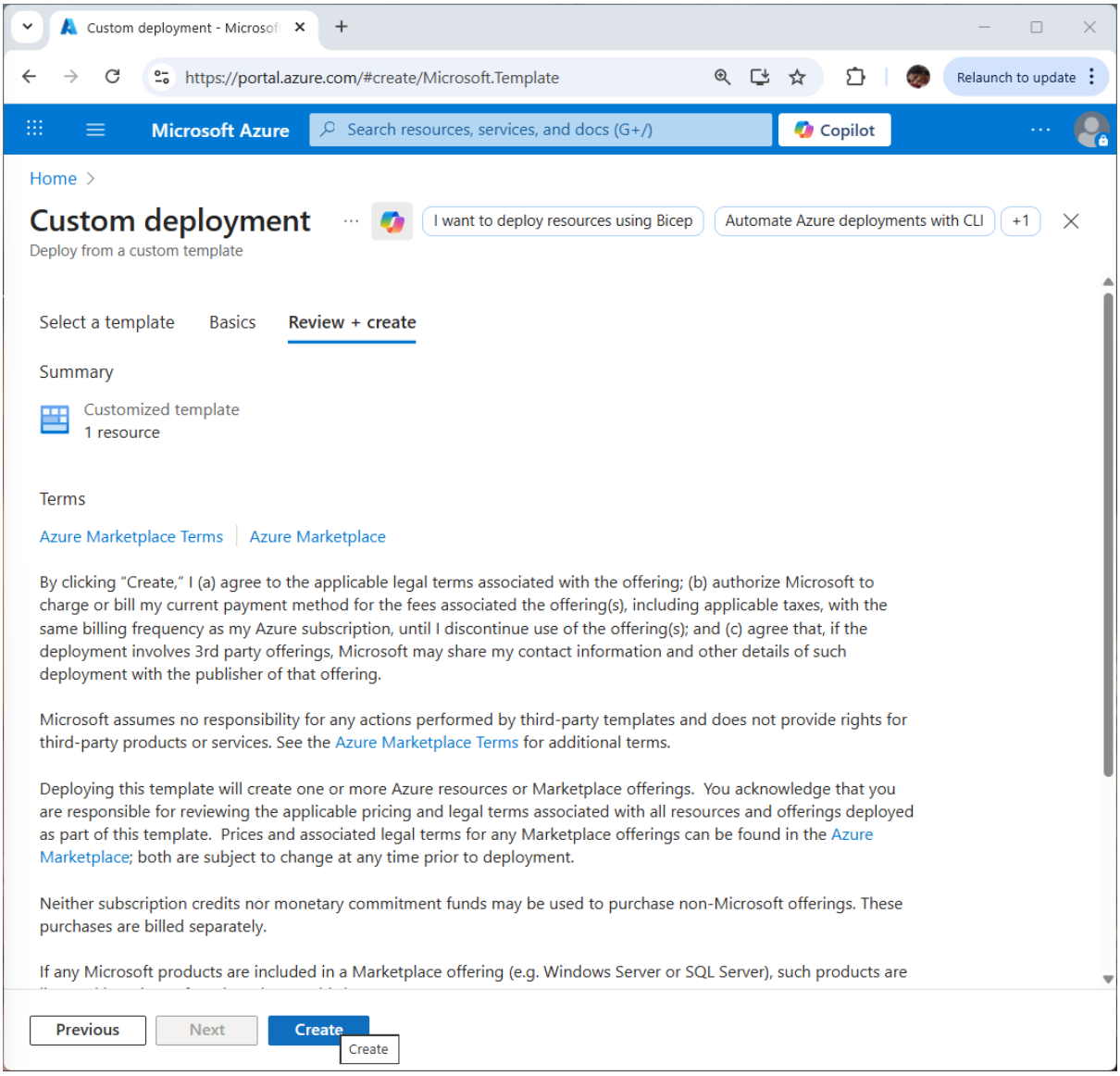
Task 2: Dataset Acquisition and Azure Resource Deployment Using IaC

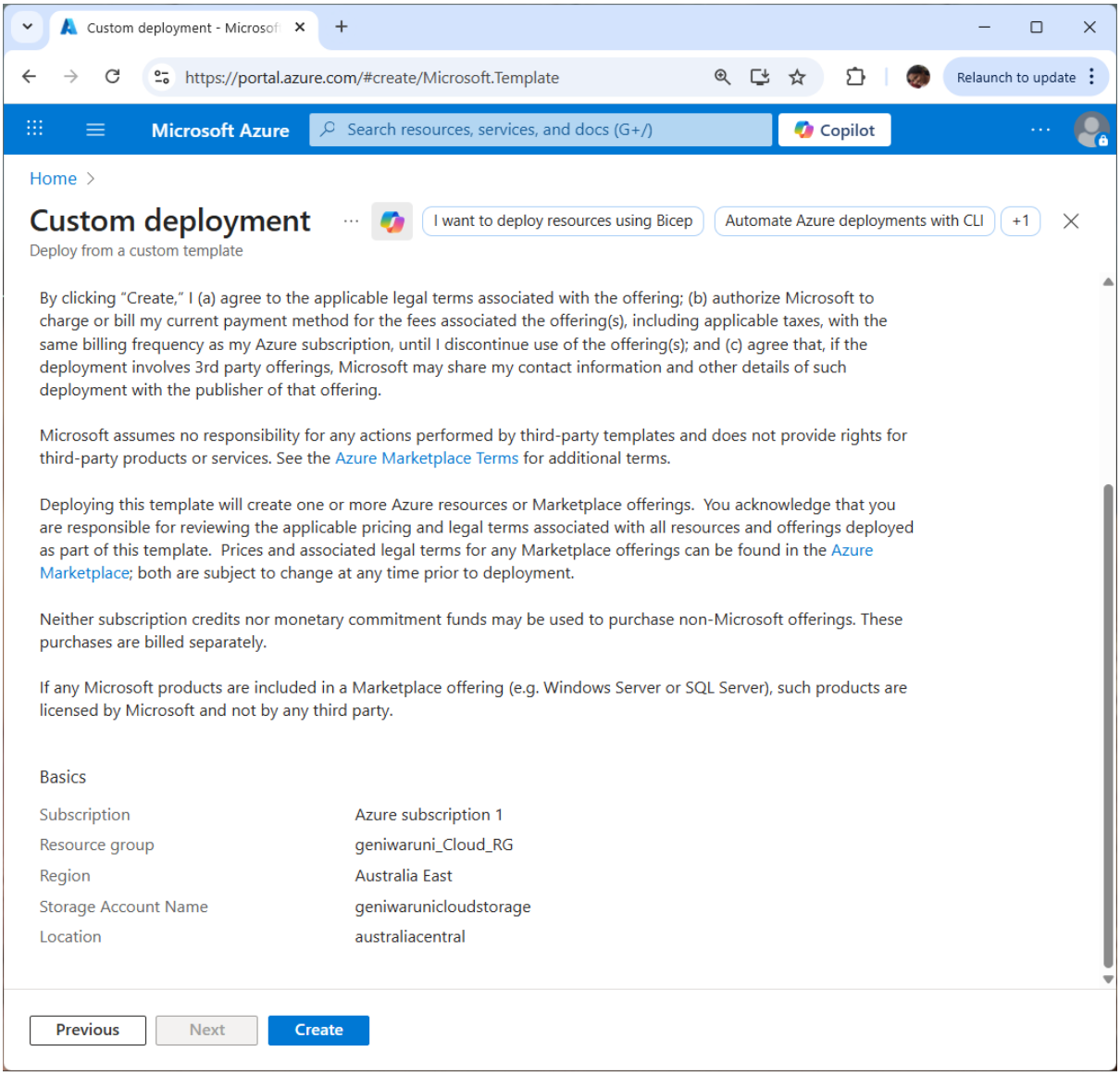
Azure Resource Deployment Using ARM Template

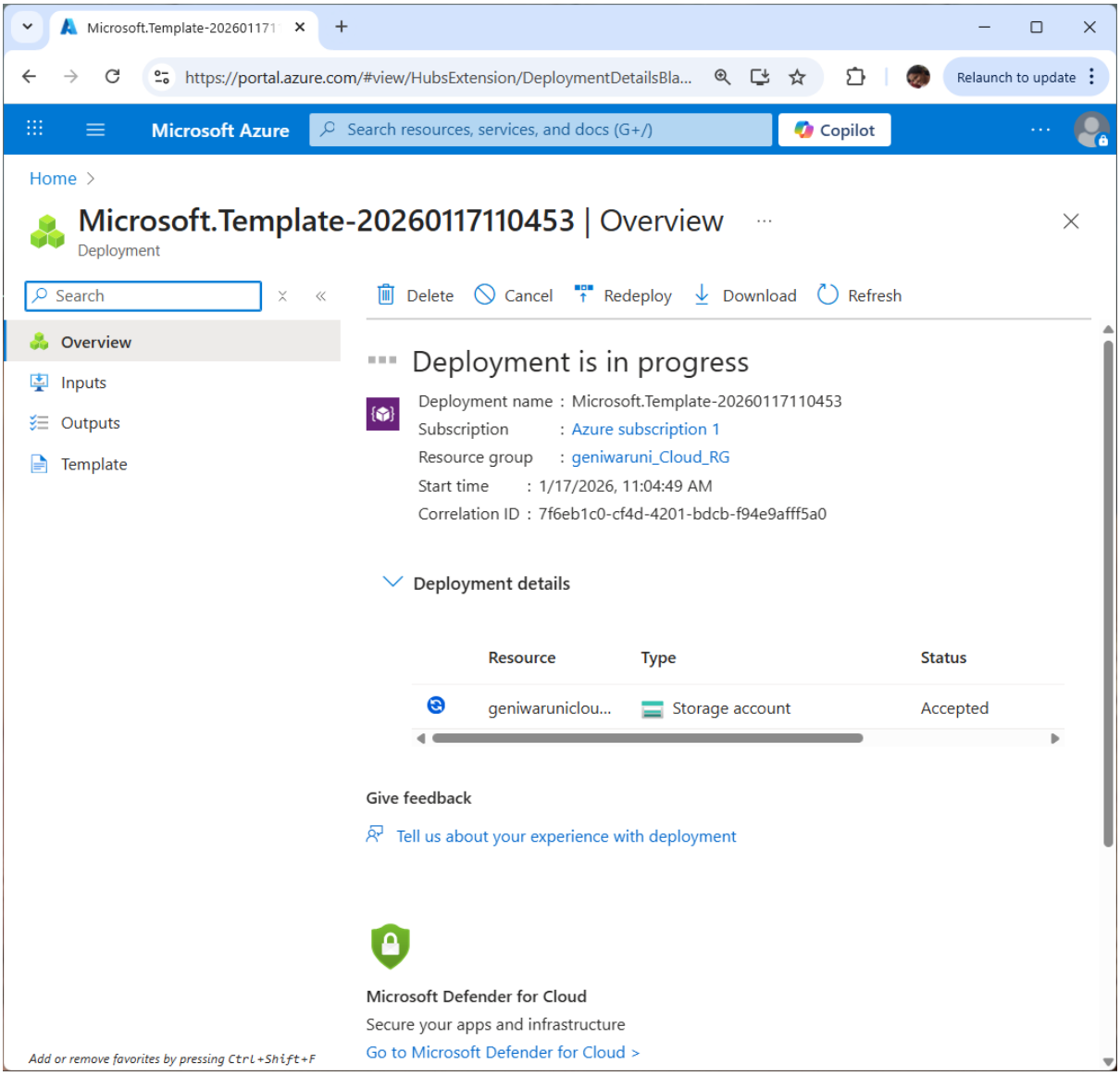
- In this task, core Azure infrastructure was deployed using an Azure Resource Manager (ARM) template to ensure consistency, repeatability, and cost control.
- A resource group named **geni_Cloud_RG** was created. Within this resource group, an Azure Data Lake Storage Gen2 account named **geniadls** was deployed. The storage account was configured with the StorageV2 type and hierarchical namespace enabled to support big data analytics.
- Resource tagging was applied to support cost governance and resource tracking.

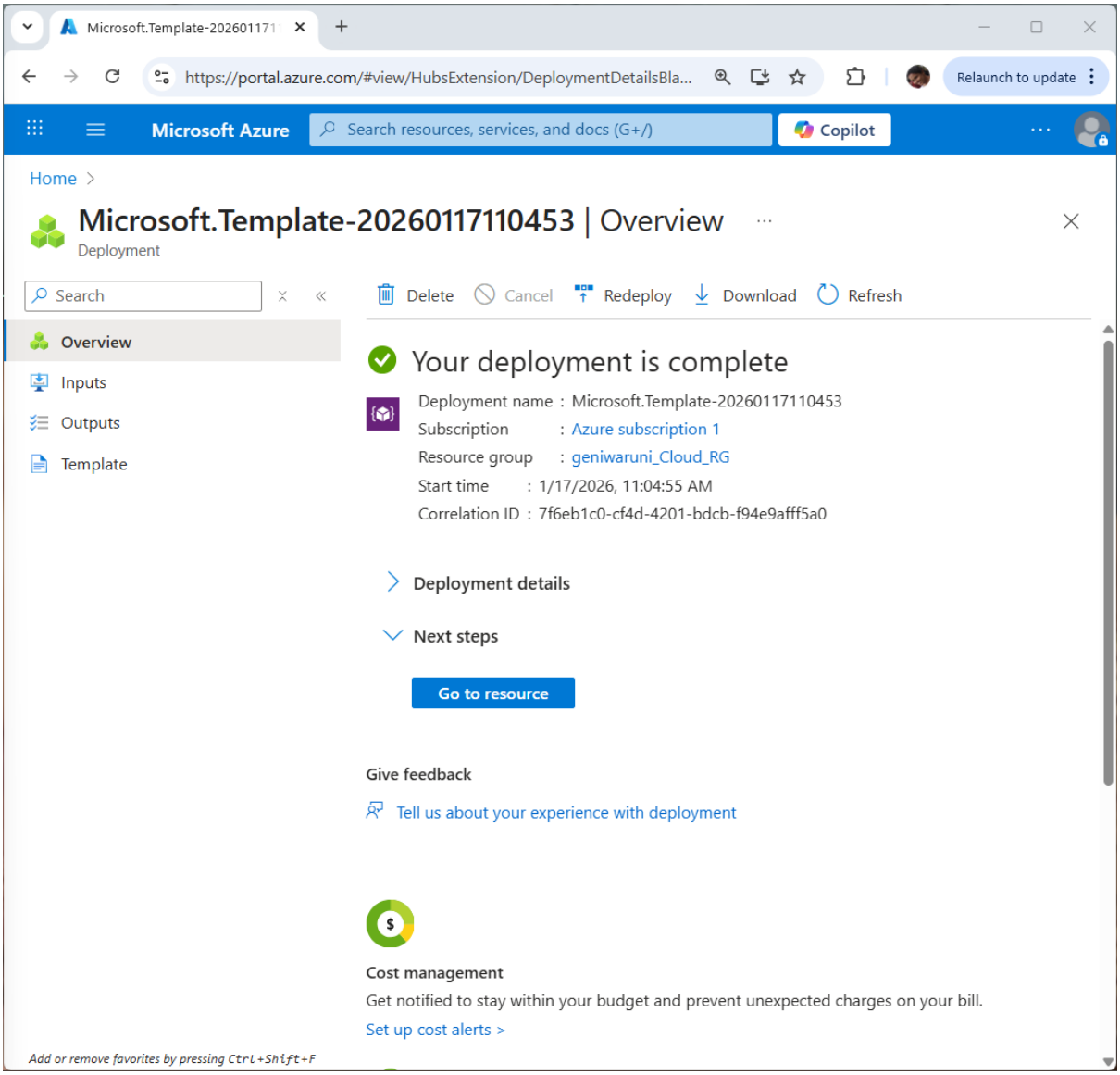


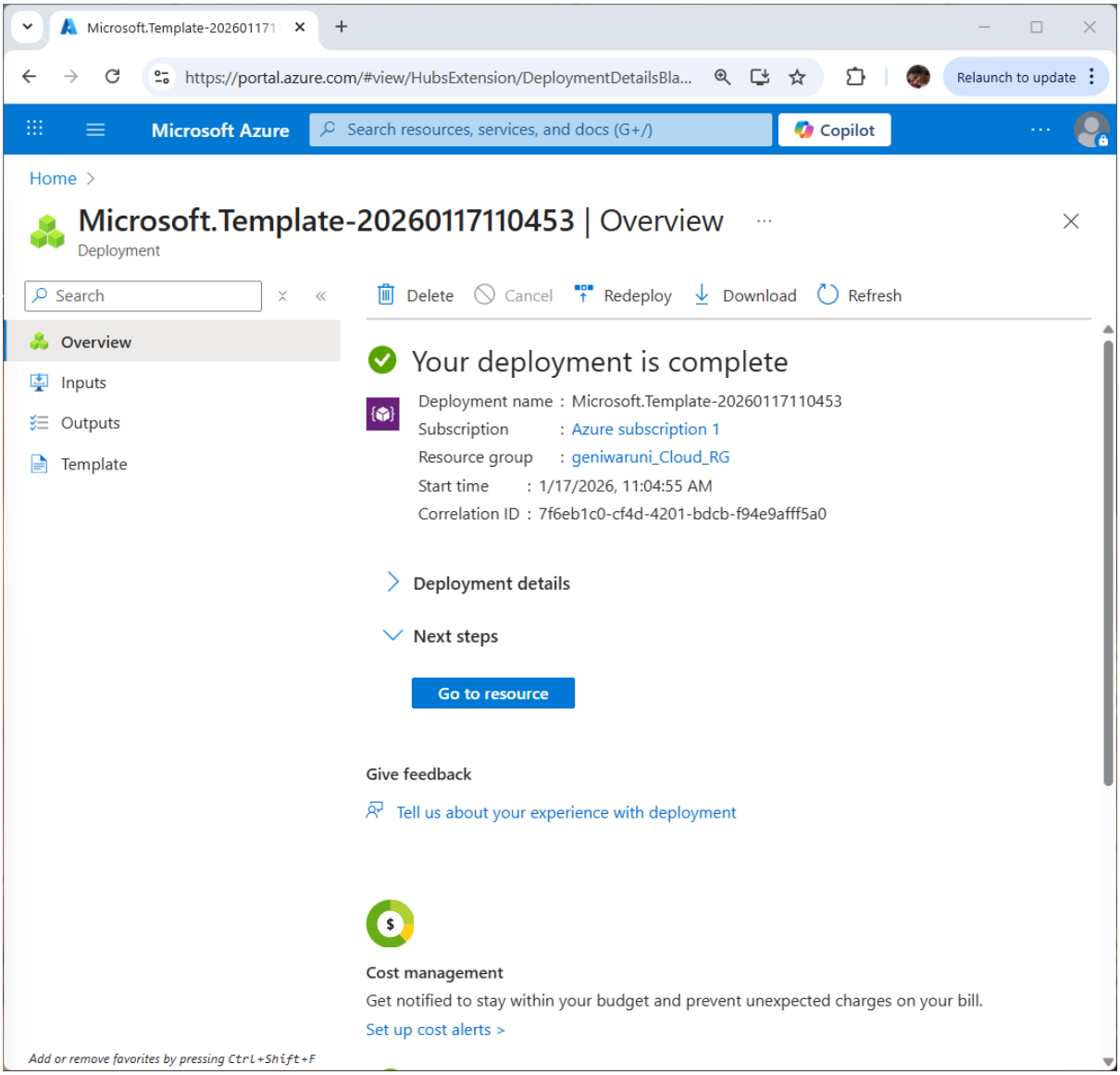

Custom deployment - Microsof
https://portal.azure.com/#create/Microsoft.Template
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Custom deployment
Deploy from a custom template
I want to deploy resources using Bicep
Automate Azure deployments with CLI
+1
New! Deployment Stacks let you manage the lifecycle of your deployments. Try it now →
Select a template
Basics
Review + create
Template
Customized template
1 resource
Edit template
Edit parameters
Visualize
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 *
(New) geniwaruni_Cloud_RG
Create new
Instance details
Region *
Australia East
Storage Account Name
geniwarunicloudstorage
Location
australiacentral
Previous
Next
Review + create
Review + create

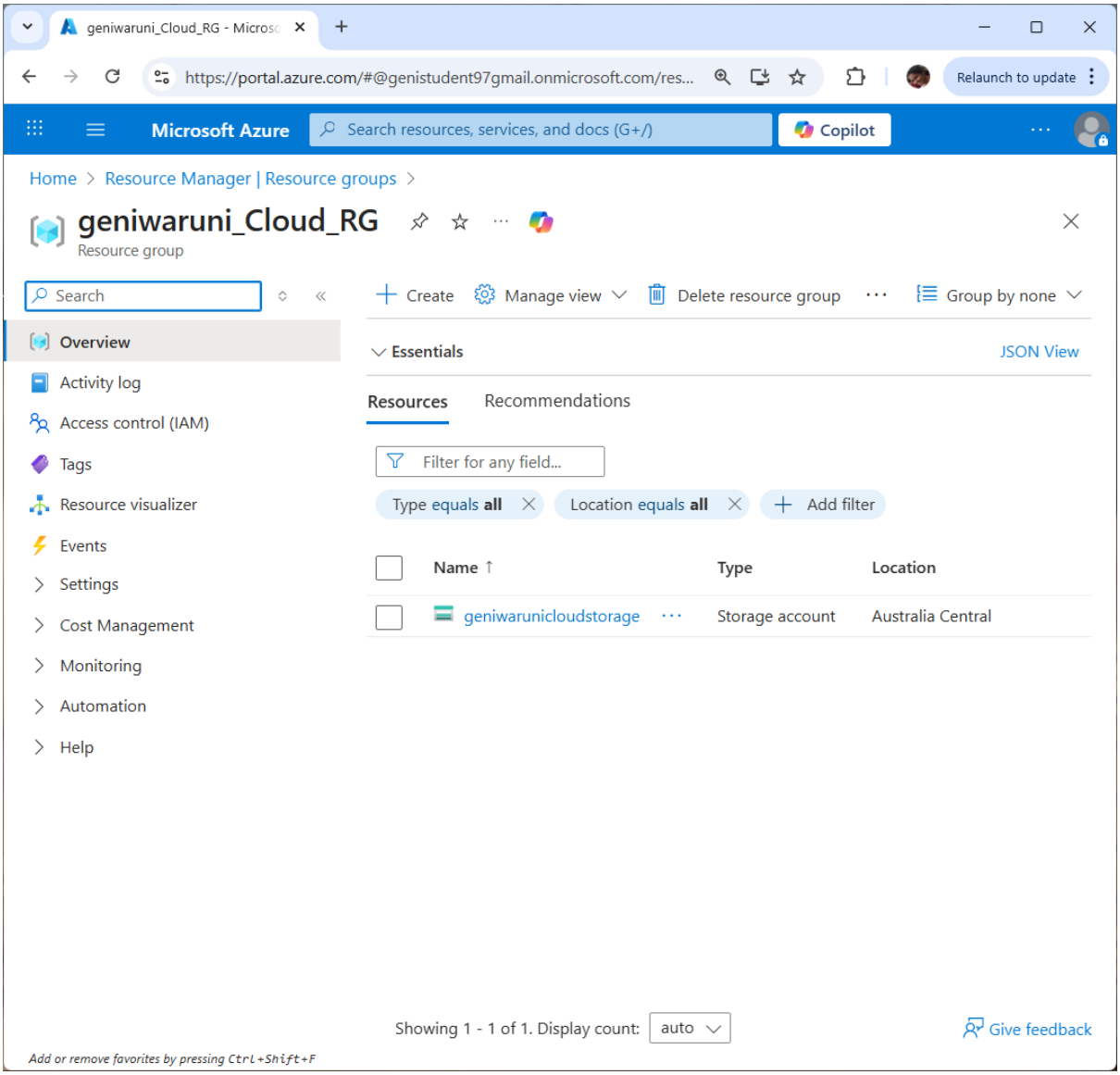

Custom deployment - Microsof
https://portal.azure.com/#create/Microsoft.Template
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Custom deployment
Deploy from a custom template
I want to deploy resources using Bicep
Automate Azure deployments with CLI
+1
Select a template
Basics
Review + create
Summary
Customized template
1 resource
Terms
Azure Marketplace Terms
Azure Marketplace
By clicking "Create," I (a) agree to the applicable legal terms associated with the offering; (b) authorize Microsoft to charge or bill my current payment method for the fees associated the offering(s), including applicable taxes, with the same billing frequency as my Azure subscription, until I discontinue use of the offering(s); and (c) agree that, if the deployment involves 3rd party offerings, Microsoft may share my contact information and other details of such deployment with the publisher of that offering.
Microsoft assumes no responsibility for any actions performed by third-party templates and does not provide rights for third-party products or services. See the Azure Marketplace Terms for additional terms.
Deploying this template will create one or more Azure resources or Marketplace offerings. You acknowledge that you are responsible for reviewing the applicable pricing and legal terms associated with all resources and offerings deployed as part of this template. Prices and associated legal terms for any Marketplace offerings can be found in the Azure Marketplace; both are subject to change at any time prior to deployment.
Neither subscription credits nor monetary commitment funds may be used to purchase non-Microsoft offerings. These purchases are billed separately.
If any Microsoft products are included in a Marketplace offering (e.g. Windows Server or SQL Server), such products are
Previous
Next
Create
Create

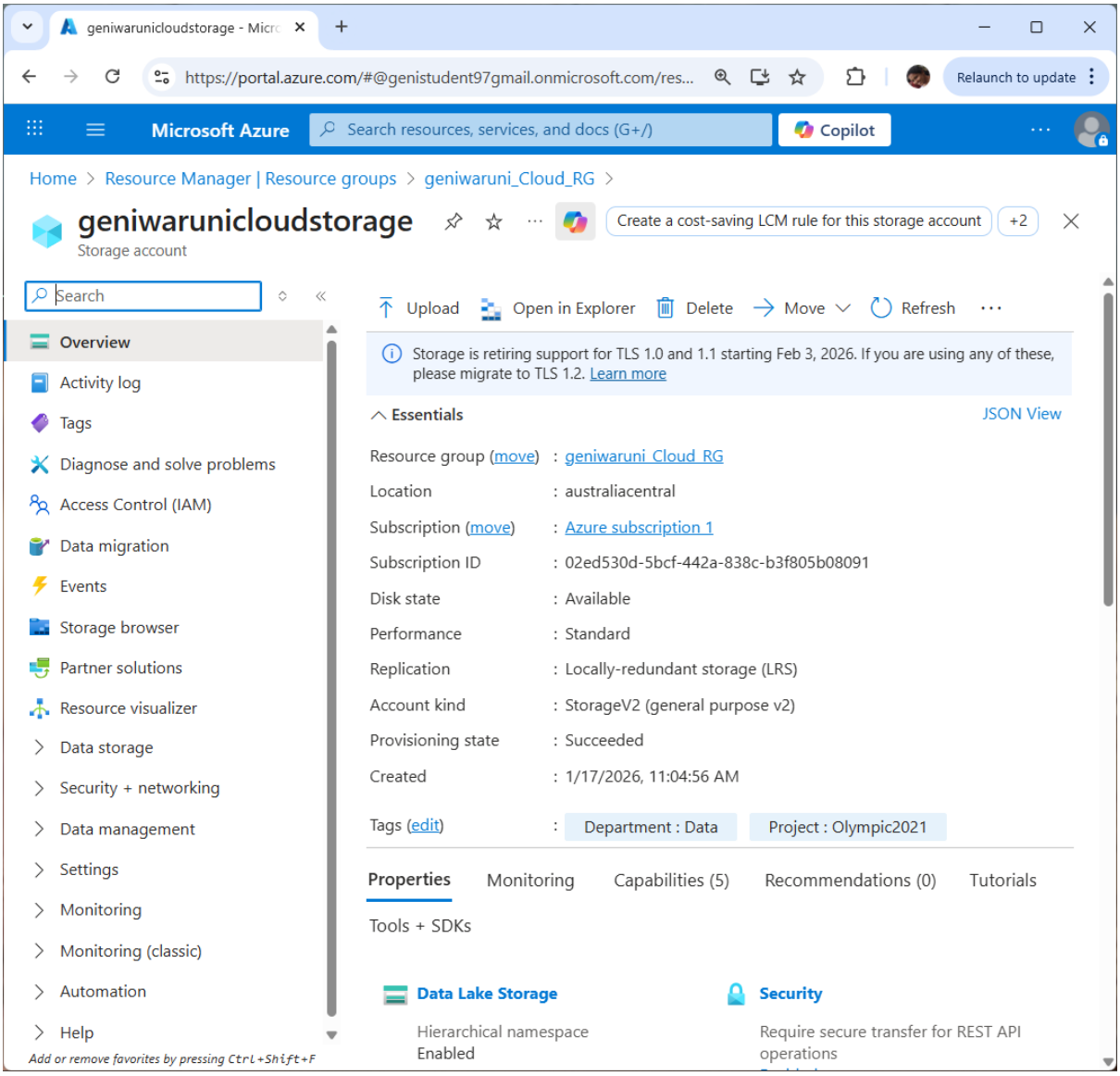

Custom deployment - Microsof
https://portal.azure.com/#create/Microsoft.Template
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Custom deployment
Deploy from a custom template
I want to deploy resources using Bicep
Automate Azure deployments with CLI
+1
By clicking "Create," I (a) agree to the applicable legal terms associated with the offering; (b) authorize Microsoft to charge or bill my current payment method for the fees associated the offering(s), including applicable taxes, with the same billing frequency as my Azure subscription, until I discontinue use of the offering(s); and (c) agree that, if the deployment involves 3rd party offerings, Microsoft may share my contact information and other details of such deployment with the publisher of that offering.
Microsoft assumes no responsibility for any actions performed by third-party templates and does not provide rights for third-party products or services. See the Azure Marketplace Terms for additional terms.
Deploying this template will create one or more Azure resources or Marketplace offerings. You acknowledge that you are responsible for reviewing the applicable pricing and legal terms associated with all resources and offerings deployed as part of this template. Prices and associated legal terms for any Marketplace offerings can be found in the Azure Marketplace; both are subject to change at any time prior to deployment.
Neither subscription credits nor monetary commitment funds may be used to purchase non-Microsoft offerings. These purchases are billed separately.
If any Microsoft products are included in a Marketplace offering (e.g. Windows Server or SQL Server), such products are licensed by Microsoft and not by any third party.
Basics
Subscription
Azure subscription 1
Resource group
geniwaruni_Cloud_RG
Region
Australia East
Storage Account Name
geniwarunicloudstorage
Location
australiacentral
Previous
Next
Create

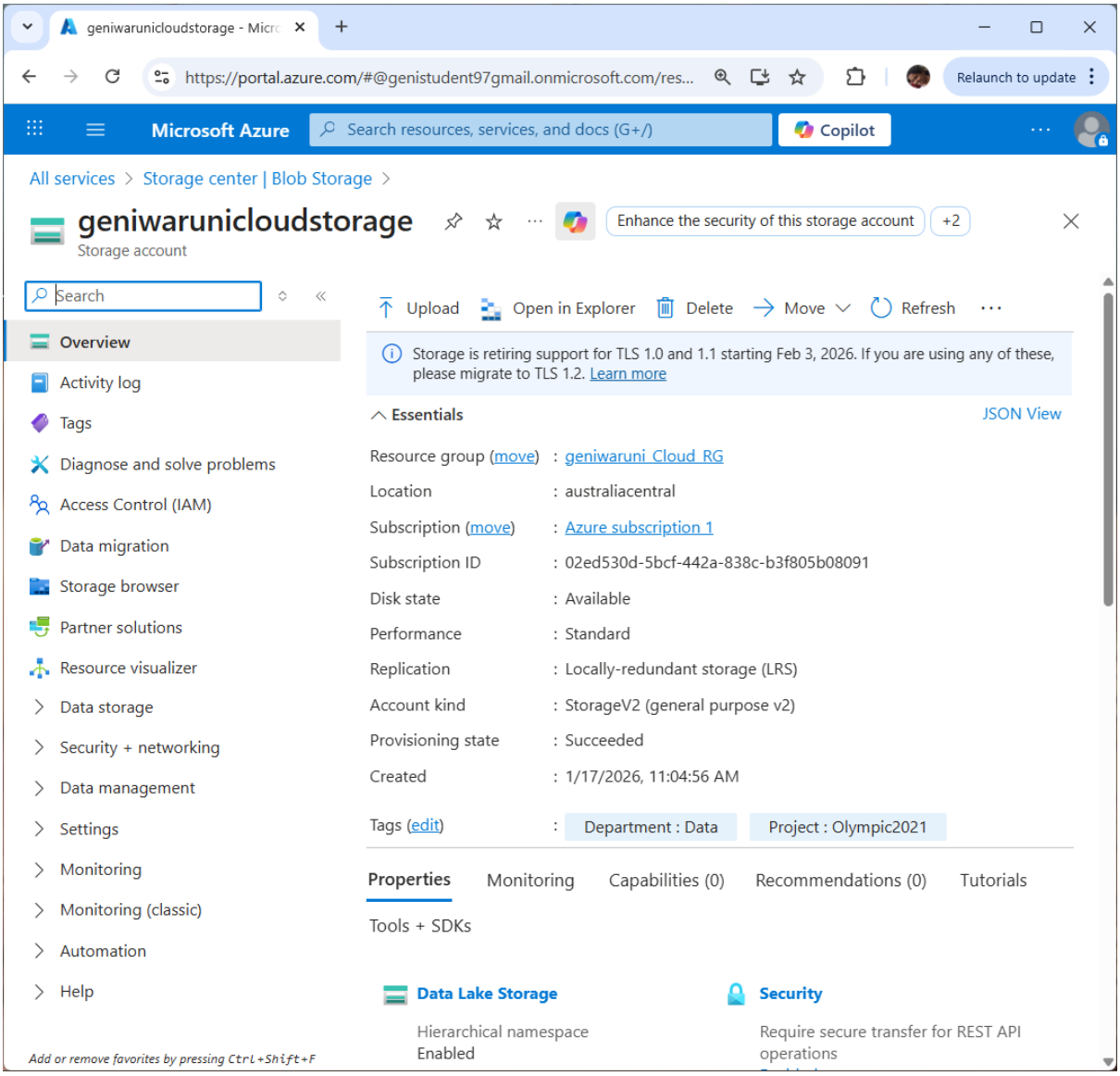

Microsoft.Template-202601171
https://portal.azure.com/#view/HubsExtension/DeploymentDetailsBla...
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Microsoft.Template-20260117110453 | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Deployment is in progress
Deployment name : Microsoft.Template-20260117110453
Subscription : Azure subscription 1
Resource group : geniwaruni_Cloud_RG
Start time : 1/17/2026, 11:04:49 AM
Correlation ID : 7f6eb1c0-cf4d-4201-bdcb-f94e9afff5a0
Deployment details
Resource
Type
Status
geniwaruniclou...
Storage account
Accepted
Give feedback
Tell us about your experience with deployment
Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >
Add or remove favorites by pressing Ctrl+Shift+F

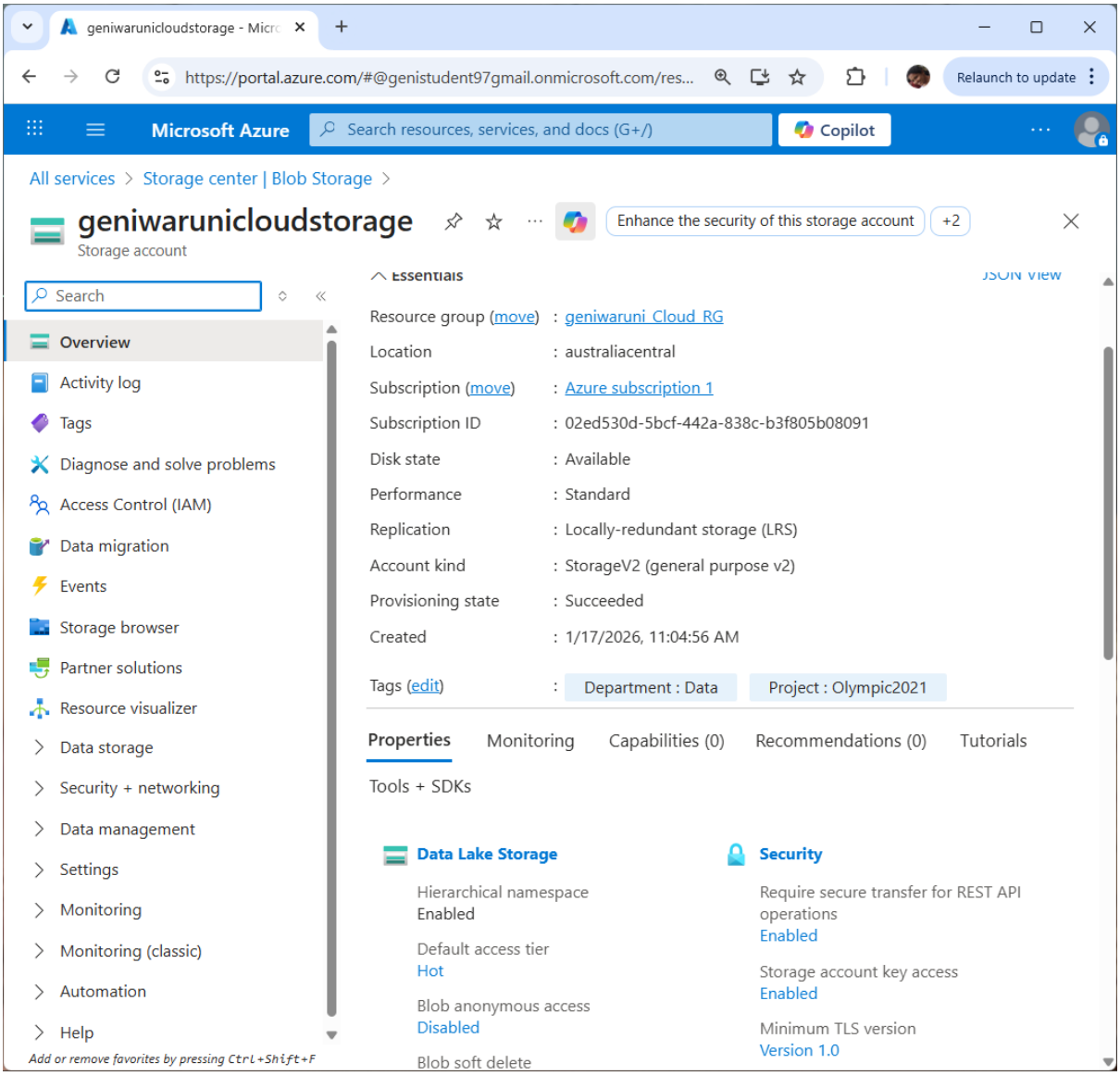

Microsoft.Template-202601171
https://portal.azure.com/#view/HubsExtension/DeploymentDetailsBla...
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Microsoft.Template-20260117110453 | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name : Microsoft.Template-20260117110453
Subscription : Azure subscription 1
Resource group : geniwaruni_Cloud_RG
Start time : 1/17/2026, 11:04:55 AM
Correlation ID : 7f6eb1c0-cf4d-4201-bdcb-f94e9afff5a0
Deployment details
Next steps
Go to resource
Give feedback
Tell us about your experience with deployment
Cost management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >
Add or remove favorites by pressing Ctrl+Shift+F


Microsoft.Template-202601171
https://portal.azure.com/#view/HubsExtension/DeploymentDetailsBla...
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Microsoft.Template-20260117110453 | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name : Microsoft.Template-20260117110453
Subscription : Azure subscription 1
Resource group : geniwaruni_Cloud_RG
Start time : 1/17/2026, 11:04:55 AM
Correlation ID : 7f6eb1c0-cf4d-4201-bdcb-f94e9afff5a0
Deployment details
Next steps
Go to resource
Give feedback
Tell us about your experience with deployment
Cost management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >
Add or remove favorites by pressing Ctrl+Shift+F


geniwaruni_Cloud_RG - Microsc
https://portal.azure.com/#@genistudent97gmail.onmicrosoft.com/res...
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home > Resource Manager | Resource groups >
geniwaruni_Cloud_RG
Resource group
Search
Overview
Activity log
Access control (IAM)
Tags
Resource visualizer
Events
Settings
Cost Management
Monitoring
Automation
Help
Create
Manage view
Delete resource group
Group by none
Essentials
JSON View
Resources
Recommendations
Filter for any field...
Type equals all
Location equals all
Add filter
Name
Type
Location
geniwarunicloudstorage
Storage account
Australia Central
Showing 1 - 1 of 1. Display count: auto
Give feedback
Add or remove favorites by pressing Ctrl+Shift+F


geniwaruniclоudstorage - Micr
https://portal.azure.com/#@genistudent97gmail.onmicrosoft.com/res...
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home > Resource Manager | Resource groups > geniwaruni_Cloud_RG >
geniwaruniclоudstorage
Storage account
Create a cost-saving LCM rule for this storage account
+2
Search
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
Data migration
Events
Storage browser
Partner solutions
Resource visualizer
Data storage
Security + networking
Data management
Settings
Monitoring
Monitoring (classic)
Automation
Help
Add or remove favorites by pressing Ctrl+Shift+F
Upload
Open in Explorer
Delete
Move
Refresh
Storage is retiring support for TLS 1.0 and 1.1 starting Feb 3, 2026. If you are using any of these, please migrate to TLS 1.2. Learn more
Essentials
JSON View
Resource group (move) : geniwaruni_Cloud_RG
Location : australiacentral
Subscription (move) : Azure subscription 1
Subscription ID : 02ed530d-5bcf-442a-838c-b3f805b08091
Disk state : Available
Performance : Standard
Replication : Locally-redundant storage (LRS)
Account kind : StorageV2 (general purpose v2)
Provisioning state : Succeeded
Created : 1/17/2026, 11:04:56 AM
Tags (edit) : Department : Data Project : Olympic2021
Properties
Monitoring
Capabilities (5)
Recommendations (0)
Tutorials
Tools + SDKs
Data Lake Storage
Hierarchical namespace
Enabled
Security
Require secure transfer for REST API operations

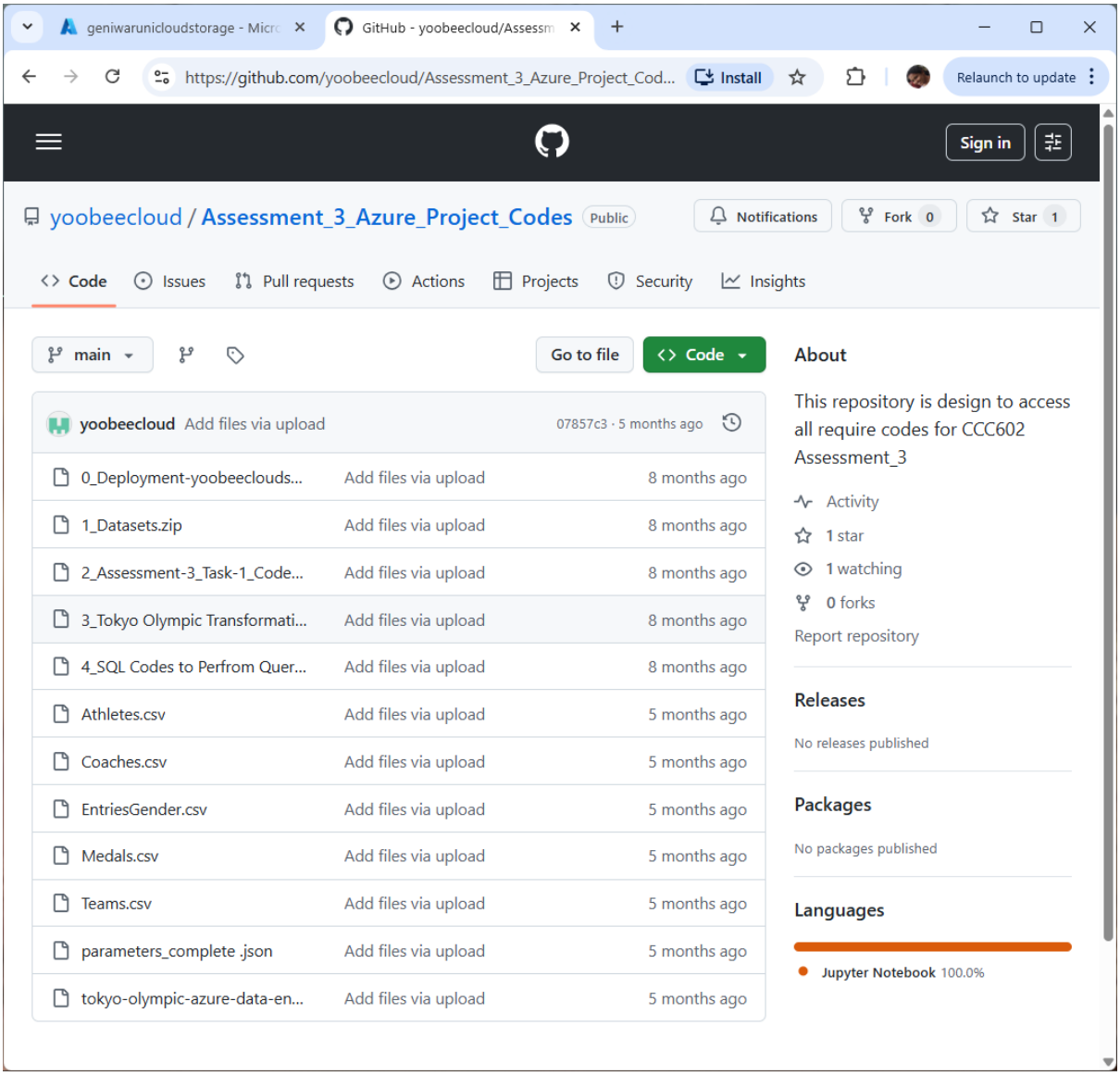

geniwaruniclоudstorage - Micr
https://portal.azure.com/#@genistudent97gmail.onmicrosoft.com/res...
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
All services > Storage center | Blob Storage >
geniwaruniclоudstorage
Storage account
Enhance the security of this storage account
+2
Search
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
Data migration
Storage browser
Partner solutions
Resource visualizer
Data storage
Security + networking
Data management
Settings
Monitoring
Monitoring (classic)
Automation
Help
Add or remove favorites by pressing Ctrl+Shift+F
Upload
Open in Explorer
Delete
Move
Refresh
Storage is retiring support for TLS 1.0 and 1.1 starting Feb 3, 2026. If you are using any of these, please migrate to TLS 1.2. Learn more
Essentials
JSON View
Resource group (move) : geniwaruni_Cloud_RG
Location : australiacentral
Subscription (move) : Azure subscription 1
Subscription ID : 02ed530d-5bcf-442a-838c-b3f805b08091
Disk state : Available
Performance : Standard
Replication : Locally-redundant storage (LRS)
Account kind : StorageV2 (general purpose v2)
Provisioning state : Succeeded
Created : 1/17/2026, 11:04:56 AM
Tags (edit) : Department : Data Project : Olympic2021
Properties
Monitoring
Capabilities (0)
Recommendations (0)
Tutorials
Tools + SDKs
Data Lake Storage
Hierarchical namespace
Enabled
Security
Require secure transfer for REST API operations



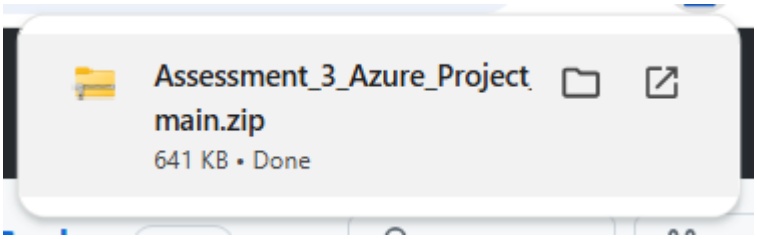
Using an ARM template for this task helped me better understand the importance of Infrastructure as Code, especially how it ensures consistent deployments and reduces manual configuration errors.

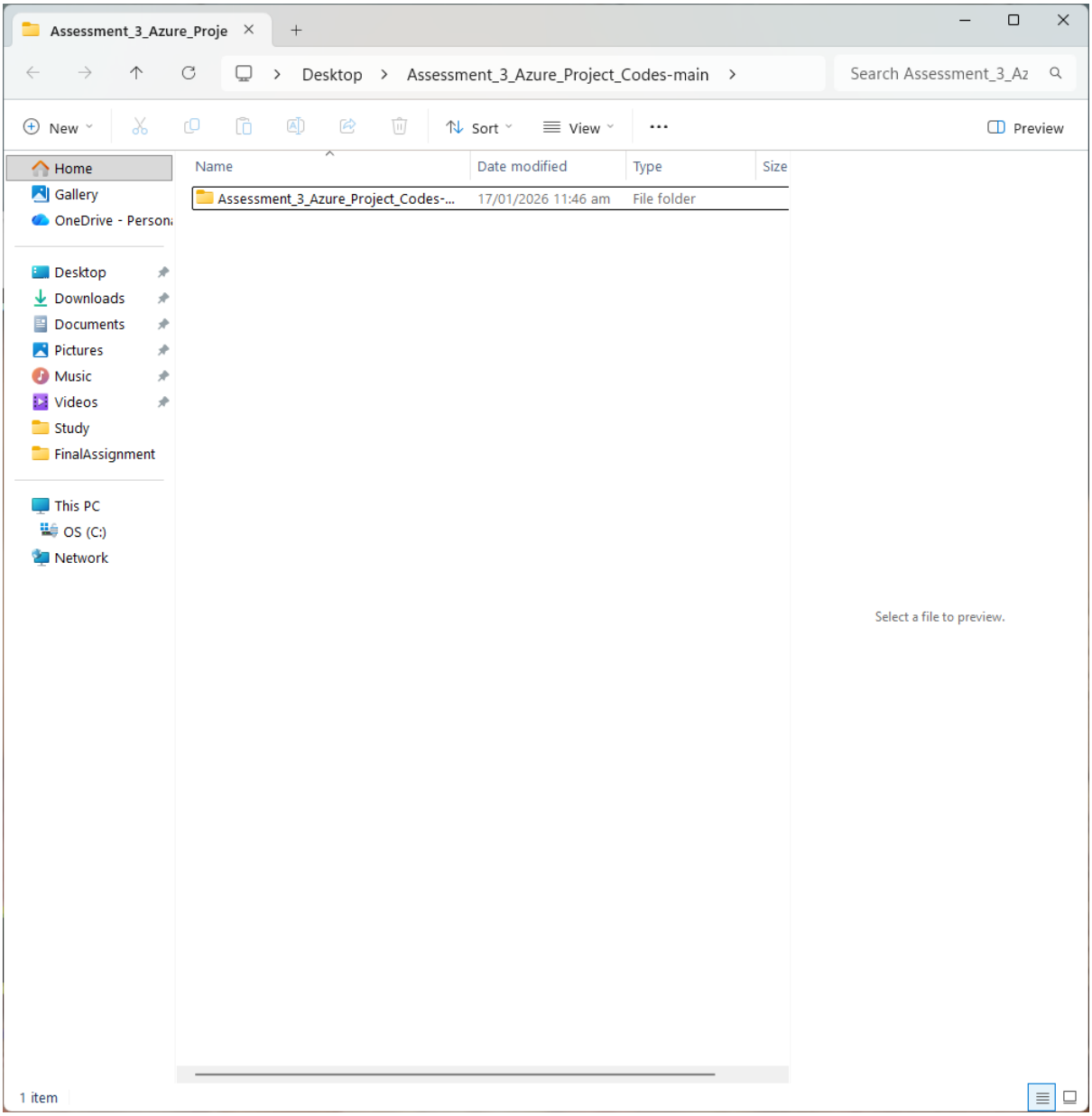
Tokyo Olympic Dataset Acquisition

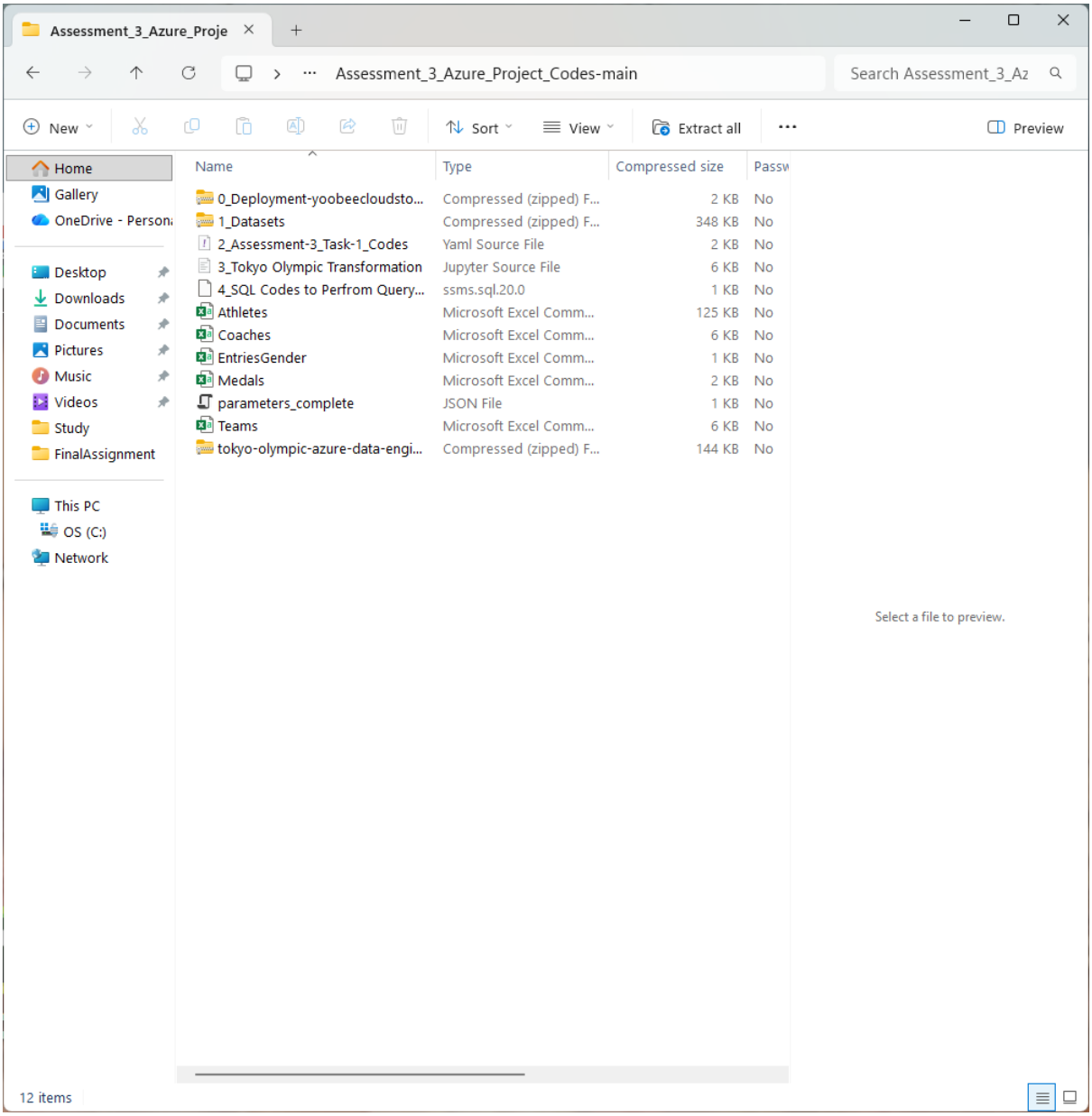
- The Tokyo Olympic dataset was downloaded from the GitHub repository provided by the tutor. The dataset contains multiple files related to athletes, medals, teams, coaches, and gender participation.
- The files were reviewed to understand their structure, file formats, and key fields before ingestion into Azure.



File Name	File Type	Description
athletes	CSV	Athlete details and country information
medals	CSV	Medal counts by country
teams	CSV	Team and discipline details
coaches	CSV	Coach information
entriesgender	CSV	Gender participation by discipline



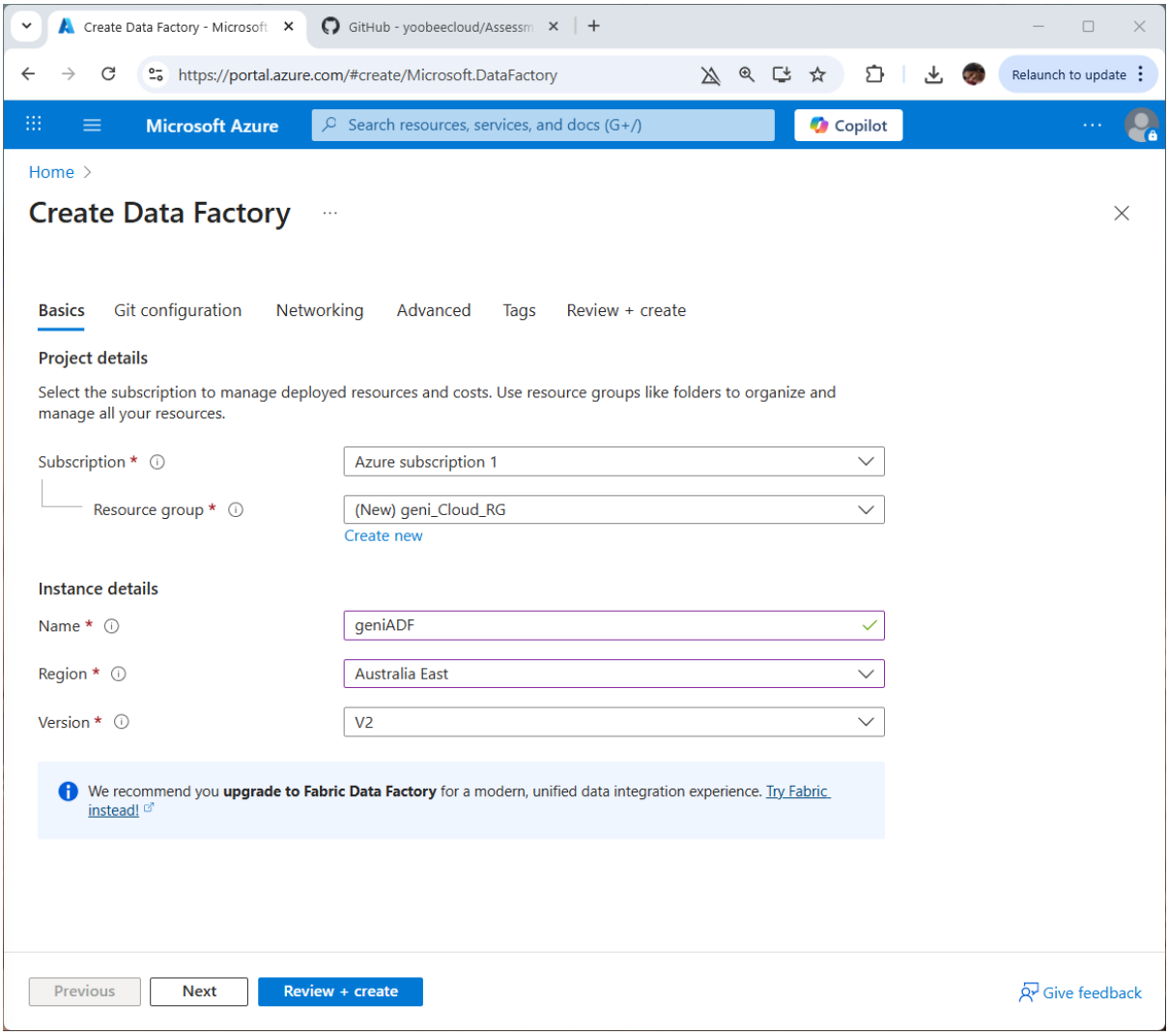

Assessment_3_Azure_Proje
Desktop
Assessment_3_Azure_Project_Codes-main
Search Assessment_3_Az
New
Sort
View
Preview
Home
Gallery
OneDrive - Personal
Desktop
Downloads
Documents
Pictures
Music
Videos
Study
FinalAssignment
This PC
OS (C:)
Network
Name
Date modified
Type
Size
Assessment_3_Azure_Project_Codes-...
17/01/2026 11:46 am
File folder
Select a file to preview.
1 item

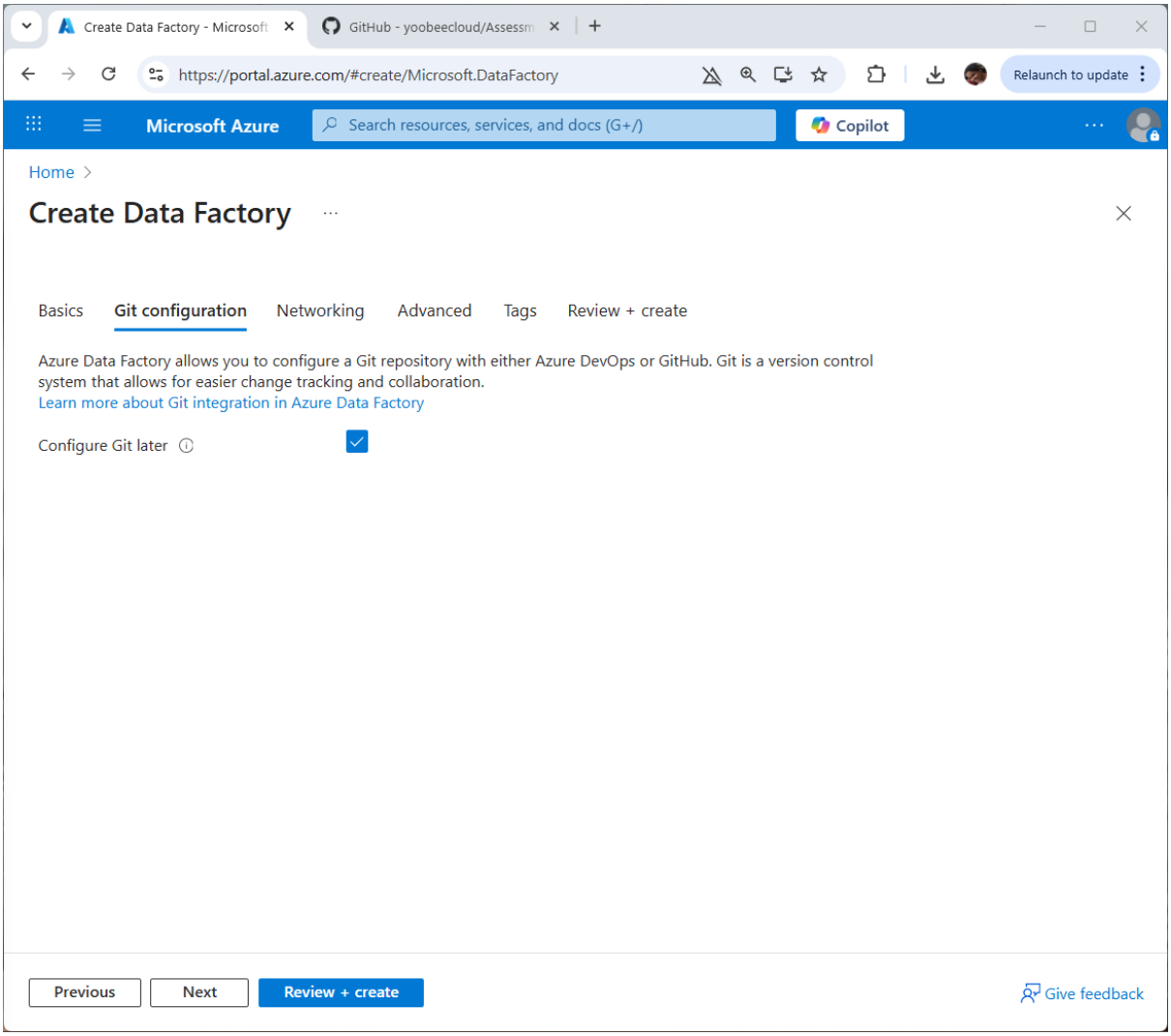

Assessment_3_Azure_Proje
Assessment_3_Azure_Project_Codes-main
Search Assessment_3_Az
New
Sort
View
Extract all
Preview
Home
Gallery
OneDrive - Personal
Desktop
Downloads
Documents
Pictures
Music
Videos
Study
FinalAssignment
This PC
OS (C:)
Network
Name
Type
Compressed size
Passw
0_Deployment-yoobeecloudsto... Compressed (zipped) F... 2 KB No
1_Datasets Compressed (zipped) F... 348 KB No
2_Assessment-3_Task-1_Codes Yaml Source File 2 KB No
3_Tokyo Olympic Transformation Jupyter Source File 6 KB No
4_SQL Codes to Perfrom Query... ssms.sql.20.0 1 KB No
Athletes Microsoft Excel Comm... 125 KB No
Coaches Microsoft Excel Comm... 6 KB No
EntriesGender Microsoft Excel Comm... 1 KB No
Medals Microsoft Excel Comm... 2 KB No
parameters_complete JSON File 1 KB No
Teams Microsoft Excel Comm... 6 KB No
tokyo-olympic-azure-data-engi... Compressed (zipped) F... 144 KB No
Select a file to preview.
12 items

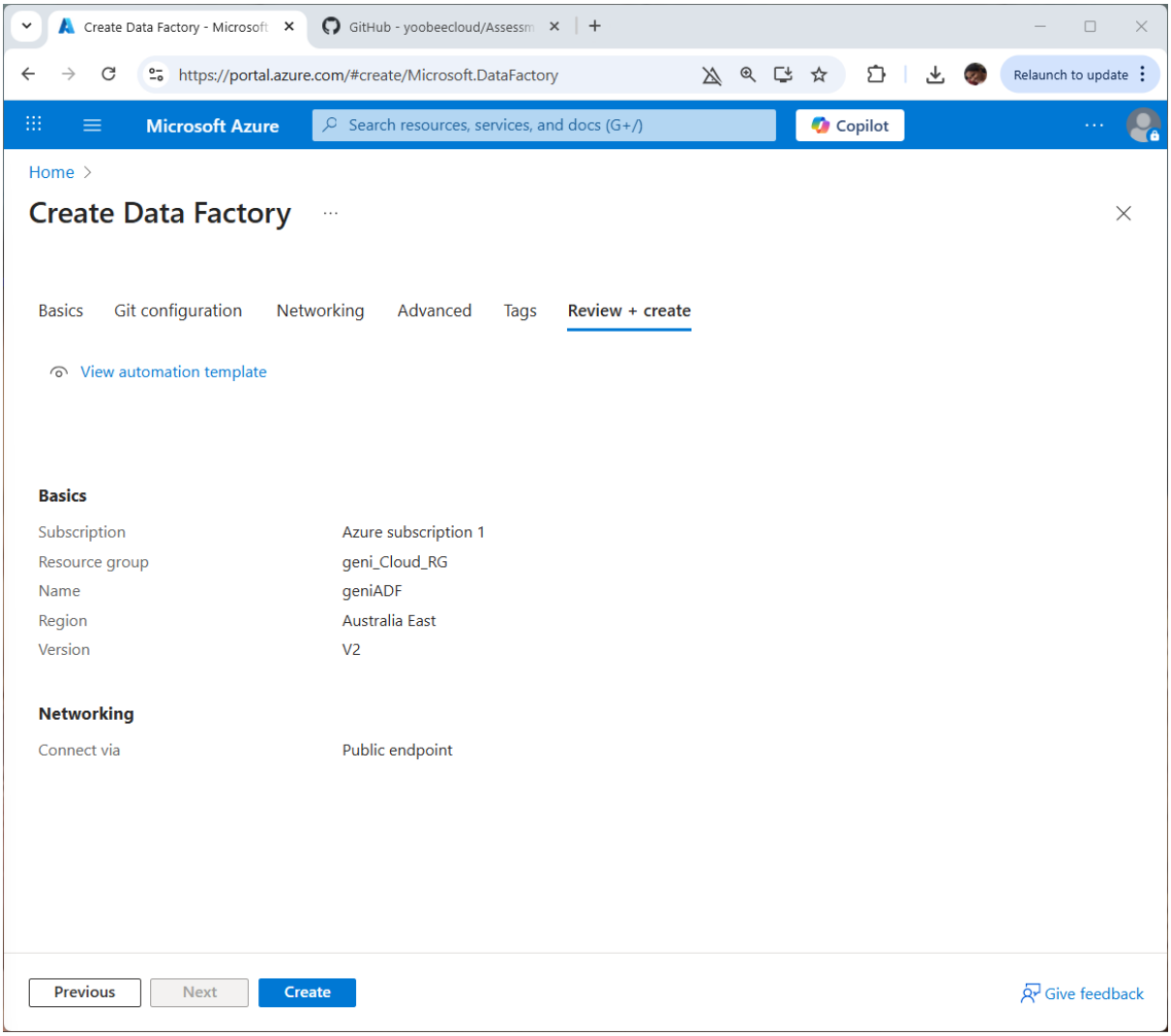
Task 3: Data Ingestion Using Azure Data Factory

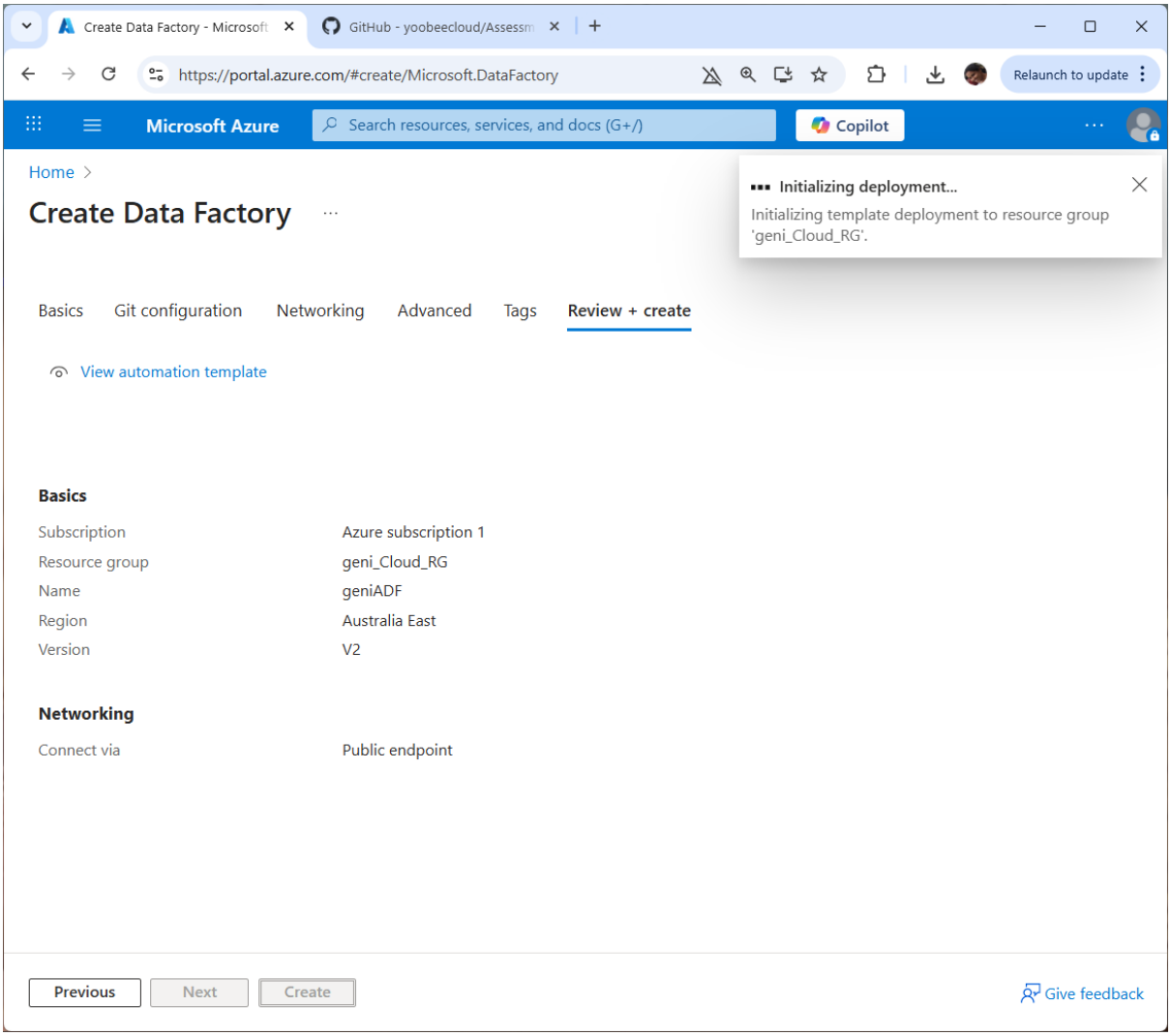
Data Ingestion Pipeline Design

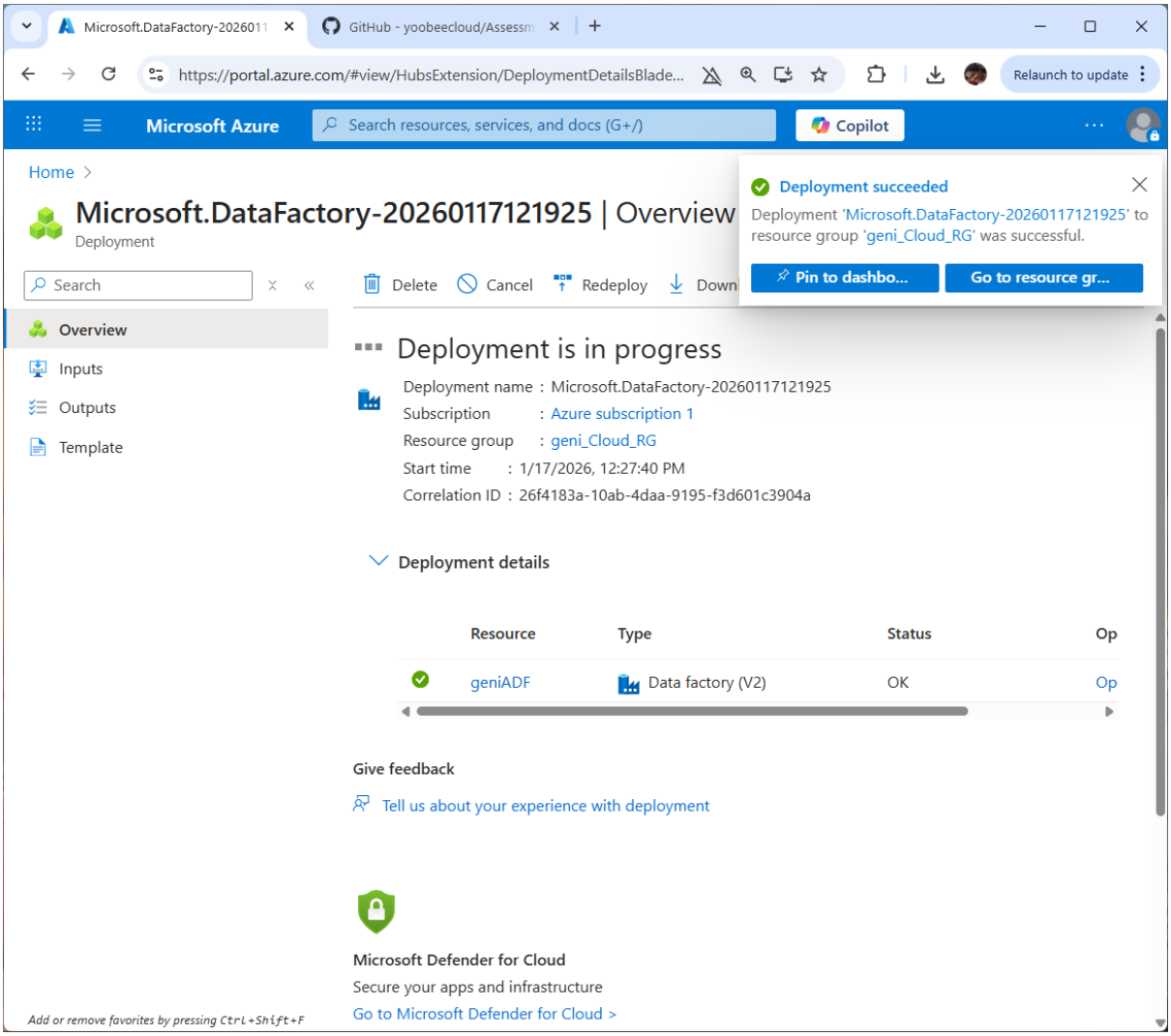
- Azure Data Factory was used to automate the ingestion of data from Azure Data Lake Storage into the /raw folder. The Copy Data activity was configured using ADLS Gen2 as both the source and sink to preserve the original data format.
- The pipeline ensured that data ingestion was repeatable and scalable.

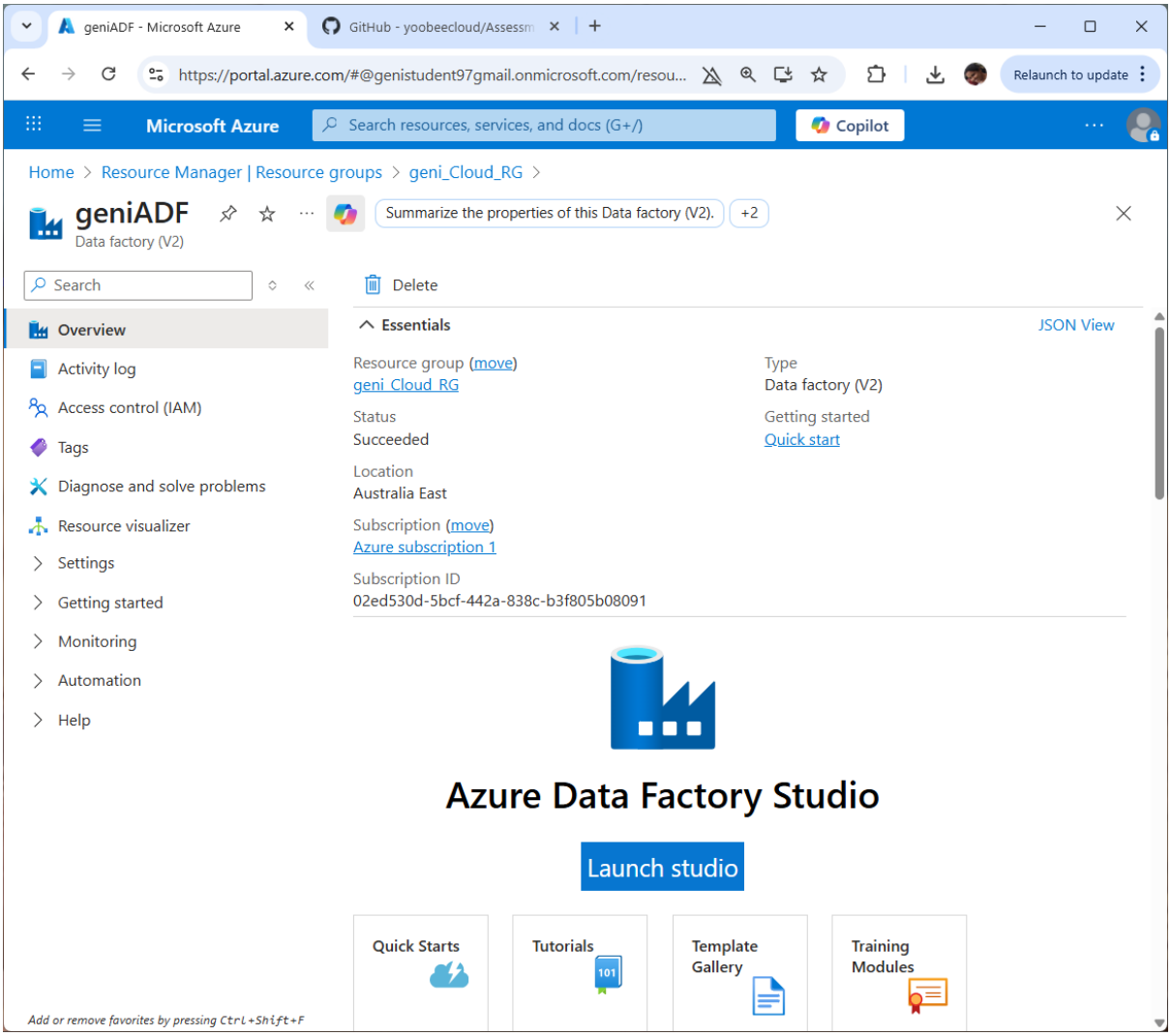


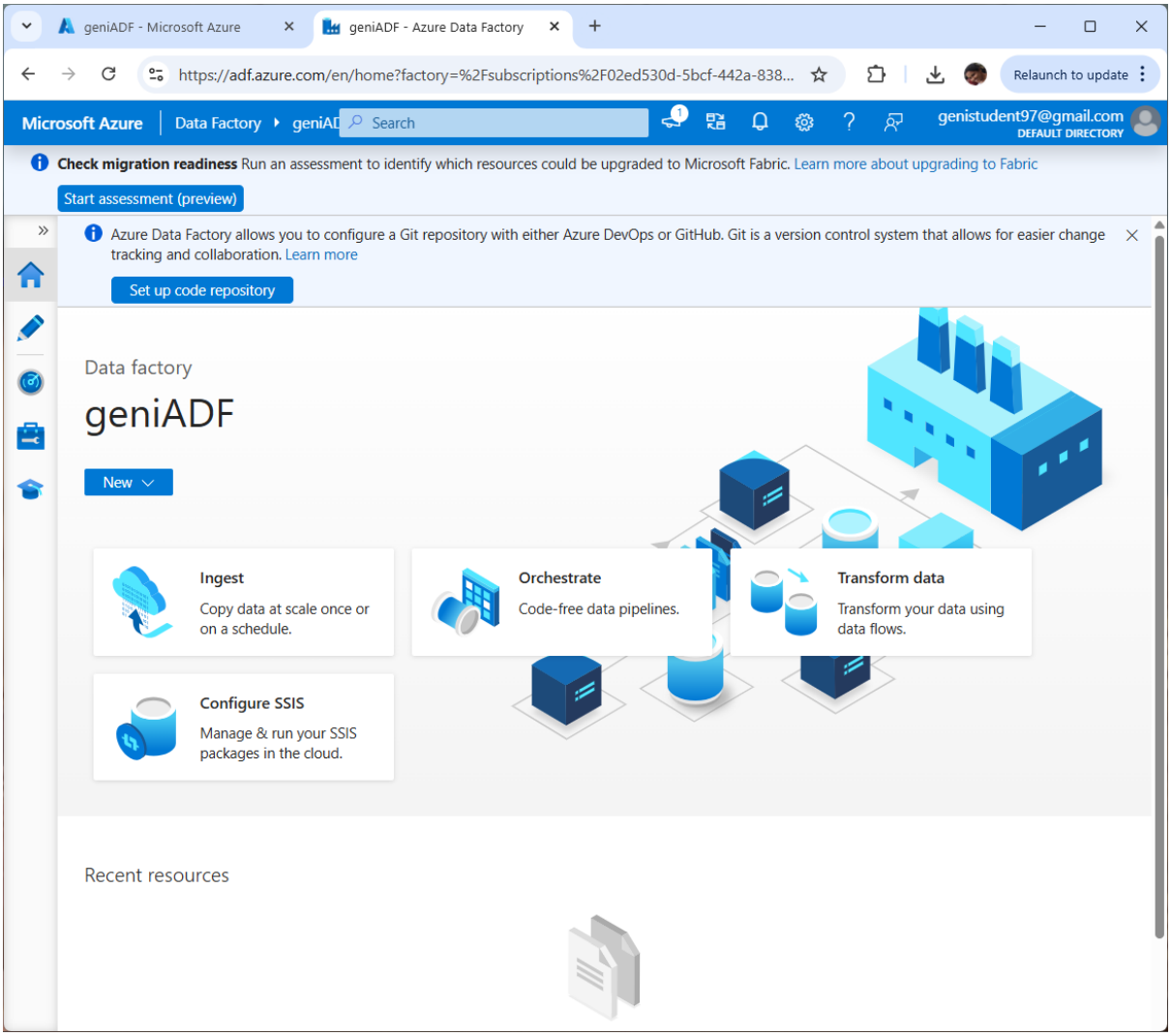

Create Data Factory - Microsoft
GitHub - yoobeecloud/Assessm
https://portal.azure.com/#create/Microsoft.DataFactory
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Create Data Factory
Basics
Git configuration
Networking
Advanced
Tags
Review + create
Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control system that allows for easier change tracking and collaboration.
Learn more about Git integration in Azure Data Factory
Configure Git later
Previous
Next
Review + create
Give feedback


Create Data Factory - Microsoft
GitHub - yoobeecloud/Assessm
https://portal.azure.com/#create/Microsoft.DataFactory
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Create Data Factory
Basics
Git configuration
Networking
Advanced
Tags
Review + create
View automation template
Basics
Subscription
Azure subscription 1
Resource group
geni_Cloud_RG
Name
geniADF
Region
Australia East
Version
V2
Networking
Connect via
Public endpoint
Previous
Next
Create
Give feedback


Create Data Factory - Microsoft
GitHub - yoobeecloud/Assessm
https://portal.azure.com/#create/Microsoft.DataFactory
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Create Data Factory
Initializing deployment...
Initializing template deployment to resource group 'geni_Cloud_RG'.
Basics
Git configuration
Networking
Advanced
Tags
Review + create
View automation template
Basics
Subscription
Azure subscription 1
Resource group
geni_Cloud_RG
Name
geniADF
Region
Australia East
Version
V2
Networking
Connect via
Public endpoint
Previous
Next
Create
Give feedback

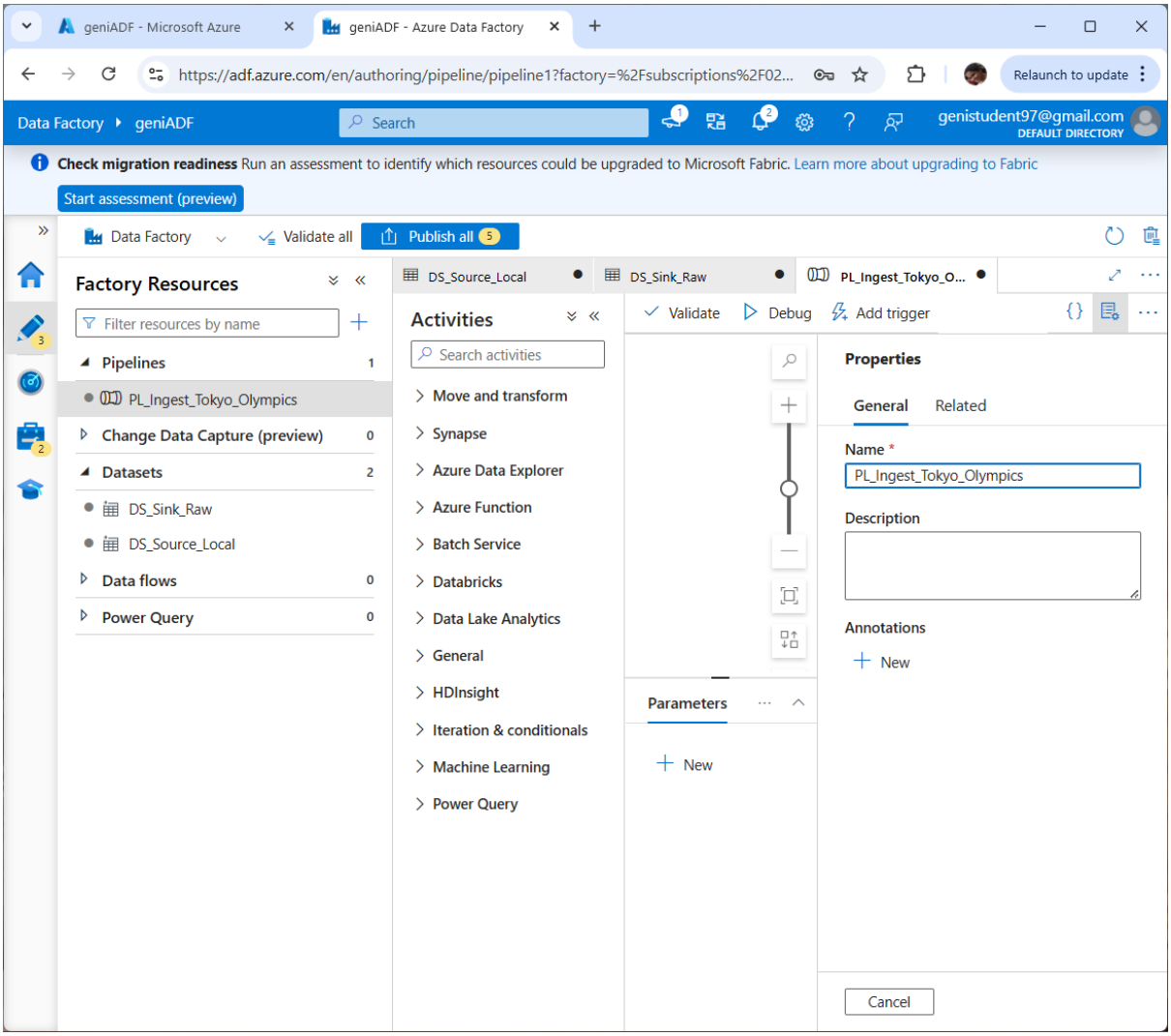

Microsoft.DataFactory-2026011
GitHub - yoobeecloud/Assessm
https://portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade...
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Microsoft.DataFactory-20260117121925 | Overview
Deployment
Deployment succeeded
Deployment 'Microsoft.DataFactory-20260117121925' to resource group 'geni_Cloud_RG' was successful.
Pin to dashbo...
Go to resource gr...
Search
Delete
Cancel
Redeploy
Down
Overview
Inputs
Outputs
Template
Deployment is in progress
Deployment name : Microsoft.DataFactory-20260117121925
Subscription : Azure subscription 1
Resource group : geni_Cloud_RG
Start time : 1/17/2026, 12:27:40 PM
Correlation ID : 26f4183a-10ab-4daa-9195-f3d601c3904a
Deployment details
Resource
Type
Status
Op
geniADF
Data factory (V2)
OK
Op
Give feedback
Tell us about your experience with deployment
Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >
Add or remove favorites by pressing Ctrl+Shift+F

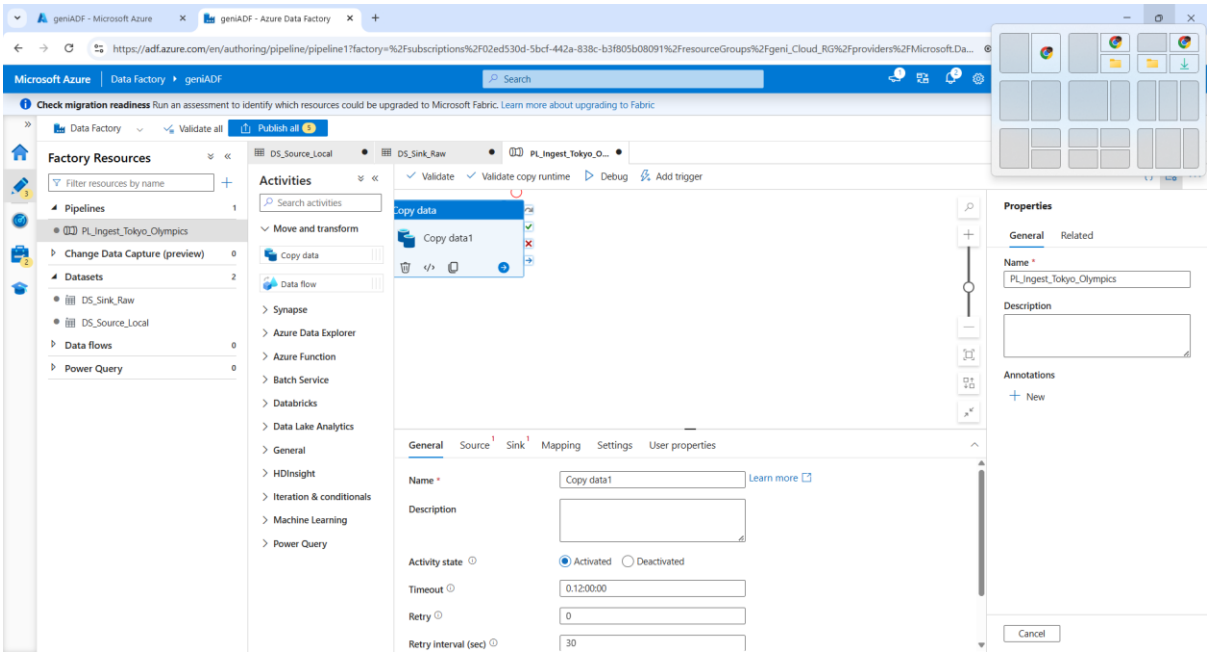

geniADF - Microsoft Azure
GitHub - yoobeecloud/Assessm
https://portal.azure.com/#@genistudent97gmail.onmicrosoft.com/resou...
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home > Resource Manager | Resource groups > geni_Cloud_RG >
geniADF
Data factory (V2)
Summarize the properties of this Data factory (V2).
+2
Search
Delete
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Resource visualizer
Settings
Getting started
Monitoring
Automation
Help
Essentials
JSON View
Resource group (move)
geni_Cloud_RG
Type
Data factory (V2)
Status
Succeeded
Getting started
Quick start
Location
Australia East
Subscription (move)
Azure subscription 1
Subscription ID
02ed530d-5bcf-442a-838c-b3f805b08091
Azure Data Factory Studio
Launch studio
Quick Starts
Tutorials
Template Gallery
Training Modules
Add or remove favorites by pressing Ctrl+Shift+F

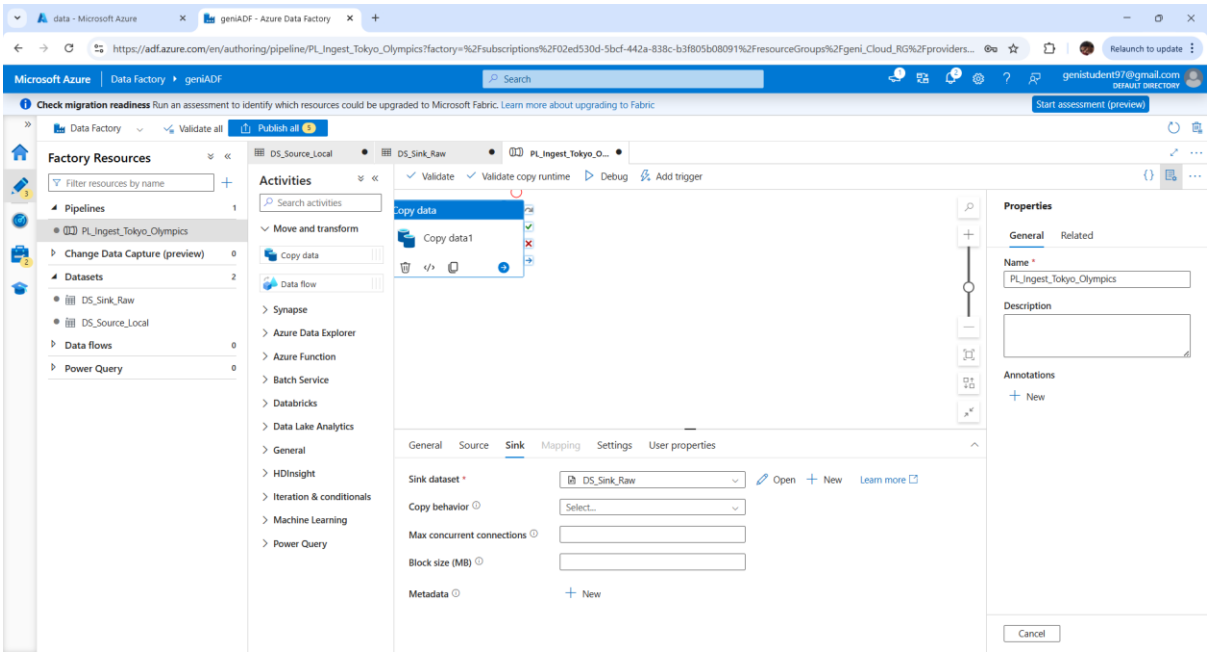


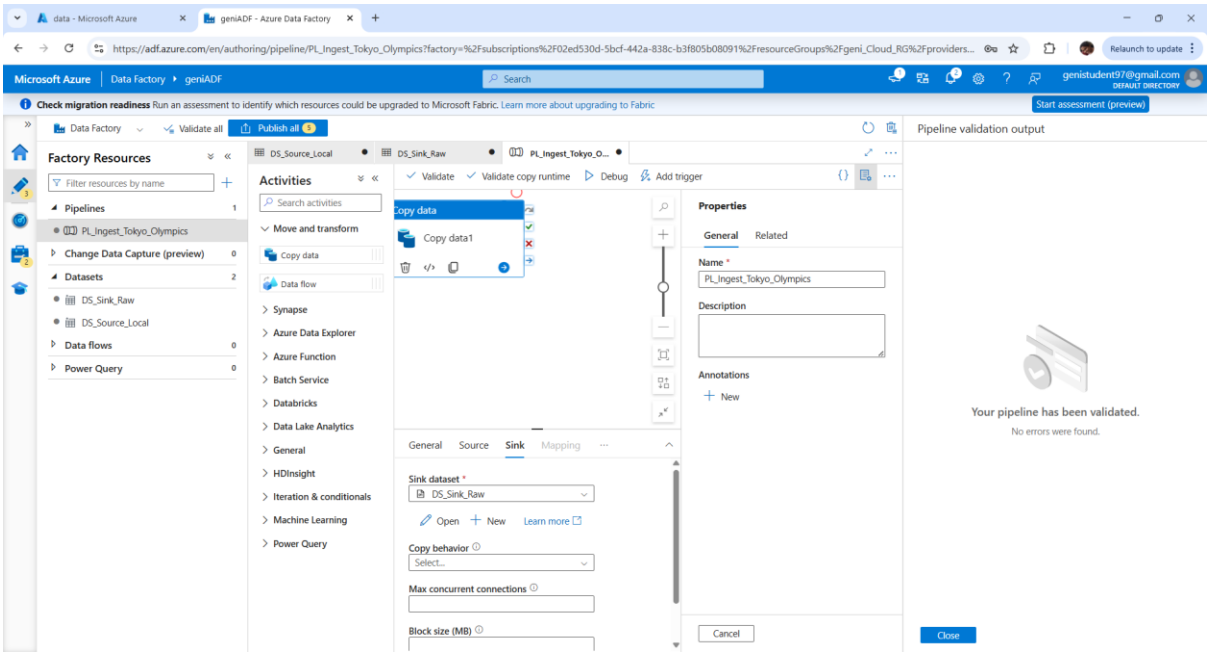
Source and Sink Configuration

- The source dataset pointed to the /source folder in the Data Lake, while the sink dataset wrote the ingested data to the /raw folder. Linked services were configured using managed identity authentication.


geniADF - Microsoft Azure
geniADF - Azure Data Factory
https://adf.azure.com/en/authoring/pipeline/pipeline1?factory=%2Fsubscriptions%2F02...
Relaunch to update
Data Factory ▸ geniADF
Search
genistudent97@gmail.com
DEFAULT DIRECTORY
Check migration readiness Run an assessment to identify which resources could be upgraded to Microsoft Fabric. Learn more about upgrading to Fabric
Start assessment (preview)
Data Factory
Validate all
Publish all 5
Factory Resources
Filter resources by name
Pipelines 1
PL_Ingest_Tokyo_Olympics
Change Data Capture (preview) 0
Datasets 2
DS_Sink_Raw
DS_Source_Local
Data flows 0
Power Query 0
DS_Source_Local
DS_Sink_Raw
PL_Ingest_Tokyo_O...
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Debug
Add trigger
Parameters
New
Properties
General
Related
Name *
PL_Ingest_Tokyo_Olympics
Description
Annotations
New
Cancel

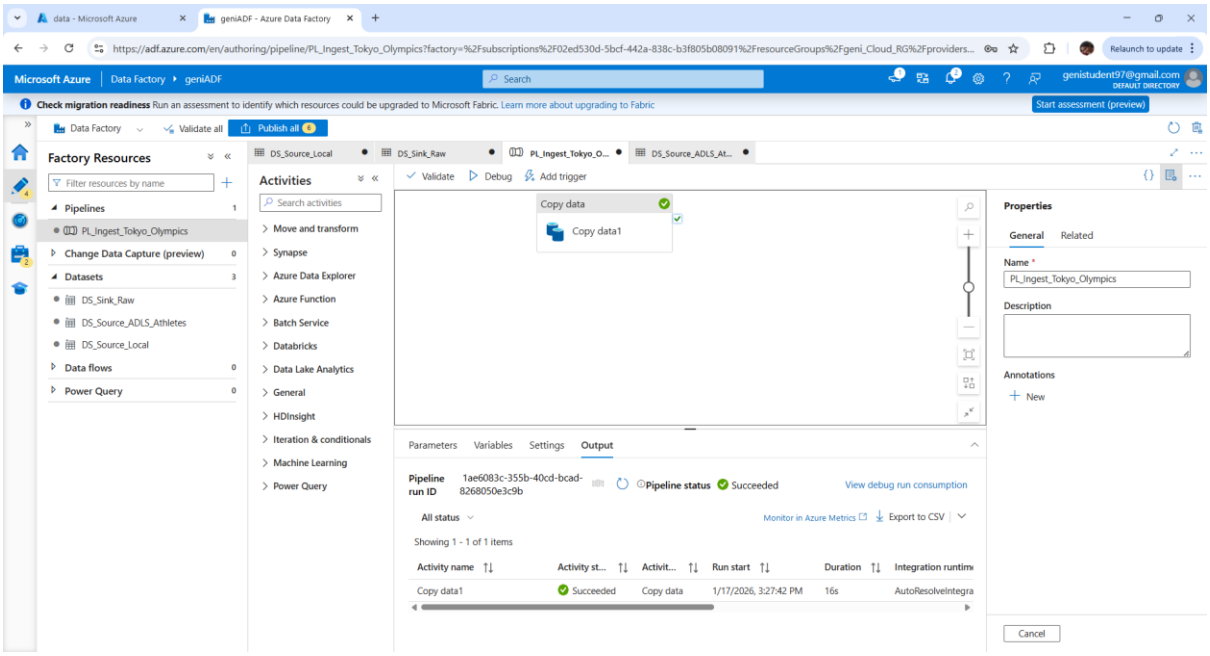

Microsoft Azure
Data Factory ▸ geniADF
Factory Resources
Activities
Validate
Validate copy runtime
Debug
Add trigger
Copy data
Copy data1
Properties
General
Related
Name *
PL_Ingest_Tokyo_Olympics
Description
Annotations
New
General
Source
Sink
Mapping
Settings
User properties
Name *
Copy data1
Learn more
Description
Activity state
Activated
Deactivated
Timeout
0.12:00:00
Retry
0
Retry interval (sec)
30
Cancel

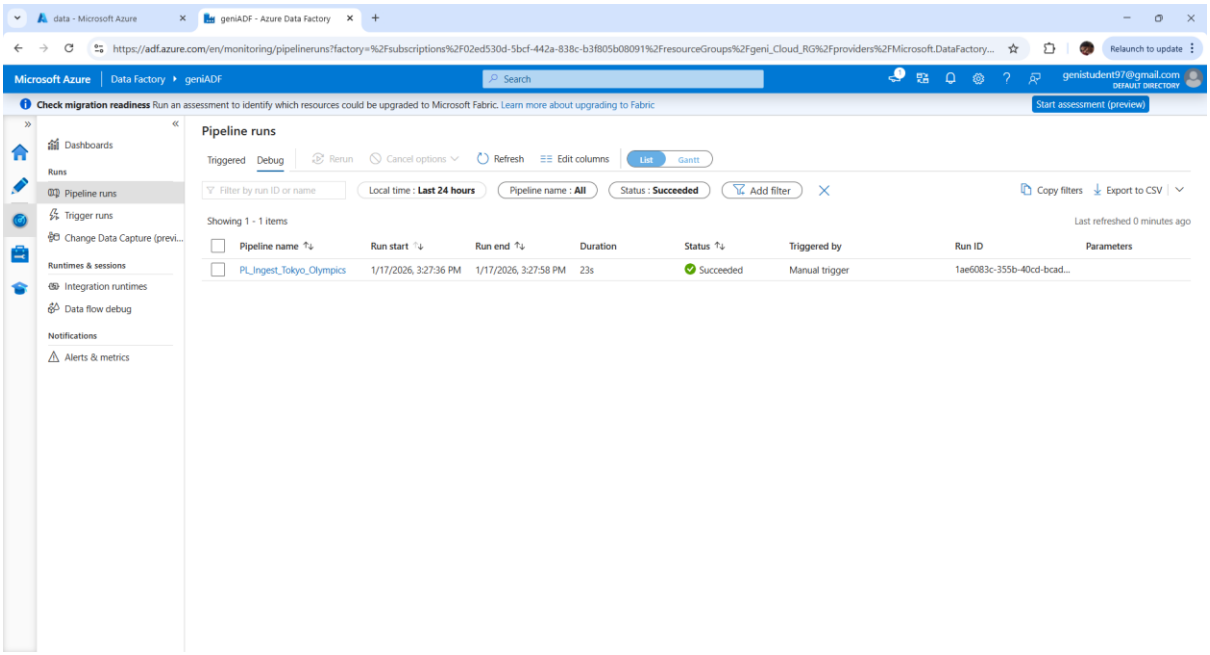


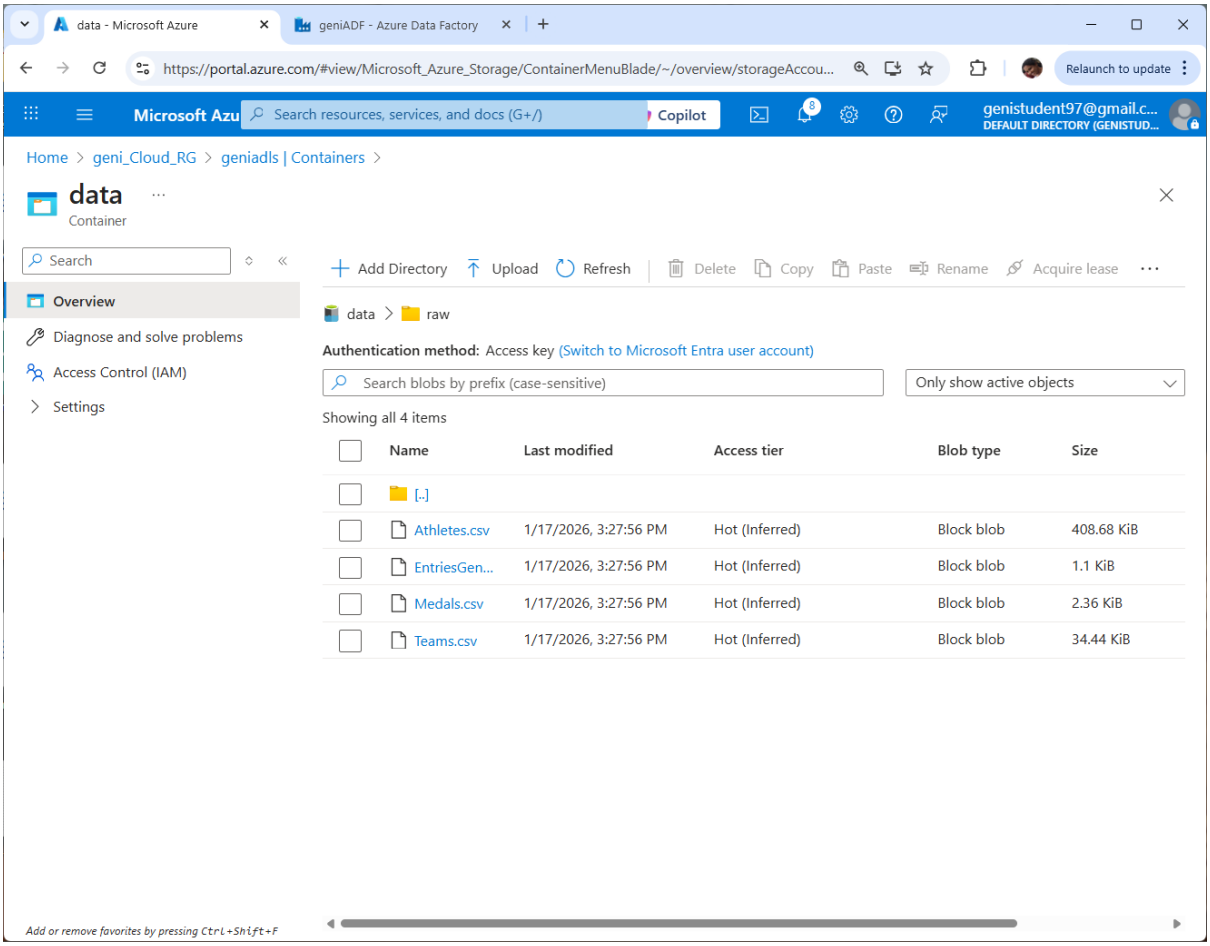


Pipeline Validation and Verification

- The pipeline was validated and executed successfully. After execution, the /raw folder was checked to confirm that all files were ingested correctly.


Microsoft Azure
Data Factory
geniADF
Factory Resources
Pipelines
PL_Ingest_Tokyo_Olympics
Datasets
DS_Sink_Raw
DS_Source_ADLS_Athletes
DS_Source_Local
Activities
Copy data
Copy data1
Properties
General
Related
Name *
PL_Ingest_Tokyo_Olympics
Description
Annotations
New
Parameters
Variables
Settings
Output
Pipeline run ID
Pipeline status
Succeeded
Activity name
Activity st...
Activit...
Run start
Duration
Integration runtime
Copy data1
Succeeded
Copy data
1/17/2026, 3:27:42 PM
16s
AutoResolveIntegra
Cancel


Microsoft Azure
Data Factory
geniADF
Pipeline runs
Triggered
Debug
Refresh
Edit columns
List
Gantt
Pipeline name
Run start
Run end
Duration
Status
Triggered by
Run ID
Parameters
PL_Ingest_Tokyo_Olympics
1/17/2026, 3:27:36 PM
1/17/2026, 3:27:58 PM
23s
Succeeded
Manual trigger



Pipeline Summary Report

Purpose of the Pipeline

- The main purpose of the Azure Data Factory pipeline was to automate the ingestion of the Tokyo Olympic dataset into Azure Data Lake Storage Gen2 in a structured and repeatable way. Instead of manually uploading files every time, the pipeline ensures that raw data is consistently moved into the /raw folder, where it can later be used for transformation and analysis. This approach follows real-world data engineering practices, where raw data must be preserved in its original format before any processing is applied.
- The pipeline also supports scalability. If more Olympic datasets or updated files are added in the future, the same pipeline can be reused without redesigning the entire process. This reduces manual effort and helps maintain data integrity across the project.

Data Flow (Source to Sink)

- The data flow begins with the Tokyo Olympic dataset stored in Azure Data Lake Storage Gen2. Initially, the files were uploaded manually into a /source folder in the data container. Azure Data Factory was then configured to use this folder as the source location.
- Within the pipeline, a **Copy Data** activity was used. The source dataset pointed to the ADLS Gen2 /source directory, and the sink dataset wrote the data into the /raw directory within the same storage account. Using ADLS Gen2 for both source and sink ensured that the data

transfer was secure, efficient, and compatible with downstream services such as Azure Databricks and Azure Synapse Analytics.

- The pipeline was validated and tested using the Debug option. After successful execution, the /raw folder was checked in the Azure Portal to confirm that all files were transferred correctly. This verification step was important to ensure that the data ingestion process worked as expected before moving to the transformation stage.

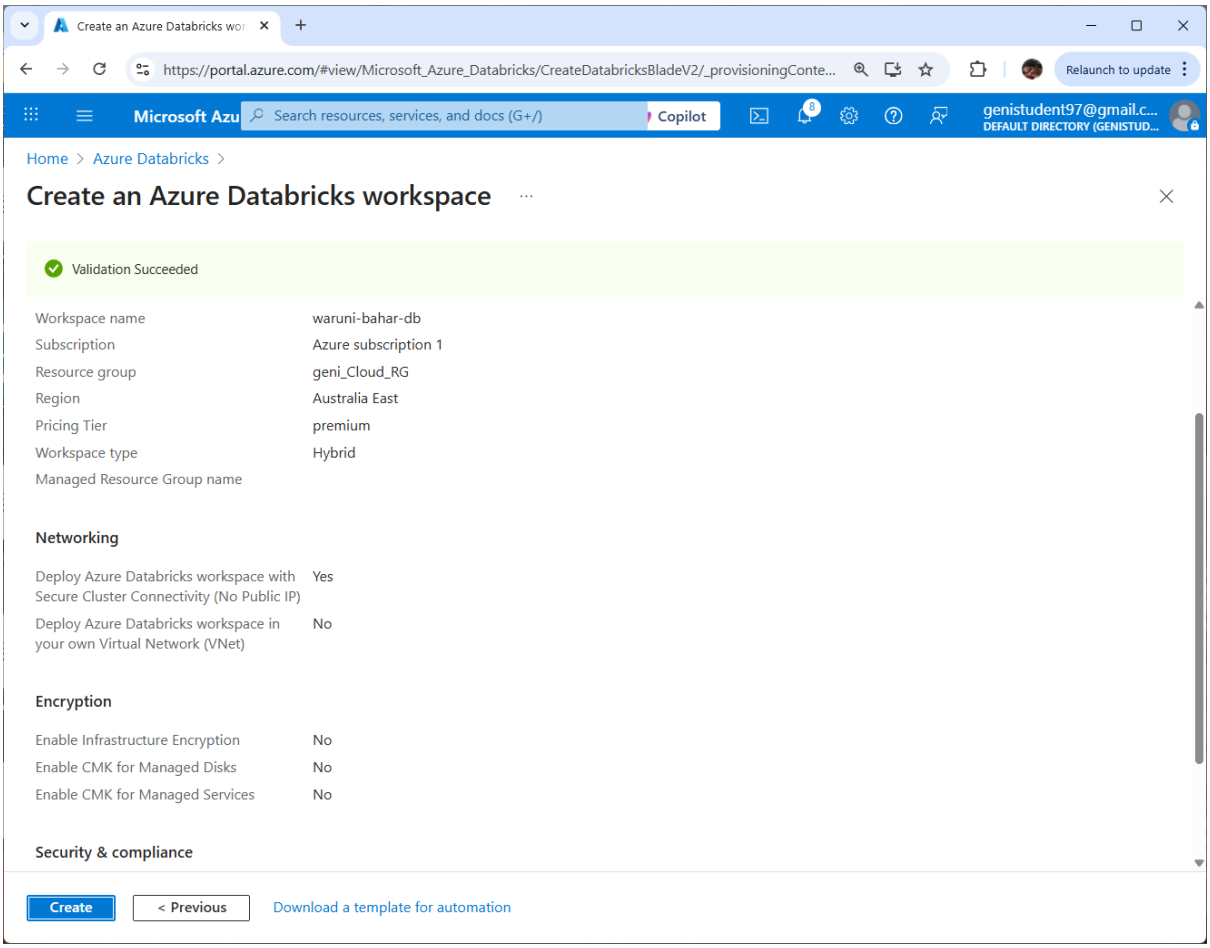
Challenges and Observations During Implementation

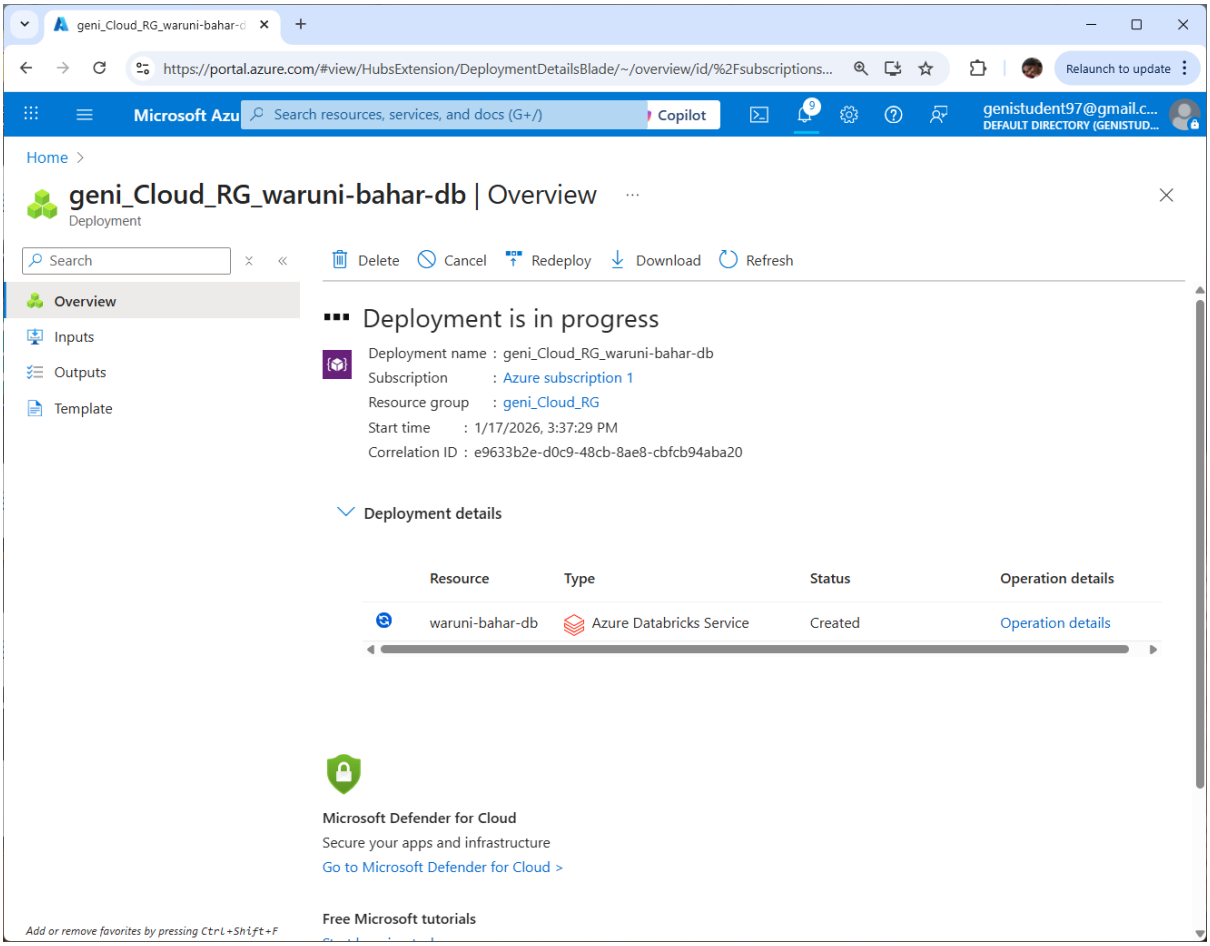
- Several challenges were encountered during the implementation of the pipeline, which helped improve understanding of real-world cloud data engineering issues.
- One of the main challenges was related to **permissions and authentication**. Initially, the pipeline failed with "Forbidden" and "AuthorizationPermissionMismatch" errors. This occurred because Azure Data Factory did not have sufficient permissions to access the Data Lake Storage account. The issue was resolved by enabling the system-assigned managed identity for Data Factory and assigning the **Storage Blob Data Contributor** role at the storage account level.
- Another challenge was confusion around using a **local file system dataset**. At first, a local (File System) dataset was configured as the source, which resulted in connection errors because Azure Data Factory cannot directly access files from a personal computer without a self-hosted integration runtime. This was resolved by uploading the files to ADLS Gen2 and using an ADLS-based dataset instead.
- Additionally, sink configuration errors occurred due to missing parameters and file paths. In some cases, leaving the file name empty in the sink dataset caused null reference errors during pipeline execution. Carefully reviewing dataset settings and explicitly defining directory paths helped resolve this issue.
- Overall, these challenges highlighted the importance of correct identity configuration, understanding how cloud services access data, and carefully validating pipeline settings. Resolving these issues provided valuable hands-on experience that closely reflects real industry scenarios.

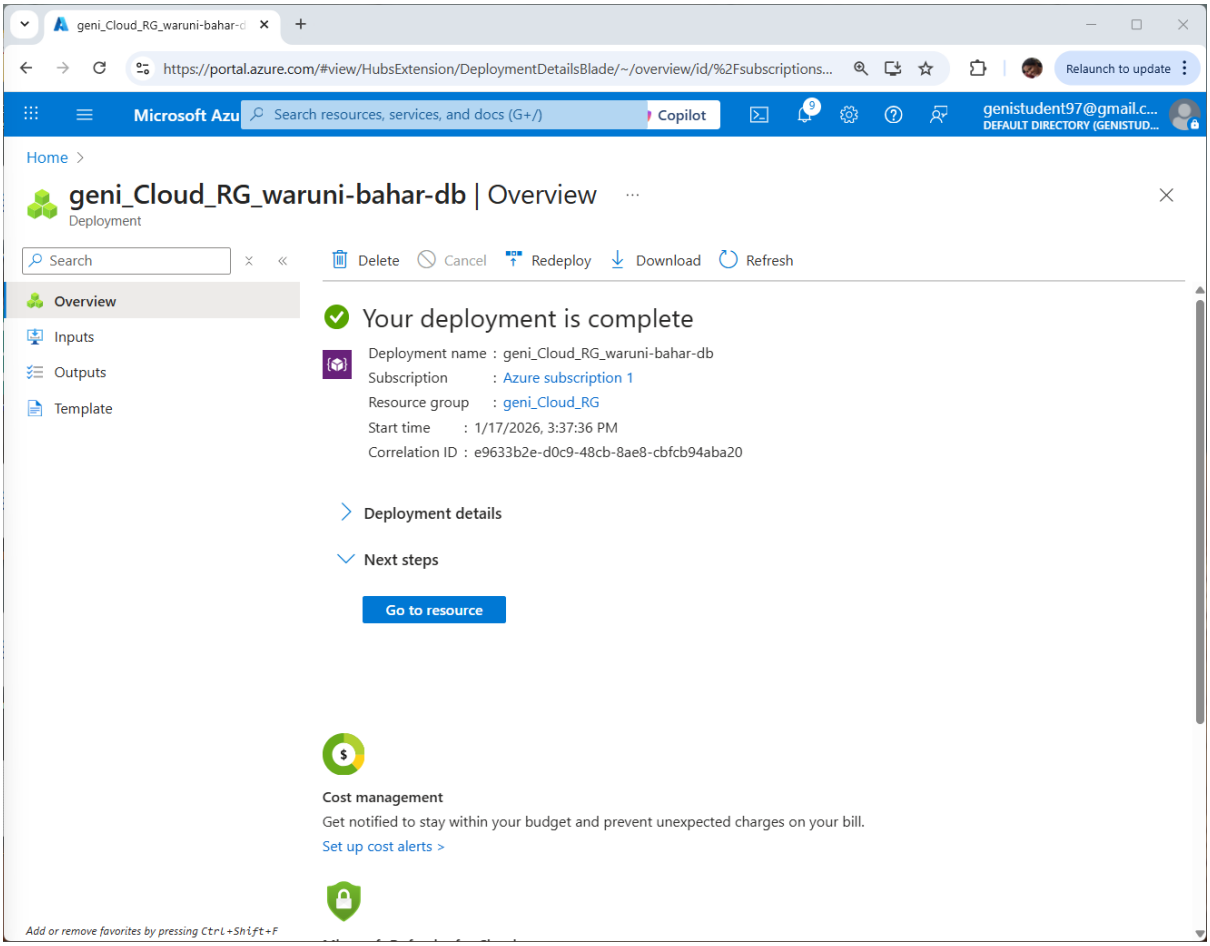
Task 4: Create an Azure Databricks Workspace and perform data transformation using code to automate the process.

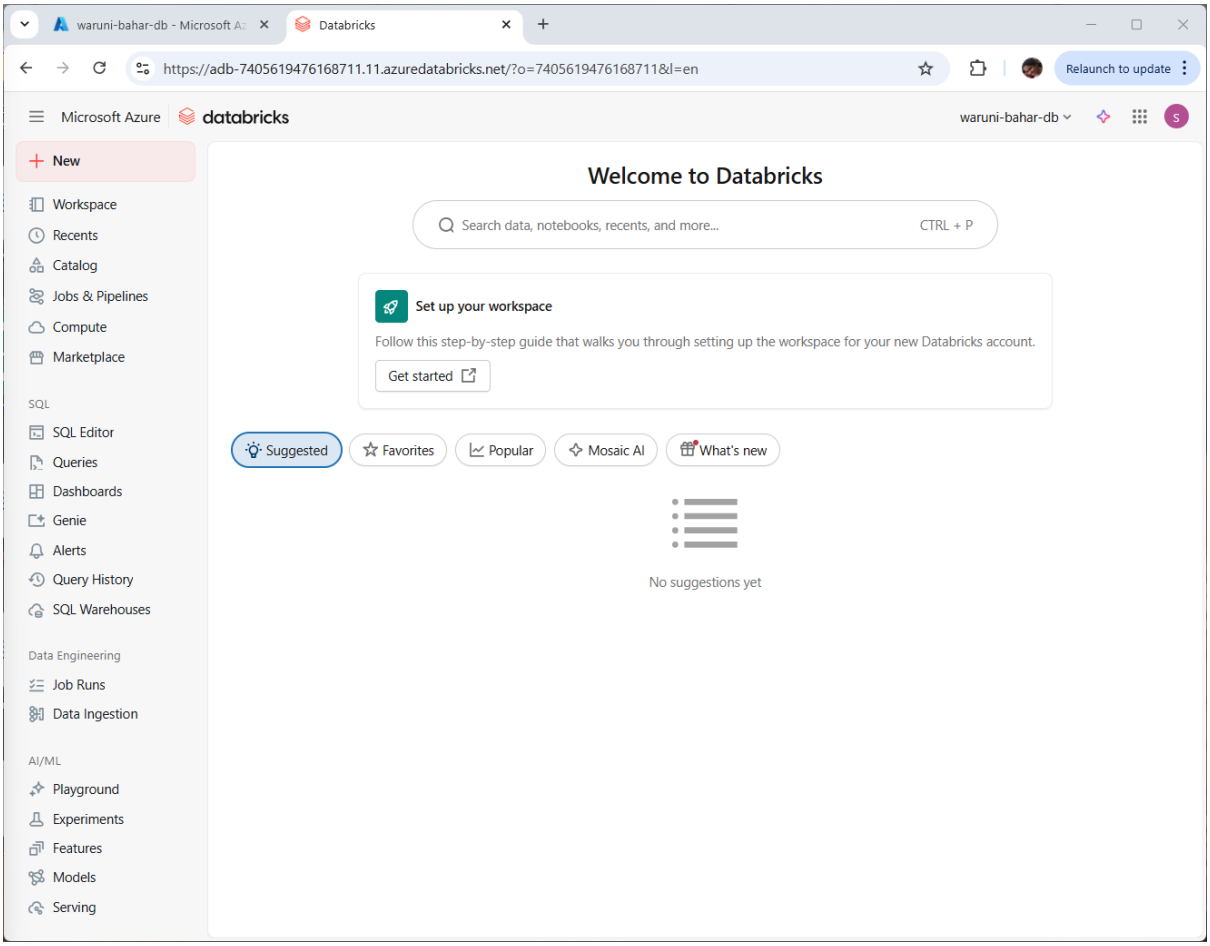
Azure Databricks Workspace and Cluster Setup

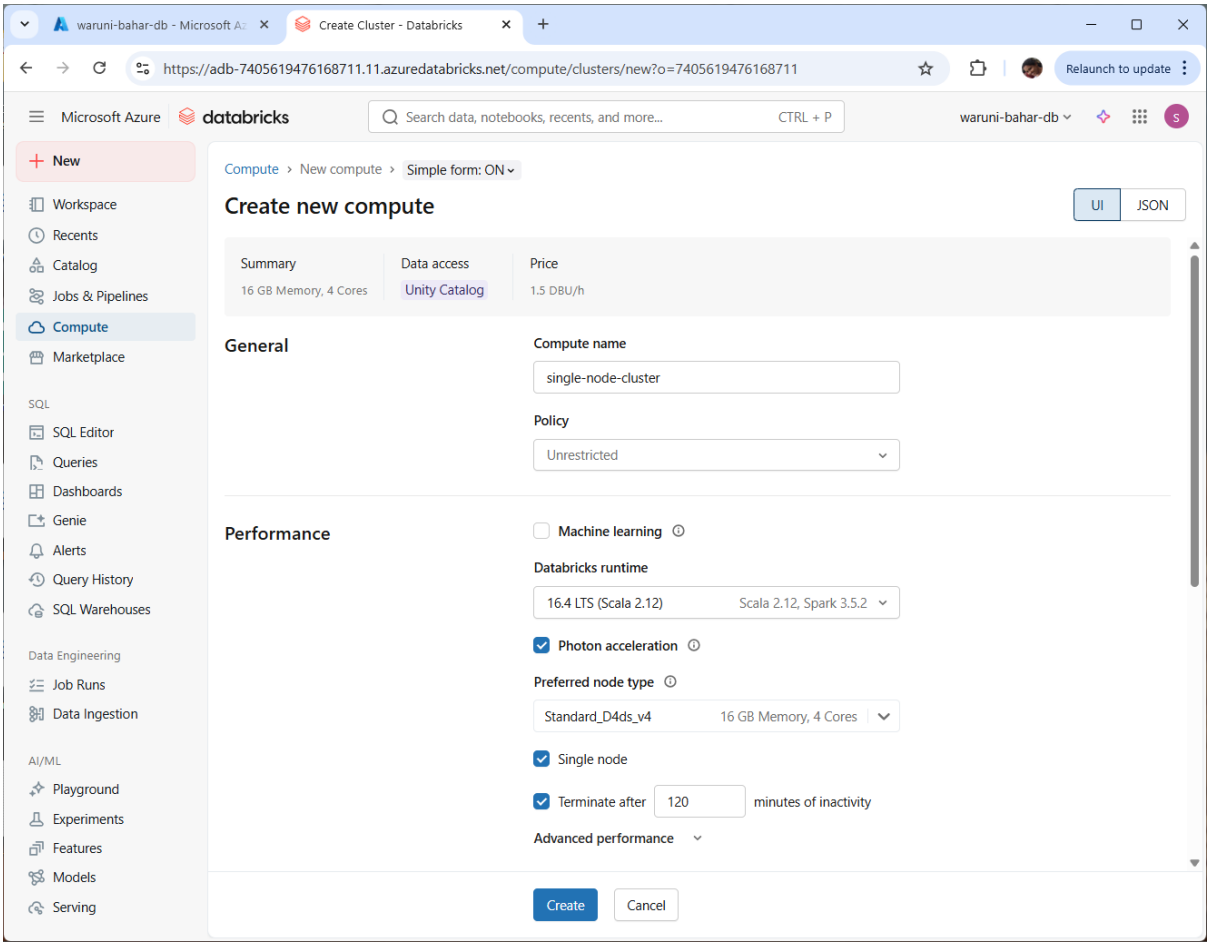
- An Azure Databricks workspace was created within the same resource group. A single-node cluster was configured with auto-termination enabled to minimise costs.

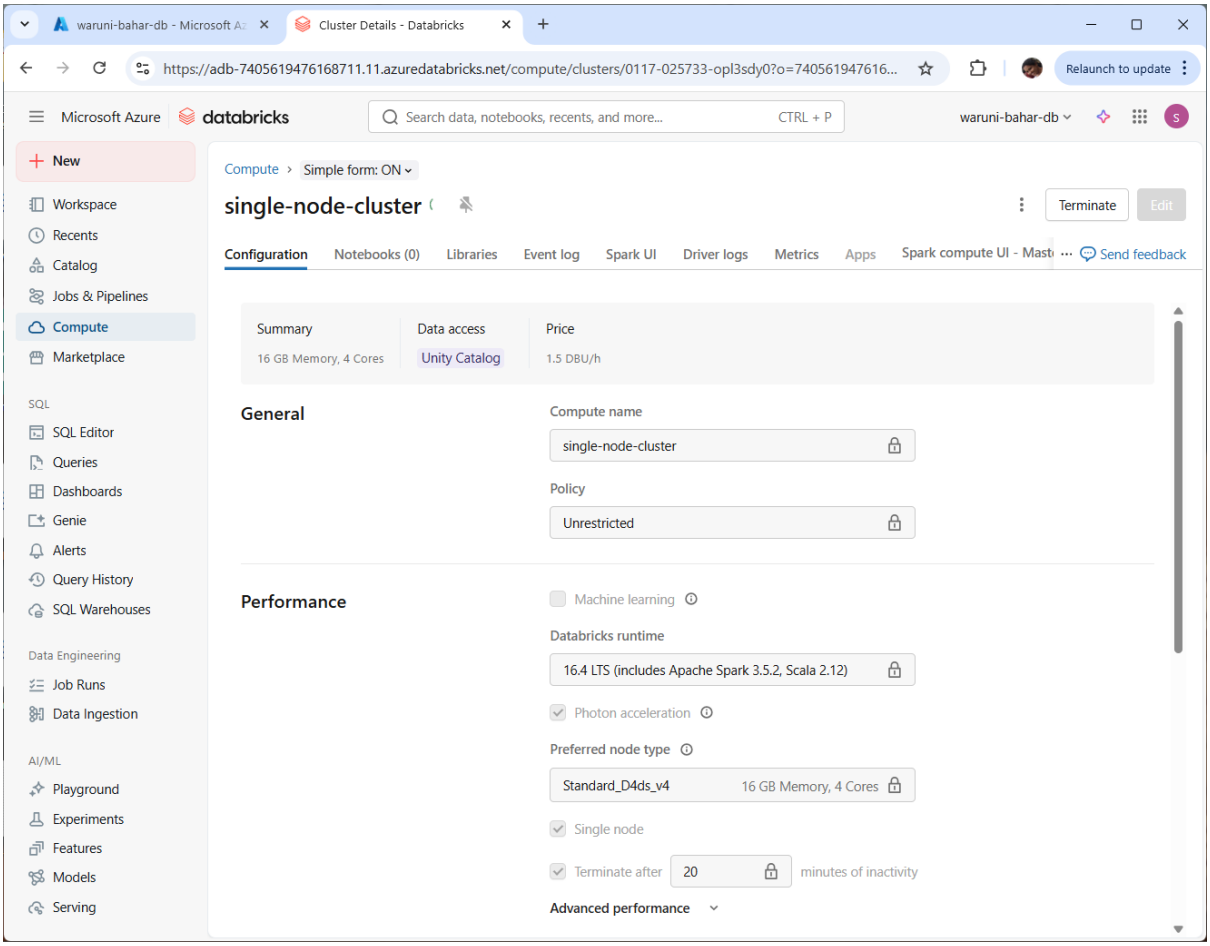



geni_Cloud_RG_waruni-bahar-d
https://portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id/%2Fsubscriptions...
Relaunch to update
Microsoft Azu
Search resources, services, and docs (G+/)
Copilot
genistudent97@gmail.c...
DEFAULT DIRECTORY (GENISTUD...
Home
geni_Cloud_RG_waruni-bahar-db | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Deployment is in progress
Deployment name : geni_Cloud_RG_waruni-bahar-db
Subscription : Azure subscription 1
Resource group : geni_Cloud_RG
Start time : 1/17/2026, 3:37:29 PM
Correlation ID : e9633b2e-d0c9-48cb-8ae8-cbfcb94aba20
Deployment details
Resource
Type
Status
Operation details
waruni-bahar-db
Azure Databricks Service
Created
Operation details
Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >
Free Microsoft tutorials
Add or remove favorites by pressing Ctrl+Shift+F


geni_Cloud_RG_waruni-bahar-d
https://portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id/%2Fsubscriptions...
Relaunch to update
Microsoft Azu
Search resources, services, and docs (G+/)
Copilot
genistudent97@gmail.c...
DEFAULT DIRECTORY (GENISTUD...
Home
geni_Cloud_RG_waruni-bahar-db | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name : geni_Cloud_RG_waruni-bahar-db
Subscription : Azure subscription 1
Resource group : geni_Cloud_RG
Start time : 1/17/2026, 3:37:36 PM
Correlation ID : e9633b2e-d0c9-48cb-8ae8-cbfcb94aba20
Deployment details
Next steps
Go to resource
Cost management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >
Add or remove favorites by pressing Ctrl+Shift+F

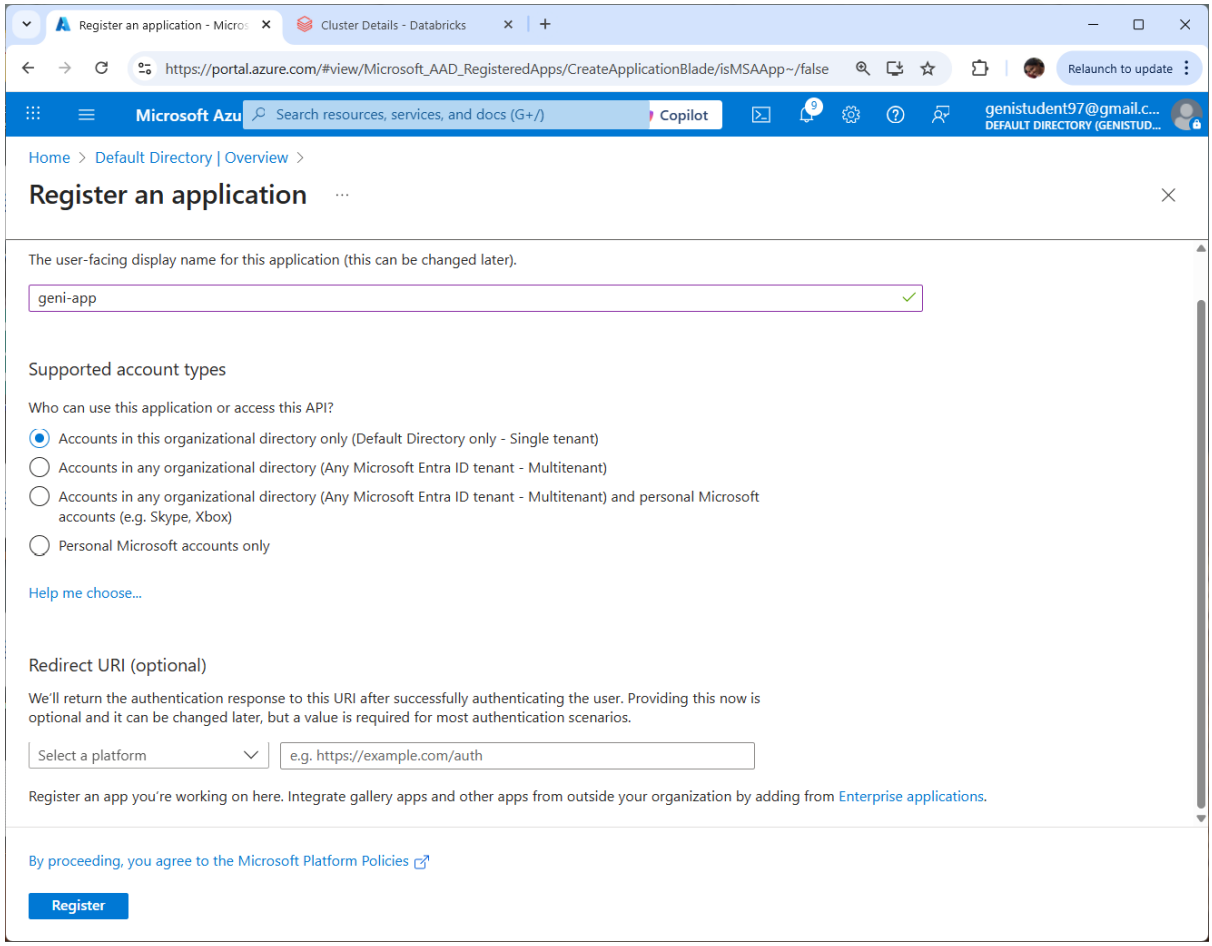

waruni-bahar-db - Microsoft A
Databricks
https://adb-7405619476168711.11.azuredatabricks.net/?o=7405619476168711&l=en
Relaunch to update
Microsoft Azure
databricks
waruni-bahar-db
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
AI/ML
Playground
Experiments
Features
Models
Serving
Welcome to Databricks
Search data, notebooks, recents, and more...
CTRL + P
Set up your workspace
Follow this step-by-step guide that walks you through setting up the workspace for your new Databricks account.
Get started
Suggested
Favorites
Popular
Mosaic AI
What's new
No suggestions yet

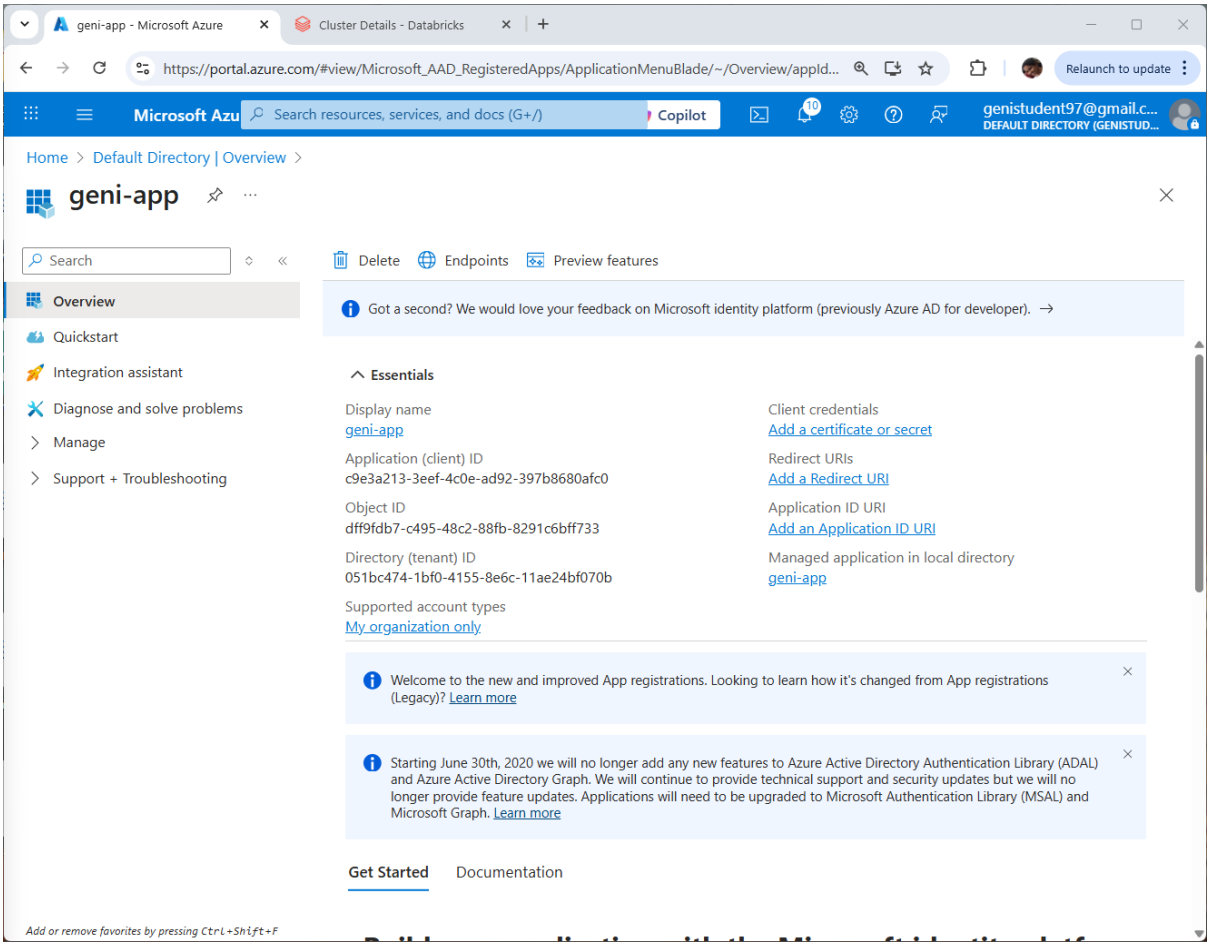

waruni-bahar-db - Microsoft A
Create Cluster - Databricks
https://adb-7405619476168711.11.azuredatabricks.net/compute/clusters/new?o=7405619476168711
Relaunch to update
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
waruni-bahar-db
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
AI/ML
Playground
Experiments
Features
Models
Serving
Compute › New compute › Simple form: ON
Create new compute
UI
JSON
Summary
16 GB Memory, 4 Cores
Data access
Unity Catalog
Price
1.5 DBU/h
General
Compute name
single-node-cluster
Policy
Unrestricted
Performance
Machine learning
Databricks runtime
16.4 LTS (Scala 2.12)
Scala 2.12, Spark 3.5.2
Photon acceleration
Preferred node type
Standard_D4ds_v4
16 GB Memory, 4 Cores
Single node
Terminate after
120
minutes of inactivity
Advanced performance
Create
Cancel

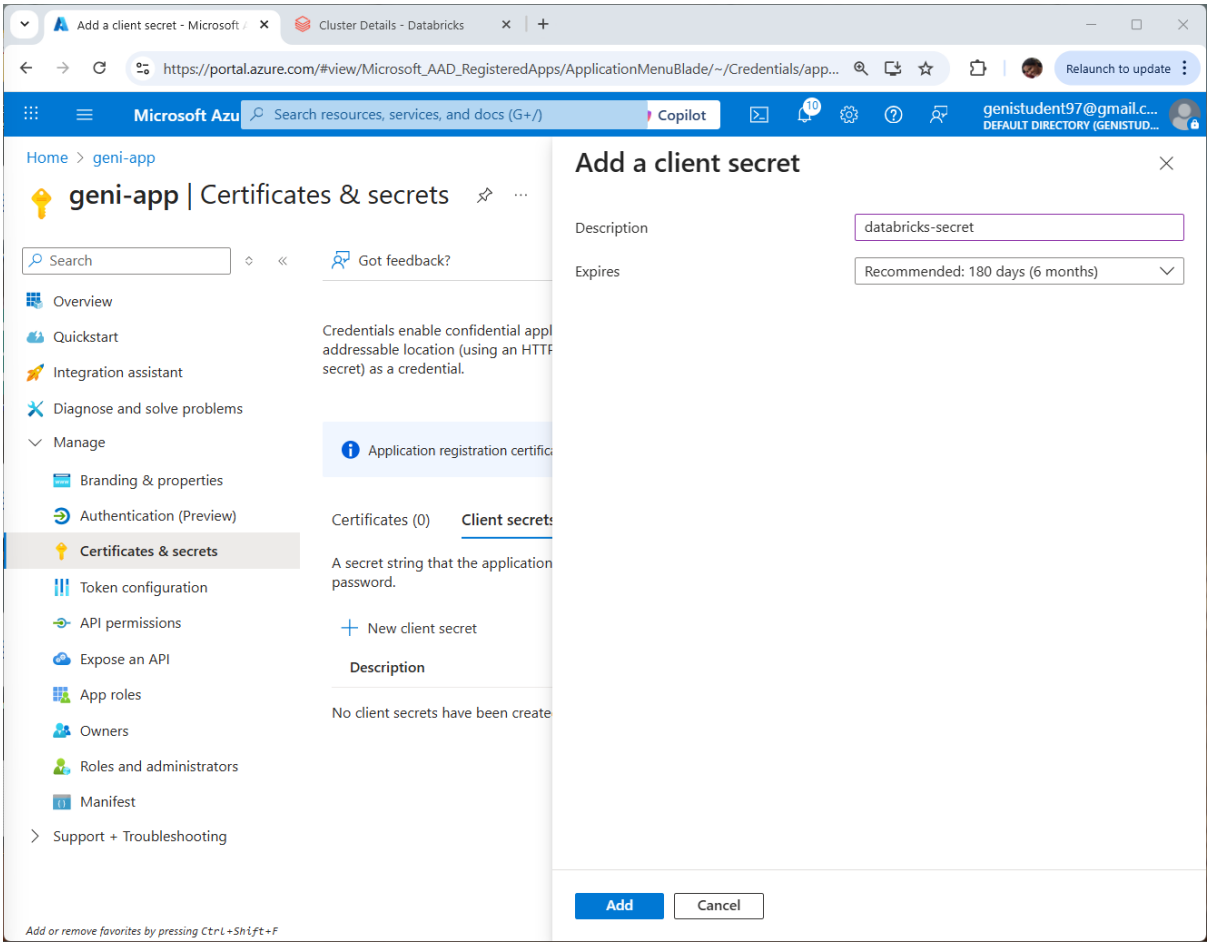


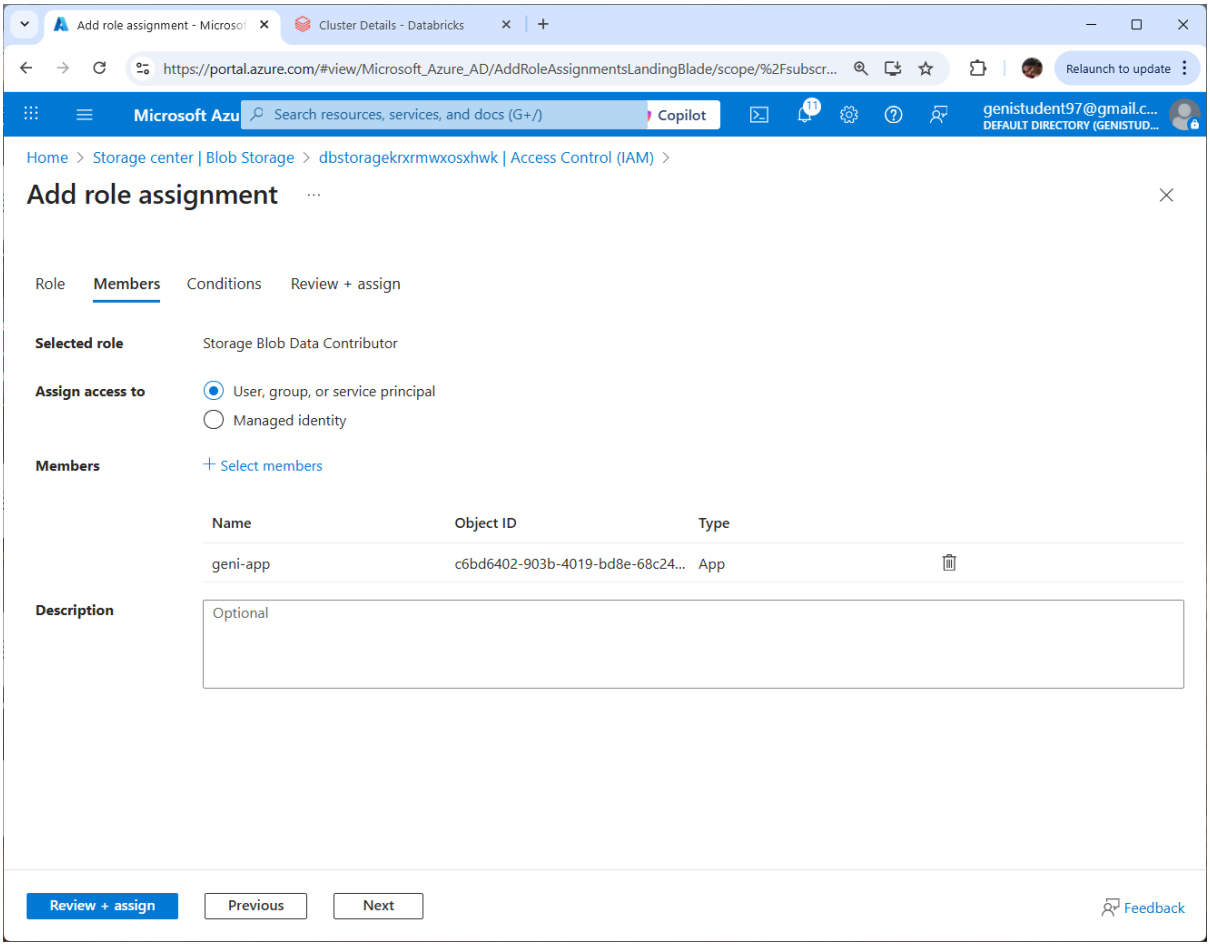
Secure Data Access and Authentication

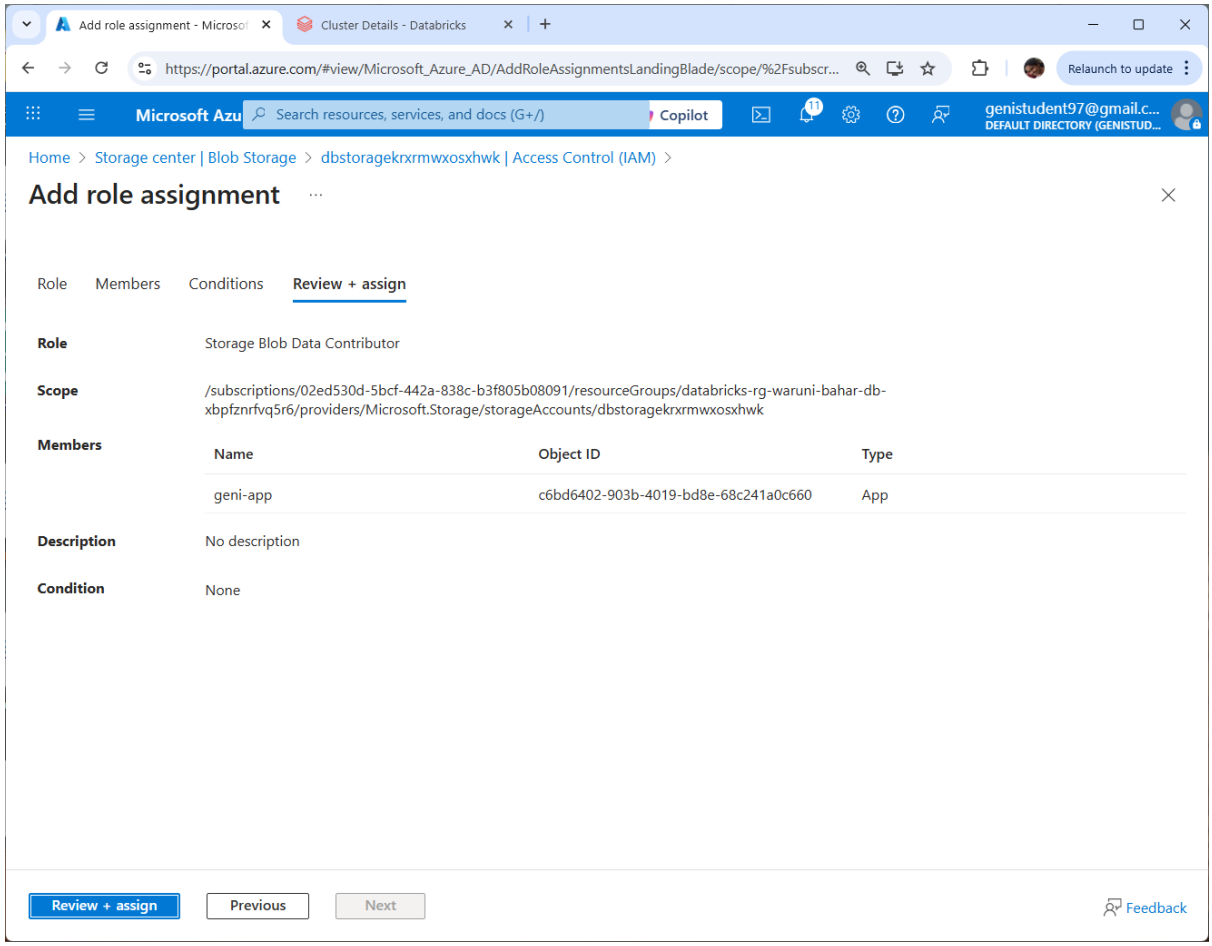
- An Azure App Registration was created to allow Databricks to securely access Azure Data Lake Storage Gen2. The app was granted the Storage Blob Data Contributor role.

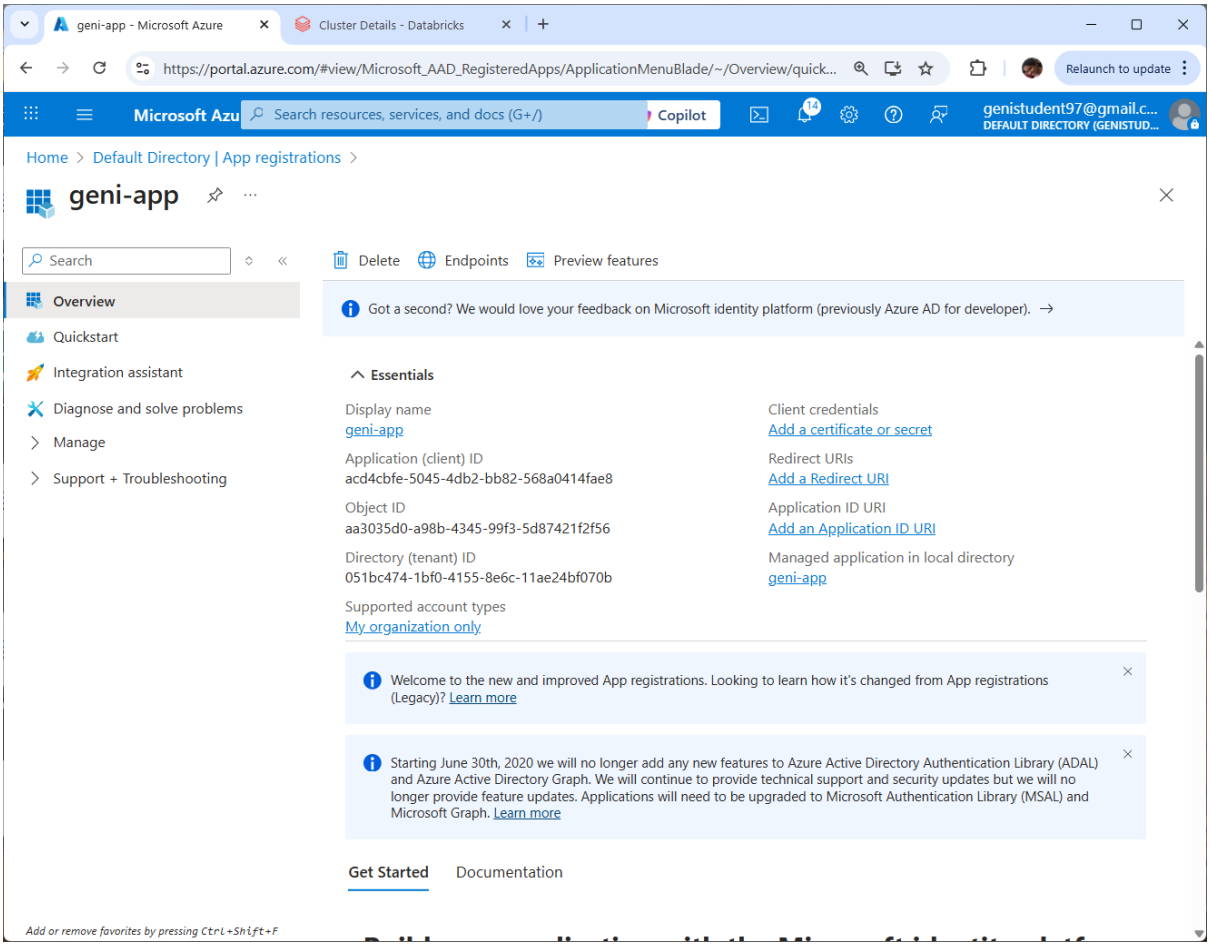

Register an application - Micros
Cluster Details - Databricks
https://portal.azure.com/#view/Microsoft_AAD_RegisteredApps/CreateApplicationBlade/isMSAApp~/false
Relaunch to update
Microsoft Azu
Search resources, services, and docs (G+/)
Copilot
genistudent97@gmail.c...
DEFAULT DIRECTORY (GENISTUD...
Home > Default Directory | Overview >
Register an application
The user-facing display name for this application (this can be changed later).
geni-app
Supported account types
Who can use this application or access this API?
Accounts in this organizational directory only (Default Directory only - Single tenant)
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)
Personal Microsoft accounts only
Help me choose...
Redirect URI (optional)
We'll return the authentication response to this URI after successfully authenticating the user. Providing this now is optional and it can be changed later, but a value is required for most authentication scenarios.
Select a platform
e.g. https://example.com/auth
Register an app you're working on here. Integrate gallery apps and other apps from outside your organization by adding from Enterprise applications.
By proceeding, you agree to the Microsoft Platform Policies
Register

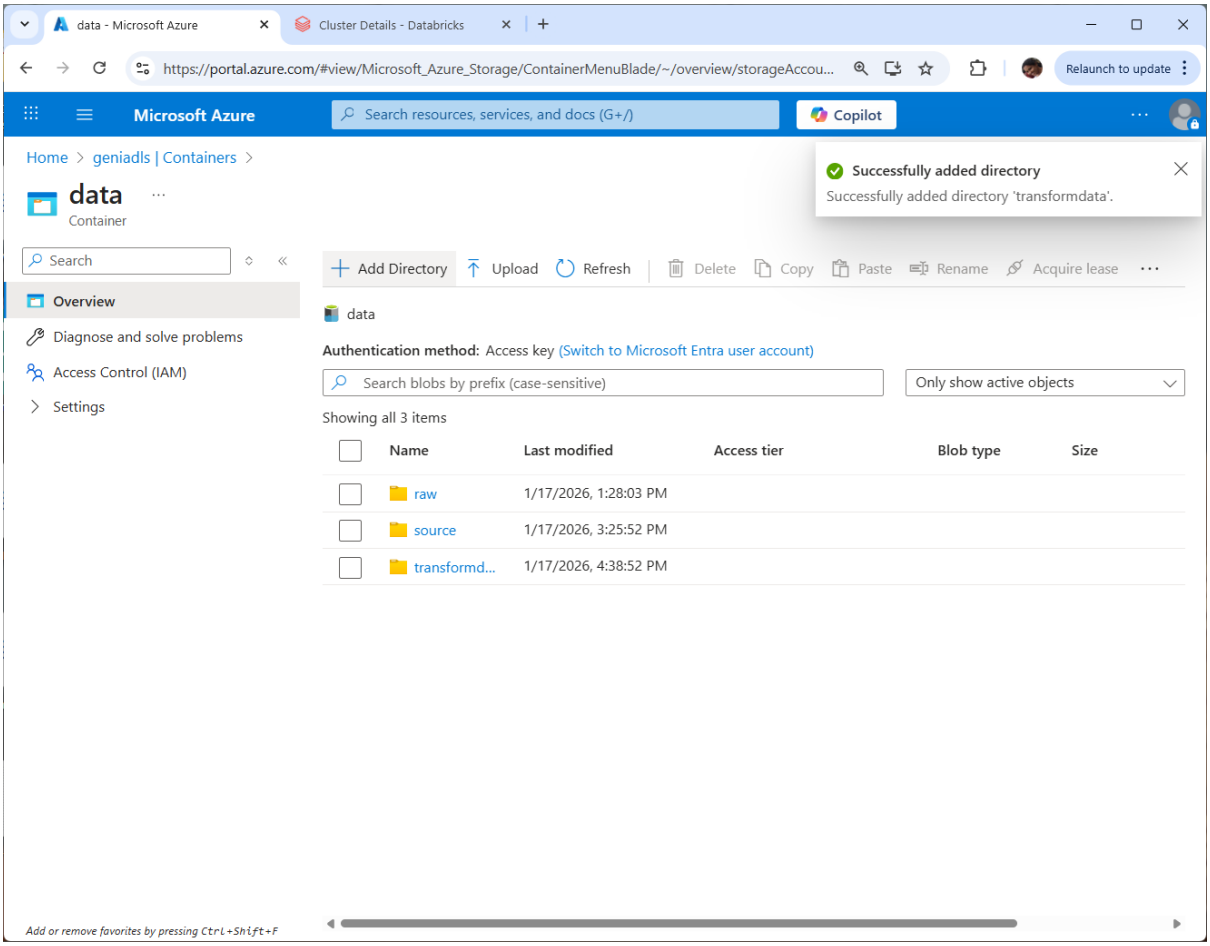

geni-app - Microsoft Azure
Cluster Details - Databricks
https://portal.azure.com/#view/Microsoft_AAD_RegisteredApps/ApplicationMenuBlade/~/Overview/appId...
Relaunch to update
Microsoft Azu
Search resources, services, and docs (G+/)
Copilot
genistudent97@gmail.c...
DEFAULT DIRECTORY (GENISTUD...
Home > Default Directory | Overview >
geni-app
Search
Overview
Quickstart
Integration assistant
Diagnose and solve problems
Manage
Support + Troubleshooting
Delete
Endpoints
Preview features
Got a second? We would love your feedback on Microsoft identity platform (previously Azure AD for developer). →
Essentials
Display name
geni-app
Application (client) ID
c9e3a213-3eef-4c0e-ad92-397b8680afc0
Object ID
dff9fdb7-c495-48c2-88fb-8291c6bff733
Directory (tenant) ID
051bc474-1bf0-4155-8e6c-11ae24bf070b
Supported account types
My organization only
Client credentials
Add a certificate or secret
Redirect URIs
Add a Redirect URI
Application ID URI
Add an Application ID URI
Managed application in local directory
geni-app
Welcome to the new and improved App registrations. Looking to learn how it's changed from App registrations (Legacy)? Learn more
Starting June 30th, 2020 we will no longer add any new features to Azure Active Directory Authentication Library (ADAL) and Azure Active Directory Graph. We will continue to provide technical support and security updates but we will no longer provide feature updates. Applications will need to be upgraded to Microsoft Authentication Library (MSAL) and Microsoft Graph. Learn more
Get Started
Documentation
Add or remove favorites by pressing Ctrl+Shift+F


Add a client secret - Microsoft
Cluster Details - Databricks
https://portal.azure.com/#view/Microsoft_AAD_RegisteredApps/ApplicationMenuBlade/~/Credentials/app...
Relaunch to update
Microsoft Azu
Search resources, services, and docs (G+/)
Copilot
genistudent97@gmail.c...
DEFAULT DIRECTORY (GENISTUD...
Home > geni-app
geni-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 appl
addressable location (using an HTTP
secret) as a credential.
Application registration certific
Certificates (0)
Client secrets
A secret string that the application
password.
New client secret
Description
No client secrets have been create
Add or remove favorites by pressing Ctrl+Shift+F
Add a client secret
Description
databricks-secret
Expires
Recommended: 180 days (6 months)
Add
Cancel


Add role assignment - Microsof
Cluster Details - Databricks
https://portal.azure.com/#view/Microsoft_Azure_AD/AddRoleAssignmentsLandingBlade/scope/%2Fsubscr...
Relaunch to update
Microsoft Azu
Search resources, services, and docs (G+/)
Copilot
genistudent97@gmail.c...
DEFAULT DIRECTORY (GENISTUD...
Home > Storage center | Blob Storage > dbstoragekrxrmwxosxhwk | 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
geni-app
c6bd6402-903b-4019-bd8e-68c24...
App
Description
Optional
Review + assign
Previous
Next
Feedback


Add role assignment - Microsof
Cluster Details - Databricks
https://portal.azure.com/#view/Microsoft_Azure_AD/AddRoleAssignmentsLandingBlade/scope/%2Fsubscr...
Relaunch to update
Microsoft Azu
Search resources, services, and docs (G+/)
Copilot
genistudent97@gmail.c...
DEFAULT DIRECTORY (GENISTUD...
Home > Storage center | Blob Storage > dbstoragekrxrmwxosxhwk | Access Control (IAM) >
Add role assignment
Role
Members
Conditions
Review + assign
Role
Storage Blob Data Contributor
Scope
/subscriptions/02ed530d-5bcf-442a-838c-b3f805b08091/resourceGroups/databricks-rg-waruni-bahar-db-xbpfznrfvq5r6/providers/Microsoft.Storage/storageAccounts/dbstoragekrxrmwxosxhwk
Members
Name
Object ID
Type
geni-app
c6bd6402-903b-4019-bd8e-68c241a0c660
App
Description
No description
Condition
None
Review + assign
Previous
Next
Feedback


geni-app - Microsoft Azure
Cluster Details - Databricks
https://portal.azure.com/#view/Microsoft_AAD_RegisteredApps/ApplicationMenuBlade/~/Overview/quick...
Relaunch to update
Microsoft Azu
Search resources, services, and docs (G+/)
Copilot
genistudent97@gmail.c...
DEFAULT DIRECTORY (GENISTUD...
Home > Default Directory | App registrations >
geni-app
Search
Overview
Quickstart
Integration assistant
Diagnose and solve problems
Manage
Support + Troubleshooting
Delete
Endpoints
Preview features
Got a second? We would love your feedback on Microsoft identity platform (previously Azure AD for developer). →
Essentials
Display name
geni-app
Application (client) ID
acd4cbfe-5045-4db2-bb82-568a0414fae8
Object ID
aa3035d0-a98b-4345-99f3-5d87421f2f56
Directory (tenant) ID
051bc474-1bf0-4155-8e6c-11ae24bf070b
Supported account types
My organization only
Client credentials
Add a certificate or secret
Redirect URIs
Add a Redirect URI
Application ID URI
Add an Application ID URI
Managed application in local directory
geni-app
Welcome to the new and improved App registrations. Looking to learn how it's changed from App registrations (Legacy)? Learn more
Starting June 30th, 2020 we will no longer add any new features to Azure Active Directory Authentication Library (ADAL) and Azure Active Directory Graph. We will continue to provide technical support and security updates but we will no longer provide feature updates. Applications will need to be upgraded to Microsoft Authentication Library (MSAL) and Microsoft Graph. Learn more
Get Started
Documentation
Add or remove favorites by pressing Ctrl+Shift+F



Data Transformation Using PySpark

- PySpark code was used to read data from the /raw folder, perform transformations such as data type casting and aggregation, and write the processed data to the /transformdata folder.
- This step ensured that data was cleaned and optimised for analytics.

Databricks notebook with transformation code

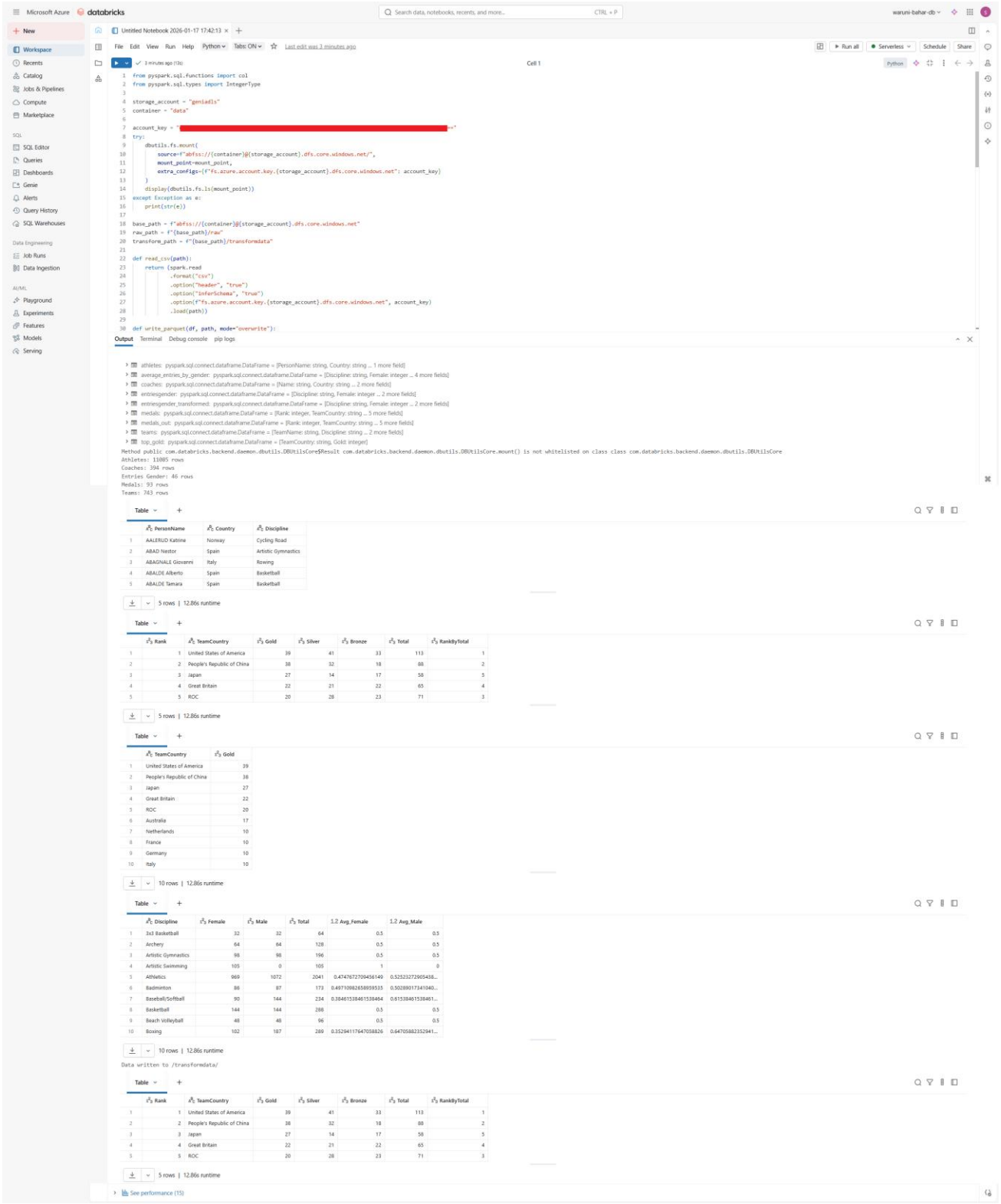
```
from pyspark.sql.functions import col
from pyspark.sql.types import IntegerType

storage_account = "geniadls"
container = "data"

account_key = " <storage-account-key-masked>"
```

```
try:
    dbutils.fs.mount(
        source=f"abfss://{container}@{storage_account}.dfs.core.windows.net/",
        mount_point=mount_point,
        extra_configs={f"fs.azure.account.key.{storage_account}.dfs.core.windows.net": account_key}
    )
    display(dbutils.fs.ls(mount_point))
except Exception as e:
    print(str(e))

base_path = f"abfss://{container}@{storage_account}.dfs.core.windows.net"
raw_path = f"{base_path}/raw"
transform_path = f"{base_path}/transformdata"

def read_csv(path):
    return (spark.read
            .format("csv")
            .option("header", "true")
            .option("inferSchema", "true")
            .option(f"fs.azure.account.key.{storage_account}.dfs.core.windows.net", account_key)
            .load(path))

def write_parquet(df, path, mode="overwrite"):
    (df.write
     .format("parquet")
     .mode(mode)
     .option(f"fs.azure.account.key.{storage_account}.dfs.core.windows.net", account_key)
     .save(path))

athletes = read_csv(f"{raw_path}/Athletes.csv")
```

```
coaches = read_csv(f"{raw_path}/Coaches.csv")
entriesgender = read_csv(f"{raw_path}/EntriesGender.csv")
medals = read_csv(f"{raw_path}/Medals.csv")
teams = read_csv(f"{raw_path}/Teams.csv")

print(f"Athletes: {athletes.count()} rows")
print(f"Coaches: {coaches.count()} rows")
print(f"Entries Gender: {entriesgender.count()} rows")
print(f"Medals: {medals.count()} rows")
print(f"Teams: {teams.count()} rows")

display(athletes.limit(5))
display(medals.limit(5))

entriesgender_transformed = (entriesgender
   .withColumn("Female", col("Female").cast(IntegerType()))
   .withColumn("Male", col("Male").cast(IntegerType()))
   .withColumn("Total", col("Total").cast(IntegerType()))
)

top_gold = medals.orderBy(col("Gold").desc()).select("TeamCountry", "Gold")
display(top_gold.limit(10))

average_entries_by_gender = (entriesgender_transformed
   .withColumn("Avg_Female", col("Female") / col("Total"))
   .withColumn("Avg_Male", col("Male") / col("Total"))
)

display(average_entries_by_gender.limit(10))
```

```
write_parquet(athletes, f"{transform_path}/athletes")
write_parquet(coaches, f"{transform_path}/coaches")
write_parquet(entriesgender_transformed, f"{transform_path}/entriesgender")
write_parquet(medals, f"{transform_path}/medals")
write_parquet(teams, f"{transform_path}/teams")

print("Data written to /transformdata/")

medals_out = (spark.read
        .format("parquet")
        .option(f"fs.azure.account.key.{storage_account}.dfs.core.windows.net", account_key)
        .load(f"{transform_path}/medals"))

display(medals_out.limit(5))
```

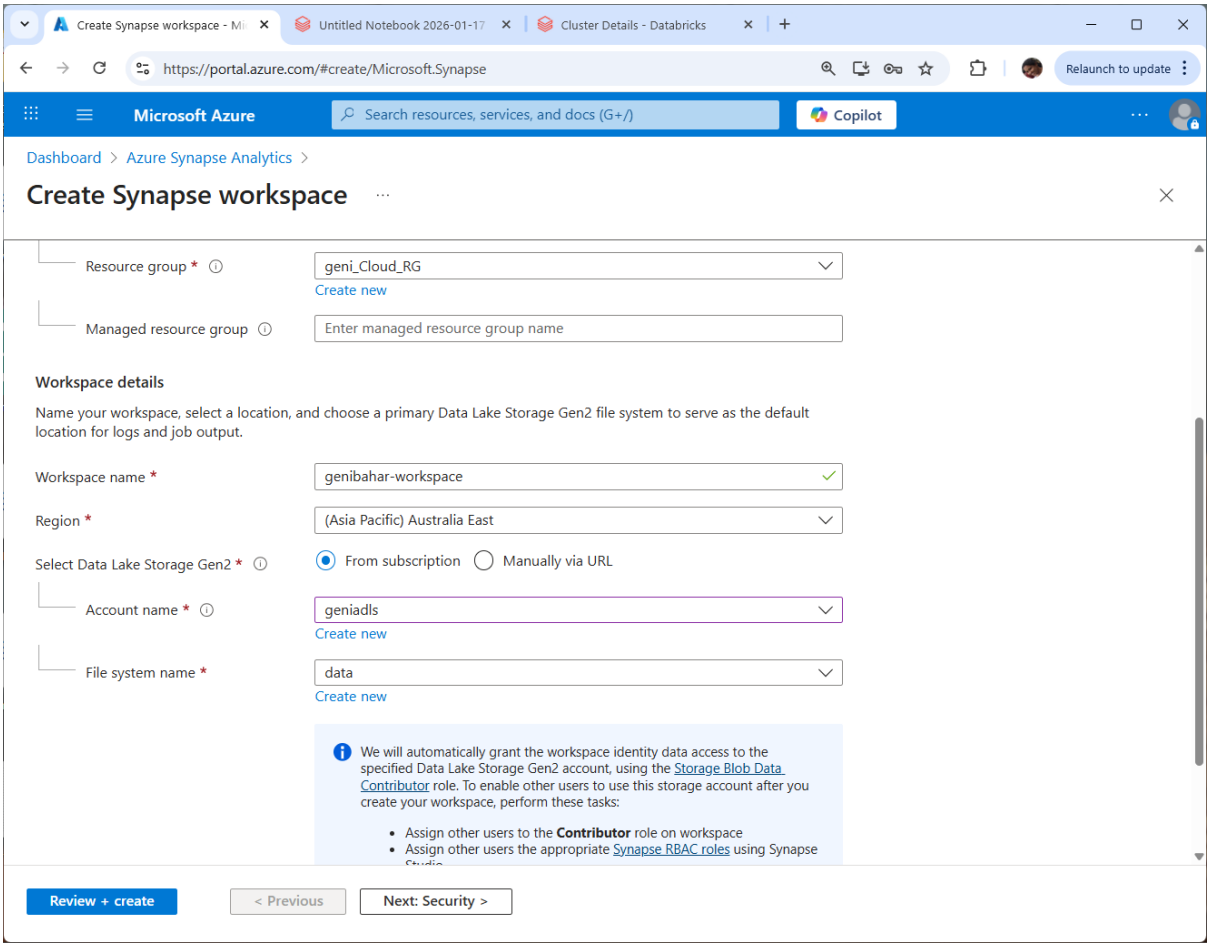
Code output

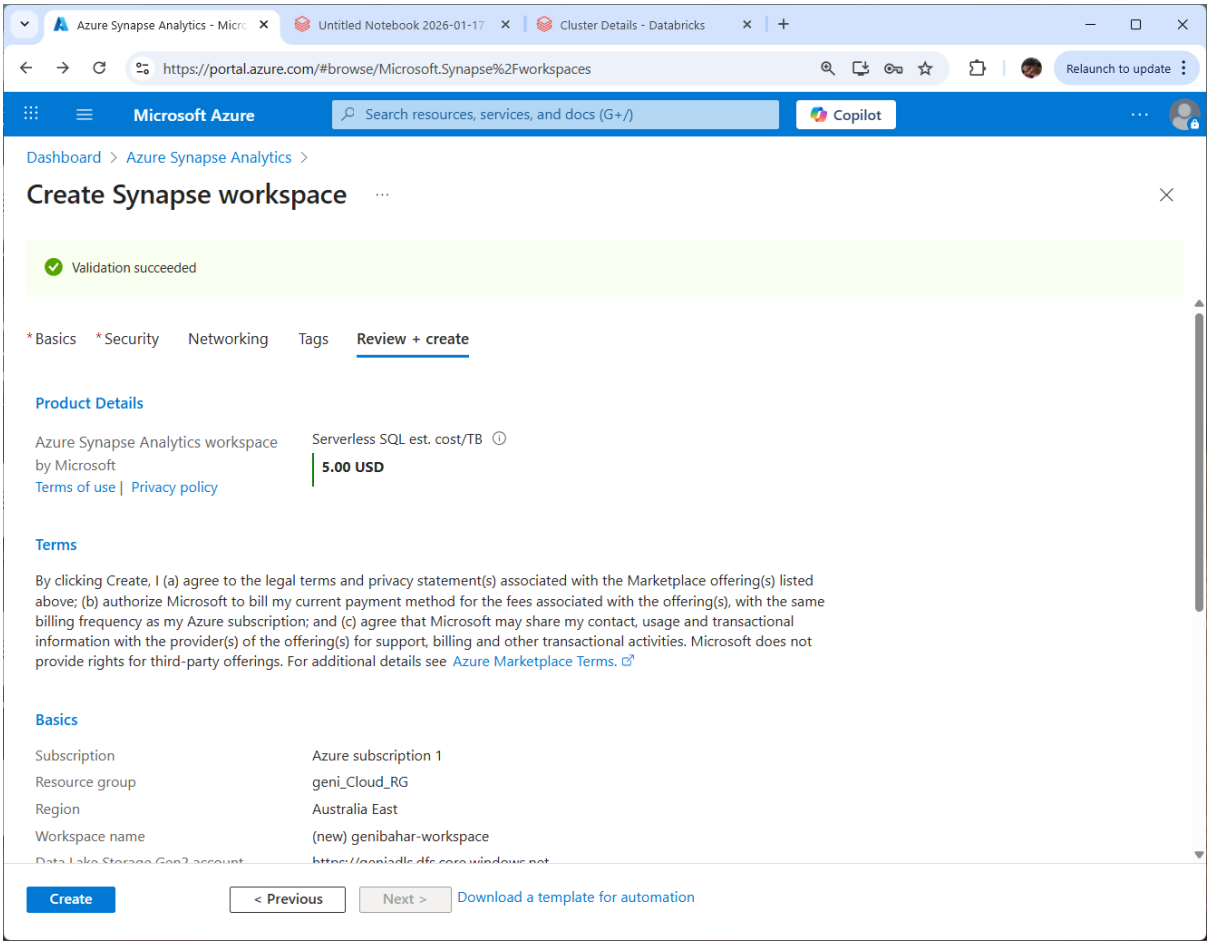

Microsoft Azure
databricks
Untitled Notebook 2026-01-17 17:42:13
Cell 1
from pyspark.sql.functions import col
from pyspark.sql.types import IntegerType
storage_account = "geniadls"
container = "data"
base_path = f"abfss://{container}@{storage_account}.dfs.core.windows.net"
raw_path = f"{base_path}/raw"
transform_path = f"{base_path}/transformdata"
def read_csv(path):
def write_parquet(df, path, mode="overwrite"):
Output
Terminal
Debug console
pip logs
Athletes: 11085 rows
Coaches: 394 rows
Entries Gender: 46 rows
Medals: 93 rows
Teams: 743 rows
Data written to /transformdata/

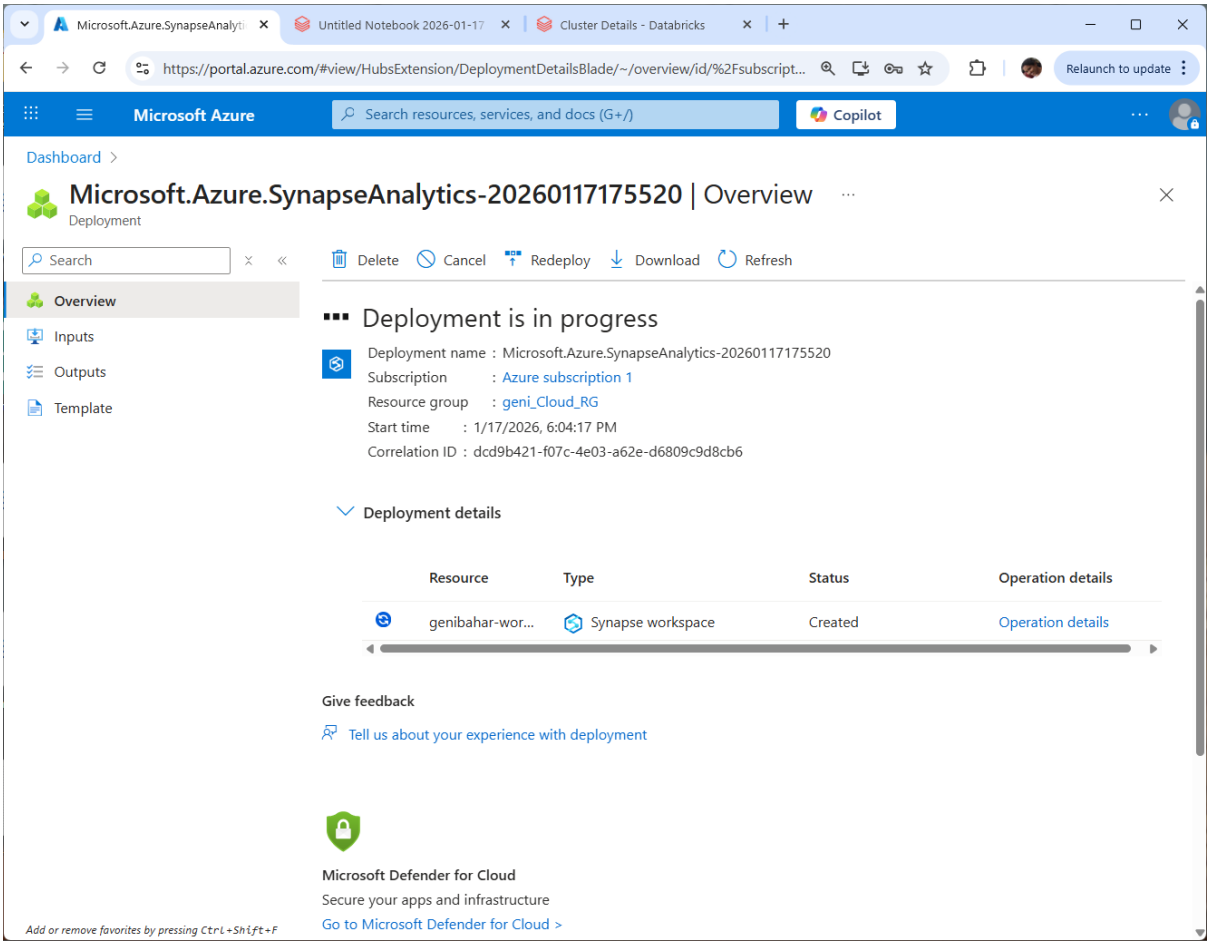
Task 5: Azure Synapse Analytics and Automated Data Integration

Azure Synapse Analytics Workspace Creation

- An Azure Synapse Analytics workspace was created in the same resource group and region. The workspace was linked to the Azure Data Lake Storage Gen2 account to enable direct access to transformed datasets.

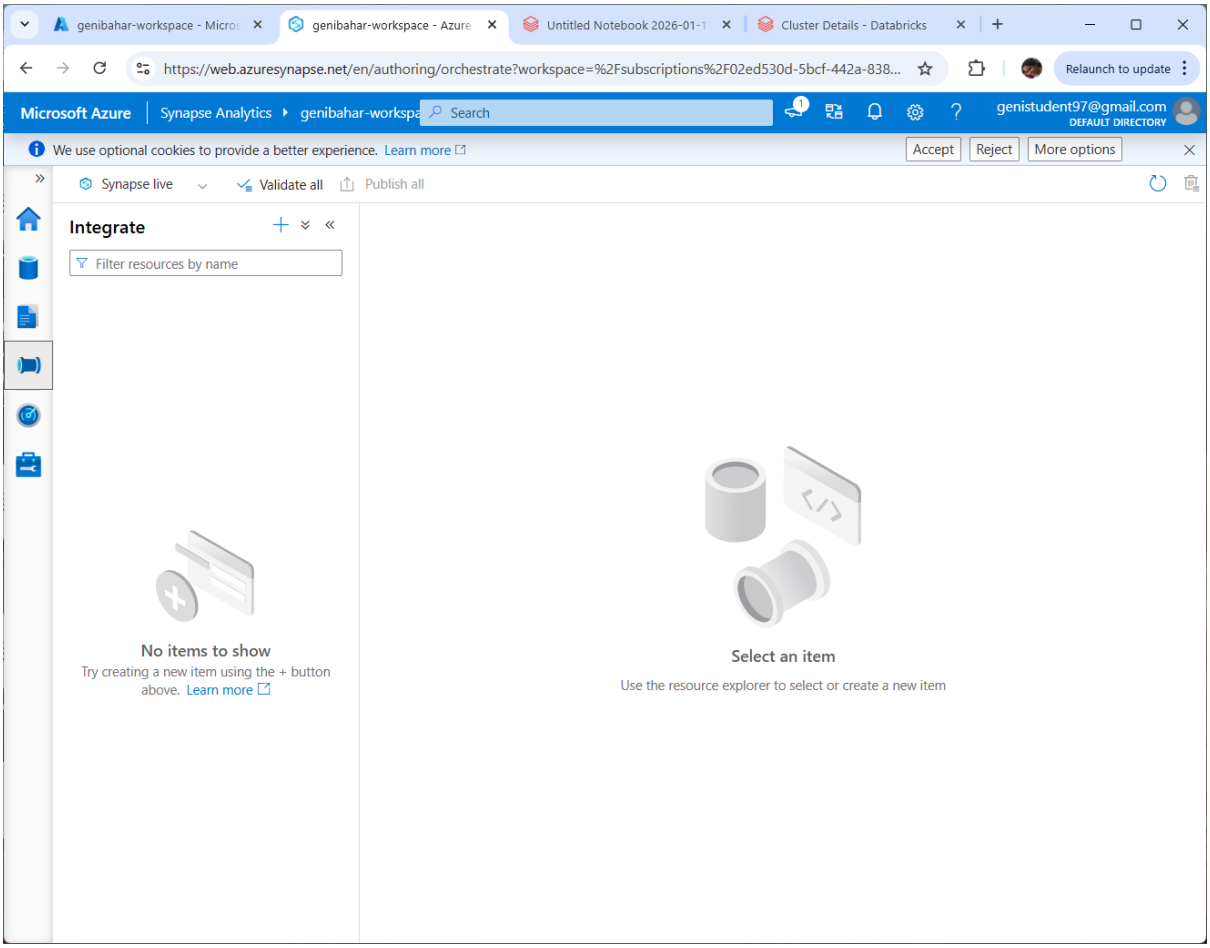


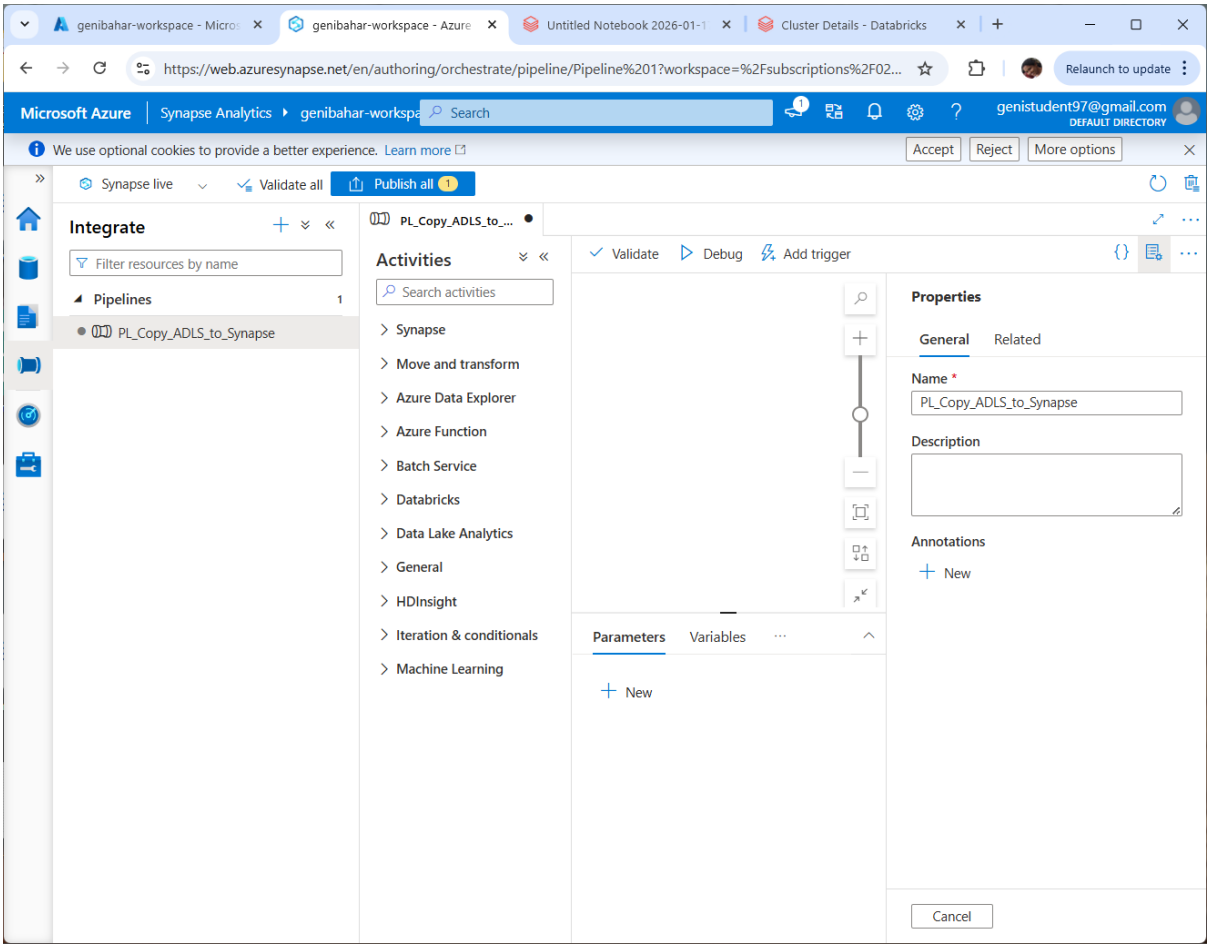

Azure Synapse Analytics - Micr
Untitled Notebook 2026-01-17
Cluster Details - Databricks
https://portal.azure.com/#browse/Microsoft.Synapse%2Fworkspaces
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Dashboard > Azure Synapse Analytics >
Create Synapse workspace
Validation succeeded
*Basics *Security Networking Tags Review + create
Product Details
Azure Synapse Analytics workspace
by Microsoft
Terms of use | Privacy policy
Serverless SQL est. cost/TB
5.00 USD
Terms
By clicking Create, I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. For additional details see Azure Marketplace Terms.
Basics
Subscription Azure subscription 1
Resource group geni_Cloud_RG
Region Australia East
Workspace name (new) genibahar-workspace
Create
< Previous
Next >
Download a template for automation

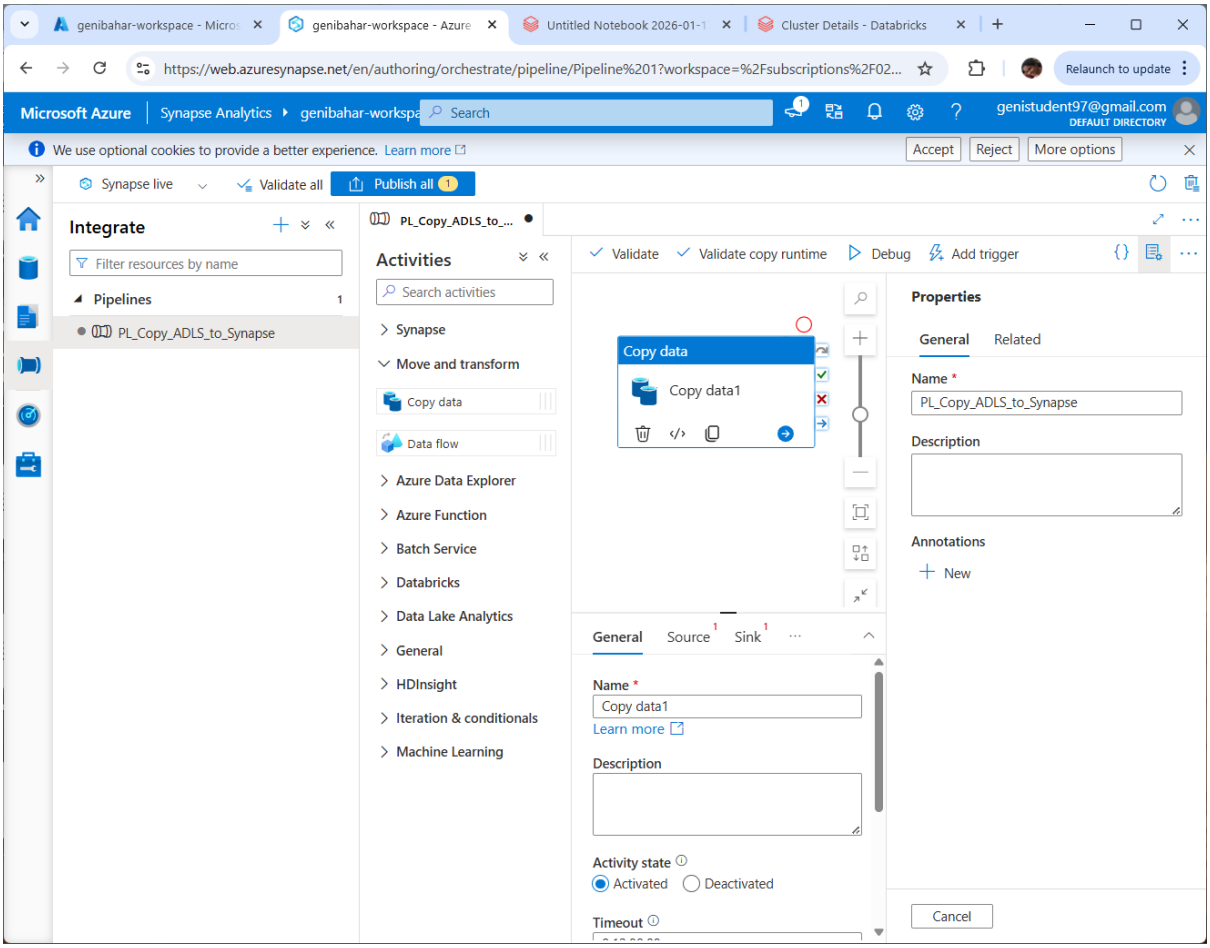


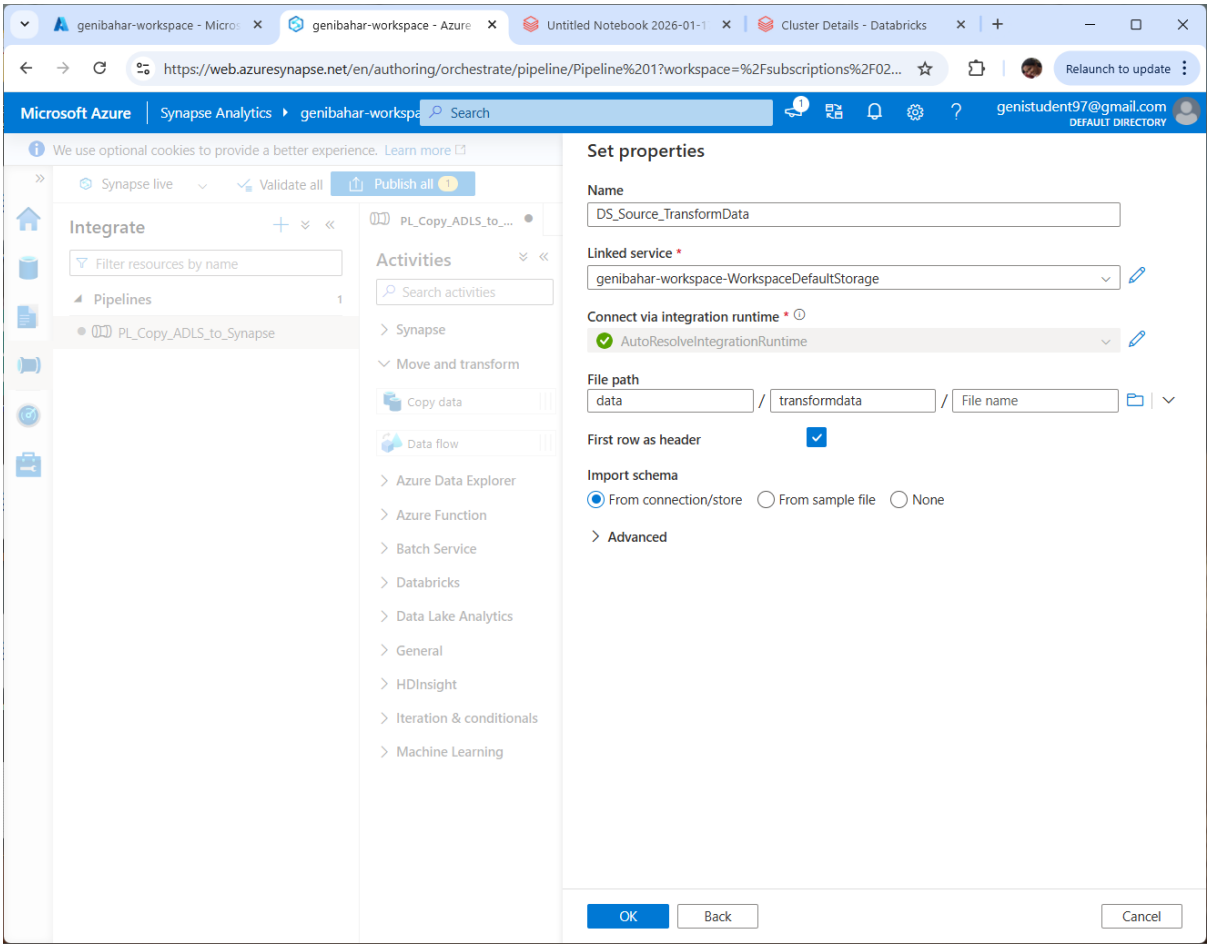
Data Integration Pipeline

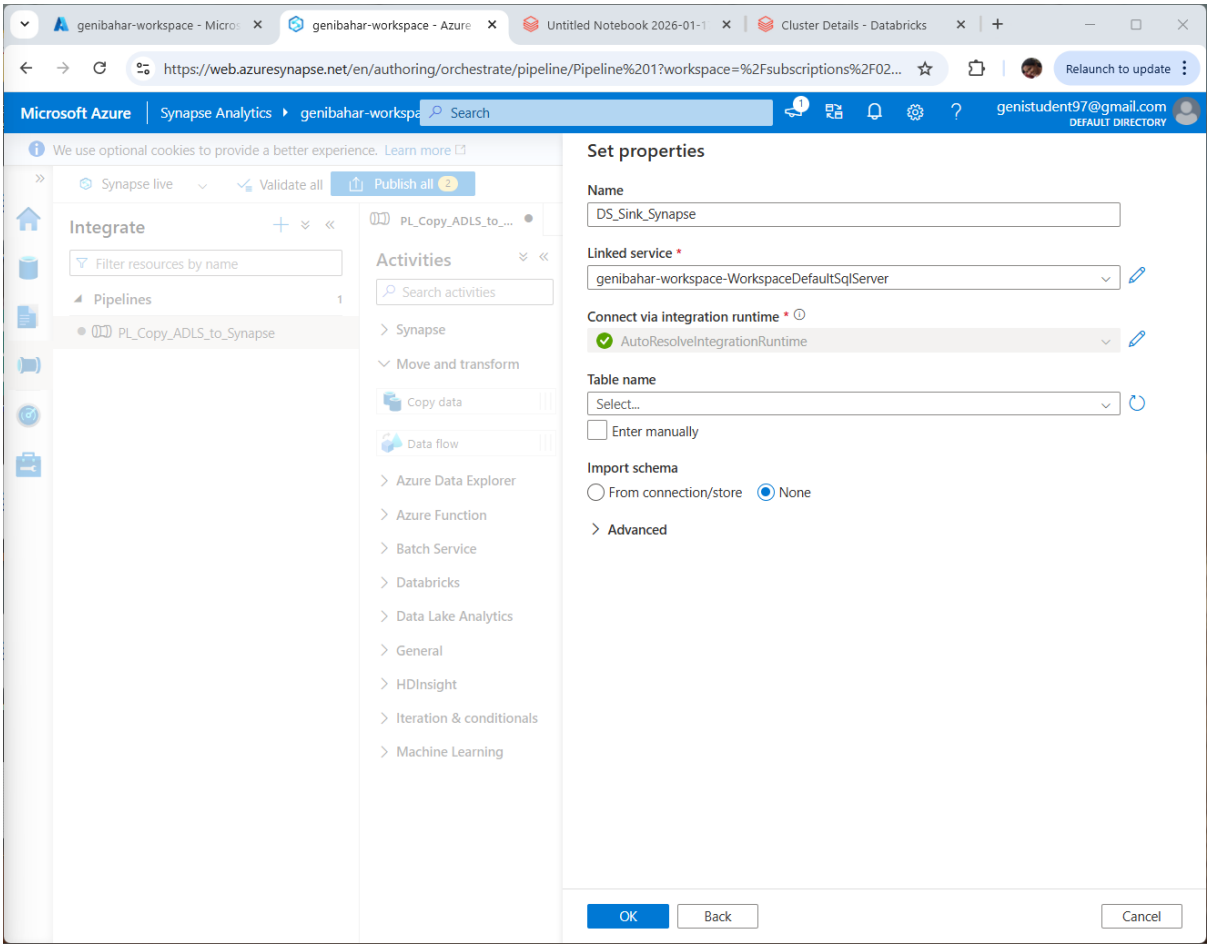
- A Synapse pipeline was designed to automate the data integration process between Azure Data Lake Storage Gen2 and Synapse Analytics. The pipeline was validated and executed successfully.

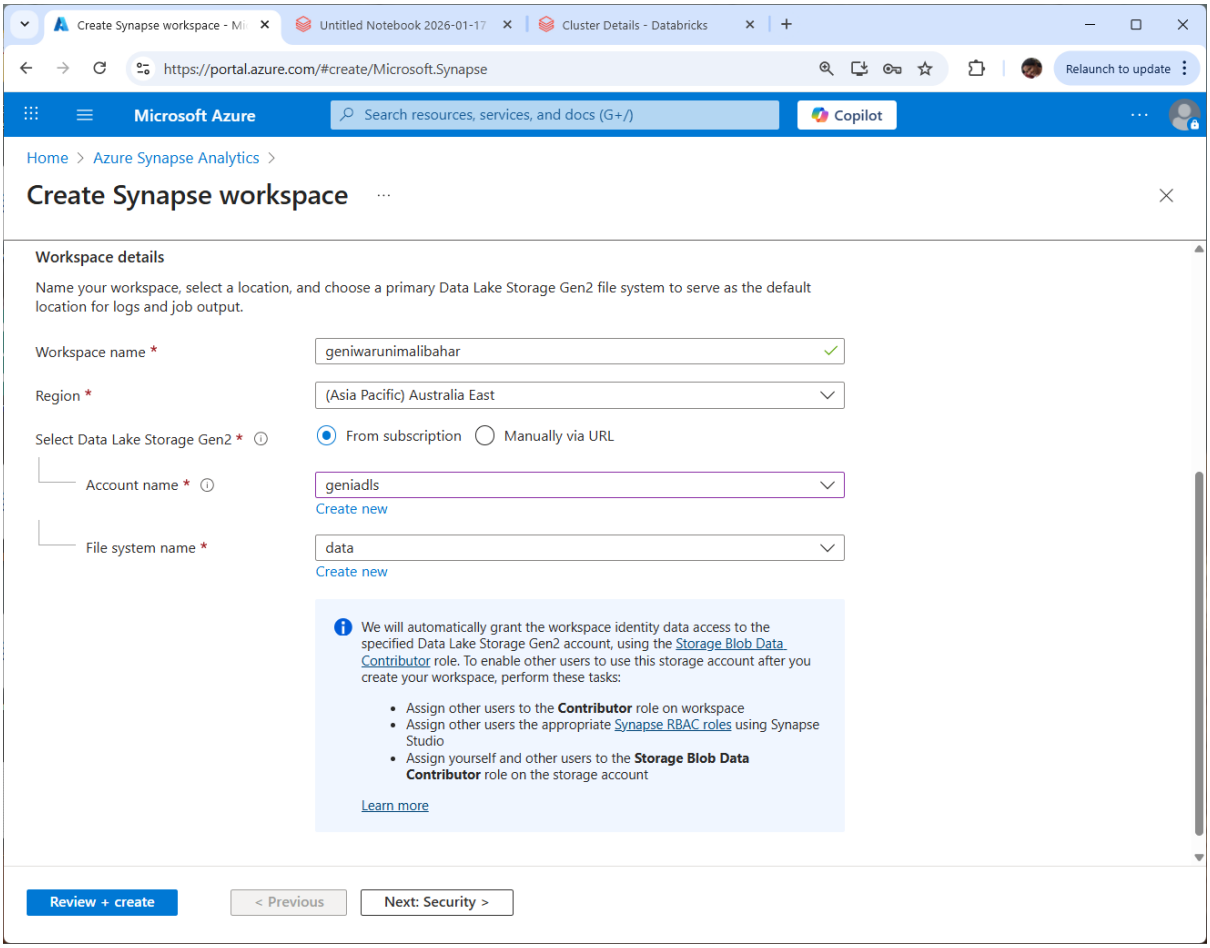

genibahar-workspace - Micros
genibahar-workspace - Azure
Untitled Notebook 2026-01-1
Cluster Details - Databricks
https://web.azuresynapse.net/en/authoring/orchestrate?workspace=%2Fsubscriptions%2F02ed530d-5bcf-442a-838...
Relaunch to update
Microsoft Azure
Synapse Analytics
genibahar-workspa
Search
genistudent97@gmail.com
DEFAULT DIRECTORY
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse live
Validate all
Publish all
Integrate
Filter resources by name
No items to show
Try creating a new item using the + button above. Learn more
Select an item
Use the resource explorer to select or create a new item

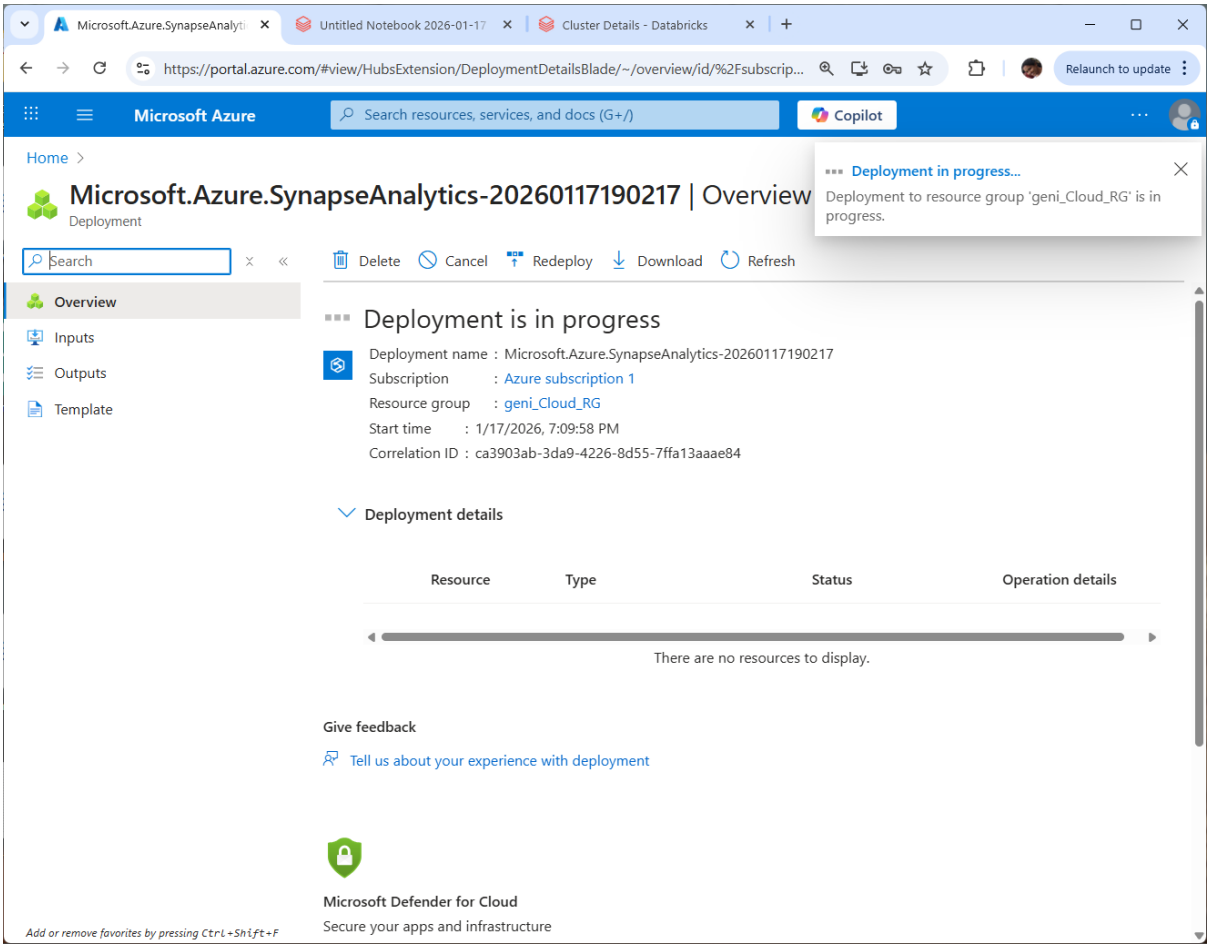

genibahar-workspace - Micros
genibahar-workspace - Azure
Untitled Notebook 2026-01-1
Cluster Details - Databricks
https://web.azuresynapse.net/en/authoring/orchestrate/pipeline/Pipeline%201?workspace=%2Fsubscriptions%2F02...
Relaunch to update
Microsoft Azure
Synapse Analytics
genibahar-workspa
Search
genistudent97@gmail.com
DEFAULT DIRECTORY
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse live
Validate all
Publish all 1
Integrate
Filter resources by name
Pipelines 1
PL_Copy_ADLS_to_Synapse
PL_Copy_ADLS_to_...
Activities
Search activities
Synapse
Move and transform
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Validate
Debug
Add trigger
Parameters
Variables
New
Properties
General
Related
Name *
PL_Copy_ADLS_to_Synapse
Description
Annotations
New
Cancel

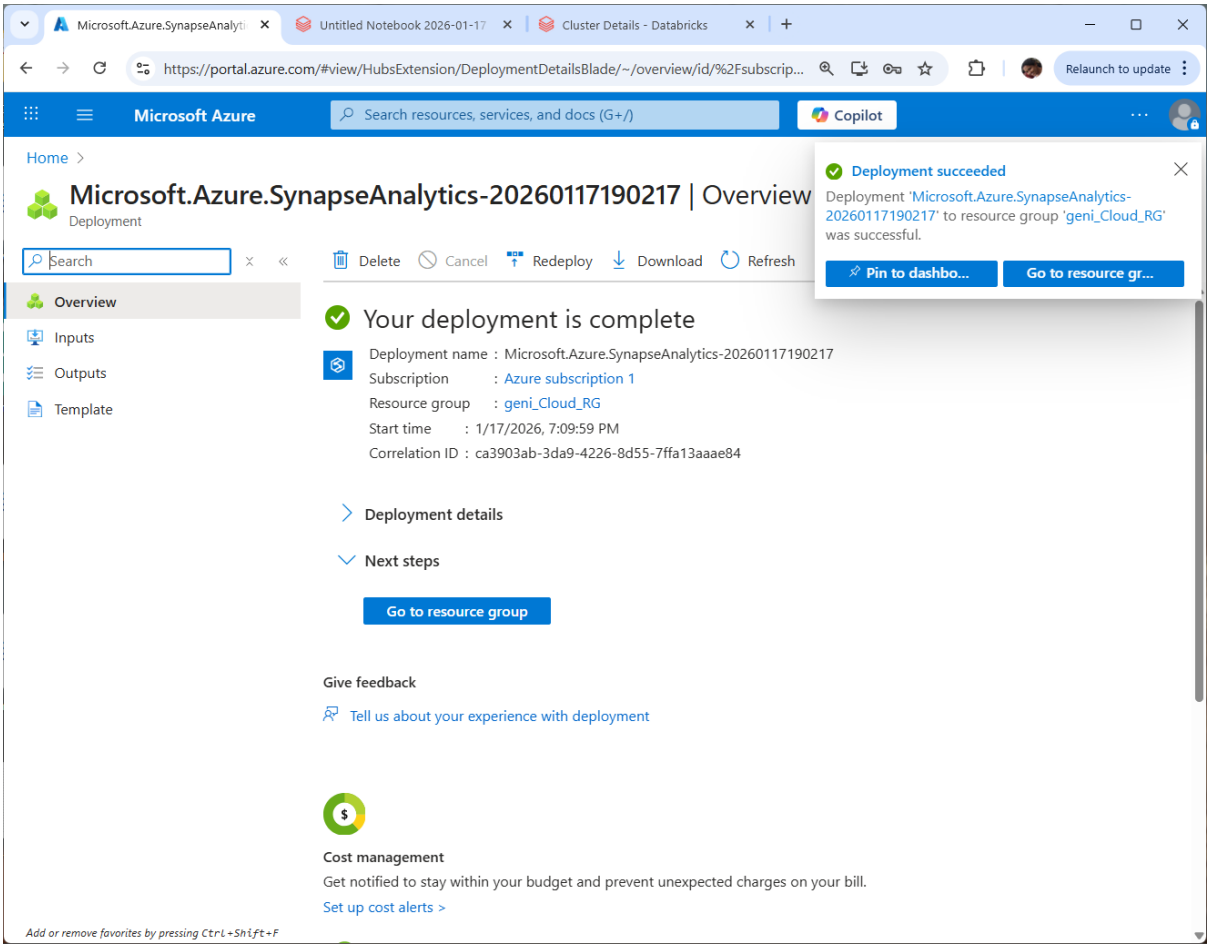

genibahar-workspace - Azure
Untitled Notebook 2026-01-1
Cluster Details - Databricks
https://web.azuresynapse.net/en/authoring/orchestrate/pipeline/Pipeline%201?workspace=%2Fsubscriptions%2F02...
Relaunch to update
Microsoft Azure
Synapse Analytics
genibahar-workspa
Search
genistudent97@gmail.com
DEFAULT DIRECTORY
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse live
Validate all
Publish all 1
Integrate
Filter resources by name
Pipelines 1
PL_Copy_ADLS_to_Synapse
PL_Copy_ADLS_to_...
Activities
Search activities
Synapse
Move and transform
Copy data
Data flow
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Validate
Validate copy runtime
Debug
Add trigger
Copy data
Copy data1
Properties
General
Related
Name *
PL_Copy_ADLS_to_Synapse
Description
Annotations
New
General
Source 1
Sink 1
Name *
Copy data1
Learn more
Description
Activity state
Activated
Deactivated
Timeout
Cancel


genibahar-workspace - Micros
genibahar-workspace - Azure
Untitled Notebook 2026-01-1
Cluster Details - Databricks
https://web.azuresynapse.net/en/authoring/orchestrate/pipeline/Pipeline%201?workspace=%2Fsubscriptions%2F02...
Relaunch to update
Microsoft Azure
Synapse Analytics
genibahar-workspa
Search
genistudent97@gmail.com
DEFAULT DIRECTORY
We use optional cookies to provide a better experience. Learn more
Synapse live
Validate all
Publish all
Integrate
Filter resources by name
Pipelines
PL_Copy_ADLS_to_Synapse
PL_Copy_ADLS_to_...
Activities
Search activities
Synapse
Move and transform
Copy data
Data flow
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Set properties
Name
DS_Source_TransformData
Linked service *
genibahar-workspace-WorkspaceDefaultStorage
Connect via integration runtime *
AutoResolveIntegrationRuntime
File path
data
transformdata
File name
First row as header
Import schema
From connection/store
From sample file
None
Advanced
OK
Back
Cancel


genibahar-workspace - Micros
genibahar-workspace - Azure
Untitled Notebook 2026-01-1
Cluster Details - Databricks
https://web.azuresynapse.net/en/authoring/orchestrate/pipeline/Pipeline%201?workspace=%2Fsubscriptions%2F02...
Relaunch to update
Microsoft Azure
Synapse Analytics
genibahar-workspa
Search
genistudent97@gmail.com
DEFAULT DIRECTORY
We use optional cookies to provide a better experience. Learn more
Synapse live
Validate all
Publish all 2
Integrate
Filter resources by name
Pipelines 1
PL_Copy_ADLS_to_Synapse
PL_Copy_ADLS_to_...
Activities
Search activities
Synapse
Move and transform
Copy data
Data flow
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Set properties
Name
DS_Sink_Synapse
Linked service *
genibahar-workspace-WorkspaceDefaultSqlServer
Connect via integration runtime *
AutoResolveIntegrationRuntime
Table name
Select...
Enter manually
Import schema
From connection/store
None
Advanced
OK
Back
Cancel

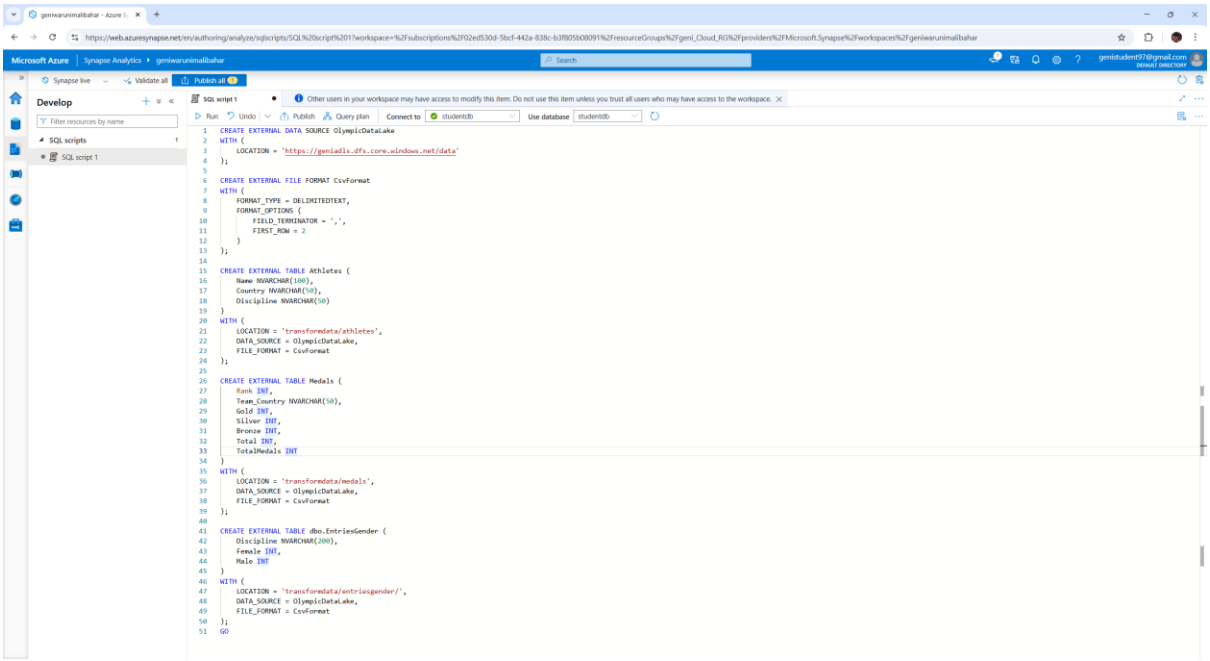

Create Synapse workspace - Mi
Untitled Notebook 2026-01-17
Cluster Details - Databricks
https://portal.azure.com/#create/Microsoft.Synapse
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home > Azure Synapse Analytics >
Create Synapse workspace
Workspace details
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 *
geniwarunimalibahar
Region *
(Asia Pacific) Australia East
Select Data Lake Storage Gen2 *
From subscription
Manually via URL
Account name *
geniadls
Create new
File system name *
data
Create new
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.SynapseAnalyti
Untitled Notebook 2026-01-17
Cluster Details - Databricks
https://portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id/%2Fsubscrip...
Relaunch to update
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Microsoft.Azure.SynapseAnalytics-20260117190217 | Overview
Deployment
Deployment in progress...
Deployment to resource group 'geni_Cloud_RG' is in progress.
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Deployment is in progress
Deployment name : Microsoft.Azure.SynapseAnalytics-20260117190217
Subscription : Azure subscription 1
Resource group : geni_Cloud_RG
Start time : 1/17/2026, 7:09:58 PM
Correlation ID : ca3903ab-3da9-4226-8d55-7ffa13aaae84
Deployment details
Resource
Type
Status
Operation details
There are no resources to display.
Give feedback
Tell us about your experience with deployment
Microsoft Defender for Cloud
Secure your apps and infrastructure
Add or remove favorites by pressing Ctrl+Shift+F



External Table Creation Using SQL

- External tables were created in Synapse using SQL scripts. These tables reference data stored in the /transformdata folder, allowing serverless querying without data duplication.

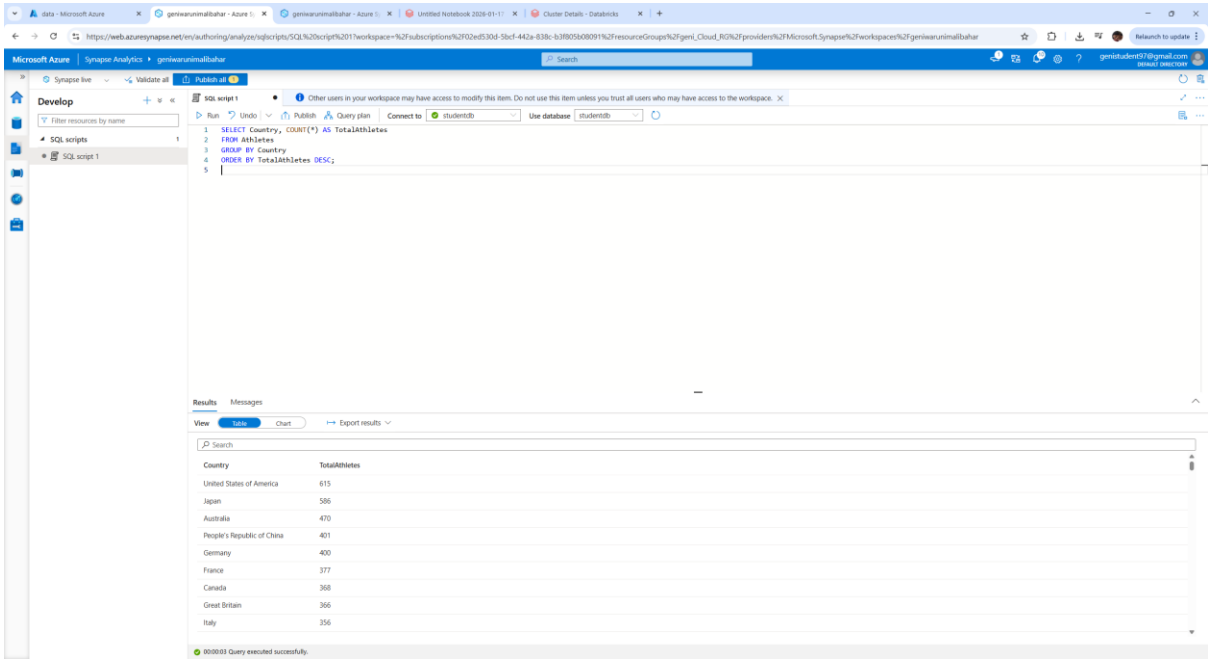


```
CREATE EXTERNAL DATA SOURCE OlympicDataLake
WITH (
    LOCATION = 'https://geniadls.dfs.core.windows.net/data'
);

CREATE EXTERNAL FILE FORMAT CsvFormat
WITH (
    FORMAT_TYPE = DELIMITEDTEXT,
    FORMAT_OPTIONS (
        FIELD_TERMINATOR = ',',
        FIRST_ROW = 2
    )
);

CREATE EXTERNAL TABLE Athletes (
    Name NVARCHAR(100),
    Country NVARCHAR(50),
    Discipline NVARCHAR(50)
)
WITH (
    LOCATION = 'transformdata/athletes',
    DATA_SOURCE = OlympicDataLake,
    FILE_FORMAT = CsvFormat
);

CREATE EXTERNAL TABLE Medals (
    Rank INT,
    Team_Country NVARCHAR(50),
    Gold INT,
    Silver INT,
    Bronze INT,
    Total INT,
    TotalMedals INT
)
WITH (
    LOCATION = 'transformdata/medals',
    DATA_SOURCE = OlympicDataLake,
    FILE_FORMAT = CsvFormat
);

CREATE EXTERNAL TABLE dbo.EntriesGender (
    Discipline NVARCHAR(200),
    Female INT,
    Male INT
)
WITH (
    LOCATION = 'transformdata/entriesgender/',
    DATA_SOURCE = OlympicDataLake,
    FILE_FORMAT = CsvFormat
);
GO
```

Data Analysis Using SQL

Several SQL queries were executed to analyse the dataset:

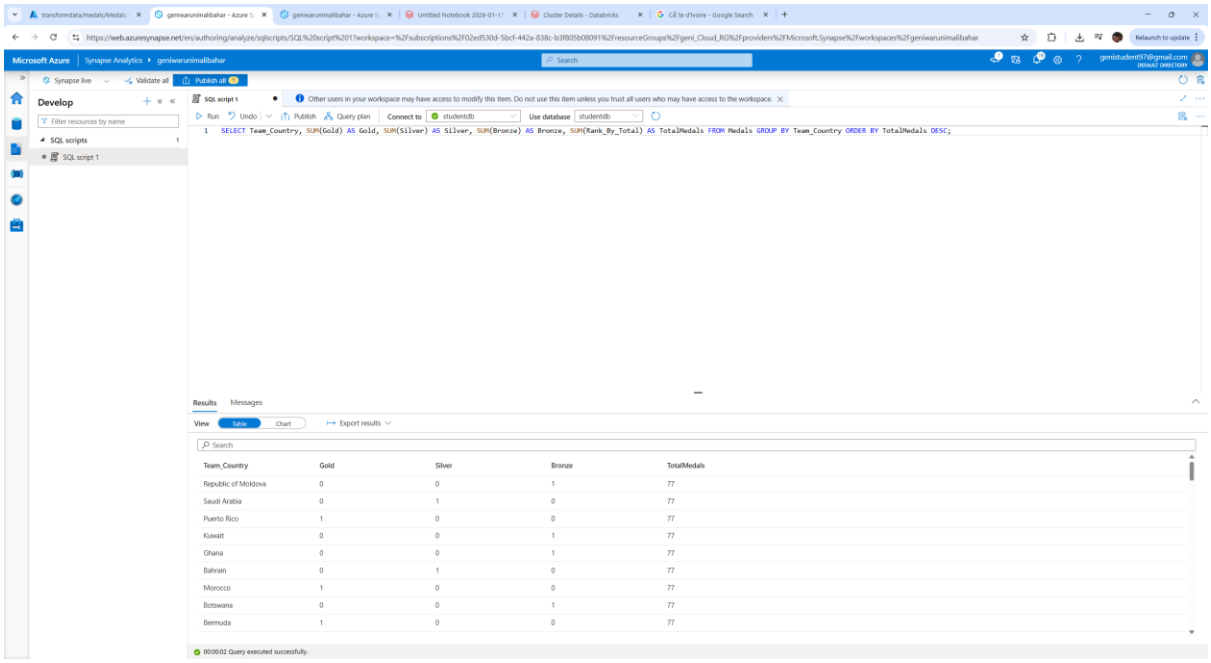
- Number of athletes by country



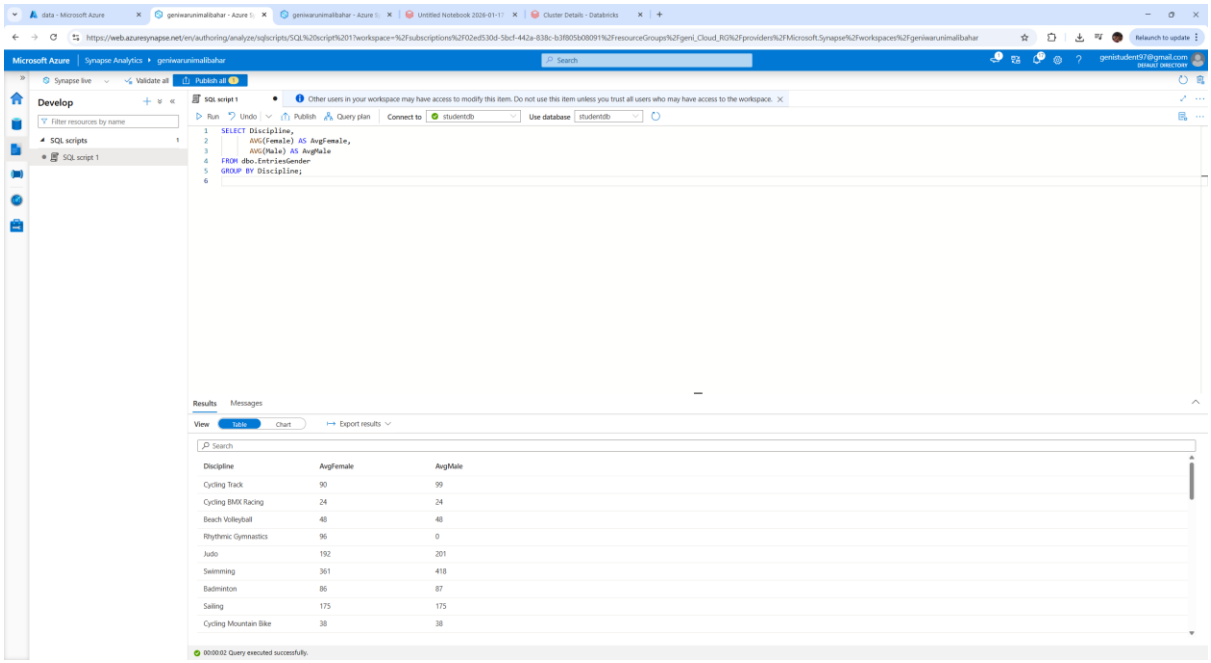
```
SELECT Country, COUNT(*) AS TotalAthletes
FROM Athletes
GROUP BY Country
ORDER BY TotalAthletes DESC;
```

Country	TotalAthletes
United States of America	615
Japan	586
Australia	470
People's Republic of China	401
Germany	400
France	377
Canada	368
Great Britain	366
Italy	356

- Total medals by country



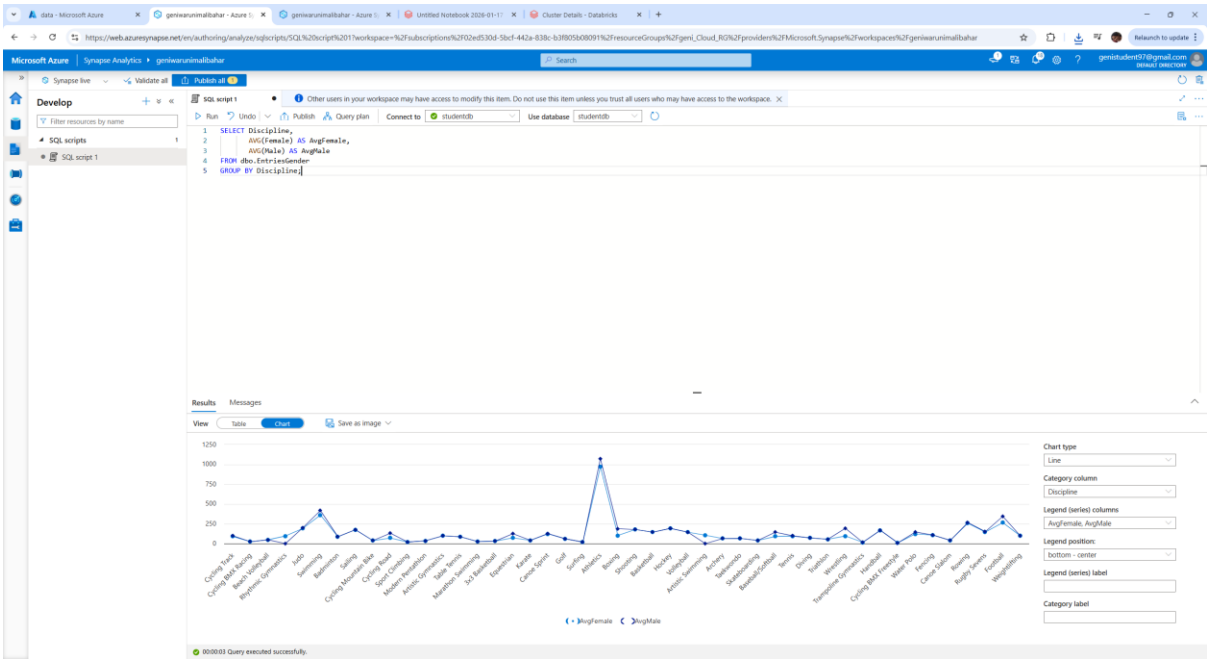
- Average gender participation by discipline



These queries provided insights into Olympic performance and participation trends.
Working with external tables in Azure Synapse helped me understand how large datasets can be analysed directly from the data lake without moving or duplicating data. The visualisations made it easier to identify trends, such as medal distribution across countries, which would be harder to see using raw query results alone.

Data Visualisation

Charts were created within Synapse Studio to visually represent the query results. These visualisations help interpret the data more effectively.


Microsoft Azure | Synapse Analytics | geniwarunimalibahar
Develop
SQL scripts
SQL script 1
SELECT Discipline,
AVG(Female) AS AvgFemale,
AVG(Male) AS AvgMale
FROM dbo.EntriesGender
GROUP BY Discipline;
Results
Messages
View
Table
Chart
Save as image
Chart type
Line
Category column
Discipline
Legend (series) columns
AvgFemale, AvgMale
Legend position:
bottom - center
Legend (series) label
Category label
AvgFemale
AvgMale
00:00:03 Query executed successfully.