Project – Serverless Data Processing

ETL Problem:

- ☑ ~~Use Azure Synapse Serverless SQL for raw data exploration directly on ADLS Gen2.~~

- ☑ ~~Trigger Azure Databricks jobs using Azure Functions for scalable transformation.~~

- ☑ ~~Code and infra managed via Azure DevOps repos and ARM templates.~~

ML Problem:

- ☑ ~~Create serverless training pipelines in Databricks with on-demand clusters.~~

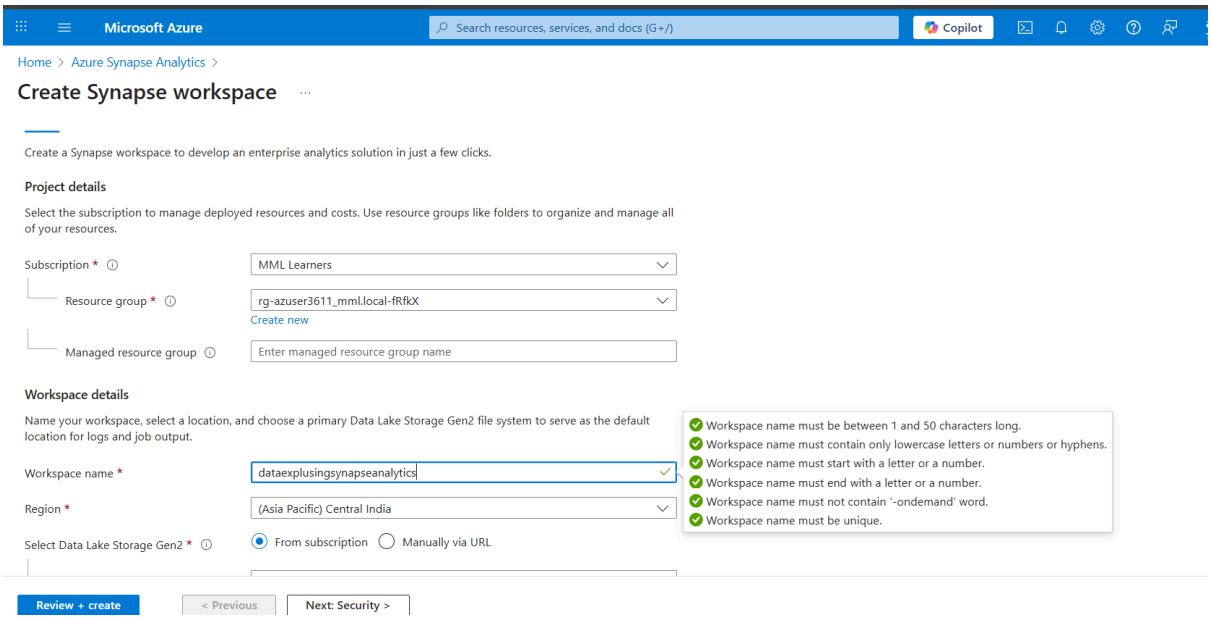
- ☑ ~~Train customer segmentation model (KMeans) using Spark MLlib.~~

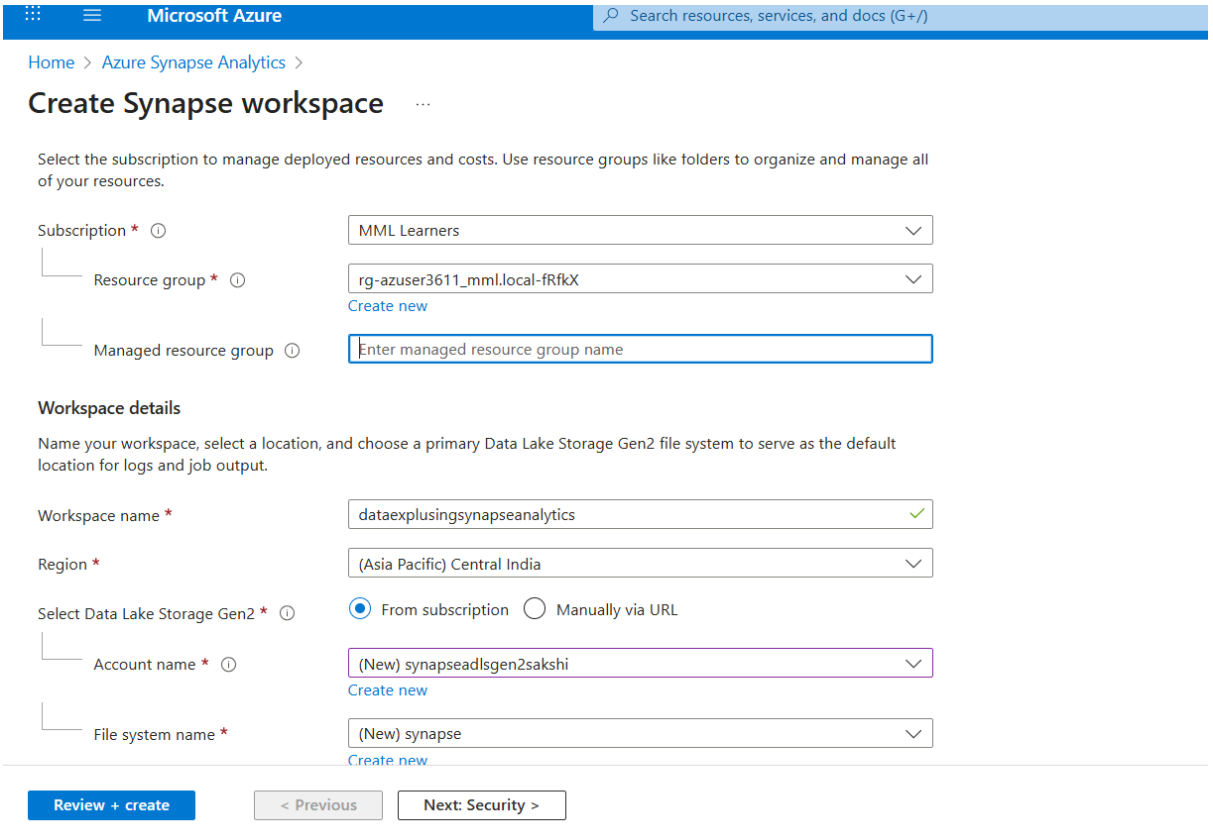
- ☐ **Deploy using CI/CD jobs .triggered via DevOps pipelines**
- ☑ ~~and monitored via MLflow.~~
- ☐ **PPT**
- ☐ **Architecture Diagram**
- ☐ **YT Video**
- ☐ **Report Refinement**
- ☐ **Docs Uploading to drive**
- ☐ **Explaining project to team members**

Architecture Diagram –

```
                    +-----------------------+
                    |   NYC Taxi Trip Data  |
                    | (CSV - ADLS Gen2)     |
                    +----------+------------+
                               |
                               v
                  +---------------------------+
                  | Azure Synapse Analytics   |
                  |  (Serverless SQL Pool)    |
                  | ↳ Data Exploration Layer  |
                  +---------------------------+
                               |
                               v
      ┌──────────────────────────────────────────┐
      │        Trigger Transformation & ML Job via DevOps        │
      └──────────────────────────────────────────┘
                               |
                               v
                    +-----------------------+
                    |    Azure DevOps       |
                    |   CI/CD Pipelines     |
                    |                       |
                    | - Repo (code, notebooks,|
                    |   ARM templates, YAML) |
                    | - Pipelines           |
                    |    ↳ Deploy Infra     |
                    |    ↳ Trigger Jobs     |
                    +-----------+-----------+
                                |
        +-----------------------+-----------------------+
        |                                               |
        v                                               v
+-----------------------+                   +-----------------------+
|   Azure Databricks    |                   |     Azure Function    |
| - Notebooks (PySpark) |                   |  (Optional HTTP trigger) |
| - Spark MLlib (KMeans) |                  |  ↳ Calls Databricks Job |
| - ETL + Training Logic |                  +-----------------------+
| - Runs on-demand      |
+----------+------------+
           |
           v
+---------------------------+
|    MLflow (Databricks)    |
|  - Log metrics, params    |
|  - Track KMeans model     |
|  - UI to compare runs     |
```

Task 1) **Data Exploration Using Azure Synapse Analytics**





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

Home > 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	MML Learners
Resource group	rg-azuser3611_mml.local-fRfkX

Create < Previous Next > Download a template for automation

Create

Microsoft Azure | Search resources, services, and docs (G+/) | Copilot | azuser3611_mml.local@... TECHADEMY LEARNING SOLUTI...

Home >

Microsoft.Azure.SynapseAnalytics-20250714093653 | Overview

Deployment

Search

Overview
Inputs
Outputs
Template

Delete Cancel Redeploy Download Refresh

Deployment is in progress

Deployment name : Microsoft.Azure.SynapseAnalytics-20250714093653
Subscription : MML Learners
Resource group : rg-azuser3611_mml.local-fRfkX
Start time : 7/14/2025, 9:50:53 AM
Correlation ID : 2ac7efb0-d530-4a06-8265-2e72f59a87f0

Deployment details

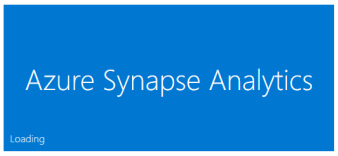
Resource	Type	Status	Operation details
synapseadlsgen2sakshi	Storage account	Accepted	Operation details

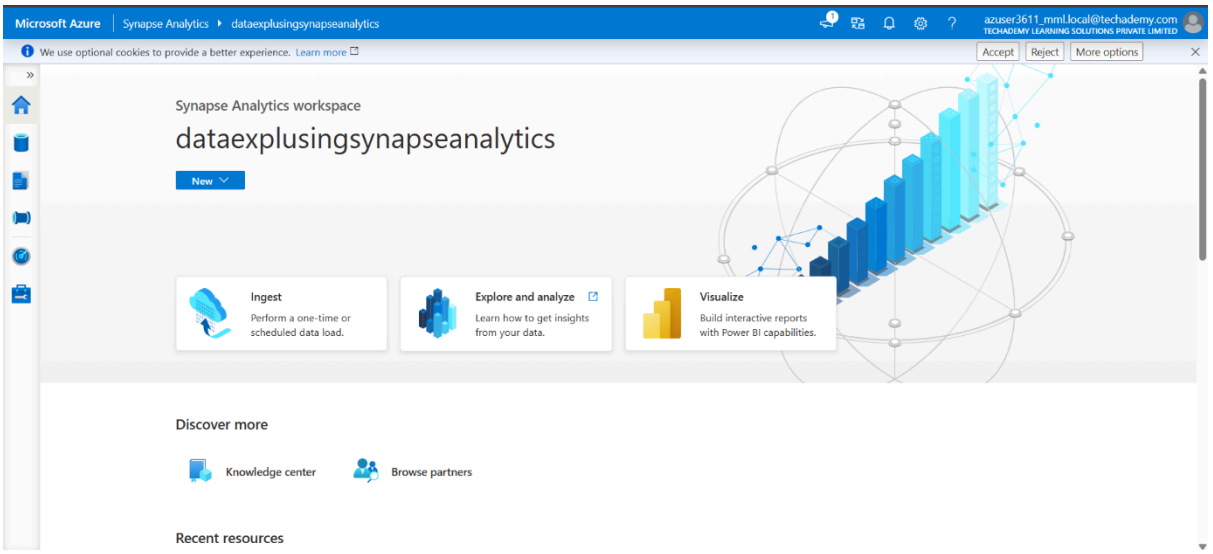
Give feedback
Tell us about your experience with deployment

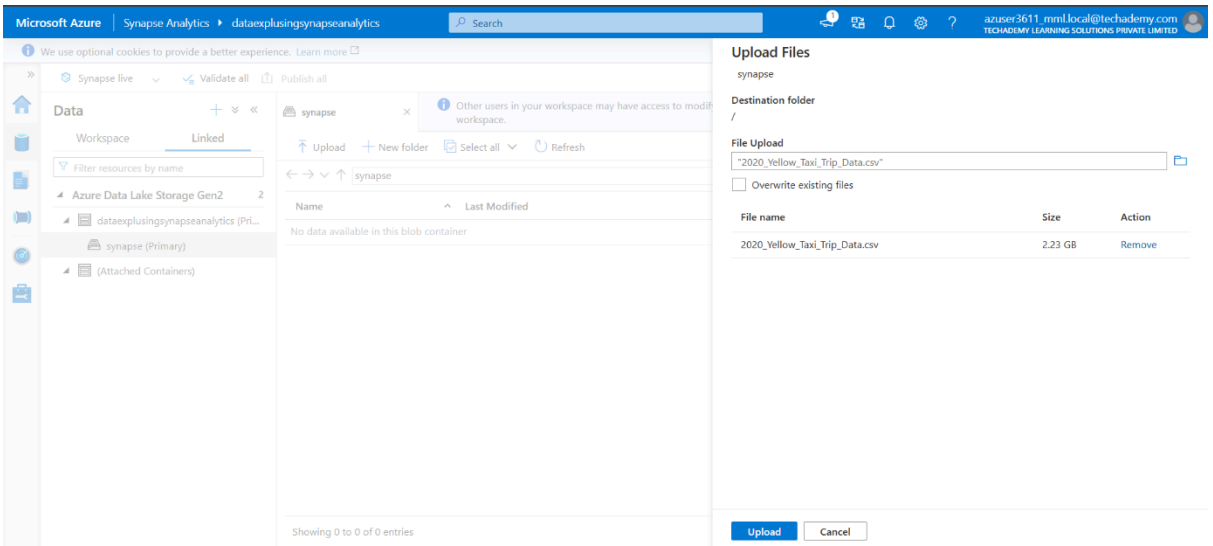
Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >

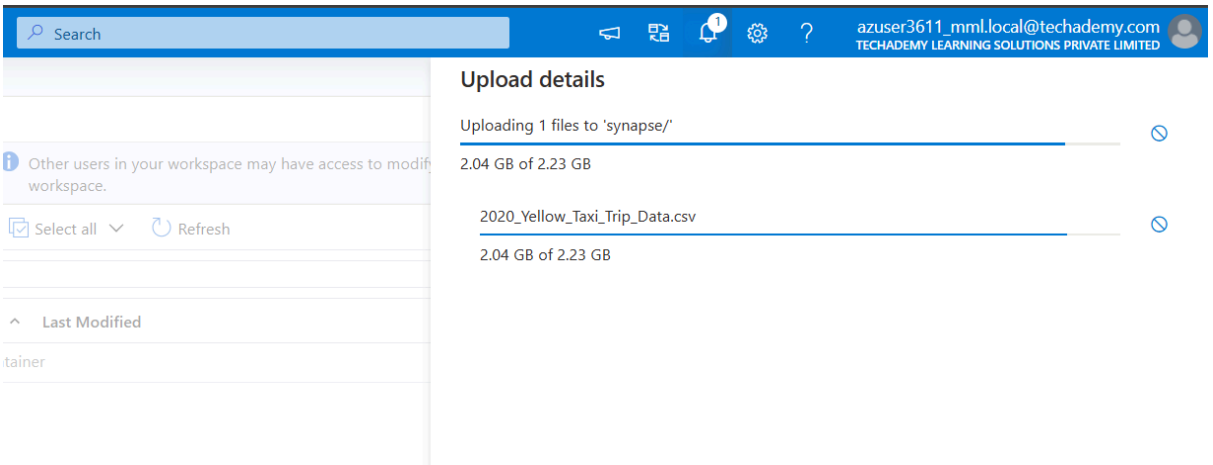
Free Microsoft tutorials
Start learning today >

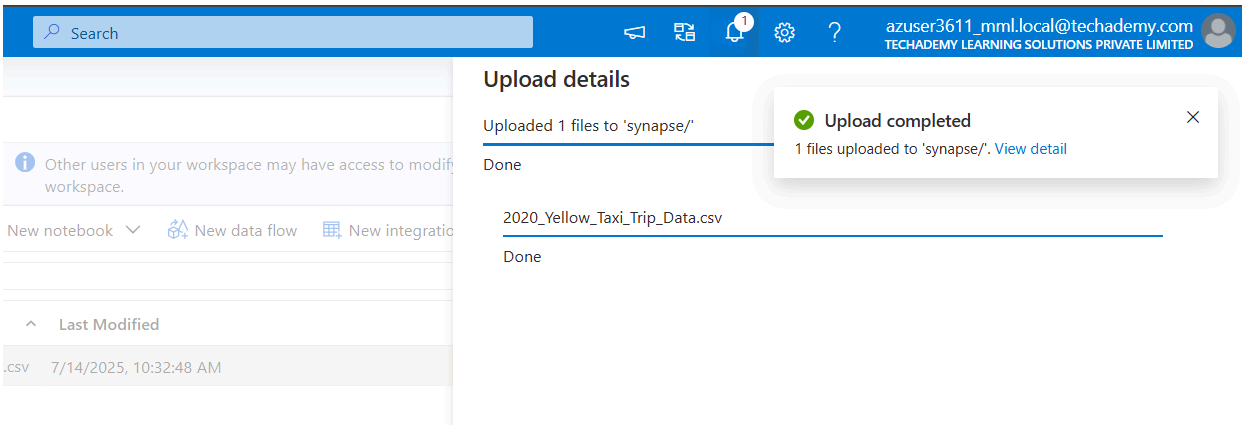
Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
Find an Azure expert >

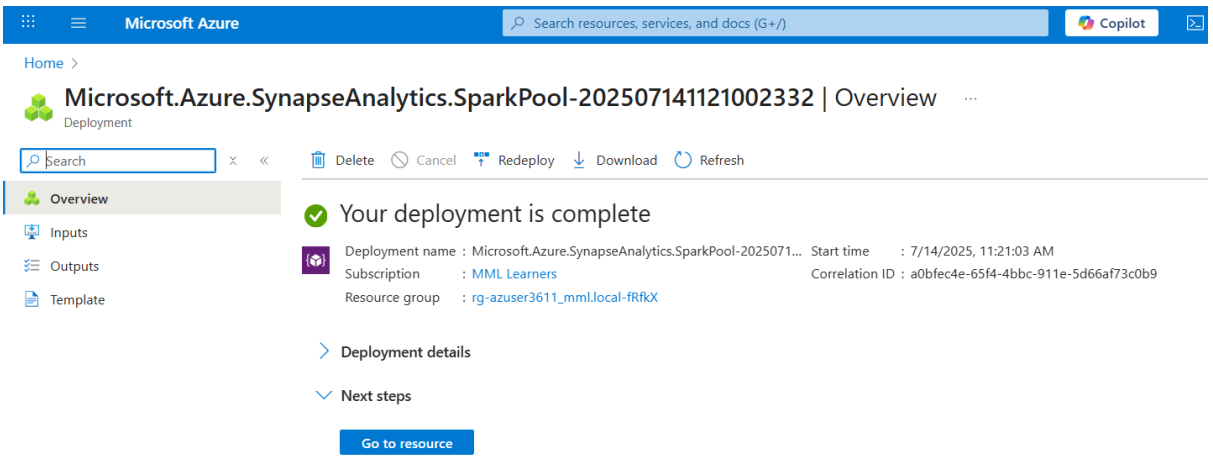

Azure Synapse Analytics
Loading

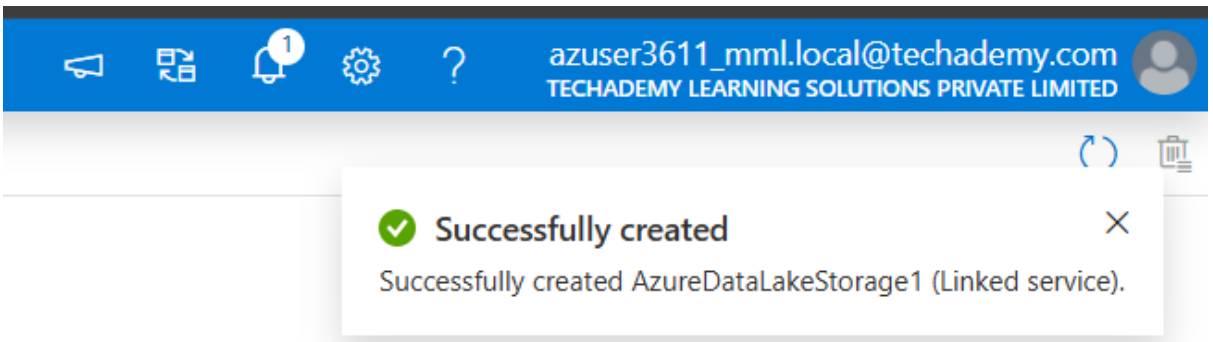

Microsoft Azure
Synapse Analytics
dataexplusingsynapseanalytics
azuser3611_mml.local@techademy.com
TECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse Analytics workspace
dataexplusingsynapseanalytics
New
Ingest
Perform a one-time or scheduled data load.
Explore and analyze
Learn how to get insights from your data.
Visualize
Build interactive reports with Power BI capabilities.
Discover more
Knowledge center
Browse partners
Recent resources

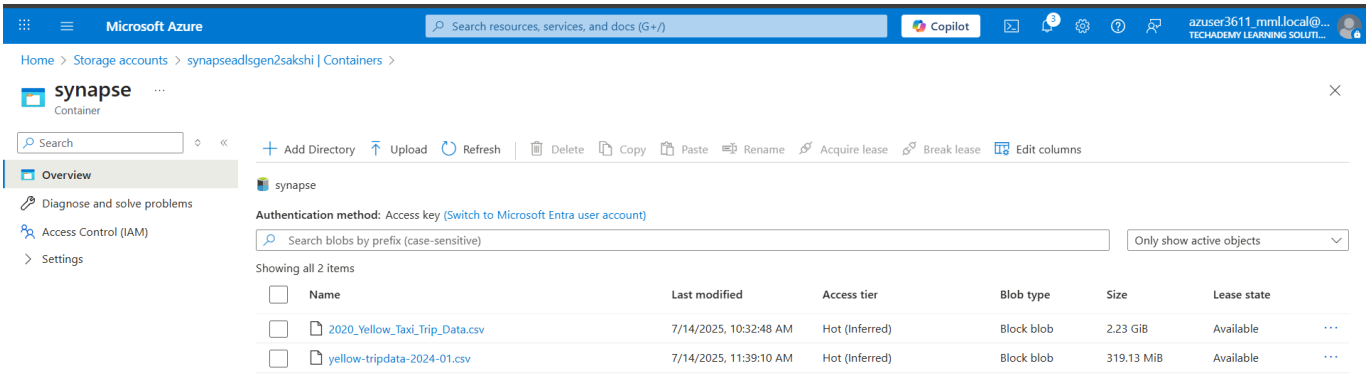

Microsoft Azure
Synapse Analytics
dataexplusingsynapseanalytics
Search
azuser3611_mml.local@techademy.com
TECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED
We use optional cookies to provide a better experience. Learn more
Synapse live
Validate all
Publish all
Data
Workspace
Linked
Filter resources by name
Azure Data Lake Storage Gen2
dataexplusingsynapseanalytics (Pri...
synapse (Primary)
(Attached Containers)
synapse
Other users in your workspace may have access to modif workspace.
Upload
New folder
Select all
Refresh
synapse
Name
Last Modified
No data available in this blob container
Showing 0 to 0 of 0 entries
Upload Files
synapse
Destination folder
/
File Upload
"2020_Yellow_Taxi_Trip_Data.csv"
Overwrite existing files
File name
Size
Action
2020_Yellow_Taxi_Trip_Data.csv
2.23 GB
Remove
Upload
Cancel


Search
azuser3611_mml.local@techademy.com
TECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED
Upload details
Uploading 1 files to 'synapse/'
2.04 GB of 2.23 GB
2020_Yellow_Taxi_Trip_Data.csv
2.04 GB of 2.23 GB
Other users in your workspace may have access to modif
workspace.
Select all
Refresh
Last Modified

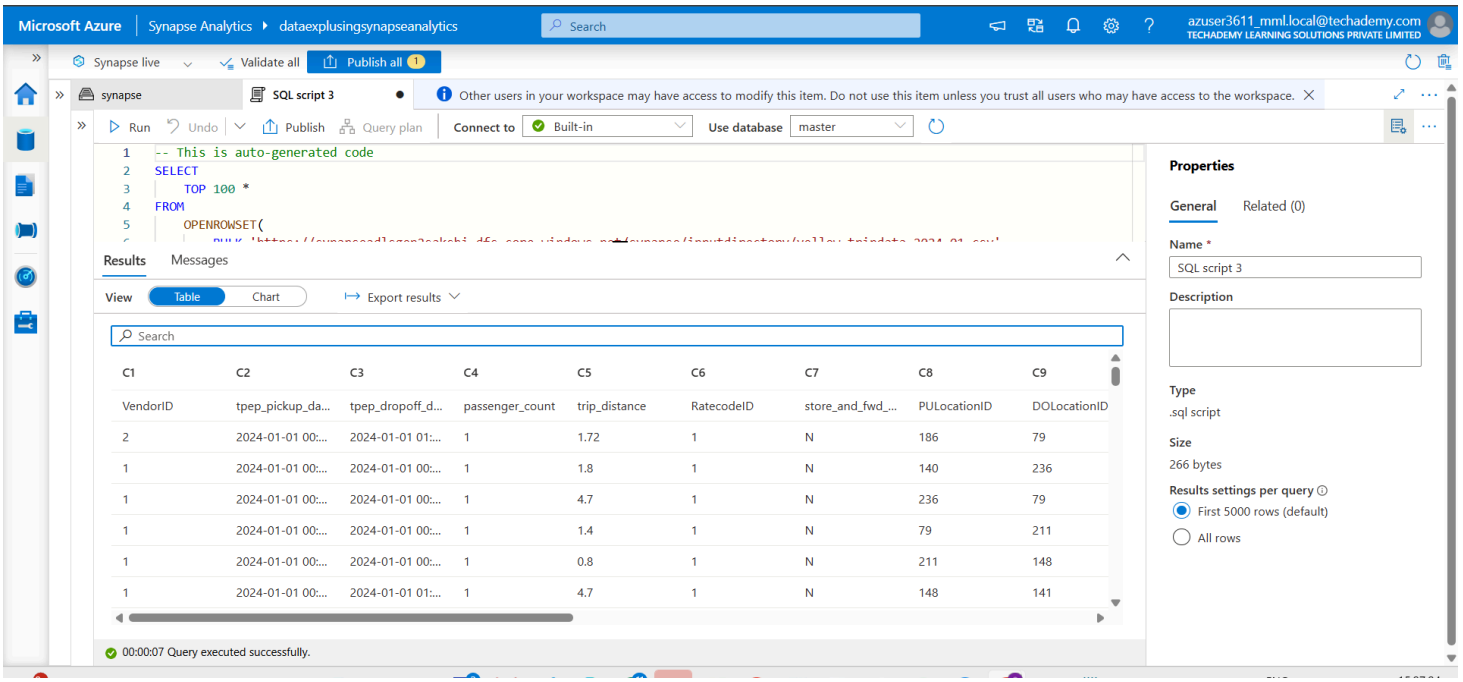

Search
azuser3611_mml.local@techademy.com
TECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED
Upload details
Uploaded 1 files to 'synapse/'
Done
Upload completed
1 files uploaded to 'synapse/'. View detail
2020_Yellow_Taxi_Trip_Data.csv
Done
Other users in your workspace may have access to modif
workspace.
New notebook
New data flow
Last Modified
.csv
7/14/2025, 10:32:48 AM


Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Microsoft.Azure.SynapseAnalytics.SparkPool-20250714112100233 2 | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name : Microsoft.Azure.SynapseAnalytics.SparkPool-2025071...
Subscription : MML Learners
Resource group : rg-azuser3611_mml.local-fRfkX
Start time : 7/14/2025, 11:21:03 AM
Correlation ID : a0bfec4e-65f4-4bbc-911e-5d66af73c0b9
Deployment details
Next steps
Go to resource

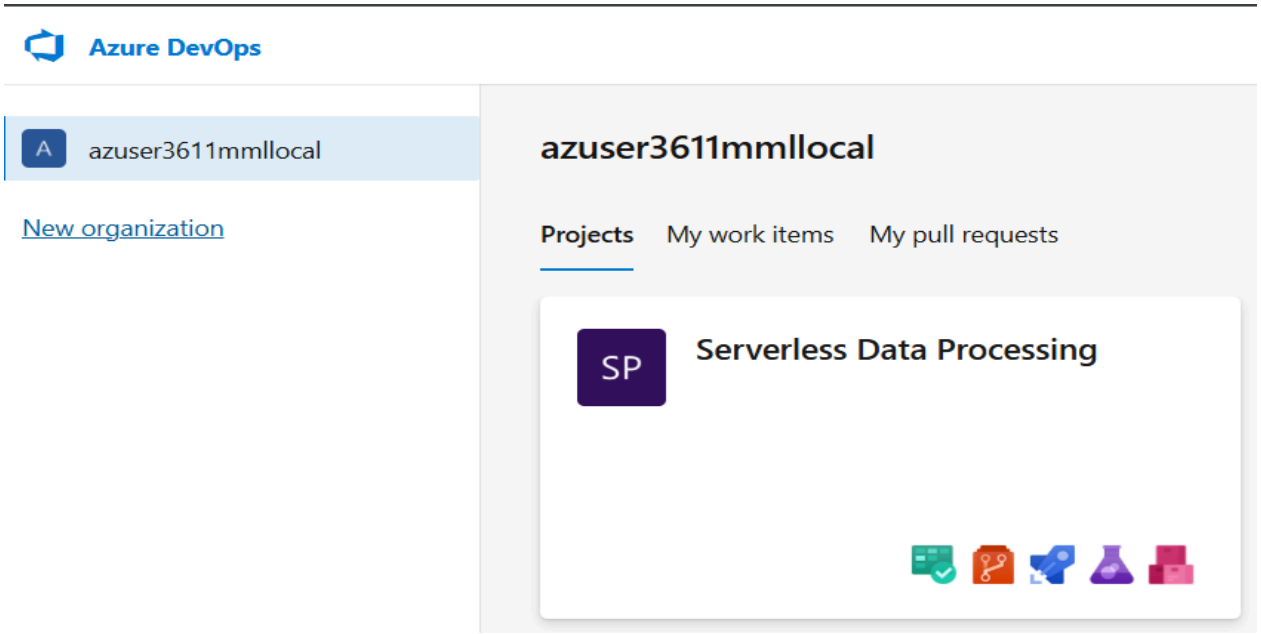


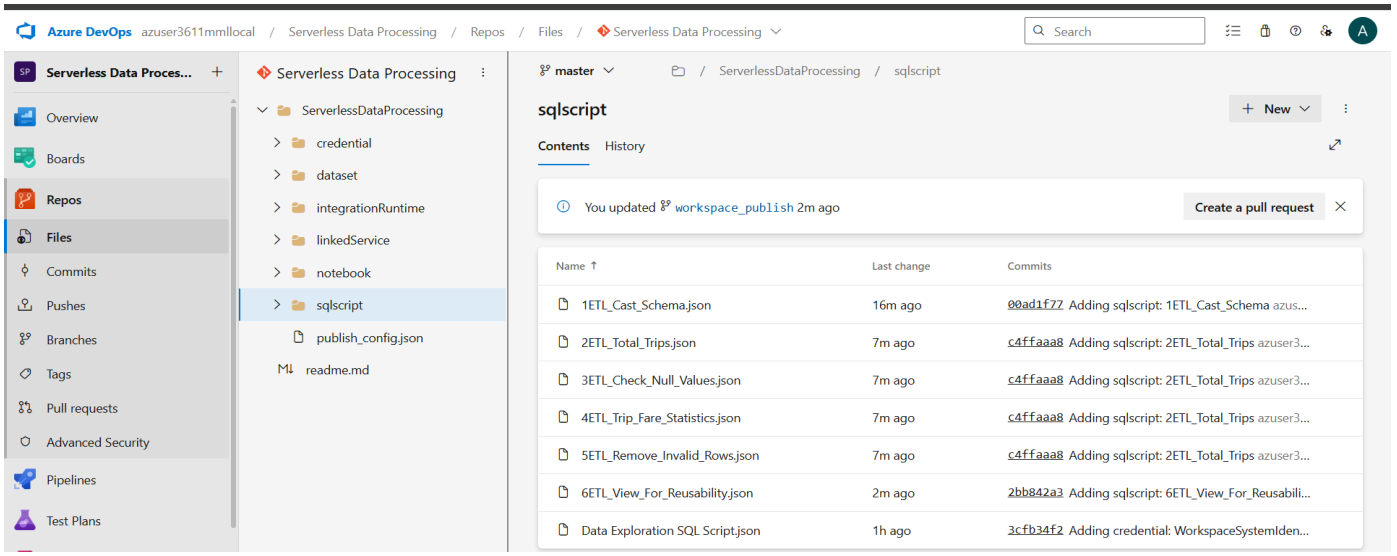


- **Fetching the first 100 rows using Synapse Serverless SQL from ADLS Gen2 dataset & ADLS Gen2 CSV file is accessible from Synapse Serverless**

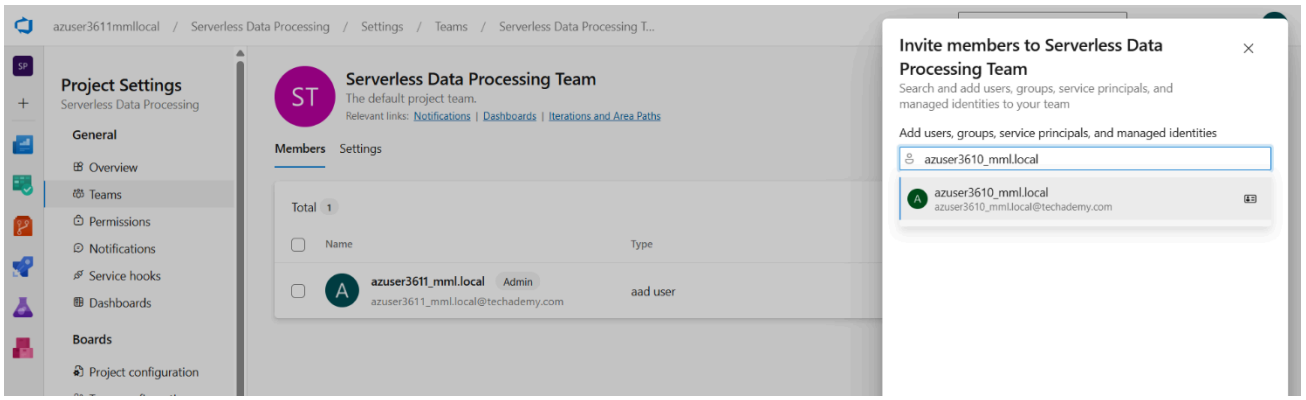


Task 2) Git Configuration - Azure DevOps & Synapse Analytics





- **Git Collaboration –**



Task 3) Azure Databricks - Machine Learning

K-Means is a machine learning algorithm used to find groups (called clusters) in data.
Each cluster groups similar rows (data points) together.

Example:

In NYC Taxi data, we might want to group trips like:

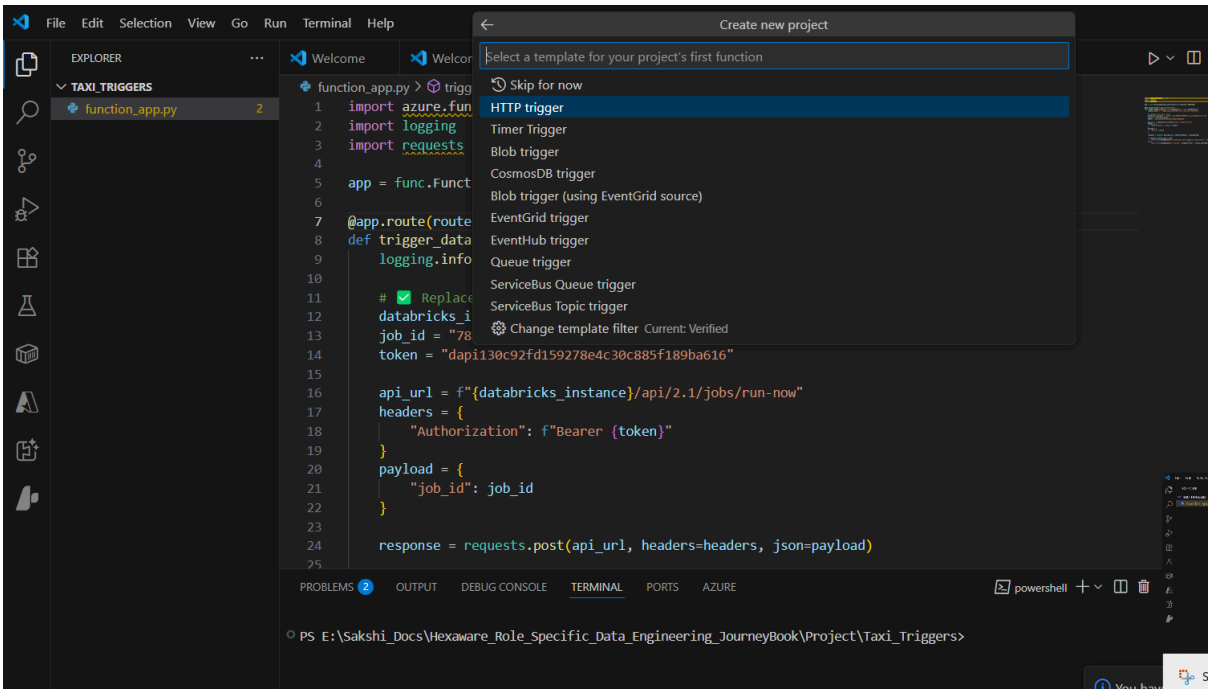
- Short trips with low fare
- Long trips with high fare
- Short trips with big tips
  These are called clusters.

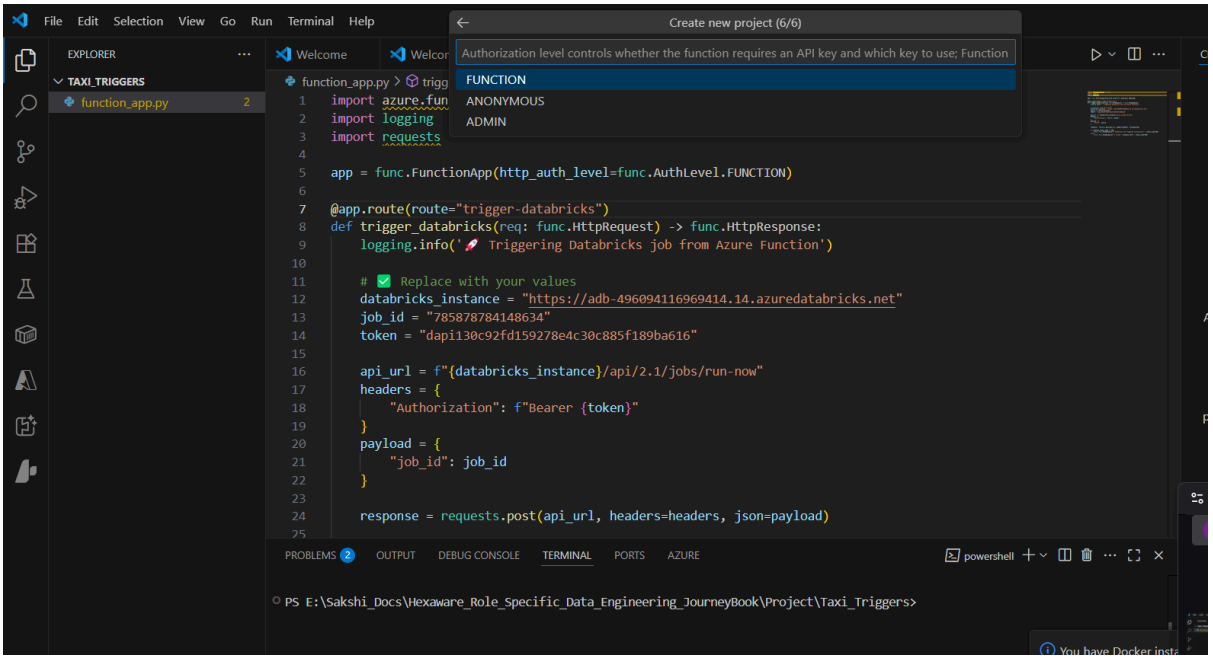
Column	Meaning
`trip_distance`	**Distance of the taxi ride (in miles)**
`fare_amount`	**Fare for the ride (in $)**
`passenger_count`	**No. of passengers during the trip**
`prediction`	**Which cluster the trip belongs to (0–3)**

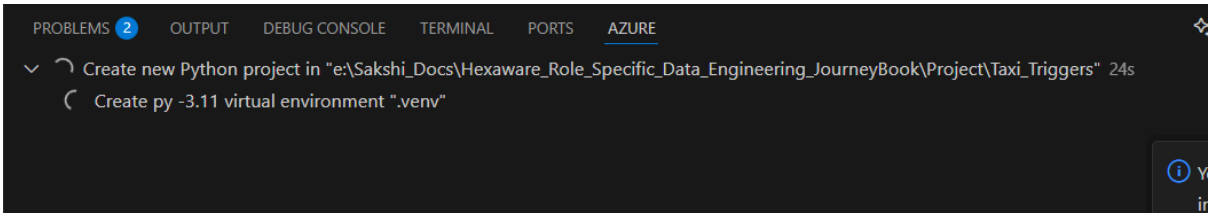
Interpreting The Results

Cluster	Approx % of Total	Likely Pattern
0	~82%	Regular trips — maybe 1–2 miles, 1 passenger, normal fare. Most common.
1	~10%	Possibly longer trips or group rides.
3	~7%	Could be airport trips or higher fare rides.

Task 4) Triggers -







File Edit Selection View Go Run Terminal Help

Taxi_Triggers

EXPLORER

TAXI_TRIGGERS
- .venv
- .vscode
- .funcignore U
- .gitignore U
- function_app.py 1, U
- host.json U
- local.settings.json
- requirements.txt U

Welcome | Welcome | function_app.py 1, U | Preview http_trigger_template.md

function_app.py > http_trigger

```python
import azure.functions as func
import logging

app = func.FunctionApp(http_auth_level=func.Au

@app.route(route="http_trigger")
def http_trigger(req: func.HttpRequest) -> fun
    logging.info('Python HTTP trigger function

    name = req.params.get('name')
    if not name:
        try:
            req_body = req.get_json()
        except ValueError:
            pass
        else:
            name = req_body.get('name')
```

Azure Functions: HTTP Trigger in Python

HTTP Trigger

The HTTP trigger lets you invoke a function with an HTTP request. You can use an HTTP trigger to build serverless APIs and respond to webhooks.

Using the Template

Following is an example code snippet for HTTP Trigger

PROBLEMS 1 | OUTPUT | DEBUG CONSOLE | TERMINAL | PORTS | AZURE

- Create new Python project in "e:\Sakshi_Docs\Hexaware_Role_Specific_Data_Engineering_JourneyBook\Project\Taxi_Triggers" Succeeded in 52s
  - Create py -3.11 virtual environment ".venv" 45s
  - Initialize Python project for VS Code 4s
  - Create New Project with HttpTrigger function 0s

File Edit Selection View Go Run Terminal Help

Taxi_Triggers

EXPLORER

TAXI_TRIGGERS
- .venv
- .vscode
- .funcignore U
- .gitignore U
- function_app.py 2, U
- host.json U
- local.settings.json
- requirements.txt U

Welcome | Welcome | Welcome | Welcome | function_app.py 2, U

function_app.py > ...

```python
import azure.functions as func
import logging
import requests

app = func.FunctionApp(http_auth_level=func.AuthLevel.FUNCTION)

@app.route(route="trigger-databricks")
def trigger_databricks(req: func.HttpRequest) -> func.HttpResponse:
    logging.info('🚀 Triggering Databricks job from Azure Function')

    # ✅ Replace with your values
    databricks_instance = "https://adb-1572507201698816.16.azuredatabricks.net"
    job_id = "902608976982792"
    token = "dapi330cd5e67647dea06b3156f7d928f3be"

    api_url = f"{databricks_instance}/api/2.1/jobs/run-now"
    headers = {
        "Authorization": f"Bearer {token}"
    }
    payload = {
```

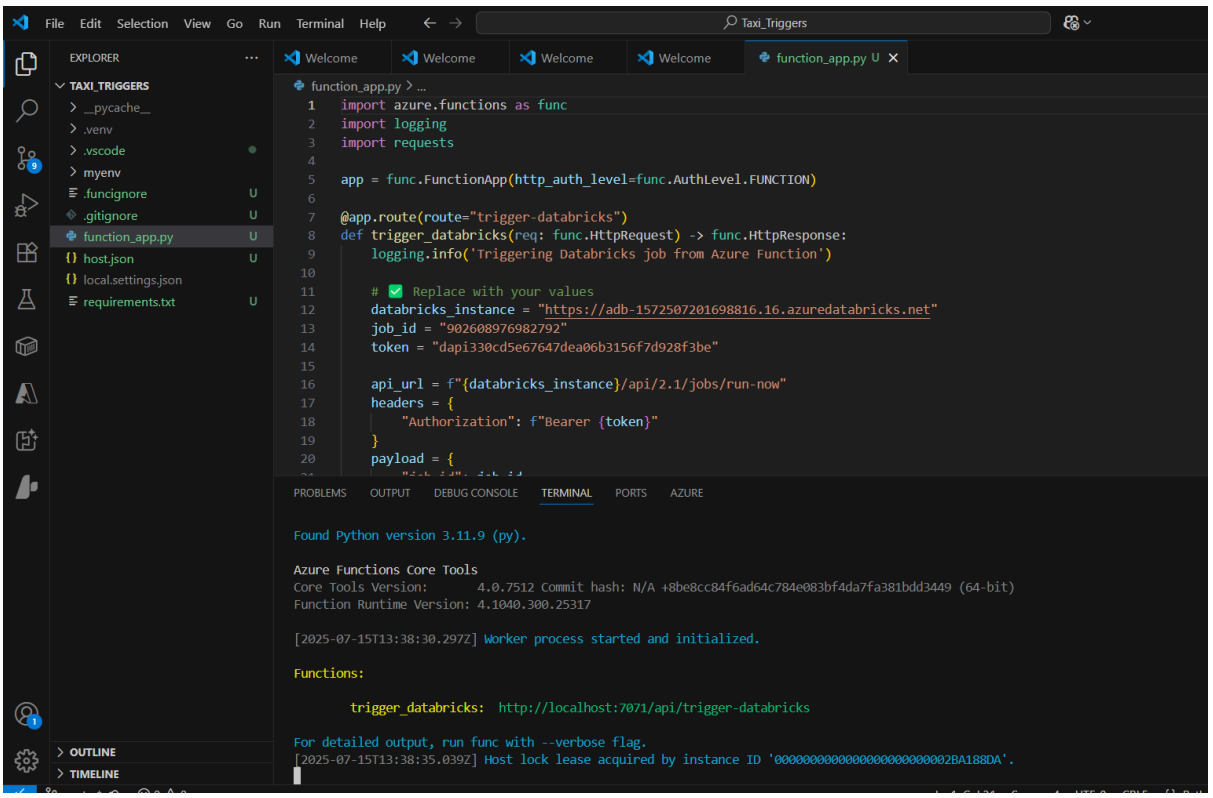
PROBLEMS 2 | OUTPUT | DEBUG CONSOLE | TERMINAL | PORTS | AZURE

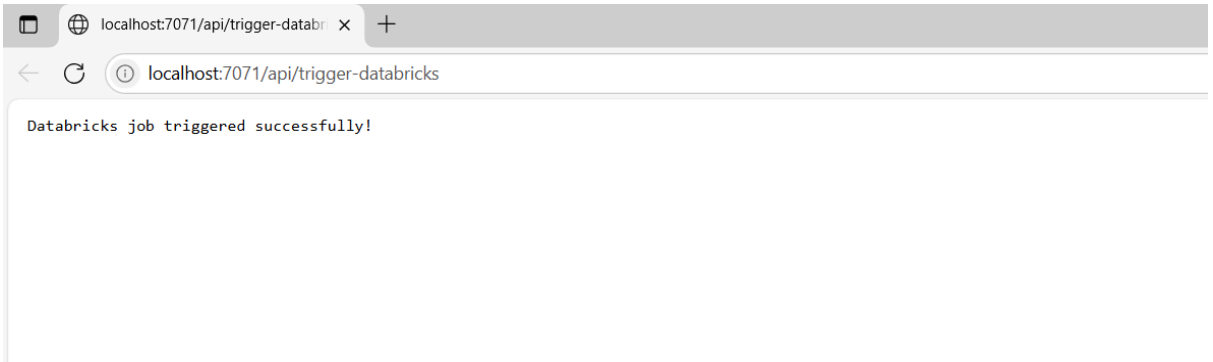
```
PS E:\Sakshi_Docs\Hexaware_Role_Specific_Data_Engineering_JourneyBook\Project\Taxi_Triggers> pip install azure-functions requests
Collecting azure-functions
  Downloading azure_functions-1.23.0-py3-none-any.whl.metadata (7.3 kB)
Collecting requests
  Downloading requests-2.32.4-py3-none-any.whl.metadata (4.9 kB)
Collecting werkzeug~=3.1.3 (from azure-functions)
  Downloading werkzeug-3.1.3-py3-none-any.whl.metadata (3.7 kB)
Collecting MarkupSafe>=2.1.1 (from werkzeug~=3.1.3->azure-functions)
  Downloading MarkupSafe-3.0.2-cp313-cp313-win_amd64.whl.metadata (4.1 kB)
Collecting charset_normalizer<4,>=2 (from requests)
  Downloading charset_normalizer-3.4.2-cp313-cp313-win_amd64.whl.metadata (36 kB)
Collecting idna<4,>=2.5 (from requests)
  Downloading idna-3.10-py3-none-any.whl.metadata (10 kB)
Collecting urllib3<3,>=1.21.1 (from re
  Downloading urllib3-2.5.0-py3-none-a
```

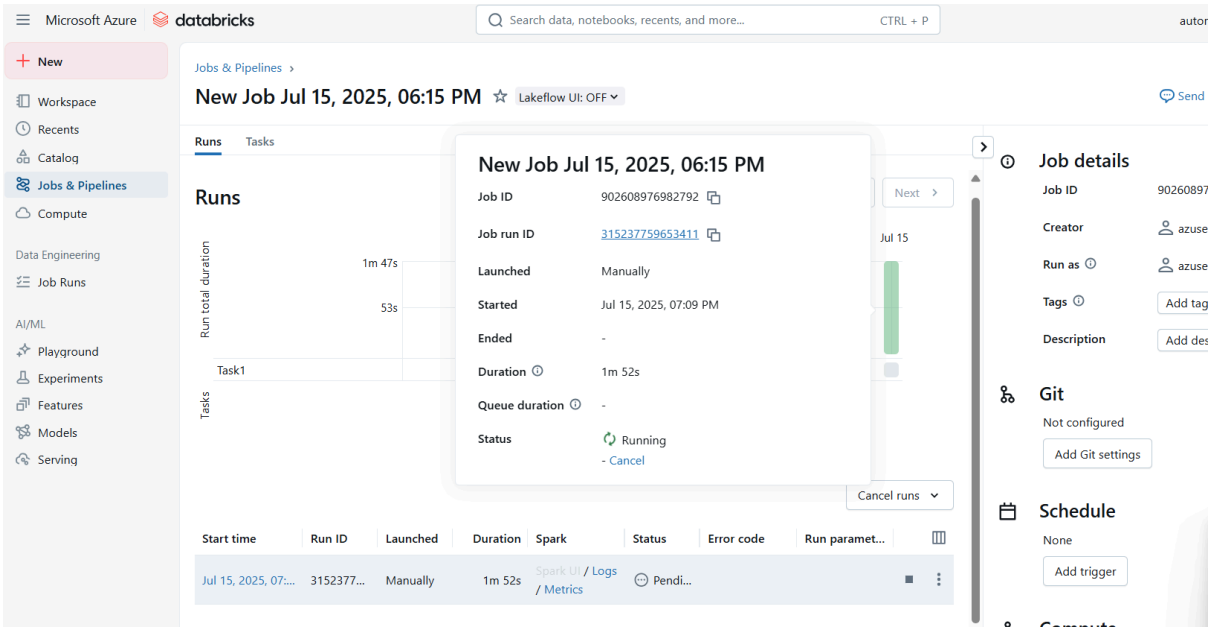
OUTLINE

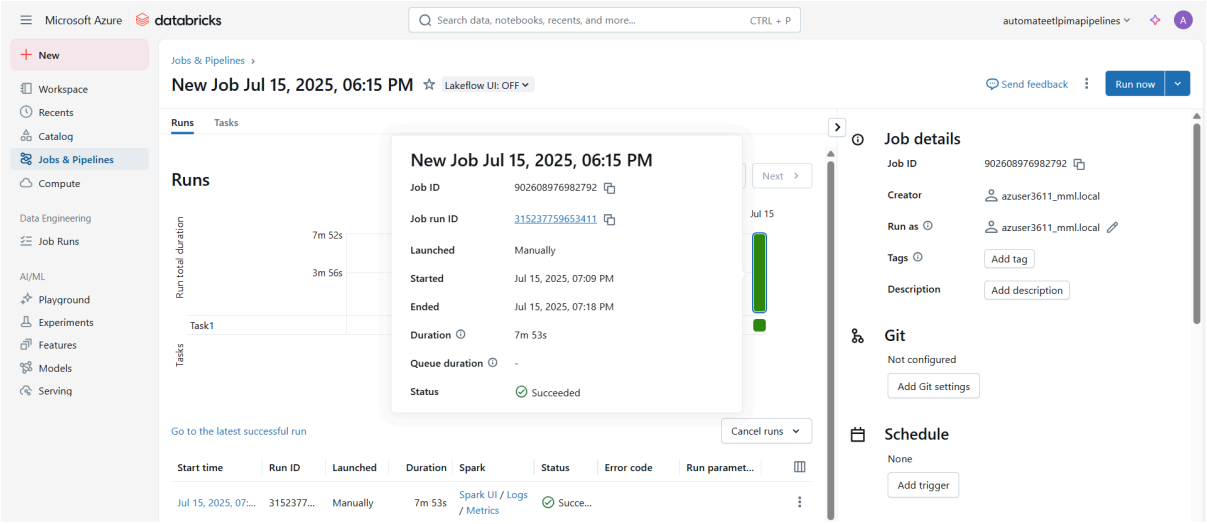
TIMELINE

meet.google.com is sharing your screen. Stop sharing Hide


File Edit Selection View Go Run Terminal Help
Taxi_Triggers
EXPLORER
TAXI_TRIGGERS
__pycache__
.venv
.vscode
myenv
.funcignore
.gitignore
function_app.py
host.json
local.settings.json
requirements.txt
OUTLINE
TIMELINE
Welcome
function_app.py U
import azure.functions as func
import logging
import requests
app = func.FunctionApp(http_auth_level=func.AuthLevel.FUNCTION)
@app.route(route="trigger-databricks")
def trigger_databricks(req: func.HttpRequest) -> func.HttpResponse:
logging.info('Triggering Databricks job from Azure Function')
# Replace with your values
databricks_instance = "https://adb-1572507201698816.16.azuredatabricks.net"
job_id = "902608976982792"
token = "dapi330cd5e67647dea06b3156f7d928f3be"
api_url = f"{databricks_instance}/api/2.1/jobs/run-now"
headers = {
"Authorization": f"Bearer {token}"
}
payload = {
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS AZURE
Found Python version 3.11.9 (py).
Azure Functions Core Tools
Core Tools Version: 4.0.7512 Commit hash: N/A +8be8cc84f6ad64c784e083bf4da7fa381bdd3449 (64-bit)
Function Runtime Version: 4.1040.300.25317
[2025-07-15T13:38:30.297Z] Worker process started and initialized.
Functions:
trigger_databricks: http://localhost:7071/api/trigger-databricks
For detailed output, run func with --verbose flag.
[2025-07-15T13:38:35.039Z] Host lock lease acquired by instance ID '000000000000000000000000002BA188DA'.


localhost:7071/api/trigger-databr
localhost:7071/api/trigger-databricks
Databricks job triggered successfully!

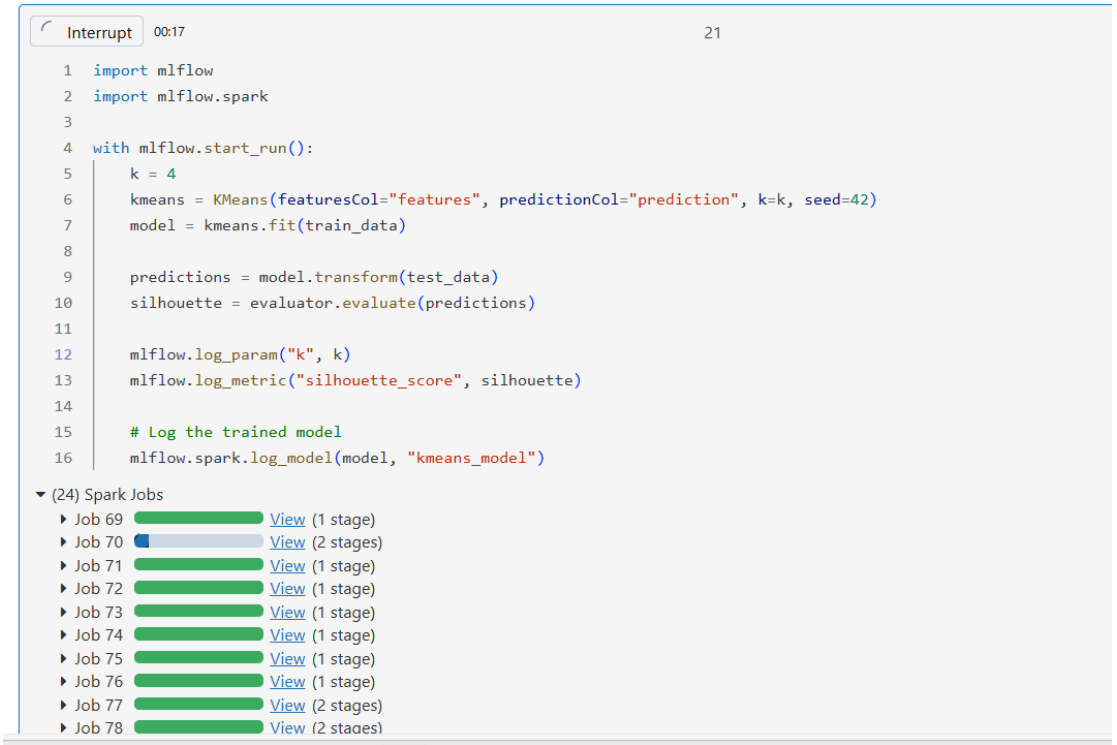

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Data Engineering
Job Runs
AI/ML
Playground
Experiments
Features
Models
Serving
Jobs & Pipelines
New Job Jul 15, 2025, 06:15 PM
Lakeflow UI: OFF
Send
Runs
Tasks
Runs
Run total duration
1m 47s
53s
Task1
Tasks
Jul 15
New Job Jul 15, 2025, 06:15 PM
Job ID 902608976982792
Job run ID 315237759653411
Launched Manually
Started Jul 15, 2025, 07:09 PM
Ended -
Duration 1m 52s
Queue duration -
Status Running
- Cancel
Cancel runs
Start time
Run ID
Launched
Duration
Spark
Status
Error code
Run paramet...
Jul 15, 2025, 07:...
3152377...
Manually
1m 52s
Spark UI / Logs / Metrics
Pendi...
Job details
Job ID
Creator
azuse
Run as
azuse
Tags
Add tag
Description
Add des
Git
Not configured
Add Git settings
Schedule
None
Add trigger



"Serverless" here doesn't mean *no servers at all*, it means:

- **You don't keep a server running all the time**
- **You only pay for what you use**
- **Resources spin up automatically when needed and shut down afterward**

Task 5) MLFlow -



Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P

New

Workspace
Recents
Catalog
Jobs & Pipelines
Compute

Data Engineering
Job Runs

AI/ML
Playground

Experiments ›

NYC_Taxi_KMeans_Training

Machine learning

Runs

MLflow 3 is now available! Try out the new features and provide feedback. Learn more

metrics.silhouette_score

Columns

Group by

Search runs using a simplified version of the SQL **WHERE** clause. Learn more

Examples:

- metrics.rmse >= 0.8
- metrics.`f1 score` < 1
- params.model = 'tree'
- attributes.run_name = 'my run'
- tags.`mlflow.user` = 'myUser'
- metric.f1_score > 0.9 AND params.model = 'tree'
- dataset.name IN ('dataset1', 'dataset2')
- attributes.run_id IN ('a1b2c3d4', 'e5f6g7h8')
- tags.model_class LIKE 'sklearn.linear_model%'
- MIN(metrics.loss) < 1
- MAX(metrics.gpu_utilization_percentage) >= 0.5

Run Name	Created	Dataset
wistful-dove-811	4 minutes ago	-

Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P

automateetlpimapipelines

New

Workspace
Recents
Catalog
Jobs & Pipelines
Compute

Data Engineering
Job Runs

AI/ML
Playground
Experiments
Features
Models
Serving

Registered Models ›

mlflow

Permissions

Use model for inference

Details

Notify me about All new activity

Created Time: Jul 15, 2025, 11:55 PM

Last Modified: Jul 15, 2025, 11:55 PM

Creator: azuser3611_mml.local@techademy.com

Description Edit

Tags

Versions All Active 0 Compare

Version	Registered at	Created by	Stage	Pending Requests	Description	Endpoints
Version 1	Jul 15, 2025, 11:55 PM	azuser3611_mml.loc...	None	-		-

Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P

New

Workspace
Recents
Catalog
Jobs & Pipelines
Compute

Data Engineering
Job Runs

AI/ML
Playground

Serving endpoints ›

Create serving endpoint

Experimenting with LLMs? Try pay-per-token Foundation Model APIs!

Model serving is only available in Premium or Enterprise workspaces. Please contact your organization admin or Databricks support.

Add served entity

Route optimization

Enable route optimization

Recommended for high throughput or low latency workloads.

Summary

Served entities

mlflow

ARM template deployment

Azure Details

Deployment scope *

Resource Group

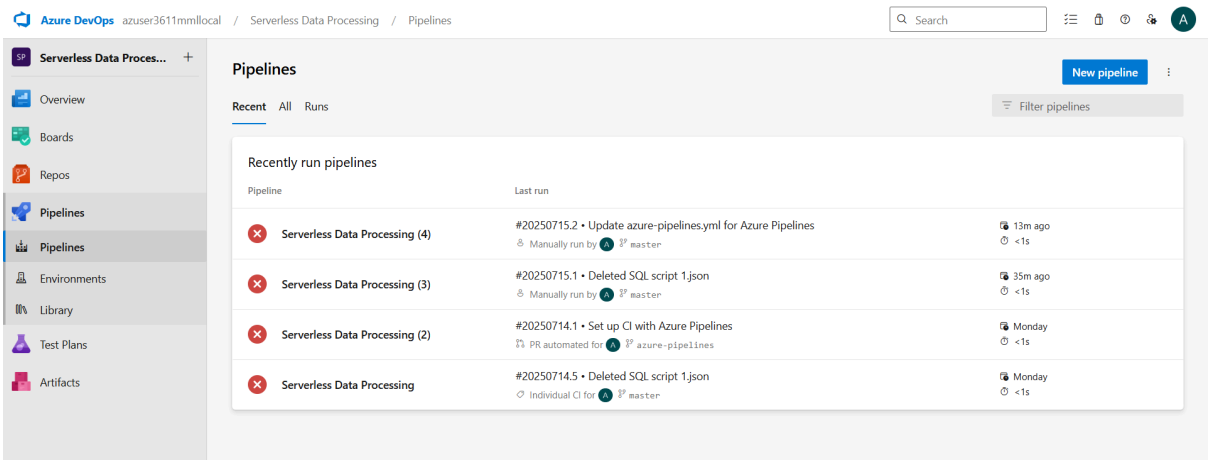
Azure Resource Manager connection *

MML Learners(2a3c6418-97b9-4d96-a2...

Failed to set Azure permission 'RoleAssignmentId: 7db1cebc-7a85-45bd-a416-726e5a8d369b' for the service principal '26222bfd-76c4-4552-b044-e6c4cb900df0' on subscription ID '2a3c6418-97b9-4d96-a24b-2c2d7633d375': error code: Forbidden, inner error code: AuthorizationFailed, inner error message The client 'azuser3611_mml.local@techademy.com' with object id '81cfa09f-100c-42af-b3f5-382077f858a8' does not have authorization to perform action 'Microsoft.Authorization/roleAssignments/write' over scope '/subscriptions/2a3c6418-97b9-4d96-a24b-2c2d7633d375/providers/Microsoft.Authorization/roleAssignme 7a85-45bd-a416-726e5a8d369b' or the scope is invalid. If access was recently granted, please refresh your credentials. Ensure that the user has 'Owner' or 'User Access Administrator' permissions on the Subscription.

About this task

Add


Azure DevOps
azuser3611mmllocal
Serverless Data Processing
Pipelines
Search
Serverless Data Proces...
Overview
Boards
Repos
Pipelines
Pipelines
Environments
Library
Test Plans
Artifacts
Pipelines
New pipeline
Recent
All
Runs
Filter pipelines
Recently run pipelines
Pipeline
Last run
Serverless Data Processing (4)
#20250715.2 • Update azure-pipelines.yml for Azure Pipelines
Manually run by
master
13m ago
<1s
Serverless Data Processing (3)
#20250715.1 • Deleted SQL script 1.json
Manually run by
master
35m ago
<1s
Serverless Data Processing (2)
#20250714.1 • Set up CI with Azure Pipelines
PR automated for
azure-pipelines
Monday
<1s
Serverless Data Processing
#20250714.5 • Deleted SQL script 1.json
Individual CI for
master
Monday
<1s