Final Project 1 Solution

1. Finding and Analyzing Data:

- Locate an open-source dataset.
- Write technical and non-technical observations about the dataset (schema, data quality, potential improvements).

The dataset I have picked for this project is from UCI Machine Learning Repository

Home - UCI Machine Learning Repository with the subject area “Climate and Environment”. Here is the link to find the data set Forest Fires - UCI Machine Learning Repository,

Chase the Fire (kaggle.com)

This predicts the burned area of forest fires in Portugal, it does not have any missing values

About the dataset and Observations:

Technical Observations:

- **Schema**: The dataset consists of 13 columns. The columns are "X", "Y", "month", "day", "FFMC", "DMC", "DC", "ISI", "temp", "RH", "wind", "rain", and "area".
- **Data Types**: The dataset contains a mix of numerical and categorical data. "X", "Y", "FFMC", "DMC", "DC", "ISI", "temp", "RH", "wind", "rain", and "area" are numerical, while "month" and "day" are categorical.
- **Missing Values:** From the given data, it appears there are no missing values. However, a more thorough check would be needed on the full dataset.
- **Data Quality:** There are no obvious inconsistencies in the data. Outliers would need to be checked for in the numerical data.

Non-Technical Observations:

- **Purpose of Data/insights**: The data seems to be related to forest fires and includes various weather and fire danger index conditions. It could be used to understand the relationship between these conditions and the occurrence or severity of forest fires.
- **Limitations/Biases**: The data seems to be limited to specific locations (as indicated by the X and Y coordinates). If these locations are not representative of the area as a whole, this could introduce bias. Also, many fires seem to have an area of 0, which could indicate an imbalance in the data.
- **Temperature**: The ‘temp’ column seems to vary quite a bit, indicating a range of weather conditions.
- **Rain**: The ‘rain’ column has a lot of zero values, suggesting that rain is not a common occurrence in these observations.

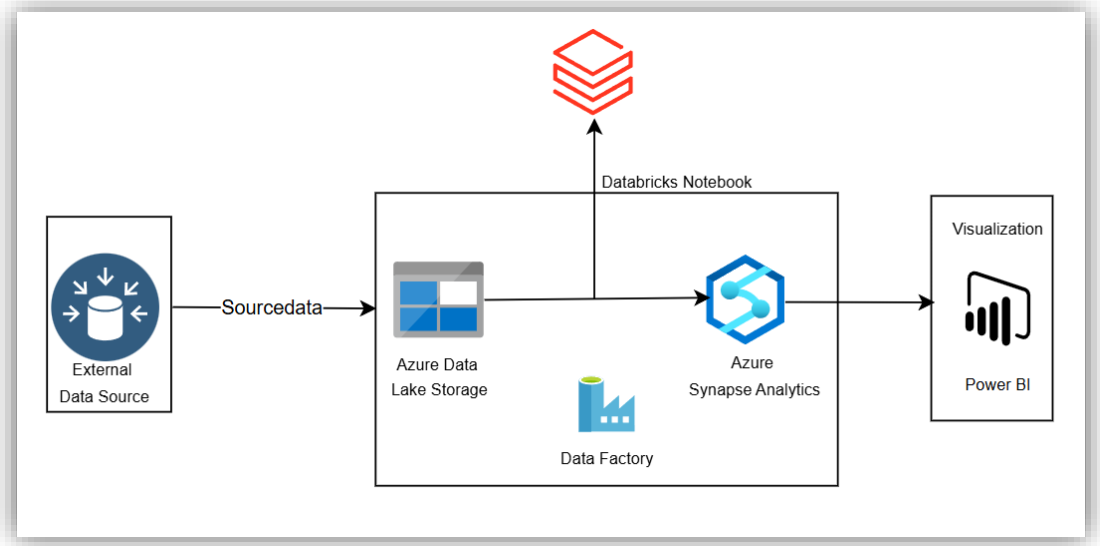
- **Area**: The ‘area’ column also has many zero values. This could indicate areas where no forest fire was observed.
- **Month and Day**: The dataset includes observations from different months and days, indicating it spans across different seasons and days of the week.

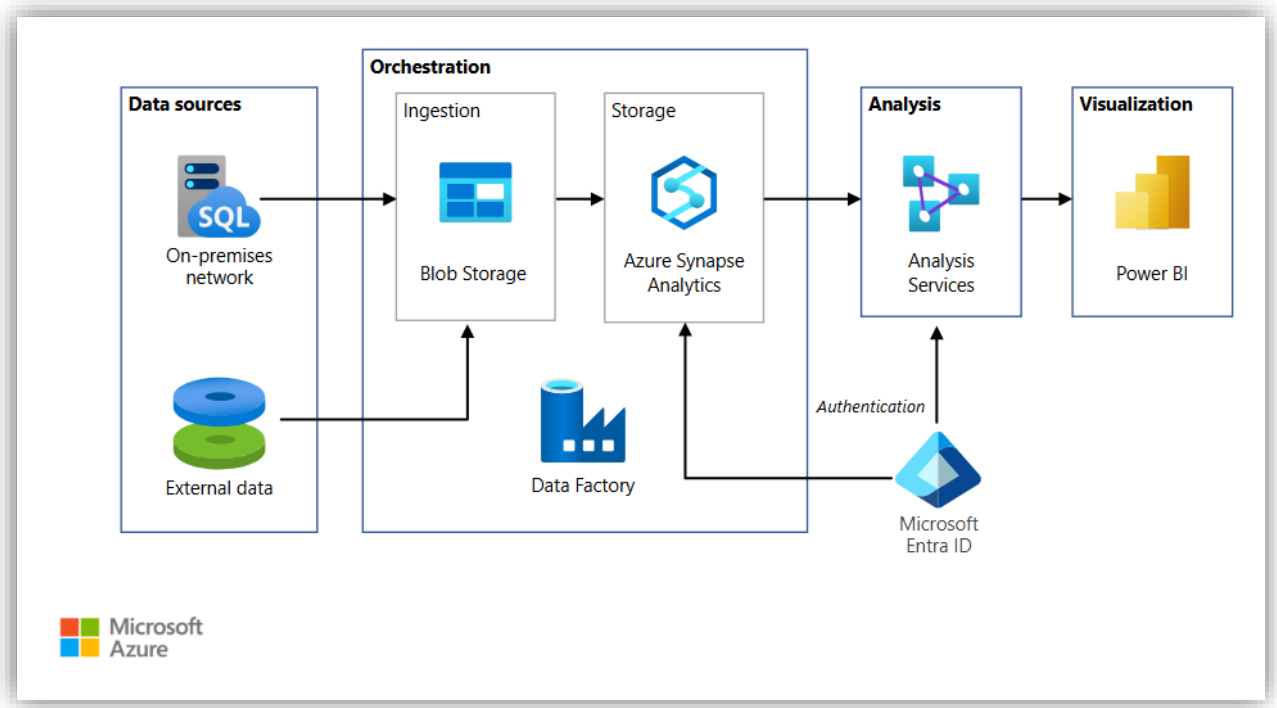
Potential Improvements:

- **Encoding**: The categorical variables (‘month’ and ‘day’) could be encoded to numerical values for use in certain types of analysis or machine learning models.
- **Outliers**: It would be beneficial to check for outliers in the numerical data which may skew analysis.
- **Feature Engineering**: New features could be created from the existing ones to provide more insights. For example, a ‘season’ feature could be engineered from the ‘month’ column.
- **Data Scaling**: The numerical data spans different ranges. Scaling the data to a standard range might be beneficial for certain types of analysis or models.
- **Adding Additional Data Sources**: Additional data such as more detailed geographical data, historical weather patterns, or human activity could enrich the dataset and provide more predictive power.

2. Architectural Diagram:

- Create an architectural diagram showing how services will be integrated.
- Specify each service's role (e.g., ADF for ingestion, ADLS for storage).

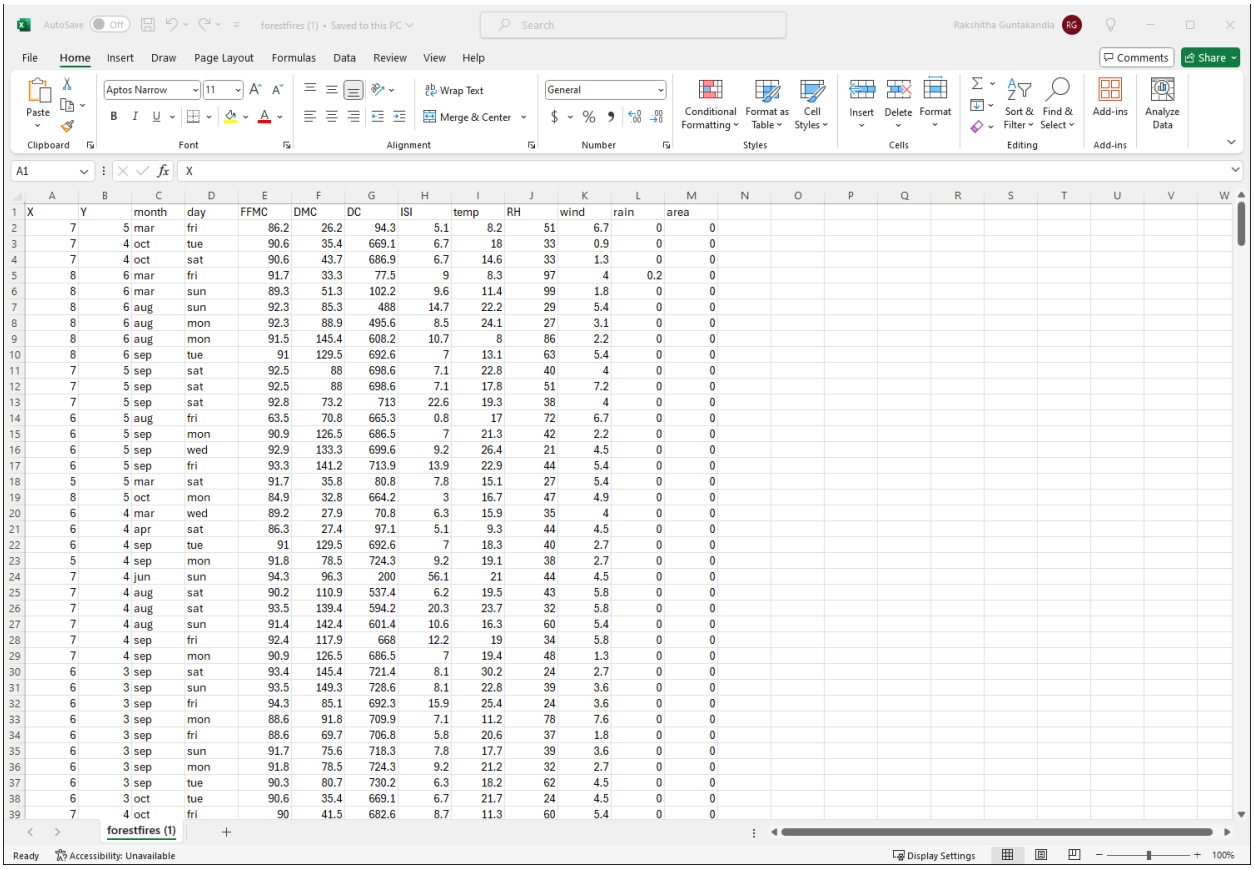




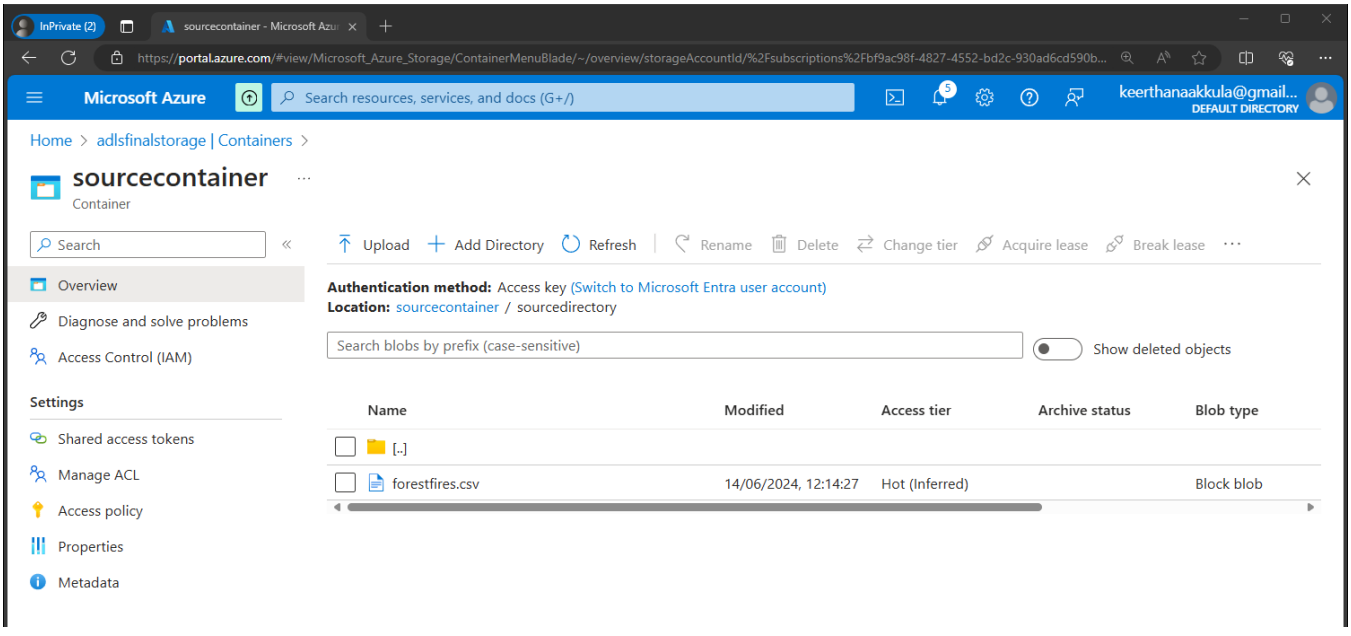

```
+------------------+          +---------------------------+          +------------------+          +---------------------+          +--------------------+
|  Data Source     |          | Azure Data Factory (ADF)  |          | Azure Data Lake  |          | Azure Synapse       |          | Power BI           |
| (Forest fire Data)|         |                           |          | Storage Gen-2    |          | Analytics           |          |                    |
+------------------+          +---------------------------+          +------------------+          +---------------------+          +--------------------+
|                  |          | - Data ingestion          |          | - Store raw data |          | - Data analysis     |          | - Dashboards       |
| - Raw data       |  -------> | - Schedule pipelines      |  ------> | - Store cleaned  |  ------->| - Run SQL queries   |  -----> | - Visualizations   |
|                  |          | - Orchestrate PySpark     |          |   data           |          | - Analytical views  |          |                    |
+------------------+          +---------------------------+          +------------------+          +---------------------+          +--------------------+
```

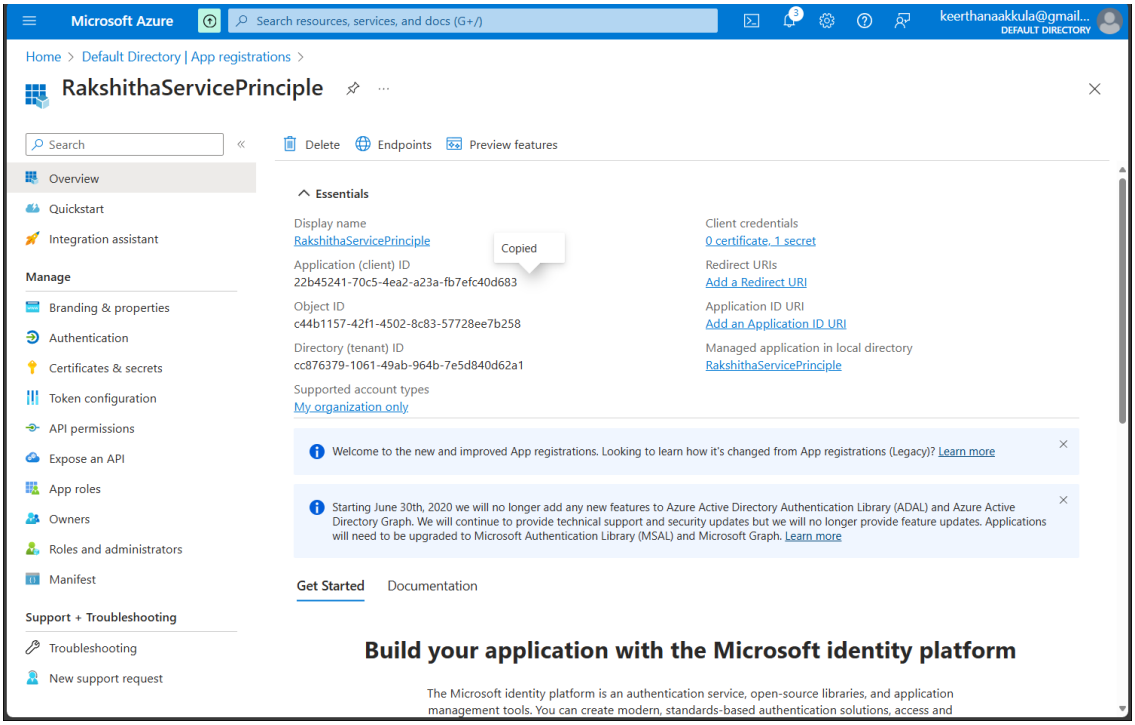
3. Data Pipeline Creation:

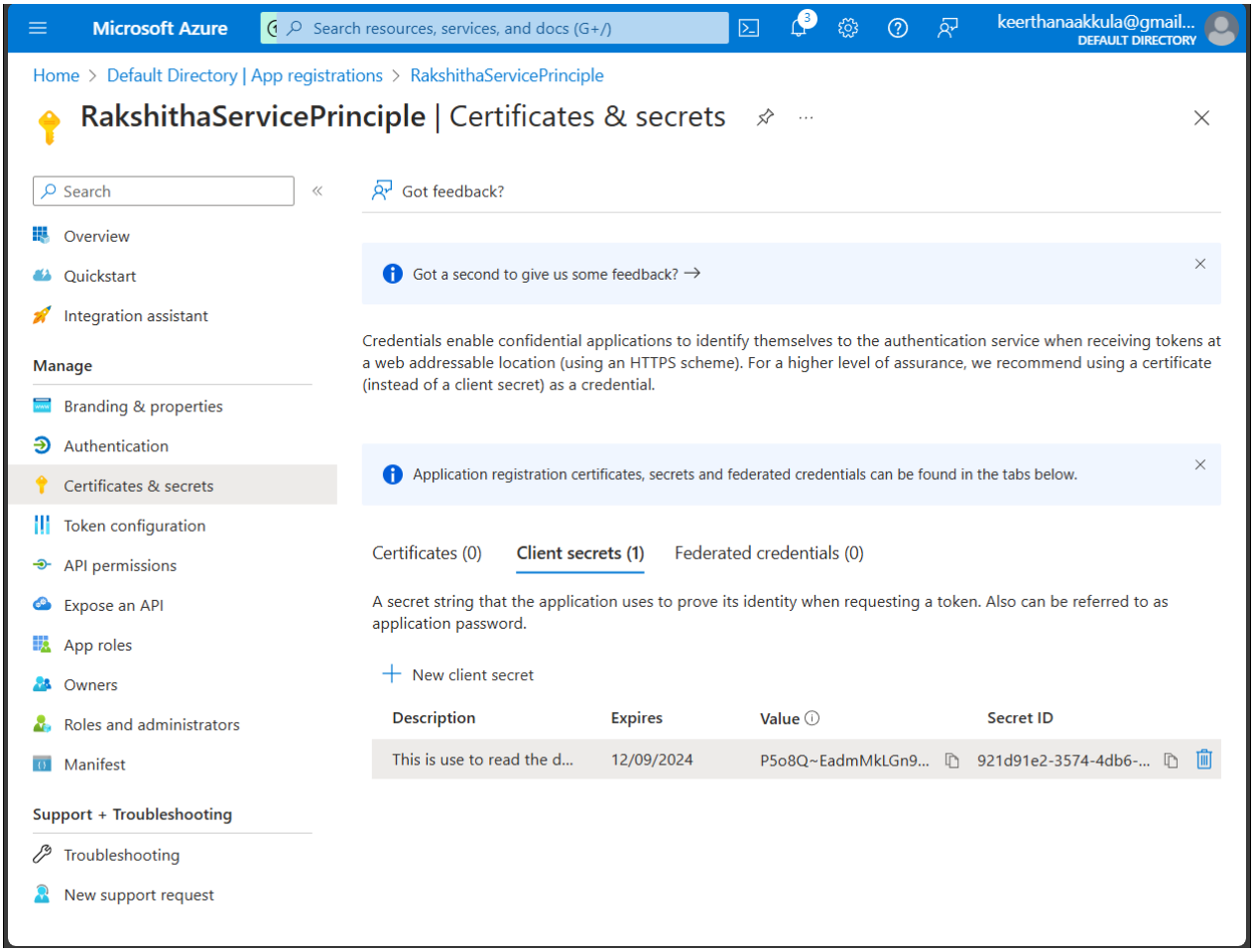
- Ingest data into the chosen source.
- Explain the choice of ingestion method.
- Store data in ADLS.
- Apply data transformations using PySpark.



X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
7	4	oct	tue	90.6	35.4	669.1	6.7	18	33	0.9	0	0
7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
8	6	mar	fri	91.7	33.3	77.5	9	8.3	97	4	0.2	0
8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0	0
8	6	aug	sun	92.3	85.3	488	14.7	22.2	29	5.4	0	0
8	6	aug	mon	92.3	88.9	495.6	8.5	24.1	27	3.1	0	0
8	6	aug	mon	91.5	145.4	608.2	10.7	8	86	2.2	0	0
8	6	sep	tue	91	129.5	692.6	7	13.1	63	5.4	0	0
7	5	sep	sat	92.5	88	698.6	7.1	22.8	40	4	0	0
7	5	sep	sat	92.5	88	698.6	7.1	17.8	51	7.2	0	0
7	5	sep	sat	92.8	73.2	713	22.6	19.3	38	4	0	0
6	5	aug	fri	63.5	70.8	665.3	0.8	17	72	6.7	0	0
6	5	sep	mon	90.9	126.5	686.5	7	21.3	42	2.2	0	0
6	5	sep	wed	92.9	133.3	699.6	9.2	26.4	21	4.5	0	0
6	5	sep	fri	93.3	141.2	713.9	13.9	22.9	44	5.4	0	0
5	5	mar	sat	91.7	35.8	80.8	7.8	15.1	27	5.4	0	0
8	5	oct	mon	84.9	32.8	664.2	3	16.7	47	4.9	0	0
6	4	mar	wed	89.2	27.9	70.8	6.3	15.9	35	4	0	0
6	4	apr	sat	86.3	27.4	97.1	5.1	9.3	44	4.5	0	0
6	4	sep	tue	91	129.5	692.6	7	18.3	40	2.7	0	0
5	4	sep	mon	91.8	78.5	724.3	9.2	19.1	38	2.7	0	0
7	4	jun	sun	94.3	96.3	200	56.1	21	44	4.5	0	0
7	4	aug	sat	90.2	110.9	537.4	6.2	19.5	43	5.8	0	0
7	4	aug	sat	93.5	139.4	594.2	20.3	23.7	32	5.8	0	0
7	4	aug	sun	91.4	142.4	601.4	10.6	16.3	60	5.4	0	0
7	4	sep	fri	92.4	117.9	668	12.2	19	34	5.8	0	0
7	4	sep	mon	90.9	126.5	686.5	7	19.4	48	1.3	0	0
6	3	sep	sat	93.4	145.4	721.4	8.1	30.2	24	2.7	0	0
6	3	sep	sun	93.5	149.3	728.6	8.1	22.8	39	3.6	0	0
6	3	sep	fri	94.3	85.1	692.3	15.9	25.4	24	3.6	0	0
6	3	sep	mon	88.6	91.8	709.9	7.1	11.2	78	7.6	0	0
6	3	sep	fri	88.6	69.7	706.8	5.8	20.6	37	1.8	0	0
6	3	sep	sun	91.7	75.6	718.3	7.8	17.7	39	3.6	0	0
6	3	sep	mon	91.8	78.5	724.3	9.2	21.2	32	2.7	0	0
6	3	sep	tue	90.3	80.7	730.2	6.3	18.2	62	4.5	0	0
6	3	oct	tue	90.6	35.4	669.1	6.7	21.7	24	4.5	0	0
7	4	oct	fri	90	41.5	682.6	8.7	11.3	60	5.4	0	0

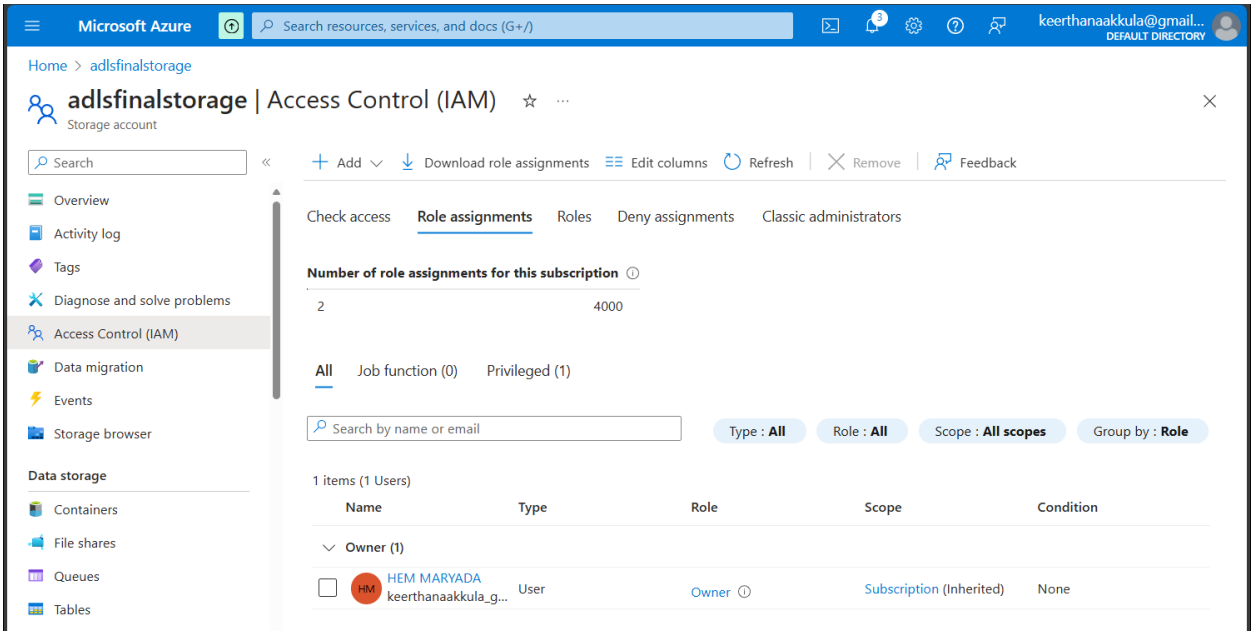


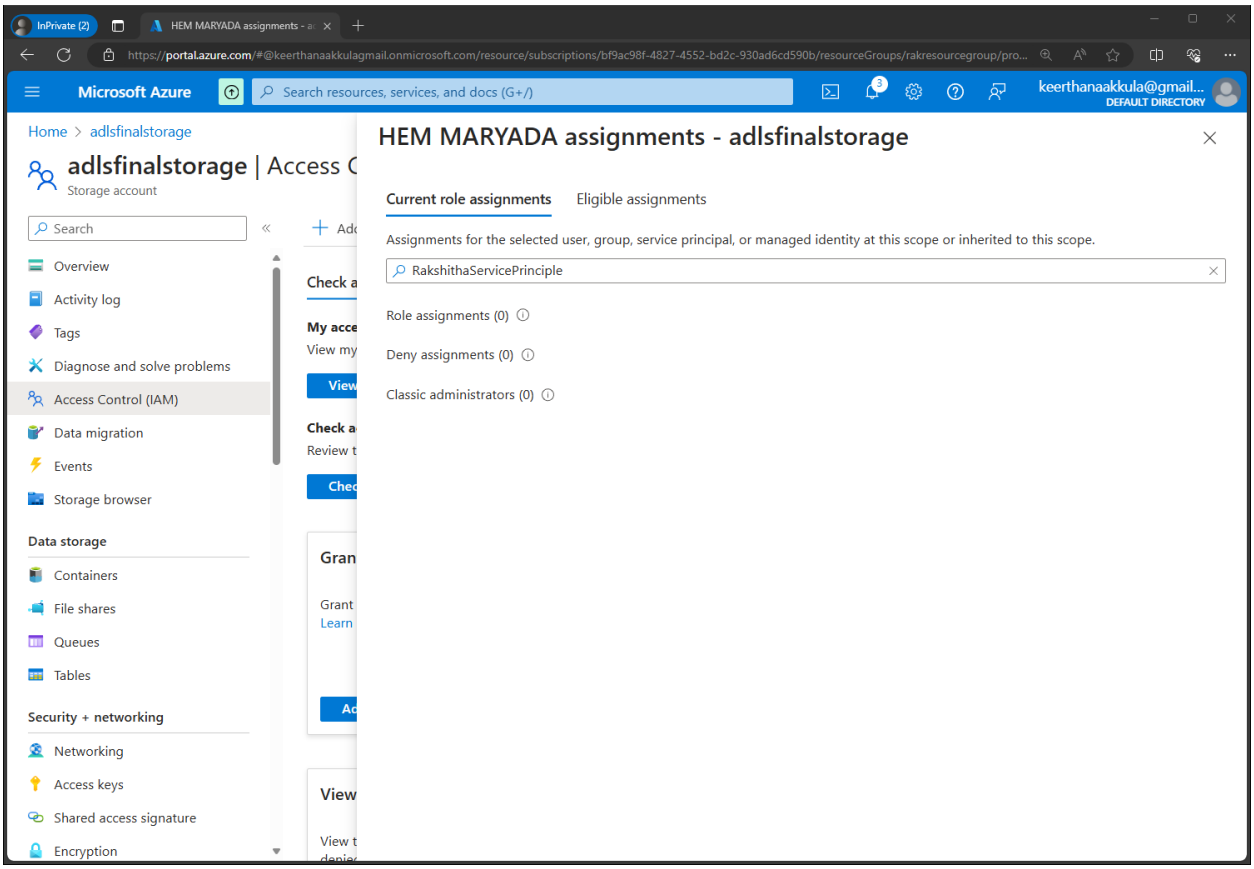

Microsoft Azure
Search resources, services, and docs (G+/)
keerthanaakkula@gmail...
DEFAULT DIRECTORY
Home > Default Directory | App registrations >
RakshithaServicePrinciple
Search
Delete
Endpoints
Preview features
Overview
Quickstart
Integration assistant
Manage
Branding & properties
Authentication
Certificates & secrets
Token configuration
API permissions
Expose an API
App roles
Owners
Roles and administrators
Manifest
Support + Troubleshooting
Troubleshooting
New support request
Essentials
Display name
RakshithaServicePrinciple
Application (client) ID
22b45241-70c5-4ea2-a23a-fb7efc40d683
Copied
Object ID
c44b1157-42f1-4502-8c83-57728ee7b258
Directory (tenant) ID
cc876379-1061-49ab-964b-7e5d840d62a1
Supported account types
My organization only
Client credentials
0 certificate, 1 secret
Redirect URIs
Add a Redirect URI
Application ID URI
Add an Application ID URI
Managed application in local directory
RakshithaServicePrinciple
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
Build your application with the Microsoft identity platform
The Microsoft identity platform is an authentication service, open-source libraries, and application management tools. You can create modern, standards-based authentication solutions, access and

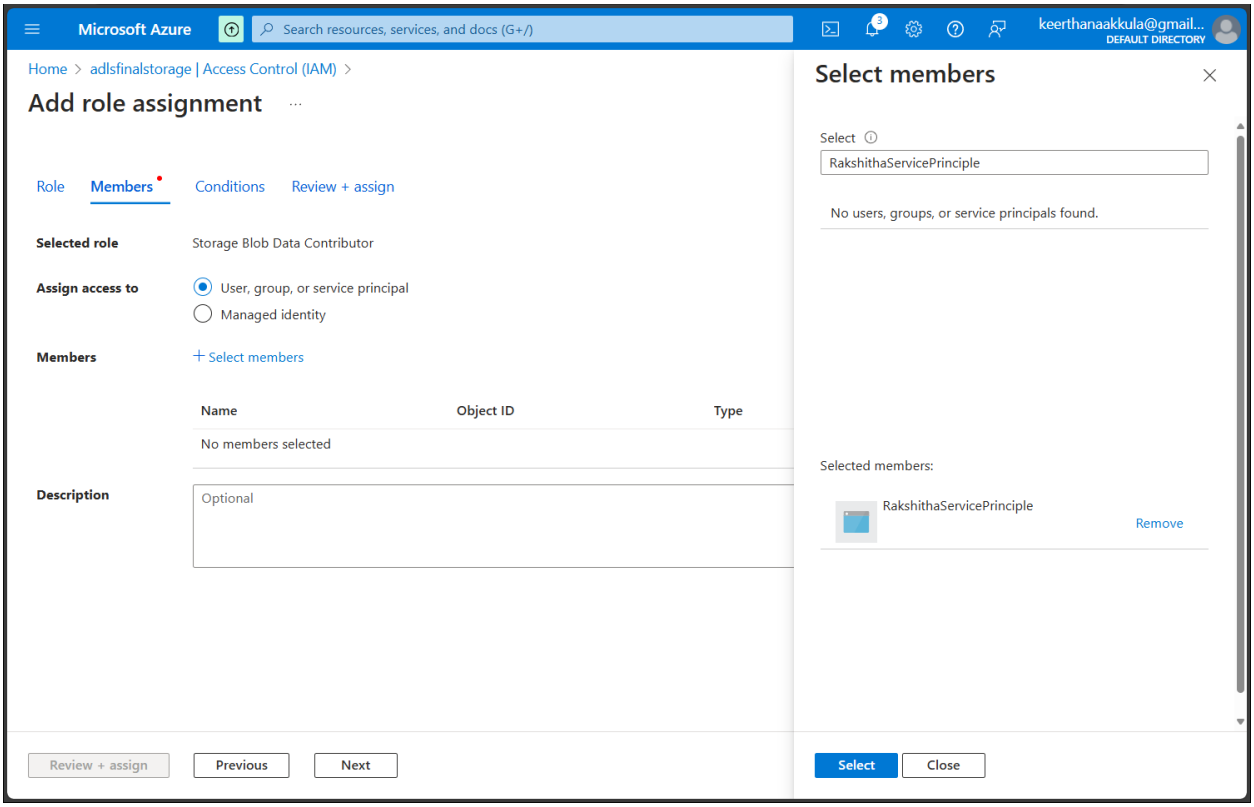

Microsoft Azure
Search resources, services, and docs (G+/)
keerthanaakkula@gmail...
DEFAULT DIRECTORY
Home > Default Directory | App registrations > RakshithaServicePrinciple
RakshithaServicePrinciple | Certificates & secrets
Search
Got feedback?
Overview
Quickstart
Integration assistant
Manage
Branding & properties
Authentication
Certificates & secrets
Token configuration
API permissions
Expose an API
App roles
Owners
Roles and administrators
Manifest
Support + Troubleshooting
Troubleshooting
New support request
Got a second to give us some feedback? →
Credentials enable confidential applications to identify themselves to the authentication service when receiving tokens at a web addressable location (using an HTTPS scheme). For a higher level of assurance, we recommend using a certificate (instead of a client secret) as a credential.
Application registration certificates, secrets and federated credentials can be found in the tabs below.
Certificates (0)
Client secrets (1)
Federated credentials (0)
A secret string that the application uses to prove its identity when requesting a token. Also can be referred to as application password.
New client secret
Description
Expires
Value
Secret ID
This is use to read the d...
12/09/2024
P5o8Q~EadmMkLGn9...
921d91e2-3574-4db6-...

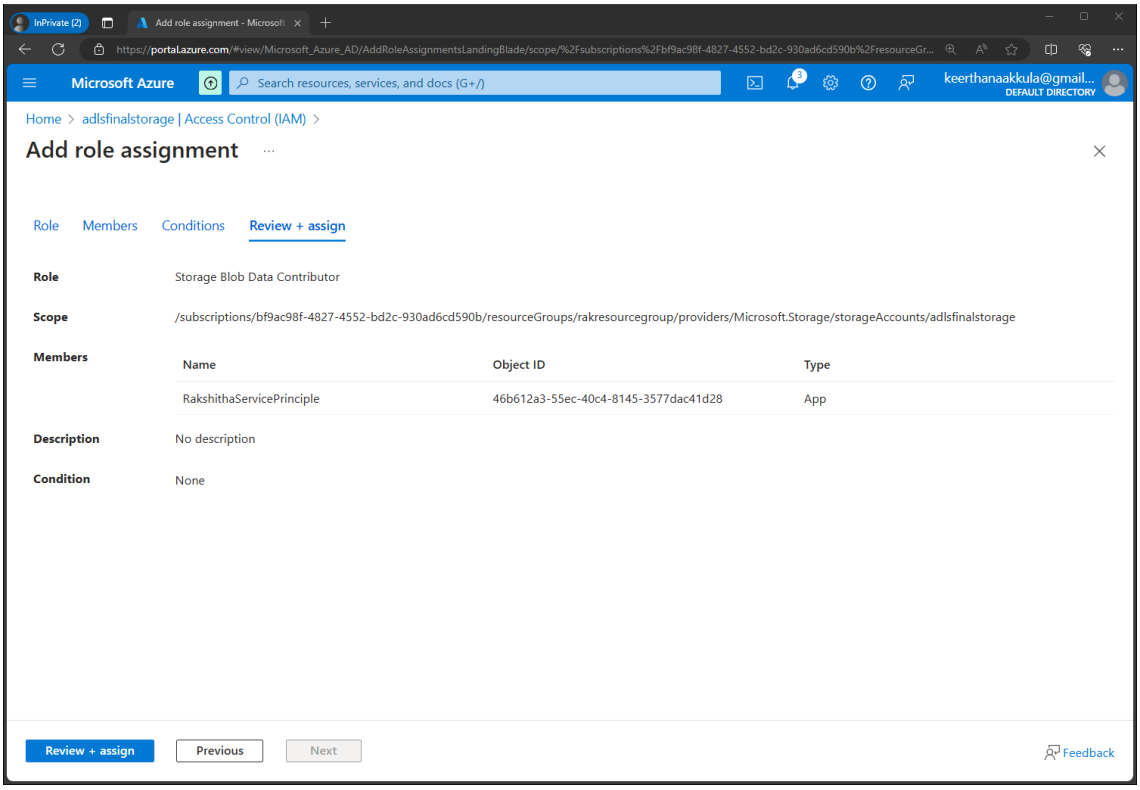
Service Principle: Client ID (username) :22b45241-70c5-4ea2-a23a-fb7efc40d683

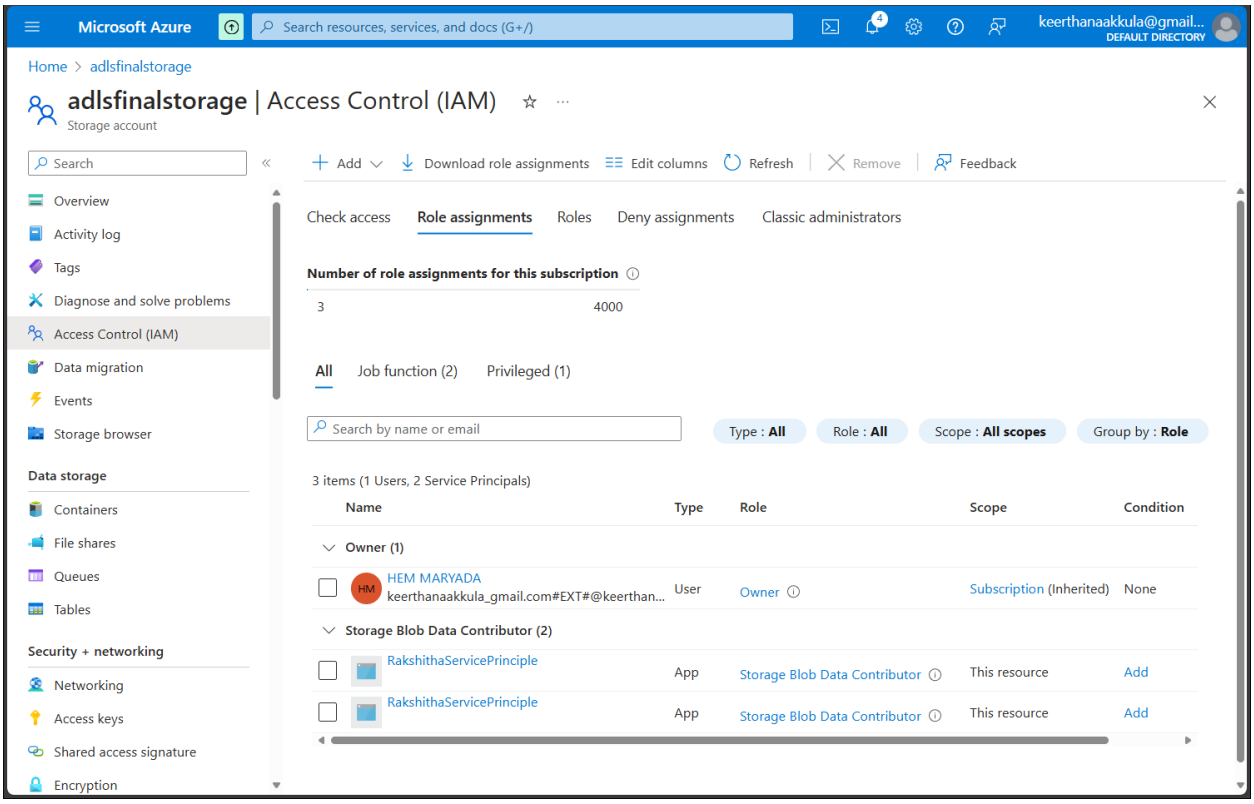
Secret value (password): P5o8Q~EadmMkLGn99uGv7St9sMw9luUUdE_kEcgg

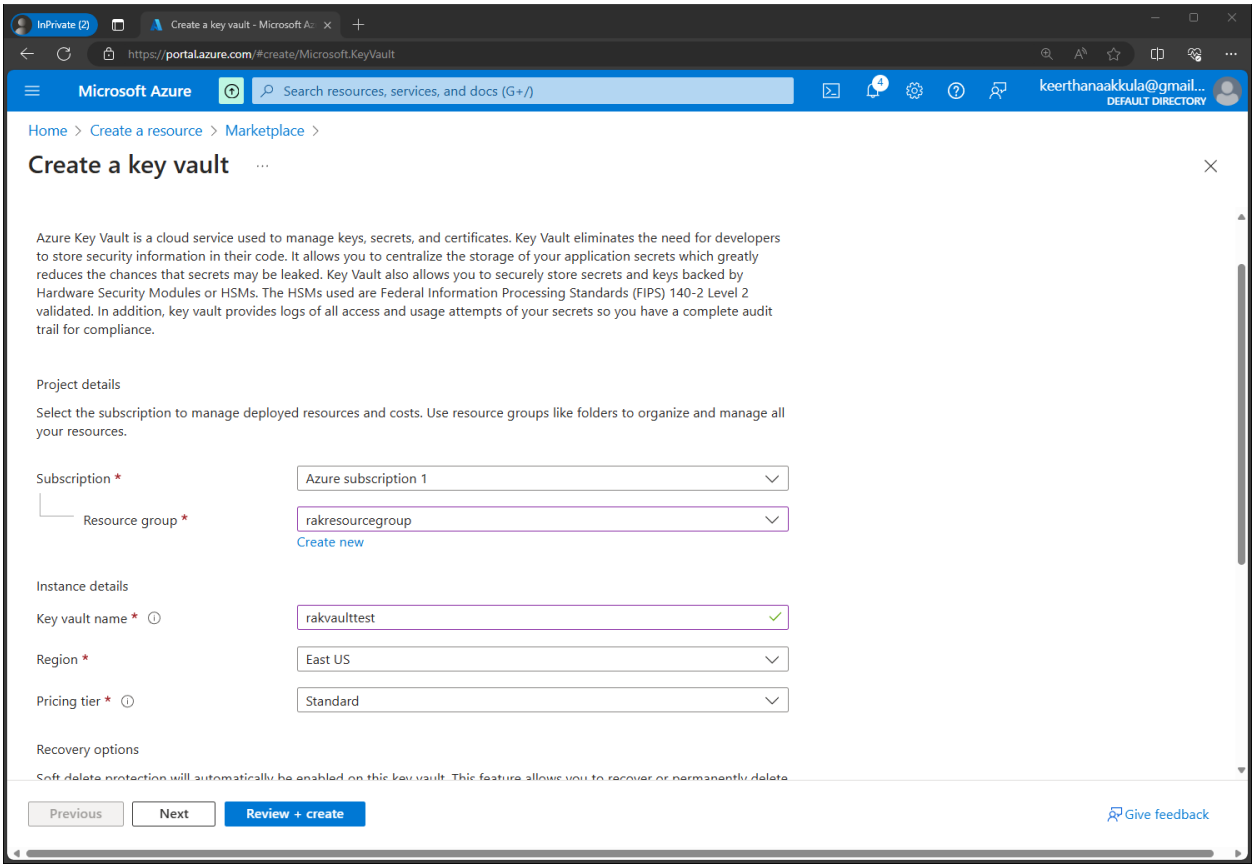


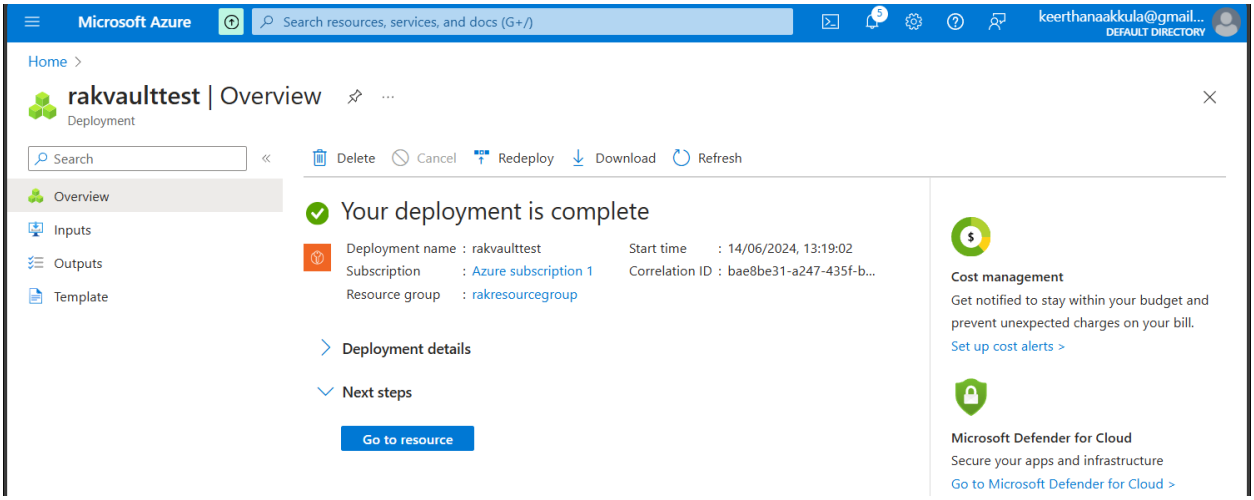


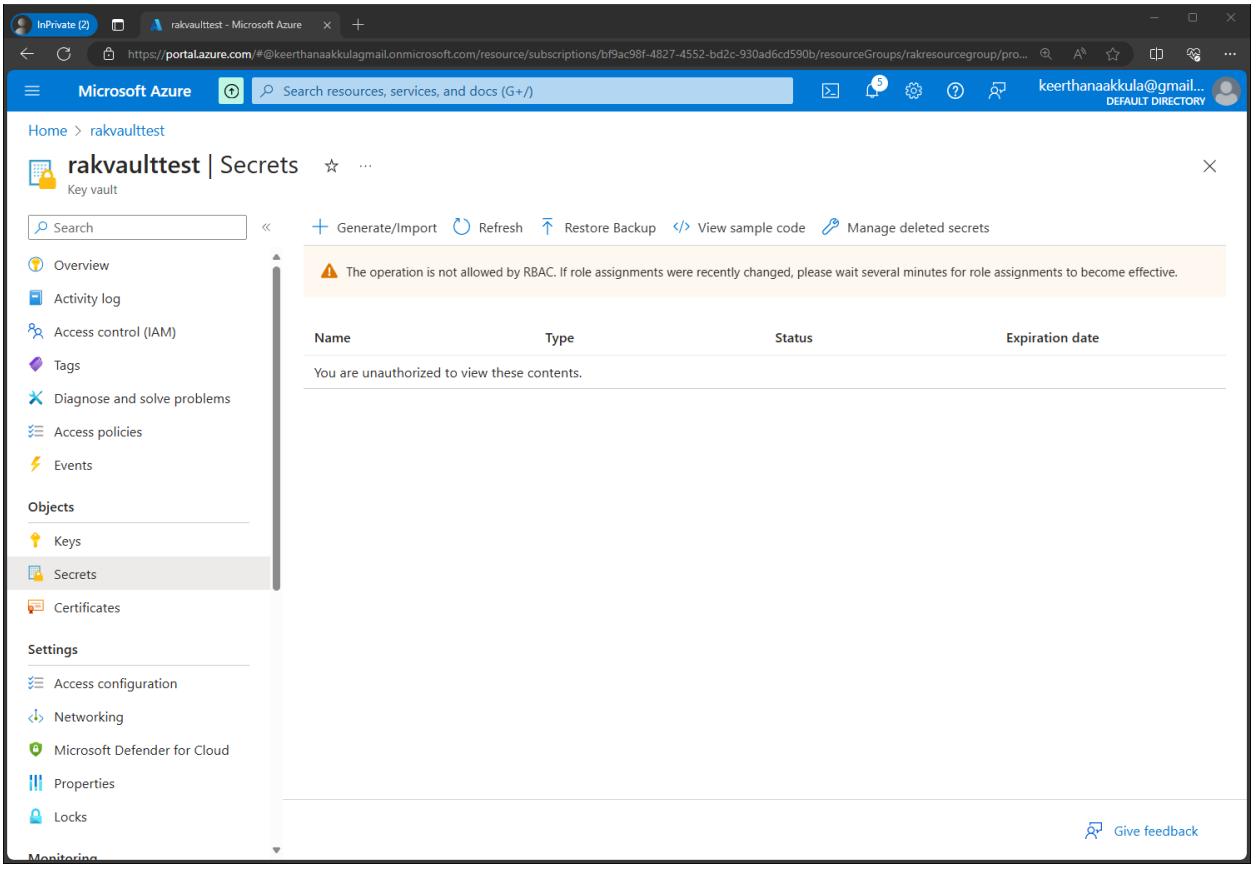

Microsoft Azure
Search resources, services, and docs (G+/)
keerthanaakkula@gmail...
DEFAULT DIRECTORY
Home > adlsfinalstorage | Access Control (IAM) >
Add role assignment
Role
Members
Conditions
Review + assign
Selected role
Storage Blob Data Contributor
Assign access to
User, group, or service principal
Managed identity
Members
+ Select members
Name
Object ID
Type
No members selected
Description
Optional
Review + assign
Previous
Next
Select members
Select
RakshithaServicePrinciple
No users, groups, or service principals found.
Selected members:
RakshithaServicePrinciple
Remove
Select
Close


InPrivate (2)
Add role assignment - Microsoft
https://portal.azure.com/#view/Microsoft_Azure_AD/AddRoleAssignmentsLandingBlade/scope/%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-930ad6cd590b%2FresourceGr...
Microsoft Azure
Search resources, services, and docs (G+/)
keerthanaakkula@gmail...
DEFAULT DIRECTORY
Home > adlsfinalstorage | Access Control (IAM) >
Add role assignment
Role
Members
Conditions
Review + assign
Role
Storage Blob Data Contributor
Scope
/subscriptions/bf9ac98f-4827-4552-bd2c-930ad6cd590b/resourceGroups/rakresourcegroup/providers/Microsoft.Storage/storageAccounts/adlsfinalstorage
Members
Name
Object ID
Type
RakshithaServicePrinciple
46b612a3-55ec-40c4-8145-3577dac41d28
App
Description
No description
Condition
None
Review + assign
Previous
Next
Feedback

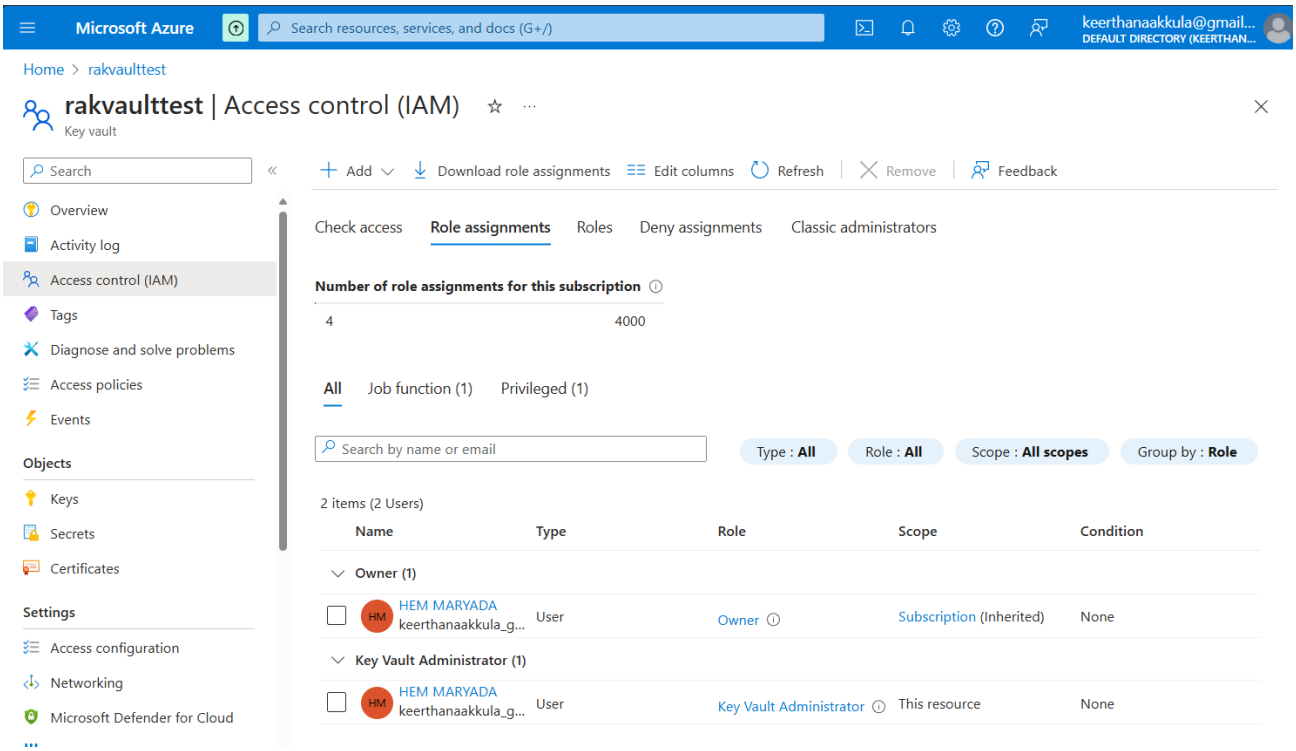

Microsoft Azure
Search resources, services, and docs (G+/)
keerthanaakkula@gmail...
DEFAULT DIRECTORY
Home > adlsfinalstorage
adlsfinalstorage | Access Control (IAM)
Storage account
Search
Add
Download role assignments
Edit columns
Refresh
Remove
Feedback
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
Data migration
Events
Storage browser
Data storage
Containers
File shares
Queues
Tables
Security + networking
Networking
Access keys
Shared access signature
Encryption
Check access
Role assignments
Roles
Deny assignments
Classic administrators
Number of role assignments for this subscription
3
4000
All
Job function (2)
Privileged (1)
Search by name or email
Type : All
Role : All
Scope : All scopes
Group by : Role
3 items (1 Users, 2 Service Principals)
Name
Type
Role
Scope
Condition
Owner (1)
HEM MARYADA
keerthanaakkula_gmail.com#EXT#@keerthan...
User
Owner
Subscription (Inherited)
None
Storage Blob Data Contributor (2)
RakshithaServicePrinciple
App
Storage Blob Data Contributor
This resource
Add
RakshithaServicePrinciple
App
Storage Blob Data Contributor
This resource
Add


InPrivate (2)
Create a key vault - Microsoft A...
https://portal.azure.com/#create/Microsoft.KeyVault
Microsoft Azure
Search resources, services, and docs (G+/)
keerthanaakkula@gmail...
DEFAULT DIRECTORY
Home > Create a resource > Marketplace >
Create a key vault
Azure Key Vault is a cloud service used to manage keys, secrets, and certificates. Key Vault eliminates the need for developers to store security information in their code. It allows you to centralize the storage of your application secrets which greatly reduces the chances that secrets may be leaked. Key Vault also allows you to securely store secrets and keys backed by Hardware Security Modules or HSMs. The HSMs used are Federal Information Processing Standards (FIPS) 140-2 Level 2 validated. In addition, key vault provides logs of all access and usage attempts of your secrets so you have a complete audit trail for compliance.
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 *
rakresourcegroup
Create new
Instance details
Key vault name *
rakvaulttest
Region *
East US
Pricing tier *
Standard
Recovery options
Previous
Next
Review + create
Give feedback



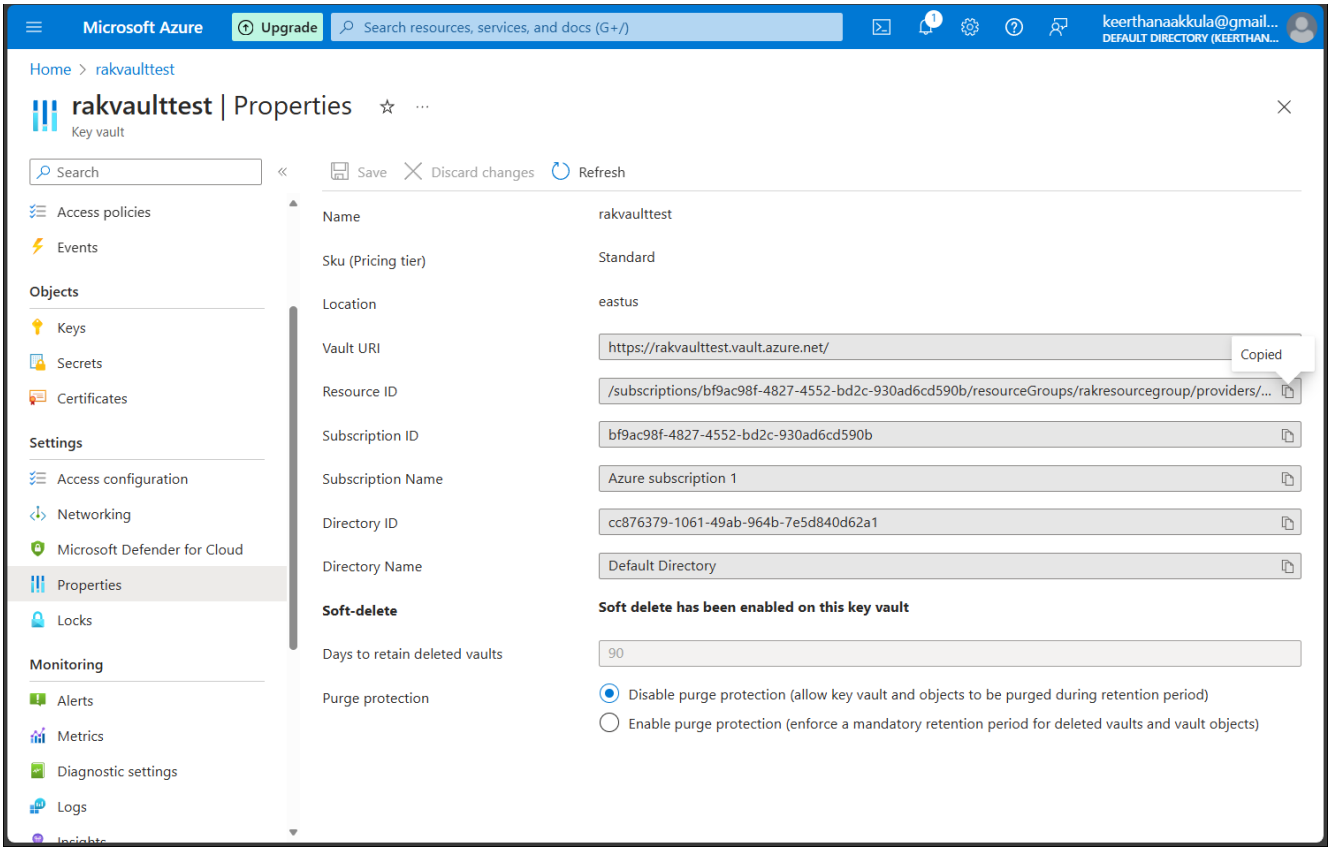


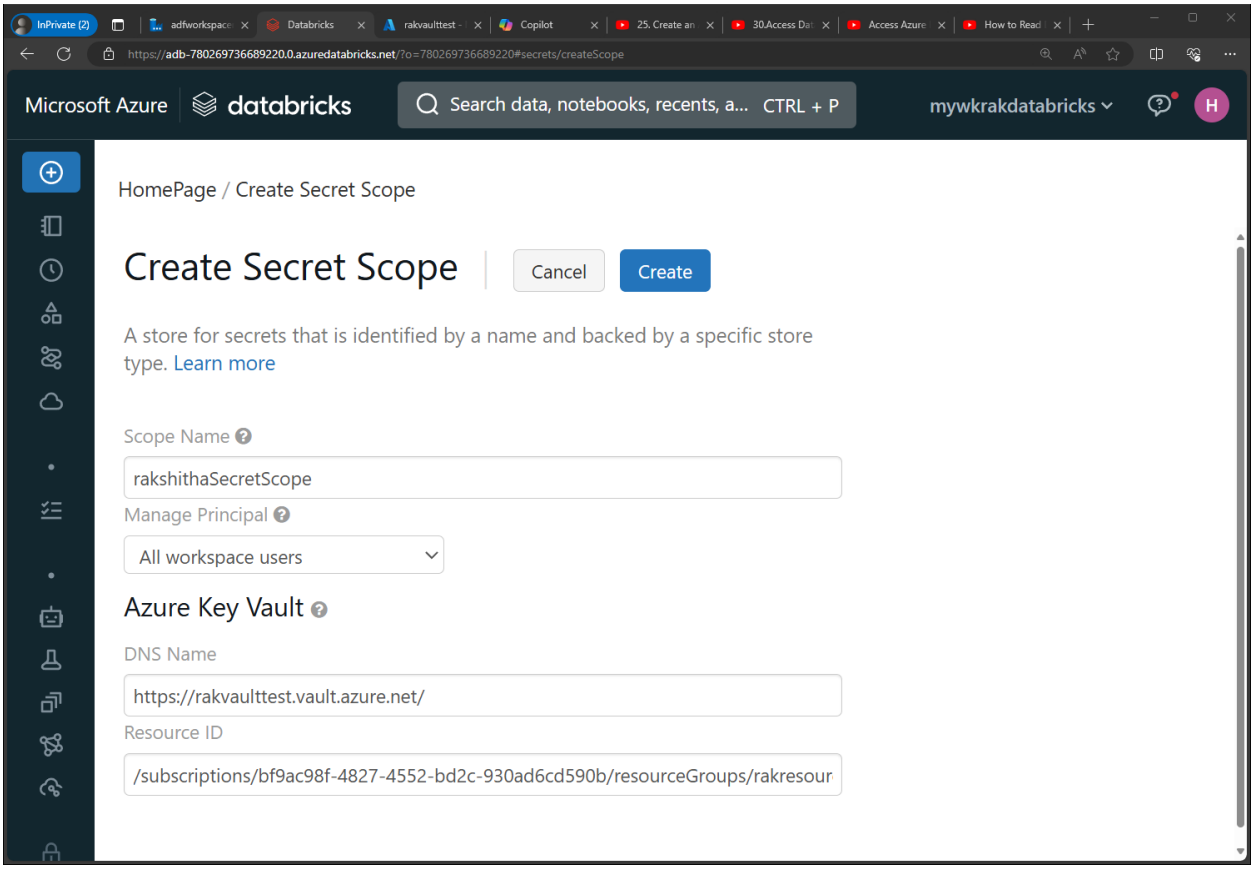
In the above screenshot as you can see that the you have to assign the admin role to this key vault account.

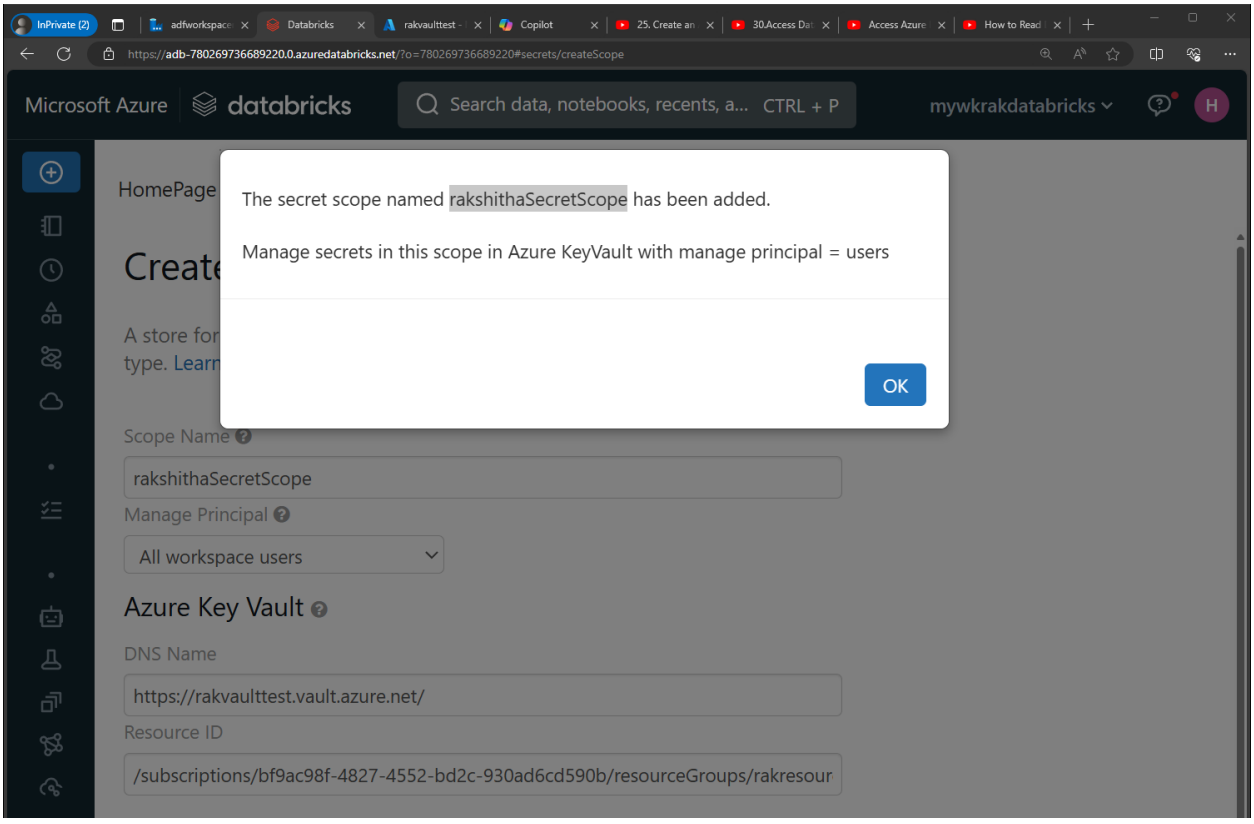


Open data bricks and open this link

https:?<databricksinstance>#secrets/createScope

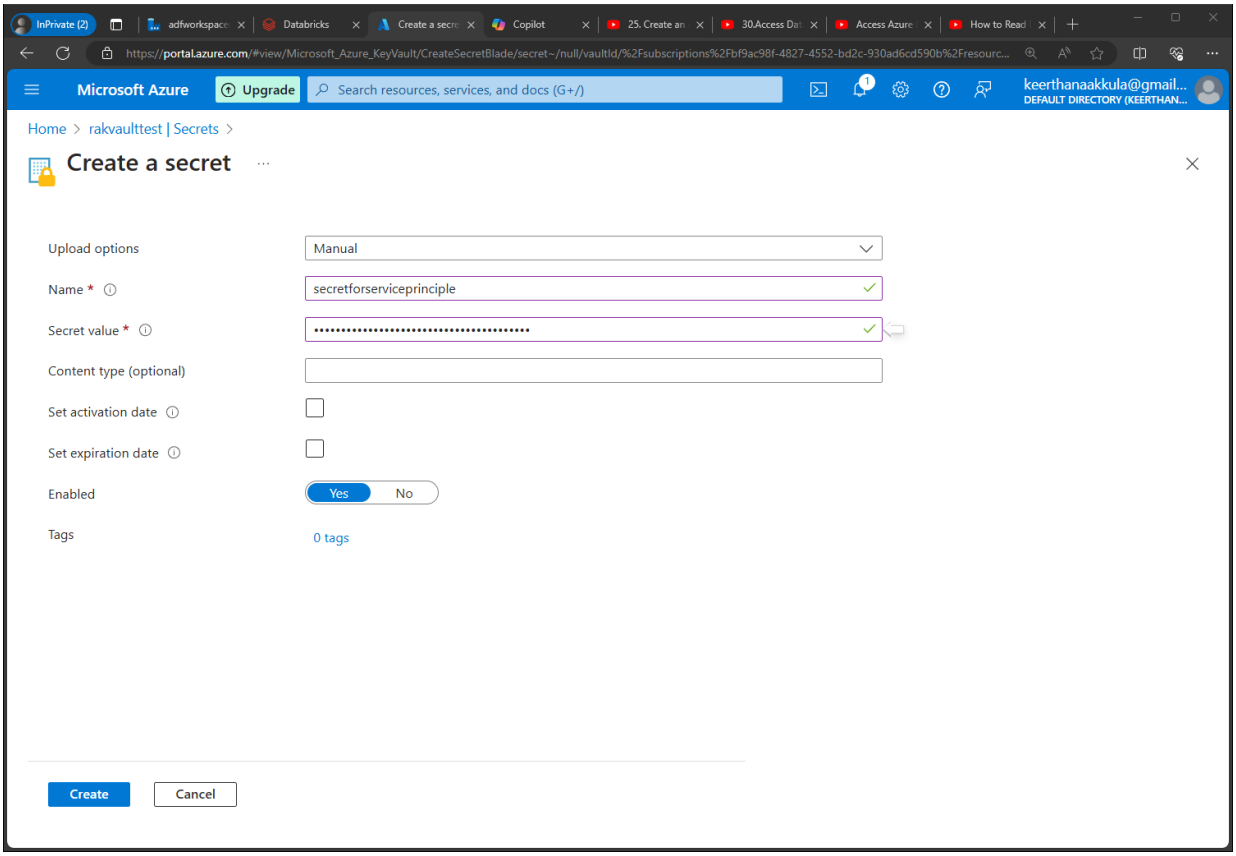


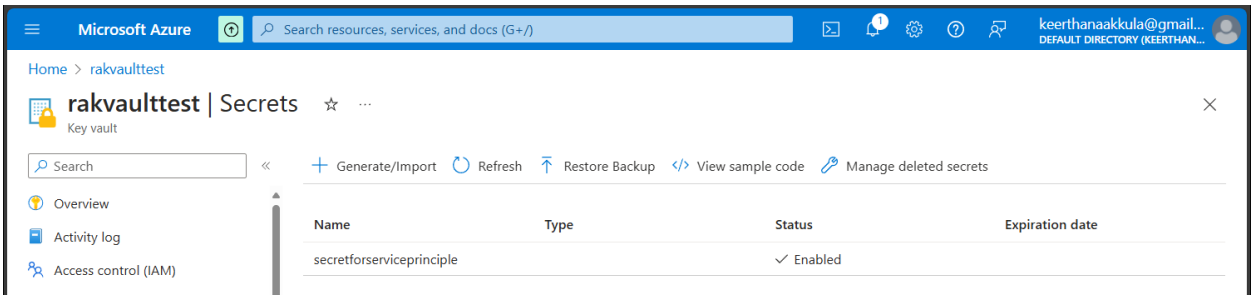

Microsoft Azure
databricks
Search data, notebooks, recents, a... CTRL + P
mywkrakdatabricks
HomePage / Create Secret Scope
Create Secret Scope
Cancel
Create
A store for secrets that is identified by a name and backed by a specific store type. Learn more
Scope Name
rakshithaSecretScope
Manage Principal
All workspace users
Azure Key Vault
DNS Name
https://rakvaulttest.vault.azure.net/
Resource ID
/subscriptions/bf9ac98f-4827-4552-bd2c-930ad6cd590b/resourceGroups/rakresour


Microsoft Azure
databricks
Search data, notebooks, recents, a... CTRL + P
mywkrakdatabricks
HomePage
The secret scope named rakshithaSecretScope has been added.
Manage secrets in this scope in Azure KeyVault with manage principal = users
OK
Scope Name
rakshithaSecretScope
Manage Principal
All workspace users
Azure Key Vault
DNS Name
https://rakvaulttest.vault.azure.net/
Resource ID
/subscriptions/bf9ac98f-4827-4552-bd2c-930ad6cd590b/resourceGroups/rakresour

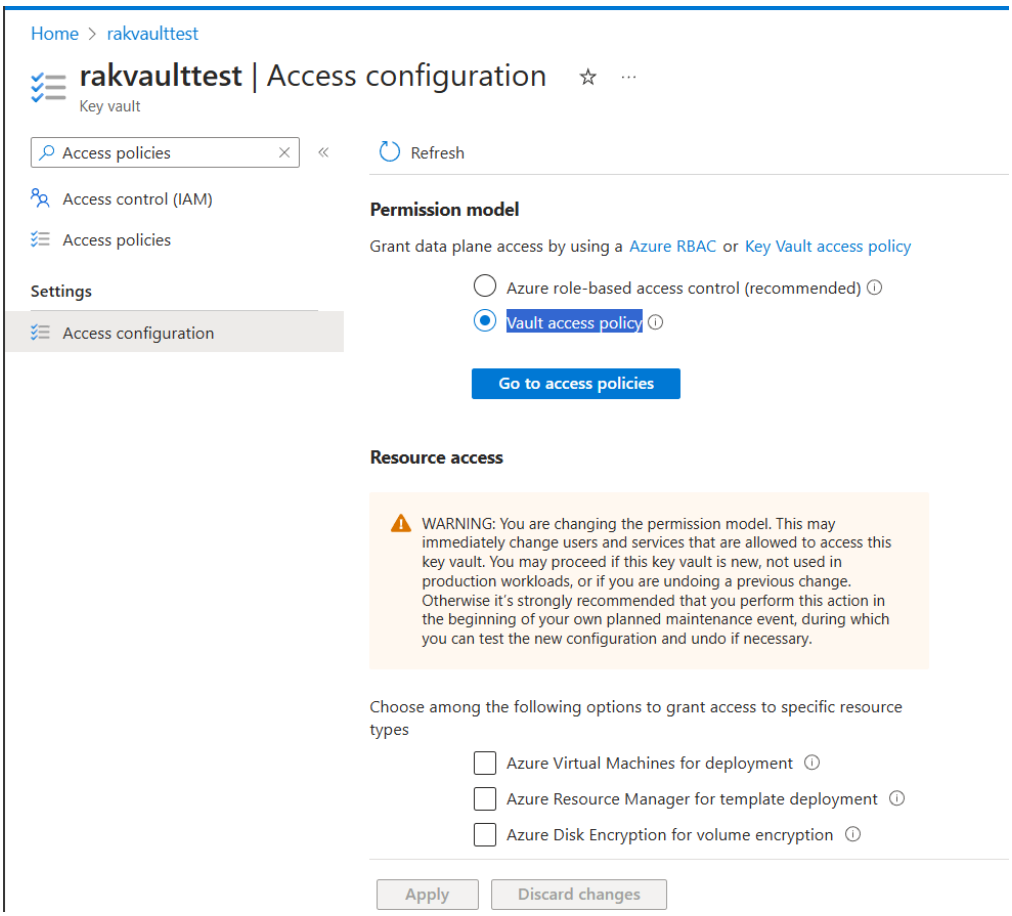
Under secret value put the Service principal name

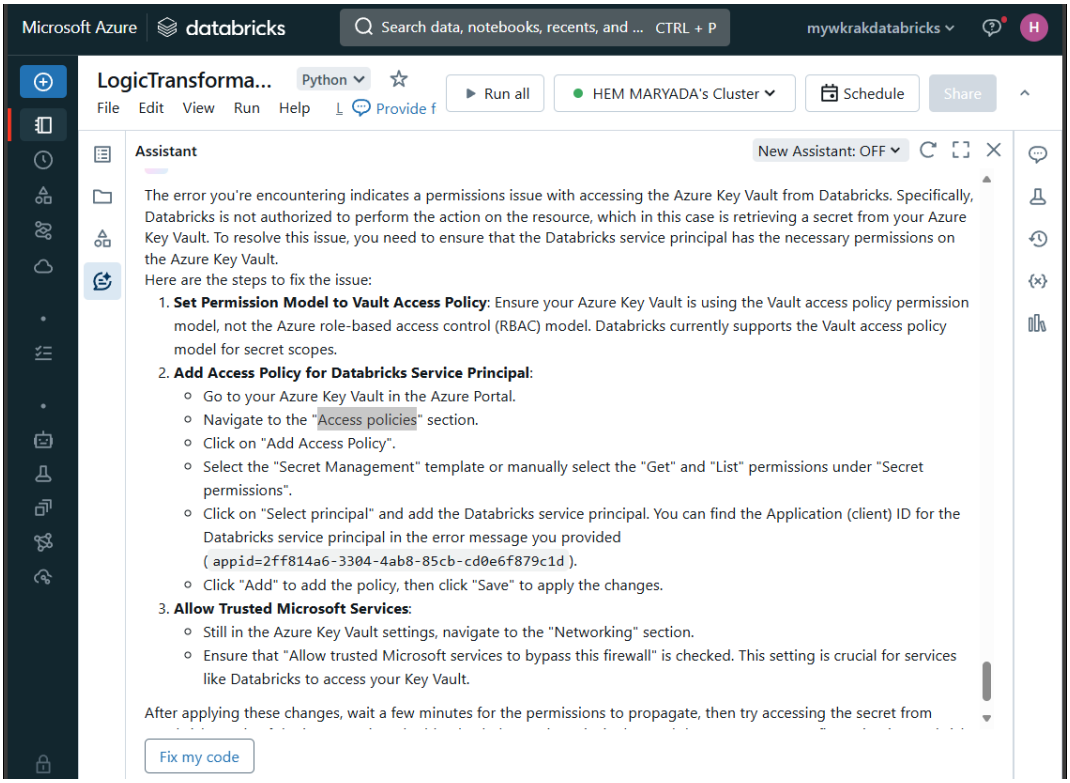
You have make sure that you the admin/permissions of that key vault to create the secret or any certification, if you not admin/have relevant permissions you can't cannt create an secret.

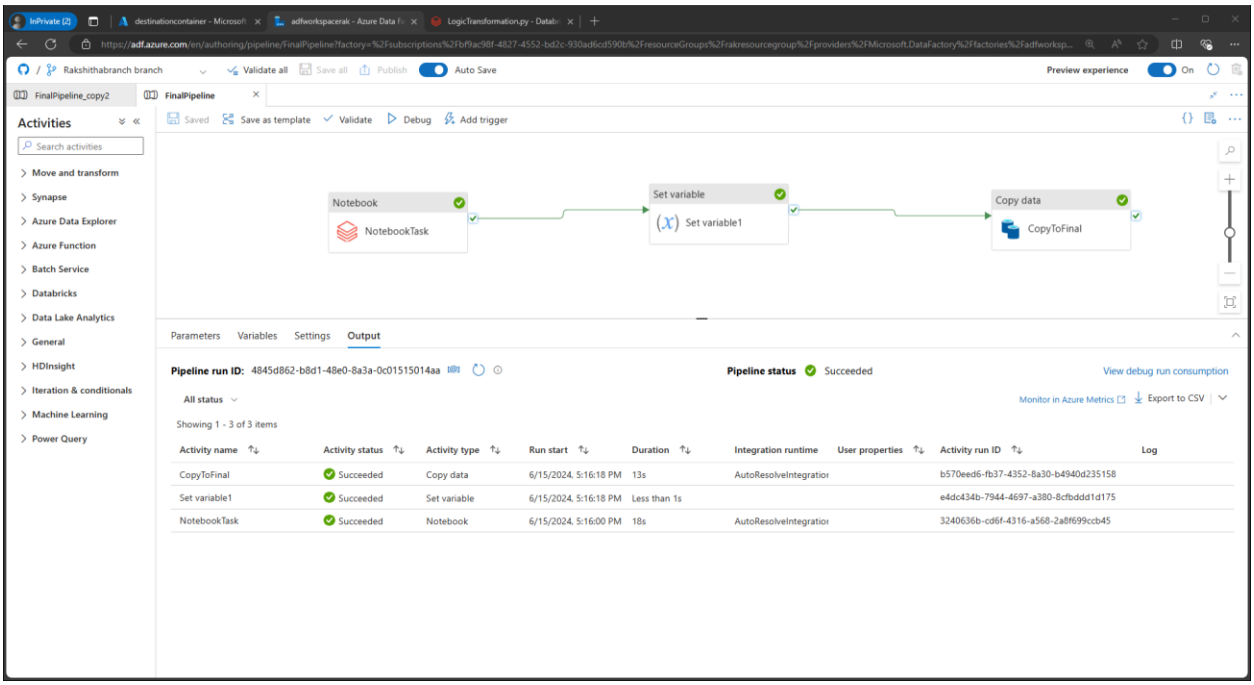


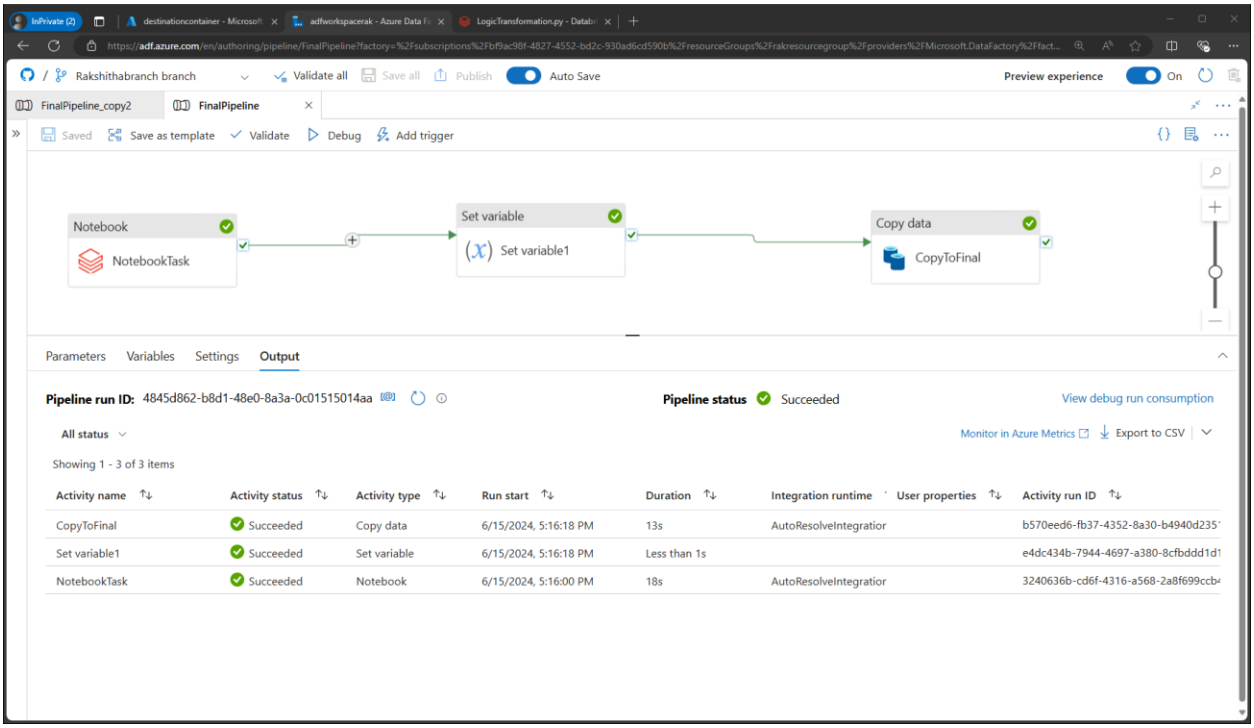


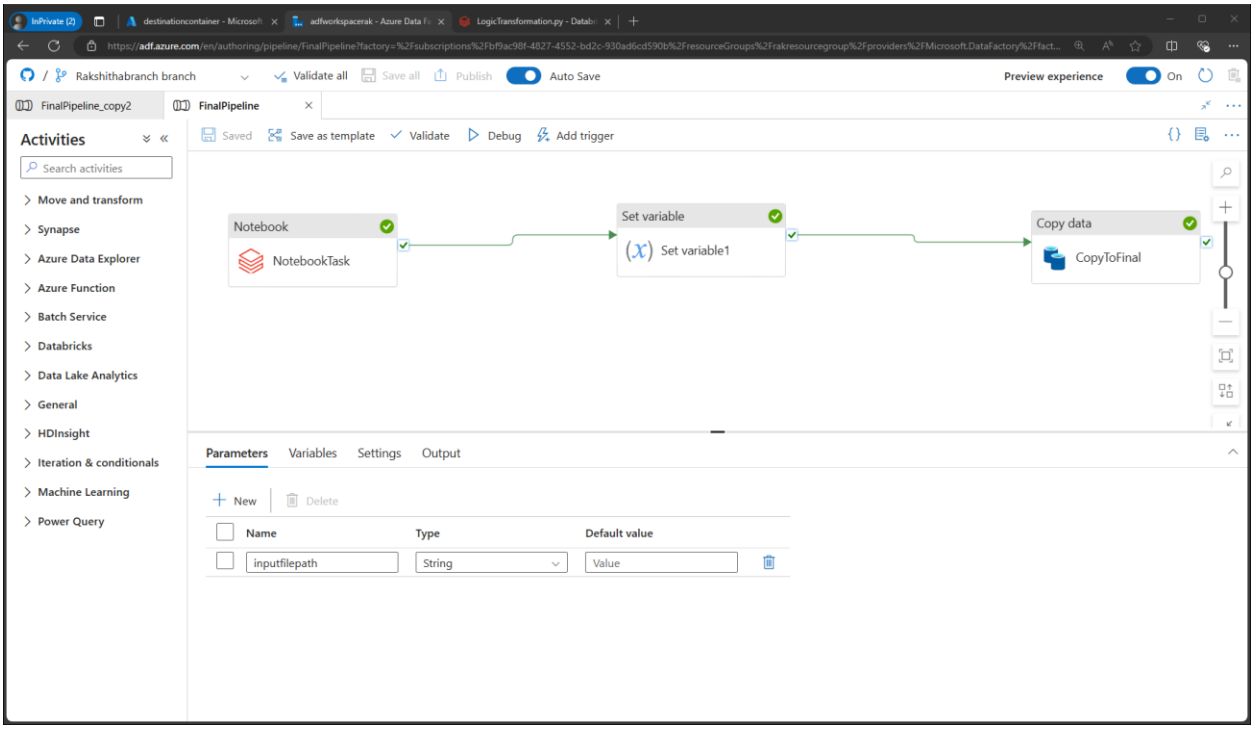
dbutils.secrets.get(scope ="", key=""). I got an error while I was running the databricks notebook. You cannt access the key vault using a role-based access you need to enable the Vault access policy from access configuration

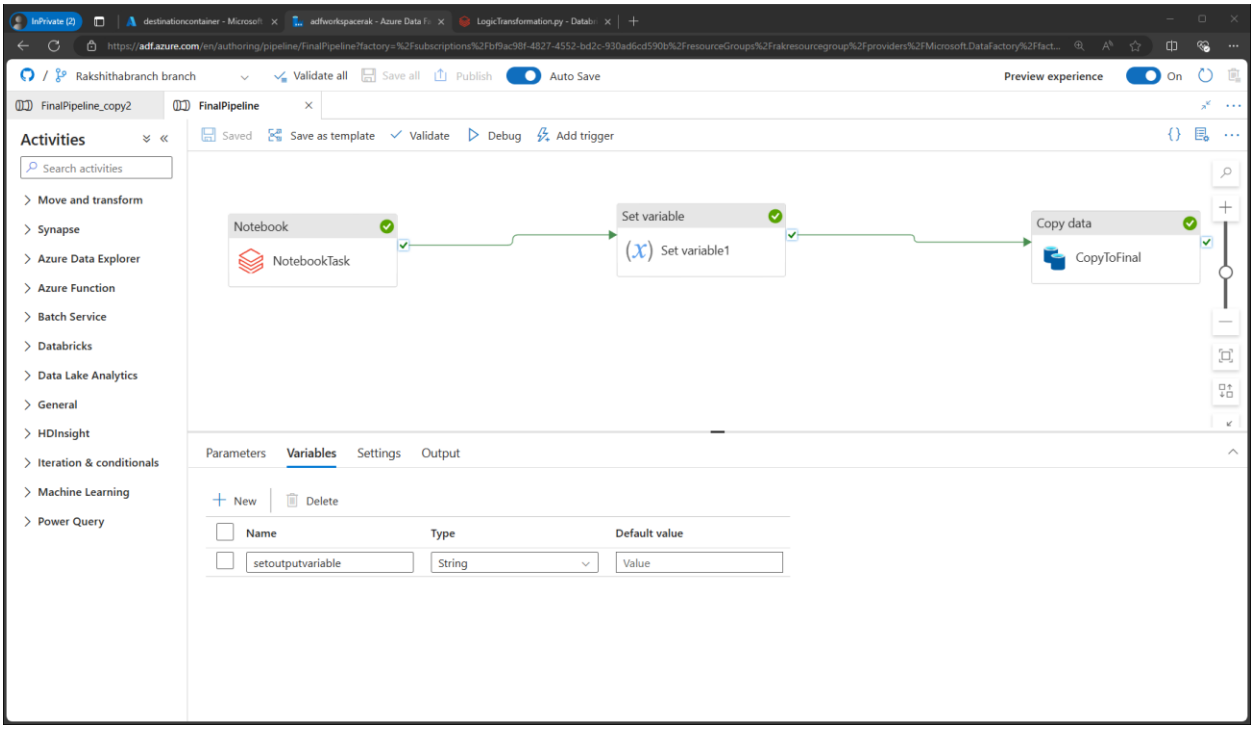

Home > rakvaulttest
rakvaulttest | Access configuration
Key vault
Access policies
Access control (IAM)
Access policies
Settings
Access configuration
Refresh
Permission model
Grant data plane access by using a Azure RBAC or Key Vault access policy
Azure role-based access control (recommended)
Vault access policy
Go to access policies
Resource access
WARNING: You are changing the permission model. This may immediately change users and services that are allowed to access this key vault. You may proceed if this key vault is new, not used in production workloads, or if you are undoing a previous change. Otherwise it's strongly recommended that you perform this action in the beginning of your own planned maintenance event, during which you can test the new configuration and undo if necessary.
Choose among the following options to grant access to specific resource types
Azure Virtual Machines for deployment
Azure Resource Manager for template deployment
Azure Disk Encryption for volume encryption
Apply
Discard changes

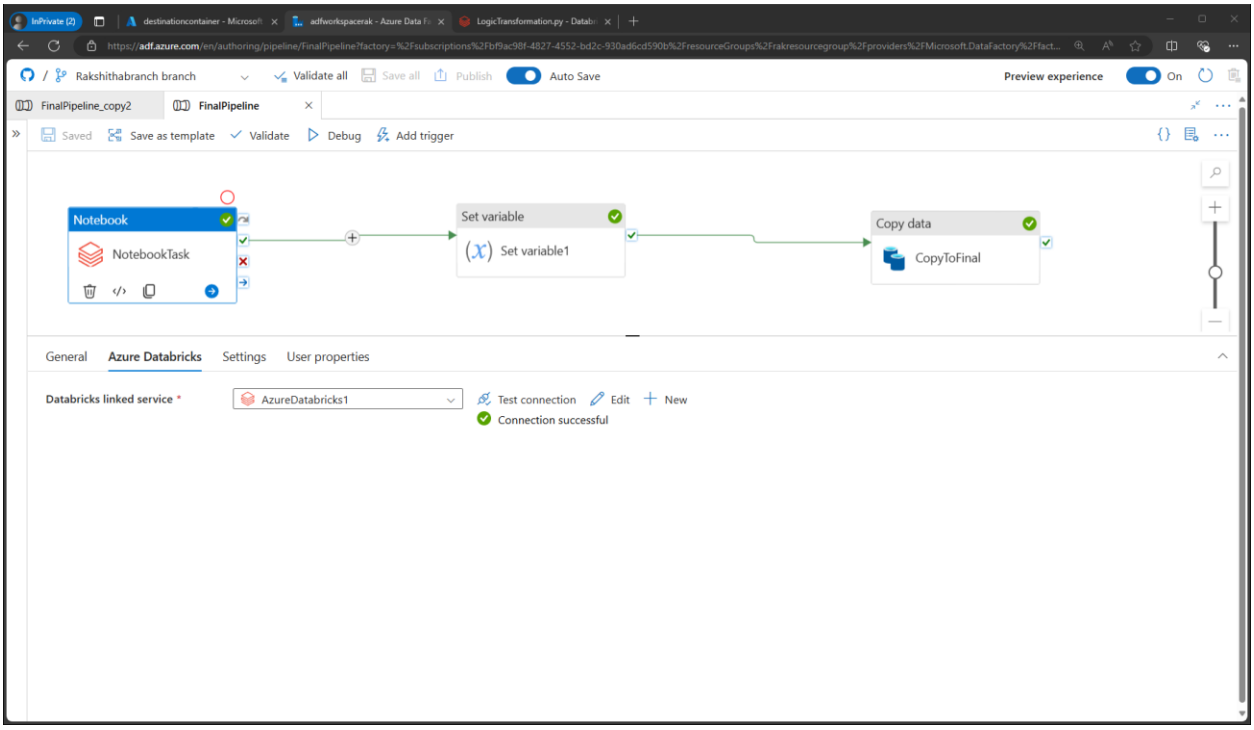

Microsoft Azure
databricks
Search data, notebooks, recents, and ... CTRL + P
mywkrakdatabricks
LogicTransforma...
Python
File Edit View Run Help
Provide f
Run all
HEM MARYADA's Cluster
Schedule
Share
Assistant
New Assistant: OFF
The error you're encountering indicates a permissions issue with accessing the Azure Key Vault from Databricks. Specifically, Databricks is not authorized to perform the action on the resource, which in this case is retrieving a secret from your Azure Key Vault. To resolve this issue, you need to ensure that the Databricks service principal has the necessary permissions on the Azure Key Vault.
Here are the steps to fix the issue:
1. Set Permission Model to Vault Access Policy: Ensure your Azure Key Vault is using the Vault access policy permission model, not the Azure role-based access control (RBAC) model. Databricks currently supports the Vault access policy model for secret scopes.
2. Add Access Policy for Databricks Service Principal:
- Go to your Azure Key Vault in the Azure Portal.
- Navigate to the "Access policies" section.
- Click on "Add Access Policy".
- Select the "Secret Management" template or manually select the "Get" and "List" permissions under "Secret permissions".
- Click on "Select principal" and add the Databricks service principal. You can find the Application (client) ID for the Databricks service principal in the error message you provided (appid=2ff814a6-3304-4ab8-85cb-cd0e6f879c1d).
- Click "Add" to add the policy, then click "Save" to apply the changes.
3. Allow Trusted Microsoft Services:
- Still in the Azure Key Vault settings, navigate to the "Networking" section.
- Ensure that "Allow trusted Microsoft services to bypass this firewall" is checked. This setting is crucial for services like Databricks to access your Key Vault.
After applying these changes, wait a few minutes for the permissions to propagate, then try accessing the secret from
Fix my code

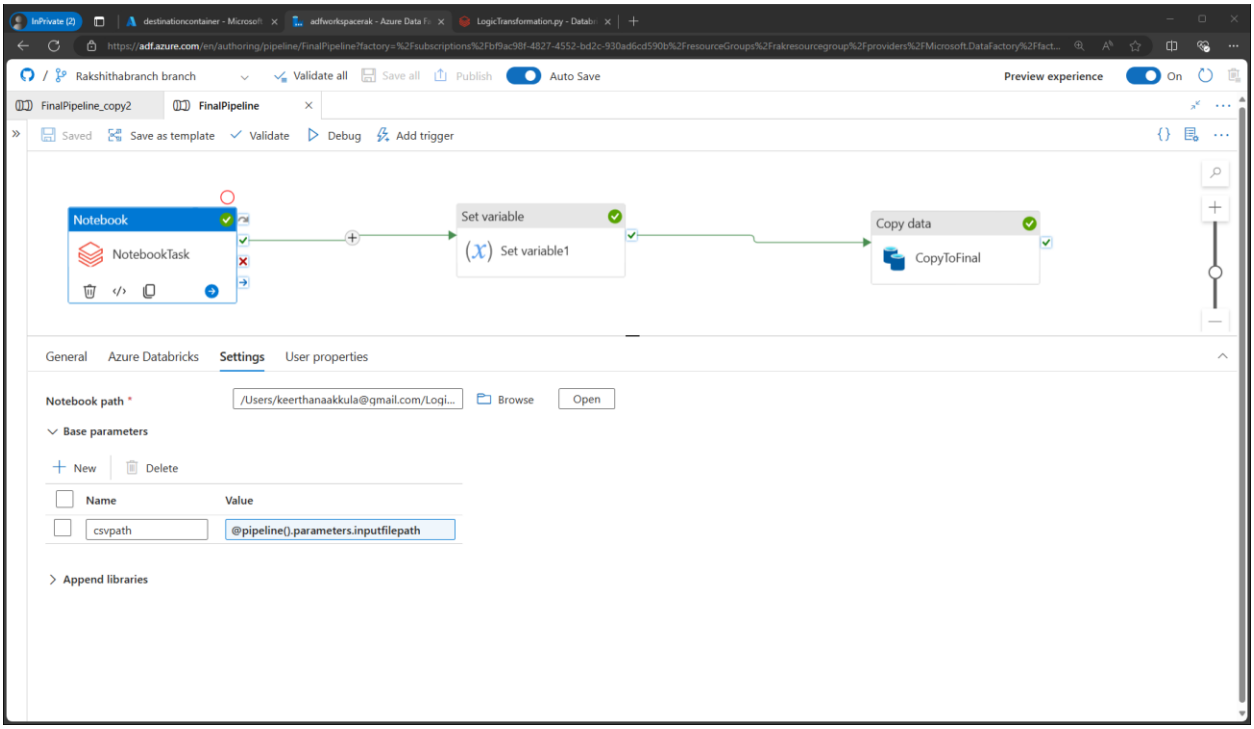

Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
FinalPipeline_copy2
FinalPipeline
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Saved
Save as template
Validate
Debug
Add trigger
Notebook
NotebookTask
Set variable
Set variable1
Copy data
CopyToFinal
Parameters
Variables
Settings
Output
Pipeline run ID: 4845d862-b8d1-48e0-8a3a-0c01515014aa
Pipeline status Succeeded
View debug run consumption
All status
Monitor in Azure Metrics
Export to CSV
Showing 1 - 3 of 3 items
Activity name
Activity status
Activity type
Run start
Duration
Integration runtime
User properties
Activity run ID
Log
CopyToFinal
Succeeded
Copy data
6/15/2024, 5:16:18 PM
13s
AutoResolveIntegratior
b570eed6-fb37-4352-8a30-b4940d235158
Set variable1
Succeeded
Set variable
6/15/2024, 5:16:18 PM
Less than 1s
e4dc434b-7944-4697-a380-8cfbddd1d175
NotebookTask
Succeeded
Notebook
6/15/2024, 5:16:00 PM
18s
AutoResolveIntegratior
3240636b-cd6f-4316-a568-2a8f699ccb45

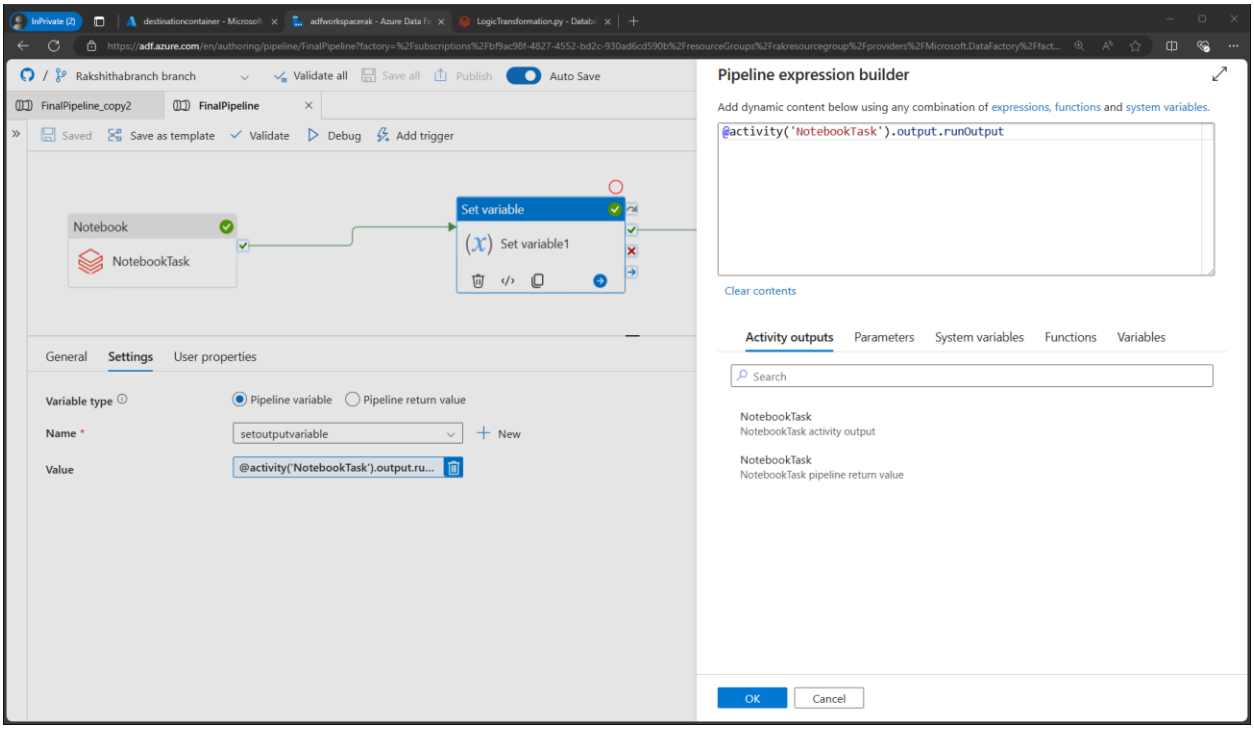

Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
FinalPipeline_copy2
FinalPipeline
Saved
Save as template
Validate
Debug
Add trigger
Notebook
NotebookTask
Set variable
Set variable1
Copy data
CopyToFinal
Parameters
Variables
Settings
Output
Pipeline run ID: 4845d862-b8d1-48e0-8a3a-0c01515014aa
Pipeline status Succeeded
View debug run consumption
All status
Monitor in Azure Metrics
Export to CSV
Showing 1 - 3 of 3 items
Activity name
Activity status
Activity type
Run start
Duration
Integration runtime
User properties
Activity run ID
CopyToFinal
Succeeded
Copy data
6/15/2024, 5:16:18 PM
13s
AutoResolveIntegratior
Set variable1
Succeeded
Set variable
6/15/2024, 5:16:18 PM
Less than 1s
NotebookTask
Succeeded
Notebook
6/15/2024, 5:16:00 PM
18s
AutoResolveIntegratior

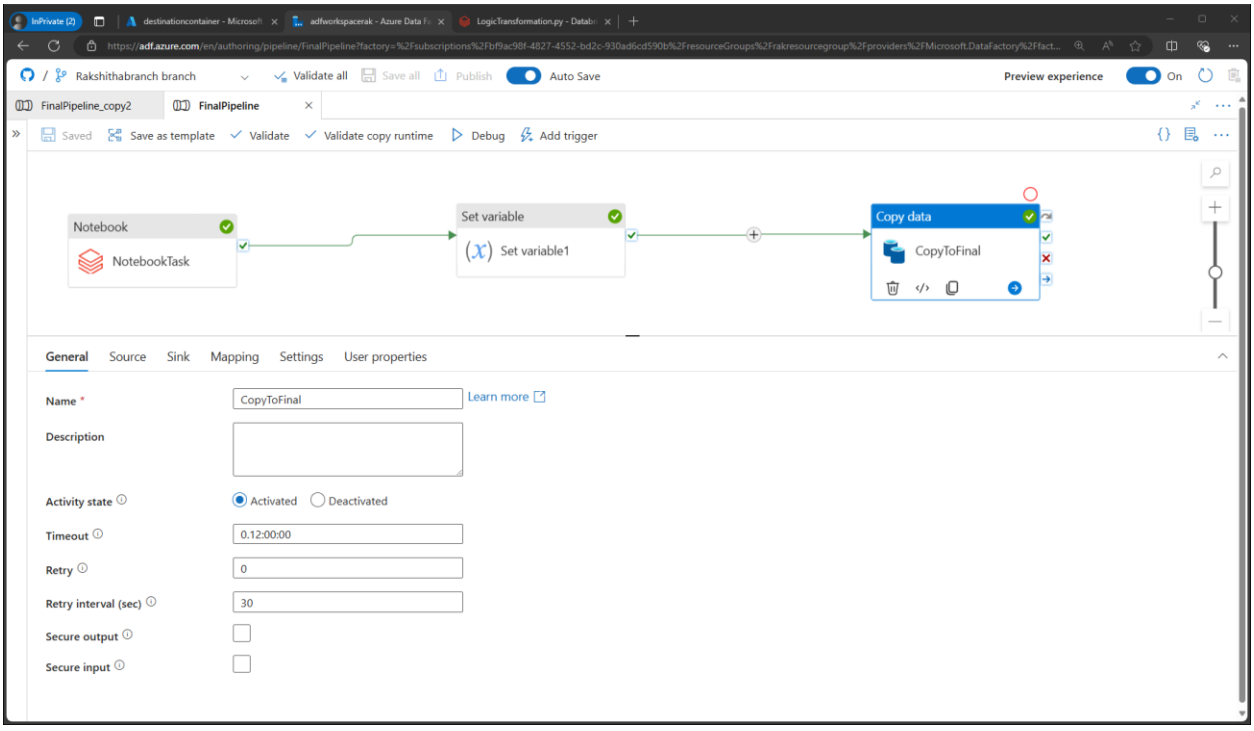

Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
FinalPipeline_copy2
FinalPipeline
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Saved
Save as template
Validate
Debug
Add trigger
Notebook
NotebookTask
Set variable
Set variable1
Copy data
CopyToFinal
Parameters
Variables
Settings
Output
New
Delete
Name
Type
Default value
inputfilepath
String
Value

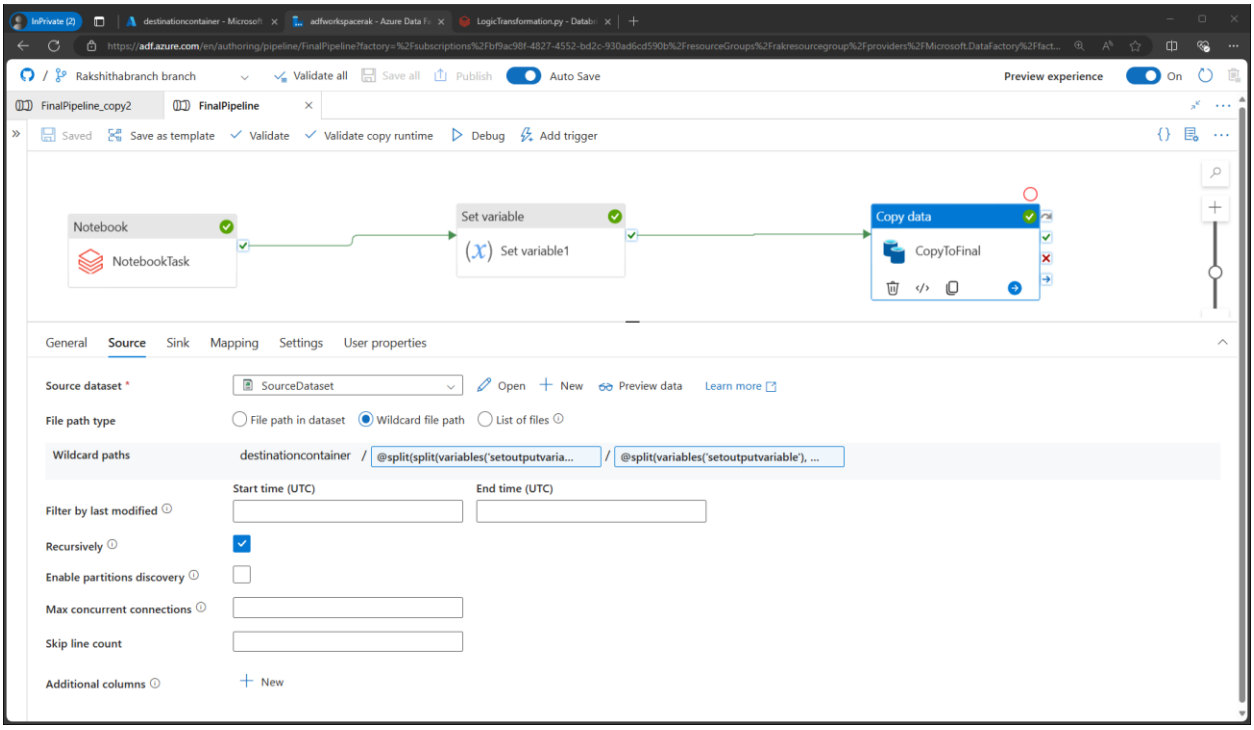

Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
FinalPipeline_copy2
FinalPipeline
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Saved
Save as template
Validate
Debug
Add trigger
Notebook
NotebookTask
Set variable
Set variable1
Copy data
CopyToFinal
Parameters
Variables
Settings
Output
New
Delete
Name
Type
Default value
setoutputvariable
String
Value

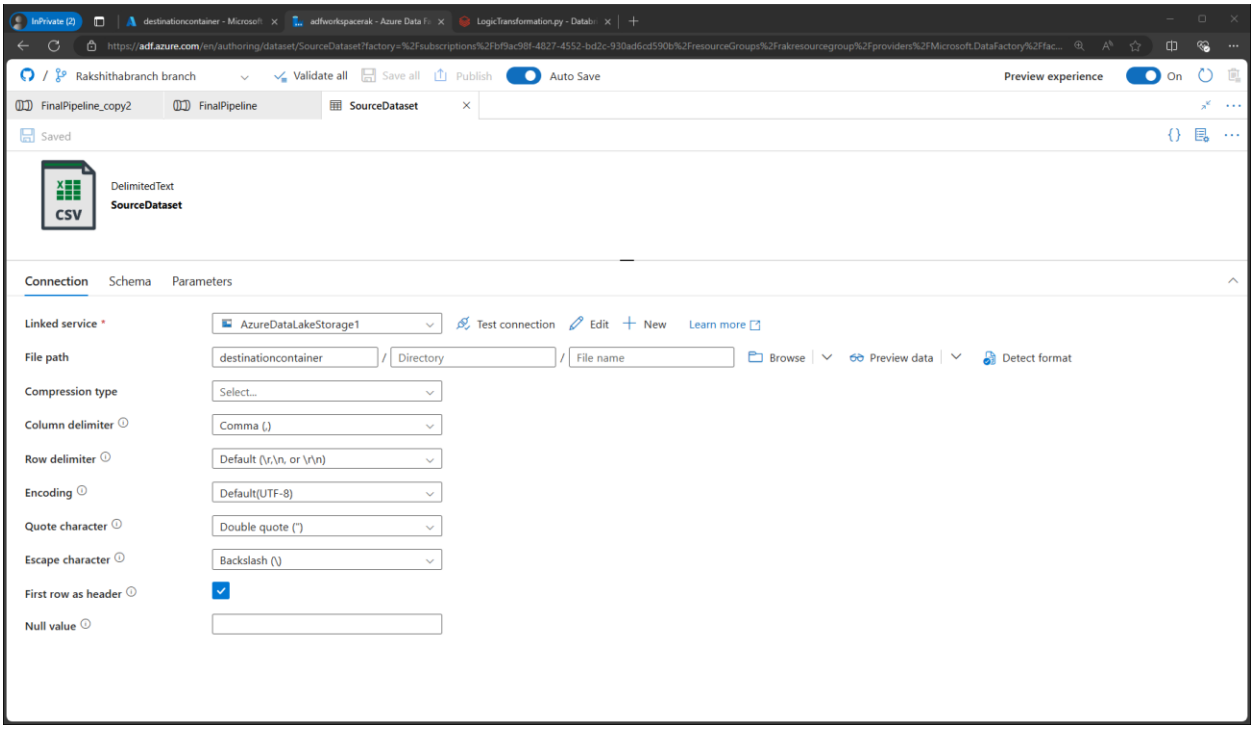

Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
FinalPipeline_copy2
FinalPipeline
Saved
Save as template
Validate
Debug
Add trigger
Notebook
NotebookTask
Set variable
Set variable1
Copy data
CopyToFinal
General
Azure Databricks
Settings
User properties
Databricks linked service *
AzureDatabricks1
Test connection
Edit
New
Connection successful

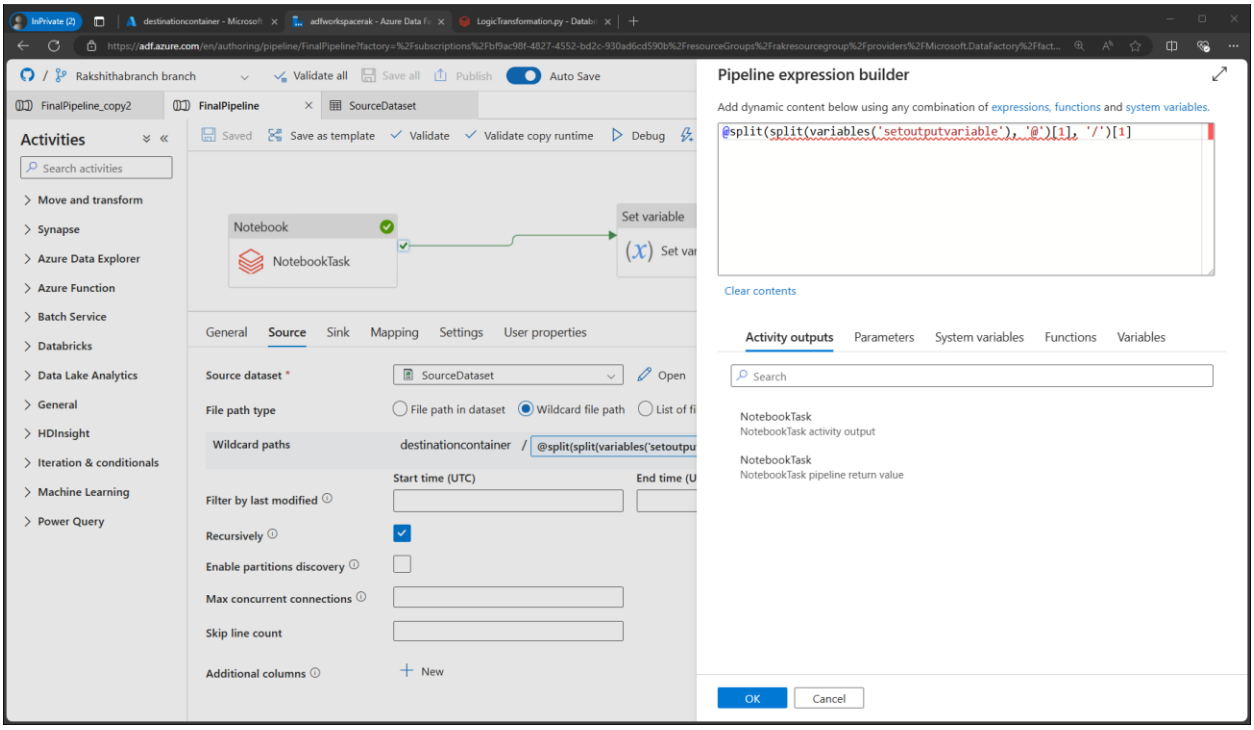

Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
FinalPipeline_copy2
FinalPipeline
Saved
Save as template
Validate
Debug
Add trigger
Notebook
NotebookTask
Set variable
Set variable1
Copy data
CopyToFinal
General
Azure Databricks
Settings
User properties
Notebook path *
/Users/keerthanaakkula@gmail.com/Logi...
Browse
Open
Base parameters
New
Delete
Name
Value
csvpath
@pipeline().parameters.inputfilepath
Append libraries

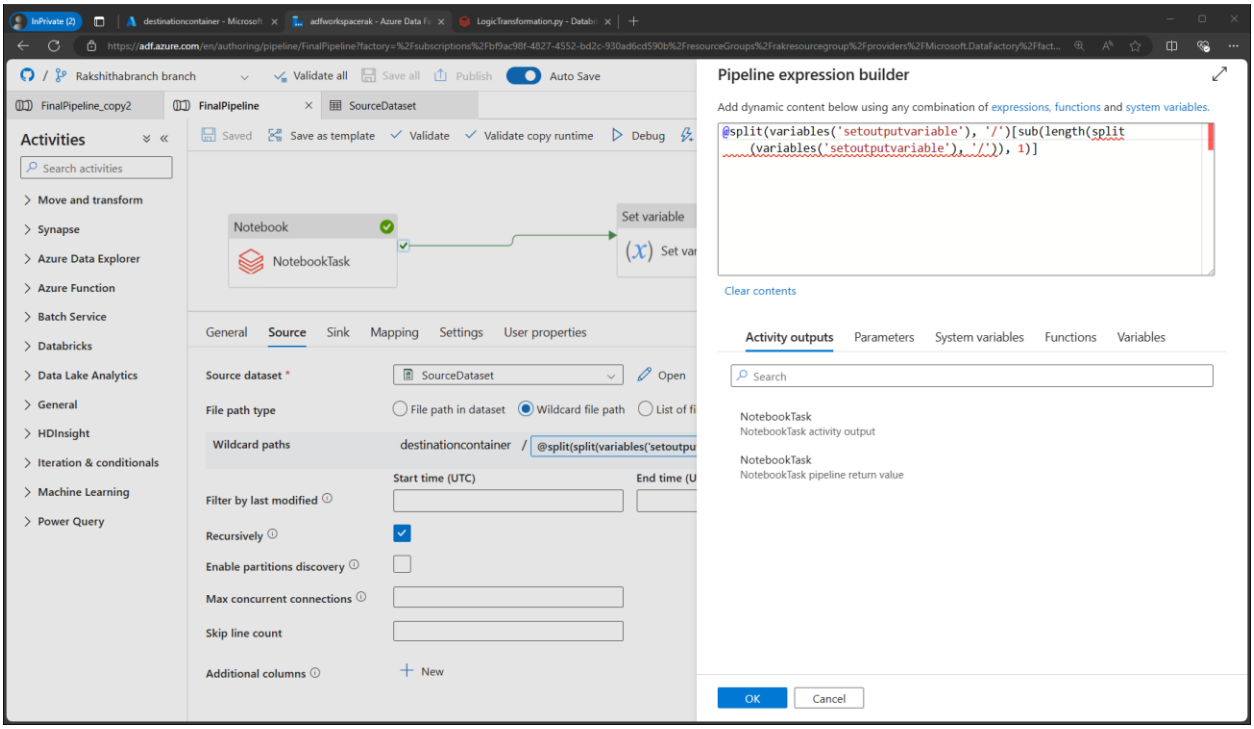

Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
FinalPipeline_copy2
FinalPipeline
Saved
Save as template
Validate
Debug
Add trigger
Notebook
NotebookTask
Set variable
Set variable1
General
Settings
User properties
Variable type
Pipeline variable
Pipeline return value
Name *
setoutputvariable
New
Value
@activity('NotebookTask').output.ru...
Pipeline expression builder
Add dynamic content below using any combination of expressions, functions and system variables.
@activity('NotebookTask').output.runOutput
Clear contents
Activity outputs
Parameters
System variables
Functions
Variables
Search
NotebookTask
NotebookTask activity output
NotebookTask
NotebookTask pipeline return value
OK
Cancel

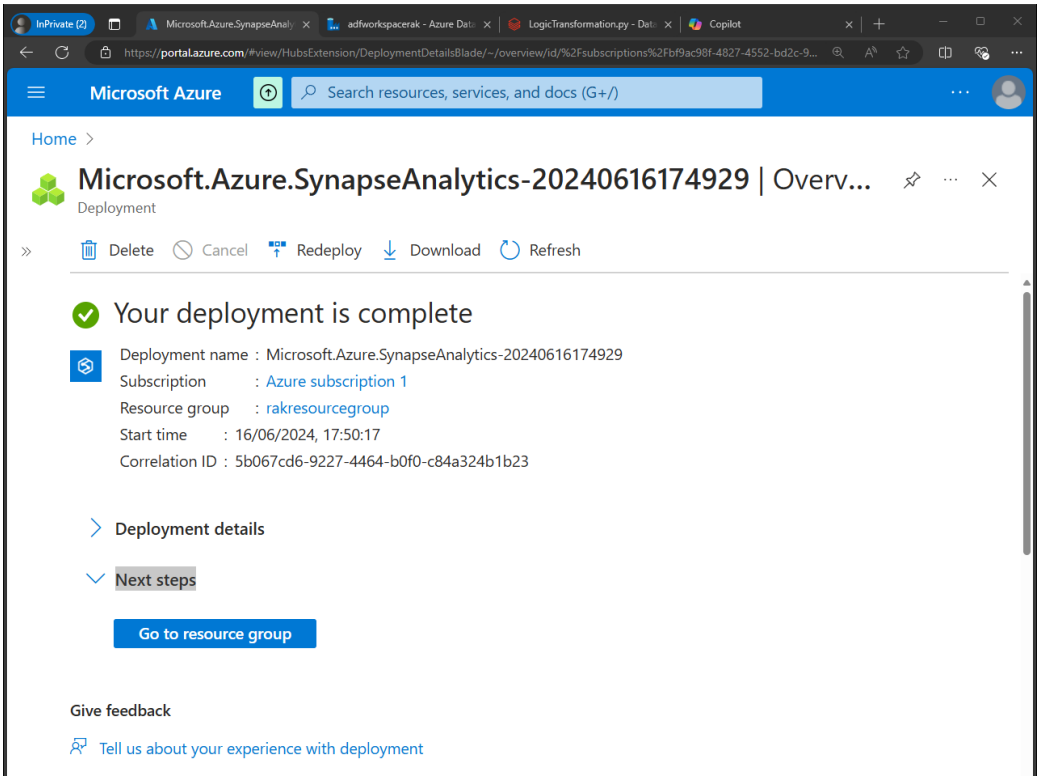

Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
FinalPipeline_copy2
FinalPipeline
Saved
Save as template
Validate
Validate copy runtime
Debug
Add trigger
Notebook
NotebookTask
Set variable
Set variable1
Copy data
CopyToFinal
General
Source
Sink
Mapping
Settings
User properties
Name *
CopyToFinal
Learn more
Description
Activity state
Activated
Deactivated
Timeout
0.12:00:00
Retry
0
Retry interval (sec)
30
Secure output
Secure input

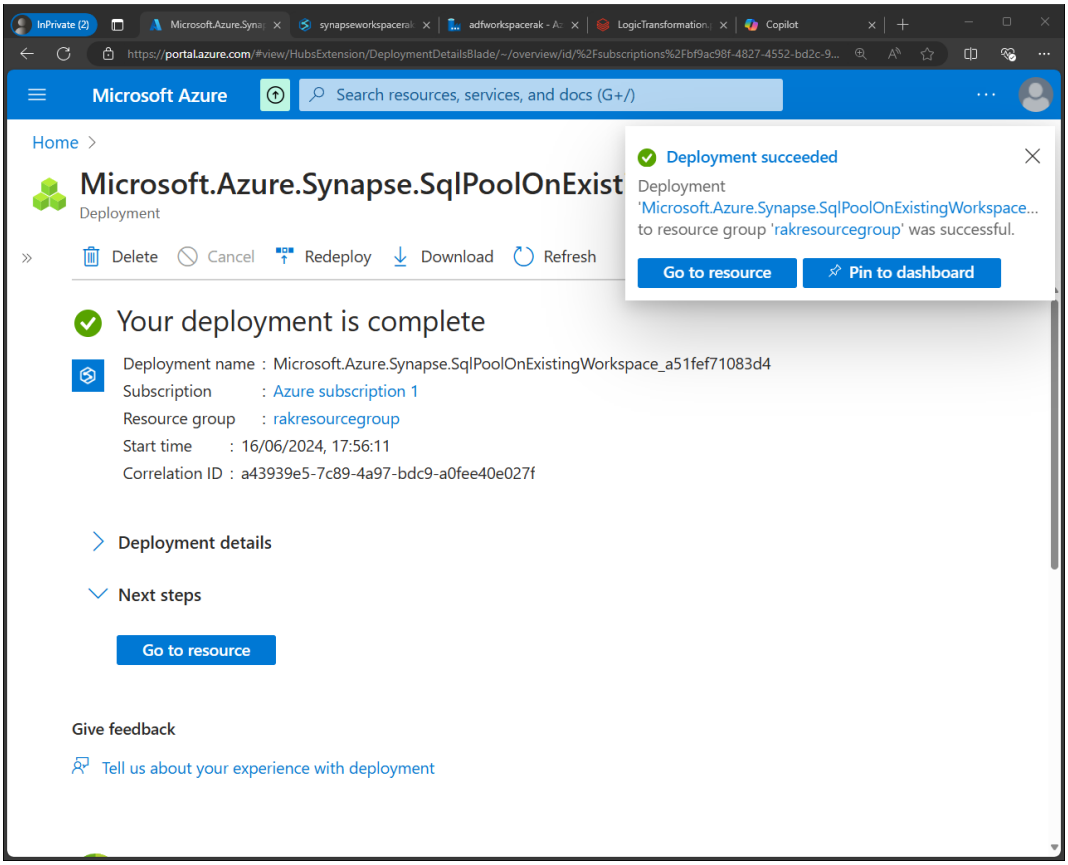

Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
FinalPipeline_copy2
FinalPipeline
Saved
Save as template
Validate
Validate copy runtime
Debug
Add trigger
Notebook
NotebookTask
Set variable
Set variable1
Copy data
CopyToFinal
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
SourceDataset
Open
New
Preview data
Learn more
File path type
File path in dataset
Wildcard file path
List of files
Wildcard paths
destinationcontainer
@split(split(variables('setoutputvaria...
@split(variables('setoutputvariable'), ...
Start time (UTC)
End time (UTC)
Filter by last modified
Recursively
Enable partitions discovery
Max concurrent connections
Skip line count
Additional columns
New


Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
FinalPipeline_copy2
FinalPipeline
SourceDataset
Saved
CSV
DelimitedText
SourceDataset
Connection
Schema
Parameters
Linked service *
AzureDataLakeStorage1
Test connection
Edit
New
Learn more
File path
destinationcontainer
Directory
File name
Browse
Preview data
Detect format
Compression type
Select...
Column delimiter
Comma (,)
Row delimiter
Default (\r,\n, or \r\n)
Encoding
Default(UTF-8)
Quote character
Double quote (")
Escape character
Backslash (\)
First row as header
Null value

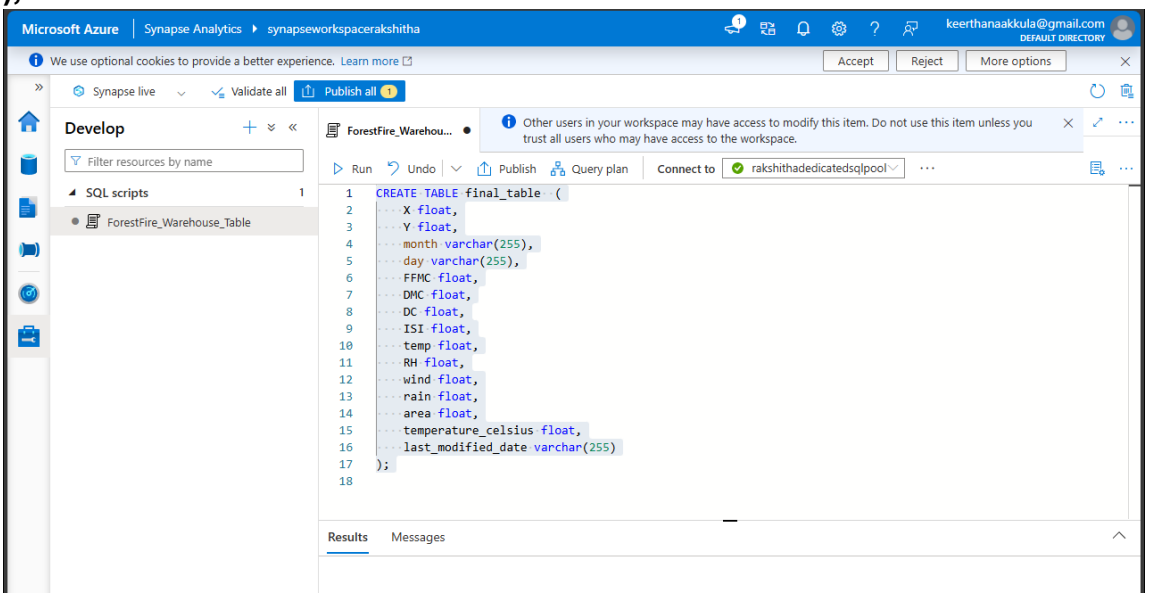

Pipeline expression builder
Add dynamic content below using any combination of expressions, functions and system variables.
@split(split(variables('setoutputvariable'), '@')[1], '/')[1]
Clear contents
Activity outputs
Parameters
System variables
Functions
Variables
Search
NotebookTask
NotebookTask activity output
NotebookTask
NotebookTask pipeline return value
OK
Cancel

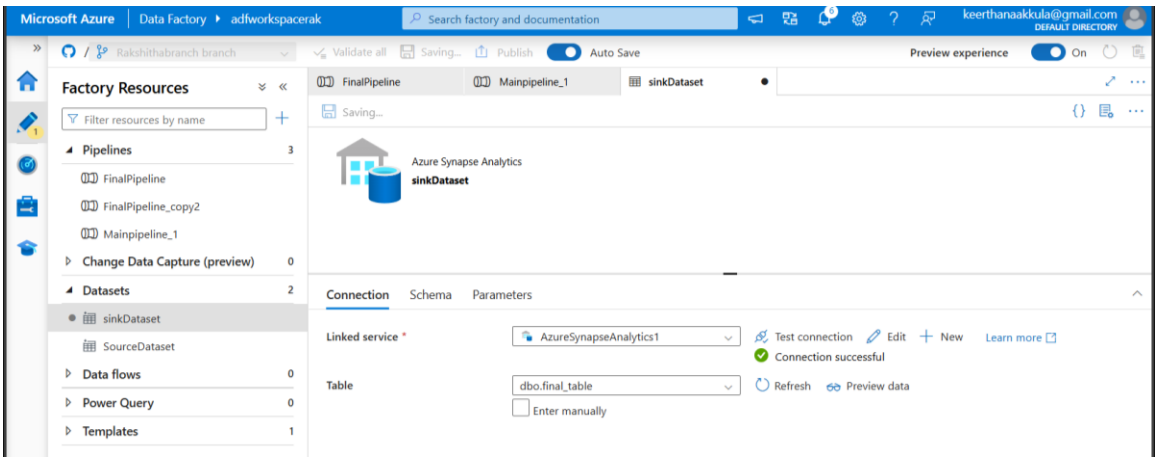

Pipeline expression builder
Add dynamic content below using any combination of expressions, functions and system variables.
@split(variables('setoutputvariable'), '/')[sub(length(split
(variables('setoutputvariable'), '/')), 1)]
Clear contents
Activity outputs
Parameters
System variables
Functions
Variables
Search
NotebookTask
NotebookTask activity output
NotebookTask
NotebookTask pipeline return value
OK
Cancel

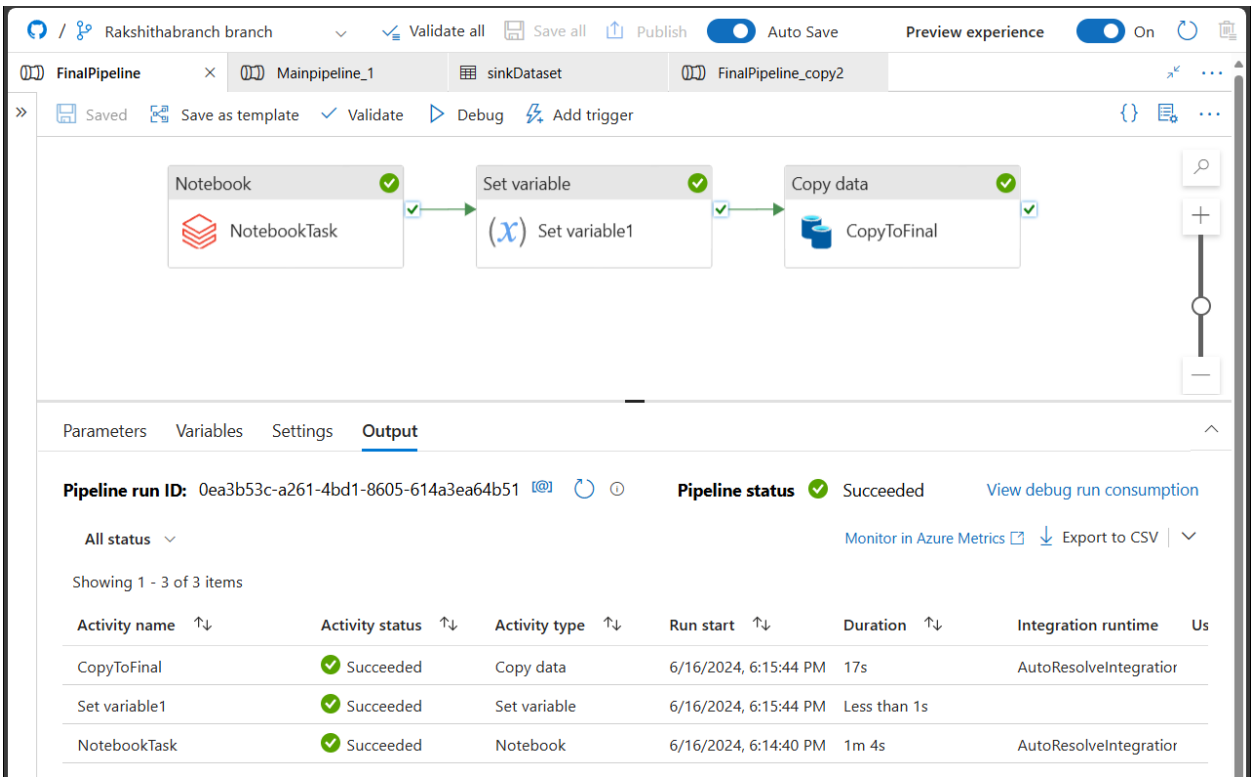

Microsoft Azure
Search resources, services, and docs (G+/)
Home
Microsoft.Azure.SynapseAnalytics-20240616174929 | Overv...
Deployment
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name : Microsoft.Azure.SynapseAnalytics-20240616174929
Subscription : Azure subscription 1
Resource group : rakresourcegroup
Start time : 16/06/2024, 17:50:17
Correlation ID : 5b067cd6-9227-4464-b0f0-c84a324b1b23
Deployment details
Next steps
Go to resource group
Give feedback
Tell us about your experience with deployment

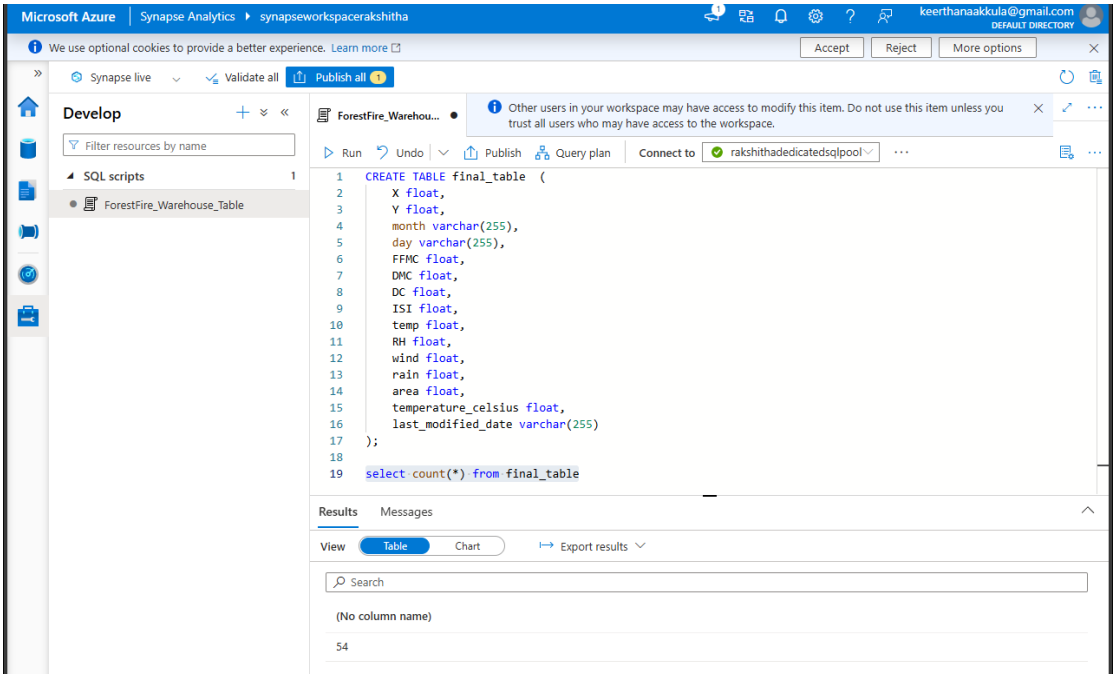

Microsoft Azure
Search resources, services, and docs (G+/)
Home
Microsoft.Azure.Synapse.SqlPoolOnExist
Deployment
Deployment succeeded
Deployment 'Microsoft.Azure.Synapse.SqlPoolOnExistingWorkspace... to resource group 'rakresourcegroup' was successful.
Go to resource
Pin to dashboard
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name : Microsoft.Azure.Synapse.SqlPoolOnExistingWorkspace_a51fef71083d4
Subscription : Azure subscription 1
Resource group : rakresourcegroup
Start time : 16/06/2024, 17:56:11
Correlation ID : a43939e5-7c89-4a97-bdc9-a0fee40e027f
Deployment details
Next steps
Go to resource
Give feedback
Tell us about your experience with deployment

```
CREATE TABLE your_table_name (
   X float,
   Y float,
   month varchar(255),
   day varchar(255),
   FFMC float,
   DMC float,
   DC float,
   ISI float,
   temp float,
   RH float,
   wind float,
   rain float,
   area float,
   temperature_celsius float,
   last_modified_date varchar (255)
);
```

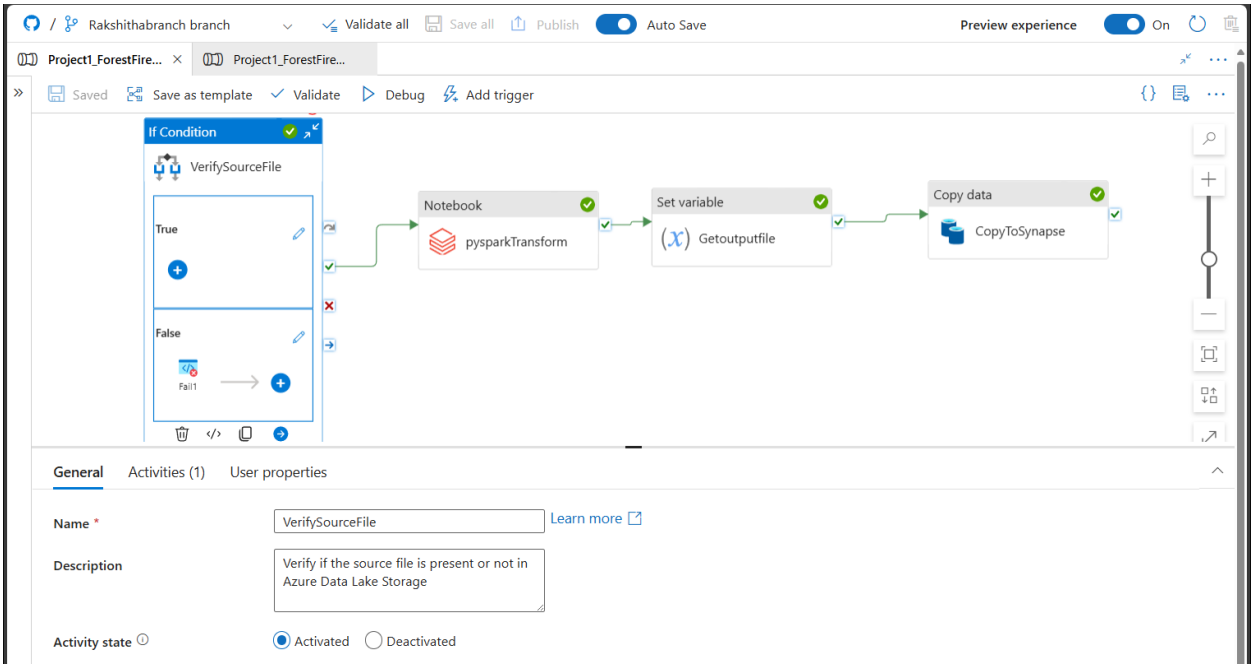


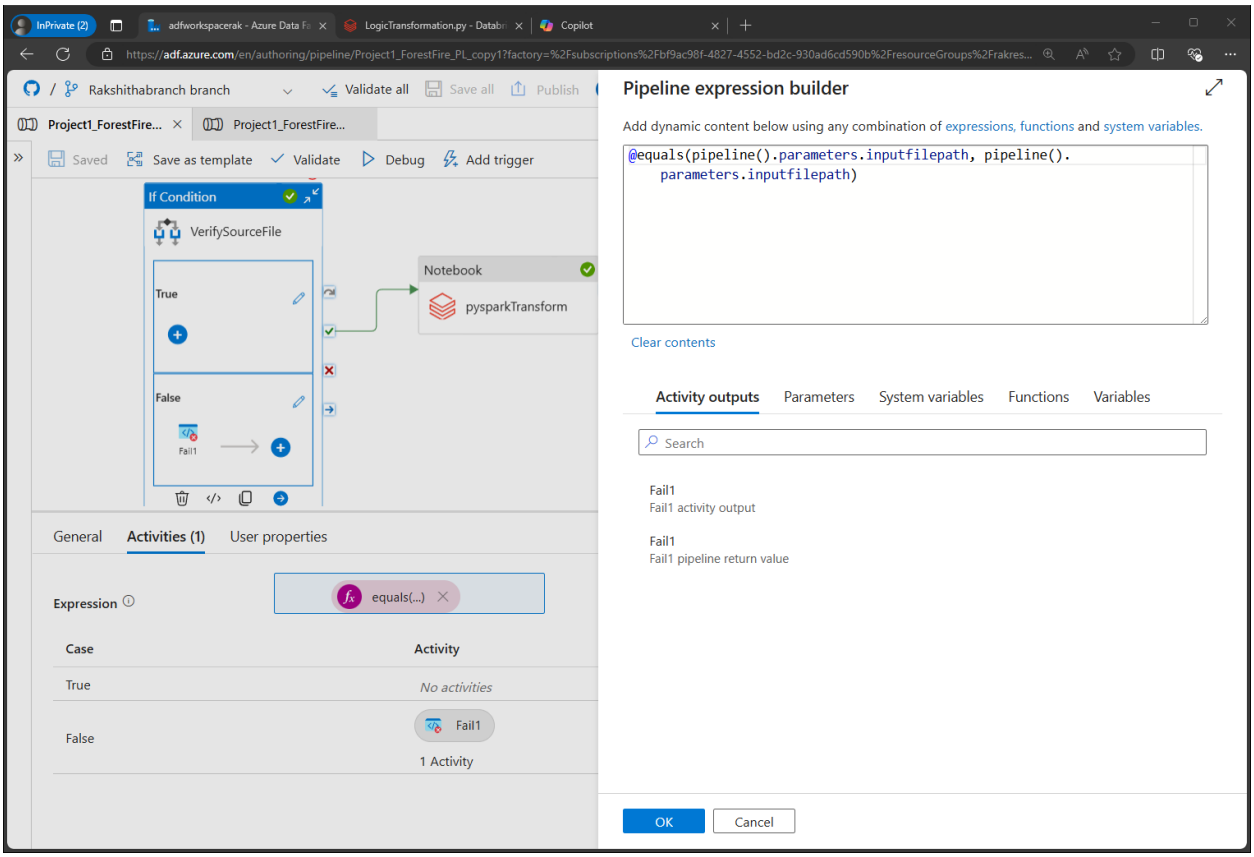


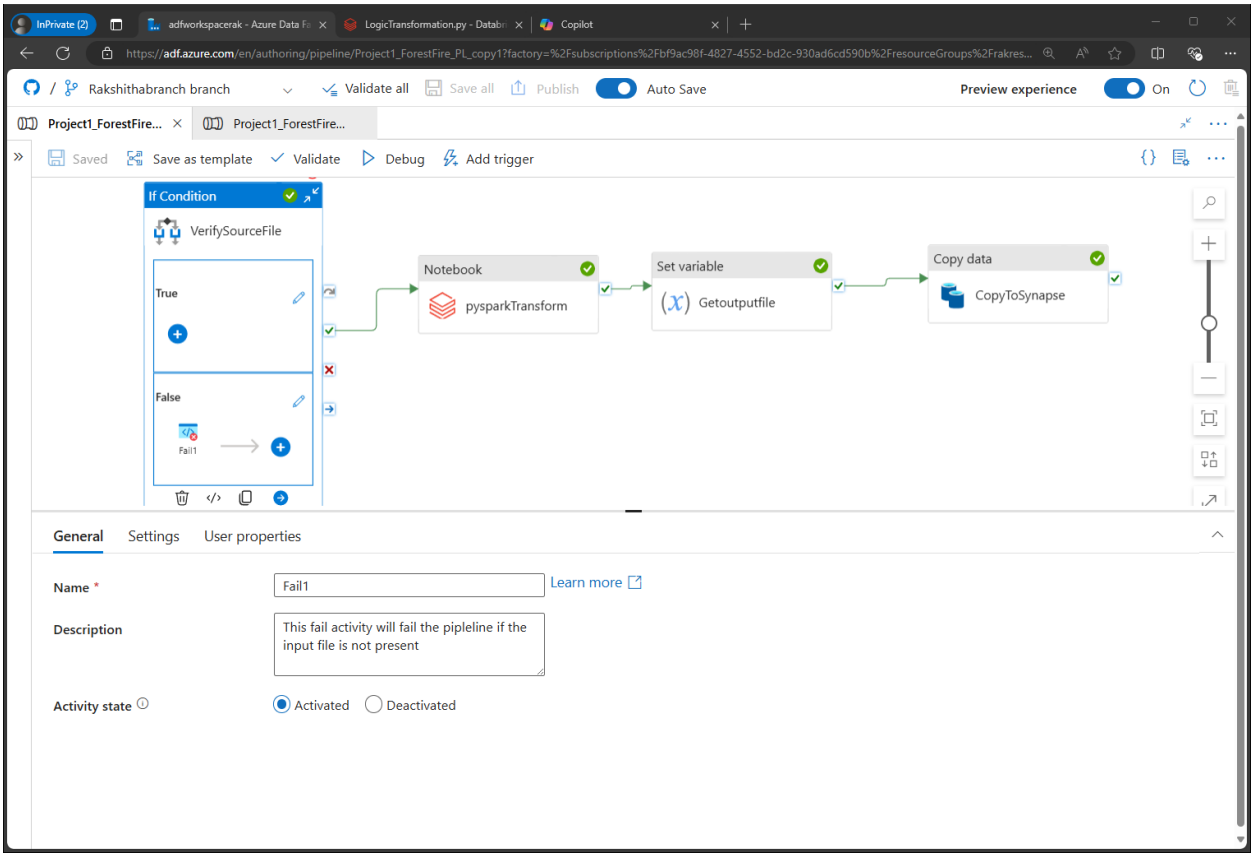


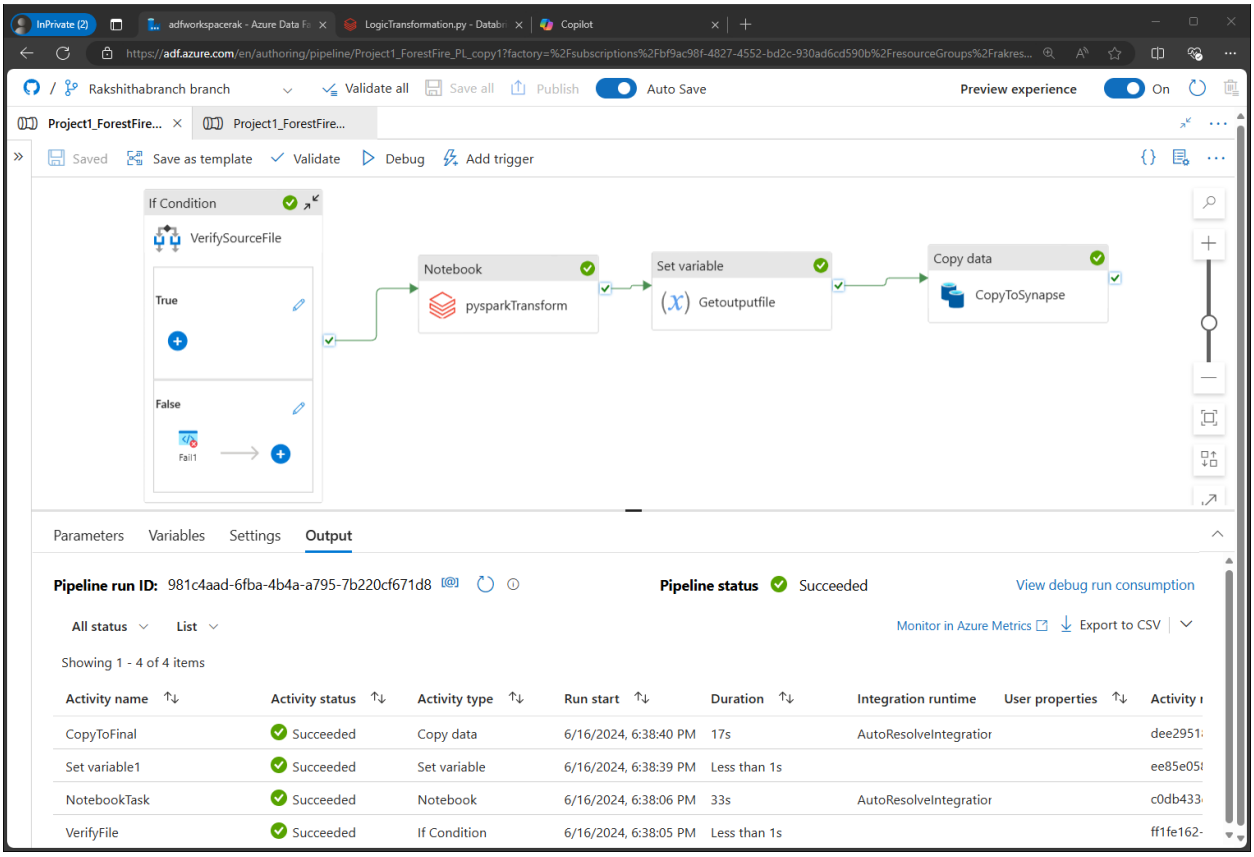


Added if-else condition to check if the source file is present or not

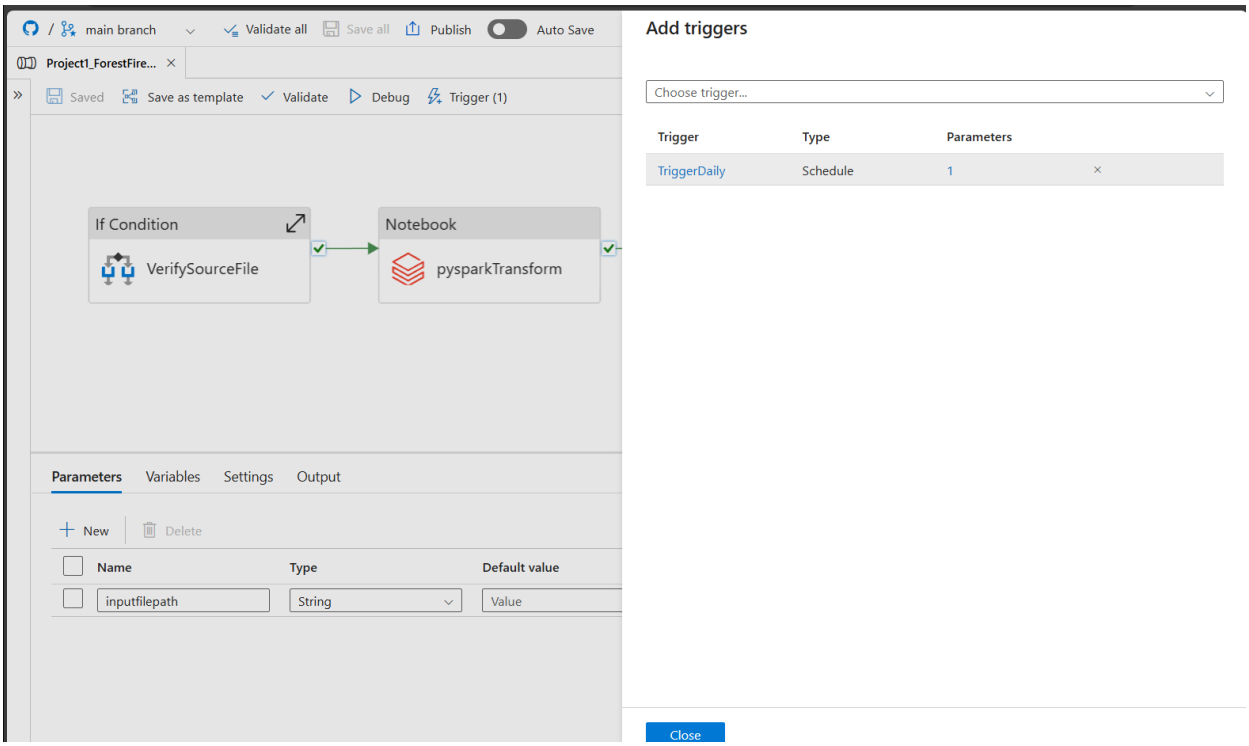

Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
Project1_ForestFire...
Project1_ForestFire...
Saved
Save as template
Validate
Debug
Add trigger
If Condition
VerifySourceFile
True
False
Fail1
Notebook
pysparkTransform
Set variable
Getoutputfile
Copy data
CopyToSynapse
General
Activities (1)
User properties
Name *
VerifySourceFile
Learn more
Description
Verify if the source file is present or not in Azure Data Lake Storage
Activity state
Activated
Deactivated

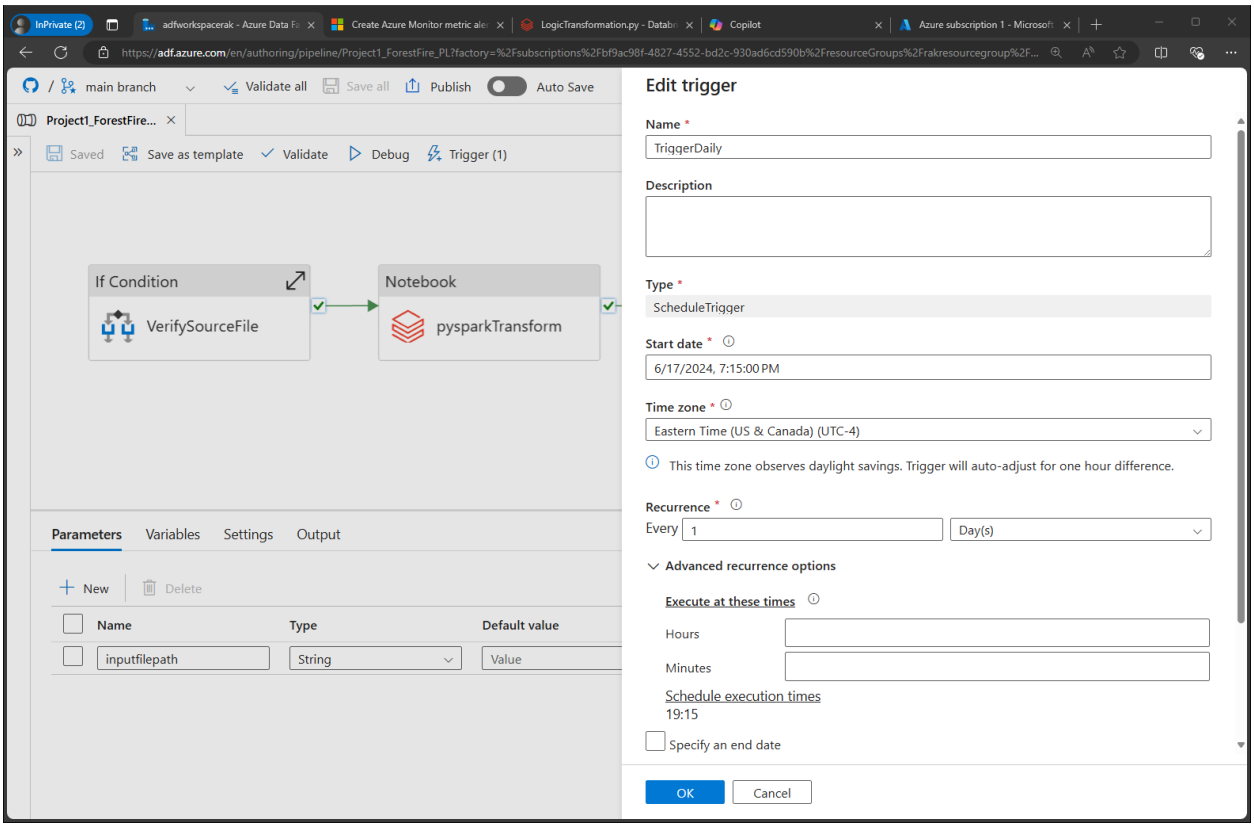

InPrivate (2)
adfworkspacerak - Azure Data Fa
LogicTransformation.py - Databr
Copilot
https://adf.azure.com/en/authoring/pipeline/Project1_ForestFire_PL_copy1?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-930ad6cd590b%2FresourceGroups%2Frakres...
Rakshithabranch branch
Validate all
Save all
Publish
Project1_ForestFire...
Project1_ForestFire...
Saved
Save as template
Validate
Debug
Add trigger
If Condition
VerifySourceFile
True
False
Fail1
Notebook
pysparkTransform
General
Activities (1)
User properties
Expression
equals(...)
Case
Activity
True
No activities
False
Fail1
1 Activity
Pipeline expression builder
Add dynamic content below using any combination of expressions, functions and system variables.
@equals(pipeline().parameters.inputfilepath, pipeline().
parameters.inputfilepath)
Clear contents
Activity outputs
Parameters
System variables
Functions
Variables
Search
Fail1
Fail1 activity output
Fail1
Fail1 pipeline return value
OK
Cancel

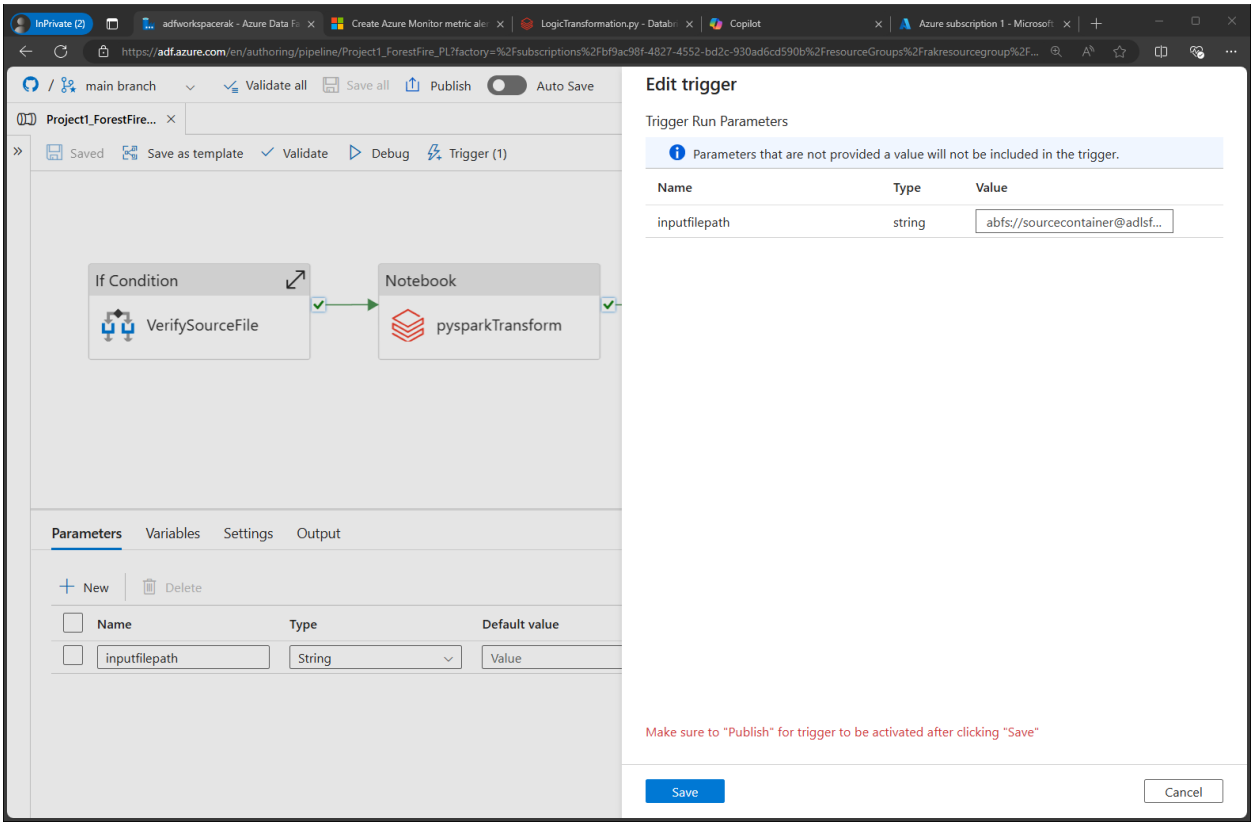

Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
Project1_ForestFire...
Saved
Save as template
Validate
Debug
Add trigger
If Condition
VerifySourceFile
True
False
Fail1
Notebook
pysparkTransform
Set variable
Getoutputfile
Copy data
CopyToSynapse
General
Settings
User properties
Name *
Fail1
Learn more
Description
This fail activity will fail the pipleline if the input file is not present
Activity state
Activated
Deactivated


Rakshithabranch branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
Project1_ForestFire...
Saved
Save as template
Validate
Debug
Add trigger
If Condition
VerifySourceFile
True
False
Fail1
Notebook
pysparkTransform
Set variable
Getoutputfile
Copy data
CopyToSynapse
Parameters
Variables
Settings
Output
Pipeline run ID: 981c4aad-6fba-4b4a-a795-7b220cf671d8
Pipeline status Succeeded
View debug run consumption
All status
List
Monitor in Azure Metrics
Export to CSV
Showing 1 - 4 of 4 items
Activity name
Activity status
Activity type
Run start
Duration
Integration runtime
User properties
CopyToFinal
Succeeded
Copy data
6/16/2024, 6:38:40 PM
17s
AutoResolveIntegratior
Set variable1
Succeeded
Set variable
6/16/2024, 6:38:39 PM
Less than 1s
NotebookTask
Succeeded
Notebook
6/16/2024, 6:38:06 PM
33s
AutoResolveIntegratior
VerifyFile
Succeeded
If Condition
6/16/2024, 6:38:05 PM
Less than 1s

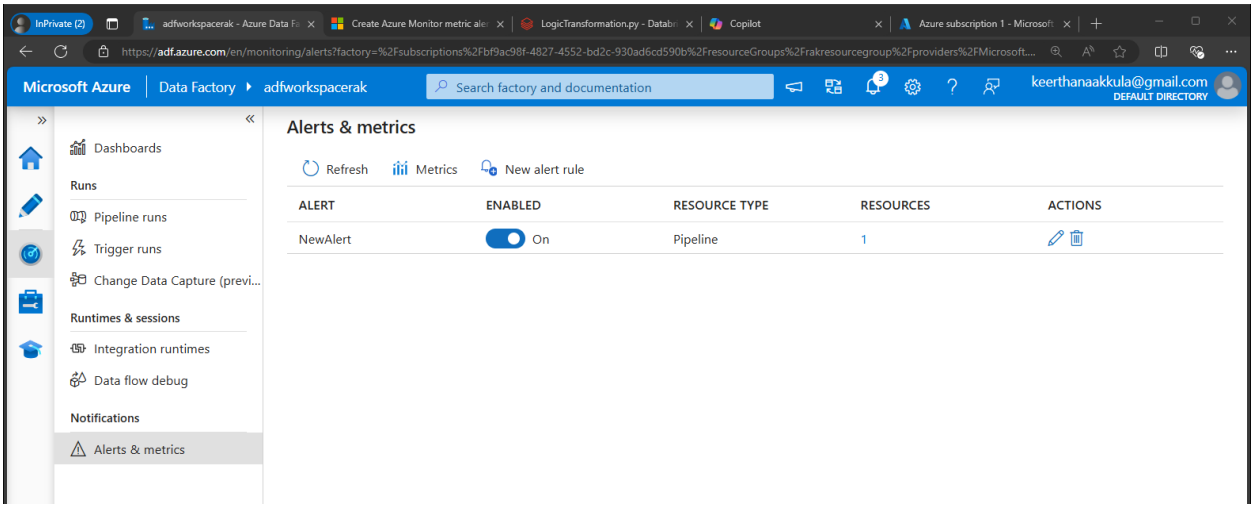
Adding schedule and alerts

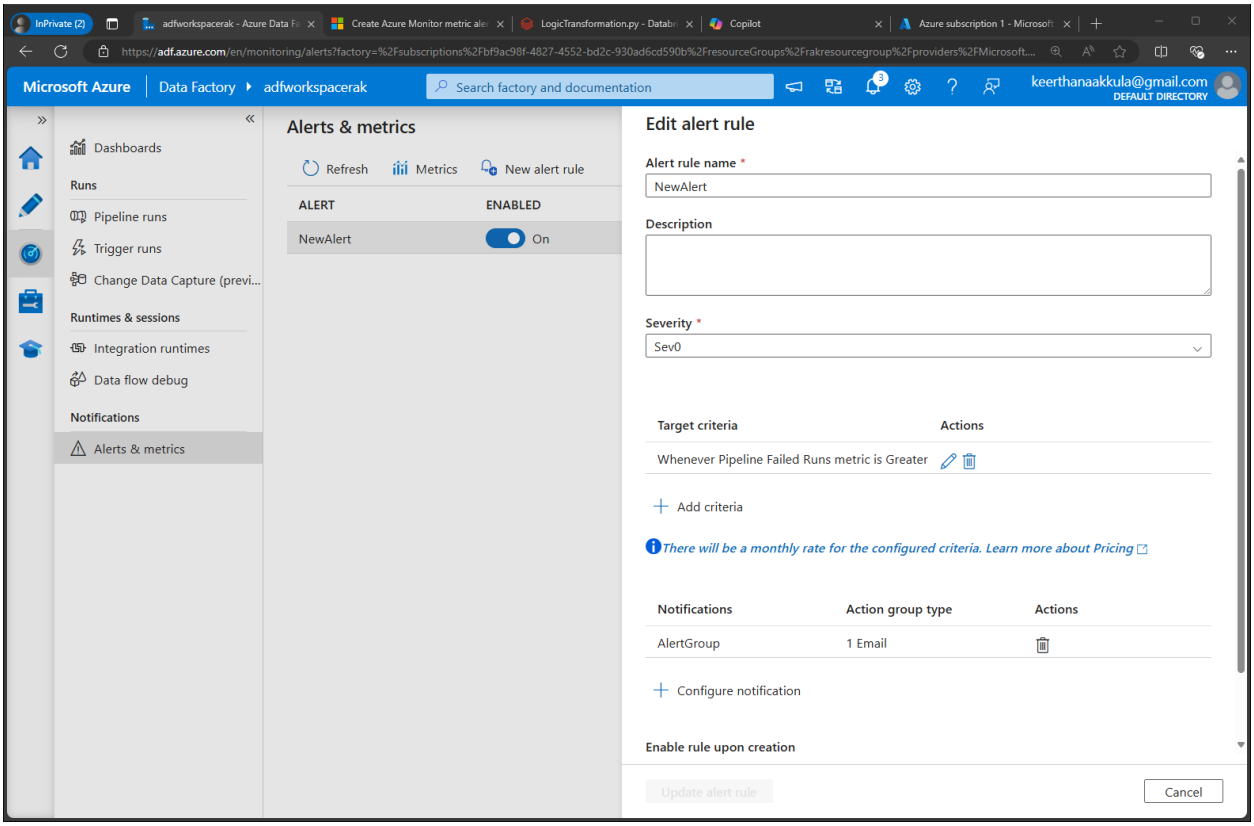


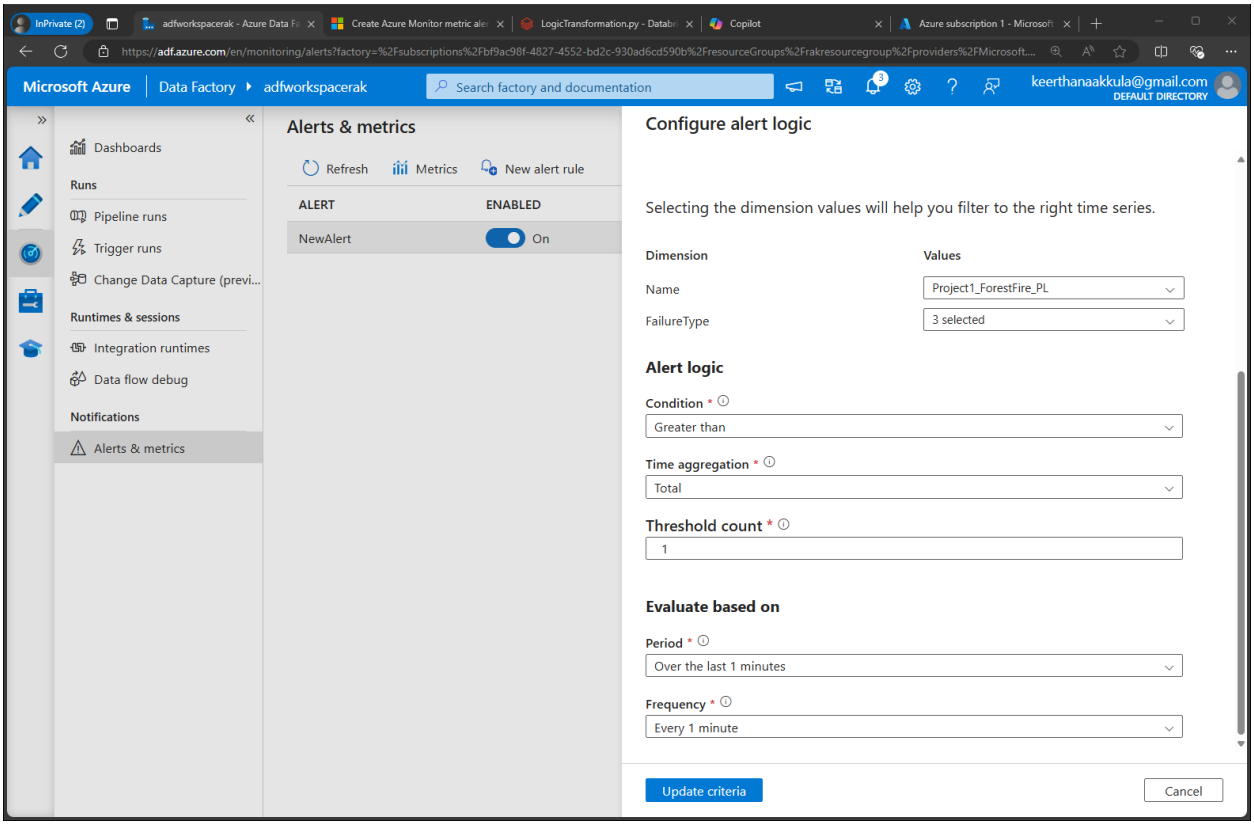



main branch
Validate all
Save all
Publish
Auto Save
Project1_ForestFire...
Saved
Save as template
Validate
Debug
Trigger (1)
If Condition
VerifySourceFile
Notebook
pysparkTransform
Parameters
Variables
Settings
Output
New
Delete
Name
Type
Default value
inputfilepath
String
Value
Edit trigger
Trigger Run Parameters
Parameters that are not provided a value will not be included in the trigger.
Name
Type
Value
inputfilepath
string
abfs://sourcecontainer@adlsf...
Make sure to "Publish" for trigger to be activated after clicking "Save"
Save
Cancel

Alerts


Microsoft Azure
Data Factory
adfworkspacerak
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Dashboards
Runs
Pipeline runs
Trigger runs
Change Data Capture (previ...
Runtimes & sessions
Integration runtimes
Data flow debug
Notifications
Alerts & metrics
Alerts & metrics
Refresh
Metrics
New alert rule
ALERT
ENABLED
RESOURCE TYPE
RESOURCES
ACTIONS
NewAlert
On
Pipeline
1


Microsoft Azure
Data Factory
adfworkspacerak
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Alerts & metrics
Refresh
Metrics
New alert rule
ALERT
ENABLED
NewAlert
On
Dashboards
Runs
Pipeline runs
Trigger runs
Change Data Capture (previ...
Runtimes & sessions
Integration runtimes
Data flow debug
Notifications
Alerts & metrics
Edit alert rule
Alert rule name *
NewAlert
Description
Severity *
Sev0
Target criteria
Actions
Whenever Pipeline Failed Runs metric is Greater
Add criteria
There will be a monthly rate for the configured criteria. Learn more about Pricing
Notifications
Action group type
Actions
AlertGroup
1 Email
Configure notification
Enable rule upon creation
Update alert rule
Cancel

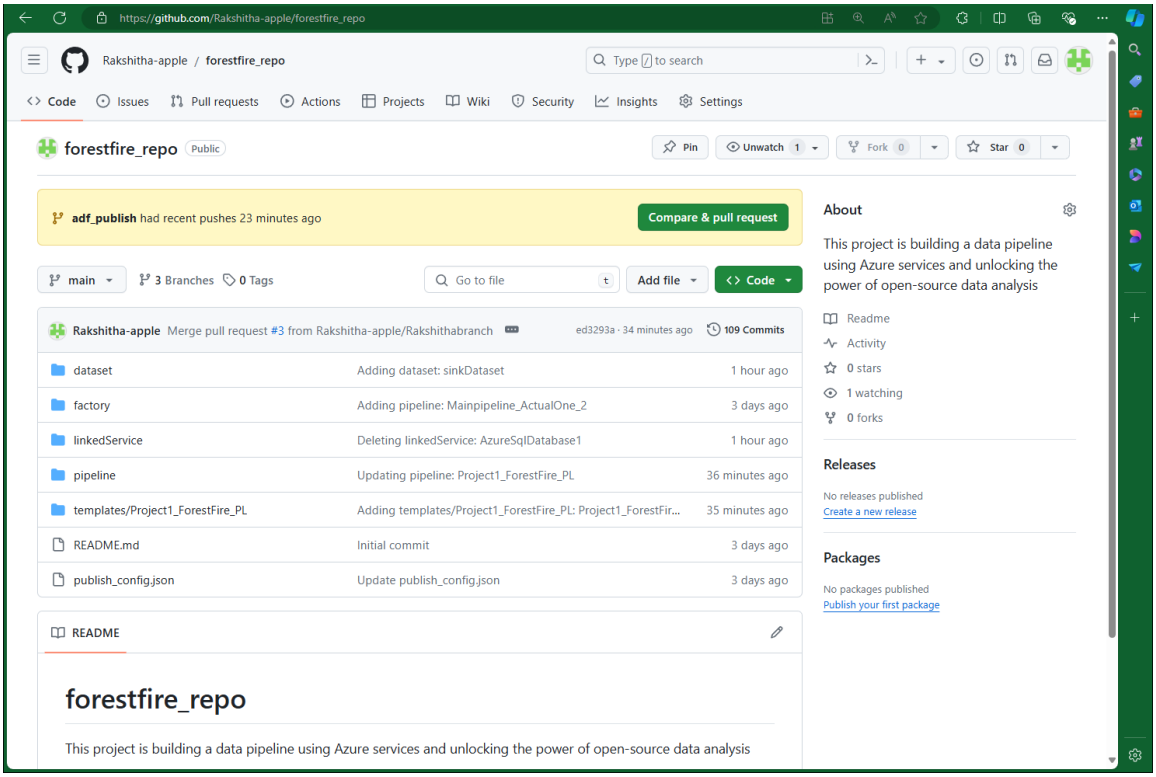

Microsoft Azure
Data Factory
adfworkspacerak
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Alerts & metrics
Refresh
Metrics
New alert rule
ALERT
ENABLED
NewAlert
On
Dashboards
Runs
Pipeline runs
Trigger runs
Change Data Capture (previ...
Runtimes & sessions
Integration runtimes
Data flow debug
Notifications
Alerts & metrics
Configure alert logic
Selecting the dimension values will help you filter to the right time series.
Dimension
Values
Name
Project1_ForestFire_PL
FailureType
3 selected
Alert logic
Condition *
Greater than
Time aggregation *
Total
Threshold count *
1
Evaluate based on
Period *
Over the last 1 minutes
Frequency *
Every 1 minute
Update criteria
Cancel

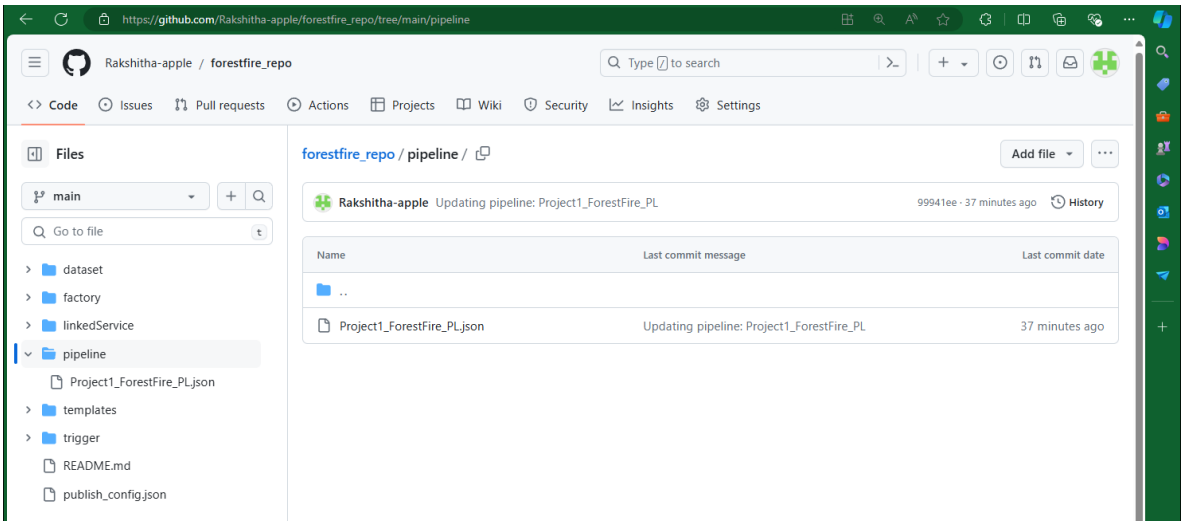
Notifications	Action group type	Actions
AlertGroup	1 Email	

+ Configure notification

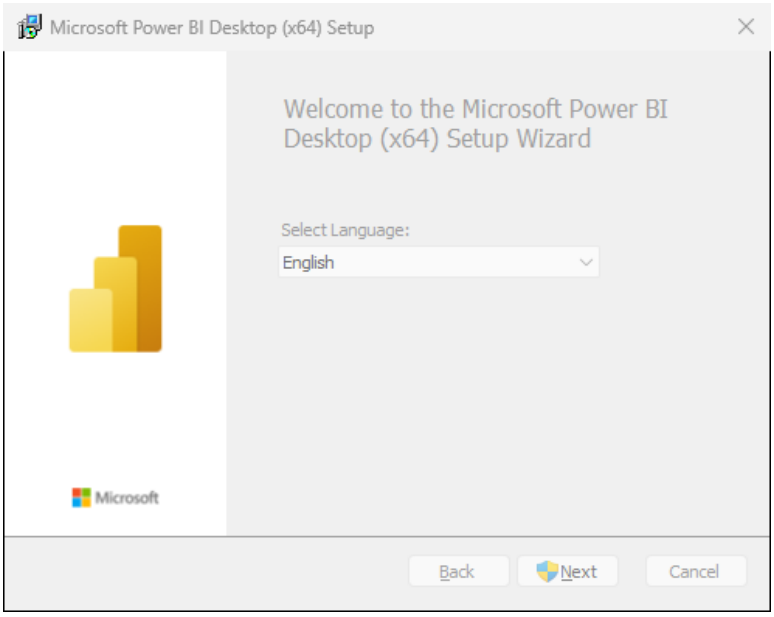
GitHub repo URL

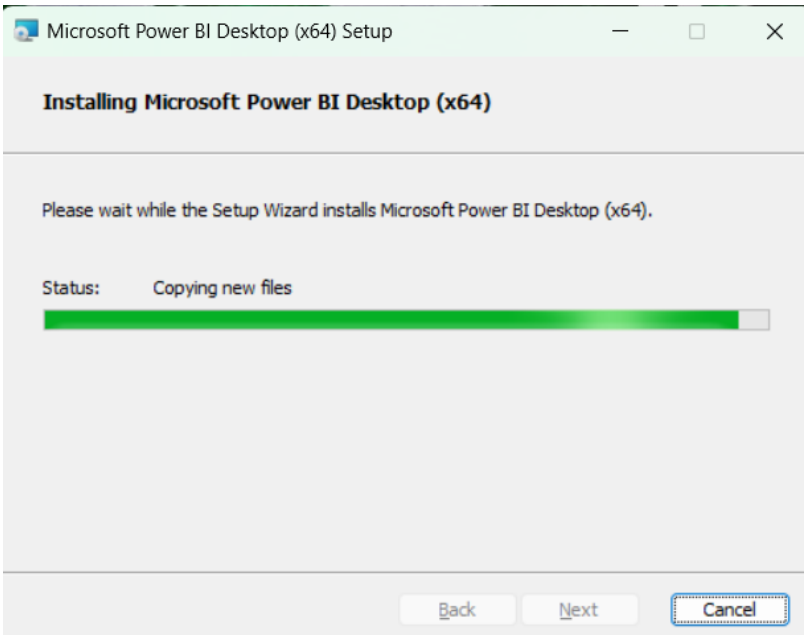
https://github.com/Rakshitha-apple/ETL_Pipleline_repo.git

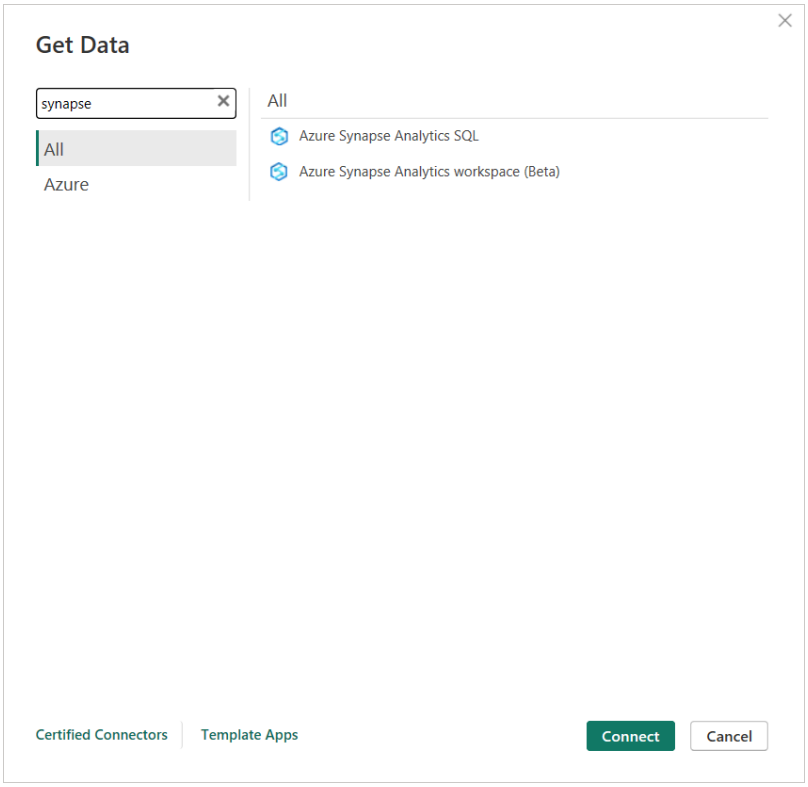


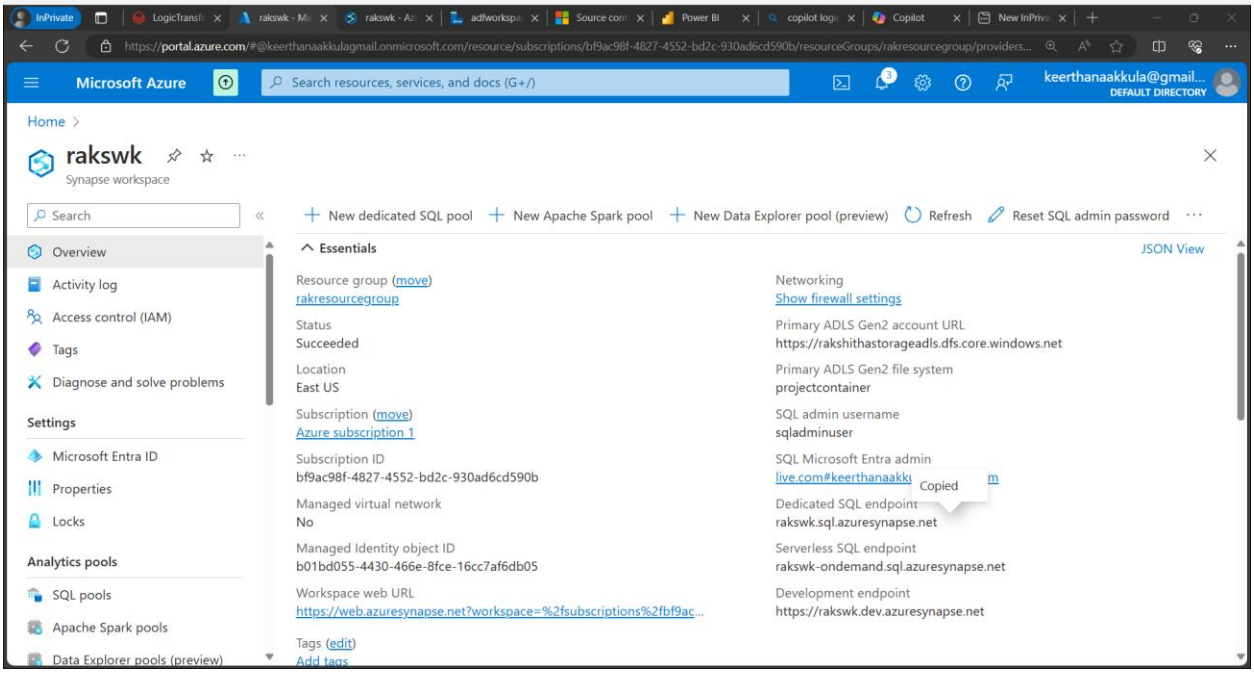


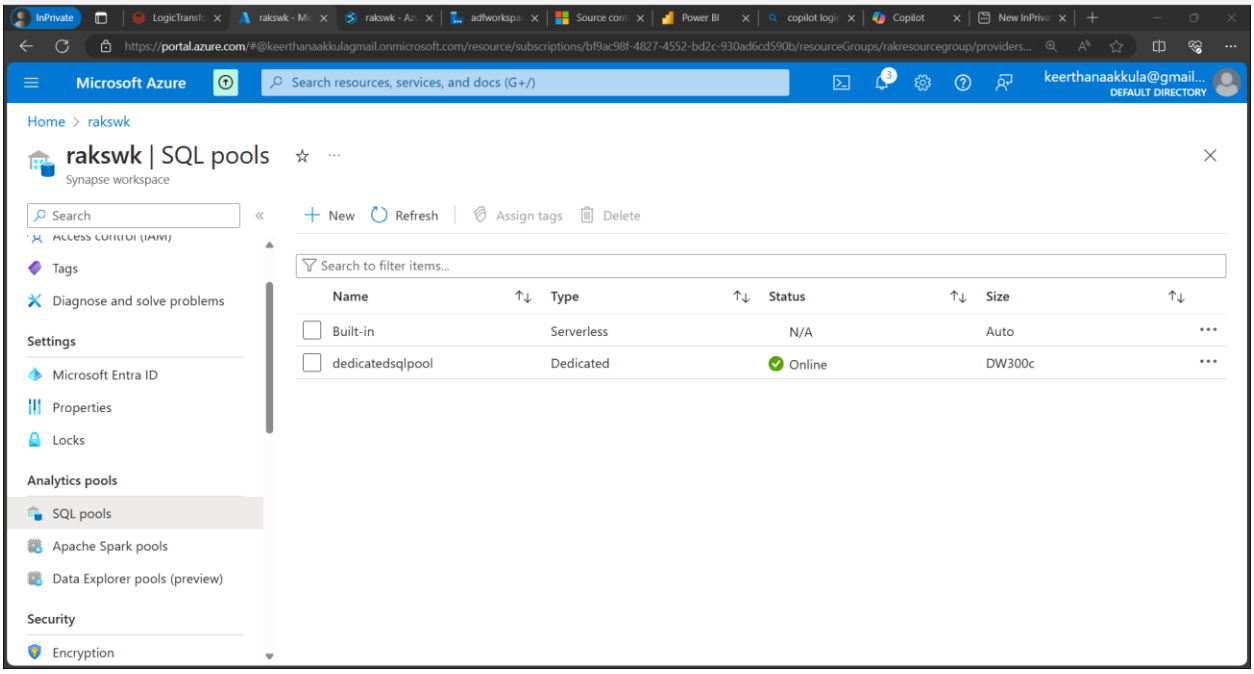
Visualize results and prepare a dashboard in Power BI.

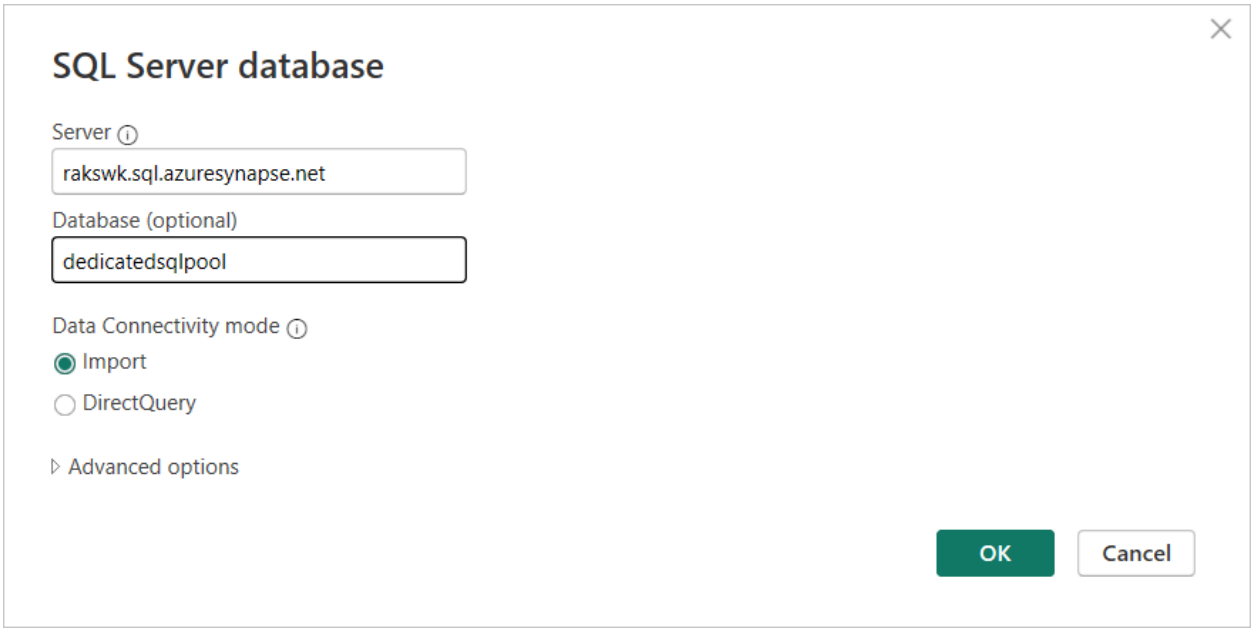


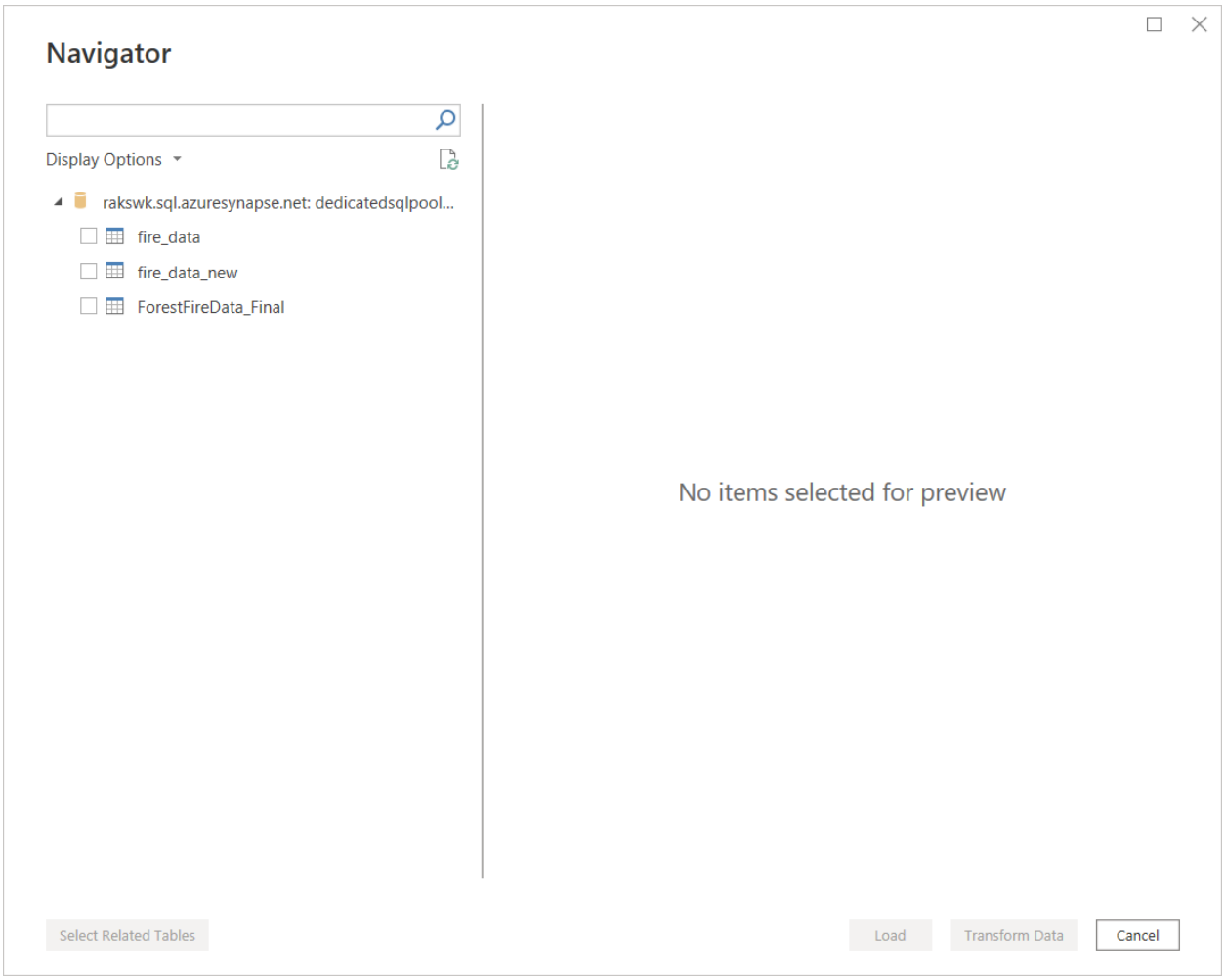


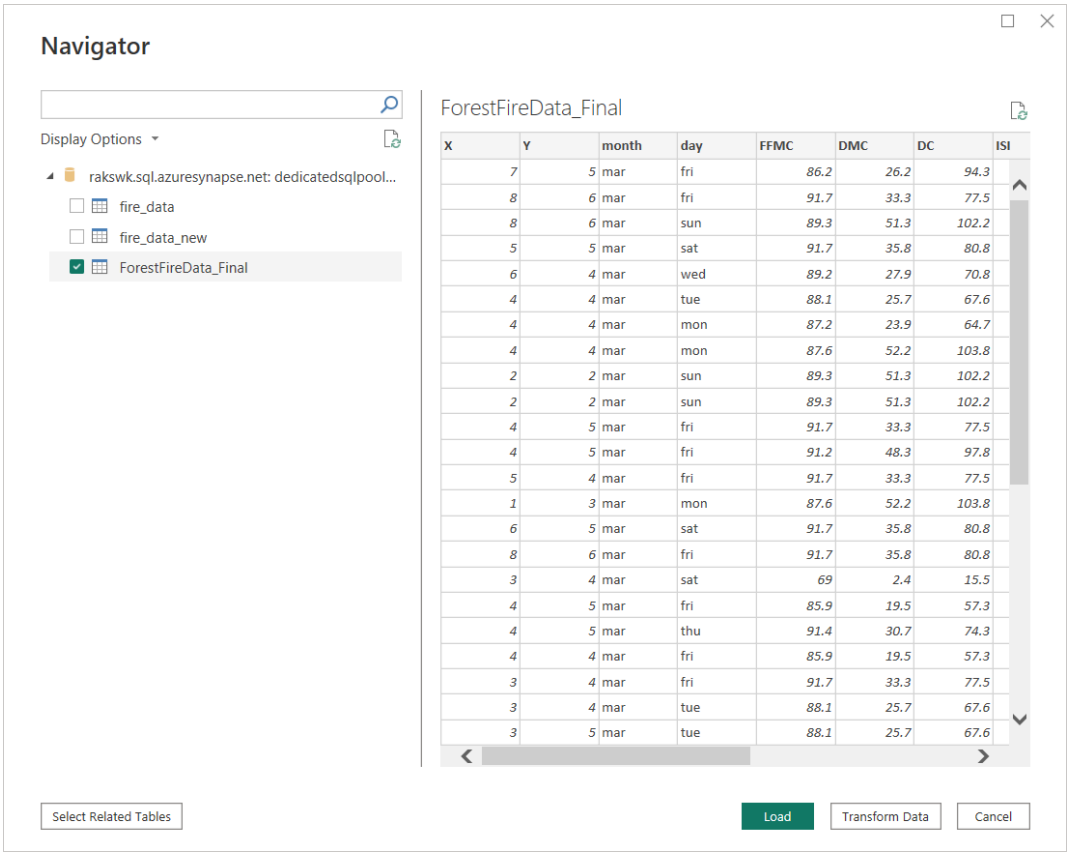

Get Data
synapse
All
Azure
All
Azure Synapse Analytics SQL
Azure Synapse Analytics workspace (Beta)
Certified Connectors
Template Apps
Connect
Cancel

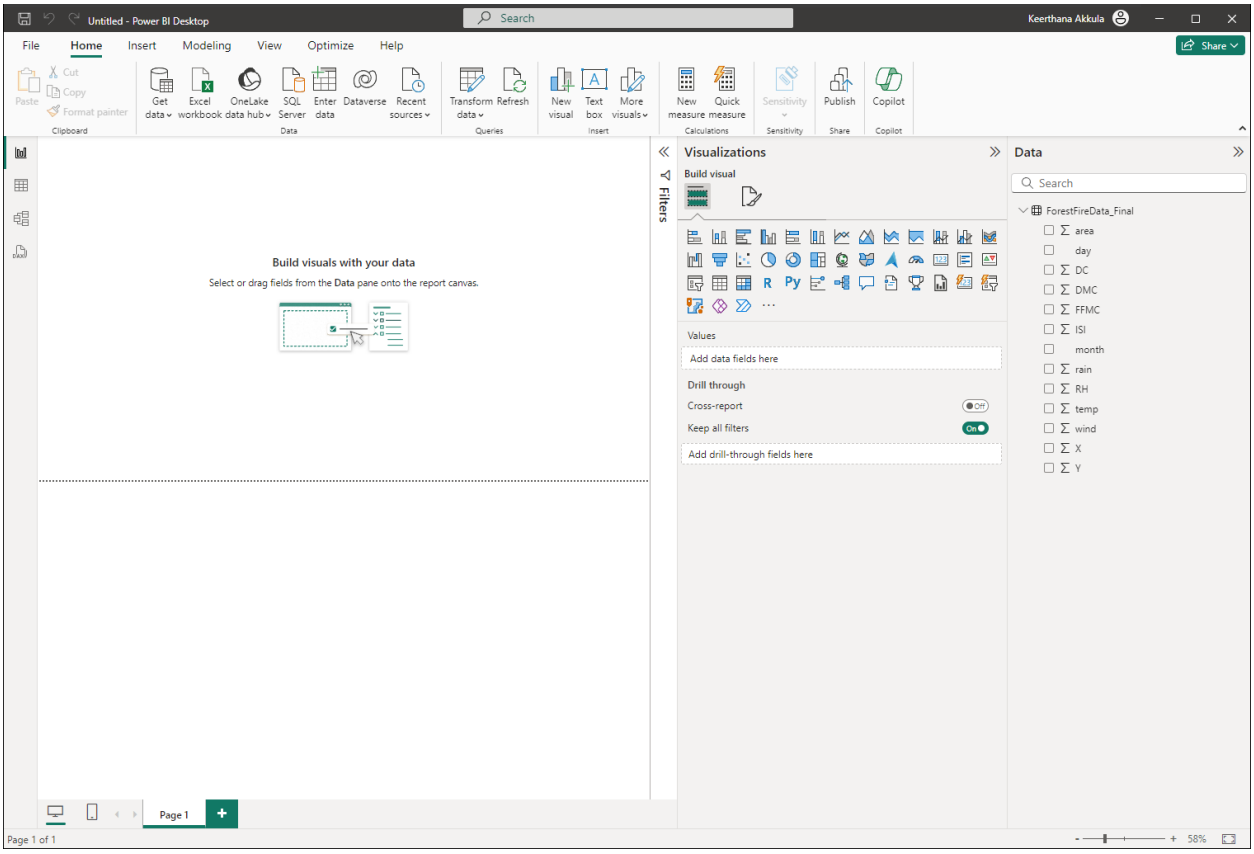

Microsoft Azure
Search resources, services, and docs (G+/)
keerthanaakkula@gmail...
DEFAULT DIRECTORY
Home
rakswk
Synapse workspace
Search
New dedicated SQL pool
New Apache Spark pool
New Data Explorer pool (preview)
Refresh
Reset SQL admin password
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Settings
Microsoft Entra ID
Properties
Locks
Analytics pools
SQL pools
Apache Spark pools
Data Explorer pools (preview)
Essentials
JSON View
Resource group (move)
rakresourcegroup
Status
Succeeded
Location
East US
Subscription (move)
Azure subscription 1
Subscription ID
bf9ac98f-4827-4552-bd2c-930ad6cd590b
Managed virtual network
No
Managed Identity object ID
b01bd055-4430-466e-8fce-16cc7af6db05
Workspace web URL
Networking
Show firewall settings
Primary ADLS Gen2 account URL
https://rakshithastorageadls.dfs.core.windows.net
Primary ADLS Gen2 file system
projectcontainer
SQL admin username
sqladminuser
SQL Microsoft Entra admin
Copied
Dedicated SQL endpoint
rakswk.sql.azuresynapse.net
Serverless SQL endpoint
rakswk-ondemand.sql.azuresynapse.net
Development endpoint
https://rakswk.dev.azuresynapse.net
Tags (edit)
Add tags

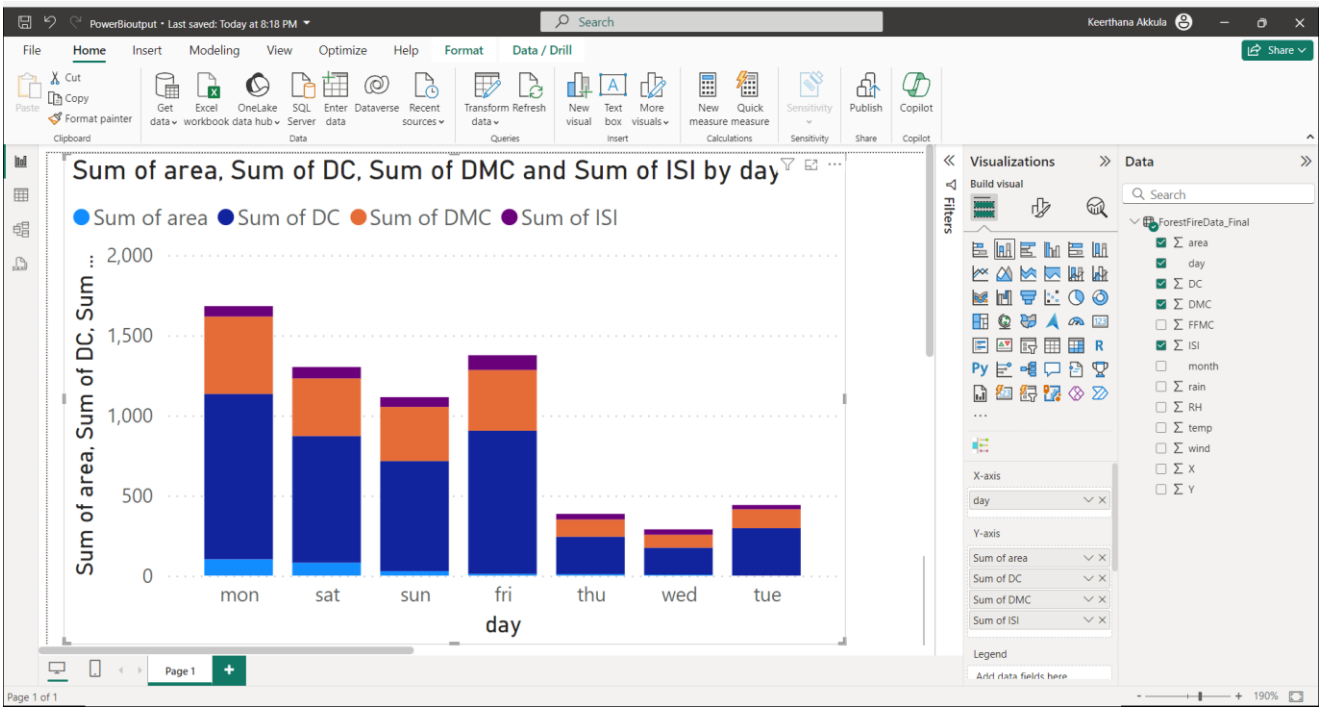

Microsoft Azure
Search resources, services, and docs (G+/)
keerthanaakkula@gmail...
DEFAULT DIRECTORY
Home > rakswk
rakswk | SQL pools
Synapse workspace
Search
New
Refresh
Assign tags
Delete
Search to filter items...
Name
Type
Status
Size
Built-in
Serverless
N/A
Auto
dedicatedsqlpool
Dedicated
Online
DW300c
Tags
Diagnose and solve problems
Settings
Microsoft Entra ID
Properties
Locks
Analytics pools
SQL pools
Apache Spark pools
Data Explorer pools (preview)
Security
Encryption

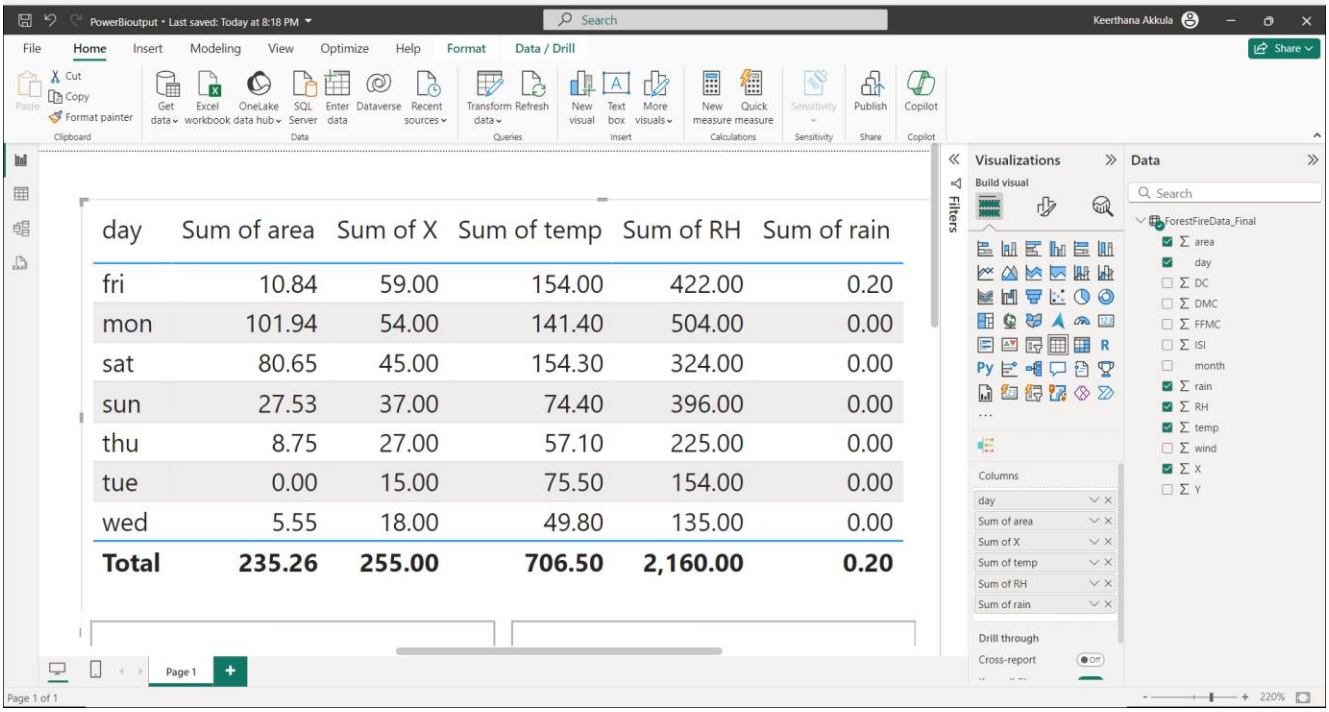

SQL Server database
Server
rakswk.sql.azuresynapse.net
Database (optional)
dedicatedsqlpool
Data Connectivity mode
Import
DirectQuery
Advanced options
OK
Cancel

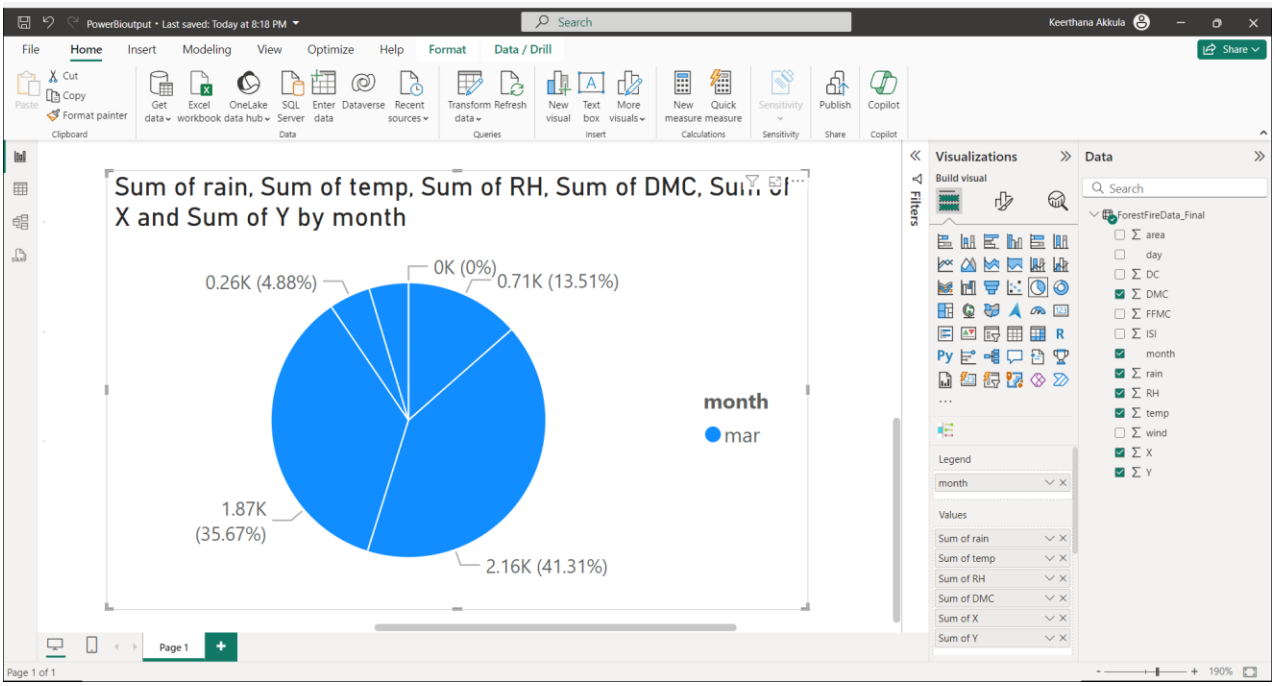

Navigator
Display Options
rakswk.sql.azuresynapse.net: dedicatedsqlpool...
fire_data
fire_data_new
ForestFireData_Final
No items selected for preview
Select Related Tables
Load
Transform Data
Cancel

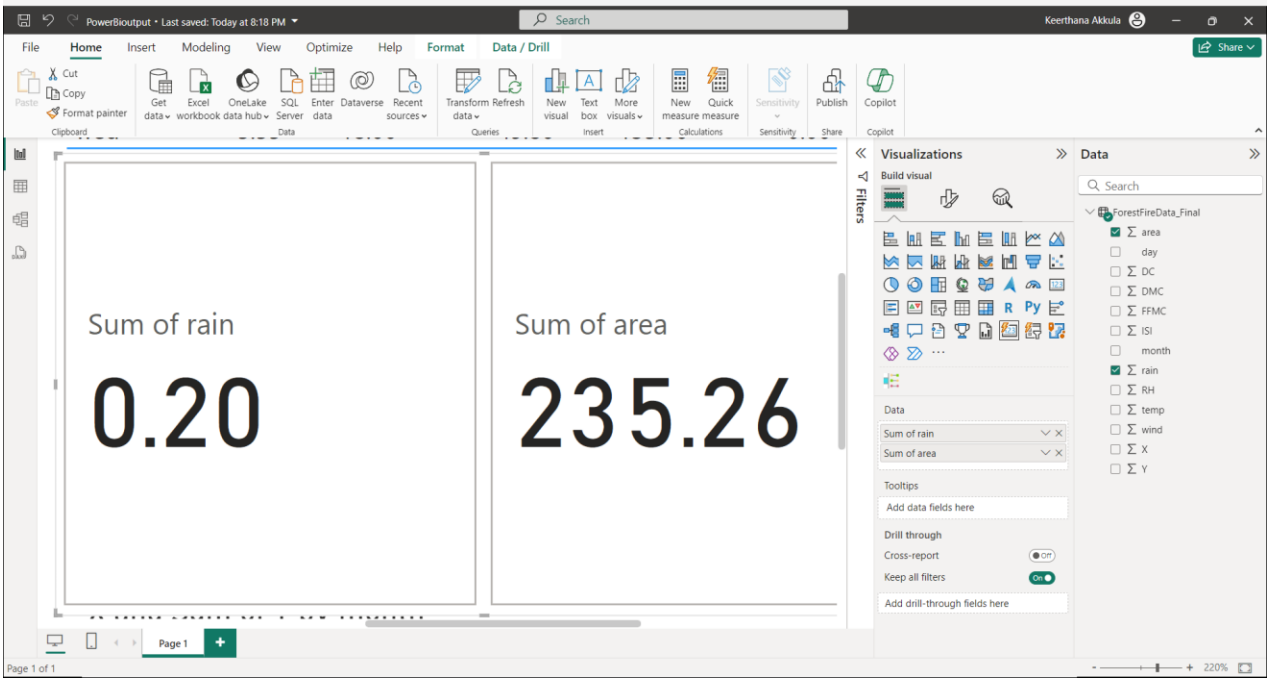

Navigator
Display Options
rakswk.sql.azuresynapse.net: dedicatedsqlpool...
fire_data
fire_data_new
ForestFireData_Final
ForestFireData_Final
X Y month day FFMC DMC DC ISI
7 5 mar fri 86.2 26.2 94.3
8 6 mar fri 91.7 33.3 77.5
8 6 mar sun 89.3 51.3 102.2
5 5 mar sat 91.7 35.8 80.8
6 4 mar wed 89.2 27.9 70.8
4 4 mar tue 88.1 25.7 67.6
4 4 mar mon 87.2 23.9 64.7
4 4 mar mon 87.6 52.2 103.8
2 2 mar sun 89.3 51.3 102.2
2 2 mar sun 89.3 51.3 102.2
4 5 mar fri 91.7 33.3 77.5
4 5 mar fri 91.2 48.3 97.8
5 4 mar fri 91.7 33.3 77.5
1 3 mar mon 87.6 52.2 103.8
6 5 mar sat 91.7 35.8 80.8
8 6 mar fri 91.7 35.8 80.8
3 4 mar sat 69 2.4 15.5
4 5 mar fri 85.9 19.5 57.3
4 5 mar thu 91.4 30.7 74.3
4 4 mar fri 85.9 19.5 57.3
3 4 mar fri 91.7 33.3 77.5
3 4 mar tue 88.1 25.7 67.6
3 5 mar tue 88.1 25.7 67.6
Select Related Tables
Load
Transform Data
Cancel


Untitled - Power BI Desktop
Search
Keerthana Akkula
File Home Insert Modeling View Optimize Help
Share
Build visuals with your data
Select or drag fields from the Data pane onto the report canvas.
Visualizations
Build visual
Values
Add data fields here
Drill through
Cross-report
Keep all filters
Add drill-through fields here
Filters
Data
Search
ForestFireData_Final
area
day
DC
DMC
FFMC
ISI
month
rain
RH
temp
wind
X
Y
Page 1
Page 1 of 1
58%


PowerBioutput • Last saved: Today at 8:18 PM
Search
Keerthana Akkula
File
Home
Insert
Modeling
View
Optimize
Help
Format
Data / Drill
Share
Sum of area, Sum of DC, Sum of DMC and Sum of ISI by day
Sum of area
Sum of DC
Sum of DMC
Sum of ISI
Sum of area, Sum of DC, Sum ...
2,000
1,500
1,000
500
0
mon
sat
sun
fri
thu
wed
tue
day
Visualizations
Build visual
Data
Search
ForestFireData_Final
area
day
DC
DMC
FFMC
ISI
month
rain
RH
temp
wind
X
Y
X-axis
day
Y-axis
Sum of area
Sum of DC
Sum of DMC
Sum of ISI
Legend
Add data fields here
Filters
Page 1
Page 1 of 1
190%

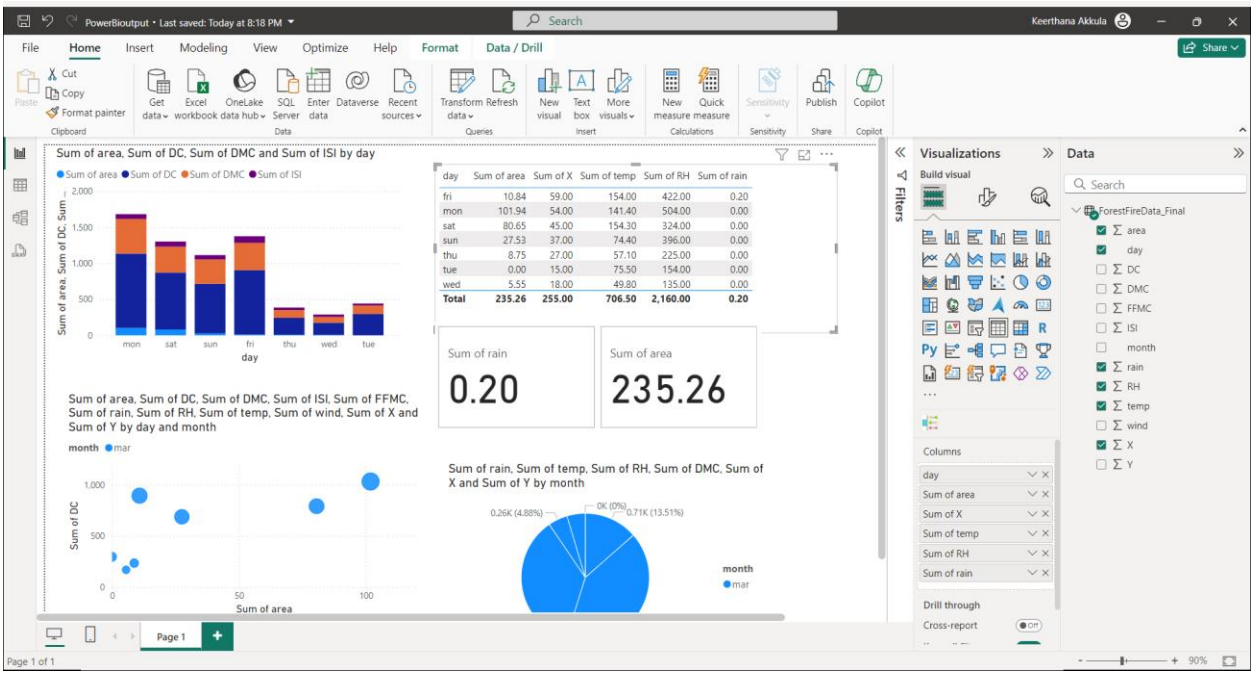

PowerBioutput • Last saved: Today at 8:18 PM
Search
Keerthana Akkula
File
Home
Insert
Modeling
View
Optimize
Help
Format
Data / Drill
Share
day
Sum of area
Sum of X
Sum of temp
Sum of RH
Sum of rain
fri 10.84 59.00 154.00 422.00 0.20
mon 101.94 54.00 141.40 504.00 0.00
sat 80.65 45.00 154.30 324.00 0.00
sun 27.53 37.00 74.40 396.00 0.00
thu 8.75 27.00 57.10 225.00 0.00
tue 0.00 15.00 75.50 154.00 0.00
wed 5.55 18.00 49.80 135.00 0.00
Total 235.26 255.00 706.50 2,160.00 0.20
Visualizations
Build visual
Data
Search
ForestFireData_Final
area
day
DC
DMC
FFMC
ISI
month
rain
RH
temp
wind
X
Y
Columns
day
Sum of area
Sum of X
Sum of temp
Sum of RH
Sum of rain
Drill through
Cross-report
Filters
Page 1
Page 1 of 1
220%

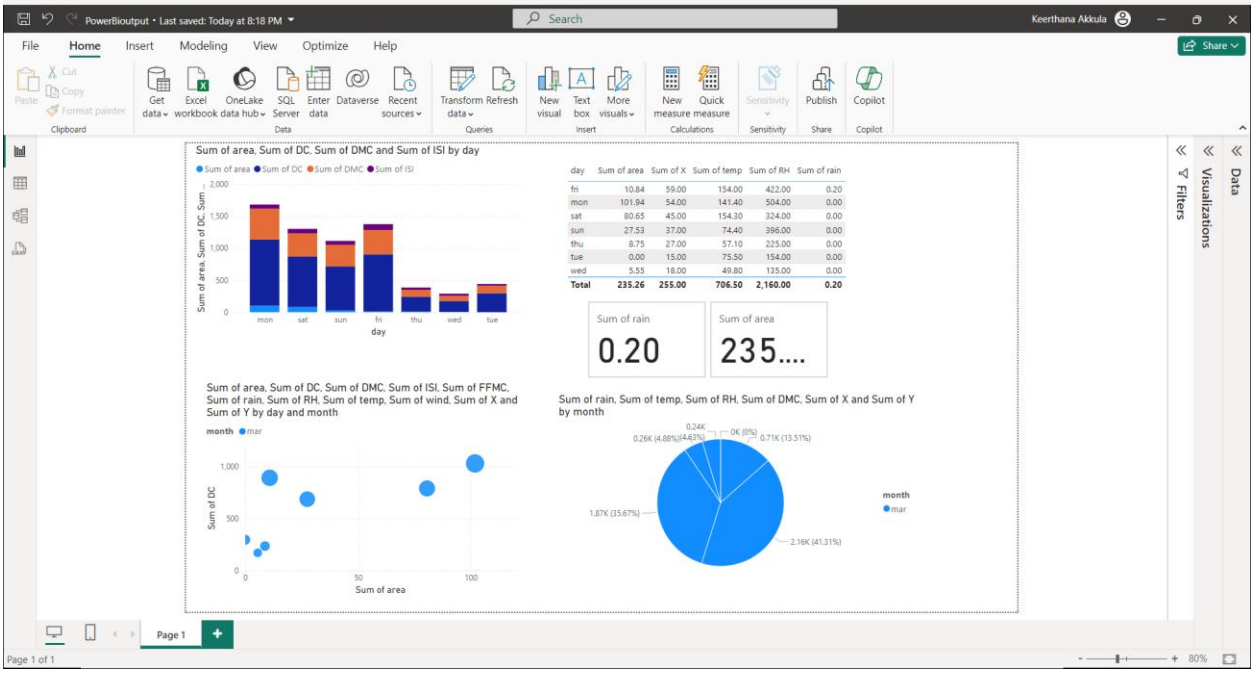




Create the visualization using the data

PowerBioutput.pbix





Conclusion:

- Document the advantages and outcomes of the data engineering process.
- Mention improvements in data quality and the benefits for the organization.

The Importance and Benefits of Data Engineering

1. Centralizing Data:
    - Data engineering helps gather and integrate data from different sources, creating a single view of the organization's operations.
    - This unified data allows key stakeholders to access the information they need easily and securely.
2. Ensuring Data Security:
    - Data engineers implement strong security measures to protect data throughout its lifecycle.
    - Proper handling and protection of data prevent unauthorized access and cyber attacks.
3. Better Decision-Making:
    - When data is accurate and well-organized, it helps analysts and executives make informed decisions.
    - Organizations can use these reliable insights to optimize processes, spot trends, and adapt to market changes.
4. Improving Data Quality:
    - Data engineering involves cleaning and transforming data to ensure its quality.
    - High-quality data leads to better analysis, more accurate predictions, and improved business results.
5. Organizational Benefits:
    1. Efficient data engineering has several advantages for organizations, including:
    2. Efficiency: Automated data pipelines save time and reduce manual work.
    3. Scalability: Well-designed systems can handle large amounts of data smoothly.
    4. Innovation: Reliable data supports advanced analytics, like AI and machine learning.
    5. Cost Savings: Effective data management reduces operational costs.
    6. Competitive Edge: Organizations that use data well gain a competitive advantage.